Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 10/11/2023

Par : **Franklin FEUKAM KOUHOUE**

Titre : **Interprétabilité des Modèles de Tarification en Actuariat: application à l'assurance automobile.**

Confidentialité : ☑ NON ☐ OUI (Durée : ☐ 1 an ☐ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de la filière

Nom :

Membres présents du jury de l'Institut des Actuaires

Entreprise : Institut Europlace de Finance

Signature :

Directeur du mémoire en entreprise :

Nom : Arthur Charpentier
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Mourad KOLLI

Signature du candidat

Franklin FEUKAM KOUHOUE

Secrétariat:

Bibliothèque:

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Avec l'avènement de nouvelles technologies couplé aux attentes sans cesse grandissantes des assurés −volonté d'hyperpersonnalisation de la relation client (assuré-assureur)−, la transformation numérique du secteur de l'assurance a beaucoup évolué au cours de ces deux dernières décennies et s'est vue fortement accélérée durant la récente crise sanitaire. Cette révolution digitale permet une collecte massive de données, offrant ainsi aux assureurs, la possibilité d'une segmentation beaucoup plus fine pour repérer les risques émergents de leurs portefeuilles et mieux ajuster leurs stratégies de gestion du risque.

Avec cet essor de données massives, se sont simultanément développés de nouveaux algorithmes puissants d'apprentissage statistique permettant d'obtenir des modèles prédictifs beaucoup plus précis que par le passé : notamment les réseaux de neurones et les modèles par agrégation, etc. Cependant, bien qu'on leur reconnaisse un meilleur pouvoir prédictif, l'intégration de ces nouveaux modèles de tarification dans la modélisation actuarielle reste modérée dans le secteur français de l'assurance. Trois principales raisons justifient ce fait : en premier lieu, le manque d'interprétabilité lié au caractère "boîte-noire" de ces modèles −ce qui s'oppose à l'exigence réglementaire de transparence du processus de tarification exigée par l'ACPR− ; en deuxième lieu, la complexité liée à la mise en œuvre de ces nouveaux modèles ; en dernier lieu, les problèmes d'éthique et de gouvernance liés à la non transparence de ces modèles. Pour lever le voile sur ces modèles et les rendre plus interprétables, depuis quelques années on assiste à une prolifération d'articles de recherche sur le sujet de l'intelligence artificielle interprétable (XAI).

Dans ce contexte, ce mémoire vise à explorer et à illustrer l'usage des différentes méthodes d'interprétabilité les plus répandues dans la littérature indispensables à l'interprétabilité des modèles opaques d'apprentissage statistique. Nous étudions également les enjeux des données télématiques dans l'amélioration de la précision des modèles tarification en assurance automobile.

Mots clefs : GLM, LocalGLMnet, Random Forest, Model Reliance (MR), SFIMP, PDP, ALE, ICE, H-statistique, Indices de Sobol, SHAP, LIME, RGPD, ACPR, XAI, Tarification automobile, Données télématiques de conduite.

Abstract

With the advent of new technologies coupled with the ever-increasing expectations of policyholders −a desire for hyper-personalization of the customer relationship (policyholder-insurer)− the digital transformation of the insurance sector has evolved significantly over the last two decades and was greatly accelerated during the recent health crisis. This digital revolution allows for massive data collection, thus offering insurers the possibility of a much finer segmentation to identify emerging risks in their portfolios and better adjust their risk management strategies.

With the rise of massive data, new powerful and self-learning algorithms have simultaneously been developed to obtain much more accurate predictive models than in the past: notably neural networks and aggregation models, etc. However, although they are recognized as having better predictive power, the integration of these new so-called machine learning models into actuarial claims modeling remains moderate in the French insurance industry. There are three main reasons for this: firstly, the lack of interpretability linked to the "black box" nature of these models −which is contrary to the regulatory requirement of transparency in the pricing process required by the ACPR, or the RGPD−; secondly, the complexity of implementing and explaining these new models −the acculturation of stakeholders still takes time− and finally, the ethical and governance problems linked to the lack of transparency of these models. To lift the veil on these black boxes and make them more interpretable, in recent years there has been a proliferation of research articles on the subject of interpretable artificial intelligence (XAI).

In this context, this thesis aims to explore and illustrate the use of the various interpretability methods most widely used in the literature, which are essential for the interpretability of opaque statistical learning models. We also study the role of telematic data in improving the accuracy of automobile insurance pricing models.

Keywords: GLM, LocalGLMnet, Random Forest, Model Reliance (MR), SFIMP, PDP, ALE, ICE, H-statistics, Sobol indices, SHAP, LIME, RGPD, ACPR, XAI, Automobile pricing, Telematics driving data.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire.

En premier lieu, j'aimerais remercier l'Institut Louis Bachelier et l'Université du Québec à Montréal de m'avoir offert l'opportunité de réaliser ce stage de fin d'étude.

Je remercie mes tuteurs Laurence Barry et Arthur Charpentier d'avoir proposé ce sujet de mémoire très passionnant et également pour la qualité de leur encadrement tout au long de la réalisation de ce mémoire. Les nombreux points et échanges m'ont été d'une aide immense.

Je voudrais remercier mon tuteur pédagogique, Nicolas Baradel, pour sa grande disponibilité et ses nombreux conseils.

Enfin, j'adresse mes sincères remerciements à tous les enseignants que j'ai eu le privilège de rencontrer durant mon parcours académique à l'ENSAE Paris et à l'ENSAE Dakar.

À Victorine et Flaubert

Note de synthèse

À l'ère du big data et de la puissance de calcul croissante, les algorithmes d'apprentissage statistique, tels que les forêts aléatoires (*"Random Forest"*), le *"Xgboost"* et les réseaux de neurones, gagnent en popularité pour la modélisation prédictive, notamment dans le domaine de l'assurance. Malgré leurs performances exceptionnelles, l'usage de ces algorithmes opaques d'apprentissage statistique en tarification reste limité dans le secteur assurantiel français. La raison principale est qu'ils sont peu interprétables. Pour résoudre ce problème, de nombreuses méthodes d'interprétabilité de modèles opaques d'apprentissage statistique ont été développées au cours de ces dernières années.

À travers le prisme de la littérature, ce mémoire explore les méthodes d'interprétabilité des modèles opaques d'apprentissage statistique et illustre l'utilisation de ces méthodes d'interprétation sur un cas de tarification automobile.

Nous utilisons la base de données publique pour la télématique des conducteurs en assurance automobile disponible sur le site web [lien] du département de mathématique de l'Université du Connecticut, aux Etats-Unis. L'ensemble de données contient 100 000 polices d'assurance pour lesquelles sont renseignées les deux informations relatives à l'expérience de sinistres du conducteur, à savoir: la fréquence et la sévérité de ses sinistres; de onze variables de risque classiques telles que la durée de la police, l'âge et le sexe du conducteur; et de trente neuf variables liées à la télématique du conducteur, incluant par exemple le nombre total de kilomètres parcourus durant la période couverture, le nombre de freinages soudains ou d'accélérations soudaines au cours de la période de couverture.

Pour ajuster la fréquence et le coût de sinistre, nous avons implémenté trois types de modèles. Un modèle intrinsèquement interprétable: le GLM; un modèle opaque: le *Random Forest*; et un modèle hybride: le LocalGLMnet (paru tout récemment en mai 2022).

Une fois ces modèles mis place, nous avons comparé leur performance prédictive sur notre base de test, à l'aide de métriques usuelles telles que la MSE, la MAE, la RMSE et la RMSE_mean. Les résultats obtenus sont présentés dans le tableau 1 pour la fréquence de sinistre et dans le tableau 2 pour la sévérité. Nous avons retenu comme meilleur modèle celui ayant la plus petite RMSE.

Pour la fréquence de sinistre, dans le tableau 1, on constate que l'ensemble des modèles mis en place ont des performances prédictives voisines. Le modèle LocalGLMnet avec loi de Poisson se démarque de très peu des deux autres modèles avec une RMSE de 0.2407691, soit un gain relatif de +0.64% par rapport au modèle GLM-Poisson et de 0.08% par rapport au *Random Forest*.

Pour l'ensemble des trois modèles de fréquence implémentés, le ratio de la fréquence moyenne prédite sur la fréquence moyenne observée dans de la base test est inférieure à 1. Cela signifie que dans la globalité, les trois modèles de fréquence implémentés sous-estiment la fréquence effective de sinistres des assurés. Le *Random Forest* avec un ratio de 0.97, est celui qui permet le mieux de se rapprocher de la fréquence moyenne effective observée au niveau de la base test.

Modèles	moy. pred./moy. réelle	MSE	MAE	RMSE	$RMSE_{mean}$
GLM-Poisson	0.8794369	0.05871528	0.08529553	0.2423124	4.959702
LocalGLMnet-Poisson	0.9115864	0.05796977	0.08565878	0.2407691	4.928114
Random Forest	0.9767824	0.05806044	0.08783279	0.2409573	4.931966

Table 1: *Comparaison de la performance prédictive des modèles de fréquence.*

En ce qui concerne les modèles de sévérité de sinistre, dès la première lecture du tableau 2, on fait un constat frappant: le modèle Random Forest est à la fois le plus prudent, avec un ratio coût moyen prédit sur coût moyen observé égal à 1.045721, et le plus précis, avec une RMSE de 1129.774, soit un gain relatif de RMSE de +0.24% par rapport au LocalGLMnet et au GLM.

Modèles	moy. pred./moy. réelle	MSE	MAE	RMSE	$RMSE_{mean}$
GLM-Tweedie	0.8177954	1285130	223.1136	1133.636	8.532343
LocalGLMnet-Tweedie	0.8809933	1282644	229.391	1132.539	8.524087
Random Forest	1.045721	1276388	248.962	1129.774	8.503275

Table 2: *Comparaison de la performance prédictive des modèles de sévérité (agrégée) mis en place.*

Par la suite, nous avons interprété les modèles de fréquence mis en place. Pour ne pas rallonger le mémoire, nous avons juste présenté les résultats de l'interprétation des modèles GLM et LocalGLMnet. Ils ont été interprétés en suivant deux approches: premièrement une approche d'interprétation basée sur le modèle; deuxièmement en adoptant une approche d'interprétation *post hoc*.

Pour l'interprétation du modèle LocalGLMnet, en ce qui concerne l'interprétation basée sur le modèle, nous avons étudié la contribution de chaque caractéristique dans le modèle, à l'aide de graphiques appropriés (confère figure 5.38). Nous avons également étudié les interactions entre les caractéristiques dans le modèle (confère figure 5.37) et évaluer l'importance des caractéristiques dans le modèle par une approche spécifique aux modèles LocalGLMnet (confère figure C.18).

Les méthodes d'interprétabilité *post hoc* implémentées pour l'interprétation des différents modèles de fréquence dans ce mémoire sont récapitulées dans le tableau 3 ci-dessous. Elles sont toutes agnostiques au modèle, donc applicables à toute sorte de modèle d'apprentissage statistique supervisé.

Méthodes	**Utilité**	**Package R**	**Référence**
Globales			
Partial Dependence Plot (PDP)	Effet marginal	iml & DALEX	Friedman *et al.* (2001)
Accumulate Locale Effect (ALE)	Effet marginal	iml & DALEX	Apley et Zhu (2020)
Model Reliance (MR)	Importance var.	vip & DALEX	Fisher *et al.* (2019)
Shapley Feature Importance (SFIMP)	Importance var.	iml	Casalicchio *et al.* (2019)
Indice de Sobol	Interactions var.	sensitivity	Sobol' (1990)
H-statistique de Friedman	Interactions var.	iml & DALEX	Friedman et Popescu (2008)
Locales			
LIME	Contribution var.	iml & DALEX	Ribeiro *et al.* (2016)
SHAP	Contribution var.	iml & DALEX	Lundberg et Lee (2017)
ICE-Plot	Effet marginal local	iml & DALEX	Goldstein *et al.* (2015)

Table 3: *Récapitulatif des méthodes d'interprétabilité implémenter dans ce mémoire dans le cas d'application à l'assurance automobile.*

Après avoir interprété les modèles suivant l'approche basée sur le modèle et l'approche *post hoc*, nous avons globalement constaté une cohérence entre les interprétations issues des deux approches, que ce soit pour le modèle GLM ou le modèle LocalGLMnet. De plus, le modèle LocalGLMnet permet d'étudier plus minutieusement les interactions entre les caractéristiques dans l'explication de la fréquence de sinistre.

Une fois les interprétations du modèle LocalGLMnet-Poisson faites, nous nous sommes servis des résultats de l'interprétation pour augmenter manuellement les relations d'interaction entre

les caractéristiques dans notre modèle GLM-Poisson de départ. L'objectif était d'améliorer la performance prédictive du modèle GLM-Poisson de départ.

Une fois notre modèle GLM-Poisson avec interactions ajoutées ajusté, nous nous en sommes servi pour réaliser des prédictions de fréquence de sinistre pour les assurés de notre base test mise de côté dès le départ.

Nous avons ensuite comparé la qualité des prédictions obtenues par le modèle GLM-Poisson initial (sans interactions) et le nouveau modèle GLM-Poisson avec interactions augmentées. Les résultats des métriques calculées sont disponibles dans le tableau 4.

On observe très clairement que le gain apporté par la prise en compte des interactions dans le modèle GLM-Poisson est infime, voir inexistant dans notre contexte.

Modèles fréquence	Métriques				Gain relatif (en %)		
	$\frac{\text{moy. pred.}}{\text{moy. réelle}}$	MSE	MAE	RMSE	MSE	MAE	RMSE
GLM initial (benchmark)	0.8794369	0.058715	0.085295	0.24231	–	–	–
GLM complexifié	0.9650467	0.058534	0.088793	0.24193	0.31	−4.10	0.15

Table 4: *Comparaison de la performance prédictive du modèle GLM fréquence initial et du modèle GLM fréquence complexifié (après l'ajout des effets non linéaires et d'interaction).*

Le second objectif de ce mémoire était d'étudier la *plus-value* des données télématiques dans la précision des modèles de tarification en assurance automobile.

Dans le tableau 5 ci-dessous, on peut facilement comparer les performances prédictives des modèles de fréquence et de sévérité lorsqu'on prend en considération les données télématiques dans la modélisation de quand on ne les prend pas en considération.

On constate bien une nette amélioration des métriques de performance prédictive lorsqu'on prend en considération les données télématiques dans la modélisation de la fréquence et du coût de sinistre.

Modèles	Métriques			Gain relatif (en %)		
	MSE	MAE	RMSE	MSE	MAE	RMSE
GLM-freq-classique (Benchmark)	0.0597765	0.09298055	0.2444923	–	–	–
GLM-freq-complet	0.05687827	0.08876985	0.2384917	4.85	4.53	2.45
RF-freq-classique	0.0593124	0.09186272	0.2435414	0.78	1.20	0.40
RF-freq-complet	0.05128279	0.08603667	0.226457	14.21	7.50	7.40
GLM-coût-classique (Benchmark)	1 301 920	250.6309	1 141.017	–	–	–
GLM-coût-complet	1 256 309	237.5303	1 120.852	3.50	5.30	1.76

Légende:
GLM-freq-classique: glm-poisson pour la fréquence, avec variables classiques uniquement (Benchmark fréq.);
GLM-freq-complet: glm-poisson pour la fréquence de sinistres, avec variables classiques et télématiques;
RF-freq-classique: random forest pour la fréquence de sinistres, avec variables classiques uniquement;
RF-freq-complet: random forest pour la fréquence de sinistres, avec variables classiques et télématiques;
GLM-coût-classique: glm-Tweedie sévérité agrégée, avec variables classiques uniquement (Benchmark sév.);
GLM-coût-complet: glm-Tweedie pour la sévérité agrégée de sinistres, avec variables classiques et télématiques.

Table 5: *Analyse de la performance des modèles avec et sans variables télématiques; au milieu du tableau sont représentés les indicateurs de performance sur la base de test; à droite sont représentés les gains relatifs par rapport au modèle GLM sans variables télématiques (appelé GLM classique).*

Nous avons montré que les variables télématiques permettent de tarifer au plus juste les contrats d'assurance automobile. Cependant, en pratique, les assureurs n'ont pas toujours le luxe de disposer de données télématiques sur leur clientèle. Ceci pour diverses raisons que nous ne détaillons pas dans ce mémoire.

Notre alternative a ensuite été de prédire les valeurs des variables télématiques à partir des caractéristiques observables chez l'assuré à la souscription du contrat, notamment : l'âge de l'assuré, l'âge de son véhicule, l'usage fait de son véhicule, son score de crédit, sa durée de couverture, le nombre annuel de miles prévus à parcourir déclarés par l'assuré à la souscription.

Une fois les différentes variables télématiques prédites, nous les avons utilisées comme prédicteurs dans nos différents modèles de sinistralité.

Soulignons que les modèles de sinistre utilisant ces données télématiques prédites sont plus performants que les modèles de sinistres lorsqu'on fait fi de ces variables télématiques prédites, et se contente d'ajuster la fréquence ou le coût de sinistre juste sur les variables de risque classiques. Cependant, la différence de performance est infime.

En définitive, dans le cadre de notre cas d'application en assurance automobile, nous avons identifié trois (03) éléments marquants:

Dans un premier temps, l'incorporation de variables télématiques améliore la précision prédictive et descriptive des modèles de tarification automobile. Dans notre étude, lorsqu'un modèle de fréquence classique de type GLM inclue à la fois les variables de risque traditionnelles ainsi que les variables télématiques, une amélioration relative d'environ 5% est observée pour la MSE et la MAE, et d'environ 2,5 % pour la RMSE. En outre, pour la modélisation de la fréquence de sinistre, lorsqu'on prend en compte les données télématiques, et qu'en plus, on utilise un modèle de fréquence opaque de type *Random Forest* capable de mieux extraire les informations contenues dans ces données télématiques, on observe un gain relatif de précision encore plus significatif: 14.21% sur la MSE, de 7.50% sur la MAE et de 7.40% sur la RMSE, par rapport au modèle GLM-fréquence n'utilisant pas de données télématiques. Cette amélioration est plus marquée lorsque l'on se focalise sur des segments spécifiques d'assurés. Par exemple, l'utilisation des variables télématiques réduit la MSE de la fréquence prédite de sinistres de plus 30% pour le segment des assurés de moins de 21 ans.

En deuxième lieu, nous avons relevé le potentiel des modèles de type LocalGLMnet comme alternative pertinente aux modèles classiques de type GLM et aux modèles d'apprentissage statistique opaques tels que les *Random Forest* en tarification automobile. Le LocalGLMnet par le biais des graphiques d'interaction permet d'étudier minutieusement les interactions entre les caractéristiques dans l'explication de la fréquence de sinistre des assurés. Dans notre étude, en plus de concurrencer le modèle *Random Forest* en termes de performance prédictive (au sens de la RMSE), les modèles LocalGLMnet se démarquent par leur caractère mi-transparent.

Enfin, notre étude met en évidence que les interprétations issues d'un modèle opaque peuvent être exploitées pour optimiser les performances du modèle GLM initial. Cependant, dans notre contexte, l'amélioration des performances de notre modèle GLM de départ était minime (gain relatif de RMSE de 0,15 %).

Cependant, bien que les résultats obtenus après l'intégration des effets d'interaction dans le modèle puissent sembler peu satisfaisants en termes de gain de précision prédictive, la méthodologie que nous avons adoptée reste adaptable à de nouvelles bases de données ou à de nouvelles expériences où les effets d'interaction jouent un rôle plus significatif sur la variable cible.

Executive summary

In the age of Big Data and increasing computing power, statistical learning algorithms such as Random Forest, Xgboost and neural networks are gaining in popularity for predictive modelling, particularly in the insurance field. for predictive modeling, particularly in the insurance sector. Despite their unprecedented performance, the use of these complex statistical learning algorithms in underwriting remains limited in the French insurance sector. The main reason is that they are difficult to interpret. To solve this problem, a number of methods for the interpretability of complex have been developed in recent years.

Through the prism of the literature, this dissertation explores methods for the interpretability of opaque statistical learning models, and illustrates the use of these interpretation methods on an automobile underwriting case.

We use the public database for driver telematics in auto insurance available on the website [link] of the Department of Mathematics at the University of Connecticut, USA. The dataset contains 100,000 insurance policies for which both information on the driver's claims experience (frequency and severity of claims), and eleven classical risk variables, such as policy duration, driver age and gender, are entered; and thirty-nine variables related to the driver's telematics, including, for example, the total number of kilometers driven during the coverage period, the number of sudden braking or acceleration events during the coverage period.

To adjust claims frequency and cost, we implemented three types of model. A natively interpretable model: the GLM; a complex model: the Random Forest; and a hybrid model: LocalGLMnet (recently released in May 2022).

Once these models had been implemented, we compared their predictive performance on our test base, using standard metrics such as MSE, MAE, RMSE and RMSE_mean. The results obtained are presented in Table 6 for loss frequency and in Table 7 for severity. The model with the lowest RMSE was selected as the best model.

For loss frequency, Table 6 shows that all the models implemented have similar predictive performances. The LocalGLMnet model with Poisson distribution distinguishes itself from the other two models, with an RMSE of 0.2407691, i.e. a relative gain of +0.64% over the GLM-Poisson model and 0.08% over the Random Forest.

For all three frequency models implemented, the ratio of predicted mean frequency to observed mean frequency at the test base is less than one. This means that, overall the three frequency models implemented underestimate the actual frequency of policyholder claims. Random Forest, with a ratio of 0.97, is the one that best approximates to the actual average frequency observed in the test base.

Models	moy. pred./moy. réelle	MSE	MAE	RMSE	$RMSE_{mean}$
GLM-Poisson	0.8794369	0.05871528	0.08529553	0.2423124	4.959702
LocalGLMnet-Poisson	0.9115864	0.05796977	0.08565878	0.2407691	4.928114
Random Forest	0.9767824	0.05806044	0.08783279	0.2409573	4.931966

Table 6: *Comparison of the predictive performance of frequency models.*

With regard to claims severity models, the first reading of Table 7, a striking observation: the Random Forest model is both the most conservative, with a ratio predicted average cost to

observed average cost equal to 1.045721 and the most accurate, with an RMSE of 1129.774, a relative gain in RMSE of +0.24% compared with LocalGLMnet and GLM.

Models	moy. pred./moy. réelle	MSE	MAE	RMSE	$RMSE_{mean}$
GLM-Tweedie	0.8177954	1285130	223.1136	1133.636	8.532343
LocalGLMnet-Tweedie	0.8809933	1282644	229.391	1132.539	8.524087
Random Forest	1.045721	1276388	248.962	1129.774	8.503275

Table 7: *Comparison of the predictive performance of the (aggregated) severity models implemented.*

We then went on to interpret the frequency models we had set up. In order not to slow down the the dissertation, we've just presented the results of the interpretation of the GLM and LocalGLMnet models. They have been interpreted following two approaches: firstly, a model-based interpretation approach; secondly, by adopting a post hoc post hoc.

For the interpretation of the LocalGLMnet model, with regard to the model-based interpretation model, we have studied the contribution of each feature in the model, using appropriate graphs (see figure 5.38). We have also studied the interactions between features in the model (see figure 5.37) and evaluate the importance of features in the model using an approach specific to LocalGLMnet models (see figure C.18).

The post hoc interpretability methods implemented for the interpretation of the various mod- frequency models are summarized in Table 8 below. They are all model-agnostic, so they can be applied to any kind of statistical learning model su- pervised.

Methods	**Utility**	**R Package**	**Reference**
Globals			
Partial Dependence Plot (PDP)	Marginal effect	iml & DALEX	Friedman *et al.* (2001)
Accumulate Locale Effect (ALE)	Marginal effect	iml & DALEX	Apley et Zhu (2020)
Model Reliance (MR)	Importance var.	vip & DALEX	Fisher *et al.* (2019)
Shapley Feature Importance (SFIMP)	Importance var.	iml	Casalicchio *et al.* (2019)
Indice de Sobol	Interactions var.	sensitivity	Sobol' (1990)
H-statistique de Friedman	Interactions var.	iml & DALEX	Friedman et Popescu (2008)
Locals			
LIME	Contribution var.	iml & DALEX	Ribeiro *et al.* (2016)
SHAP	Contribution var.	iml & DALEX	Lundberg et Lee (2017)
ICE-Plot	Local marginal effect	iml & DALEX	Goldstein *et al.* (2015)

Table 8: *Summary of the interpretability methods implemented in this dissertation for the automobile insurance application.*

Having interpreted the models according to the model-based approach and the post hoc approach, we found overall consistency between the interpretations derived from the two approaches, for both the GLM and LocalGLMnet models. What's more, the LocalGLMnet allows us to study in greater detail the interactions between characteristics in explaining of loss frequency.

Once the interpretations of the LocalGLMnet-Poisson model had been made, we used interpretation results to manually augment the interaction relationships between features features in our original GLM-Poisson model. The aim was to improve the predictive performance of the original GLM-Poisson model.

Once we had fitted our GLM-Poisson model with added interactions, we used it to make claims frequency predictions for the policyholders in our test base, set aside at the outset. from the outset.

We then compared the quality of the predictions obtained by the initial GLM-Poisson model model (without interactions) and the new GLM-Poisson model with augmented interactions. The results are shown in Table 9.

We can clearly see that the benefit of taking interactions into account in the GLM-Pisson model is minimal, if not non-existent in our context. GLM-Poisson model is minimal, or even non-existent in our context.

Frequency models	Metrics				Relative gain (en %)		
	$\frac{\text{moy. pred.}}{\text{moy. réelle}}$	MSE	MAE	RMSE	MSE	MAE	RMSE
initial GLM (benchmark)	0.8794369	0.058715	0.085295	0.24231	–	–	–
complexified GLM	0.9650467	0.058534	0.088793	0.24193	0.31	−4.10	0.15

Table 9: *Comparison of the predictive performance of the initial GLM frequency model and the complexified GLM frequency model (after adding non-linear and interaction effects).*

The second objective of this thesis was to study the added value of telematics data in the accuracy of automobile insurance pricing models.

In Table 10 below, the predictive performance of the frequency and severity models frequency and severity models when telematics data is taken into account in the modelling data into the model than when they are not.

There is a clear improvement in predictive performance metrics when telematics data are telematics data is taken into account when modeling claims frequency and cost loss frequency.

Models	Metrics			Relative Gain (en %)		
	MSE	MAE	RMSE	MSE	MAE	RMSE
GLM-freq-classique (Benchmark)	0.0597765	0.09298055	0.2444923	–	–	–
GLM-freq-complet	0.05687827	0.08876985	0.2384917	4.85	4.53	2.45
RF-freq-classique	0.0593124	0.09186272	0.2435414	0.78	1.20	0.40
RF-freq-complet	0.05128279	0.08603667	0.226457	14.21	7.50	7.40
GLM-coût-classique (Benchmark)	1 301 920	250.6309	1 141.017	–	–	–
GLM-coût-complet	1 256 309	237.5303	1 120.852	3.50	5.30	1.76

Legend:
GLM-freq-classique: glm-poisson for frequency, with classical variables only (Benchmark fréq.);
GLM-freq-complet: glm-poisson for claims frequency, with classic and telematic variables;
RF-freq-classique: random forest for claims frequency, with classic variables only;
RF-freq-complet: random forest for claims frequency, with classic and telematics variables;
GLM-coût-classique: glm-Tweedie aggregated severity, with classic variables only (Benchmark sév.);
GLM-coût-complet: glm-Tweedie for aggregate claims severity, with classic and telematics variables.

Table 10: *Performance analysis of models with and without telematics variables; in the middle of the table are the performance indicators on the test basis; on the right are the relative gains compared to the GLM model without telematics variables (called "GLM classique").*

We've shown that telematic variables can be used to price motor insurance contracts as accurately as possible. In practice, however, insurers do not always have the luxury of telematics data

on their customers. There are a number of reasons for this, which we won't go into detailed in this report.

Our alternative was then to predict the values of the telematics variables on the basis of the policyholder's observable characteristics at the time of taking out the contract, in particular: the policyholder's age, the age of his vehicle, the use made of his vehicle, his credit score, the duration of his coverage, the expected annual number of miles to be covered as declared by the policyholder at the time of taking out the contract, etc. The values of the telematics variables were then predicted on the basis of the policyholder's observable characteristics at the time of taking out the contract.

Once the various telematics variables had been predicted, we used them as predictors in our various claims models.

It's worth noting that claims models using these predicted telematics data perform better than claims models that ignore these predicted telematics variables, and simply adjust claims frequency or cost on conventional risk variables. However, the difference in performance is minute.

Finally, in the context of our automotive insurance case study, we have identified three (03) salient features:

Firstly, the incorporation of telematics variables improves the predictive and descriptive accuracy of automobile pricing models. In our study, when a classical GLM-type frequency model is extended by including traditional risk variables as well as telematics variables, a relative improvement of around 5% is observed on MSE and MAE, and of around 2.5% with regard to RMSE. In addition, for claims frequency modelling, when telematics data is taken into account, and in addition to this, using a complex Random Forest frequency model capable of better extracting the information contained in these telematics data, we achieve an even more significant relative gain in accuracy 14.21% on the MSE, 7.50% on the MAE and 7.40% on the RMSE, compared to the GLM-frequency model not using telematic data. This improvement is even more marked when we focus on specific policyholder segments. For example, the use of telematics variables reduces the MSE of the predicted claims frequency by more than 30% on the segment of policyholders under 21 years of age.

Secondly, we have identified the potential of LocalGLMnet models as a relevant alternative to classical GLM models and complex statistical learning models such as Random Forest in car pricing. Using interaction graphs, LocalGLMnet enables us to study in detail the interactions between characteristics in explaining policyholders' claims frequency. In our study, in addition to competing with the Random Forest model in terms of predictive performance (in the sense of RMSE), LocalGLMnet models stand out for their semi-transparent nature.

Thirdly, our study shows that the interpretations derived from a complex model can be exploited to optimize the performance of the initial GLM model. However, in our context, the performance improvement of our original GLM model was minimal (relative RMSE gain of 0.15%).

However, although the results obtained following the integration of interaction effects into the model may seem unsatisfactory in terms of gains in predictive accuracy, the methodology we have adopted remains adaptable to new databases or new experiments where interaction effects play a more significant role on the target variable.

Table des matières

Introduction

A l'ère du *big data* et de la progression exponentielle de la puissance de calcul, les algorithmes complexes d'apprentissage statistique représentent des outils à fort potentiel pour la modélisation prédictive en industrie.

Les modèles de type *bagging* (Random Forest), de type *boosting* (Xgboost) et de réseaux de neurones sont désormais très prisés dans de nombreux domaines, notamment celui de l'assurance. La démocratisation des modèles récents d'apprentissage statistique est due à leur grande performance prédictive.

Cependant, dans le secteur assurentiel français, le recours aux algorithmes complexes d'apprentissage statistique comme modèle de tarification de la sinistralité reste très modéré. Trois contraintes majeures justifient ce constat :

• La *contrainte d'interprétabilité* : le caractère opaque de ces modèles se heurte frontalement aux exigences réglementaires de transparence du processus de tarification instituées par l'Autorité de contrôle prudentiel et de résolution (ACPR) au niveau de France et par le Règlement général sur la protection des données (RGPD) à l'échelle européenne.

• Les *contraintes opérationnelles* : la complexité de ces algorithmes nécessite une expertise technique spécifique, ce qui peut entraîner des difficultés de mise en œuvre au sein d'équipes actuarielles traditionnelles.

• Les *contraintes éthiques* : les algorithmes utilisées conduisent souvent à une hyperindividualisation du risque et replacent au centre le risque de discrimination à la souscription. Ce qui est contraire au principe de mutualisation.

Pour lever ces contraintes, au cours de ces dernières années, de nombreux articles scientifiques proposant des méthodes d'interprétabilité des algorithmes d'apprentissage statistique ont vus le jour.

Dans ce mémoire, à travers le prisme de la littérature, nous présenterons les méthodes d'interprétabilité les plus pertinentes et les plus utilisées à l'heure actuelle par les chercheurs et les praticiens.

Nous mettrons en place un cas d'application en tarification automobile à partir d'une base de données publiques. Nous implémenterons trois types de modèles pour ajuster la fréquence et le coût de sinistre : un modèle GLM, un modèle de Forêt aléatoire (*Random Forest*) et un modèle LocalGLMnet. L'objectif poursuivi dans ce cas d'application est triple :

• Premièrement, nous étudions la valeur ajoutée des données télématiques dans l'amélioration de la précision des modèles de tarification en assurance automobile.

• Deuxièmement, mettons en compétitions le modèle LocalGLMnet tout récent, paru en mai 2022, avec les deux autres modèles mis en oeuvre. En cas de bonnes performances du LocalGLMnet, la finalité serait d'encourager les actuaires à l'utiliser dans leurs processus de tarification. D'autant plus qu'il s'agit d'un modèle au moins *semi-interprétable*.

• Enfin, nous utilisons les outils d'interprétabilité développés dans ce mémoire pour lever le voile sur les modèles opaques mis en place. Nous nous servons des interprétations issues de ces modèles opaques (relations d'interaction détectées entre les caractéristiques, ou effets non-linéaires de certaines caractéristiques) pour enrichir le modèle GLM-fréquence de départ, dans l'optique d'améliorer sa performance prédictive.

Chapitre 1

Généralités sur l'apprentissage statistique

1.1 Apprentissage statistique

1.1.1 Approche générale de l'apprentissage supervisé

Dans cette sous-section nous nous appuyons sur les notes de cours d'*Apprentissage Statistique* de Dalalyan (2018).

Nous observons une base de données composée de n couples $Z_i = (X_i, Y_i)$ que nous supposons être des réalisations indépendantes d'une même loi $\mathbb{P}$ inconnue. On notera $Z_i = (X_i, Y_i) \sim \mathbb{P}$ indépendante et identiquement distribués. Les X_i appartiennent à un espace $\mathcal{X}$ et s'appellent les entrées ou les caractéristiques. Typiquement, $\mathcal{X} = \mathbb{R}^p$ ($p \in \mathbb{N}^*$). Les Y_i appartiennent à un espace $\mathcal{Y}$, et s'appellent les sorties. Typiquement, $\mathcal{Y}$ est fini ou $\mathcal{Y}$ est un sous-ensemble de $\mathbb{R}$. Le but de l'apprentissage supervisé est de prévoir la sortie Y associée à toute nouvelle entrée X, où il est sous-entendu que la paire (X, Y) est une nouvelle réalisation de la loi $\mathbb{P}$, cette réalisation étant indépendante des réalisations précédemment observées.

Une fonction de prédiction est une fonction (mesurable) de $\mathcal{X}$ dans $\mathcal{Y}$. Dans ce qui suit, nous supposons que toutes les quantités que nous manipulons sont mesurables. Notons $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ l'ensemble de toutes les fonctions de prédiction. La base de données $Z_1, ..., Z_n$ est appelée ensemble d'apprentissage. Un algorithme d'apprentissage est une fonction qui à tout ensemble d'apprentissage renvoie une fonction de prédiction, c'est-à-dire une fonction de l'union $\bigcup_{n=1}^{\infty} \mathcal{Z}^n$ dans $\mathcal{F}(\mathcal{X}, \mathcal{Y})$, avec $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$. C'est un estimateur de la "meilleure" fonction de prédiction, où le terme meilleure se définit par une fonction de perte.

Soit $l(y, y')$ la perte encourue lorsque la sortie observée est y et la valeur prédite y'. La fonction $l : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$ est appelée fonction de perte. Les deux exemples les plus fréquemment utilisés sont $l(y, y') = \mathbb{I}(y \neq y')$ pour les problèmes de classification et $l(y, y') = |y - y'|^q$, où $q \geq 1$ est un réel fixe, pour les problèmes de régression. Dans le second cas on parle de problème de régression $\mathbb{L}^q$ (lorsque $q = 2$, la tâche d'apprentissage s'appelle aussi régression aux moindres carrés ordinaire).

Ainsi, pour une fonction de prédiction $g : \mathcal{X} \longrightarrow \mathcal{Y}$ son risque ou erreur de généralisation est mesurée par :

$$\mathcal{R}_{\mathbb{P}}(g) = \mathbb{E}_{\mathbb{P}}[l(Y, g(X))] \tag{1.1}$$

La "meilleure" fonction de prédiction est une fonction de $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ minimisant le risque $\mathcal{R}_{\mathbb{P}}$; c'est-à-dire une fonction telle que :

$$g^*_{\mathbb{P}} \in \underset{g \in \mathcal{F}(\mathcal{X}, \mathcal{Y})}{\operatorname{argmin}} \mathcal{R}_{\mathbb{P}}(g) \tag{1.2}$$

Une telle fonction $g^*_{\mathbb{P}}$ n'existe pas nécessairement. Cependant, pour les fonctions de pertes usuelles, notamment celles que nous considérons par la suite, le problème d'optimisation 1.2 admet au moins une solution. Cette "meilleure" fonction est appelée fonction *oracle* ou *prédicteur de Bayes*. Elle dépend de la probabilité inconnue $\mathbb{P}$, et par conséquent, est inconnue, car $\mathbb{P}$ elle-même

est inconnue .

– En régression aux moindres carrés, lorsque X, $Y \in L^1$ (sont intégrables), une fonction oracle est :

$$\eta^*_{\mathbb{P}}(x) = \mathbb{E}_{\mathbb{P}}(Y|X = x) = \int_{\mathcal{Y}} y \, d\mathbb{P}_{Y|X}(y|x)$$

– En classification, les fonctions oracles sont les fonctions $g^*_{\mathbb{P}}$ satisfaisant :

$$g^*_{\mathbb{P}}(x) \in \underset{y \in \mathcal{Y}}{argmax}\, \mathbb{P}(Y = y|X = x), \text{ pour } x \in \mathcal{X}.$$

Lorsque $\mathcal{Y} = \{0, 1\}$, la fonction $x \longmapsto \mathbb{I}_{g^*_{\mathbb{P}}(x) > \frac{1}{2}}$ est une fonction oracle pour la classification binaire : on l'appelle le *classifieur de Bayes*.

1.1.2 Minimisation du risque empirique, compromis biais-variance et notion de sur-apprentissage

Rappelons tout d'abord que le risque d'une fonction de prédiction $g : \mathcal{X} \longrightarrow \mathcal{Y}$ est défini par $\mathcal{R}_{\mathbb{P}}(g) = \mathbb{E}_{\mathbb{P}}[l(Y, g(X))]$. Le but de tout algorithme d'apprentissage est de trouver une fonction de prédiction dont le risque est aussi faible que possible, autrement dit, aussi proche que possible du risque des prédicteurs oracles. Or, étant donné que la distribution réelle $\mathbb{P}$ de la variable aléatoire génératrice des observations est inconnue, la fonction de risque $\mathcal{R}_{\mathbb{P}}$ et les prédicteurs oracles sont donc inconnus. Néanmoins, le risque $\mathcal{R}_{\mathbb{P}}(g)$ est estimé par son équivalent empirique :

$$\hat{\mathcal{R}}_n(g) = \frac{1}{n}\sum_{i=1}^{n} l(Y_i, g(X_i))$$

En admettant que $\mathbb{E}_{\mathbb{P}}[l(Y, g(X))^2] < +\infty$ (ce qui est généralement le cas dans la pratique, car le plus souvent la variable aléatoire cible Y est à support fini ou a une distribution voisine de celle d'une gaussienne), alors par la loi forte des grands nombres et le théorème centrale limite on affirme que :

$$\hat{\mathcal{R}}_n(g) \underset{n \longrightarrow \infty}{\longrightarrow} \mathcal{R}_{\mathbb{P}}(g) \text{ (p.s.)}, \quad \sqrt{n}\{\hat{\mathcal{R}}_n(g) - \mathcal{R}_{\mathbb{P}}(g)\} \underset{n \longrightarrow \infty}{\longrightarrow} \mathcal{N}(0, \mathbb{V}ar[l(Y, g(X))]) \text{ (en } \mathcal{D})$$

Ainsi, de ces résultats de convergence, il en ressort que pour toute fonction de prédiction g, la quantité aléatoire $\hat{\mathcal{R}}_n(g)$ effectue des déviations en $O(\frac{1}{\sqrt{n}})$ autour de sa moyenne $\mathcal{R}_{\mathbb{P}}(g)$. Nous avions pour objectif de minimiser $\mathcal{R}_{\mathbb{P}}$. Cependant, puisque cette dernière quantité est inconnue mais "correctement" approchée par son correspondant empirique $\hat{\mathcal{R}}_n$ qui est connu, il va de soi de considérer les algorithmes d'apprentissage qui minimisent le risque empirique :

$$\hat{g}_{n,\mathcal{G}} = \underset{g \in \mathcal{G}}{argmin}\, \hat{\mathcal{R}}_n(g)$$

Avec $\mathcal{G}$ un sous-ensemble de $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ (par exemple, l'espace des hyperplans de $\mathbb{R}^p$ pour la régression *linéaire*). Dans la pratique, la recherche de la solution optimal $\hat{g}_{n,\mathcal{G}}$ s'obtient par les méthodes de *descente de gradient* ou de *descente de gradient stochastique* (présentée dans la sous-section 2.3, du chapitre 2).

Il n'est généralement pas intéressant de prendre $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$, tout entier parce que :

• D'une part, cela mène généralement au sur-apprentissage, dans la mesure où le risque de l'algorithme solution du programme d'optimisation peut être très inférieure au risque réel (même si la taille de l'ensemble d'apprentissage tend vers l'infini). En pratique, il faut prendre $\mathcal{G}$ suffisamment

grand pour pouvoir approcher une multitude de fonctions, mais pas aussi grand au point de mener au sur-apprentissage.

• D'autre part, cela pourrait parfois être très coûteux d'un point du vue informatique (avec des gigantesques temps de calcul).

En notant $g^*_{\mathbb{P},\mathcal{G}} = \underset{g\in\mathcal{G}}{argmin}\,\mathcal{R}_{\mathbb{P}}(g)$ une fonction minimisant le risque théorique sur $\mathcal{G}$, on a le résultat naturel suivant :

$$\mathcal{R}_{\mathbb{P}}(\hat{g}_{n,\mathcal{G}}) \geq \mathcal{R}_{\mathbb{P}}(g^*_{\mathbb{P},\mathcal{G}}) \geq \mathcal{R}_{\mathbb{P}}(g^*_{\mathbb{P}})$$

Plus précisément, l'excès de risque de $\hat{g}_{n,\mathcal{G}}$ se décompose en deux termes positifs, appelés erreur stochastique (liée à l'estimation) et erreur systématique ou biais (lié à la restriction du choix de la fonction de prédiction dans le sous ensemble $\mathcal{G}$, plutôt que dans $\mathcal{F}(\mathcal{X},\mathcal{Y})$ tout entier) :

$$\mathcal{R}_{\mathbb{P}}(\hat{g}_{n,\mathcal{G}}) - \mathcal{R}_{\mathbb{P}}(g^*_{\mathbb{P}}) = \underbrace{\mathcal{R}_{\mathbb{P}}(\hat{g}_{n,\mathcal{G}}) - \mathcal{R}_{\mathbb{P}}(g^*_{\mathbb{P},\mathcal{G}})}_{\text{erreur stochastique}} + \underbrace{\mathcal{R}_{\mathbb{P}}(g^*_{\mathbb{P},\mathcal{G}}) - \mathcal{R}_{\mathbb{P}}(g^*_{\mathbb{P}})}_{\text{erreur systématique}}$$

Plus $\mathcal{G}$ est grand, plus le biais est faible, mais plus l'erreur stochastique est en général grand et inversement. Il y a donc un compromis à trouver dans le choix de $\mathcal{G}$. Ce compromis porte le nom de dilemme "biais-variance", où le terme variance provient du terme de l'erreur stochastique compte tenu de la variabilité de l'échantillon d'apprentissage que nous supposons dans notre formalisme être réalisation de variables indépendantes et identiquement distribuées. Il indique dans quelle mesure notre classificateur change si nous l'entraînons sur un ensemble d'apprentissage différent. Ainsi, une grande variance est évocatrice de sur-apprentissage. Le biais quant à lui est inhérent à la classe de modèle $\mathcal{G}$ choisie.

Sur la figure 1.1, sur l'image de gauche, nous avons une visualisation graphique du biais et de la variance à l'aide d'un diagramme en œil de boeuf. Nous supposons que le centre de la cible (en rouge) soit un modèle qui prédit parfaitement les valeurs correctes. Au fur et à mesure que nous nous éloignons du centre de la cible, nos prédictions deviennent de moins en moins bonnes. L'on observe clairement que le meilleur modèle est celui qui a à la fois un faible biais (faible erreur systématique), et une faible variance (faible erreur stochastique).

À l'inverse, les modèles les moins bons sont ceux ayant un biais et une variance élevés. Sur le graphique de droite de la figure 1.1, on voit que le point idéal pour tout modèle est le niveau de complexité auquel le carré du biais égalise la variance du modèle. Notons ici que la complexité du modèle est une notion étroitement liée au choix de $\mathcal{G}$: en effet, la complexité du modèle s'accroît généralement avec la taille et/ou la nature des éléments de $\mathcal{G}$. En pratique, le choix de $\mathcal{G}$ est limité à quelques familles usuelles de types de fonctions (modèles ou algorithmes) parmi lesquelles : l'ensemble des applications linéaires, XGBoost, Random Forest, CART, etc.

1.1.3 Principe de validation croisée

La validation croisée est une technique intellectuellement très intéressante. En apprentissage statistique, juger de la qualité d'un modèle sur les données qui ont servi à le construire ne permet en rien de savoir comment le modèle se comportera sur des nouvelles données (ceci fait allusion à l'erreur stochastique dont il était question dans la partie précédente). Il s'agit du problème dit de "généralisation". L'approche classique consiste alors à séparer l'échantillon initial (de taille n) en deux :

• Une partie qui servira à entraîner le modèle (la base d'apprentissage de taille m) ;

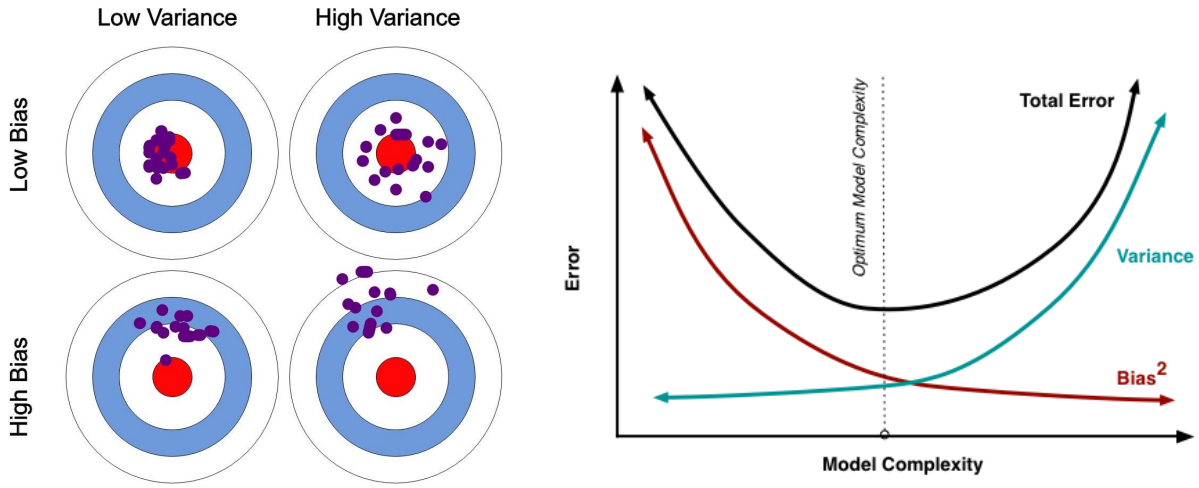


FIGURE 1.1 : *Illustration graphique du biais et de la variance (à gauche), contribution du biais et de la variance à l'erreur totale en fonction de la complexité du modèle (à droite) (issues de [Toward Data Science] et de [Durivaux]).*

• Une autre partie qui servira à tester le modèle (la base de test de taille $n - m$).

Cette dernière permet alors de mesurer un "vrai" risque prédictif. Ainsi, la validation croisée est généralement utilisée pour optimiser les paramètres ou hyperparamètres du modèle, de sorte à trouver les paramètres optimaux du modèle : c'est-à-dire ceux qui minimisent l'erreur de prédiction sur l'échantillon de validation. Elle permet ainsi d'éviter le sur-apprentissage.

Il existe différentes méthodes de validation croisée. Les trois les plus utilisées sont les suivantes :

• *Leave One Out Cross Validation* ;

• *k-folds cross validation* (avec k généralement égale à 5 ou 10) ;

• La *Méthode Holdout*.

Pour une présentation détaillée de ces différentes méthodes, nous suggérons de consulter Hastie *et al.* (2009). La figure 1.2 illustre sommairement le principe du *k-folds cross validation*, qui est celui le plus répandu en pratique, notamment pour $k = 5$ ou $k = 10$ et que nous utiliserons par la suite.

1.1.4 Choix des métriques d'évaluation de la performance des modèles

Une fois les modèles ajustés, il faut calculer leur performance de généralisation sur la base de test, afin de les comparer entre eux et retenir enfin, le(s) meilleur(s) d'entre eux.

Suivant qu'on se trouve face à un problème de classification ou de régression, la palette de choix des métriques diffère.

• En régression, les métriques traditionnellement utilisées sont la MSE (Moyenne des carrées des écarts entre la prédiction et la valeur réelle attendue) et la MAE (moyenne des écarts absolus entre la prédiction et la valeur réelle attendue absolue).

La MSE et la MAE sont définies par :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{f}(x^{(i)})\right)^2 \text{ et } MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y^{(i)} - \hat{f}(x^{(i)})\right|, \tag{1.3}$$

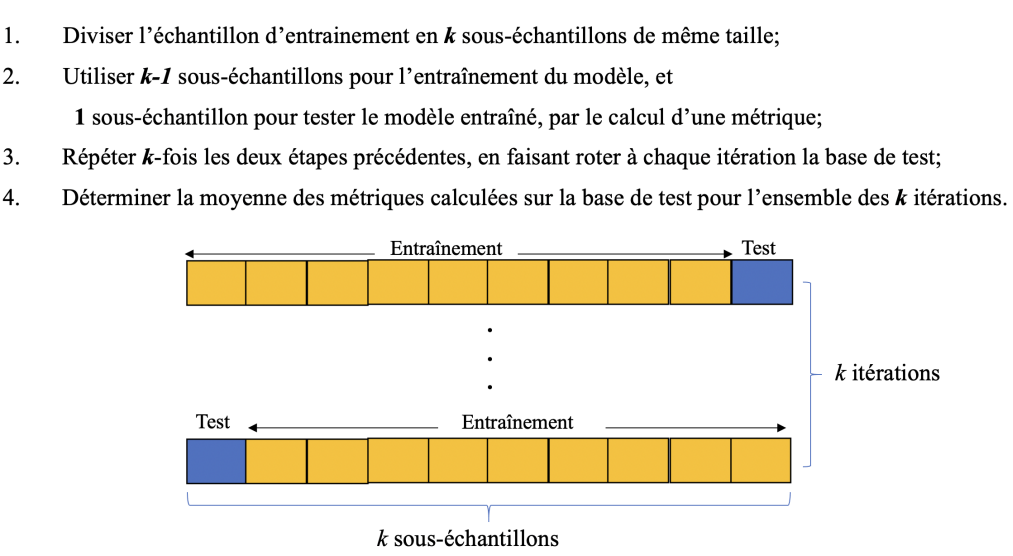


FIGURE 1.2 : *Procédure de la méthode de validation croisée à k plis (k-folds cross validation, en anglais).*

où $\hat{f}(x^{(i)})$ est la prédiction que le modèle renvoie pour la $i-$ème observation de la base de test.

La MSE et la MAE seront faibles si les réponses prédites sont très proches des réponses réelles, et sera grande si, pour certaines des observations, les réponses prédites et les réponses réelles diffèrent substantiellement.

Dans la comparaison entre les modèles, il peut arriver que l'un soit meilleur que l'autre pour l'une de ces métrique sans toutefois l'être pour l'autre.

Dans ce mémoire, la métrique que nous utiliserons pour l'évaluation de la performance des modèles sera basée sur la RMSE. En effet, cette dernière présente l'avantage de pénaliser les fortes erreurs de prédiction relativement à la MAE, ce qui coïncide bien avec les exigences de la tarification en assurance. Plus précisément, nous utiliserons une version normalisée de la RMSE, notée $RMSE_{mean}$ et définie par :

$$RMSE_{mean} = \frac{RMSE}{\bar{y}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{f}(x^{(i)})\right)^2}}{\frac{1}{n}\sum_{i=1}^{n} y^{(i)}} \tag{1.4}$$

L'avantage de faire recours à une telle version normalisée de la MSE est qu'elle permet d'obtenir une métrique sans unité : ce qui rend possible et plus rigoureuse la comparaison de performance entre différents modèles.

• En classification binaire, la métrique classiquement utilisée est le AUC (Area under curve) qui correspond à l'aire en dessous de la courbe ROC. Plus généralement la matrice de confusion, l'*Accuracy* ou la $F-$mesure sont les métriques utilisées pour les problèmes de classification à plus de deux classes.

En ce qui concerne les modèles de classification, c'est à dire les modèles qui seront utilisés dans le chapitre 5 pour prédire les valeurs des variables télématiques, la métrique retenue est l'AUC.

L'AUC permet de résumer la courbe ROC en un seul nombre : l'aire sous cette courbe. Il est

égal à 100% pour un modèle parfait et à 50% pour un modèle non-informatif. La performance d'un modèle d'apprentissage statistique se situe donc entre ces deux valeurs, et mieux l'AUC se rapproche 100% mieux le modèle est performant.

1.2 Transparence en actuariat

Dans le contexte spécifique de l'assurance, l'utilisation de l'apprentissage statistique pour classer et tarifer le risque à partir de données observées soulève des questions d'équité et de discrimination assurantielle (Barry et Charpentier (2022)).

Il existe des contraintes réglementaires qui visent à encadrer l'utilisation de l'apprentissage statistique dans le monde de l'assurance. En France, l'Autorité de contrôle prudentiel et de résolution (ACPR) veille à la protection des consommateurs et exige que l'assureur puisse justifier de manière détaillée et exacte toutes les décisions prises pour le calcul d'une prime d'assurance. Les tarifs doivent être explicites, notamment pour éviter le risque de discrimination. La lutte contre les préjugés discriminatoires est un point de contrôle majeur de l'ACPR (Fliche et Yang (2018)).

A titre d'exemple, depuis 2012, l'ACPR n'autorise plus la prise en compte du *genre* comme critère de tarification en assurance automobile.

Un autre point important est que le caractère boîte-noire des algorithmes d'apprentissage statistique se heurtent aux contraintes d'opérationnalité. En effet, bien au delà des contraintes réglementaires, il est fondamental pour l'assureur de connaître le plus finement possible, les règles de décision des algorithmes qu'il utilise pour mesurer le risque des assurés. Cela lui permet de mieux cerner (identifier) les potentiels facteurs de risque pour, *in fine*, mieux le gérer. Or, avec l'utilisation des algorithmes opaques d'apprentissage statistique, les règles de décision du modèle ne sont pas "directement" à portée de vue de l'assureur.

Chapitre 2

Modèles de tarification sujets à l'interprétation

Puisque nous traitons de l'interprétabilité des modèles de tarification dans ce mémoire, il est question dans ce chapitre de présenter quelques uns de ces modèles qui requiert de l'interprétabilité. Les modèles qui seront présentés dans ce chapitre sont ceux qui seront utilisés dans le chapitre 5 pour la tarification automobile. Ensuite, nous utiliserons les méthodes d'interprétabilités qui seront introduites au chapitre 4 afin d'expliquer les tarifs issus de ces modèles.

2.1 Un Modèle "transparent" : le modèle linéaire généralisé (GLM)

Étant donné qu'il s'agit dans cette section de présenter le modèle linéaire généralisé, une bonne pratique serait de commencer par une présentation au moins sommaire du modèle de régression linéaire. Ceci, afin de fluidifier la compréhension du modèle GLM, qui n'est rien d'autre qu'une extension du modèle linéaire. Nous renvoyons le lecteur en annexe A pour la présentation du modèle linéaire.

2.1.1 Formalisation

Les modèles linéaires généralisés ont été introduits en statistique au cours de la seconde moitié du 20^{eme} siècle par Nelder et Wedderburn (1972). Comme dans le cadre du modèle de régression linéaire, le principe de base de tout modèle linéaire généralisé consiste à conserver la structure linéaire (somme pondérée des caractéristiques), cependant le terme "généralisé" vient du fait que, par le biais d'une fonction de lien bien choisie, l'on s'autorise des distributions non gaussiennes de la variable cible. Plus précisément, si l'on désigne par Y la variable cible et $X = (X_1, \cdots, X_p)^\top$ les caractéristiques, le GLM suppose que :

$$Y|X \sim \mathcal{L}oi(\mu(X)), \quad \text{où} \quad m[\mu(X) = \mathbb{E}(Y|X)] = X^\top \beta \tag{2.1}$$

Avec $\mathcal{L}oi$ une loi paramétrique permettant de modéliser au mieux la variable d'intérêt Y conditionnellement aux caractéristiques X et m une fonction explicite convenablement choisie, appelée *fonction de lien*.

- Si Y est binaire alors, $\mathcal{L}oi$= loi de Bernouilli ;
- Lorsque Y est à valeurs dans $\{0, \cdots, K\}$ ($K \in \mathbb{N}$), $\mathcal{L}oi$= loi multinomiale ;
- Lorsque Y est à valeurs dans $\mathbb{N}$, on utilise généralement une loi de Poisson ;
- Lorsque Y est à valeurs dans $\mathbb{R}^+$, on utilise généralement une loi gamma, inverse gaussienne, etc.

2.1.2 Famille exponentielle

La famille exponentielle regroupe les lois de probabilités dont la fonction de masse de probabilité (ou la densité) est donnée par :

$$t(y,\theta) = h(y)\,\exp\left(\sum_{j=1}^{k}\eta_j(\theta)T_j(y) - B(\theta)\right), \quad y \in \mathcal{Y}, \tag{2.2}$$

où $\eta_j(\cdot)$, $T(\cdot) : \mathbb{R}^k \longmapsto \mathbb{R}^k$, $h(\cdot) : \mathbb{R}^d \longmapsto \mathbb{R}^k$ la fonction de base, $B(\cdot) : \mathbb{R}^k \longmapsto \mathbb{R}$ et $\theta \in \Theta \subset \mathbb{R}^k$.

Dans le cadre des modèles linéaires généralisés, on considère une classe beaucoup plus simple à un (01) paramètre (c'est-à-dire $k = 1$) et $T(y) = y$. Les fonctions B et h sont notées différemment.

Dans le cadre des GLM, la fonction de masse de probabilité (ou la densité) est donnée par :

$$\ln t(y,\theta) = \frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi), \quad y \in \mathcal{Y} \subset \mathbb{R}, \tag{2.3}$$

où $\mathcal{Y}$ est le support de la variable aléatoire, typiquement $\{0,1\}$, $\mathbb{N}$ ou $\mathbb{R}^+$; a, b, c trois fonctions différentiables propres à la loi ; θ le paramètre d'intérêt et ϕ le paramètre de dispersion. On note alors $Y \sim \mathcal{F}_{\text{exp}}(\theta,\, \phi,\, a,\, b,\, c)$.

Les moments du premier et du deuxième ordre se calculent facilement par :

$$\mathbb{E}(Y) = \mu = b'(\theta), \quad \mathbb{V}ar(Y) = a(\phi)b''(\theta);$$

2.1.3 Caractérisation des GLMs à partir de la famille exponentielle

Un modèle linéaire généralisé est caractérisé par trois hypothèses :

1. *Une loi de probabilité* : les $(y_i)_{1\leq i\leq n}$ sont supposés indépendants et y_i suit une loi $\mathcal{F}_{exp}(\theta_i,\, \phi,\, a,\, b,\, c)$, où θ_i est le paramètre d'échelle, ϕ le paramètre de dispersion ;

2. *Une fonction déterministe* : le vecteur des caractéristiques $X = (x_j)_{1\leq j\leq p}$ fournit le prédicteur linéaire $\eta = X^\top\beta$;

3. *Une fonction lien* $m : \mathbb{R} \longmapsto \bar{\mathcal{Y}}$ monotone, différentiable et inversible telle que : $\mathbb{E}(y_i) = m^{-1}(\eta_i = X^{(i)\top}\beta)$, pour tout $i \in \{1, \cdots, n\}$.

Notons que les paramètres θ_i sont liés au prédicteur linéaire η_i par la relation :

$$\mu_i = \mathbb{E}(y_i) = b'(\theta_i) = m^{-1}(\eta_i = X^{(i)\top}\beta)$$

C'est par cette relation qu'apparaît la structure transparente des GLMs. En effet le paramètre β_j explique directement comment la caractéristique individuelle x_j influence le prédicteur linéaire $\eta(X)$ et par conséquent la valeur attendue $\mu(X)$ de Y.

2.1.4 Interprétation du modèle linéaire généralisé

Une fois le modèle bien spécifié, l'estimation des paramètres β et du paramètre de dispersion ϕ se fait soit par la méthode du maximum de vraisemblance, soit par la méthode des moindres carrés pondérés itérés tel que décrit par Denuit et Charpentier (2005).

Effet marginal des variables explicatives

Une fois les paramètres estimés, l'actuaire aimerait connaître l'effet individuel de chaque caractéristique dans le calcul de sa variable cible afin d'être en mesure d'expliquer les différences de

résultats entre individus. Dans le cadre des GLMs, l'interprétation des coefficients $\hat{\beta}$ estimés des variables explicatives dépend de la fonction de lien m utilisé à l'étape de spécification du modèle.En guise d'illustration, désignons par $\hat{y}$ la valeur prédite par notre modèle pour un individus ayant les caractéristiques $x = (x_1, \cdots, x_p)$, alors :

$$\hat{y} = m^{-1}(x^\top \hat{\beta}) = h(x^\top \hat{\beta}) = h\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \, x_j\right), \quad \text{avec } h = m^{-1}$$

Désignons par $\tilde{x} = (x_1, \cdots, x_{j_0} + 1, \cdots, x_p)$, le même vecteur que x sauf pour une composante quelconque $j_0 \in \{1, \cdots, p\}$ où $\tilde{x}_{j_0} = x_{j_0} + 1$. Alors la prédiction pour ce nouveau vecteur de caractéristiques est :

$$\tilde{y} = h\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \, x_j + \hat{\beta}_{j_0}\right), \quad \text{avec } h = m^{-1}$$

L'effet marginal de la caractéristique x_{j_0} sur la cible correspond donc à la "différence" entre $\hat{y}$ et $\tilde{y}$ conséquente à la variation de x_{j_0} d'une unité supplémentaire.

– Si $m = Id$ (fonction identité), alors $h = Id$ et $\tilde{y} - \hat{y} = \hat{\beta}_{j_0}$: on dit que la caractéristique x_{j_0} a un effet additif sur la variable cible ;

– Si $m = \ln$ (logarithme), alors $h =$ exponentielle et $\dfrac{\tilde{y}}{\hat{y}} = \exp(\hat{\beta}_{j_0})$: dans ce cas que x_{j_0} a un effet multiplicatif sur la variable cible.

Importance des variables dans le GLM

Du fait des propriétés asymptotiques de l'estimateur du maximum de vraisemblance, comme dans le cas du modèle de régression linéaire (c.f. section A.1), l'importance d'une variable peut être définie sur la base un test significativité locale. La statistique utilisée est la suivante (statistique de Wald) :

$$w = \left(\frac{\hat{\beta}_j}{\sqrt{\tilde{J}(\beta)_{jj}}}\right)^2, \text{ où } \tilde{J}(\beta)_{jj} \text{ est le } j^{ieme} \text{ élément diagonal de la matrice } \tilde{J}(\beta) = (\tilde{I}(\beta))^{-1}$$

avec $\tilde{I}(\beta)$ la contrepartie empirique de la matrice d'information de Fisher associée à l'estimation des paramètres β.

Ainsi, plus w est élevée plus la variable est significative et donc importante dans l'explication de la variable d'intérêt.

2.1.5 Limites du modèle GLM

Certes, le modèle linéaire généralisé est actuellement le plus répandu dans la littérature actuarielle et l'industrie de l'assurance de par son caractère intuitif, transparent et son haut degré d'interprétabilité.

Cependant, de par sa faible complexité, il demeure limité dans la modélisation du comportement des assurés. Notamment, ils n'incluent pas les potentiels termes d'interactions entre les différentes variables explicatives pourtant parfois jugés très déterminants dans l'explication du comportement des assurés.

Pour palier à cette limite, les actuaires font généralement recours à des techniques manuelles. Afin de cerner l'influence non linéaire d'une variable continue sur la variable cible, une approche

consiste généralement à transformer la variable continue en question en une variable catégorielle, via un découpage en classes d'intervalles.

Une autre démarche, plus classique, consiste à substituer cette variable par une transformation (par exemple, polynomiale ou sinusoïdale). Quant aux termes d'interaction, ils sont construits et greffés au modèle manuellement.

Toutefois, ces approches semblent peu convaincantes. En effet, l'ajout à la main des termes de degrés supérieurs ou d'interaction impose des choix arbitraires de transformations de variables qui peuvent parfois s'avérer erronées. En outre, cela conduit le plus souvent à une prolifération (non optimale) des paramètres à estimer, engendrant inutilement d'énormes coûts algorithmiques.

L'innovation des modèles récents d'apprentissage statistique est qu'ils permettent dans une certaine mesure de détecter et capturer intrinsèquement les interactions entre les caractéristiques de manière beaucoup plus robustes que celles construites à la main.

Nous allons à présent introduire quelques modèles d'apprentissage statistique −parmi tant d'autres− dont la structure permet, *a priori*, de mieux prendre en compte les potentiels effets non-linéaires et d'interactions des variables explicatives sur la variable cible.

2.2 Un Modèle complexe : les Forêts aléatoires

2.2.1 Algorithme des forêts aléatoires

Les forêts aléatoires ont été introduite par Breiman (2001). Son principe consiste à construire une grande collection d'arbres dé-corrélés, puis à faire la moyenne de leur prédiction pour obtenir la prédiction finale.

Les forêts aléatoires sont populaires et sont mises en œuvre dans une variété de progiciels. Dans cette partie, on s'intéresse au fonctionnement de l'algorithme *Random Forest*. Ce modèle sera utilisé dans notre application actuarielle au chapitre 5 pour modéliser la fréquence et la sévérité des sinistres.

L'algorithme s'articule comme suit : soit B le nombre d'arbres utilisés dans la forêt aléatoire, soit y la variable cible, X la matrice des variables explicatives. On considère l'ensemble d'apprentissage que l'on a à notre disposition : $Z = \left(x^{(i)}, y_i\right)_{i=1}^{n}$, avec $x^{(i)} = (x_j^{(i)})_{1\leq j\leq p}$:

Entrées de l'algorithme :

- $COLLECTION$: une liste initialement vide, qui contiendra les différents arbres de la forêt aléatoire.
- N : la taille de l'échantillon boostrap sur lequel un arbre individuel sera entraîné. Par défaut $N = 63.20\%$ de la taille de la base d'entraînement dans le package $randomForest$ du logiciel R (version 4.7-1.1).
- n_{min} : la taille minimale des noeuds terminaux des différents arbres constituant la forêt aléatoire. Pour la classification, la valeur par défaut de n_{min} est 1. Pour la régression, la valeur par défaut de n_{min} est 5.
- m : nombre de variables échantillonnées pour le partitionnement de chaque noeud de l'arbre constituant la forêt aléatoire. Pour la classification, la valeur par défaut de m est la partie entière de $\sqrt{p}$. Pour la régression, la valeur par défaut de m est la partie entière de $\frac{p}{3}$.

Instructions de l'algorithme :

– Pour $b = 1$ à B :
  – Construire un échantillon *boostrap* Z^* de taille N à partir des données d'entraînement Z.
  – Ajuster l'arbre T_b sur les données *boostrapées* Z^*, en répétant récursivement les étapes suivantes pour chaque noeud terminal, jusqu'à ce que la taille minimale du noeud n_{min} soit atteinte.
    – Choisir m variables au hasard parmi les p variables explicatives.
    – Choisir la meilleur variable et le point de séparation parmi les m variables.
    – Diviser le noeud en deux noeuds fils sur la base de la meilleur variable et du meilleur point de séparation sélectionnés à l'étape précédente.
  – Ajouter T_b dans $COLLECTION$

Sorties de l'algorithme :

– En sortie de la boucle *for*, obtient $COLLECTION = \{T_b\}_{b=1}^{B}$;
Pour obtenir la prédiction d'un nouveau point de caractéristiques x, on procède comme suit :

– **Pour la régression** : $\hat{f}^B_{random\,forest}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

– **Pour la classification** : désignons par $\hat{C}_b(x)$ la classe prédite par le b^{eme} arbre de la liste $COLLECTION$: $\hat{C}^B_{random\,forest}(x) = mode\left(\left\{\hat{C}_b(x)\right\}_{b=1}^{B}\right)$.

2.2.2 Forêt aléatoire et surapprentissage

Lorsque le nombre de caractéristiques est important, mais que la fraction des caractéristiques pertinentes est faible, les forêts aléatoires risquent d'être peu performantes pour lorsque m est petit. En effet, à chaque fractionnement, la probabilité que les caractéristiques pertinentes soient sélectionnées sera faible. Par conséquent la majorité des arbres obtenus seront faiblement significatifs du phénomène modélisé. En pratique les principaux hyperparamètres su modèle *random forest*, à savoir : B et m sont déterminés par validation croisée. Les autres hyperparamètres sont généralement laissés à leur valeur défaut.

2.2.3 Avantages et inconvénients des forêts aléatoires

Quelques avantages

Comme avantages, nous notons que :

– Le principe de fonctionnement des forêts aléatoires est intuitif. Ce qui rend sa manipulation relativement facile.

– Sa mise en oeuvre est possible à partir de nombreux progiciels (R, Python notamment).

– Étant basé sur un principe de construction d'arbres dé-corrélés entre eux, ils permettent d'extraire un maximum d'informations dans les données, ce qui permet de réaliser des gains de précision prédicitve dans de nombreuses circonstances.

– Étant reconstitué à partir d'arbres de décision, les *random forest* permettent également de capturer les potentielles interactions entre les caractéristiques dans l'explication ou la prédiction de la variable cible.

Quelques inconvénients

Comme inconvénients, nous pouvons relever :

– Le caractère boîte noire de cet algorithme. Ce qui le rend difficilement interprétable au premier abord. Cependant, les méthodes d'interprétabilité introduites au chapitre 4 seront désormais utilisées pour lever le voile sur ces algorithmes de forêt aléatoire, dans l'optique de mieux comprendre les relations apprises par le modèle sur les données d'entraînement.

– La durée d'ajustement du modèle est généralement importante et peut aller jusqu'à plusieurs heures suivant la taille de la base d'entraînement. Cette limite se fait davantage ressentir à l'étape d'optimisation des hyperparamètres du modèle.

2.3 Un modèle "hybride" paru en 2022 : le LocalGLMnet

Dans cette section, nous nous appuyons principalement sur le récent article de Richman et Wüthrich (2022) consacré à l'introduction des modèles LocalGLMnet. Nous qualifions ce modèle d'hybride parce qu'il est partiellement transparent mais aussi, partiellement également boîte noire. Ce modèle sera utilisé dans notre cas d'application actuarielle du chapitre 5 pour la modélisation de la fréquence et de la sévérité agrégée de sinistres en tarification automobile. Considéré comme modèle hybride, nous l'interpréterons d'une part par une approche d'interprétation basée sur modèle, et d'autre part nous l'interpréterons à l'aide des outils d'interprétabilité *post hoc* indépendants du modèle développés au chapitre 4. Une telle démarche nous permettra de confronter la cohérence entre des interprétations basées sur le modèle de celles issues des méthodes d'interprétabilité *post hoc* (présentées au chapitre 4).

2.3.1 Généralités sur les LocalGLMnet

Les modèles d'apprentissage profond communément appelés *réseaux de neurones profonds* permettent d'obtenir des modèles de régression compétitifs généralement plus précis que les modèles statistiques classiques tels que les modèles linéaires généralisés, et même les modèles de *"machine learning"* tels que les forêts aléatoires. Ce succès repose sur le fait que ces modèles réalisent une ingénierie des caractéristiques en interne, afin d'extraire un maximum d'informations de ces caractéristiques pour une tâche de prédiction donnée.

Cependant les prédictions issues de ces modèles sont le plus souvent difficiles à interpréter ou à expliquer. Pour palier à cette limite de transparence, de nombreuses approches sont proposées dans la littérature :

• L'une des plus populaires parmi ces approches est celle qui consiste à mener des interprétations *post hoc* des prédictions issues du modèle. C'est cette approche que nous développons dans le chapitre 4.

• Une autre approche consiste à concevoir des modèles hybrides qui exploite à la puissance prédictive des réseaux de neurones tout gardant des traits de transparence. C'est dans cette seconde perspective que s'inscrit les modèles LocalGLMnet. Les LocalGLMnet sont basés sur une architecture novatrice qui conserve la structure linéaire transparente des modèles GLMs, à la différence que contrairement au cas des modèles GLMs où les coefficients des prédicteurs sont constants, dans les modèles LocalGLMnet les coefficients prédicteurs sont eux-même fonctions de l'ensemble des prédicteurs. Les coefficients des prédicteurs linéaires sont ajustés grâce à des modèles non-linéaires de type réseaux de neurones profonds.

Des approches similaires au LocalGLMnet ont été étudiées dans la littérature actuarielle dans le contexte de la prévision de la mortalité, par exemple Perla *et al.* (2021) qui estiment les coefficients d'un modèle de type Lee-Carter en utilisant des réseaux neuronaux convolutifs et récurrents.

2.3.2 Formalisation mathématique

L'architecture des modèles LocalGLMnet est obtenue suite à une combinaison spécifique entre les modèles de type GLM et des modèles de type réseaux de neurones (notamment les *feed-forward neural (FFN)*). Ainsi, pour présenter les modèles LocalGLMnets, il est nécessaire de présenter au préalable les modèles GLM et FFN. La présentation des GLM a déjà été faite plus haut dans ce chapitre, à la sous-section 2.1. Présentons maintenant les réseaux de neurones profonds de type *Feed-Forward Neural*.

Réseau neuronal à anticipation (FFN) entièrement connecté

Un réseau neuronal à anticipation (FFN) entièrement connecté s'appuie sur l'ingénierie des informations basée sur les caractéristiques $x = (x_1, \cdots, x_p),\ p \in \mathbb{N}^*$ à partir de transformations non-linéaires. La $m-$ième couche FFN du réseau neuronal à anticipation (FFN) entièrement connecté est définie par l'application :

$$\begin{aligned} z^{(m)} : \mathbb{R}^{q_{m-1}} &\longrightarrow \mathbb{R}^{q_m} \\ t &\longmapsto z^{(m)}(t) = \left(z_1^{(m)}(t), \cdots, z_{q_m}^{(m)}(t)\right)^\top \end{aligned} \tag{2.4}$$

avec $q_{m-1},\ q_m \in \mathbb{N}^*$ et $q_0 = p$; $t = \left(t_1, \cdots, t_{q_{m-1}}\right)^\top \in \mathbb{R}^{q_{m-1}}$, et $z_j^{(m)}(t),\ 1 \leq j \leq q_m$ le $j-$ième neurone de la couche m définie par :

$$z_j^{(m)}(t) = \Phi_m\left(\omega_{0,\,j}^{(m)} < \omega_j^{(m)},\, t >\right) = \Phi_m\left(\omega_{0,\,j} + \sum_{l=1}^{q_{m-1}} \omega_{l,\,j}^{(m)} t_l\right) \tag{2.5}$$

où $\Phi_m : \mathbb{R} \longrightarrow \mathbb{R}$ désigne la fonction d'activation choisie pour la $m-$ième couche $z^{(m)}$; $\omega_j^{(m)} = \left(\omega_{l,\,j}^{(m)}\right)_{1 \leq l \leq q_{m-1}}^\top \in \mathbb{R}^{q_{m-1}}$ et une constante $\omega_{0,\,j} \in \mathbb{R}$. Pour rester fidèle aux notations de la littérature, on utilise le produit scalaire $< a, b >= a^\top b$, de sorte que le terme $x^\top \beta$ utilisé auparavant devient ici $< \beta, x >$ ou $< x, \beta >$.

La définition d'une couche étant à présent faite, un réseau neuronal à anticipation (FFN) de profondeur $d \in \mathbb{N}$ noté $z^{(d:1)}$ est une composition de d couches FFN. Plus précisément :

$$\begin{aligned} z^{(d:1)} : \mathbb{R}^p &\longrightarrow \mathbb{R}^{q_d} \\ x &\longmapsto z^{(d:1)}(x) = \left(z^{(d)} \circ \cdots \circ z^{(1)}\right)(x) \end{aligned} \tag{2.6}$$

Définition du modèle LocalGLMnet

L'idée fondamentale des modèles LocalGLMnet consiste à préserver autant que possible la structure linéaire des GLMs de manière à pouvoir évaluer la manière dont les caractéristiques individuelles x_j de x influencent la fonction de régression. On laisse les paramètres β_j dépendre des caractéristiques x (les β_j sont également appelés *poids d'attention*).

Soit $x \in \mathbb{R}^p$ le vecteur des caractéristiques, désignons par Y la variable cible, on suppose que conditionnellement à x, la moyenne de Y, notée μ, satisfait l'hypothèse de régression suivante :

$$x \longmapsto m(\mu) = m(\mu(x)) = \beta_0 + < \beta,\, x >= \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{2.7}$$

où $m : \mathbb{R} \longrightarrow \mathbb{R}$ est une fonction de liaison strictement monotone et continue.

La particularité des modèles LocalGLMnets par rapport aux modèles GLMs est que dans les

LocalGLMnets on fait l'hypothèse que les coefficients β_j de la régression çi-dessus (Equation 2.7) sont dépendants des caractéristiques x. Pour modéliser les poids d'attention $\beta_j(x)$, $1 \leq j \leq p$, on se donne un réseau de neurones à anticipation (FFN) de profondeur $d \in \mathbb{N}$ dont les dimensions d'entrée et de sortie sont égales à p (i.e., $q_0 = q_p = p$) :

$$\begin{aligned} &\beta : \mathbb{R}^p \longrightarrow \mathbb{R}^p \\ &x \longmapsto \beta(x) = z^{(d:1)}(x) = \left(z^{(d)} \circ \cdots \circ z^{(1)}\right)(x) \end{aligned} \tag{2.8}$$

avec $z^{(d:1)}$ définit telle qu'à l'équation 2.6.

Le LocalGLMnet est donc défini par la décomposition additive suivante (après application d'une fonction de liaison m strictement monotone et continue) :

$$x \longmapsto m(\mu(x)) = \beta_0 + < \beta(x),\, x >= \beta_0 + \sum_{j=1}^{p} \beta_j(x) x_j \tag{2.9}$$

2.3.3 Ajustement du modèle LocalGLMnet

Méthodes de descente de gradient stochastique (SGD)

Comme pour la plus part des algorithmes d'apprentissage statistique profond, le modèle LocalGLMnet peut être ajusté par des méthodes de descente de gradient stochastique.

L'objectif de la méthode de descente de gradient (classique) est de trouver un minimum d'une fonction de plusieurs variables le plus rapidement possible. Elle est basée sur l'idée selon laquelle le vecteur opposé au gradient indique une direction vers des plus petites valeurs de la fonction. Pour minimiser la fonction, il suffit alors de suivre d'un pas la direction opposée à son gradient et d'itérer le processus jusqu'à l'obtention d'un minimum local "raisonnable". La méthode de descente de gradient stochastique (SGD) est une façon d'optimiser les calculs de la descente de gradient afin d'être encore plus rapide relativement à la descente de gradient classique.

On considère des données $\left(x^{(i)},\, y_i\right)_{i=1}^{n}$, avec $x^{(i)} \in \mathbb{R}^p$ des covariables et $y_i \in \mathbb{R}$ la variable cible. L'on désire trouver une fonction $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ qui ajuste au mieux y sur le vecteur de covariables x. On introduit une fonction d'erreur E qui mesure l'*erreur totale* du modèle, c'est-à-dire, la somme des carrés des écarts individuels E_i, entre la sortie attendue y_i et la sortie produite $f(x^{(i)})$:

$$E = \sum_{i=1}^{n} \underbrace{\left(y_i - f(x^{(i)})\right)^2}_{E_i}.$$

L'objectif est de déterminer la fonction f qui minimise l'erreur E.

L'erreur totale E est une fonction de $p \in \mathbb{N}^*$ paramètres $a_1, \cdots, a_p$: il s'agit des paramètres qui définissent l'expression de f.

Pour minimiser E suivant les paramètres $a_1, \cdots, a_p$, avec la méthode de descente de gradient stochastique classique, on procède comme suit :

Etape 1− On part d'un point initial : $P_0 = (a_1, \cdots, a_p) \in \mathbb{R}^p$;

Etape 2− On calcule ensuite (par récurrence) : $P_{k+1} = P_k - \delta \nabla E(P_k) = P_k - \delta grad\, E(P_k)$, $k = 0, 1, 2 \cdots, N$, avec N plus ou moins grand, de sorte que $\nabla E(P_N)$ soit suffisamment faible et δ le taux d'apprentissage (optimalement choisi par la méthode de la *validation croisée*).

Pour effectuer l'étape 2 ci-dessus, il faut calculer $\nabla E(\cdot)$ en tout point P_k, $k = 0, 1, 2 \cdots, N$.

Et comme $\nabla E(\cdot) = \sum_{i=1}^{n} \nabla E_i(\cdot)$, ceci revient à calculer une somme de n termes à chaque itération $k \in \{1, \cdots, N\}$, ce qui peut être assez lourd pour des échantillons de grandes tailles (n assez grand). C'est donc à ce moment qu'entre en jeu la méthode de la descente de gradient stochastique qui vise à réduire les temps de calcul, tout conservant une bonne qualité d'optimisation.

Principe de la descente de gradient stochastique

L'idée de la descente de gradient stochastique est de considérer un seul gradient ∇E_i à chaque itération k, à la place de ∇E. Ce qui revient à modifier l'épate 2 de la méthode du gradient classique en :

$$P_{k+1} = P_k - \delta \nabla E_I(P_k)$$

où I est une variable aléatoire uniforme sur $\{1, \cdots, n\}$ (correspondant au numéro d'une instance de donnée) qui à chaque itération k, choisir un seul i aléatoirement dans $\{1, \cdots, n\}$ pour le calcul de P_{k+1}. Ainsi, au lieu de calculer un gros gradient (celui de E tout entier), l'on se ramène au calcul d'un gradient beaucoup plus simple (celui d'un seul E_i). Toutefois, il est important de noter qu'en pratique, bien que la méthode SGD réduit la complexité du gradient à calculer à chaque itération, elle nécessite néanmoins un nombre d'itérations N relativement plus important afin d'assurer la convergence des solutions obtenues.

Remarque : pour une présentation plus détaillée de la méthode SGD consulter Goodfellow *et al.* (2016).

2.3.4 Interprétation du modèle LocalGLMnet

Comment interpréter les coefficients résultant de l'estimation de l'équation (2.9) ?

Pour répondre à cette question sélectionnons arbitrairement une composante $\beta_j(x)x_j$, j fixé dans $\{1, \cdots, p\}$ de l'équation (2.9) et interprétons la.

Les interprétations se font suivant différents cas de figures :

– Si $\beta_j(x) \equiv \beta_j \neq 0$, alors le poids d'attention $\beta_j(x)$ ne dépend pas des caractéristiques x. On retrouve une interprétation similaire au cas classique des GLMs.

– Si $\beta_j(x) \equiv 0$ alors la caractéristique j n'a pas une influence "significative" sur la variable cible Y.

– Lorsque $\beta_j(x) = \beta_j(x_j)$, le modèle révèle que nous avons un terme $\beta_j(x_j)x_j$ qui n'interagit pas avec les autres caractéristiques.

De manière générale, pour un j quelconque fixé dans $\{1, \cdots, p\}$, on peut analyser l'effet des différentes composantes de x sur $\beta_j(x)$. Si $\beta_j(x)$ ne montre aucune sensibilité par rapport aux composantes différentes de x_j, alors nous n'avons pas d'interactions et dans le cas contraire, nous en avons. Afin d'extraire ces informations d'interactions, nous considérons les gradients :

$$\nabla \beta_j(x) = \left(\frac{\partial}{\partial x_1} \beta_j(x), \cdots, \frac{\partial}{\partial x_p} \beta_j(x) \right)^\top \in \mathbb{R}^p \tag{2.10}$$

La $j-$ème composante de $\nabla \beta_j(x)$ explore si nous avons un effet linéaire de x_j sur la variable cible ou non.

• On dira qu'il y a effet linéaire de x_j sur Y, si et seulement si $\frac{\partial}{\partial x_j} \beta_j(x) = 0$.

• Les composantes du gradient de la forme $\frac{\partial}{\partial x_k}\beta_j(x)$, avec $k \neq j$ quantifient les forces d'interaction entre la caractéristique x_j et les autres caractéristiques.

Il faut noter que l'analyse des interactions se fait par le biais du tracé des diagrammes des fonctions de gradients ci-dessus. Nous illustrerons l'interprétation de ces interactions entre les caractéristiques dans notre cas d'application au chapitre 5, lors de l'interprétation du modèle LocalGLMnet mis en place pour l'ajustement de la fréquence de sinistres.

– Une fois les poids d'attention $\hat{\beta}_j(x)$ tous estimés pour chaque caractéristique d'entrée, $1 \leq j \leq p$, nous pouvons définir une notion d'importance des variables par le calcul de la mesure suivante :

$$VI_j = \frac{1}{n}\sum_{i=1}^{n} |\hat{\beta}_j(x^{(i)})| \tag{2.11}$$

où $\mathcal{D} = \left(x^{(i)}\right)_{i=1}^{n}$ désigne la matrice de covariables du jeu de données d'entraînement. Les variables x_j sont au préalable standardisées (centrées et réduites) à l'étape d'ajustement du modèle. Plus la valeur VI_j est élevée, plus la caractéristique x_j contribue globalement à la formation des prédictions.

2.3.5 LocalGLMnet et présélection des variables

L'un des atouts des modèles LocalGLMnet est qu'ils possèdent une procédure de présélection des variables explicatives qui leur est propre.

Le principe consiste à mettre en place une procédure rigoureuse de test d'hypothèses pour chaque $j = 1, \cdots, p$:

$$H_0 : \hat{\beta}_j(x) = 0 \text{ contre } H_1 : \hat{\beta}_j(x) \neq 0 \tag{2.12}$$

Supposons que nous disposons de p variables explicatives $X = (X_1, \cdots, X_p) \in \mathbb{R}^p$. Nous souhaitons identifier parmi ces variables, celles qui ont un effet significatif dans l'ajustement de notre variable cible Y : c'est à dire les variables X_j pour lesquelles $\beta_j(x) \neq 0$ $(1 \leq j \leq p)$, dans la régression de Y sur les X à l'aide d'un modèle LocalGLMnet.

Richman et Wüthrich (2022) (auteurs du modèle LocalGLMnet) proposent de suivre la procédure suivante :

– Étendre l'ensemble des caractéristiques X en ajoutant une variable X_{p+1} qui soit complètement indépendante du vecteur de variable X et qui n'ait a priori pas d'incidence sur la variable cible Y.

– Une fois X_{p+1} obtenue (par exemple à l'aide d'un générateur de nombre aléatoire), la prochaine étape consiste à ramener à la même échelle l'ensemble des $p+1$ variables (on peut par exemple les centrer et les réduire).

– Ensuite, on calibre notre LocalGLMnet sur $X_+ = (X, X_{p+1})$ et on obtient les coefficients : $\hat{\beta}_1(x_+^{(i)}), \cdots, \hat{\beta}_{p+1}(x_+^{(i)})$, pour $i = 1, \cdots, n$. Puisqu'initialement, X_{p+1} a été minutieusement choisi de manière à n'avoir aucune influence dans la régression, alors on s'attend à ce que la moyenne et la variance empirique de $\hat{\beta}_{p+1}(x_+)$ soient proches de 0, c'est-à-dire :

$$\begin{cases} \bar{b}_{p+1} = & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \hat{\beta}_{p+1}(x_+^{(i)}) \approx 0 \quad \text{et} \\ \\ \hat{s}_{p+1} = & \sqrt{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(\hat{\beta}_{p+1}(x_+^{(i)}) - \bar{b}_{p+1}\right)^2} \approx 0 \end{cases} \tag{2.13}$$

– L'hypothèse nulle du test 2.12 sera rejetée avec une confiance de $1-\alpha$, ($\alpha \in (0,\,1)$), pour la variable j, si la proportion de $\hat{\beta}_j(x_+^{(i)})$, $1 \leq i \leq n$, contenue dans l'intervalle centrée I_α est substantiellement plus petite que $1-\alpha$, avec :

$$I_\alpha = [q_{\mathcal{N}}(\alpha/2) \cdot \hat{s}_{p+1},\, -q_{\mathcal{N}}(\alpha/2) \cdot \hat{s}_{p+1}] \text{ l'intervalle de confiance à } 1-\alpha \text{ de } \hat{\beta}_{p+1}(x_+). \tag{2.14}$$

Dans le cas d'application du chapitre 5 la variable de contrôle X_{p+1} a été générer aléatoirement suivant une distribution uniforme sur $[-\sqrt{3},\,\sqrt{3}]$.

Chapitre 3

Notion d'interprétabilité en apprentissage statistique

"Les explications sont la monnaie dans laquelle nous échangeons nos croyances." Lombrozo, 2006.

Les prédictions obtenues par les modèles d'apprentissage statistique, par exemple avec les réseaux de neurones artificiels sont généralement très précises. Cependant, les informations relatives à la prise de décision de ces modèles sont opaques pour l'utilisateur. Or, dans un domaine aussi sensible que celui des assurances, la précision prédictive n'est pas le seul enjeu à prendre en considération : il est également nécessaire d'être en mesure d'expliquer la prise de décision issues des modèles utilisés. Ce chapitre est structuré en quatre sections :

• Tout d'abord, nous présentons quelques raisons de la recherche d'interprétabilité des algorithmes d'apprentissage statistique ;

• Ensuite, nous donnons une définition de la notion *d'interprétabilité* en apprentissage statistique ;

• Enfin, dans les deux dernières sections nous dressons un état de l'art des différentes méthodes d'interprétabilité de modèles d'apprentissage statistique *supervisé*, tout en mettant un accent particulier sur les méthodes les plus utilisées à l'heure actuelle.

3.1 Raisons de la recherche d'interprétabilité

Les modèles d'apprentissage statistique ne sont plus jugés sur la seule base de leur performance prédictive, mais également sur leur niveau d'interprétabilité. Dans cette section nous présentons les raisons qui motivent la recherche d'interprétabilité des modèles d'apprentissage statistique. Nous nous appuyons principalement sur les articles de Lipton (2018) et Burkart et Huber (2021) consacrés à la présentation de la notion d'interprétabilité des modèles complexes d'apprentissage statistique supervisé.

3.1.1 Confiance

L'une des principales raisons de la recherche d'interprétabilité est d'évaluer dans quelle mesure l'on peut faire *"confiance"* aux prédictions issues du modèle. Mais qu'est-ce que la confiance dans notre contexte ? S'agit-il simplement de s'assurer que le modèle prédira de manière correcte ? Si tel est le cas, un modèle suffisamment précis sur un jeu de données test (RMSE, AUC, MAE, etc. élevée) devrait être digne de confiance et l'interprétabilité ne servirait à rien. La confiance en un modèle est une notion plus étendue. Elle ne se limite pas au pouvoir prédictif du modèle, mais il est également question de comprendre le processus sous-jacent à la prise de décision du modèle.

En assurance automobile par exemple, au delà d'être précis dans le calcul de la prime, l'assureur doit également s'assurer que son modèle n'utilise pas des variables discriminatoires (comme le sexe)

dans le calcul de la prime. Pour se faire, il a besoin d'interpréter son modèle pour mieux comprendre son fonctionnement.

3.1.2 Causalité

L'interprétabilité peut également viser à étudier la causalité entre les phénomènes. On aimerait parfois se faire une idée de l'importance des variables d'entrée dans la formation du résultat obtenu en sortie. Les modèles d'apprentissage statistique sont parfois utilisés par des chercheurs et praticiens à des fins heuristiques dans l'espoir de déduire des propriétés ou de générer des hypothèses sur le monde réel, qui pourront ensuite être tester expérimentalement. Pour se faire, il est nécessaire dans ces situations de comprendre les relations sous-jacentes apprises par le modèle.

Cependant, l'étude de l'inférence causale à partir de l'interprétation des modèles d'apprentissage statistique doit se faire avec beaucoup du recul. Les règles de décision des modèles d'apprentissage statistiques sont construites sur la base d'associations déduites de calculs de corrélation et d'entropie croisée. Ce qui ne reflètent pas toujours des relations de causalité au sens strict du terme (corrélation $\neq$ causalité).

3.1.3 Transférabilité

Les règles de décision apprises par le modèle doivent être transférables aux nouvelles données non vues par le modèle. Caruana *et al.* (2015) illustre cette propriété de transférabilité par un exemple simple. Ils considèrent un modèle entraîné à prédire la probabilité de décès suite à une pneumonie. Le modèle attribue un risque de décès moins important aux patients souffrant d'asthme. Ce qui est paradoxale lorsqu'on sait que l'asthme est un facteur de vulnérabilité aux maladies respiratoires. Or, ce faible risque de décès chez les patients atteints d'asthme était tout simplement dû au fait qu'ils recevaient déjà un traitement plus agressif du fait de leur fragilité. Si ce modèle était utilisé pour lutter contre la pneumonie dans une nouvelle population, les personnes atteintes d'asthme recevraient un traitement moins agressif, ce qui invaliderait le modèle. En interprétant le modèle, on vérifie ainsi si les propriétés apprises sont généralisables.

Considérons un autre exemple, cette fois en assurance. Une fois qu'un modèle de tarification est industrialisé, il convient de garantir que sa pertinence reste valide dans le temps. Si la qualité ou la définition d'une variable évolue au cours du temps (par exemple, l'apparition sur le marché de nouvelles marques de voitures, alors que le modèle n'a été entraîné que sur des marques existantes avant une certaine date), le pouvoir prédictif et la qualité des décisions prises par le modèle peuvent fortement se dégrader. Il est important de connaître ce sur quoi le modèle s'appuie pour faire ses prédictions afin de savoir dans quelle mesure il reste pertinent.

3.1.4 Caractère informatif

Kim *et al.* (2015) et Huysmans *et al.* (2011) soutiennent que dans certaines situations, les modèles d'apprentissage automatique supervisé peuvent être utilisés tout simplement, pour fournir des informations supplémentaires et utiles aux décideurs humain s. Considérerons par exemple un lycéen en classe terminale dans un établissement de la place qui cherche à obtenir des conseils de la part de son conseiller d'orientation. L'élève aimerait savoir quelle formation lui scierait le mieux après son baccalauréat. Le conseiller pourrait se contenter de lui proposer une formation sans justification, mais ce n'est pas forcément très convaincant, même si l'élève a conscience que le conseiller est raisonnablement expérimenté.

Or, admettons que le conseiller se fait assister par un un modèle d'apprentissage statistique pour donner ses recommandations d'orientation aux élèves, alors au-delà de simplement prédire

les potentielles choix de formation adaptées aux caractéristiques de l'élève, le conseiller pourrait désiré de faire mieux en fournissant en plus à l'élève, des éléments justificatifs ou explicatifs de sa décision. Pour cela le conseiller devrait faire recours à l'interprétabilité des modèles.

Selon Selvaraju *et al.* (2016), la compréhension du modèle de prédiction et des facteurs sous-jacents permet aux experts du domaine de comparer le modèle de prédiction aux connaissances existantes du domaine. En actuariat par exemple, elle permettrait à l'assureur de mieux cerner le comportement des assurés et donc les nouveaux *risques émergents* de son portefeuille, rendant ainsi possible une meilleure gestion du risque.

Le caractère informatif des algorithmes d'apprentissage statistique est davantage mis en avant, d'autant plus que connaître les raisons d'une certaine décision est devenu un besoin sociétal (Goodman et Flaxman (2017)). A l'heure actuelle, dans les pays de l'Union Européenne, toute personne qui est affectée par une décision automatisée peut faire usage du droit à l'explication détaillée et compréhensible de la décision en question. D'où une nécessité de l'interprétabilité.

3.1.5 Prise de décision juste et éthique

Aujourd'hui, nous laissons des programmes informatiques approuver des octrois de crédit, calculer certaines primes de risque en assurance, filtrer des candidats à un emploi etc. Certains tribunaux vont jusqu'à déployer des algorithmes informatisés pour prédire le risque de récidive[1] (c.f. Chouldechova (2017)). Il est probable que cette tendance ne fera que s'accélérer au fil du temps. Ce qui remet au premier plan les problématiques de justice et d'éthique.

Face à cette problématique, de nouvelles réglementations (RGPD) ont vus le jour dans l'Union européenne. Elles proposent que les individus affectés par des décisions algorithmiques aient un droit à l'explication (Goodman et Flaxman (2017)).

Dans le contexte spécifique de l'assurance en France, il existe d'autres contraintes réglementaires liées à l'utilisation des modèles d'apprentissage automatique pour la prise de décision. L'Autorité de Contrôle Prudentiel et de Résolution (ACPR) veille à la protection des consommateurs et exige que l'assureur puisse justifier de manière détaillée et exacte toutes les décisions prises dans le calcul de la prime. En effet, la lutte contre les préjugés discriminatoires est un point de contrôle majeur pour l'ACPR, Fliche et Yang (2018).

À titre d'exemple l'ACPR n'autorise pas la prise en compte du *genre* comme critère de tarification en assurance automobile. L'interprétabilité peut ainsi servir d'indicateur à l'évaluation des critères difficilement quantifiables des modèles d'apprentissage statistique telle que leur équité.

3.2 Définition de la notion d'interprétabilité : présentation du cadre PDR

Le concept d'interprétabilité est à la fois important mais difficile à préciser. La définition du terme *interprétabilité* en apprentissage statistique ne fait pas l'objet d'un consensus. Tant les motivations de l'interprétabilité sont nombreuses et les objets à interpréter divers (des modèles linéaires aux réseaux de neurones). Dans cette section nous apportons des éléments de réponse aux questions suivantes :

- Que signifie interpréter un modèle ?
- Quelle méthode d'interprétation utiliser pour quel problème ou quel public particulier ?

[1] c'est-à-dire la probabilité qu'un individu retombe dans un comportement criminel

3.2.1 Deux grandes classes de méthodes d'interprétabilité

• Cycle de vie d'un projet d'apprentissage automatique

De prime abord, pour mieux fixer les idées commençons par situer le processus d'interprétation dans le cycle de vie d'un projet de science de données. À quelle(s) étape(s) du processus de construction du modèle interviennent l'interprétation ?

La figure 3.1 présente une description schématique du cycle de vie d'un projet de science de données. L'interprétabilité entre jeu dans les phases de modélisation et d'analyse post hoc.

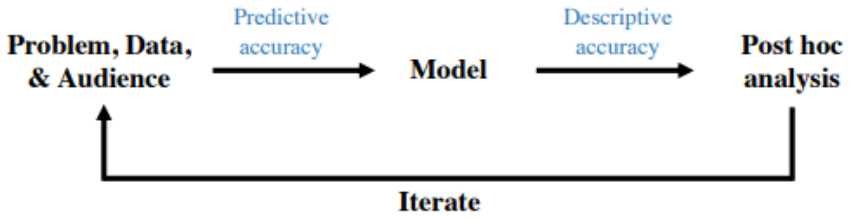


FIGURE 3.1 : *Vue d'ensemble des différentes étapes (texte en noir) du cycle de vie d'un projet de science des données où l'interprétabilité est importante (issue de l'article Murdoch* et al. *(2019)).*

On distingue deux grandes classes de méthodes d'interprétation de modèle :

• Méthodes d'interprétation basées sur le modèle

Dès l'étape de modélisation, sur la base du problème à résoudre et des caractéristiques des données disponibles, le praticien sélectionne un ou plusieurs modèle(s) candidat(s) et les ajustent. A ce stade, entre déjà en jeu des considérations d'interprétabilité liées au choix entre les modèles plutôt simples, relativement faciles à interpréter, et les modèles plus complexes, de type boîtes noires, plus à même de s'adapter aux données mais a priori, difficile à interpréter. A ce niveau, on parle *d'interprétabilité basée sur le modèle.*

• Méthodes d'interprétation post hoc

A l'étape de l'analyse post hoc, le praticien interroge ou analyse chacun des modèles ajustés. Contrairement aux méthodes d'interprétabilité basées sur le modèle, où l'interprétabilité est directement intégrée au processus de modélisation, les outils utilisés pour mener des interprétations post hoc interviennent après la modélisation. L'outil ou la méthode d'interprétation choisie est fonction du problème étudié et du public à qui est destiné les informations issues de l'interprétation.

3.2.2 Présentation du cadre PDR : précision prédictive, descriptive et pertinence

Le cadre PDR (précision prédictive, précision descriptive, pertinence) est étroitement lié au cycle de vie d'un projet de science de données. À l'issue d'un projet de conception d'un modèle d'apprentissage statistique, les résultats obtenus devraient être le plus fidèles possible au processus sous-jacent que le praticien essaie de comprendre. Cependant, il existe généralement des bruits (erreurs) qui altèrent l'exactitude des résultats obtenus en sortie. Ces erreurs peuvent provenir de deux sources : soit au niveau de l'approximation des associations contenues dans les données par un modèle inapproprié (précision prédictive), soit au niveau de l'explication des relations apprises par le modèle (précision descriptive). Une interprétation est digne de confiance d'autant plus qu'elle maximise ces deux dimensions de précision (prédictive & descriptive).

3.2.2.1 Précision prédictive

D'après l'article de Murdoch *et al.* (2019), la *précision prédictive* renvoie au degré auquel une méthode d'interprétation capture avec exactitude les relations sous-jacentes contenues dans les données. Elle renseigne sur la capacité d'un algorithme à s'ajuster au processus que le praticien cherche à comprendre.

En apprentissage automatique, l'évaluation de la performance prédictive d'un modèle se fait par le biais de métriques telles que le *taux d'erreur* pour les problèmes de classification (et courbe ROC pour la classification binaire), l'*erreur quadratique moyenne* et ses dérivées pour les problèmes de régression. Ces dernières métriques sont calculées sur plusieurs échantillons de données test pour s'assurer de la fiabilité et de la robustesse de la précision prédictive.

Par ailleurs, suivant la problématique étudiée, l'on peut parfois être emmené à dépasser le cadre d'un indicateur de précision moyenne tel que l'AUC ou la RMSE, pour s'intéresser au degré de précision dans certaines classes spécifiques d'intérêt. L'évaluation de la précision prédictive, peut également se faire par comparaison de la distribution de valeurs prédites et celle des valeurs attendues.

3.2.2.2 Précision descriptive

La *précision descriptive* renvoie au "degré" auquel une méthode d'interprétation capture objectivement les relations apprises par un modèle d'apprentissage automatique donné (Murdoch *et al.* (2019)).

Généralement, les méthodes d'interprétation basées sur le modèle fournissent une représentation fiable des relations apprises par le modèle. Ainsi, l'approche d'interprétabilité basée sur le modèle permet d'obtenir une *précision descriptive* élevée, mais au détriment de la *précision prédictive*. *A contrario*, l'approche d'interprétabilité post hoc permet généralement d'obtenir une *précision prédictive* élevée, cela, contre une *précision descriptive* faible.

En pratique en fonction du problème et du public visé, le praticien est appelé à choisir une approche d'interprétabilité de manière à assurer un arbitrage raisonnable entre le niveau de précision prédictive et de précision descriptive. Ce qui introduit la troisième dimension du cadre PDR : la pertinence.

3.2.2.3 Pertinence

Pour fixer le choix d'une approche d'interprétabilité, les critères de précision ne suffisent pas, les informations extraites doivent également être pertinentes. Par exemple dans le contexte de la météorologie, un agriculteur et un météorologue peuvent vouloir chacun des informations différentes d'un même modèle[2].

La pertinence joue souvent un rôle majeur dans la détermination du niveau de compromis entre précision prédictive et la précision descriptive. Par exemple, lorsque l'interprétabilité est nécessaire pour s'assurer du caractère équitable des prédictions du modèle, la précision descriptive sera plus favorisée. En revanche, la précision prédicitve sera mise en avant dans des situations où la connaissance du processus sous-jacent n'est pas prioritaire.

À présent que le cadre PDR d'évaluation des méthodes d'interprétation de modèles d'apprentissage automatique a été posé, les deux sections suivantes présenteront successivement les propriétés caractéristiques des méthodes d'interprétation basées sur le modèle, et celles des méthodes *post*

[2]Par exemple l'agriculteur aimerait savoir avec quelle probabilité il gèlera dans sa zone d'activité au prochain printemps ; le météorologue quant à lui est à la recherche des inférences causales pour identifier les facteurs climatiques favorisant la formation de gel.

hoc. Cependant, avant d'y arriver, commençons par présenter les spécificités de ces deux classes de méthodes d'interprétabilité selon le cadre PDR.

La figure 3.2 synthétise bien ces spécificités. Les méthodes basées sur le modèle et *post hoc* visent toutes les deux à améliorer la prédiction descriptive du modèle, cependant, seules les méthodes basées sur le modèle sont susceptibles d'affecter la précision prédictive. En effet, l'interprétabilité basée sur le modèle implique l'utilisation d'un modèle plus simple pour ajuster les données, ce qui peut parfois affecter négativement la précision prédictive, mais donne une précision descriptive plus élevée. L'interprétabilité post hoc implique l'utilisation de méthodes pour extraire des informations d'un modèle entraîné (sans effet sur la précision prédictive).

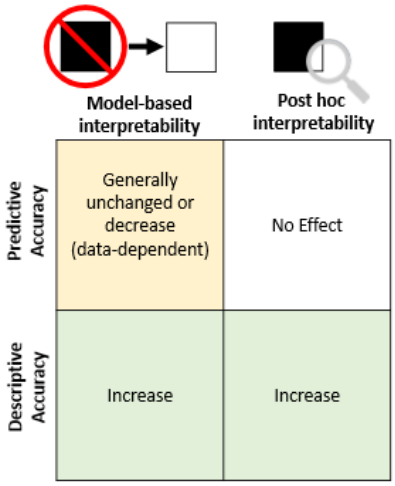


FIGURE 3.2 : *Impact des méthodes d'interprétabilité sur les précisions descriptives et prédictives (issue de l'article Murdoch* et al. *(2019)).*

3.3 Interprétabilité basée sur le modèle (IBM)

À présent nous allons examiner comment les considérations d'interprétabilité entre en jeu dans l'étape de modélisation du cycle de vie d'un projet d'apprentissage automatique (cf. figure 3.1). Comme mentionné dans l'article de Murdoch *et al.* (2019), le principal défi de l'interprétabilité basée sur le modèle consiste à proposer des modèles suffisamment simples (de par leur architecture) pour être facilement compris par un grand public, tout en restant suffisamment sophistiqués pour ajuster correctement le processus sous-jacent. Par ailleurs, la sélection d'un modèle doit tenir compte de l'ensemble du cadre PDR. La précision prédictive doit être prioritaire car elle conditionne la fiabilité des deux autres critères : si elle est mauvaise, toute l'analyse ultérieure sera suspecte (Breiman (2001) ; Freedman (1991)). Toutefois, l'objectif principal de ces méthodes reste la recherche de précision descriptive.

Présentons maintenant quelques propriétés souhaitées des méthodes d'interprétabilité basées sur le modèle.

3.3.1 Sparsité ou parcimonie

Lorsque le praticien estime que le processus d'intérêt qu'il souhaite comprendre est fonction d'un ensemble épars de signaux (caractéristiques), il peut imposer la sparsité à son modèle en

limitant le nombre de paramètres non nuls aux variables les plus "importantes"[3] dans l'ajustement du modèle.

Le principe de sparsité est étroitement lié à la théorie du rasoir d'Ockham[4] : *principe philosophique qui stipule que la solution la plus simple est souvent la meilleure*. Une fois prise en considération, la parcimonie permet d'obtenir une meilleure précision descriptive, car généralement, plus c'est simple, mieux c'est compréhensible. Par ailleurs, la sparsité pourrait également permettre d'améliorer la précision prédictive, par exemple en évitant le problème de surapprentissage d'une part ; et d'autre part, en fournissant des informations plus pertinentes (en effet, trop d'informations peut parfois tuer l'information ; il faut alors veiller à se limiter à celles qui sont vraiment utiles).

En apprentissage automatique, les méthodes d'incorporation de la sparsité dans un modèle varient d'une famille de modèles à l'autre.

(i) En ce qui concerne les modèles linéaires et linéaires généralisés, les méthodes permettant d'obtenir la sparsité utilisent le plus souvent :

– La régularisation de type Lasso ;

– Ou encore les algorithmes de sélection de variables tels que le *forward-stepwise* ou le *backward-stepwise* basés sur la minimisation des critères AIC, BIC, ou la maximisation du R^2-ajusté (pour les modèles linéaires gaussiens), Hastie *et al.* (2009).

– Ou encore une approche basée sur des méthodes combinatoires (« branch-and-bound »), confère le package *leaps* sur R. Cette approche a connu des progrès récents, Bertsimas et Dunning (2016) utilisant une approche de "mixed integer linear programming".

(ii) En ce qui concerne les autres modèles complexes d'apprentissage automatique (par exemple, les réseaux de neurones, il est plus difficile d'introduire intrinsèquement la sparsité au modèle. Ainsi, pour repérer les variables les plus "essentielles", les praticiens utilisent généralement en amont :

– Soit un modèle linéaire afin d'identifier le sous-groupe de variables les plus significatives ;

– Ou alors un arbre de classification/régression (CART) qui permet hiérarchisation des variables explicatives selon leur pouvoir de discrimination mesuré par le critère de *l'importance d'une variable* (c.f. Malot-Tuleau (2006)).

Toutefois, il faut noter que dans les deux cas, il s'agit de méthodes approchées, notamment celui de l'utilisation du critère de l'importance des variables. En effet, l'utilisation de ce dernier présente trois limites majeures : il se pose le problème du seuillage de l'importance des variables (quelles sont les variables que l'on va conserver et celles que l'on va éliminer ?) ; D'autre part, la notion d'importance des variables ne prend pas en compte la corrélation entre les variables. Par conséquent, même en fixant un seuil à l'importance des variables, on ne retient pas obligatoirement le plus petit paquet de variables discriminantes ; Enfin, l'instabilité de l'algorithme CART induit également l'instabilité de l'importance des variables.

(iii) Il existe également des approches agnostiques au modèle, comme par exemple, l'algorithme CSA (contribution-selection algorithm), qui est un algorithme de sélection des variables basée sur les valeurs de Shapley en Théorie des Jeux. Cet algorithme fut introduit dans l'article de Cohen *et al.* (2007). L'algorithme CSA estime de manière itérative l'utilité des caractéristiques et les sélectionne en conséquence, en utilisant soit la sélection forward, soit l'élimination backward. Il

[3] Importante au sens de leur potentialité à améliorer la précision prédictive.

[4] Le rasoir d'Ockham tient son nom du frère franciscain anglais Guillaume d'Ockham (v. 1285 - 9 avril 1347), philosophe et logicien.

peut optimiser diverses mesures de performance sur des données non vues telles que l'AUC, MSE, MAE. Une comparaison empirique avec plusieurs autres méthodes de sélection de caractéristiques existantes montre que la variante d'élimination à rebours (backward) de la CSA conduit aux résultats de classification les plus précis sur un ensemble d'ensembles de données (Cohen *et al.* (2007)).

3.3.2 Simulabilité

D'après Murdoch *et al.* (2019) un modèle est dit "simulable" si un humain à qui l'interprétation est destinée est capable de simuler l'ensemble de son processus de prise décision, et ce, dans un laps de temps "raisonnable". Cette contrainte imposée au modèle ne peut être possible que lorsque le nombre de caractéristiques est faible, c'est-à-dire lorsque le modèle est parcimonieux. Les modèles de *règles de décision* et les *arbres de décision* sont généralement considérés comme simulables en raison de leur processus de décision hiérarchique. Les modèles linéaires généralisés sont également considérés comme tels en raison de la modélisation des prédictions sous forme de somme pondérée qui rend transparent la façon dont les prédictions sont produites. La simulabilité accroît considérablement la précision descriptive des méthodes IBM et est étroitement liée à la sparsité et à la structure du modèle considéré. Ainsi, les méthodes de parcimonie (Lasso, forward-stepwise, etc.) amélioreraient, au moins indirectement, la simulabilité.

Toutefois, étant donné que la notion de simulabilité impose une limite "raisonnable" de temps de calcul par l'humain et compte tenu des limites de la cognition humaine, nous aboutissons donc au résultat qu'aucune famille de modèle d'apprentissage automatique n'est intrinsèquement simulable (ni les modèles linéaires, ni les règles de décision, ni les arbres de décision). En effet, dès lors que la dimension devient suffisamment élevée, les listes de règles de décision deviennent peu maniables, de même pour les arbres de décision profonds. On peut également citer le cas des modèles GLM couramment utilisés en tarification actuarielle, où le nombre de coefficients peut parfois être de l'ordre des centaines, rendant ainsi compliquée la simulabilité, donc l'interprétation.

3.3.3 Modularité ou décomposabilité

D'après l'article de Murdoch *et al.* (2019) un modèle est *modulaire* si une ou plusieurs parties significatives de son processus de prédiction peuvent être interprétées indépendamment. Quant à Lipton (2018), un modèle est modulaire si chaque partie dudit modèle admet une explication intuitive. Cette caractéristique de modularité correspond à la propriété d'*intelligibilité* décrite par Lou *et al.* (2012). Par exemple, chaque noeud d'un arbre de décision peut correspondre à une description en texte clair et plus ou moins pertinent. En outre, les modèles additifs généralisés satisfont généralement la propriété de modularité. Toutefois, nous devons utiliser cette notion de décomposabilité avec beaucoup de recul : en effet, les poids d'un modèle linéaire peuvent sembler intuitifs, mais ils peuvent parfois être instables (biais de variables omises ou incluses).

Cette propriété accroît considérablement la performance descriptive et pertinence des méthodes IBM. Pour l'illustrer, reprenons l'exemple du traitement de la pneumonie. Le modèle entraîné a appris que le fait d'être asthmatique est associé à un risque plus faible de mourir d'une pneumonie. Pourtant en réalité, c'est l'inverse qui se produit : on sait que les patients asthmatiques ont un risque plus élevé de mourir d'une pneumonie. Raison pour laquelle, dans les données recueillies, tous les patients asthmatiques ont reçu des soins agressifs, ce qui a heureusement permis de réduire leur risque de mortalité par rapport à la population générale ; Ainsi, si ce modèle d'apprentissage automatique avait été utilisé sans interprétation les patients atteints asthmatiques atteints de pneumonie auraient été dépriorisés pour le traitement, ce qui aurait accru leur probabilité de mou-

rir. Heureusement, l'utilisation d'un modèle modulaire (arbre de décision) a permis aux chercheurs d'identifier et de corriger des erreurs comme celle-ci.

3.3.4 Ingénierie des caractéristiques

Le fait de disposer de caractéristiques plus informatives simplifie la relation qui doit être apprise par le modèle, ce qui accroît le plus souvent la précision descriptive et même prédictive tout en améliorant la pertinence des interprétations. Il existe une variété d'approches pour construire des caractéristiques interprétables. Les unes reposent sur l'expertise existante du praticien dans le domaine : dans de nombreux domaines individuels, les connaissances des experts peuvent être utilisées pour construire des ensembles de caractéristiques utiles à l'élaboration de modèles prédictifs, descriptifs et pertinents.

Les autres approches d'ingénierie des caractéristiques reposent sur les idées tirées de l'analyse exploratoire des données et méthodes d'apprentissage non supervisé telle que le regroupement (*clustering*). En outre, les méthodes de réduction de la dimensionnalité sont aussi des outils efficaces d'ingénierie de caractéristiques. Elles se concentrent sur la recherche d'une représentation des données qui soit moins dimensionnelle que les données originales (ce qui renoue avec la propriété de parcimonie évoquée plus haut). Parmi les méthodes de réduction de la dimension, les plus répandues sont l'analyse en composantes principales, l'analyse des correspondances multiples et l'analyse des corrélation canonique (Saporta (2006) ; Escofier et Pagès (1998)). Ainsi, l'utilisation de caractéristiques intuitives et en nombre réduit peut permettre non seulement d'améliorer la précision descriptive et la pertinence des interprétations, mais aussi d'accroître l'efficacité prédictive en réduisant le nombre de paramètres à ajuster, ce qui limite le risque de sur-apprentissage.

3.4 Interprétabilité post hoc

Comme le montre la figure 3.1 du cycle de vie d'un projet d'apprentissage automatique, l'interprétabilité post hoc, contrairement à l'interprétabilité basée sur le modèle intervient une fois le modèle ajusté et les prédictions faites.

Un avantage de cette méthode d'interprétabilité est que nous pouvons interpréter des modèles opaques, sans avoir à sacrifier la performance prédictive (confère figure 3.2). En effet, les méthodes d'interprétation post hoc se concentrent essentiellement sur l'amélioration de la précision descriptive et de la pertinence. Pour se faire, une variété de méthodes d'interprétabilité post hoc ont été développées pour décrire fidèlement et pertinemment le modèle ajusté et les résultats de prédictions.

Comme le souligne Murdoch *et al.* (2019), les méthodes d'interprétation *post hoc* se divisent en deux catégories principales : les interprétations au niveau de la prédiction (interprétations locales) et au niveau de l'ensemble de données (interprétations globales). Les interprétations locales se concentrent sur l'explication de quelques prédictions individuelles faites par le modèle, telles que les caractéristiques et les interactions qui ont conduit à la prédiction particulière.

Ainsi, les explications générées ne sont valables que pour l'instance x en question et son "voisinage proche". Les approches globales quant à elles se focalisent sur les relations générales apprises par le modèle. Toutefois, ces deux approches demeurent étroitement liées, en ce sens que les informations utiles extraites au niveau global fournissent le plus souvent des indications au niveau local.

Dans cette section, nous présentons globalement et succinctement les formes d'interprétabilité post hoc, aussi bien à l'échelle globale que locale. Dans le chapitre suivant, nous reviendrons sur

les détails des méthodes les plus sollicitées en pratique.

3.4.1 Interprétation globale ou interprétation au niveau du jeu de données

L'interprétation au niveau global s'intéresse aux relations générales apprises par un modèle. Par exemple, les relations qui sont pertinentes pour une classe spécifique d'entrées ou une sous-population particulière. Les approches d'explications globales, peuvent être subdivisées en deux sous-classes : les méthodes d'interprétation par *substitution* et les méthodes de *génération* d'interprétations.

3.4.1.1 Modèles de substitution globaux

Comme mentioné dans Henelius *et al.* (2014), un modèle de substitution traduit le modèle initial en un modèle approximatif. Cette technique est utilisée lorsque le modèle initial n'est pas interprétable par lui-même, c'est-à-dire chaque fois qu'il s'agit d'une boîte noire.

Le principe de ces méthodes est simple et intuitif : un modèle interprétable est construit par-dessus la boîte noire de manière à mimer toutes les prédictions de la boîte noire avec une grande précision prédictive, ce qui permet de comprendre la boîte noire. Le processus d'ajustement d'un modèle global de substitution est illustré par la figure 3.3. Le modèle de substitution peut être appliqué à une classe de modèles spécifique ou il peut être agnostique au modèle.

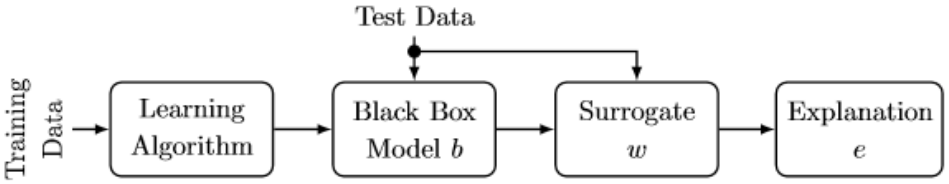


FIGURE 3.3 : *Processus d'ajustement d'un modèle de substitution global ω (issue de Burkart et Huber (2021)).*

Les classes de modèles de substitution couramment utilisées dans la pratique s'appuient généralement sur :

– Le modèle de régression linéaire ou linéaire généralisé, et leurs variantes (avec régularisation Lasso, ridge, etc.) ;

– Les modèles arbres de décision (algorithme CART) ;

– Les modèles de règles de décision.

Les méthodes de substitution basées sur les règles de décision sont subdivisées en trois approches : l'approche pédagogique (indépendante du modèle), l'approche de décomposition (spécifique au modèle) et l'approche hybride combinant les deux premières.

L'approche pédagogique perçoit le modèle de prédiction comme une boîte noire et utilise la relation entre les entrées et les sorties pour extraire des règles de décision interprétables. L'approche de décomposition quant à elle utilise la structure interne du modèle de prédiction sous-jacent pour en extraire un modèle de règles de décision facilement interprétable (confère figure 3.4). Une vue d'ensemble des différentes méthodes de substitution globale actuellement disponibles dans la littérature est présentée dans le tableau B.1 (en annexe).

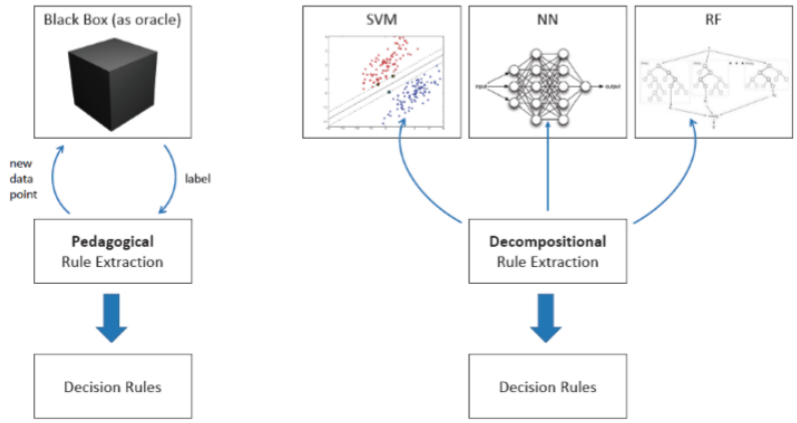


FIGURE 3.4 : *Différence entre les approches d'extraction de règles pédagogiques et décompositionnelles (issue de Burkart et Huber (2021)).*

3.4.1.2 Méthodes de génération d'explications globales

Dans cette section, nous décrivons les approches qui peuvent directement générer une explication globale. La différence avec les modèles de substitution est que l'explication est directement déduite du modèle de la boîte noire sans l'intermédiaire d'un substitut global, comme le montre la figure 3.5.

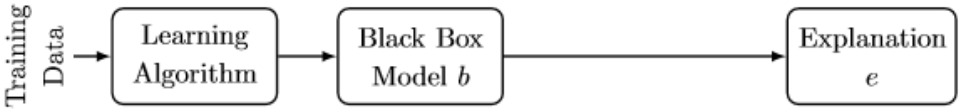


FIGURE 3.5 : *Processus de génération d'explication globale e (issue de Burkart et Huber (2021)).*

□ Interaction et importance des caractéristiques

Le score d'importance d'une variable au niveau d'un modèle tente de capturer la contribution de la variable, dans un ensemble de données, à la formation des prédictions. Dans la littérature, diverses méthodes ont été développées pour calculer le score d'importance des variables dans de nombreux modèles, notamment dans les *arbres de décision* et *forêts aléatoires* avec le développement des *RF Feature Importance* (Hall *et al.* (2017)).

Cette notion d'importance spécifique aux modèles utilisant les *arbres de décision*, tels que les *forêts aléatoires* avait déjà été introduit plutôt par Breiman (2001). Par ailleurs, d'autres méthodes ont été développées pour les réseaux de neurones par Olden *et al.* (2004) mais également des méthodes indépendantes au modèle ont été mises en place (nous abordons les plus usuelles dans la sous-section 4.2.2 du chapitre 4).

Une autre notion importante est la mise en évidence des interactions entre les caractéristiques. Cette notion est d'autant importante en ce sens que les modèles complexes d'apprentissage automatique imbriquent généralement plusieurs relations non linéaires entre-elles et apprennent ainsi des interactions complexes entre les caractéristiques.

À cet effet, pour extraire les interactions importantes entre les caractéristiques, les chercheurs et praticiens ont développées de nombreuses méthodes spécifiques aux différentes familles de modèles d'apprentissage automatique. Par exemple, pour les forêts aléatoires (Kumbier *et al.* (2018) ; Basu *et al.* (2018)) et les réseaux de neurones (Tsang *et al.* (2017) ; Abbasi-Asl et Yu (2017)). Par

ailleurs, il existe également plusieurs autres méthodes de détection des interactions indépendantes du modèle : par exemple, la H-statistique de Friedman (Friedman et Popescu (2008)) ou l'indice de Sobol (Sobol' (1990)). Une présentation plus détaillée de ces dernière méthodes sera faite dans la sous-section 4.2.3 du chapitre 4.

□ **Visualisation**

Lorsqu'on travaille sur des ensembles de données de grande dimension, il est parfois difficile de comprendre rapidement les relations complexes apprises par le modèle. Dans de pareilles circonstances, la présentation des résultats joue un rôle crucial pour une meilleure compréhension.

Les chercheurs ont développés un certain nombre d'outils de visualisations différents qui aident à fluidifier la compréhension des relations apprises par un modèle. Par exemple, en ce qui concerne les modèles linéaires ou linéaires généralisés régularisés (Lasso, Ridge), le graphique des trajectoires des coefficients de régression montre comment la variation du paramètre de régularisation affecte les coefficients ajustés. En ce qui concerne les réseaux de neurones, il existe également plusieurs outils spécifiques à leur visualisation. De nombreuses autres méthodes de visualisation agnostiques au modèle ont été développées par des chercheurs. Entre autres, nous avons les graphiques de dépendance partielle (PDP), qui présentent la prédiction moyenne de la boîte noire lorsqu'une seule variable varie dans sa plage de valeurs possibles. L'idée sous-jacente étant de montrer comment les variations de la variable concernée affectent la prédiction du modèle en général (Goldstein *et al.* (2015)) ; nous avons également les graphiques des effets locaux accumulés (ALE) qui sont développés dans la sous-section 4.2.1 du chapitre 4.

□ **Analyses des prédictions et des résidus empiriques**

Dans la pratique, la précision prédictive des modèles d'apprentissage automatique est généralement évaluée à partir de métriques telles que le taux d'erreur pour les problèmes de classification et l'erreur quadratique moyenne pour les problèmes de régression. Cependant, ces métriques renseignent uniquement sur la performance moyenne des modèles.

Dans certaines situations, il peut être utile de creuser davantage en examinant non seulement la précision moyenne, mais aussi la distribution des prédictions afin de peaufiner notre compréhension du modèle. L'analyse des erreurs peut recéler beaucoup d'autres informations. Par exemple, lorsqu'on a à choisir parmi plusieurs modèles, le tracé des graphiques de résidus en fonction des prédictions et/ou des variables explicatives peut permettre d'identifier le modèle qui ajuste le mieux le phénomène : le modèle le mieux ajusté serait celui pour lequel le nuage des points {prédiction, résidus} ou {variables explicatives, résidus} est le mieux dispersé, c'est-à-dire, celui qui révèle le mieux une absence de tendance apparente sur les graphiques.

3.4.2 Interprétation locale ou interprétation au niveau des prédictions

Les explicateurs locaux permettent la deuxième manière de générer des explications. Le principe est le suivant : étant donné une prédiction d'un modèle (boîte noire ou interprétable), l'explicateur local fournit des informations relatives aux processus de prédiction qui ne sont valables que pour l'instance particulière concernée et ne peuvent pas être généralisées à l'ensemble du modèle.

Comme les méthodes d'interprétation au niveau du modèle (globales), les méthodes post hoc locales peuvent également être subdivisées en deux sous-classes : celle constituée des méthodes de substitution locales, et celle des méthodes de génération d'explications locales.

L'ensemble de ces méthodes visent, d'une manière ou d'une autre à quantifier la contribution des caractéristiques responsables de la prédiction que nous souhaitons expliquer.

Notons que l'importance des variables au niveau local est généralement plus informative que les scores d'importance de variables au niveau global. En effet, les modèles non-linéaires sont par nature hétérogènes, ainsi, la contribution d'une variable peut varier d'une instance à l'autre en raison des potentielles interactions avec d'autres variables. Par conséquent, les méthodes locales permettent d'accroître davantage la précision descriptive des modèles opaques et peuvent souvent servir de d'indicateur pour examiner des critères difficilement quantifiables tels que l'équité, la causalité, la fiabilité, etc. du modèle boîte noire.

3.4.2.1 Modèles de substitution locaux

La figure 3.6 illustre sommairement le processus de mise en place d'un modèle de substition local.

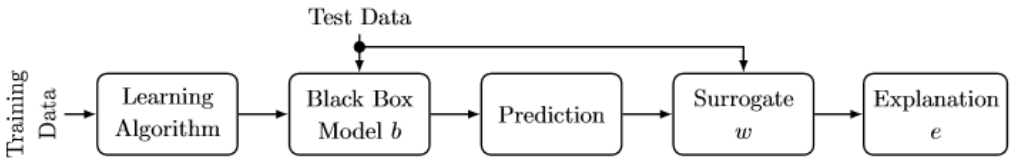


FIGURE 3.6 : *Processus d'ajustement d'un modèle de substitution local ω (issue de Burkart et Huber (2021)).*

Les approches locales de substitution reposent essentiellement sur les modèles interprétables tels que les *arbres et règles de décision*, le *modèle linéaire ou linéaire généralisé*.

Parmi les plus répandus actuellement, on peut citer les méthodes LIME (Local Interpretable Model-Agnostic Explanations) et SHAP (SHapley Additive exPlanation). LIME a été introduite par les auteurs Ribeiro *et al.* (2016). Ils essaient de décrire une prédiction particulière faite par n'importe quel modèle de classification ou de régression.

La prédiction à interpréter est alors représentée sous la forme d'un modèle d'explication, par exemple un modèle linéaire, un arbre de décision ou une liste de règles.

Pour illustrer un cas d'usage de la méthode LIME, considérons un modèle de forêt aléatoire ajusté sur les données portant sur le [diabète] de Kaggle. Pour un patient particulier le modèle prédit qu'il sera diabétique avec une probabilité de 73%. Ainsi, à l'aide de la méthode LIME illustré sur la figure 3.7 ci-dessous, l'on retrouve clairement que les raisons principales de cette prédiction sont que le taux de glucose du patient en question est supérieur à 99 et sa tension artérielle est supérieure à 70 −sur la droite, nous pouvons observer les caractéristiques réelles du patient.

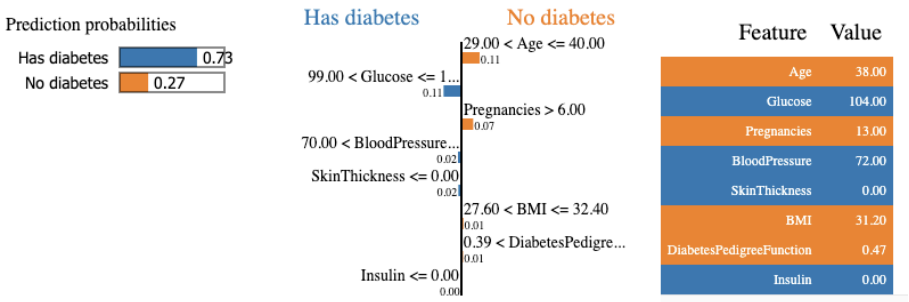


FIGURE 3.7 : *Explicabilité de LIME pour une seule instance*

L'approche SHAP et ses variantes (kernelSHAP ou linearSHAP) dûe à Lundberg et Lee (2017), vise à extraire la contribution des caractéristique pour une prédiction donnée. En entrée, elle

prend le modèle ajusté et l'instance pour laquelle nous voulons une explication. Comme mentionné dans l'article Burkart et Huber (2021), les contributions des variables (encore appelées *valeurs de shapley*) sont calculées en marginalisant tour à tour sur chaque variable pour analyser comment le modèle se comporte en son absence. Un avantage de l'approche SHAP est qu'elle repose sur une solide base mathématique (la théorie des jeux coopératifs) et est assez intuitive, bien que souvent coûteuse en temps de calcul pour des problèmes de grande dimension. Une présentation plus détaillée de cette approche sera faite dans le chapitre 4.

3.4.2.2 Approches de génération d'explications locales

La figure 3.8 résume la démarche de mise en oeuvre d'un processus de génération d'interprétations locales.

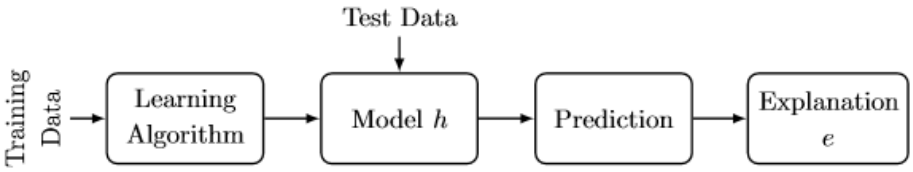


FIGURE 3.8 : *Processus de génération d'explication locale e (issue de Burkart et Huber (2021))*

□ Méthodes d'évaluation de la saillance par attribution d'un score d'importance aux caractéristiques

Comme mentionné dans Burkart et Huber (2021), les méthodes de saillance relient en général une prédiction particulière du modèle au vecteur de caractéristiques, en rangeant ces dernières suivant un ordre décroissant ou croissant de leur pouvoir explicatif.

Il existe une multitude de méthode de saillance, certaines étant spécifiques aux classes particulières de modèles complexes et d'autres indépendantes du modèle. Cependant, dans le chapitre 4, un accent particulier sera mis sur les diagrammes d'espérance conditionnelle individuelle (en anglais, *Individual Conditional Expectation-ICE*) qui sont une extension des PDP couramment utilisées en pratique Goldstein *et al.* (2015).

□ Autres méthodes de génération d'explications locales

Outre l'évaluation de la saillance, il existe d'autres méthodes de générations d'interprétations locales. Par exemple :

(i) les *méthodes contrefactuelles* : pour une prédiction que l'on veut expliquer, les méthodes contrefactuelles consistent à rechercher dans l'espace des caractéristiques, une instance "proche" de cette prédiction du point de vue des caractéristiques d'entrée, mais qui mène à la prédiction d'une classe ou valeur différente. Cependant, les instances contrefactuelles souffrent le plus souvent d'un faible degré d'interprétabilité, car pour une seule prédiction, il peut exister une multitude de contrefactuelles différents, surtout dans des contextes de haute dimensionnalité. Il devient alors difficile de synthétiser l'information.

(ii) *Méthodes d'explication basées sur les ancres* : La méthode des ancres explique les prédictions individuelles de tout modèle de classification boîte noire en trouvant une règle de décision qui "ancre" suffisamment la prédiction. On dit qu'une règle ancre une prédiction si les changements dans les valeurs des autres caractéristiques en dehors de celles qui constituent la règle n'affectent pas la prédiction. Nous développons cette méthode d'interprétabilité locale agnostique au modèle plus bas, dans le chapitre 4.

En définitive, il ressort que la notion d'interprétabilité est complexe à définir et il existe un large éventail de méthodes d'interprétation de modèles. Le cadre PDR présenté dans ce chapitre fournir trois *désidératas* primordiaux des méthodes d'interprétabilité : la précision prédictive, l'exactitude descriptive et la pertinence, la pertinence étant jugée par rapport à un public humain précis à qui est destiné les explications.

En outre, pour mieux structurer le déluge de méthodes d'interprétation, nous avons présenté une catégorisation des techniques existantes en deux grandes catégories : les méthodes d'interprétation basées sur le modèle (IMB) et les méthodes d'interprétation post hoc.

Chacune de ces deux catégories sont à leur tour subdivisées en sous-catégories. Dans le chapitre suivant, nous présentons plus en détails les méthodes d'interprétabilité post-hoc les plus populaires à l'heure actuelle aussi bien dans le monde de l'entreprise que dans celui de la recherche.

Chapitre 4

Méthodes d'interprétabilité post hoc, agnostiques au modèle

Les méthodes d'interprétation des modèles d'apprentissage statistique sont nombreuses. Elles sont regroupées en deux catégories :

- Les méthodes d'interprétation basées sur l'architecture du modèle ;
- Les méthodes d'interprétation *post hoc*, basées sur les prédictions issues du modèle.

Les méthodes basées sur le modèle visent à interpréter le modèle en s'appuyant sur son architecture. Les méthodes *post hoc* quant à elles visent à interpréter les mécanismes sous-jacents au modèle en partant des prédictions générées par le modèle. Compte tenu de la complexité de l'architecture des modèles d'apprentissage statistique "récents" (Random Forest, XGboost, etc.), l'approche d'interprétation *post hoc* est la mieux appropriée pour leur interprétation.

Dans ce chapitre, nous présenterons les méthodes d'interprétation *post hoc*, comme une panacée à l'utilisation des modèles complexes d'apprentissage statistique récents, notamment dans des domaines sensibles tel que celui des assurances.

4.1 Catégorisation de méthodes d'interprétation post hoc

4.1.1 Grandes catégories des méthodes d'interprétation post hoc

L'ensemble des méthodes d'interprétation *post hoc* existantes est vaste. Pour faciliter leur présentation Burkart et Huber (2021) proposent de les catégoriser suivant trois dimensions (confère figure 4.1) :

– *Les méthodes locales vs les méthodes globales* : dans le premier cas l'interprétabilité n'est valable que localement pour une seule instance de données et extensible dans un voisinage restreint de cette instance de données. Or, avec les méthodes globales l'interprétation est valable pour l'ensemble du modèle ;

– *Les méthodes d'interprétation spécifiques au modèle vs les méthodes d'interprétation agnostiques (indépendantes) au modèle* : dans la première famille, il s'agit des méthodes qui ne fonctionnent que pour une classe fermée de modèles, et dans la seconde famille, il s'agit des méthodes applicables à toutes les classes de modèles (ou presque) ;

– *Les méthodes indépendantes des données vs les méthodes dépendantes de données* : dans le premier cas, le mécanisme pour générer des interprétations ne nécessite pas de données supplémentaires, contrairement aux méthodes du second cas.

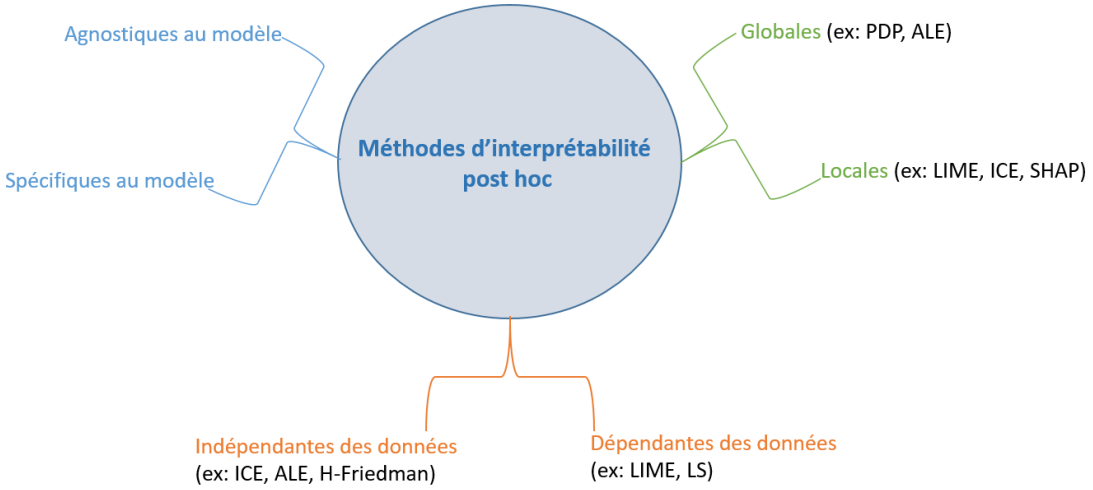


FIGURE 4.1 : *Catégorisation des méthodes d'interprétabilité post hoc selon trois dimensions.*

4.1.2 Méthodes d'interprétation explorées dans ce mémoire

Dans ce mémoire, nous nous limitons à la présentation des méthodes d'interprétation *post hoc*, agnostiques au modèle (aussi bien globales que locales, indépendantes aux données que dépendantes des données). Notre restriction à cette catégorie de méthodes (méthodes *post hoc et agnostiques au modèle*) se justifie tout simplement par le fait que ces méthodes permettent d'interpréter un large éventail de modèles d'apprentissage statistique. Rendant, ainsi possible la comparaison de plusieurs modèles sous la base de mêmes outils.

Les méthodes d'interprétation présentées dans ce chapitre seront mis en oeuvre dans le cas d'application du chapitre 5 pour interpréter les modèles complexes de fréquence de sinistre automobile mis en place.

4.2 Méthodes d'explication globales

Dans cette section, nous présentons les méthodes d'interprétation globales répandues dans la littérature en les structurant en trois sous-catégories, selon l'information qu'elles permettent d'extraire dans le modèle :

– Tout d'abord, les méthodes permettant d'évaluer *l'effet des caractéristiques* sur la variable à prédire ;

– Ensuite, celles permettant de quantifier *l'importance des caractéristiques* dans la formation des prédictions du modèle ;

– Enfin, les méthodes permettant d'analyser la *force d'interaction entre les caractéristiques* dans la formation des prédictions du modèle.

Notons que ces trois sous-catégories de méthodes d'interprétabilité ne sont pas mutuellement exclusives. Il existe des méthodes d'interprétation qui permettent à la fois d'évaluer l'effet des variables, de mesurer la force d'interaction entre les caractéristiques et même de calculer l'importance des variables dans le modèle. C'est notamment le cas des diagrammes ALE qui peuvent servir à mesurer l'effet des caractéristiques sur la variable prédite tout en faisant ressortir la force d'interaction entre ces caractéristiques dans la prédiction de la variable cible.

4.2.1 Analyse de l'effet global des caractéristiques sur la variable cible : PDP, ALE-Plot

Dans de nombreux domaines d'application de l'apprentissage supervisé, notamment en actuariat, comprendre et visualiser les effets des variables prédictives sur la réponse prédite est d'une importance capitale. C'est notamment le cas lorsque la finalité de la modélisation est explicative plutôt que prédictive. Même lorsque l'objectif de l'apprentissage est purement prédictif, la compréhension des effets des prédicteurs sur la variable cible peut être assez importante pour s'assurer de la fiabilité du modèle ou pour identifier des phénomènes contre-intuitifs.

Afin d'analyser l'effet global d'une caractéristique sur la variable prédite, plusieurs méthodes ont été développées par des chercheurs. Les plus populaires d'entre elles sont la méthode PDP et la méthode ALE. Elles fournissent une visualisation graphique des relations apprises par le modèle entre les différentes variables explicatives et la variable cible.

4.2.1.1 Graphiques de dépendances partielles (PDP)

□ Principe général et formalisation mathématique

Les graphiques de dépendances partielles (en anglais *Partial Dependences Plots−PDP*) sont des outils particulièrement utiles dans la compréhension des relations apprises par un modèle d'apprentissage supervisé. Ils ont été introduit par le statisticien américain Friedman (2001).

Le principe est simple : le PDP trace le changement de la valeur moyenne prédite lorsque la ou les caractéristiques spécifiées varient sur leur distribution marginale. Autrement dit, le graphique de dépendance partiel (PDP) a pour objectif de montrer l'effet marginal d'une ou de plusieurs variables explicatives (généralement pas plus de deux) sur la prédiction faite par le modèle.

Désignons par $X = (X_j)_{1\leq j\leq p}$ le vecteur des caractéristiques et Y la variable d'intérêt à prédire, le tout défini sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ et dont une réalisation échantillonnale est donnée par les n observations $\left(x^{(i)}, y^{(i)}\right)_{1\leq i\leq n}$ (avec n, $p \in \mathbb{N}^*$). Soit f la vraie fonction d'approximation de Y à partir des prédicteurs X et $\hat{f}$ son estimateur. Soit $S \subset \{1, \cdots, p\}$ le sous-ensemble des indices de caractéristiques dont on veut analyser l'effet sur les prédictions, et C son complémentaire ($C = \{1, \cdots, p\} \setminus S$). La fonction de dépendance partielle de f en les caractéristiques S est définie par :

$$f_S(x_S) = \mathbb{E}_{X_C}[f(x_S, X_C)] = \int f(x_S, X_C) d\mathbb{P}(X_C), \quad \text{pour tout } x_S \in \mathcal{X}_S \tag{4.1}$$

Ainsi, à chaque sous-ensemble de prédicteurs S est associée une fonction de dépendance partielle notée f_S. Cependant, étant donné que dans la pratique ni le vrai modèle f ni la vraie distribution jointe $d\mathbb{P}(X_C)$ des X_C n'est généralement connu la fonction de dépendance partielle définie par l'équation (4.1) est approximée par :

$$\hat{f}_S(x_S) = \hat{\mathbb{E}}_{X_C}[\hat{f}(x_S, X_C)] = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x_S, x_C^{(i)}\right), \quad \text{pour tout } x_S \in \mathcal{X}_S^{\mathcal{D}_a} \tag{4.2}$$

Où $\mathcal{X}_S^{\mathcal{D}_a}$ désigne l'ensemble des valeurs prises par le vecteur aléatoire X_S sur la base de données d'apprentissage $(\mathcal{D}_a)$.

Pour le sous-ensemble d'indices S de caractéristiques et un jeu de données d'entraînement $\left(x^{(i)}, y^{(i)}\right)_{1\leq i\leq n}$, l'approximation de la fonction de dépendance partielle $\hat{f}_S$ associée à S est évaluée pour chaque réalisation x_S de X_S observée au sein de l'échantillon de données d'entraînement. Il résulte un ensemble de n couples de la forme $\left\{\left(x_S^{(i)}, \hat{f}_S(x_S^{(i)})\right)\right\}_{1\leq i\leq n}$.

Ainsi, lorsque $cardinal(S) = 1$ *ou* 2, Friedman *et al.* (2001) propose de représenter le nuage de points constitués par ces n couples de valeurs. Le graphique résultant de la jointure des points du nuage est appelé graphique de dépendance partielle (PDP) associé au vecteur de caractéristiques X_S.

□ Mesure de l'importance des variables basée sur les PDP

À la page 1 217 de l'article originale introduisant les PDP, Friedman (2001) introduit déjà la notion *d'importance relative des caractéristiques* basée sur les fonctions de dépendances partielles. Son approche sera reprise quelques années plus tard par Greenwell *et al.* (2018). L'intuition de base est la suivante :

Un graphique de PDP plat indique que la variable (ou le groupe de variables) associé n'est pas important dans l'explication de la variable cible. À l'inverse, plus le graphique varie, plus la variable (ou le groupe de variables) associée est important.

Dans ce mémoire, nous ne développerons pas davantage la notion d'importance des caractéristiques basée sur les PDP. La raison est simple : en effet, la mesure de l'importance des caractéristiques basée sur les PDP est peu robuste dans lorsqu'il existe des interactions entre les caractéristiques, ce qui est généralement le cas en pratique.

□ Avantages des PDP

Les PDP possèdent de nombreux atouts. Entre autres, on peut relever :

– *Leur caractère intuitif* : la fonction de dépendance partielle associée à une variable X en un point $x \in X(\Omega)$ fixé, représente la prédiction moyenne que l'on obtiendrait si l'on contraignait toutes les observations de l'échantillon d'entraînement à posséder la valeur x pour la variable X, les autres caractéristiques restant inchangées.

– *La clarté des interprétations* : en cas d'absence d'interactions entre les caractéristiques, le PDP associé à une caractéristique (ou à un groupe de caractéristiques) montre parfaitement comment la prédiction moyenne de variable cible varie lorsque la caractéristique (ou le groupe de caractéristiques) est modifiée. L'interprétation des PDP correspond alors à une interprétation causale au niveau du modèle −même si pas nécessairement causale au niveau du monde réel. Les PDP renseignent donc clairement sur le comportement moyen du modèle.

– *Mise en oeuvre facile* : un autre avantage qu'on reconnaît aux PDP est qu'ils sont faciles à implémenter.

□ Inconvénients des PDP

Bien que les graphiques PDP aient gagnés en popularité depuis leur apparition, compte tenu de leur caractère assez intuitif, ils souffrent également de plusieurs limites :

– Tout d'abord, il repose sur l'hypothèse d'indépendance entre les prédicteurs. Ce qui n'est pas toujours vérifiée en pratique.

Supposons par exemple que l'on soit dans un contexte d'assurance automobile et que l'on veut prédire la fréquence de sinistre d'un assuré à partir de son âge et de l'ancienneté de son permis de conduire. Pour le calcul du PDP associé à l'âge du conducteur, pour un niveau d'âge fixé, par exemple 18 ans, on fait la moyenne des prédictions sous la distribution marginale de l'ancienneté du permis (suivant l'équation 4.2). Ce calcul pourrait inclure des permis d'ancienneté supérieure à 18 ans, ce qui est irréaliste pour une personne de 18 ans (car il n'est naturellement pas possible qu'un permis de conduire soit plus âgé que son détenteur).

Ainsi, lorsque les caractéristiques sont corrélées, dans le calcul des PDP nous faisons intervenir de nouveaux points de données irréalistes, ce qui peut conduire à des interprétations fallacieuses.

Cette incapacité des PDP à prendre compte les corrélations entre les caractéristique dans les interprétations fournies est corrigée par les graphiques des effets locaux cumulés (Accumulated Local Effects-ALE) que nous présenterons dans la sous-section suivante.

− Un deuxième inconvénients des PDPs est qu'ils peuvent dans certaines circonstances masquer les effets hétérogènes d'une caractéristique −puisqu'ils sont basés sur des calculs de moyennes. Les courbes ICE présentées plus bas dans ce chapitre permettent de palier à cette limite des PDP.

− Une autre limite moins cité que les deux précédentes est que les PDP sont fortement associés aux données d'entraînement et ne reposent pas sur des procédures de tests statistiques. Ce qui nécessite d'être très prudent quant à la généralisation des interprétations obtenues au monde réel.

4.1.1.2 Graphiques des effets locaux accumulés (ALE plots)

□ Motivation

Supposons que nous avons ajusté un modèle d'apprentissage supervisé pour approximer $\mathbb{E}(Y|X = x) = f(x)$, avec $X = (X_1, \cdots, X_p)$. Ici, pour simplifier la présentation des concepts, limitons-nous au cas $p = 2$. Nous souhaitons visualiser et comprendre la dépendance partielle des prédictions $\hat{y} = \hat{f}(x_1, x_2)$ suite aux variations marginales des valeurs x_1 ou x_2 des variables X_1 et X_2 respectivement. Nous pouvons faire recours aux PDPs abordés dans la section précédente. Cependant, comme évoqué plus haut, il se pose un véritable problème avec cette dernière approche lorsque les variables explicatives sont fortement corrélées. En guise d'illustration, supposons par exemple que les caractéristiques X_1 et X_2 soient telles que :

$$X_1 = U + \varepsilon_1,\ X_2 = U + \varepsilon_2 \tag{4.3}$$

$$\text{avec } U,\ \varepsilon_1,\ \varepsilon_2\ iid; U \sim U(0,1);\ \varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, 0.05^2)$$

Alors, $(X_1,\ X_2)$ suit une distribution uniforme le long du segment de droite $x_1 = x_2$, augmentée de deux bruits gaussiens. Ensuite 200 observations (X_1, X_2) sont générées à partir de ce modèle définie par le système d'équation (4.3). La fonction de dépendance partielle est calculée pour une valeur de $x_1 = 0.3$.

La figure 4.2a montre que le PDP dans le cas de données très corrélées effectue une extrapolation assez sévère au-delà de l'enveloppe des données d'apprentissage. Cela se met en évidence sur la figure par la densité marginale de X_2 qui est assez dispersée autour des données d'apprentissage (pointillés noir étalés le long de la première bissectrice).

Une alternative intuitive pour contourner ce problème d'extrapolation sévère serait de faire recours aux M-plots qui à la place de la densité marginale de X_2, utilise plutôt la densité de X_2 conditionnellement à $X_1 = 0.3$. Cette fois, le problème semble être corrigé. En effet, comme le montre la figure 4.2b, la densité conditionnelle est beaucoup plus concentrée autour des données d'apprentissage. Sauf qu'il se pose maintenant un problème beaucoup plus sérieux avec l'utilisation des M-plots : celui du biais de variable omise. Pour mieux illustrer le problème posé par le M-plot, commençons par présenter sa formulation mathématiquement.

Le M-plot de l'effet de X_1 correspond au tracé du graphe de la fonction de x_1 définie par :

$$f_{1,M}(x_1) = \mathbb{E}[f(X_1, X_2)|X_1 = x_1] = \int f(x_1, X_2)\, d\mathbb{P}(X_2|X_1 = x_1),\ \text{ pour tout } x_1 \in \mathcal{X}_1 \tag{4.4}$$

Une estimation empirique de $f_{1,M}$ est donnée par :

$$\hat{f}_{1,M}(x_1) = \frac{1}{n(x_1)} \sum_{i \in N(x_1)} f(x_1, x_2^{(i)}) \tag{4.5}$$

$N(x_1) \subset \{1, \cdots, n\}$ est le sous-ensemble d'indices des observations pour lesquelles leur valeur $x_1^{(i)}$ pour la variable X_1 tombe dans un petit voisinage de x_1 convenablement sélectionné, et $n(x_1)$ est le nombre total d'observations par les $x_1, \cdots, x_n$ qui tombent dans ledit voisinage.

Par l'équation (4.4) on observe bien que calculer la fonction $f_{1,M}(x_1)$ revient à régresser la variable cible Y sur X_1 tout en ignorant la variable X_2. Par conséquent, si Y dépend de X_1 et X_2, avec X_1 et X_2 fortement corrélées, $f_{1,M}(x_1)$ reflétera intrinsèquement l'effet des deux variables, pourtant ce qui nous intéresse c'est d'isoler uniquement l'effet de X_1 sur Y. Les M-plot sont donc inappropriés lorsque les caractéristiques sont fortement corrélées entre elles. Ils souffrent du biais de variable omise (confère D'Haultfoeuille (2021)).

Le graphique des effets locaux accumulés (ALE) introduit une nouvelle méthode d'évaluation des effets principaux et d'interaction des caractéristiques sur la variables cible qui surmonte les limites précédentes rencontrées avec les diagrammes PDP et les diagrammes M-Plot.

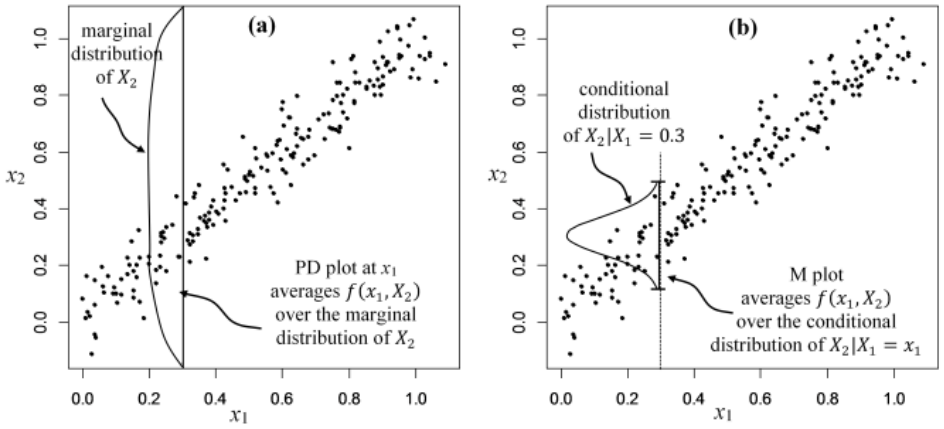


FIGURE 4.2 : *Illustration du calcul de (a) $f_{1,PD}(x_1)$ et (b) $f_{1,M}(x_1)$ à $x_1 = 0.3$ (issue de l'article de Apley et Zhu (2020)).*

□ Principe général de l'ALE

La méthode ALE a été introduit par Apley et Zhu (2020). Afin de simplifier la présentation du principe de l'ALE, sans nuire à la généralité considérons le cas $p = 2$ et supposons que la fonction f pour laquelle on souhaite calculer l'ALE est différentiable (c'est généralement le cas en pratique).

Le diagramme ALE d'une caractéristique X_1 s'obtient par le tracé d'une fonction définie sur $X_1(\Omega)$.

Pour une valeur $x_1^* \in X_1(\Omega)$ fixée, on procède comme suit :

– On calcule d'abord l'effet local de la caractéristique X_1 au voisinage de x_1^* sur la fonction de prédiction f. Pour ce faire on évalue la fonction $f^1 \equiv \frac{\partial f}{\partial x_1}$ en tout point $(x_1 = z_1, x_2)$, avec à chaque fois z_1 fixé et x_2 parcourant $X_2(\Omega)$;

– Par la suite, on calcule la moyenne pondérée de ces effets locaux suivant la distribution

conditionnelle $\mathbb{P}(X_2|X_1 = z_1)$:

$$\mathbb{E}\left(\frac{\partial f}{\partial x_1}(z_1,\, X_2)\right) = \int_\Omega \frac{\partial f}{\partial x_1}(z_1,\, X_2)\, d\mathbb{P}(X_2|X_1 = z_1) \tag{4.6}$$

– Enfin, nous intégrons les effets locaux moyens pour toutes les valeurs de z_1 inférieures ou égales à x_1^*.

$$\int_{-\infty}^{x_1^*} \mathbb{E}\left(\frac{\partial f}{\partial x_1}(z_1,\, X_2)\right) dz_1 \tag{4.7}$$

Le recours à la densité conditionnelle dans le calcul de l'effet local moyen évite l'extrapolation sévère requise dans les graphiques de PDP. En outre, en calculant les effets locaux moyens et en les accumulant jusqu'à x_1^*, par opposition au calcul de la moyenne directe de f conditionnellement à $X_1 = x_1^*$, l'ALE permet d'éviter le biais de variable omise contenu dans les M-plots en cas de forte corrélation entre les caractéristiques.

Notons dès à présent qu'en plus de calculer les effets de premiers ordre d'une caractéristique sur la variable dépendante, l'ALE permet également de calculer les effets de second ordre, et par conséquent peut servir d'outil de mesure de l'interaction entre les caractéristiques. Nous y reviendrons dans le section suivante.

□ Formalisation mathématique de l'ALE de premier et de second ordre

D'entrée de jeu, notons qu'il est possible de calculer les fonctions ALE d'ordre 3 ou d'ordre supérieure, cependant, elles ne sont pas couramment utilisées en pratique car elles sont difficilement représentables graphiquement. Pour cette raison nous nous limitons dans ce mémoire à la définition des effets de premier (ou effets principaux) et de second ordre.

• Notations préliminaires à la définition

– Nous posons $\mathcal{X} \equiv X(\Omega)$ et nous supposons que $\mathcal{X}$ est un compact de $\mathbb{R}^p$;
Pour chaque $j \in \{1, \cdots, p\}$, nous posons $\mathcal{X}_j \equiv X_j(\Omega) = [x_{min,\,j},\ x_{max,\,j}]$;

– Pour chaque $K = 1,\, 2, \cdots$, et $j \in \{1, \cdots, p\}$, nous posons $\mathcal{P}_j^K = \{z_{k,j}^K : k = 0,\, 1, \cdots,\, K\}$ une partition de $\mathcal{S}_j$ en K intervalles tels que $z_{0,j}^K = x_{min,\,j}$ et $z_{K,j}^K = x_{max,\,j}$;

– Nous notons $\delta_{j,K} \equiv \underset{1\leq k\leq K}{max} \left|z_{k,j}^K - z_{k-1,j}^K\right|$ le pas de la partition $\mathcal{X}_j$;

– Pour chaque $x \in \mathcal{X}_j$, nous définissons $k_j^K(x)$ l'index de l'intervalle de $\mathcal{P}_j^K$ dans lequel se trouve x, c'est à dire le $k \in \{1, \cdots,\, K\}$ pour lequel $x \in \left(z_{k-1,j}^K, z_{k,j}^K\right]$;

– Enfin, nous notons $X_{\setminus j} = (X_k : k = 1,\, 2, \cdots,\, p,\, k \neq j)$, c'est-à-dire le vecteur des $p-1$ prédicteurs privé du $j - ieme$ prédicteur X_j.

• Définition de l'ALE de premier ordre

Considérons un j quelconque dans $\{1,\, 2, \cdots, p\}$ et supposons que la suite des partitions $\{\mathcal{P}_j^K\}_{K\geq 1}$ est telle que $\underset{K\longrightarrow+\infty}{lim} \delta_{j,K} = 0$.

Lorsque $f(.)$ est telle que la limite définie en (4.8) existe et est indépendante de toute les partitions $\mathcal{P}_j^K,\ K \geq 1$, l'ALE de premier ordre non-centré de la caractéristique X_j, est défini pour tout $x_j \in \mathcal{X}_j$ par :

$$g_{j,ALE}(x_j) \equiv \lim_{K \longrightarrow +\infty} \sum_{k=1}^{k_j^K(x_j)} \mathbb{E}\left[f(z_{k,j}^K,\, X_{\setminus j}) - f(z_{k-1,j}^K,\, X_{\setminus j}) \mid X_j \in \left(z_{k-1,j}^K,\, z_{k,j}^K\right]\right] \tag{4.8}$$

Si en plus, $f(.)$ satisfait aux conditions suivantes :

– $f(.)$ différentiable par rapport à x_j ;

– $f^j(.) \equiv \frac{\partial f(.)}{\partial x_j}$ continue sur $\mathcal{X}$;

– $\mathbb{E}[f^j(X_j,\, X_{\setminus j})|X_j = z_j]$ continu en z_j sur $\mathcal{X}_j$

Alors :

$$g_{j,ALE}(x_j) = \int_{x_{min,\,j}}^{x_j} \mathbb{E}[f^j(X_j,\, X_{\setminus j}) \,|\, X_j = z_j]\, dz_j \tag{4.9}$$

L'ALE de premier ordre centré est celui qui est généralement utilisé en pratique. Il est défini par :

$$f_{j,ALE}(x_j) \equiv g_{j,ALE}(x_j) - \mathbb{E}_{X_j}[g_{j,ALE}(X_j)] \tag{4.10}$$

Dans la suite sauf mention contraire, le terme ALE sera utilisé pour désigner l'ALE centré.

• Définition de l'ALE de second-ordre

Soient $\{j, l\} \subseteq \{1, \cdots, d\}$ la paire d'indice associée aux caractéristiques X_j et X_l, et $\{\mathcal{P}_j^K\}_{K\geq 1}$, $\{\mathcal{P}_l^K\}_{K\geq 1}$ deux suites de partitions de $\mathcal{S}_j$ et $\mathcal{S}_l$ respectivement telles que : $\lim_{K \longrightarrow +\infty} \delta_{j,K} = \lim_{K \longrightarrow +\infty} \delta_{l,K} = 0$. Lorsque $f(.)$ est telle que la limite définie en (4.11) existe et est indépendante de toute suite de partitions $\{\mathcal{P}_j^K\}_{K\geq 1}$ et $\{\mathcal{P}_l^K\}_{K\geq 1}$, nous définissons l'ALE de second-ordre non-centré du couple de caractéristiques $(X_j,\, X_l)$, en tout point $(x_j,\, x_l) \in \mathcal{X}_j \times \mathcal{X}_l$, par :

$$h_{\{j,l\},\,ALE}(x_j, x_l) = \lim_{K \longrightarrow +\infty} \sum_{k=1}^{k_j^K(x_j)} \sum_{m=1}^{k_l^K(x_l)} \mathbb{E}\left[\Delta_f^{\{j,l\}}(K,\, k,\, m;\, X_{\setminus\{j,l\}}) \mid X_j \in \left(z_{k-1,j}^K, z_{k,j}^K\right],\, X_l \in \left(z_{m-1,l}^K, z_{m,l}^K\right]\right], \tag{4.11}$$

Avec

$$\begin{aligned}\Delta_f^{\{j,l\}}(K,\, k,\, m\; x_{\setminus\{j,l\}}) = &\left[f\left(z_{k,j}^K,\, z_{m,l}^K,\; x_{\setminus\{j,l\}}\right) - f\left(z_{k-1,j}^K,\, z_{m,l}^K,\; x_{\setminus\{j,l\}}\right)\right] \\ &- \left[f\left(z_{k,j}^K,\, z_{m-1,l}^K,\; x_{\setminus\{j,l\}}\right) - f\left(z_{k-1,j}^K,\, z_{m-1,l}^K,\; x_{\setminus\{j,l\}}\right)\right]\end{aligned} \tag{4.12}$$

Si en plus, la fonction f satisfait les condition suivantes (ce qui sera le plus souvent le cas en pratique) :

– $f(.)$ différentiable en (x_j, x_l) ;

– $f^{\{j,l\}}(.) \equiv \frac{\partial^2 f(x_j, x_l, x_{\setminus\{j,l\}})}{\partial x_j\, \partial x_l}$ continue en $(x_j,\, x_l)$;

– $\mathbb{E}\left[f^{\{j,l\}}(x_j, x_l,\, x_{\setminus\{j,l\}}) \,|\, X_j = z_j,\, X_l = z_l,\right]$ est continu en z_j sur $\mathcal{X}_j \times \mathcal{X}_l$

Alors :

$$h_{\{j,l\},ALE}(x_j, x_l) = \int_{x_{min,\,l}}^{x_l} \int_{x_{min,\,j}}^{x_j} \mathbb{E}\left[f^{\{j,l\}}(x_j, x_l,\, x_{\setminus\{j,l\}}) \,|\, X_j = z_j,\, X_l = z_l,\right] dz_j dz_l. \tag{4.13}$$

Enfin, la fonction ALE de second-ordre centrée des effets de $(X_j,\, X_l)$, notée $f_{\{j,l\},ALE}$ est définie par :

$$f_{\{j,l\},ALE}(x_j, x_l) \equiv g_{\{j,l\},ALE}(x_j, x_l) - \mathbb{E}\left[g_{\{j,l\},ALE}(X_j, X_l)\right] \tag{4.14}$$

Avec,

$$\begin{aligned} g_{\{j,l\},ALE}(x_j, x_l) \quad \equiv h_{\{j,l\},ALE}(x_j, x_l) &- \int_{x_{min,\,j}}^{x_j} \mathbb{E}\left[\frac{\partial h_{\{j,l\},ALE}(X_j,X_l)}{\partial x_j} \mid X_j = z_j\right] dz_j \\ &- \int_{x_{min,\,l}}^{x_l} \mathbb{E}\left[\frac{\partial h_{\{j,l\},ALE}(X_j,X_l)}{\partial x_l} \mid X_l = z_l\right] dz_l \end{aligned} \tag{4.15}$$

□ **Procédure d'estimation empirique de $f_{j,ALE}$ et $f_{\{j,l\},ALE}$**

La procédure d'estimation de $f_{j,ALE}$ peut se résumer en deux grandes étapes :

Étape 1

Premièrement, on remplace la suite de partitions dans (4.8) (ou dans (4.11)) par une partition appropriée de notre échantillon de données $\left\{x_J^{(i)} : i = 1, \cdots, n\right\}$, avec $J = j$ (ou $J = \{j, l\}$). On découpe cette plage d'échantillon en une partition constitué de petits intervalles (ou cellules rectangulaires dans le cas $J = \{j, l\}$).

Pour chaque $j \in \{1, 2, \cdots, p\}$, nous considérons une partition $\{N_j(k) = (z_{k-1,j}, z_{k,j}] : k = 1, 2, \cdots, K\}$ suffisamment fine de l'échantillon $\left\{x_j^{(i)} : i = 1, \cdots, n\right\}$ en K intervalles, avec $z_{k,j}$ le quantile d'ordre $\frac{k}{K}$ de la distribution empirique $\left\{x_j^{(i)} : i = 1, \cdots, n\right\}$, $k = 0, 2, \cdots, K$.

Ce dernier choix garantit d'avoir suffisamment d'individus dans chaque intervalle, mais présente l'inconvénient d'avoir des intervalles de tailles très variables, notamment si les queues de la distribution sont épaisses.

Pour chaque $k = 1, 2, \cdots, K$, on dénombre le nombre d'observation $n_j(k)$ de la base d'entraînement $\left\{x_j^{(i)} : i = 1, \cdots, n\right\}$ qui tombe dans $N_j(k)$, bien entendu, $\sum_{k=1}^{K} n_j(k) = n$.

• **Étape 2**

Ensuite, on remplace l'espérance conditionnelle dans (4.8) (ou dans (4.11)) par la moyenne empirique de f sur les $\{x_{\setminus J}^{(i)} : i = 1, \cdots, n\}$, conditionné à ce que $x_J^{(i)}$ tombe dans l'intervalle (ou la cellule rectangulaire) correspondant(e) de la partition.

Une fois ces deux étapes achevées, on obtient finalement un estimateur $\hat{g}_{j,\,ALE}(.)$ de l'effet de premier-ordre non-centré en sommant les moyennes empiriques conditionnelles jusqu'à x, et aussi l'estimateur de l'ALE de premier ordre (centré) $\hat{f}_{j,\,ALE}(.)$:

$$\hat{g}_{j,\,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{\{i:\, x_j^{(i)} \in N_j(k)\}} \left[f(z_{k,j},\, x_j^{(i)}) - f(z_{k-1,j},\, x_j^{(i)})\right],\ \forall\, x \in (z_{0,j},\, z_{K,j}] \tag{4.16}$$

Par suite,

$$\begin{aligned} \hat{f}_{j,\,ALE}(x) \quad &= \quad \hat{g}_{j,\,ALE}(x) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \hat{g}_{j,\,ALE}(x_j^{(i)}) \\ &= \quad \hat{g}_{j,\,ALE}(x) - \tfrac{1}{n}\textstyle\sum_{k=1}^{K} n_j(k)\hat{g}_{j,\,ALE}(z_{k,j}), \\ &\qquad \forall\, x \in (z_{0,j},\, z_{K,j}] \end{aligned} \tag{4.17}$$

La figure 4.3 illustre parfaitement cette procédure de partitionnement pour l'estimation de l'ALE

de premier-ordre de la caractéristique X_1, dans le cas $d = 2$ (avec $n = 30$ observations, $K = 5$ intervalles).

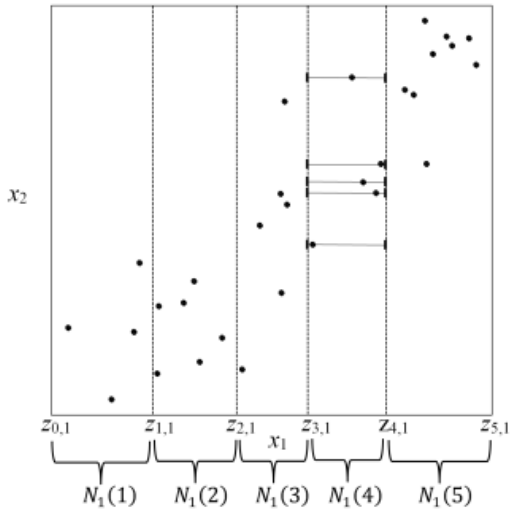


FIGURE 4.3 : *Illustration de la procédure de calcul de la fonction ALE de premier ordre de* X_1, $f_{1,ALE}$ *(issue de l'article de Apley et Zhu (2020)).*

En ce qui concerne l'estimation de $f_{\{j,l\},ALE}$, la procédure est similaire à celle de $f_{j,ALE}$, sauf que cette fois on partitionne $\left\{(x_j^{(i)},\, x_l^{(i)}) :\, i = 1,\, 2, \cdots,\, n\right\}$ en K^2 cellules rectangulaires suffisamment fine, obtenu par le produit cartésien des partitions unidimensionnelles de $\left\{x_j^{(i)} :\, i = 1,\, 2, \cdots,\, n\right\}$ et $\left\{x_l^{(i)} :\, i = 1,\, 2, \cdots,\, n\right\}$ en K intervalles. La figure 4.4 illustre la procédure de découpage de l'échantillon bidimensionnelle $\left\{(x_j^{(i)},\, x_l^{(i)}) :\, i = 1,\, 2, \cdots,\, n\right\}$ en K^2 rectangles.

Pour estimer $f_{\{j,l\},ALE}$, nous estimons d'abord $g_{\{j,l\},ALE}$ définie en (4.15) qui à son tour nécessite premièrement l'estimation de $h_{\{j,l\},ALE}$ définie en (4.13).

Désignons par $N_{\{j,l\}}(k,m) = N_j(k) \times N_l(m) = (z_{k-1,j}, z_{k,j}] \times (z_{m-1,l}, z_{m,l}]$ et $n_{\{j,l\}}(k,m)$ le nombre d'observations de la base d'entraînement à l'intérieur du rectangle $N_{\{j,l\}}(k,m)$ de telle sorte que : $\sum_{k=1}^{K}\sum_{m=1}^{K} n_{\{j,l\}}(k,m) = n$

$$\hat{h}_{\{j,l\},\,ALE}(x_j,\, x_l) = \sum_{k=1}^{k_j(x_j)}\sum_{m=1}^{k_l(x_l)} \frac{1}{n_{\{j,l\}}(k,m)} \sum_{i:\, x_{\{j,l\}}^{(i)} \in N_{\{j,l\}}(k,m)} \Delta_f^{\{j,l\}}(K,\, k,\, m;\, x_{\{j,l\}}^{(i)}) \tag{4.18}$$

Avec,

$$\begin{aligned}\Delta_f^{\{j,l\}}(K,\, k,\, m\; x_{\setminus\{j,l\}}^{(i)}) = &\left[f\left(z_{k,j},\, z_{m,l},\, x_{\setminus\{j,l\}}^{(i)}\right) - f\left(z_{k-1,j},\, z_{m,l},\, x_{\setminus\{j,l\}}^{(i)}\right)\right] \\ &- \left[f\left(z_{k,j},\, z_{m-1,l},\, x_{\setminus\{j,l\}}^{(i)}\right) - f\left(z_{k-1,j},\, z_{m-1,l},\, x_{\setminus\{j,l\}}^{(i)}\right)\right]\end{aligned} \tag{4.19}$$

Puis,

$$\begin{aligned}\hat{g}_{\{j,l\},\,ALE}(x_j,\,x_l) &= \hat{h}_{\{j,l\},\,ALE}(x_j,\,x_l) \\ &\quad -\sum_{k=1}^{k_j(x_j)} \frac{1}{n_j(k)} \sum_{\{i\,:\,x_j^{(i)} \in N_j(k)\}} \left[\hat{h}_{\{j,l\},\,ALE}(z_{k,j},\,x_l^{(i)}) - \hat{h}_{\{j,l\},\,ALE}(z_{k-1,j},\,x_l^{(i)})\right] \\ &\quad -\sum_{m=1}^{k_l(x_l)} \frac{1}{n_l(m)} \sum_{\{i\,:\,x_l^{(i)} \in N_l(m)\}} \left[\hat{h}_{\{j,l\},\,ALE}(x_j^{(i)},\,z_{m,l}) - \hat{h}_{\{j,l\},\,ALE}(x_j^{(i)},\,z_{m-1,l})\right]\end{aligned} \tag{4.20}$$

Enfin,

$$\begin{aligned}\hat{f}_{\{j,l\},\,ALE}(x_j,\,x_l) &= \hat{g}_{\{j,l\},\,ALE}(x_j,\,x_l) - \tfrac{1}{n}\sum_{i=1}^{n} \hat{g}_{\{j,l\},\,ALE}(x_j^{(i)},\,x_l^{(i)}) \\ &= \hat{g}_{\{j,l\},\,ALE}(x_j,\,x_l) - \tfrac{1}{n}\sum_{k=1}^{K}\sum_{m=1}^{K} n_{\{j,l\}}(k,m)\,\hat{g}_{\{j,l\},\,ALE}(z_{k,j},\,x_{m,l})\end{aligned} \tag{4.21}$$

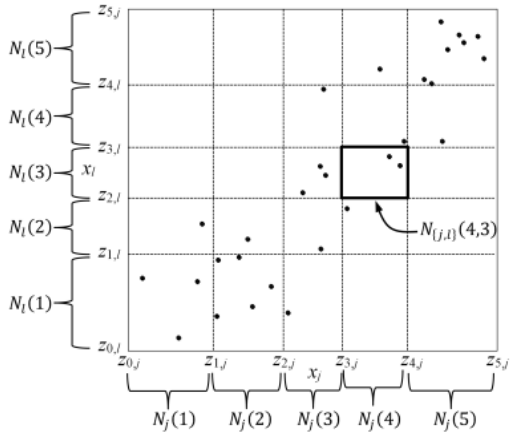


FIGURE 4.4 : *Illustration de la procédure de calcul de la fonction ALE de second ordre de* $(X_1,\,X_2)$*,* $f_{\{1,2\},ALE}$ *(issue de l'article Apley et Zhu (2020)).*

Les théorèmes 3 et 4 de l'annexe A de l'article de Apley et Zhu (2020) démontrent la convergence de ces estimateurs des fonctions ALE de premier et second ordre.

□ ALE pour les variables catégorielles

La méthode ALE nécessite que la caractéristique pour laquelle on mesure l'effet sur la variable cible soit ordonnée. En effet, comme le montre l'expression définie en (4.7), la méthode accumule les effets dans une certaine direction.

Pour calculer l'ALE d'une variable catégorielle disposant d'un ordre naturel entre ses modalités, la démarche est similaire à celle des variables numériques. En ce qui concerne les variables catégorielles ne disposant pas d'un ordre naturel entre les modalités, une solution consiste à créer ou imposer un ordre entre celles-ci sur base d'un calcul de distance entre les modalités.

Pour ce faire, on fixe arbitrairement une modalité de référence, puis on calcule la distance entre la modalité de référence et les autres modalités. Une fois les distances calculées, on ordonne les modalités suivant l'ordre croissant de leur proximité à la modalité de référence.

□ Mauvaises et bonnes façons d'interpréter les tracés ALE

– Le diagramme ALE s'interprète de la même manière que le diagramme PDP. De plus, le fait de centrer la formule permet d'interpréter l'ALE comme l'effet d'une variable sur la prédiction en comparaison à la prédiction moyenne sur l'ensemble des données d'apprentissage. Ainsi, pour une caractéristique X_j, si en un certain point $x = (x_j, x_{\setminus j})$ avec $x_j = 2$, l'ALE de premier-ordre vaut -1, cela signifie que lorsque la j-ième variable vaut 2, alors la valeur de prédiction est inférieure de 1, en comparaison à la prédiction moyenne du modèle.

– Lorsque la fonction f est additive, c'est-à-dire lorsque f peut se mettre sous la forme $f(x) = \sum_{j=1}^{d} f_j(x_j)$, on montre facilement que : $f_{j,\,ALE}(.) = f_{j,\,PDP}(.)$.

– Cependant, lorsque les caractéristiques sont très corrélées, une mauvaise façon d'interpréter le diagramme ALE de X_j serait de dire : si nous fixons $X_{\setminus j} = a$ et faisons varier X_j sur toute sa plage, alors l'ALE de X_j est le reflet de l'évolution des prédictions conséquente aux variations de X_j. Car, étant donné la forte corrélation entre les caractéristiques, lorsque X_j varie, maintenir $X_{\setminus j}$ fixe n'est pas réaliste.

□ **Avantages et inconvénients de l'approche ALE**

• **Avantages**

– *Les tracés ALE sont sans biais* : à l'inverse des PDP, l'ALE ne présente pas de biais et gère beaucoup plus efficacement les relations de dépendances entre variables explicatives. Il est conseillé d'utiliser les ALE plutôt que les PDP ;

– *Coût de calculs informatiques* : le tracé de l'ALE est computationnellement plus rapide à mettre en oeuvre que le tracé des PDP ;

– *facile à interpréter* : les tracés ALE sont centrés. La valeur à chaque point de la courbe ALE est la différence entre la prédiction en ce point et la prédiction moyenne sur l'ensemble de la base d'apprentissage. Cela rend l'interprétation facile.

• **Inconvénients**

– *Complexité* : l'ALE-plot est plus complexe à implémenter et moins intuitive que la méthode PDP.

– *Limité dans sa capacité à détecter les effets hétérogènes d'une caractéristique sur la variable cible* : contrairement aux PDP, il n'est pas possible de tracer plusieurs ALE-plot conditionnellement à une seule instance. Il n'y a pas d'équivalent ICE pour les ALE-plot (nous présentons les ICE plus bas). L'ALE est donc relativement limité dans la détection des phénomènes d'hétérogénéité.

4.2.2 Evaluation de l'importance globale des caractéristiques : les méthodes MR et SFIMP

4.2.2.1 Model Reliance (MR) : mesure de l'importance des caractéristiques par permutation

Les outils d'évaluation de l'importance des variables décrivent dans quelle mesure les caractéristiques contribuent à la précision prédictive d'un modèle de prédiction.

Une approche récente de mesure de l'importance des variables a émergée au cours de ces dernières années. Elle diffère des notions d'importance de variables abordées dans le cadre des modèles de régression linéaire et d'arbres de décision.

L'idée de cette nouvelle approche est de considérer qu'une variable est d'autant plus importante que l'erreur de prédiction du modèle augmente après avoir permuté les valeurs de cette variable considérée.

Dans cette section, nous nous appuyons principalement sur l'article de Fisher *et al.* (2019) et le livre de Molnar (2020).

□ Principe général

Le principe est le suivant : *nous mesurons l'importance d'une caractéristique en calculant l'augmentation de l'erreur de prédiction du modèle conséquente à une permutation des valeurs de cette caractéristique.*

Lorsque la prédiction d'un modèle est grandement modifiée lorsqu'on bouleverse les valeurs d'une caractéristique, cela signifie que le modèle est sensible aux variations des valeurs de ladite caractéristique. La caractéristique joue donc un rôle prépondérant dans le modèle. Inversement, une variable dont la modification des valeurs n'impacte que peu la précision du modèle est considérée comme moins importante.

□ Formalisation mathématique

– Soit $Y \in \mathcal{Y}$ la variable cible, $X = (X_1, \cdots, X_p) \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ le vecteur des caractéristiques.

– Soit $\left(x^{(i)}, y^{(i)}\right)_{1 \leq i \leq n}$ une réalisation échantillonnale de (X, Y) composé de n observations (avec $n, p \in \mathbb{N}^*$).

– Désignons par $f \in \mathcal{F}$ notre modèle d'intérêt et $\hat{f}$ son estimateur. Ici $\mathcal{F}$ représente la famille de modèles choisir pour l'ajustement de Y sur X (par exemple la famille des modèles GLM ou la famille des GAMs, ou alors la famille des forêts aléatoires, etc.)

Soit $L : \mathcal{F} \times \mathcal{Z} \longrightarrow \mathbb{R}^+$, avec $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}$, la fonction de perte.

L pourrait être la perte quadratique définie par : $L(f,\ (y,\ x)) = (y - f(x))^2$ pour des problèmes de régression, ou la perte charnière (hinge loss) $L_h(f,\ (y, x)) = (1 - yf(x))_+$ pour des problèmes de classification.

• Perte commutée

Pour décrire dans quelle mesure la précision d'un modèle de prédiction f dépend d'une caractéristique X_j ($j \in \{1, \cdots, p\}$), Fisher *et al.* (2019) introduisent la notion de *perte commutée*.

Supposons que l'on dispose de $Z^{(a)} = (Y^{(a)},\ X^{(a)})$ et $Z^{(b)} = (Y^{(b)},\ X^{(b)})$ deux sous-échantillons indépendants et de même distribution que $Z = (Y, X)$, la *perte commutée* associée à X_j est définie par :

$$e_{j,\,com}(f) \equiv \mathbb{E}\left[L\left(f,\ (Y^{(b)},\ X_j^{(a)},\ X_{\setminus j}^{(b)}\right)\right] \tag{4.22}$$

L'expression *"commutée"* vient du fait que dans l'expression de $e_{j,\,com}(f)$, nous utilisons les variables $\left(Y^{(b)},\ X_{\setminus j}^{(b)}\right)$ de $Z^{(b)}$ et la variable X_j de $Z^{(a)}$: tout se passe comme si on avait commuté $X_j^{(b)}$ et $X_j^{(a)}$.

Une autre façon d'interpréter la quantité $e_{j,\,com}(f)$ est de la voir comme la perte attendue de f lorsque du bruit est ajouté à la caractéristique X_j de telle sorte que X_j devient complètement non informative de Y.

Après avoir calculé $e_{j,\,com}(f)$, on la compare à la perte attendue standard lorsqu'aucune des caractéristiques n'est perturbée. Désignons par $e_{j,\,orig}(f)$ cette perte standard. Elle est définie par :

$$e_{j,\,orig}(f) \equiv \mathbb{E}\left[L\left(f,\,(Y,\,X_j,\,X_{\setminus j}\right)\right] \tag{4.23}$$

• Ratio de dépendance au modèle (MR)

On définit le ratio suivant communément appelé *model relaiance* (en anglais) ou *dépendance au modèle* (en français). Il permet de mesurer l'importance de la variable X_j dans la précision du modèle f :

$$MR_j(f) \equiv \frac{e_{j,\,com}(f)}{e_{j,\,orig}(f)} \equiv \frac{\text{perte attendue du modèle } f \text{ lorsqu'on introduit du bruit dans } X_j}{\text{perte attendue standard du modèle } f \text{ (sans aucun bruit)}} \tag{4.24}$$

– Une $MR_j(f) > 1$ signifie que la caractéristique X_j a une grande importance pour la précision du modèle f.

Par exemple, une $MR_j(f) = 3$ signifie que le modèle est fortement dépendant de X_j, en ce sens qu'une perturbation des valeurs de X_j triple l'erreur commise par le modèle.

– Une $MR_j(f) = 1$ signifie que le modèle ne dépend pas de X_j, en ce sens qu'un brouillage des valeurs de X_j n'a aucune incidence sur la précision du modèle f.

– Une $MR_j(f) < 1$ est difficile à interpréter. Cela voudrait dire qu'une variable aléatoire explique mieux Y que la caractéristique X_j.

La comparaison entre $e_{j,\,com}(f)$ et $e_{j,\,orig}(f)$ peut également se faire en utilisant la différence au lieu du ratio. C'est-à-dire qu'on pourrait également définir MR par :

$$MR_{j,\,difference}(f) = e_{j,\,com}(f) - e_{j,\,orig}(f)$$

Dans le package *vip* de R dédié à la mesure de l'importance des caractéristiques, le MR est par défaut calculé en utilisant la différence plutôt que le ratio. Pour passer au ratio, il suffit de spécifier à l'argument *type* de la fonction *vi-permut* la valeur "ratio".

• Estimation empirique du ratio $MR_j(f)$

Pour obtenir un estimateur $\hat{MR}_j(\hat{f})$ de $MR_j(f)$, il suffit de remplacer grossièrement $e_{j,\,com}(f)$ et $e_{j,\,orig}(f)$ par leur correspondant empirique :

$$\hat{e}_{j,\,orig}(\hat{f}) \equiv \frac{1}{n}\sum_{i=1}^{n} L\left\{\hat{f}, \left(y^{(i)}, (x_j^{(i)}, x_{\setminus j}^{(i)})\right)\right\} \tag{4.25}$$

$$\hat{e}_{j,\,com}(\hat{f}) \equiv \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{k\neq i} L\left\{\hat{f}, \left(y^{(k)}, (x_j^{(i)}, x_{\setminus j}^{(k)})\right)\right\}$$

Dans l'estimateur de $e_{j,\,com}(f)$ ci-dessus, nous avons agrégé toutes les combinaisons possibles des valeurs observées pour $(Y, X_{\setminus j})$ et X_j, en excluant les paires qui sont effectivement observées dans l'échantillon original. Cependant, cet agrégation peut être prohibitive du point de vue informatique en raison de la taille de l'échantillon. Pour réduire le temps de calcul, un autre estimateur de $e_{j,\,com}(f)$ peut être utilisé :

$$\hat{e}_{j,\,divide}(\hat{f}) \equiv \frac{1}{2\lfloor n/2 \rfloor}\sum_{i=1}^{\lfloor n/2 \rfloor}\left[L\left\{\hat{f};\left(y^{(i)},\, x_j^{(i+\lfloor n/2 \rfloor)},\, x_{\setminus j}^{(i)}\right)\right\} + L\left\{\hat{f};\left(y^{(i+\lfloor n/2 \rfloor)},\, x_j^{(i)},\, x_{\setminus j}^{(i+\lfloor n/2 \rfloor)}\right)\right\}\right] \tag{4.26}$$

$\hat{e}_{j,\,divide}(\hat{f})$ consiste à diviser l'ensemble de données en deux et d'échanger les valeurs de la caractéristique X_j des deux moitiés.

Remarque : les estimateurs $\hat{e}_{j,\,orig}(\hat{f})$, $\hat{e}_{j,\,comm}(\hat{f})$, $\hat{e}_{j,\,divide}(\hat{f})$ appartiennent à la classe des U-statistiques, bien connue en Statistique pour leurs bonnes propriétés. Ainsi sous des hypothèses mineures, ces estimateurs sont tous sans biais et asymptotiquement normaux (confère Hoeffding et Robbins (1948)).

□ MR et inférence causale

La mesure de l'importance des variables basée sur les permutations via le calcul du $MR(f)$ se rapporte à étudier la façon dont le modèle se comporte suite à une intervention sur les caractéristiques. Ceci se rapproche de l'objectif de l'inférence causale qui consiste à étudier comment une intervention sur une variable modifiera le résultats sur une variable cible donnée.

Soit X_j ($j \in \{1, \cdots, p\}$) la caractéristique dont on veut évaluer l'effet causal sur une variable Y. Supposons sans nuire à la généralité que X_j est binaire et prend ses valeurs dans $\{0, 1\}$. Par soucis d'alignement avec la littérature sur l'inférence causale, nous renommons provisoirement nos variables :

$T := X_j$ un indicateur de traitement binaire ;

$C := X_{\setminus j}$ un ensemble de variable de contrôle ;

Y : la variable d'intérêt.

On a la relation, $Y = Y_0(1-T)+Y_1T$, où Y_1 et Y_0 représentent les résultats potentiels lorsqu'on est soumis ou non au traitement respectivement.

Posons $f_0(t, c) \equiv \mathbb{E}(Y|T = t,\, C = c)$ la fonction d'espérance conditionnelle que l'on cherche à approximer à l'aide de notre modèle d'apprentissage statistique.

Posons $CATE(c) \equiv \mathbb{E}(Y_1 - Y_0|C = c)$ l'effet marginal moyen conditionnel du traitement T.

Sous les deux hypothèses suivantes qui garantissent la bonne définition et l'identifiabilité de f_0 et $CATE$:

$$\begin{aligned} H_1 :&\ \ (Y_1,\, Y_0) \perp T \mid C \quad \text{(Indépendance conditionnelle du traitement ou sélection ignorable)} \\ H_2 :&\ \ 0 < \mathbb{P}(T = 1 \mid C = c) < 1,\, \forall c \quad \text{(Condition de positivité ou d'équilibrage)} \end{aligned} \tag{4.27}$$

L'importance du traitement (T) dans le modèle f_0 mesuré par le ratio $MR_T(f_0)$ se décompose comme suit :

$$MR_T(f_0) = 1 + \frac{Var(T)}{\mathbb{E}_{T,C}\left[Var(Y \mid T, C)\right]} \sum_{t\in\{0,1\}} \left\{\mathbb{E}(Y_1 - Y_0 \mid T = t)^2 + Var(CATE(C) \mid T = t)\right\} \tag{4.28}$$

De l'équation (4.28), on peut déduire que :

– Lorsque $Var(T) > 0$, un effet marginal moyen du traitement de magnitude élevée ($\mathbb{E}(Y_1 - Y_0 \mid T = t)^2$ élevé) implique un accroissement de $MR_T(f_0)$, toutes choses égales par ailleurs. Ainsi, sous H_1 et H_2, plus une variable a un degré élevé de causalité sur la cible Y, mieux son $MR(f_0)$ est élevé et par conséquent mieux elle contribue à la précision prédictive du modèle f_0.

– Cependant, la réciproque n'est pas vérifiée. Une $MR(f_0)$ n'implique pas nécessairement que la caractéristique en question a un effet causal de magnitude élevée. On observe par exemple que le $MR(f_0)$ croît mécaniquement avec la valeur $Var(T)$. Pourtant cette dernière n'a rien à voir avec la causalité de T sur Y.

□ **Avantages et limites de la méthode** MR

• **Avantages**

Parmi les avantages, on peut relever :

– *Une interprétation facile* : l'importance d'une variable suivant cette approche est l'augmentation de l'erreur de prédiction conséquente au brouillage des informations apportées par la variable concernée ;

– *Grandeur sans unité* : l'utilisation du MR calculé avec le ratio est une grandeur sans unité. Ce qui rend possible la comparaison de l'importance des caractéristiques entre plusieurs modèles ;

– *Vitesse de calculs* : la mesure de l'importance des caractéristiques basée sur les permutations possède également des avantages en terme de temps de calculs. En effet, elle ne nécessite pas d'entraîner le modèle plus d'une fois. Contrairement, à d'autres approches qui suggèrent de supprimer une caractéristique du modèle, puis de ré-entraîner le modèle, dans l'optique de comparer l'erreur du modèle avec et sans la caractéristique d'intérêt : ce qui nécessite parfois beaucoup de temps ;

– L'importance des caractéristique basée sur les permutations est en partie reliée à l'effet causal de la caractéristique ;

– La mesure de l'importance des caractéristiques basée sur la statistique MR est plus ou moins vraisemblable en ce sens qu'elle prend intrinsèquement en compte les interactions entre la caractéristique d'intérêt et les autres caractéristiques dans le calcul de MR.

• **Limites**

– Le niveau d'importance d'une caractéristique n'a pas une interprétation causale directe. Autrement dit un MR élevé n'implique pas nécessairement que la caractéristique concernée a un effet causal important sur la variable cible ;

– Si les caractéristiques sont très corrélées entre elles, l'importance d'une caractéristique peut être biaisée par des instances de données irréalistes.

4.2.2.2 Importance des variables basée sur les valeurs de Shapley

□ **Principe**

La mesure de l'importance des caractéristiques basée sur les valeurs de Shapley est une alternative à la mesure de l'importance des caractéristiques basée sur les permutations.

Cette approche de mesure de l'importance des variable s'intitule *Shapley Feature IMPortance (SFIMP)*. Son principe est le suivant :

1– Dans un premier temps, on calcule la différence de performance du modèle entre le scénario où toutes les covariables sont utilisées et celui où toutes les covariables sont ignorées.

2– Dans un second dans temps, on répartit *"équitablement"* la différence de performance obtenue à l'issue de l'étape 1, entre les différentes caractéristiques. La répartition équitable en question s'inspire de l'utilisation d'outils de partage des gains développés en Théorie des Jeux coopératifs.

□ **Formalisation**

Désignons par $GE(\hat{f},\, \mathcal{P}) = \mathbb{E}[L(\hat{f},\, (Y,\, X))]$ l'erreur de généralisation du modèle ajusté $\hat{f}$ sur un échantillon généré par un processus de loi $\mathcal{P}$.

Soit $\hat{GE}(\hat{f},\, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} L(\hat{f},\, (y^{(i)},\, x^{(i)})$ un estimateur empirique de $GE(\hat{f},\, \mathcal{P})$ obtenu à partir d'un échantillon de données $\mathcal{D}$ de n réalisations de $(Y,\, X)$.

Soit $S \subseteq \{X_1, \cdots,\, X_p\}$, on appelle *fonction caractéristique* de la coalition des variables de S la quantité définie par :

$$v_{GE}(S) = \hat{GE}_S(\hat{f}, \mathcal{D}) - \hat{GE}_{\emptyset}(\hat{f}, \mathcal{D}) \tag{4.29}$$

avec

$$\hat{GE}_S(\hat{f}, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{n-1} \sum_{k \neq i}^{n} L(\hat{f},\, (y^{(i)},\, x_S^{(i)},\, x_{\setminus S}^{(k)})$$

La fonction caractéristique $v_{GE}(S)$ mesure la différence de performance du modèle lorsqu'on utilise uniquement les caractéristiques dans S et lorsqu'on utilise aucune caractéristique du tout pour la prédiction de Y.

La contribution marginale d'une variable X_j $(j \in \{1, \cdots,\, p\})$ à une coalition de variables S est donnée par :

$$\begin{aligned} \Delta_j(S) \quad &= v_{GE}(S \cup \{j\}) \,-\, v_{GE}(S) \\ &= \hat{GE}_{S\cup\{j\}}(\hat{f}, \mathcal{D}) \,-\, \hat{GE}_S(\hat{f}, \mathcal{D}) \end{aligned}$$

Alors la mesure $SFIMP$ de la variable X_j est définie par :

$$\begin{aligned} \hat{\phi}_j(v_{GE}) \quad &= \frac{1}{p!}\sum_{\pi \in \Pi} \Delta_j(B_j(\pi)) \\ &= \frac{1}{p!}\sum_{\pi \in \Pi} \left[\hat{GE}_{B_j(\pi)\cup\{j\}}(\hat{f}, \mathcal{D}) \,-\, \hat{GE}_{B_j(\pi)}(\hat{f}, \mathcal{D})\right]. \end{aligned} \tag{4.30}$$

où

– Π : l'ensemble de toutes les permutations possible de l'ensemble $\{1, \cdots,\, p\}$ des index des caractéristiques ;

– π : un élément de Π, c'est-à-dire une permutation des index de caractéristiques ;

– $B_j(\pi)$ est l'ensemble des index de caractéristiques qui apparaissent avant la j-ième caractéristique dans π.

Par exemple, pour $p = 4$, si nous considérons la caractéristique d'index $j = 4$ et la permutation $\pi = \{2,\, 3,\, \mathbf{4},\, 1\}$, alors $B_4(\pi) = \{2,\, 3\}$. Cela suppose qu'au préalable on attribue un rang à chaque des caractéristiques.

□ **Propriétés de $\hat{\phi}_j$**

L'importance des variables basée sur les valeurs de Shapley (SFIMP) satisfait les quatre propriétés souhaitables suivantes :

1– *Efficacité* : $\sum_{j=1}^{d}\phi_j = v_{GE}(\{X_1, \cdots, X_p\})$. Toutes les valeurs SFIMP s'additionnent pour donner $v_{GE}(\{X_1, \cdots, X_p\})$ qui correspond à la différence de performances prédictives lorsque toutes les variables sont utilisées et lorsque toutes les variables sont ignorées.

Cela nous permet alors de calculer la proportion d'importance expliquée pour chaque caractéristique X_j en utilisant la formule suivante : $\dfrac{\phi_j}{\sum_{j=1}^{p}\phi_j}$;

2– *Symétrie* : si $v_{GE}(S \cup \{X_j\}) = v_{GE}(S \cup \{X_k\})$ pour tout $S \subseteq \{X_1, \cdots, X_p\} \setminus \{X_j, X_k\}$, alors $\phi_j = \phi_k$.

C'est-à-dire que deux caractéristiques X_j et X_k qui ont des contributions marginales égales pour toutes les coalitions possibles, ont nécessairement la même valeur SFIMP ;

3– *Propriété factice* : si $v_{GE}(S \cup \{j\}) = v_{GE}(S)$ pour tout $S \subseteq \{X_1, \cdots, X_d\}$, alors $\phi_j = 0$.

La valeur SFIMP de la covariable X_j est nulle lorsque sa contribution marginale ne change pas quelque soit la coalition S à laquelle elle est ajoutée.

4– *Additivité* : $\phi_j(v_{GE} + w_{GE}) = \phi_j(v_{GE}) + \phi_j(w_{GE})$. L'importance de X_j dans $v_{GE} + w_{GE}$ est égale à la somme des importances individuelles de X_j issues des deux mesures de performances distinctes v_{GE} et dans w_{GE}.

Similairement, toute multiplication de la mesure de performance avec une constante positive α n'affecte pas le classement des caractéristiques. Autrement dit, $\phi_j(\alpha \cdot v_{GE}) = \alpha \cdot \phi_j(v_{GE})$. $\phi_j(\cdot)$ est donc linéaire.

□ **Avantages et inconvénients des SFIMP**

• **Avantages**

– *Fondement théorique rigoureux* : cette approche de mesure de l'importance des variables basée sur les valeurs de Shapley repose sur une théorie mathématique solide (la théorie des jeux coopératifs).

– Les valeurs $SFIMP$ permettent d'avoir une interprétation globale du fonctionnement du modèle. Ces valeurs prennent également en considération les interactions entre les caractéristiques dans le calcul des valeurs SFIMP ϕ_j.

– L'une des particularités des valeurs $SFIMP$ est qu'elles permettent de mieux calculer l'importance des variables même dans des situations de forte corrélation entre les caractéristiques car elles considèrent la contribution marginale d'une caractéristique aux différentes coalitions.

• **Inconvénients**

– L'inconvénient qui revient le plus souvent est le coût de calcul des valeurs $SFIMP$ à partir de la formule (4.30) lorsque le nombre p de caractéristiques ou la taille n de la base de données de test $\mathcal{D}$ sont élevés.

Pour palier à cette limite, Casalicchio *et al.* (2019) propose l'algorithme suivant pour le calcul des valeurs de $SFIMP$.

Dans cet algorithme, d'une part le nombre $p!$ de l'équation (4.30) est remplacé par un entier $m_{feat} \ll p!$.

D'autre part, pour le calcul de l'erreur de généralisation, nous considérons $m_{obs} \in \mathbb{N}^*$ permutations aléatoires parmi l'ensemble des $n!$ permutations possibles $\{\tau_1, \cdots, \tau_{n!}\}$ de l'ensemble des n observations de $\mathcal{D}$.

TABLE 4.1 : Algorithme d'approximation de la contribution ϕ_j de la caractéristique X_j à la performance globale du modèle

variables d'entrée : m_{feat}, m_{obs}, $\hat{f}$, L, $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{1 \leq i \leq n}$
pour $k \in \{1, \cdots, m_{feat}\}$ **faire :**
  choisir une permutation aléatoire des indices des p caractéristiques $\pi \in \Pi$
  poser $S = B_j(\pi)$
  initialiser : $\hat{GE}_{S,perm} = 0$, $\hat{GE}_{S\cup\{j\},perm} = 0$
  pour $l \in \{1, \cdots, m_{obs}\}$ **faire :**
    choisir une permutation aléatoire des indices d'observations $\tau \in \{\tau_1, \cdots, \tau_{n!}\}$
    $\hat{GE}_{S,perm} = \hat{GE}_{S\cup\{j\},perm} + \frac{1}{n}\sum_{i=1}^{n} L\left(\hat{f}, (y^{(i)}, x_S^{(i)}, x_{\setminus S}^{(\tau(i))})\right)$
    $\hat{GE}_{S\cup\{j\},perm} = \hat{GE}_{S\cup\{j\},perm} + \frac{1}{n}\sum_{i=1}^{n} L\left(\hat{f}, (y^{(i)}, x_{S\cup\{j\}}^{(i)}, x_{\setminus S\cup\{j\}}^{(\tau(i))})\right)$
  calculer la contribution marginale pour la caractéristique j dans l'itération k :
  $\Delta_j^{(k)}(S) = \frac{1}{m_{obs}}(\hat{GE}_{S\cup\{j\},perm} - \hat{GE}_{S,perm})$
retourner : $\hat{\phi}_j = \frac{1}{m_{feat}}\Delta_j^{(k)}(S)$

4.2.3 Évaluation de la force d'interaction entre les caractéristiques : H-statistique de Friedman, Indices de Sobol

Soit f une fonction de plusieurs variables définie en $x = (x_1, \cdots, x_p) \in \mathbb{R}^p$.

Comme mentionné dans Friedman et Popescu (2008), on dira que f possède des interactions en x_j et x_k si la différence de la valeur de $f(x)$ résultant du changement de x_j dépend de la valeur de x_k.

– Lorsque X_j et X_k sont des caractéristiques aléatoires numériques, l'existence d'interactions entre-elles peut se traduire par la relation :

$$\mathbb{E}_X\left[\frac{\partial^2 f(X)}{\partial x_j \partial x_k}\right]^2 > 0$$

– Pour des variables catégorielles, les dérivés partielles sont remplacer par des différences finies.

– En absence d'interactions entre x_j et x_k, la fonction f peut s'écrire comme la somme de deux fonctions : l'une ne dépendant pas de x_j et l'autre ne dépendant pas de x_k. Soit, $f(x) = a(x_{\setminus j}) + b(x_{\setminus k})$.

– Dans le cas où x_j ne possède aucune interaction avec les autres variables, f peut s'exprimer comme suit : $f(x) = h(x_j) + g(x_{\setminus j})$ et on dit que f est additive en x_j.

– De manière similaire, on dira qu'il existe une interaction entre l variables numériques $x_{j_1}, \cdots, x_{j_l}$, $(l > 2)$ si :

$$\mathbb{E}_X\left[\frac{\partial^l f(X)}{\partial x_{j_1} \cdots \partial x_{j_l}}\right]^2 > 0$$

Au cours de ces dernières années, plusieurs méthodes ont été développées pour évaluer l'existence des effets d'interaction dans les modèles prédictifs. Dans cette section nous présentons celles

d'entre-elles les plus populaires.

4.2.3.1 Détection d'interactions à partir H-statistique de Friedman

□ Principe général

Cette approche de détection de l'interaction entre les variables dans un modèle fut introduite en 2008, par Friedman et Popescu (2008). Elle est essentiellement basée sur les propriétés de la fonction de dépendance partielle *centrée*.

– Lorsque deux caractéristiques X_j et X_k n'interagissent pas, la fonction de dépendance partielle de la paire $X_S = \{X_j, X_k\}$ se décompose comme la somme de fonctions de dépendance partielle individuelle de chacune des deux variables :

$$PD_{jk}(x_j, x_k) = PD_j(x_j) + PD_k(x_k) \tag{4.31}$$

– Lorsque X_j n'interagit avec aucune des autres caractéristiques, on a plutôt :

$$f(x) = PD_j(x_j) + PD_{\setminus j}(x_{\setminus j}) \tag{4.32}$$

Les propriétés ci-dessus des fonctions de dépendance partielle sont utilisées pour construire des statistiques afin d'évaluer les effets d'interaction entre caractéristiques.

On distingue deux types d'interaction entre les caractéristiques :

• *L'interaction bi-directionnelle* qui nous indique si et dans quelle mesure deux caractéristiques du modèle interagissent l'une avec l'autre ;

Pour le test de la présence d'interaction bi-directionnelle nous utilisons la statistique :

$$H_{jk}^2 = \frac{\sum_{i=1}^{N} \left[\hat{PD}_{jk}\left(x_j^{(i)}, x_k^{(i)}\right) - \hat{PD}_j\left(x_j^{(i)}\right) - \hat{PD}_k\left(x_k^{(i)}\right)\right]^2}{\sum_{i=1}^{N} \hat{PD}_{jk}^2\left(x_j^{(i)}, x_k^{(i)}\right)} \tag{4.33}$$

Cette quantité mesure la proportion de variance de $\hat{PD}_{jk}(x_j^{(i)}, x_k)$ non capturée par $\hat{PD}_j(x_j) + \hat{PD}_k(x_k)$.

Elle est nulle en cas d'absence d'interaction entre les caractéristiques X_j et X_k. Inversement une valeur proportionnellement plus grande indique un effet d'interaction plus fort entre les caractéristiques concernées.

• *L'interaction totale* qui nous indique si et dans quelle mesure une entité interagit dans le modèle avec toutes les autres entités.

Pour la mesure de l'interaction totale, on utilise la statistique :

$$H_j^2 = \frac{\sum_{i=1}^{N} \left[\hat{f}\left(x^{(i)}\right) - \hat{PD}_j\left(x_j^{(i)}\right) - \hat{PD}_{\setminus j}\left(x_{\setminus j}^{(i)}\right)\right]^2}{\sum_{i=1}^{N} \hat{f}^2\left(x^{(i)}\right)} \tag{4.34}$$

Cette quantité est non nulle dans la mesure où X_j interagit avec au moins une autre variable dans le modèle.

En pratique, en analysant l'ensemble des valeurs $\{H_j : 1 \leq j \leq d\}$, on peut identifier les caractéristiques pour lesquelles il existe au moins un effet d'interaction avec les autres caractéristiques. Et s'étant donné une caractéristique X_j admettant des effets d'interaction, il suffit d'examiner les valeurs $\{H_{jk} : k \neq j\}$ pour identifier les variables avec lesquelles elle interagit le plus.

□ **Avantages et inconvénients des H-statistique de Friedman**

• **Avantages**

Comme avantages, nous pouvons relever que :

– Interprétation intuitive : l'interaction basée sur la statistique H de Friedman est définie comme la part de variance du modèle expliquée par l'interaction concernée.

– La statistique H est comparable non seulement entre les caractéristiques, mais également d'un modèle à l'autre. Car de par sa formulation comme ratio de deux quantités de même dimension, elle est sans dimension.

– La statistique H est à mesure de détecter toute sorte d'interaction entre les caractéristiques peu importe leur complexité.

• **Limites**

– L'une des premières limites que nous pouvons soulever est que les H-statistiques peuvent parfois repérer des interactions fallacieuses entre les caractéristiques. Ces interactions fallacieuses peuvent typiquement se produire en cas de forte corrélation entre les caractéristiques.

– Il n'est pas évident de trancher si une interaction est significativement supérieure à 0. Pour s'en assurer, il est nécessaire de faire recours à un test d'hypothèse statistique, ce qui n'est pas encore disponible dans la littérature, à notre connaissance.

– La statistique H n'est pas bornée et peut parfois excéder la valeur 1 : il est donc difficile de dire quand H est suffisamment grand pour que nous puissions considérer une interaction comme forte.

– Enfin, nous pouvons évoquer le gros coût de calcul des valeurs de la statistique H.

4.2.3.2 Détection d'interactions à partir des indices de Sobol

□ **Principe général**

Le calcul des indices de Sobol s'appuie sur la théorie de l'analyse des sensibilités et les principales contributions de Sobol' (1990).

Ces indices permettent d'analyser numériquement la structure d'une fonction non linéaire.

Le principe de calcul des indices de Sobol s'inspire essentiellement du principe de la décomposition de la variance réalisée en analyse de la variance (*ANOVA decomposition*).

La force d'interaction $V_{jj'}$ entre deux caractéristiques X_j et $X_{j'}$ est mesurée en effectuant la différence entre la variance du modèle ajusté sur la paire de variables concernée (c'est-à-dire $\mathbb{V}ar\left(\mathbb{E}(Y \mid X_j,\, X_{j'})\right)$) et la somme des variances des modèles ajustés individuellement sur chacune des deux variables (c'est-à-dire $\mathbb{V}ar\left(\mathbb{E}(Y \mid X_j)\right) + \mathbb{V}ar\left(\mathbb{E}(Y \mid X_k)\right)$). Soit,

$$V_{jj'} = \mathbb{V}ar\left(\mathbb{E}(Y \mid X_j,\, X_{j'})\right) - \left(\underbrace{\mathbb{V}ar\left(\mathbb{E}(Y \mid X_j)\right)}_{V_j} + \underbrace{\mathbb{V}ar\left(\mathbb{E}(Y \mid X_k)\right)}_{V_{j'}} \right)$$

□ **Formalisation mathématique et interprétation des indices de Sobol**

Commençons par énoncer le lemme de la décomposition de la variance telle que formulé par Efron et Stein (1981).

• Lemme de décomposition ANOVA

Toute variable aléatoire $S\left(X_1, \cdots, X_p\right)$ fonction de p variables aléatoires indépendantes $X_1, \cdots, X_p$ et telles que $\mathbb{E}(S^2) < +\infty$, peut se décomposer comme suit :

$$\begin{aligned} S(X_1, \cdots, X_p) = \quad & \mu \quad + \quad \sum_{j=1}^{p} A_j(X_j) + \sum_{1 \leq j < j' \leq p} B_{j,j'}(X_j, X_{j'}) \\ + \quad & \sum_{1 \leq j < j' < j'' \leq p} C_{j,j',j''}(X_j, X_{j'}, X_{j''}) + \cdots + H(X_1, X_2, \cdots, X_p) \end{aligned} \tag{4.35}$$

où tous les $2^d - 1$ termes aléatoires du membre droit de l'égalité (4.35) sont centrés et mutuellement non-corrélées (orthogonalité). En conséquence cette propriété d'orthogonalité, la décomposition de S ci-dessus est *unique* (confère Sobol' (1990)).

avec,

$$\begin{cases} \mu & = \mathbb{E}[S] \text{ (moyenne globale de } S); \\ A_j(x_j) & = \mathbb{E}[S \mid X_j = x_j] - \mu, \text{ (effet principal de la variable } X_j \text{ dans } S); \\ B_{j,j'}(x_j, x_{j'}) & = \mathbb{E}[S \mid X_j = x_j, X_{j'} = x_{j'}] - \mathbb{E}[S \mid X_j = x_j] - \mathbb{E}[S \mid X_{j'} = x_{j'}] + \mu; \\ & \quad \text{(effet d'interaction de second ordre entre les variables } X_j, X_{j'} \text{ dans } S); \\ \vdots & \quad \vdots \\ H(x_1, x_2, \cdots, x_p) & = \mathbb{E}[S \mid X_1 = x_1, \cdots, X_p = x_p] - \cdots . \end{cases}$$

• Théorème de décomposition de Sobol de la variance

La variance du modèle $f(X) = \mathbb{E}[Y \mid X_1, \cdots, X_p]$ à entrées $X_1, \cdots, X_p$ indépendantes s'obtient en remplaçant $S(X)$ par $f(X) = \mathbb{E}[Y \mid X_1, \cdots, X_p]$ dans l'équation (4.35), puis en appliquant la variance de part et d'autre des membres l'égalité obtenue. On obtient :

$$V \equiv \mathbb{V}ar(f(X)) = \left[\sum_{j=1}^{p} V_j\right] + \left[\sum_{1 \leq j < j'' \leq p} V_{jj''}\right] + \cdots + V_{1 \cdots p} \tag{4.36}$$

avec

$$\begin{cases} V_j & = \mathbb{V}ar(\mathbb{E}(Y \mid X_j)) \\ V_{jj'} & = \mathbb{V}ar(\mathbb{E}(Y \mid X_j, X_{j'})) - V_j - V_{j'} \\ V_{j,j',j''} & = \mathbb{V}ar(\mathbb{E}(Y \mid X_j, X_{j'}, X_{j''})) - V_{jj'} - V_{jj''} - V_{j'j''} - V_j - V_{j'} - V_{j''} \\ & \cdots \\ V_{1 \cdots d} & = V - \left[\sum_{1 \leq j \leq d} V_j\right] - \left[\sum_{1 \leq j < j'' \leq d} V_{jj''}\right] - \cdots - \left[\sum_{1 \leq j_1 < \cdots < j_{d-1} \leq d} V_{j_1 \cdots j_{d-1}}\right] \end{cases}$$

L'équation (4.36) est appelée *décomposition de Sobol de la variance du modèle* f.

• Définition : indices de sensibilité d'ordre k

À partir de la décomposition de Sobol (4.36) ci-dessus, on définit les indices de sensibilité d'ordre k ($k \in \{1, \cdots, p\}$) :

$$S_{i_1 \cdots i_k} = \frac{V_{i_1 \cdots i_k}}{V}$$

$S_{i_1 \cdots i_k}$ exprime la proportion de la variance du modèle qui est expliquée par l'interaction entre les caractéristiques $X_{i_1}, \cdots, X_{i_k}$. Il s'agit de la sensibilité du modèle aux variables $X_{i_1}, \cdots, X_{i_k}$ qui n'est pas prise en compte par leur effets d'interaction d'ordre inférieur à k.

En pratique le calcul des indices de Sobol de à tous les ordres occasionne de gigantesque coûts de calcul informatique, notamment lorsque p est grand. À cet effet, Homma et Saltelli (1996) ont introduit la notion d'*indice de Sobol d'effet total*.

L'*indice de Sobol d'effet total* mesure l'effet total d'une caractéristique donnée dans le modèle, en incluant toutes les synergies possibles entre cette caractéristique et les autres caractéristiques.

• Définition : indice de Sobol d'effet total ou indice de sensibilité total

L'*indice de sensibilité total* noté S_{T_j} dû à la variable X_j est défini comme la somme de tous les indices de sensibilité relatifs à la variable X_j, soit :

$$S_{T_j} = 1 - S_j^e$$

où S_j^e désigne la somme de tous les termes d'interactions $S_{j_1 \cdots j_l}$ ($1 \leq l \leq p$) dans lesquels l'indice j n'apparaît pas. Par exemple pour $p = 3$, $S_{T_1} = 1 - S_1^e$, avec $S_1^e = S_2 + S_3 + S_{23}$.

□ Procédure d'estimation des indices de Sobol

En pratique, il existe plusieurs techniques permettant d'estimer les indices de Sobol de premier ordre. L'estimation des indices d'ordres supérieurs s'obtiennent de manière analogue et de façon itérative (par exemple, l'estimation des indices du second ordre requiert l'estimation préalable des indices d'ordre 1).

Nous présenterons ici l'estimation des indices de Sobol par la méthode de Monte Carlo. Cette approche d'estimation fut proposée par Sobol (2001). Une autre approche pour estimer les indices de Sobol est la méthode FAST qui repose sur la décomposition de Fourier du modèle ($\hat{f}$) à expliquer. Pour la présentation détaillée de cette approche, nous renvoyons à Tissot et Prieur (2012).

• Estimation par la méthode de Monte Carlo

L'approche d'estimation par la méthode de Monte Carlo est basée sur la réécriture de l'indice

$$S_j = \frac{\mathbb{V}ar[\mathbb{E}(Y \mid X_j)]}{\mathbb{V}ar[\mathbb{E}(Y \mid X)]} = \frac{\mathbb{V}ar[\mathbb{E}(Y \mid X_j)]}{\mathbb{V}ar[f(X)]}$$

avec $X = (X_1, \cdots, X_p)$ et $f(X) = \mathbb{E}(Y \mid X)$ comme suit :

$$S_j = \frac{Cov(f(X),\, f(\tilde{X}^*))}{\mathbb{V}ar[f(X)]} \tag{4.37}$$

où $\tilde{X}^* = (X_1^*, \cdots, X_{j-1}^*, X_j, X_{j+1}^* \cdots, X_p^*)$, avec X_k^* une réplication indépendante et de même

loi que X_k (pour tout $k \neq j$).

Une fois cette réécriture posée, par des simulations de Monte Carlo on génère X^* une copie de X. On procède comme suit :

1− Considérons deux échantillons de taille n indépendants et identiquement distribués issus de $\mathbb{P}_X$ généré par la méthode de Monte Carlo,

$$X = \left\{x^{(i)} = (x_1^{(i)}, \cdots, x_p^{(i)})\right\}_{i=1}^n \quad \text{et} \quad X^* = \left\{x^{*(i)} = (x_1^{*(i)}, \cdots, x_p^{*(i)})\right\}_{i=1}^n$$

2− Ensuite on calcule :

$$\left\{f(x^{(i)}) = f(x_1^{(i)}, \cdots, x_p^{(i)})\right\}_{i=1}^n \text{ et } \left\{f(\tilde{x}^{*(i)}) = f(x_1^{*(i)}, \cdots, x_{j-1}^{*}{}^{(i)}, x_j^{(i)}, x_{j+1}^{*}{}^{(i)}, \cdots, x_p^{*(i)})\right\}_{i=1}^n$$

3− On remplace la covariance et la variance dans l'équation (4.37) par leur correspondant empirique et on obtient alors un estimateur de S_j :

$$\hat{S}_j = \frac{\left[\frac{1}{n}\sum_{i=1}^n f\left(x^{(i)}\right) f\left(\tilde{x}^{*(i)}\right)\right] - \left[\frac{1}{n}\sum_{i=1}^n f\left(x^{(i)}\right)\right]\left[\frac{1}{n}\sum_{i=1}^n f\left(\tilde{x}^{*(i)}\right)\right]}{\left[\frac{1}{n}\sum_{i=1}^n f\left(x^{(i)}\right)^2\right] - \left[\frac{1}{n}\sum_{i=1}^n f\left(x^{(i)}\right)\right]^2} \tag{4.38}$$

En remarquant que $\mathbb{E}[f(X)] = \mathbb{E}[f(\tilde{X}^*)]$, nous pouvons remplacer l'estimateur de $\mathbb{E}[f(\tilde{X}^*)]$ par celui de $\mathbb{E}[f(X)]$, et l'estimateur de S_j ci-dessus devient tout simplement :

$$\hat{S}_j = \frac{\left[\frac{1}{n}\sum_{i=1}^n f\left(x^{(i)}\right) f\left(\tilde{x}^{*(i)}\right)\right] - \left[\frac{1}{n}\sum_{i=1}^n f\left(x^{(i)}\right)\right]^2}{\left[\frac{1}{n}\sum_{i=1}^n f\left(x^{(i)}\right)^2\right] - \left[\frac{1}{n}\sum_{i=1}^n f\left(x^{(i)}\right)\right]^2} \tag{4.39}$$

4− Une fois les indices de Sobol de premier ordre tous estimés, l'on peut également estimer les indices S_u, d'ordre supérieur ou égale à 2.

On commence par réécrire :

$$S_u = \frac{\mathbb{V}ar[\mathbb{E}(Y \mid (X_j : j \in u))] + \sum_{v \subset u}(-1)^{|u|-|v|}\,\mathbb{V}ar[\mathbb{E}(Y \mid (X_j : j \in v))]}{\mathbb{V}ar[f(X)]}$$

comme suit :

$$S_u = \frac{\mathbb{V}ar[\mathbb{E}(Y \mid (X_j : j \in u))]}{\mathbb{V}ar[f(X)]} + \sum_{v \subset u}(-1)^{|u|-|v|}\,\frac{\mathbb{V}ar[\mathbb{E}(Y \mid (X_j : j \in v))]}{\mathbb{V}ar[f(X)]} \tag{4.40}$$

Ainsi, pour estimer chaque ratio de variances dans la formule de S_u (4.40), il suffit donc de considérer deux n-échantillons :

$$\left\{x^{(i)} = \left(x_v^{(i)}, x_{\setminus v}^{(i)}\right)\right\}_{i=1}^n \quad \text{et} \quad \left\{x^{*(i)} = \left(x_v^{*(i)}, x_{\setminus v}^{*}{}^{(i)}\right)\right\}_{i=1}^n, \quad \forall v \subseteq u$$

On calcule ensuite :

$$\left\{f(x^{(i)}) = f(x_v^{(i)}, x_{\setminus v}^{(i)})\right\}_{i=1}^n \quad \text{et} \quad \left\{f(\tilde{x}^{*(i)}) = f(x_v^{(i)}, x_{\setminus v}^{*}{}^{(i)})\right\}_{i=1}^n, \quad \forall v \subseteq u$$

Enfin, pour obtenir $\hat{S}_u$ on remplace chaque $\dfrac{\mathbb{V}ar[\mathbb{E}(Y \mid (X_j : j \in v))]}{\mathbb{V}ar[f(X)]}$, $(v \subseteq u)$ dans l'équation (4.40) par son correspondant empirique.

• Remarques

– Les indices obtenus par calcul numérique sont des estimations non-biaisées et convergent ;

– Risque d'imprécision suivant la qualité et la taille de l'échantillonnage : (i) Pour des facteurs peu influents, l'estimation de l'indice peut être négatif ; (ii) l'estimation de l'indice de sensibilité total S_{T_j} peut être inférieure à celle de l'indice de premier ordre S_j.

– Pour toutes ces raisons, il es donc nécessaire d'évaluer la précision des estimation faite, par répétition ou bootstrap.

□ Avantages et inconvénients des indices de Sobol

• Avantages

– *Interprétation intuitive* : l'indice de Sobol représente la part de variance du modèle expliquée par une caractéristique ou un groupe de caractéristiques. En outre, l'interprétation des indices de Sobol est très intuitive, car grâce à la décomposition fonctionnelle de la variance, leur somme est égale à 1, et étant tous positifs, plus l'indice sera grand (proche de 1), plus la variable (ou le groupe des variables) aura de l'importance. Elles paraissent donc, dans une certaine mesure, mieux interprétables que les H-statistique de Friedman qui sont quant à elle non bornées.

– *Grandeur sans unité* : l'indice de Sobol d'une caractéristique est une grandeur sans unité, elle peut donc être comparé d'un modèle à l'autre.

– Disponibilité d'une librairie R nommée *sensitivity* permettant de calculer les indices de Sobol.

• Inconvénients

– *Coût informatique élevé* : le nombre total d'indice de sensibilité (tout ordre compris) dans un problème à p caractéristiques est de $2^p - 1$. Par exemple pour $p = 10$, l'algorithme doit calculer $2^{10} - 1 = 1023$ indices de Sobol. Or, en actuariat par exemple, l'on a souvent affaire à des variables catégorielles ayant plusieurs modalités et lors la modélisation, chacune de ces modalités est traitée comme une variable à part entière. Ce qui peut rapidement faire exploser la valeur de $2^p - 1$.

– Le calcul des indices de Sobol repose sur l'hypothèse d'indépendance entre les caractéristiques. Ce qui est rarement vérifiée en pratique.

Chastaing *et al.* (2012) ont introduit les *indices de Sobol généralisée* qui sont mieux adaptée à l'analyse de sensibilité globale lorsque l'hypothèse d'indépendance entre les caractéristiques est violée. Cependant, nous les approfondirons pas dans ce mémoire.

4.3 Méthodes d'explication locales

Dans cette section nous présentons les outils les plus populaires à l'heure actuelle pour la génération des interprétations au niveau des instances individuelles du jeu de données.

4.3.1 Analyse de l'effet local des caractéristiques : boîte à outils ICE

Les diagrammes ICE ont été introduit par Goldstein *et al.* (2015). L'algorithme ICE (en anglais, *Individual Conditional Expectation*) est un outil de visualisation du modèle estimé par n'importe

quel algorithme d'apprentissage automatique supervisé.

Les diagrammes ICE étendent les diagrammes de dépendance partielle (PDP) de Friedman *et al.* (2001). Les tracés ICE affinent le tracé de dépendance partielle (PDP) en représentant graphiquement la relation fonctionnelle entre la réponse prédite et un ensemble de caractéristiques pour chaque instance individuel du jeu de données.

Elles permettent de mettre en évidence les effets d'hétérogénéité des caractéristiques sur la variable prédite.

Procédure de la méthode ICE

Considérons la matrice $n \times p$ de n réalisations du vecteur caractéristiques $X = \left\{(x_S^{(i)}, x_C^{(i)})\right\}_{i=1}^{n}$ avec $S \subseteq \{1, \cdots, p\}$ l'ensemble des caractéristiques pour lesquelles l'on souhaite tracer les courbes ICE et $C = \{1, \cdots, p\} \setminus S$ (en pratique $Card(S) = 1$). Soit $\hat{f}$ le modèle à interpréter. La procédure de tracé du diagramme ICE pour le sous-ensemble de caractéristiques S se résume en l'algorithme suivant :

► **Algorithme ICE**

---

1 : L'algorithme prend en entrée la matrice X de caractéristiques constituée des données d'entraînement du modèle $\hat{f}$. La matrice X est d'ordre $n \times p$. En entrée, l'on spécifie également le sous-ensemble S d'index de caractéristiques pour lesquelles l'on souhaite tracer les n courbes ICE $\hat{f}_S^{(1)}, \cdots, \hat{f}_S^{(n)}$.
2 : pour i allant de 1 à n **faire :**
3 : $\quad \hat{f}_S^{(i)} \longleftarrow 0_{n\times 1}$
4 : $\quad x_C \longleftarrow X[i, C]$ $\quad \sharp$ on fixe x_C pour l'instance i
5 : $\quad$ **pour** l allant de 1 à n **faire :**
6 : $\quad\quad x_S \longleftarrow X[l, S]$
7 : $\quad\quad \hat{f}_S^{(i)}[l] \longleftarrow \hat{f}([x_S, x_C])$
8 : $\quad$ **fin pour**
9 : fin pour
10 : on obtient en sortie les : $\hat{f}_S^{(1)}, \cdots, \hat{f}_S^{(n)}$
11 : pour chaque $i = 1, \cdots, n$: on trace la courbe $\mathcal{C}^{(i)}$ des points $\left\{\left(x_S^{(l)}, \hat{f}_S^{(i)}[l]\right)\right\}_{l=1}^{n}$.

ICE centré ou c-ICE : une première variante des courbes ICE

Généralement, les courbes ICE commencent à des niveaux différents (car les $\hat{f}_S^{(i)}[1]$ ne sont pas tous égaux lorsqu'on fait varier i).

Par conséquent les courbes individuelles sont généralement empilées les unes sur les autres. Ce qui peut masque l'effet hétérogène de certaines caractéristiques dans le modèle.

Pour accroître la lisibilité du diagramme, une solution consiste à centrer toutes les courbes à un certain point initial et à n'afficher que l'écart dans la prédiction en partant de ce point. Le nouveau diagramme obtenu à l'issue de cette procédure de remise à niveau est appelé le *ICE centré (c-ICE).*

Pour chaque courbe $\mathcal{C}^{(i)}$ dans le diagramme ICE, son correspondant dans le diagramme c-ICE est obtenu suite au tracé de la courbe de la fonction :

$$\hat{f}_{cent}^{(i)}(\cdot) = \hat{f}^{(i)}(\cdot) - \mathbf{1}\hat{f}(x^*, x_C^{(i)}) = \hat{f}(\cdot, x_C^{(i)}) - \mathbf{1}\hat{f}(x^*, x_C^{(i)}) \tag{4.41}$$

où **1** est un vecteur de 1 avec le nombre approprié de dimensions (généralement une ou deux), x^* est le point d'ancrage choisi (généralement le minimum des valeurs de x_S, de sorte que toutes les courbe $\mathcal{C}_{cent}^{(i)}$ prennent leur origine en zéro (0)).

En procédant de la sorte, on obtient un tracé qui met bien en évidence les éventuels effets d'hétérogénéité x_S sur $\hat{f}$.

ICE dérivé ou d-ICE : une seconde variante des courbes ICE

Une autre variante du diagramme ICE permettant de détecter visuellement la présence d'effets d'interaction entre les caractéristiques est le *diagramme de la dérivée partielle de $\hat{f}$ par rapport à x_S*.

Supposons le cas où il n'existe pas d'interaction entre les sous-ensemble de covariables X_S et X_C dans le modèle ajusté $\hat{f}$. Alors, $\hat{f}(x) = \hat{f}(x_S, x_C) = g(x_S) + h(x_C)$ et par conséquent :

$$\frac{\partial \hat{f}(x)}{\partial x_S} = g'(x_S) \text{ (Lorsque card(S)=1)} \tag{4.42}$$

Dans ce cas, toutes les courbes $\mathcal{C}^{(i)}$ du diagramme ICE ont la même forme (puisque ayant la même pente $g'(x_S)$ en chaque point x_S) et sont juste disposées les unes au dessus des autres suivant la valeur de l'intercept $h(x_C^{(i)})$.

Cependant, étant donné le caractère généralement touffu des diagrammes ICE, il est difficile d'identifier visuellement est-ce que les courbes ont toutes les mêmes pentes ou non, raison pour laquelle, l'on choisit plutôt de représenter directement les courbes ICE d'un estimateur de la dérivée partielle de $\hat{f}$ par rapport à x_S $\left(\dfrac{\partial \hat{f}}{\partial x_S}\right)$.

Le diagramme résultant est appelé *diagramme ICE dérivé* en abrégé *d-ICE*. Il s'obtient par exécution de l'algorithme suivant qui retourne en sortie les dérivées $d\hat{f}_S^{(1)}, \cdots, d\hat{f}_S^{(n)}$ des n fonctions ICE :

► **Algorithme d-ICE**

---

1 : L'algorithme prend en entrée la matrice X des caractéristiques. X est de taille $n \times p$ et composée de données d'entraînement du modèle $\hat{f}$. Les fonctions $\hat{f}_S^{(1)}, \cdots, \hat{f}_S^{(n)}$ issues de l'algorithme ICE. Un opérateur D qui permet de générer une approximation numérique de la dérivée des fonctions.

2 : **pour** i allant de 1 à n **faire :**
3 : $\quad d\hat{f}_S^{(i)} \longleftarrow 0_{n\times 1}$
4 : $\quad x_C \longleftarrow X[i, C]$ $\quad \sharp$ on fixe x_C pour l'instance i
5 : $\quad\quad$ **pour** l allant de 1 à n **faire :**
6 : $\quad\quad\quad x_S \longleftarrow X[l, S]$
7 : $\quad\quad\quad d\hat{f}_S^{(i)}[l] \longleftarrow D\left[\hat{f}_S^{(i)}(x_S)\right]$
8 : $\quad\quad$ **fin pour**
9 : **fin pour**
10 : on obtient en sortie les : $d\hat{f}_S^{(1)}, \cdots, d\hat{f}_S^{(n)}$
11 : pour chaque $i = 1, \cdots, n$: on trace la courbe $\mathcal{C}_{der}^{(i)}$ des points $\left\{\left(x_S^{(l)}, d\hat{f}_S^{(i)}[l]\right)\right\}_{l=1}^{n}$.

□ **Avantages et inconvénients**

• **Avantages**

Les avantages des diagrammes ICE sont nombreux. Citons-en quelques uns :

– Leur caractère intuitif : chaque ligne représente pour chaque individu, l'évolution de la prédiction lorsque les valeurs de la caractéristique varient ;

– Les diagrammes ICE sont également extrêmement rapides à mettre en place ;

– Les courbes ICE mettent en évidence les effets hétérogènes des caractéristiques sur la variable cible ;

– Les courbes ICE peuvent également permettre de visualiser les effets d'extrapolations dans le modèle ajusté ;

– Les graphiques d-ICE permettent de mettre en évidence les éventuelles régions d'interaction des caractéristiques ;

– On peut comparer les graphiques ICE de plusieurs modèles différent pour une caractéristique donnée.

– Disponibilité d'un package R entièrement dédié à la mise en oeuvre des courbes ICE : le package *ICEbox*.

• **Inconvénients**

Comme inconvénients, nous pouvons relever que :

– Bien que le fait de retenir une courbe par individu permet de mettre en évidence les effets hétérogènes des caractéristiques, cela conduit parfois à des graphiques potentiellement trop chargés et peu lisibles. Dans de pareilles situations, une solution pourrait être de tracer les courbe ICE uniquement pour un échantillon réduit d'observations.

– Les courbes ICE ne peuvent rigoureusement afficher qu'une seule caractéristique à la fois ($Card(S) = 1$).

– Comme avec les PDP, si la caractéristique d'intérêt est fortement corrélée à une autre caractéristique, certains points des courbes individuelles $\mathcal{C}^{(i)}$ peuvent être des points irréalistes.

En cas de potentielle forte corrélation entre les caractéristiques, il est conseillé de faire recours en priorité à la méthode ALE pour évaluer l'effet d'une caractéristique sur la variable cible.

4.3.2 Modèles de substitution locaux : LIME, LS

□ **Principe général**

Les modèles de substitution locaux sont des modèles interprétables qui sont utilisés pour expliquer les prédictions individuelles des modèles d'apprentissage statistique complexe.

La recette pour former un modèle de substitution local est assez simple et se rapproche plus ou moins de celle des modèles de substitution globaux décrit en annexe B.1. Il suffit de suivre dans l'ordre les cinq (05) étapes suivantes :

1 : *Choix de l'instance de données à interpréter* : sélectionner l'instance d'intérêt pour laquelle l'on souhaite avoir une explication de la prédiction.

2 : *Échantillonnage des données pour l'ajustement du modèle de substitut local* : ensuite, l'on perturbe le jeu de données en ajoutant un bruit aléatoire (échantillonné suivant la loi $\mathcal{N}(0, 1)$) aux valeurs de l'ensemble des caractéristiques. On détermine ensuite les prédictions de ces nouveaux points à l'aide du modèle à interpréter.

3 : *Pondération des instances échantillonnées* : on pondère les instances de l'échantillon obtenu à l'étape précédente à l'issue de la perturbation du jeu de données initial. Les poids accordés sont proportionnelles à la proximité entre les instances et l'instance à interpréter (fixé à l'étape 1).

4 : *Ajustement du modèle de substitut local* : on entraîne un modèle interprétable pondéré sur l'ensemble de données obtenu à l'étape 2, avec comme variable cible, les prédictions obtenues à l'étape 2 et en utilisant les poids calculés à l'étape 3.

5 : *Interprétation du modèle* Enfin, on interprète la prédiction de notre instance d'intérêt générée par le modèle complexe en interprétant le modèle local entraîné à l'étape 4.

□ Formalisation mathématique

Mathématiquement, les modèles de substitution locaux avec contrainte d'interprétabilité peuvent être exprimés comme suit :

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\ \mathcal{L}(f,\, g,\, \pi_x) + \Omega(g) \tag{4.43}$$

Autrement dit, le modèle d'explication $\xi(x)$ pour l'instance individuelle x est un modèle $g \in G$ (avec G une classe de modèles potentiellement interprétables, par exemple, les modèles linéaires, les arbres de décisions, ou les règles de décisions) qui minimise une perte $\mathcal{L}$ (par exemple, la perte quadratique) de manière à garantir à la fois l'interprétabilité et la fidélité locale.

De plus, g doit être assez simple pour rester suffisamment interprétable, pour cette raison on impose la contrainte de complexité $\Omega(g)$ au programme d'optimisation. Par exemple, on peut choisir $\Omega(\cdot)$ comme le nombre de coefficients non nuls (contrainte de régularisation) ou alors, la profondeur de l'arbre (lorsque G = classe des arbres de décision).

Quant à f, il s'agit du modèle à expliquer. Enfin, $\pi_x(z)$ est une fonction de pondération qui tient compte de la proximité entre les instances z obtenues à l'issue de la perturbation des données et l'instance d'intérêt x : elle permet de définir le voisinage de x.

La plupart des méthodes de substitution locales disponibles dans la littérature diffèrent essentiellement au niveau de l'étape 3 lors de la construction des poids.

L'une des méthodes de substitution locales les plus utilisées aujourd'hui est la méthode LIME (Local Interpretable Model-agnostic Explanations) parue en 2016. Elle permet d'expliquer les prédictions individuelles de n'importe quel modèle aussi complexe soit-il, d'une manière interprétable et relativement fidèle au modèle à interpréter.

Cependant, bien qu'étant très intuitive la méthode LIME ne circonscrit pas toujours le "bon" voisinage à utiliser pour générer des explications locales robustes. A cet effet, la méthode LS (Local Surrogate) viendra deux ans plus tard (2018) après la parution tenter d'améliorer la méthode LIME en proposant une nouvelle méthodologie pour le choix d'un voisinage plus approprié pour la générations d'explications locales plus robustes et fiables.

4.3.2.1 LIME : Local Interpretable Model-agnostic Explanations

□ Spécificités de LIME

Pour obtenir la méthode LIME, il suffit de remplacer dans l'équation 4.43 :

− G par la classe des modèles linéaires, c'est-à-dire g sous la forme $g(z^{'}) = \omega_g \cdot z^{'}$;

− $\pi_x(z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right)$ un noyau de lissage exponentielle pour définir le voisinage, avec D une fonction de distance appropriée (par exemple, la *similarité cosinus* pour des données textuelles, ou la distance *euclidienne* pour des données numériques) ; σ le paramètre de largeur du noyau ;

– $\mathcal{L}$: l'erreur quadratique pondérée par π_x définie par :

$$\mathcal{L}(f,\, g,\, \pi_x) = \sum_{z,\, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z') \right)^2 \tag{4.44}$$

où $\mathcal{Z}$ est l'échantillon obtenu après la perturbation du jeu de données à l'étape 2 telle que décrit en 4.3.2.

– En ce qui concerne le choix optimal de la fonction de contrainte de complexité, il varie suivant la nature de la tâche de classification.

Pour la classification de données numériques, Ω pourrait correspondre à une contrainte de régularisation de type LASSO. Pour des données textuelles, $\Omega(g) = \infty \mathbb{I}\{\|\omega_g\|_0 > K\}$.

La figure 4.5 ci dessous présente l'intuition de LIME. Le modèle complexe f est représentée par l'arrière-plan bleu/rose. La croix rouge en gras représente l'instance que l'on souhaite expliquer. L'algorithme LIME échantillonne les instances tout autour de ce point. La ligne pointillée représente le modèle linéaire appris et qui est localement fidèle au modèle f au voisinage de l'instance à interpréter.

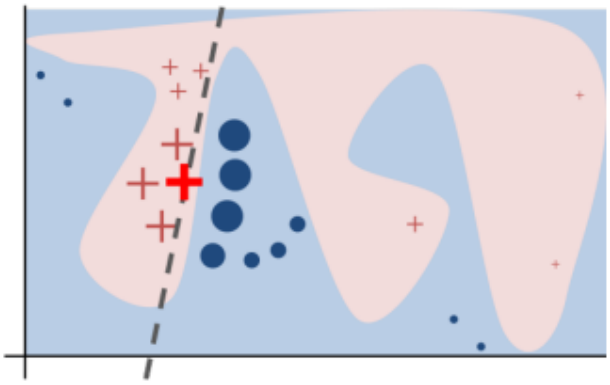


FIGURE 4.5 : *Illustration du principe de fonctionnement de la méthode LIME (Issue de Ribeiro et al. (2016)).*

□ Avantages et limites de LIME

• Avantages

– LIME est l'une des rares méthodes d'interprétabilité qui fonctionne à la fois pour les données tabulaires, textuelles et les données images ;

– La flexibilité dans le choix du type de modèle de substitution à considérer : arbre de décision, modèles linéaires, etc.

– Implémentation facile grâce à la disponibilité des packages *LIME* ou *iml* sous R.

• Limites

Les limites de la méthode LIME sont nombreuses :

– *Manque de fiabilité* : Slack *et al.* (2020) démontrent que les techniques d'explications *post hoc* qui s'appuient sur des perturbations des caractéristiques telles que LIME ne sont pas fiables. Ils illustrent comment les explications de LIME peuvent être manipulées par le *data scientist* pour

cacher les biais. Cette possibilité de manipulation des explications issues de LIME remet en cause leur fiabilité.

– *Faible robustesse des explications* : une autre limite évoquée par Alvarez-Melis et Jaakkola (2018) est la question de la robustesse des explications issue de LIME. Ils montrent que pour deux instances assez voisines, leurs explications issues LIME ne sont pas toujours semblables.

– Zhang *et al.* (2019) reviennent également sur ce problème de robustesse des explications basées sur la méthode LIME. Ils évoquent quelques sources d'incertitude à l'origine de la faible robustesse de la méthode LIME à savoir :

(i) Le caractère aléatoire et peu rigoureux de la procédure d'échantillonnage. Les points de données sont échantillonnés à partir d'une distribution gaussienne en ignorant la corrélation initiale existante entre les entités. Ce qui pourrait conduire à des points de données improbables qui peuvent ensuite être utilisés pour entraîner le modèle de substitution locale. Ce qui biaise les explications issues du modèle.

(ii) La grande sensibilité des explications au choix du paramètre σ défini ci-dessus ; Le choix de ce paramètre est identifié comme le plus gros problème de LIME.

4.3.2.2 LS : Local Surrogate

Les limites de la méthode LIME liées à la difficulté du choix du bon voisinage à considérer pour la construction du modèle de substitution local ont données lieu à l'introduction d'une nouvelle classe de modèles de substitution locaux appelée *Local Surrogate (LS)*. L'approche LS est très similaire à l'approche LIME dans leur principe de fonctionnement. La seule différence se situe au niveau de l'étape 2 de l'algorithme de construction du modèle de substitut local présenté en 4.3.2.

Dans cette section, nous nous appuyons principalement sur Laugel *et al.* (2018), dans lequel les auteurs montrent l'impact majeur du choix de la stratégie d'échantillonnage adéquate pour générer les données de substitution locale sur la qualité de l'approximation du modèle de substitution en s'appuyant sur la méthode LIME, avant même d'introduire la nouvelle méthode : Local surrogate. Plus précisément, ils montrent que le problème de LIME n'est pas lié à une simple paramétrisation (σ) permettant de contrôler l'étendue du voisinage échantillonné : par exemple, il montre que centrer l'échantillonnage sur l'instance de la prédiction à expliquer peut ne pas être le meilleur endroit pour approximer la limite de décision de la boîte noire.

L'intuition de cette nouvelle approche est la suivante : elle vise à déterminer le voisinage de l'instance à expliquer en dehors duquel le modèle de substitut local n'est plus assez précis pour expliquer le modèle complexe. Une fois le voisinage bien délimité, on échantillonne les données dans ce voisinage pour ajuster le modèle de substitution local. Une fois le modèle ajusté, on l'interprète pour obtenir les explications de la prédiction de notre instance d'intérêt.

□ Procédure de construction du voisinage local avec l'approche LS

Étant donnée l'instance x pour laquelle l'on souhaite expliquer la prédiction issue de la boîte noire f, Laugel *et al.* (2018) propose de suivre la démarche suivante :

– Premièrement on détermine l'instance x_{bord} la plus proche de x, mais dont la valeur prédite diffère de celle x. Cela revient à résoudre le programme d'optimisation suivant :

$$x_{bord} = \underset{z \in X_x}{argmin} \|x - z\| \quad (\text{avec } X_x = \{a \in X,\, f(a) \neq f(x)\}) \tag{4.45}$$

En pratique la résolution de ce programme d'optimisation se fait grâce à l'algorithme *Growing-Spheres* introduit par Laugel *et al.* (2017).

Le principe de cet algorithme est le suivant : on cherche la solution itérativement dans des hypersphères centrées en x et de rayons croissant. Pour chaque rayon fixé, l'on génère des points z dans l'hypersphère, puis on calcule leur prédiction avec le modèle de boîte noire f et vérifie ceux qui tombent de l'autre coté de la frontière (c'est-à-dire les instances z telles que $b(z) \neq b(x)$) et on choisit x_{bord} (confère figure 4.6).

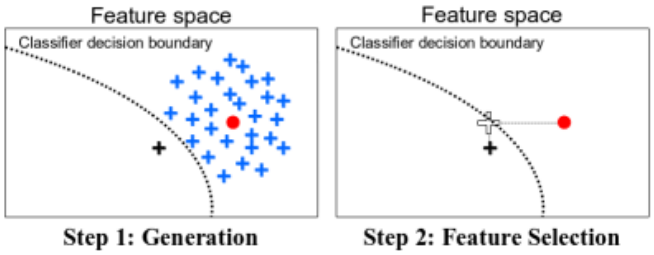


FIGURE 4.6 : *Illustration des sphères de croissance : Le cercle rouge représente l'observation à interpréter, les signes plus les observations générées par les sphères de croissance (bleu pour les alliés, noir pour les ennemis). Le plus blanc est l'ennemi final x_{bord} utilisé pour générer des explications (Issue de Laugel* et al. *(2017)).*

– Une fois le choix de x_{bord} fait, l'échantillon d'entraînement du modèle de substitution local est obtenu en procédant au tirage aléatoire de n instances suivant la loi uniforme dans la sphère $\mathcal{S}$ de rayon r_x et de centre x_{bord}. Désignons par X_{s_x} cet échantillon.

– On entraîne enfin le modèle de substitution local sur le jeu de donnée (X_{s_x}, Y_{s_x}), avec pour covariables les X_{s_x} et pour variable cible $Y_{s_x} = f(X_{s_x})$.

4.3.3 SHAP : SHapley Additive exPlanations

Cette section est basée sur l'article Lundberg et Lee (2017). Nous avons précédemment abordé les valeurs de Shapley dans la sous-section 4.2.2 pour la mesure de l'importance globale des variables. Dans cette section nous les utilisons dans le cadre de la génération d'explications locales. Les valeurs de Shapley permettent d'apporter des éléments de réponses aux questions de savoir : comment les différentes caractéristiques influent-elles sur les résultats de la prédiction pour une instance individuelle donnée ? Quelles sont les variables les plus importantes qui influent sur les résultats de la prédiction pour une instance spécifique donnée ?

□ De la théorie des jeux et des valeurs de Shapley à l'apprentissage automatique explicable

Commençons d'abord par présenter le principe de la théorie des jeux afin de mieux comprendre comment elle est utilisée pour expliquer les prédictions des modèles d'apprentissage statistique.

La théorie des jeux est la branche des mathématiques appliquées consacrée à l'étude théorique des interactions sociales entre des acteurs en concurrence dans un contexte stratégique, chacun des acteurs désirant maximiser ses gains. Ainsi, un jeu est donc la donnée (i) d'un ensemble d'agents en concurrence encore appelés "joueurs" ; (ii) des stratégies admissibles ; (iii) des gains associés à chaque stratégie. On parle de jeu coopératif lorsque les joueurs peuvent former des coalitions entre eux afin de maximiser leur gain mutuel avant de se le repartir équitablement selon la contribution de chaque joueur à la formation du gain. Les valeurs de Shapley donnent une formule fiable qui permet de calculer la contribution au jeu de chaque joueur.

Dans le contexte de l'apprentissage statistique, les valeurs des caractéristiques d'une instance de données servent de membres à la coalition. Autrement dit, par analogie à la théorie des jeux, les caractéristiques représentent les "joueurs". Les valeurs de Shapley permettent de repartir le "gain" (valeur de la prédiction) entre les caractéristiques de manière équitable (suivant leur contribution à la formation de la valeur prédite).

□ Valeur de Shapley

Soient f notre modèle d'apprentissage statistique et une instance $x = (x_1, \cdots, x_p)$ dont on souhaite avoir les explications de la prédiction par f. La valeur de Shapley de la variable explicative X_j, $j \in \{1, \cdots, p\}$ est définie comme la contribution marginale de la variable X_j à la prédiction parmi toutes les "coalitions" concevables. Elle définie par la formule :

$$\phi_j(\Delta^x) = \sum_{S \subseteq F \setminus \{X_j\}} \frac{|S|!(p - |S| - 1)!}{p!} \times \underbrace{[\Delta^x(S \cup \{X_j\}) - \Delta^x(S)]}_{\text{contribution marginale}} \tag{4.46}$$

avec, F l'ensemble des caractéristiques et $\Delta^x(\cdot)$ telle que pour tout $S = \{X_{i1}, \cdots, X_{is}\} \subseteq \{X_1, \cdots, X_p\}$ on a :

$$\Delta^x(S) = \underbrace{\mathbb{E}\left[f(X_1, \cdots, X_p \mid X_{i1} = x_{i1}, \cdots, X_{is} = x_{is}\right]}_{:= f_x(S)} - \underbrace{\mathbb{E}\left[f(X_1, \cdots, X_p\right]}_{:= f_x(\emptyset)} \tag{4.47}$$

Il s'agit de l'amélioration de la prédiction causé par l'observation des variables explicatives contenues dans S par rapport à la situation où on observe aucune variable explicative du tout.

Remarques

– S'étant donné une coalition S ne contenant pas la caractéristique X_j, la contribution marginale évalue la mesure dans laquelle le modèle change (s'améliore ou se détériore) lorsqu'on ajoute la caractéristique X_j à la coalition S.

– La valeur SHAP ($\phi_j(x)$) est la contribution marginale moyenne d'une caractéristique X_j sur toutes les coalitions possibles.

– Plus la valeur de $\phi_j(x)$ est élevée en valeur absolue, plus est grande l'influence que la variable X_j a sur la prédiction de l'instance x.

– Lorsque le signe de $\phi_j(x)$ est positif alors la variable X_j contribue à augmenter la valeur de la prédiction $f(x)$, sinon elle contribue à la diminuer.

– La somme des contributions des variables représente la différence entre la prédiction du modèle et la prédiction attendue du modèle sans aucune information sur les réalisations des variables explicatives.

– À partir des valeurs de Shapley, on peut mesurer les effets d'interaction de second ordre entre les caractéristiques en utilisant la formule suivante :

$$\phi_{j,k}(x) = \sum_{S \subseteq F \setminus \{X_j, X_k\}} \frac{|S|!(p - |S| - 2)!}{2(p-1)!} \times \Delta^x_{j,k}(S) \tag{4.48}$$

avec

$$\Delta^x_{j,k}(S) = f_x(S \cup \{j, k\}) - f_x(S \cup \{j\}) - f_x(S \cup \{k\}) + f_x(S) \tag{4.49}$$

□ **Approche d'estimation des valeurs de Shapley**

Plusieurs approches ont été proposé dans la littérature pour l'estimation des valeurs de Shapley :

(a) KernelSHAP

KernelSHAP estime pour une instance x les contributions de chaque caractéristique à la formation de la valeur prédite $f(x)$. KernelSHAP a été développé par Lundberg et Lee (2017) et elle s'inspire de l'approche d'estimation de LIME avec un choix particulier de pondération.

(b) TreeSHAP

Proposé par Lundberg et al. (2018), TreeSHAP est une approche d'estimation des valeurs SHAP basée sur les arbres de décision, les forêts aléatoires et les arbres boostés par gradient. TreeSHAP a été présenté comme une alternative rapide à KernelSHAP.

(c) LinearSHAP

Pour les modèles linéaires, lorsqu'on suppose l'indépendance des caractéristiques d'entrée, les valeurs SHAP peuvent être directement déduite à partir des coefficients de poids du modèle. On montre que pour $f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$, la valeur SHAP de l'instance x pour la caractéristique X_j est égale à :

$$\phi_j(x) = \beta_j(x_j - \mathbb{E}[X_j]) \text{ et } \phi_0(x) = \beta_0 \tag{4.50}$$

En somme, dans ce chapitre nous avons présenté les aspects théoriques des méthodes d'interprétabilité *post hoc* les plus populaires à l'heure actuelle. Dans le chapitre 5 nous mettrons en oeuvre ces méthodes sur un cas pratiques d'assurance automobile afin d'interpréter les modèles complexes d'apprentissage statistique qui seront mis en place pour la modélisation de la fréquence de sinistre.

Chapitre 5
Application à la tarification automobile

Dans ce chapitre, nous nous plaçons dans un contexte de tarification automobile. Nous calculons la prime pure à partir des données historiques de sinistres mises à notre disposition. Pour la modélisation de la fréquence et du coût de sinistre, nous avons mis en place trois types de modèles : un modèle de type GLM, un modèle de forêt aléatoire et un modèle de type LocalGLMnet (famille de modèles assez récente, parue en 2022).

L'objectif de ce chapitre est double :

– Premièrement, étant donné que le jeu de données historiques de sinistres mis à notre disposition contient des variables télématiques liées aux comportements de conduite des assurés, nous étudions dans quelle mesure la prise en compte de ces données télématiques dans la modélisation de la sinistralité permet d'améliorer la précision des modèles.

– Dans un second temps, nous revenons à la problématique principale du mémoire : celle de l'interprétabilité. Nous utilisons les outils d'interprétabilité présentés au chapitre 4 pour interpréter l'un des modèles complexes mis en place pour la modélisation de la fréquence de sinistre. L'objectif est de recueillir autant d'informations sur le fonctionnement sous-jacent de ce modèle complexe de manière à rendre son processus de prédiction aussi transparent que celui d'un modèle GLM.

5.1 Données de l'étude

5.1.1 Source des données

Nous utilisons la base de données synthétique pour la télématique des conducteurs en assurance automobile créé par So *et al.* (2021) disponible sur le site web [lien] du département de mathématique de l'université du Connecticut, aux Etats-Unis.

L'ensemble de données contient 100 000 polices d'assurance pour lesquelles sont renseignées les deux informations relatives à l'expérience de sinistres du conducteur, à savoir : la fréquence et la sévérité de ses sinistres ; de onze variables de risque classiques telles que la durée de la police, l'âge et le sexe du conducteur ; et de trente neuf variables liées à la télématique du conducteur, incluant par exemple le nombre total de kilomètres parcourus durant la période couverture, le nombre de freinages soudains ou d'accélérations soudaines au cours de la période de couverture.

La construction du jeu de données se base sur des données réelles acquises auprès d'un assureur de la place. Ce dernier proposait à ses clients, depuis 2013, un programme *d'assurance basée sur l'utilisation* (en anglais *Usage-based insurance, UBI*). Il s'agit d'un type de police d'assurance automobile qui exploite l'émergence des nouvelles technologies avancées de collecte et de traitement de données massives pour améliorer la tarification, de sorte que le coût de l'assurance soit directement lié à l'utilisation du véhicule.

La période d'observation des données réelles à partir desquelles a été construite la présente base de données s'étend entre 2013 et 2016 (inclus), pour 70 000 polices observées.

La méthodologie de constitution du jeu de données final repose sur la méthode SMOTE. Cette méthode a été développée par Chawla *et al.* (2002). Elle consiste en deux étapes :

– Partant des 70 000 données réelles observées, on sur-échantillonne les observations minoritaires pour les différentes variables explicatives dans l'optique d'obtenir un nouvel échantillon plus grand. On obtient de nouvelles observations pour les variables explicatives, mais ces observations n'ont pas de valeurs pour les variables cibles de notre expérience.

– Pour déterminer les valeurs des variables cibles pour ces observations sur-échantillonnées, on fait recours à une imputation par régression. À partir des 70 000 données réelles, pour chaque variable cibles, on ajuste un modèle de régression neuronale de la variable cible sur les variables explicatives. On se sert ensuite de ce modèle de régression pour prédire les valeurs des variables cibles des observations sur-échantillonnées.

Une description plus détaillée de la méthodologie de construction de cette base de données est disponible dans l'article de So *et al.* (2021). Notons néanmoins qu'après avoir comparer le jeu de données réel et le jeu des données synthétique final, So *et al.* (2021) parviennent à la conclusion que le jeu de données synthétique imite exceptionnellement bien l'ensemble de données réelles. Ils suggèrent alors qu'on peut utiliser ces données synthétiques pour entraîner des modèles statistiques à la place de données réelles sans grandes crainte de perte d'objectivité des résultats obtenus.

5.1.2 Présentation des variables contenues dans le jeu de données

Le fichier de données à notre disposition contient au total 52 variables. Ces dernières peuvent être présentées en trois groupes :

(*a*) **Variables classiques**

– *Car.age* : l'âge du véhicule assuré, avec des valeurs comprises dans l'intervalle [-2, 20]. Les valeurs négatives sont rares et correspondent aux modèles de véhicule les plus récents, car leur achat peut parfois prendre jusqu'à deux ans à l'avance ;
– *Years.noclaims* : nombre d'années sans sinistre, avec des valeurs dans [0, 79] ;
– *Region* : statut de la région où habite le conducteur : Rural/ Urbain ;
– *Insured.sex* : sexe du conducteur assuré (Homme/Femme) ;
– *Martial* : état matrimonial du conducteur assuré (Célibataire/Marié) ;
– *Credit.score* : points de crédit du conducteur assuré ;
– *Car.use* : usage fait du véhicule (Privé, Commun, Agriculture, Commercial) ;
– *Insured.age* : l'âge du conducteur assuré (en année), avec des valeurs dans [16, 103] ;
– *Annual.miles.drive* : Miles annuels prévus à parcourir déclarés par le conducteur ;
– *Territory* : il fait référence au code d'emplacement territorial du véhicule, qui comporte 55 étiquettes dans {11, 12, 13, ..., 91} ;
– *Duration* : correspond à la période pendant laquelle l'assuré est couvert (en jours), avec des valeurs dans [27, 366].

(*b*) **Variables de télématique**

– *Annual.pct.drive* : pourcentage annualisée du temps passé sur la route. C'est-à-dire le nombre de jours de l'année pendant lesquels un assuré utilise un véhicule divisé par 365, ses valeurs sont comprises dans [0, 1] ;
– *Pct.drive.**xxx*** : pourcentage du jour de conduite ***xxx*** de la semaine : lundi/mardi/.../dimanche ;
– *Left.turn.intensity**xx*** : nombre de virages à gauche par 1000 miles avec intensité : 08/09/10/11/12 ;
– *Right.turn.intensity**xx*** : nombre de virages à droite par 1000 miles avec intensité : 08/09/10/11/12 ;
– *Total.miles.driven* : distance totale parcourue en miles ;

– *Avgdays.week* : nombre moyen de jours utilisés par semaine ;
– *Pct.drive.**x**hrs* : pourcentage de conduite en $\boldsymbol{x}$ heures : 2 heures/3 heures/4 heures ;
– *Pct.drive.rush**xx*** : pourcentage de conduite pendant les heures de pointe en xx : matinée/soirée ;
– *Pct.drive.wk**xxx*** : pourcentage de conduite pendant $\boldsymbol{xxx}$: la semaine/le weekend ;
– *Brake.**xx**miles* : nombre de freinages brusques d'intensité : 6/8/9/11/12/14 $mph_{/s}$ par *1000 miles* ;
– *Accel.**xx**miles* : nombre d'accélérations brusques d'intensité : 6/8/9/11/12/14 $mph_{/s}$ par *1000 miles*.

(*c*) **Variables de sinistralité**

– *AMT-Claim* : montant agrégé des réclamations pendant la période d'observation, avec des valeurs dans [0, 104 074.89] ;

– *NB-Claim* : nombre de sinistres de l'assuré pendant la période d'observation , avec des valeurs dans $\{0, 1, 2, 3\}$.

Remarque : parmi les variables télématiques, les groupes des variables *Pct.drive.**xxx*** et celui des variables *Pct.drive.wk**xxx*** sont des *variables dites de composition*. Cela signifie que la somme des sept variables *Pct.drive.**xxx*** (à savoir : *Pct.drive.lundi*, ..., *Pct.drive.dimanche*) ou des deux variables *Pct.drive.wk**xxx*** (à savoir *Pct.drive.wkday* et *Pct.drive.wkend*) vaut toujours 100%, pour chaque observation.

5.2 Analyses préliminaires des données de l'étude

5.2.1 Nettoyage de la base de données

A ce stade, il est question de s'assurer en amont de la qualité des données à notre disposition. De leur qualité dépendront la pertinence de tous les résultats et analyses qui seront obtenus par la suite.

Plus précisément, il s'agit de détecter et de traiter la présence d'éventuelles valeurs manquantes, aberrantes ou erronées au niveau du jeu de données de l'étude. Ignorer les valeurs manquantes peut parfois mener à des estimateurs fortement biaisés. Par ailleurs, les valeurs aberrantes, lorsqu'elles ne sont pas correctement traitées, elles peuvent mal conduire les analyses statistiques. Les valeurs erronées correspondent par exemple aux erreurs de saisie dans la base de données.

Pour notre jeu de données, aucune valeur manquante n'a été détectée pour l'ensemble des 52 variables. En ce qui concerne, les valeurs aberrantes, hormis la variable *Car.age* correspondant à l'âge du véhicule assuré, pour laquelle on observe des valeurs d'âge négatives, aucune autre valeur aberrante n'a été détectée. Comme évoqué précédemment, ces valeurs négatives correspondent aux modèles de véhicule les plus récents, car leur achat peut parfois prendre jusqu'à deux ans à l'avance. Ainsi, étant donné qu'il s'agit pour la plupart de police non active, et d'autant plus qu'elles ne représentent que moins de 2 % (précisément, 1.942 %) de la base de données, nous supprimerons ces lignes et travaillerons avec les 98 058 polices restantes.

5.2.2 Analyse et traitement des sinistres extrêmes

Une seconde étape primordiale à une bonne calibration de la prime pure consiste à identifier et traiter les montants de sinistres extrêmes. Pour ce faire, sur la figure 5.1, nous avons représenté la part des plus petits sinistres non-nuls dans la charge totale.

Sur cette figure 5.1, on constate des faits notables :

– 50 % des plus petits sinistres ne représentent que 12.55 % de la charge totale ;

– 75 % des plus petits sinistres représentent 33.30 % de la charge totale ;

– 90 % des plus petits sinistres représentent 57.02 % de la charge totale de sinistre.

Ces différents constats sont inéluctablement révélateurs d'une forte queue de distribution des montants de sinistres.

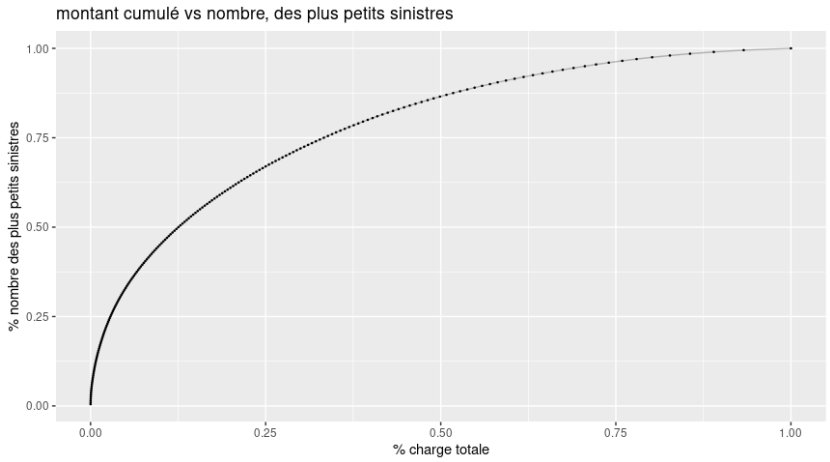


FIGURE 5.1 : *Part des plus petits sinistres dans la charge totale.*

Le tableau 5.1 résume également les informations concernant la distribution des montants de sinistres strictement positifs sur la période d'étude.

– On constate que 97.5% des montants de sinistres strictement positifs ont une valeur inférieure à 16 500 dollars, tandis qu'on dispose d'un seul sinistre dont le montant va au-delà de 100 000 dollars.

– De plus, en revenant à la figure 5.1, parmi les polices sinistrées (celles dont le montant de sinistre est strictement positif), les 5 % plus grands montants de sinistre représentent jusqu'à 30 % de la charge totale de sinistre.

Ces faits confirmeraient bien l'existence de potentiels de sinistres extrêmes[1] au sein de notre portefeuille de polices d'assurance.

min	quantile 25 %	médiane	moyenne	quantile 75%	quantile 97.5%	quantile 99%	max
0.77	768.37	1 964.53	3 457.63	3 972.35	16 330.89	25 502.39	104 074.89

TABLE 5.1 : *Résumé de la distribution des montants de sinistres strictement positifs.*

Lors de la modélisation du coût des sinistres nous pouvons utiliser une seule famille de loi sur l'ensemble des montants (y compris les extrêmes). Cependant, de par la sous-représentativité des sinistres extrêmes au sein du jeu de données, ils ne seront pas correctement pris en compte par de tels modèles. Or, il s'agit d'événements certes rares mais dont la survenance est très coûteuse à l'assureur : par conséquent, il est nécessaire ou même indispensable de minutieusement les prendre

[1]Sinistres extrêmes : il s'agit de sinistres rares mais dont la survenance conduit à des pertes financières importantes pour l'assureur

en considération dans nos modèles de calcul de prime. Ce qui nous poussent à dépasser le cadre d'une modélisation simpliste et d'étudier plus en détail la prise en compte des sinistres extrêmes dans nos modèles de sévérité.

Face aux sinistres extrêmes, pour améliorer la calibration des modèles de sévérité, les actuaires font généralement recours à deux méthodes :

• Une première solution rigoureuse consiste à enlever les montants atypiques, modéliser les sinistres attritionnels à l'aide d'un GLM standard, et modéliser les sinistres atypiques à l'aide d'une loi appropriée, généralement une loi de Pareto généralisée (GPD)[2].

Dans notre étude, la comparaison des quantiles empiriques des montants de sinistres contre les quantiles théoriques de la loi de Pareto standard (figure 5.2 a) et de la loi exponentielle (figure 5.2 b) révèle que les montants de sinistres auraient des queues moins épaisses que celle d'une loi de Pareto standard de même moyenne, et se rapprocherait plus d'une distribution exponentielle[3] de même moyenne.

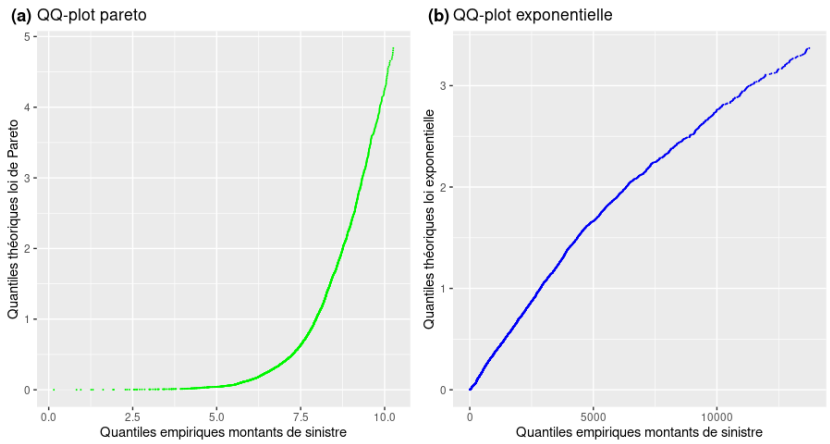


FIGURE 5.2 : *QQ-plots distribution de Pareto (a), distribution exponentielle (b).*

Dans la pratique, l'étape la plus difficile dans la modélisation des sinistres extrêmes est le choix d'un seuil à partir duquel les sinistres sont considérés comme extrêmes : si le seuil utilisé est très faible cela permet d'augmenter le nombre de sinistres extrêmes, mais l'approximation par la loi de Pareto généralisée est mauvaise, car cette loi ne charge que de grandes valeurs ; Inversement, utiliser un seuil trop élevé limite le nombre de données extrêmes, ce qui altère la qualité des estimateurs de la loi GPD à considérer. Le choix du seuil est donc important et doit trouver un juste équilibre entre ces deux extrêmes.

Heureusement pour nous, la théorie des valeurs extrêmes (TVE) fournit des procédures rationnelles et scientifiques pour l'estimation du seuil à considérer.

Pour déterminer le seuil à considérer, la TVE préconise d'utiliser le graphique des *dépassements moyens de seuil*. Le principe de ce graphique est le suivant : si les dépassements de seuils suivent une GPD, on doit avoir une approximation linéaire au delà du seuil (Robert (2018)).

La courbe des dépassements moyens de seuil (*mean excess plot* en anglais) que l'on obtient avec nos données de sinistres est représentée sur la figure 5.3. D'après cette figure, un seuil approprié serait 5 500 dollars, ce qui semble être une limite acceptable pour l'assureur. Seules 627 sinistres

[2]La GPD est la loi dédiée à la modélisation des dépassements de seuil.

[3]Rappelons néanmoins qu'une loi exponentielle correspond à un cas particulier de loi de Pareto. Elle est généralement utiliser pour modéliser les phénomènes sans mémoire.

ont un montant supérieur à ce seuil, soit moins de 0.7 % de l'ensemble des polices exposées sur la période d'étude.

Compte tenu de ce très faible effectif de données extrêmes, nous optons finalement pour la seconde option de prise en compte des sinistres extrêmes qui est l'écrêtement des sinistres.

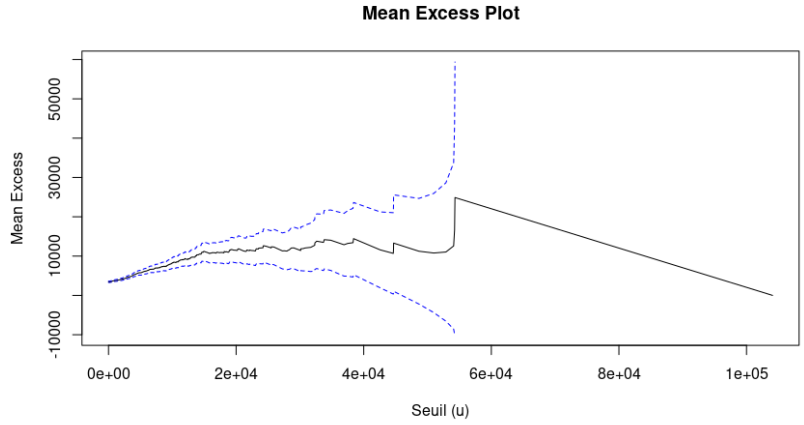


FIGURE 5.3 : *Mean Excess Plot des montants de sinistres en trait plein (en pointillés bleus, il s'agit des bornes de l'intervalle à 95%).*

• Écrêtement des sinistres : cela revient à inflater l'ensemble des montants en répartissant uniformément la charge surcrête. Il s'agit là d'une opération de mutualisation des sinistres extrêmes entre les différents assurés. Les montants de sinistres après écrêtement notés $\tilde{S}_i$ deviennent :

$$\tilde{S}_i = \min(S_i, u) + \frac{\Pi_u}{n}, \ i = 1, \cdots, n;$$

où Π_u est la charge surcrête définie par :

$$\Pi_u = \sum_{i=1}^{n} \max(S_i - u, 0)$$

n le nombre d'observation de notre base de données à savoir 98 058 et $u = 5\,500$ dollars.

5.2.3 Présélection des variables explicatives

Pour l'actuaire détenir une grande quantité d'informations sur l'assuré ou sur l'objet assuré, lui permet généralement de mieux appréhender les potentiels facteurs ou comportements de risque. Cependant, il arrive parfois que :

• Certaines ou plusieurs des variables utilisées dans un modèle de tarification ne soient pas associées à la sinistralité. L'inclusion de telles variables non pertinentes entraîne une complexité inutile du modèle résultant.

• Dans les modèles linéaires le problème de la multicolinéarité des variables ne fait plus débat depuis longtemps (Farrar et Glauber (1967)). En apprentissage statistique, la présence de multicolinéarité entre les caractéristiques n'a certes pas d'effets majeurs sur la performance prédictive des modèles, mais pourrait poser problème au niveau de la pertinence des interprétations, notamment au niveau de l'interprétation de l'importance des variables ou aussi au niveau de l'interprétation des graphiques PDP, ALE-plot, etc. (confère chapitre 4).

En supprimant les variables inutiles ou redondantes nous pouvons obtenir un modèle avec une meilleure performance prédictive (évitant le sur-apprentissage), plus facile à ajuster d'un point de vue de temps de calcul, plus facile à interpréter et plus pertinent.

Pour faire un premier tri et exclure les variables non-pertinentes, nous avons utilisé les deux approches suivantes pour la sélection des variables :

• *Une approche non-supervisée* : nous avons calculé la *matrice des liaisons*[4] des variables deux à deux. Pour des paires de variables "fortement" corrélées, nous avons retenu au plus une seule d'entre les deux.

• *Une approche supervisée* : à l'aide de l'approche de sélection des variables basée sur les modèles LocalGLMnet (présentée au chapitre 2), nous avons procédé à un second tri de variables. Cette approche présente l'avantage de sélectionner les variables sur la base d'une procédure de test d'hypothèse de significativité statistique. La finalité de ce test étant de retenir uniquement les variables significativement associées à la sinistralité (fréquence ou sévérité de sinistres).

5.2.3.1 Présélection non-supervisée des variables explicatives : matrice des liaisons

En statistique, il est facile de mesurer l'intensité de la liaison entre deux variables par un indicateur. Dans notre étude, puisque nous disposons de variables des deux types (qualitative et quantitative), une extension de la matrice des corrélations est la matrice des liaisons dans laquelle, à l'intersection de la ligne j et de la colonne k, on trouve :

– le carré du *coefficient de corrélation* calculé entre les deux variables lorsqu'elles sont toutes les deux quantitatives ;

– le carré du *rapport de corrélation* entre les deux variables, lorsqu'elles sont de types hétérogènes ;

– le V *de Cramer* entre les deux variable, lorsqu'elles sont toutes les deux qualitatives.

L'ensemble de ces indicateurs sont compris dans l'intervalle [0, 1]. Plus la valeur est proche de 1, plus la liaison entre les deux variables est forte, et plus elle se rapproche de 0 moins la liaison est forte.

La figure 5.4 résume les liaisons entre les paires de variables de notre jeu de données. Elle met en évidence les groupes de variables corrélées au sein du jeu de données :

• Tout d'abord, le groupe des variables télématiques relatives au nombre d'accélérations brusques (*Accel.**xx**miles*) et celles relatives au nombre de freinages brusques (*Brake.**xx**miles*).

• Les variables relatives au nombres de virages à droite par 1000 miles (*R.turn.intensity**xx***) sont également fortement corrélées entre-elles. De même que les variables relatives au nombre de virages à gauche par 1000 miles (*L.turn.intensity**xx***).

• Il se dégage également une forte corrélation entre le nombre d'année de sans sinistre (*Years.noclaims*) et le type d'usage du véhicule (*Car.age*).

A présent, afin d'éviter les informations redondantes, il faut retenir une seule variable par groupe de variables corrélées. Pour ce faire, nous avons procédé à une classification hiérarchique ascendante de la matrice de liaison.

[4]La matrice des liaisons est une extension de la matrice des corrélations au cas où l'on dispose simultanément de variables quantitatives et de variables qualitatives.

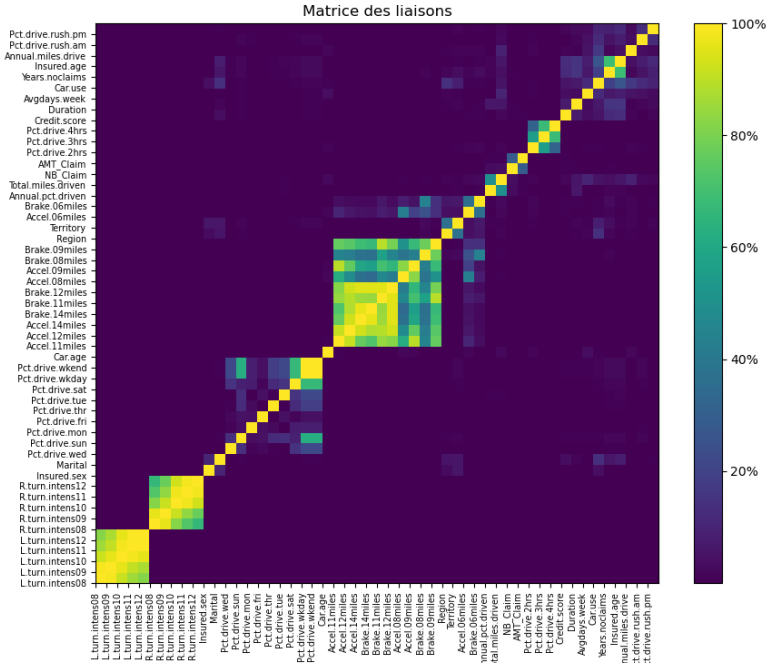


FIGURE 5.4 : *Matrice de liaisons entre les variables du jeu de données.*

Le principe est de définir une distance entre deux variables à partir de leur corrélation, typiquement :

$$d(X_j, X_k) = 1 - \rho^2(X_j, X_k),$$

avec $\rho^2(X_j, X_k)$ la valeur de l'indicateur de liaison approprié entre les variables X_j et X_k, suivant leur nature (qualitative ou quantitative). Les variables "proches" dans la hiérarchie sont alors celles le plus corrélées. On trace ensuite le dendrogramme associé à cette classification hiérarchique ascendante et on choisit le seuil de la corrélation minimale au-dessus de laquelle deux variables seront considérées comme redondantes (fortement liées).

La figure 5.5 illustre les résultats obtenus avec nos données. Le seuil de regroupement choisi est 20%. Autrement dit les variables sont considérées comme redondantes au-delà de 80% de corrélation. L'axe des abscisses représente à la distance entre les variables. Pour une paire de variable donnée, plus est elle faible mieux les variables sont corrélées entre-elles.

Au final, on extrait 13 variables redondantes qui ne seront pas utilisées par la suite, à savoir :

Pct.drive.wkend, *Accel.09miles*, *Accel.12miles*, *Brake.11miles*, *Brake.12miles*, *Brake.14miles*, *Left.turn.intensity09*, *Left.turn.intensity10*, *Left.turn.intensity11*, *Left.turn.intensity12*, *Right.turn.intensity09*, *Right.turn.intensity11*, *Right.turn.intensity12*.

Autrement dit, dans chacun des groupes de corrélation observés précédemment sur la matrice de corrélation (figure 5.4), on a retenu une seule variable représentative par groupe. Ces variables se rapportent toutes à la télématique du conducteur et concernent ses nombres d'accélérations et de freinages brusques, son nombre de virages (à gauche comme à droite) et sa proportion de conduite du week-end durant la semaine.

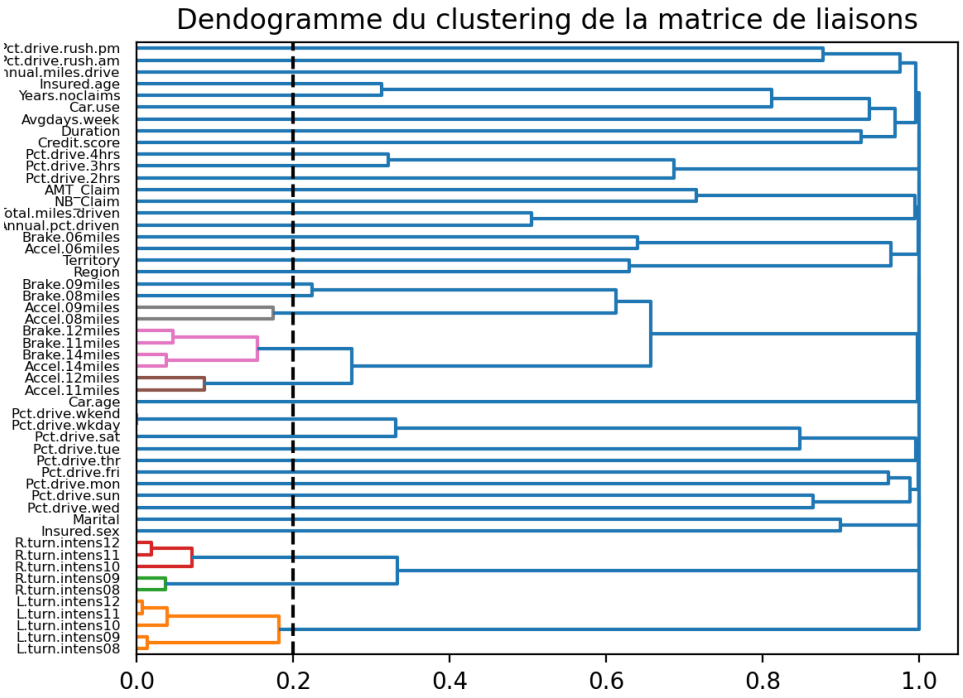


FIGURE 5.5 : *Dendrogramme avec un seuil de regroupement entre les variables à 20%.*

5.2.3.2 Présélection supervisée des variables explicatives : présélection basée sur un modèle LocalGLMnet

Cette deuxième approche de sélection des variables est plus orientée que la première. Elle vise à ne retenir que les variables qui sont effectivement associées au phénomène que nous souhaitons modéliser ; Dans notre cas, elle met évidence les variables explicatives importante dans la modélisation de la sinistralité (fréquence et coût).

Nous avons ajusté des modèles LocalGLMnet sur le montant et la fréquence de sinistres. À ce niveau l'objectif principal n'est pas de calibrer le modèle le plus performant possible sur nos données, mais juste de se faire une idée sur les variables importantes dans la modélisation de la sinistralité. Nous calculons l'intervalle de confiance I_α (voir équation 2.14 au chapitre 2) pour un niveau de significativité $\alpha = 0,1\%$. L'intervalle résultant $I_{0,1\%}$ est illustré par les bandes de couleur bleu ciel sur les différents graphiques des figures 5.6 et C.1.

□ Présélection des variables continues et des variables binaires

• Commençons par analyser les poids d'attention $\hat{\beta}_j(x)$ des variables continues et des variables binaires.

Le ratio de couverture au dessus de chaque graphique de la figure 5.6 correspond à la proportion de points $\hat{\beta}_j(x_+^{(i)})$ $(1 \leq i \leq 10\,000)$ qui tombe dans l'intervalle I_α ; autrement dit la proportion de points $\hat{\beta}_j(x_+^{(i)})$ qui sont considérés comme nuls parmi les 10 000 cas calculés).

Nous rejetterons l'hypothèse nulle $H_0 : \hat{\beta}_j(x) = 0$, uniquement pour les variables dont le ratio de couverture est "substantiellement" plus petit que 99.9 % (par exemple pour les ratio inférieur à 85%).

Les variables ayant un ratio de couverture supérieur à 85 % pour le montant de sinistre (confère

figure 5.6) et la fréquence de sinistre (confère figure C.1) sont au nombre de 12. Elles seront toutes mise à l'écart dans la suite. Il s'agit des variables :

Insured.sexe, *Martial*, *Region*, *Years.noclaims*, *Pct.drive.tue*, *Pct.drive.wed*, *Pct.drive.fri*, *Pct.drive.3hrs*, *Brake.06miles*, *Right.turn.intensity.08*, *Right.turn.intensity.10*, *Accel.08miles*.

Nous excluons aussi la variable *Duration* correspondant à la période pendant laquelle l'assuré est couvert (en jours), car en soi, il ne s'agit pas en soi d'un facteur de risque.

• Une autre approche complémentaire pour la sélection des variables est de se fier au diagramme d'importance des variables résultant des modèles LocalGLMnet ajustés. La figure 5.8 résume cette information pour notre jeu de données.

Toutefois, nous rappelons qu'à ce stade, les modèles ajustés n'ont pas été scrupuleusement optimisée, il s'agissait tout simplement d'avoir une première idée sur le niveau d'importance des différentes caractéristiques. À cet effet, il se pourrait qu'une fois l'optimisation faite, l'ordre d'importance des variables soit légèrement modifié.

Le critère de sélection basée sur l'importance des variables (confère figure 5.8) nous mène aux mêmes conclusions que celles de la sélection des variables basée sur le test d'hypothèse de significativité statistique, ce qui est plutôt rassurant.

□ Présélection des caractéristiques catégorielles non binaires

En ce qui concerne les variables catégorielles avec plus de deux modalités, à savoir l'usage fait du véhicule (*Car.use*) et l'indicateur anonymisé du territoire de conduite (*Territory*), analysons leur effet sur la sinistralité.

Au préalable, nous réalisons un *encodage à chaud*[5] (*one-hot encoding*, en anglais) de chacune de ces variables afin de mieux mettre en évidence l'effet individuel de chaque modalité sur la variable cible.

Le critère de sélection reste similaire au cas des caractéristiques continues ou binaires : pour une caractéristique donnée, mieux les boxplots sont à l'extérieure de la bande délimitée par les traits en bleu ciel, mieux la caractéristique est importante pour la modélisation de la variable cible.

En définitive, à l'issue des deux tris effectués (non-supervisé et supervisé), nous retenons 23 variables explicatives pour la suite de l'étude dont :

– Six (06) variables classiques : l'âge de l'assuré (*Insured.age*), son score de crédit (*Credit.score*), l'âge de son véhicule (*Car.age*), l'indicateur anonymisé de son territoire de conduite (*Territory*), l'usage fait du véhicule (*Car.use*) et le nombre de miles annuels prévus à parcourir déclarés par le conducteur (*Annual.miles.drive*)

– Dix sept (17) variables télématiques reparties entre le volume de conduite de l'assuré (*Total.miles.driven, Annual.pct.driven, Avgdays.week*), les jours de conduite dans la semaine (*Pct.drive.mon, Pct.drive.thr, Pct.drive.sat, Pct.drive.sun*), la durée au volant (*Pct.drive.2hrs, Pct.drive.4hrs*), le fait de conduire en matinée ou en soirée (*Pct.drive.rush.am, Pct.drive.rush.pm*) et son comportement au volant (*Brake.08miles, Brake.09miles, Accel.06miles, Accel.11miles, Accel.14miles, Left.turn.intensity.08*).

[5] L'encodage à chaud consiste à éclater la variable catégorielle initiale, en autant de variable indicatrice que ses modalités.

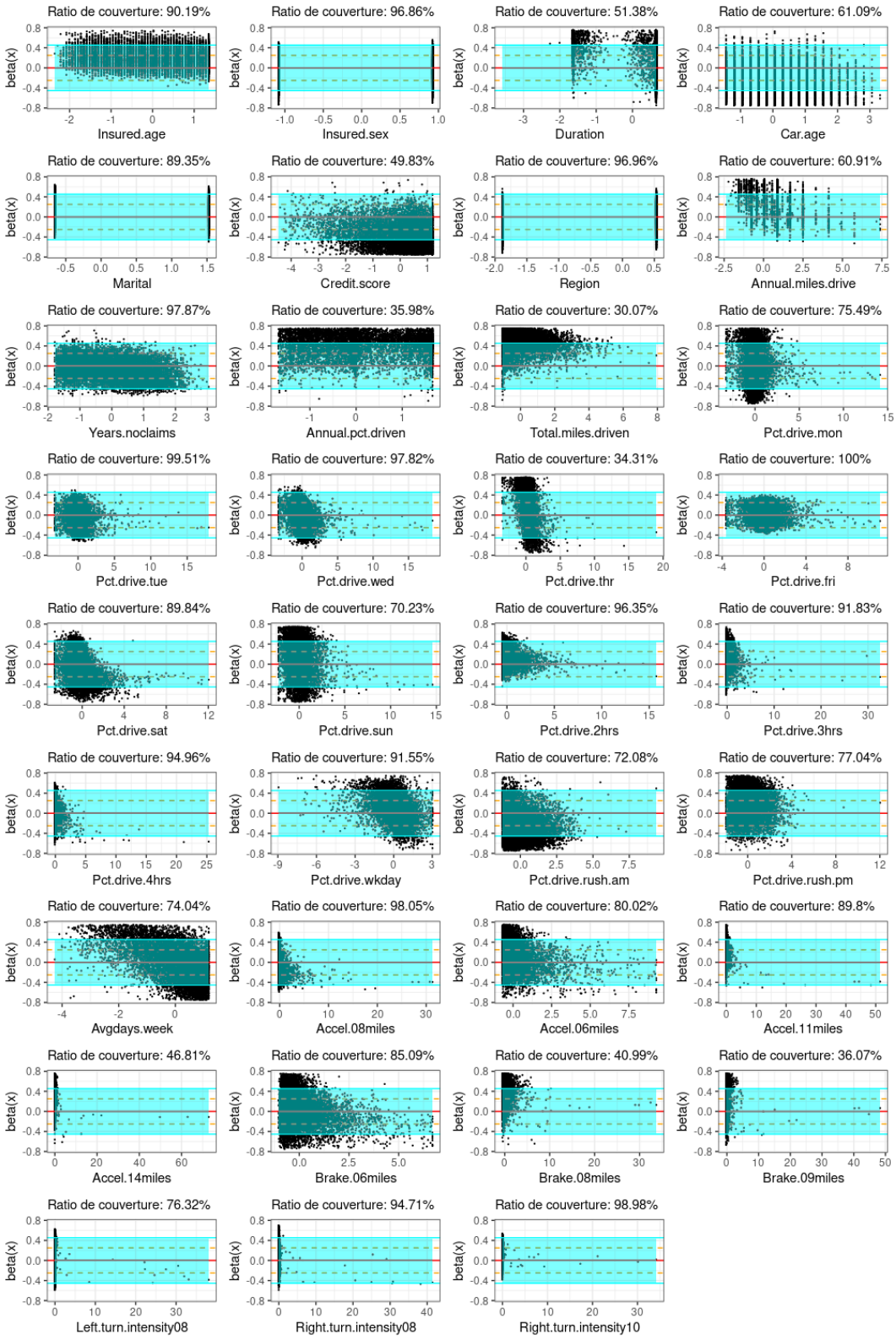


FIGURE 5.6 : *Poids d'attention* $\hat{\beta}_j(x_+^{(i)})$ *sur le montant de sinistres, des caractéristiques continues et binaires pour* 10 000 *instances* $x_+^{(i)}$ *de la base de test sélectionnées aléatoirement ; la bande bleu clair indique la zone de confiance* I_α *au niveau de signification* $\alpha = 0,1\%$ *du test d'hypothèse nulle* $H_0 : \hat{\beta}_j(x_+) = 0$.

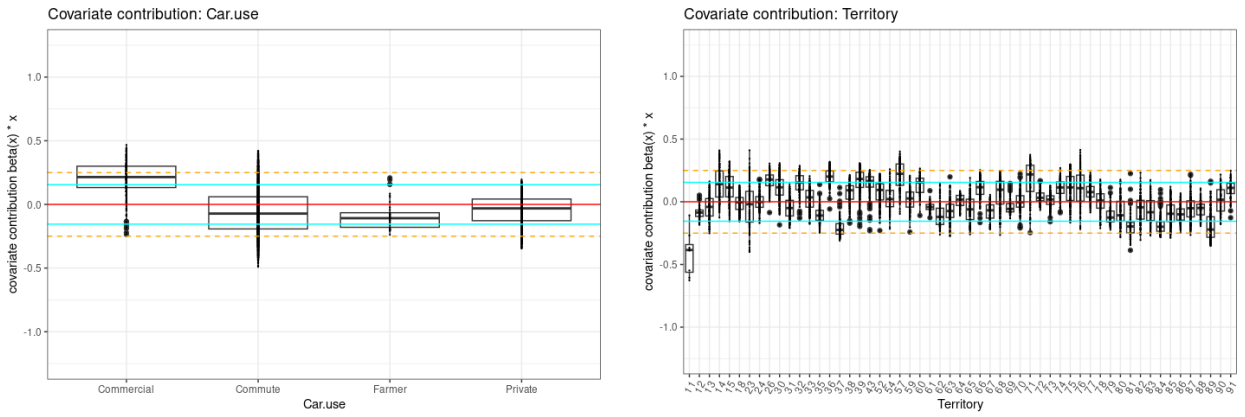


FIGURE 5.7 : *Boxplot des contributions $\hat{\beta}_j(x_+)$ des composantes catégorielles du type d'usage du véhicule (Car.use) et du territoire de conduite de l'assuré (Territory) sur la fréquence de sinistre ; la région à l'intérieur des lignes bleu clair indique la zone de confiance I_α au niveau de signification $\alpha = 0,1\%$ du test d'hypothèse nulle $H_0 : \hat{\beta}_j(x_+) = 0$.*

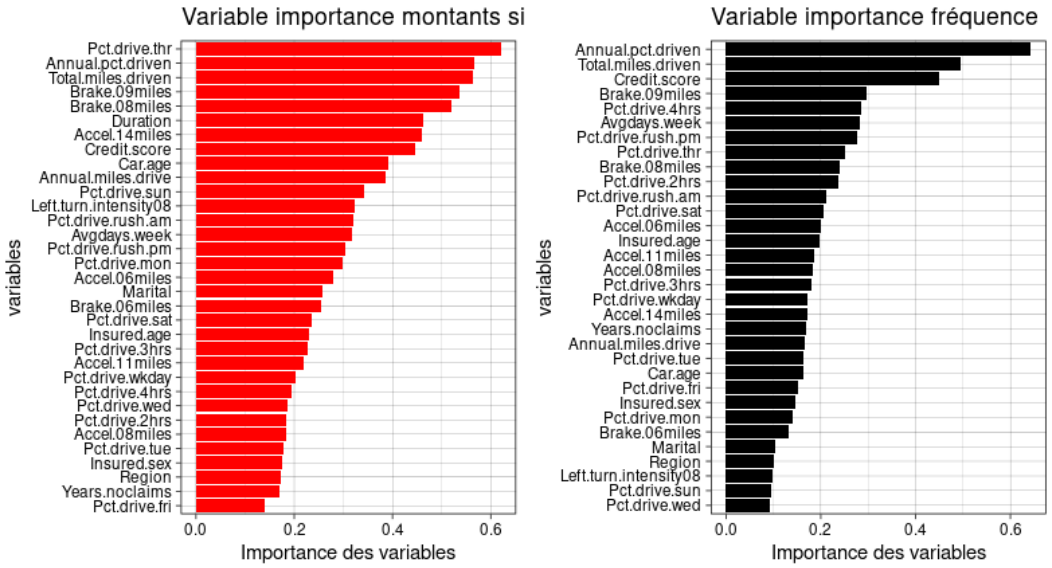


FIGURE 5.8 : *Importance des variables sur le montant agrégé des sinistres (à gauche) et sur la fréquence des sinistres (à droite) calculer avec la formule des VI_j définie au chapitre 2, par l'équation 2.11.*

5.2.4 Analyse descriptive des données de l'étude

Dans cette partie, il est question de mener une analyse descriptive des données à partir de visualisations graphiques afin de mieux appréhender les variables explicatives.

Nous avons au préalable convenablement regroupé les variables continues en classes d'intervalles, en s'assurant de l'homogénéité des classes vis-à-vis des variables de sinistralité, tout en garantissant que chaque classe contienne un nombre suffisant d'observations. Les variables catégorielles résultantes de cette transformation portent leur nom initial de variable augmenté du suffixe "G" (par exemple, après segmentation de *Car.age* la variable obtenue s'intitule *Car.ageG*).

Le tableau C.1 en annexe C répertorie l'ensemble des bornes retenues lors du découpage des variables explicatives continues. En outre, la variable *Territory* qui présentait initialement 55 modalités a été regroupé en trois zones : la $zone_A$ qui regroupe les territoires d'indicateurs (13, 14, 15, 24, 52, 62, 69, 81, 82, 83, 90, 91) ; la $zone_B$ (11, 18, 23, 26, 30 à 32, 35 à 39, 43, 54, 59, 60, 66 à 68, 70 à 73,76 à 78, puis 85 à 89) et la $zone_C$ (12, 33, 57, 61, 63, 64, 65, 74, 75).

L'idée de regrouper les différentes caractéristiques en classes d'intervalles se justifie tout simplement par le fait que les modèles linéaires généralisés impliquent une monotonie de l'effet des caractéristiques sur la prédiction, ce qui n'est que très rarement vérifié en pratique.

La figure 5.9 et les onze autres figures disponibles à l'annexe C résument la sinistralité selon les différentes variables explicatives disponibles dans notre jeu de données. Après analyse de ces graphiques, on relève des faits notables :

• De prime abord, dans notre base de données les véhicules à usage commun sont les plus représentés (50%), suivis des véhicules à usage privée (46%) ; les véhicules à usage commercial et agricole sont les moins représentées avec une proportion de 3% et 1% respectivement. La fréquence de sinistre est la plus élevée chez les véhicules à usage commercial et la moins élevée chez les véhicules à usage agricole. Cependant, les montants agrégé de sinistres sont généralement plus élevés pour les véhicules à usage commun.

• Nous remarquons également que la fréquence de sinistre semble "proportionnelle" au nombre de virage à gauche par 1000 miles avec une intensité (accélération) de 08 mph/s (miles par heure par seconde) : plus le nombre de virage avec une intensité de 08 mph/s augmente, plus la fréquence de sinistre est élevée (voir figure C.5 en annexe C).

• On remarque également une tendance croissante entre le nombre de miles parcourus et la fréquence de sinistre : ceux qui roulent le plus sont ceux qui enregistrent les plus grandes fréquences de sinistre. Cependant, en cas de sinistre, ce sont les assurés ayant le moins roulé qui présentent des coûts agrégés les plus élevés (voir figure C.6 en annexe C).

• En outre, à mesure que le nombre d'accélérations brusques d'intensité 06 mph/s augmente, la fréquence et le montant agrégé des sinistres augmentent (voir figure C.4 en annexe C).

• Il en va de même pour le nombre de freinages brusques d'intensité 08 ou 09 mph/s : les assurés qui freinent brusquement avec de telles intensités, sont généralement ceux là qui enregistrent les fréquences et coûts agrégés de sinistres les plus élevés (voir figure C.3 en annexe C).

• Parmi les quatre jours de conduite pré-sélectionnés, peu importe le jour de la semaine, l'évolution de la sinistralité est quasiment la même.

• Quoi que pour des raisons éthiques le score de crédit ne constitue pas une variable tarifaire par excellence, on observe néanmoins, une relation négative entre le score de crédit de l'assuré et son niveau de sinistralité : sur la figure C.6, on observe bien qu'à mesure que le score de crédit augmente, le fréquence et le coût agrégé de sinistre diminuent.

5.2.5 Évaluation de la performance prédictive des modèles : Formation des bases de données d'entraînement et de test

Afin de s'assurer que les différents modèles de prédiction mis en place dans la suite de ce mémoire généralisent bien les phénomènes pour lesquels ils ont été conçu, nous séparerons la base de données initiale en deux parties : 85 % des données pour l'entraînement et 15% pour le test.

Pour la formation des bases de données d'entraînement et de test, nous nous sommes arrangés de garantir en amont un certain équilibre entre la composition de ces deux bases. De manière à ce

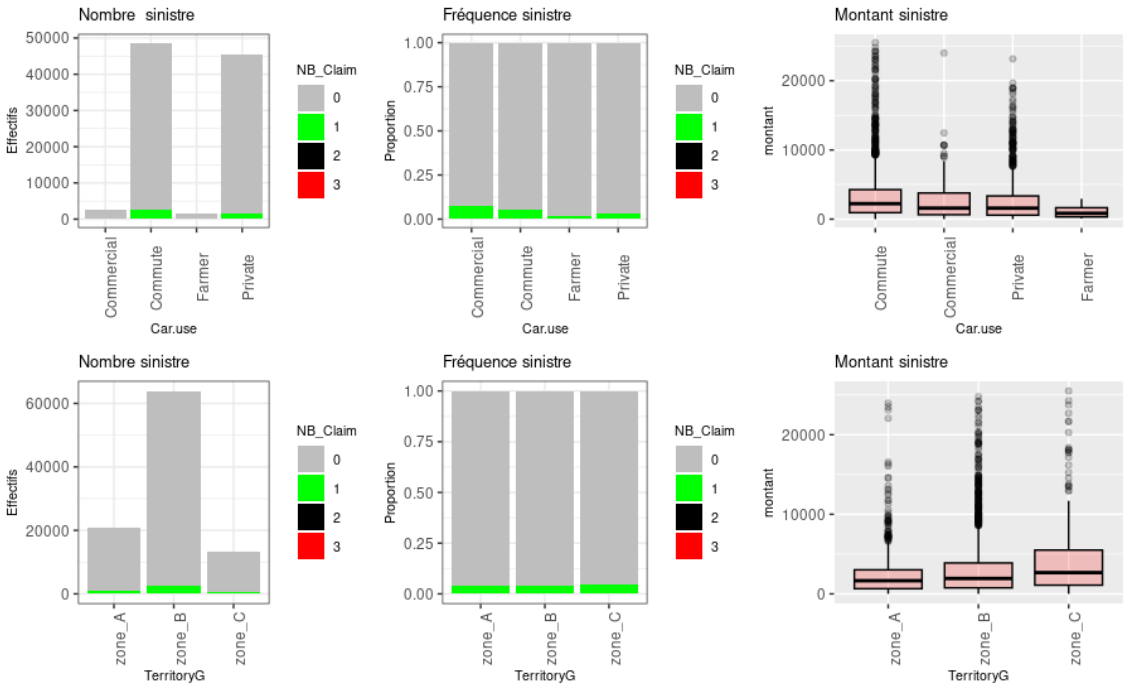


FIGURE 5.9 : *Analyse de la sinistralité selon les variables Car.use et Territory.*

que la base d'apprentissage reflète au mieux la base initiale (en proportion) et que la base de test à son tour reflète au mieux la base d'entraînement. Pour ce faire, la procédure de construction des bases d'entraînement et de test s'est inspirée de la méthode *d'échantillonnage équilibré* introduit par Tillé (2010). Pour sa mise en oeuvre, nous avons utilisé la fonction *samplecube* du package R [sampling].

Le tableau 5.2 récapitule une évaluation de la proximité entre la base d'entraînement, la base de test. Nous constatons que les deux bases sont quasiment identiques en répartition (en moyenne ou en proportion) pour les variables centrales de l'étude. Ainsi, l'évaluation de la performance de généralisation de nos modèles à partir de la base de test sera bien objective.

Nous rappelons que la variable cible dans nos modèles de fréquence sera $\frac{NB\text{-}Claim \times 366}{Duration}$ qui correspond au nombre de sinistre observé rapporté à la période couverture. Pour éviter que la fréquence annuelle de sinistre de certains assuré ayant une période de couverture très courte n'explose suite de cette transformation, nous avions envisagé de maintenir dans notre jeu de données uniquement les polices dont la période de couverture était d'au moins 182 jours (une demi année).

Après vérification, nous nous sommes rendu compte que dans notre jeu de données, seules 75 polices sur 98 058 polices au total ont une période de couverture inférieure à 182 jours et pour l'ensemble de ces polices, le nombre totale de sinistre observé sur leur période de couverture est nul, donc nous ne courrons pas de risque d'explosion de la quantité $\frac{NB\text{-}Claim \times 366}{Duration}$. Finalement, on converse donc toutes nos polices.

	Base entière	Base train	Base test	
Variables	valeur moyenne	valeur moyenne	valeur moyenne	Déviation relative bases train vs test (%)
Duration	313.60	313.60	313.60	0
Insured.age	51.43	51.43	51.43	0
Car.age	5.77	5.77	5.77	0
Credit.score	801	801	801	0
Total.miles.driven	4848.50	4848.50	4848.45	0
Annual.pct.driven	0.50	0.50	0.50	0
NB-Claim	0.04	0.04	0.04	0
AMT-Claim	133.30	133.40	132.90	-0.37
Car.use				
Commercial	2.60%	2.60%	2.60%	0
Commute	49.70%	49.70%	49.70%	0
Farmer	1.40%	1.40%	1.40%	0
Private	46.30%	46.30%	46.30%	0
TerritoryG				
Zone-A	21.40%	21.40%	21.40%	0
Zone-B	65.20%	65.20%	65.20%	0
Zone-C	13.40%	13.40%	13.40%	0
Nombre d'observations	98 058	83 350	14 708	

TABLE 5.2 : *Récapitulatif de l'évaluation de la proximité entre la base d'apprentissage et la base de test pour les variables explicatives de premier ordre.*

5.3 Données télématiques et optimisation tarifaire

Dans cette section, notre objectif est double :

- **Objectif 1 : évaluation de la plus-value des données télématiques dans la modélisation de la sinistralité**

Dans un premier temps, il est question d'évaluer les potentielles améliorations de la précision prédictive dues à l'usage des données télématiques comme variables explicatives dans la modélisation de la fréquence et de la sévérité agrégée de sinistre.

- **Objectif 2 : prédiction des données télématiques**

Dans un second temps, étant donné que l'assureur ne dispose pas toujours des informations télématiques sur l'ensemble de ses clients ou sur ses souscripteurs, nous essayerons de prédire les valeurs des données télématiques annuelles pour les assurés de notre portefeuille, à partir des informations disponibles et pertinentes. Par la suite, nous intégrerons les prédictions obtenues dans la modélisation de la fréquence et du coût agrégé des sinistres et analysons les éventuelles améliorations conséquentes à cet enrichissement de la base de données.

Dans la suite de ce mémoire nous travaillerons principalement sur les variables télématiques convenablement transformées en variables binaires. Le tableau 5.3 fournit les résultats obtenus à l'issue de cette nouvelle transformation des variables télématiques en variables binaires.

Nous avons retenu pour la suite uniquement les variables télématiques qui se sont avérées pertinentes (discriminantes) dans l'explication de la sinistralité, au vu des résultats de l'analyse descriptive.

Variables \ Classes	classe 1	classe 2
Accel.06milesB	[0, 30]	]30, +[
Accel.11milesB	0	]0, +[
Accel.14milesB	0	]0, +[
Brake.08milesB	[0, 10]	]10, +[
Brake.09milesB	[0, 5]	]5, +[
Left.turn.int.08B	[0, 150]	]150, +[
Total.miles.drivB	[0, 4500]	]4500, +[
Avgdays.weekB	[0, 4]	]4, 7]
Annual.pct.drivB (%)	[0, 50]	]50, 100]
Pct.drive.2hrsB (%)	0	]0, 100]
Pct.drive.4hrsB (%)	0	]0, 100]
Pct.drive.rush.amB (%)	[0, 5]	]5, 100]
Pct.drive.rush.pmB (%)	[0, 10]	]10, 100]

TABLE 5.3 : *Résumé des treize (13) variables télématiques retenues et des classes obtenues après leur transformation en variables binaires.*

5.3.1 Données télématiques : une valeur ajoutée pour la précision des modèles de sinistre

Cette partie est structurée en deux sous-parties : dans la première, nous étudions la plus value des données télématiques dans la modélisation de la fréquence de sinistre ; dans la seconde nous évaluons leur valeur ajoutée dans la modélisation de la sévérité agrégée des sinistres.

5.3.1.1 Évaluation de la valeur ajoutée des données télématiques dans la modélisation de la fréquence de sinistre

Nous présentons dans cette partie les résultats de quatre modèles de fréquence tous mis en place dans l'optique d'évaluer l'apport des variables télématiques dans la précision des modèles.

Dans ces différents modèles, la variable cible est le nombre de sinistres de l'assuré rapporté à la durée d'exposition de son contrat (représentée dans notre base par la variable *Duration*).

Dans notre jeu de données, la fréquence moyenne de sinistre est 0.04484081 et sa variance est de 0.0475013 : on est en situation d'équidispersion de la fréquence de sinistre. Pour ce faire, nous utiliserons un modèle GLM poissonnien pour la modélisation de la fréquence de sinistre.

Dans cette sous-partie, étant donné que notre finalité est de savoir est-ce que l'ajout des données télématiques améliorent ou pas la précision de la modélisation de la fréquence de sinistre, nous mettrons plus en avant les résultats de performance prédictive des différents modèles ajustés, à l'aide des différentes métriques présentées dans la sous-section 1.1 du chapitre 1.

Les modèles ont été mis en oeuvre sous le logiciel R : avec le package *stats* pour les GLM et le package *randomForest* pour le random forest.

(a) Modèle GLM avec uniquement des variables tarifaires classiques

Commençons par ajuster un *modèle Benchmark*, utilisant les variables tarifaires classiques disponibles dans notre jeu de données à savoir : *TerritoryG, Car.useG, Car.ageG, Insured.ageG, Annual.miles.driveG*. Les sorties du modèle estimé sont récapitulées dans le tableau 5.4.

Les informations qui nous intéressent le plus sont celles relatives à la performance prédictive de nos modèles. Pour cette raison, nous calculons les métriques de précision prédictive du présent modèle sur notre échantillon de test. Les résultats obtenus sont reportés dans le tableau 5.5.

	Estimate
(Intercept)	-2.60^{***}
	(0.28)
TerritoryGzone_B	0.00
	(0.04)
TerritoryGzone_C	0.04
	(0.06)
Car.useCommute	-0.25^{**}
	(0.09)
Car.useFarmer	-0.97^{***}
	(0.25)
Car.usePrivate	-0.27^{**}
	(0.09)
Car.ageG[3, 5]	-0.10^{*}
	(0.04)
Car.ageG[6, 9]	-0.35^{***}
	(0.05)
Car.ageG[10, +[	-0.75^{***}
	(0.06)
Insured.ageG[21, 30]	-0.22
	(0.23)
Insured.ageG[31, 40]	-0.50^{*}
	(0.23)
Insured.ageG[41, 50]	-0.36
	(0.23)
Insured.ageG[51, 60]	-0.30
	(0.23)
Insured.ageG[61, 70]	-0.51^{*}
	(0.23)
Insured.ageG[71, 80]	-0.71^{**}
	(0.24)
Insured.ageG[81, +]	-0.52
	(0.28)
Credit.scoreG]750, 800]	-0.25^{***}
	(0.05)
Credit.scoreG]800, 850]	-0.64^{***}
	(0.05)
Credit.scoreG]850, 900]	-0.70^{***}
	(0.05)
Annual.miles.driveG]6000, 7500]	0.87^{***}
	(0.14)
Annual.miles.driveG]7500, 12 500]	0.93^{***}
	(0.14)
Annual.miles.driveG]12 500, +[	1.15^{***}
	(0.15)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

TABLE 5.4 : *Résumé GLM Poisson (lien logarithmique) avec variables tarifaires classiques.*

Déviance	MSE	MAE	RMSE	$RMSE_{mean}$
25405	0.0597765	0.09298055	0.2444923	5.004322

TABLE 5.5 : *Résultats de performance prédictive du GLM fréquence avec variables tarifaires classiques (modèle Benchmark).*

(b) Modèle GLM avec variables tarifaires classiques + variables télématiques

A présent nous passons à l'étape suivante qui consiste à ajuster de nouveau un modèle GLM poissonien sur la fréquence de sinistre. Cette fois, nous utilisons en plus des variables classiques de l'étape précédente, les variables télématiques. Les sorties du modèle ajusté sont disponibles dans le tableau C.2 en annexe C.

Les métriques de performance prédictive de ce modèle sur notre échantillon de test, sont récapitulées dans le tableau 5.6.

Déviance	MSE	MAE	RMSE	$RMSE_{mean}$
22720	0.05687827	0.08876985	0.2384917	4.881498

TABLE 5.6 : *Résultats de performance prédictive du GLM fréquence avec variables tarifaires traditionnelles augmentées des variables télématiques.*

Les modèles linéaires généralisés sont généralement mis en avant dans la modélisation assurantielle du fait de leur structure simple et interprétable. Cependant, cette interprétabilité se fait au détriment de leur précision prédictive.

Pour évaluer optimalement le gain de précision prédictive apporté par les données télématiques, il serait plus intéressant d'ajuster la fréquence de sinistres à l'aide d'un modèle plus sophistiqué capable d'extraire un maximum d'informations dans les données.

Pour ce faire, nous avons privilégié les modèles de type *Random Forest* (confère chapitre 2).

(c) Modèle Random forest avec uniquement des variables tarifaires classiques

Dans un algorithme de forêt aléatoire, on dispose au total de 5 principaux hyperparamètres. Puisque nous souhaitons ajuster le "meilleur" modèle *Random Forest* pour la modélisation de la fréquence, il est nécessaire d'optimiser ces hyperparamètres, afin de déterminer leur valeur qui minimiserons l'erreur de prédiction sur la base de test.

Nous nous limitons à l'optimisation des deux hyperparamètres centraux du *Random Forest* et laissons tous les autres à leur valeur par défaut.

Il s'agit des hyperparamètres : *ntrees*, le nombre d'arbres à retenir pour l'agrégation finale et, *mtry*, le nombre de variables échantillonnées pour la construction de chaque noeud.

• Commençons par optimiser le nombre de variable à échantillonner à chaque noeud ($mtry$). Pour ce faire, fixons le nombre d'arbres ($ntrees$) à 150 et faisons varier $mtry$ dans l'ensemble $\{1, 2, 3, 4, 5, 6\}$ (le nombre total de variables explicatives dans ce cas est de 7, donc $mtry$ ne saurait excéder 7).

Le choix du paramètre optimal est basé sur une approche par validation croisée de type $5-folds$. La métrique retenue pour le calcul de l'erreur de validation est la RMSE. Le graphique de 5.10 montre clairement que pour un nombre d'arbres ($ntrees$) égale à 150, le nombre optimal de variables à échantillonner pour la formation des noeuds des différents arbres de la forêt aléatoire est $mtry^* = 6 variables$. Car c'est pour $mtry = 6$ que la RMSE des données de validation est en générale minimale (de même pour la MAE).

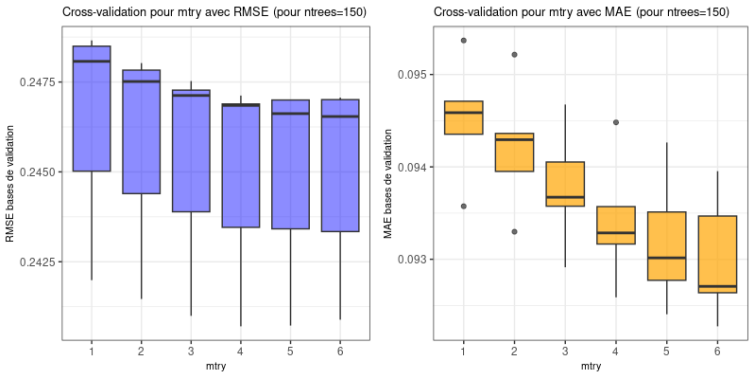


FIGURE 5.10 : *Etape 1 : Optimisation du paramètre mtry dans le modèle RF avec variables classiques uniquement et ntrees fixé à 150.*

• À présent, nous fixons $mtry$ à 6 et nous cherchons à déterminer la valeur optimale de $ntrees$. Pour cela, nous attribuons tour à tour à $ntrees$, les valeurs $\{20, 50, 100, 150, 180, 200, 250, 300\}$ et examinons celle pour laquelle, le modèle commet le moins d'erreur sur les échantillons de validation.

La lecture du graphique 5.11 révèle qu'un bon choix pour $ntrees$ serait 20.

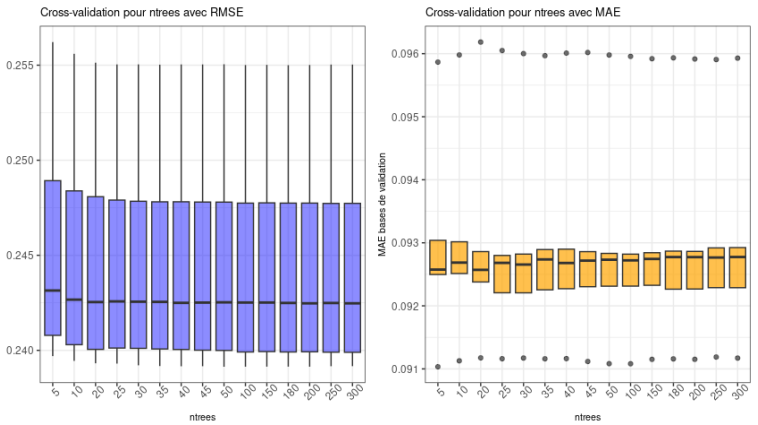


FIGURE 5.11 : *Etape 2 : Optimisation du paramètre ntrees dans le modèle RF avec variables de risque classiques (pour le meilleur mtry retenu à l'étape 1, voir figure 5.10).*

Nous retenons comme modèle final celui ayant pour hyperparamètres $mtry^* = 6$, $ntrees^* = 20$ et les valeurs par défaut pour les autres hyperparamètres.

Les performances prédictives du modèle final sur notre échantillon de test sont détaillées dans le tableau 5.7.

MSE	MAE	RMSE	$RMSE_{mean}$
0.0593124	0.09186272	0.2435414	4.984857

TABLE 5.7 : *Résultats de performance prédictive du RF fréquence avec uniquement les variables tarifaires traditionnelles.*

À ce stade, il est légitime de se poser la question de savoir : qu'en sera-il des performances d'un modèle *Random Forest* avec les variables classiques augmentées des variables télématiques ?

(d) Modèle Random forest avec des variables tarifaires usuelles + variables téléma tiques

Dans cette sous-section nous ajustons la fréquence de sinistre à partir des variables de risque classiques et des variables télématiques. Le type de modèle utilisé est le *Random Forest*.

Afin de tirer le meilleur parti de notre modèle, nous optimisons au préalable ses hyperparamètres centraux comme précédemment. Le volume de la base de données ayant significativement augmenté suite à l'ajout des treize (13) variables télématiques, pour réduire le temps d'ajustement de notre modèle *Random Forest*, nous avons convenablement réduit la base d'entraînement à un échantillon de taille 40 milles observations.

• Pour l'hyperparamètre *mtry*, comme le montre le graphique 5.12, la valeur optimale, c'est-à-dire celle pour laquelle la RMSE sur les échantillons de validation est généralement moindre est $mtry^* = 6$.

• À présent, ayant fixé $mtry = 6$, nous déterminons le nombre d'arbre *ntree* pour lequel l'erreur de validation du modèle est minimale. À cet effet, nous faisons varier les valeurs *ntree* dans l'ensemble {10, 20, 30, 40, 50, 80, 100, 150, 200}. Comme le montre la figure 5.13, la valeur finalement retenue est : $ntrees^* = 100$.

Pour ces valeurs optimales à savoir $ntrees^* = 100$ et $mtry^* = 6$, les performances prédictives sur notre base de test du modèle *Random Forest* finalement ajusté sont répertoriées dans le tableau 5.8.

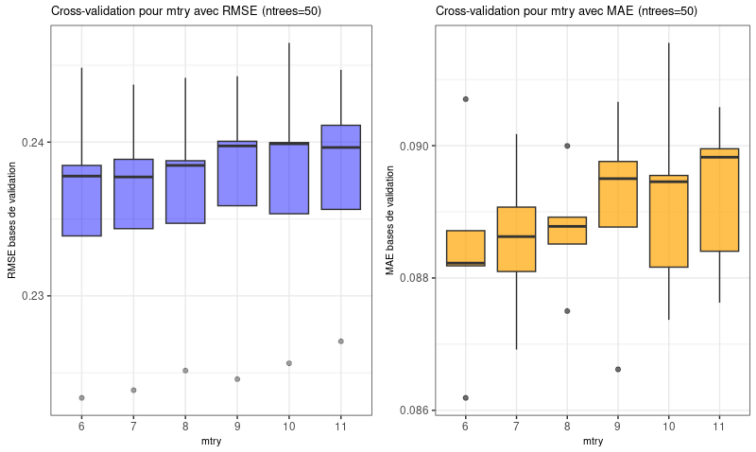


FIGURE 5.12 : *Etape 1 : Optimisation du paramètre mtry dans le modèle RF avec variables de risque classiques et variables télématiques, pour ntrees initialement fixé à 50.*

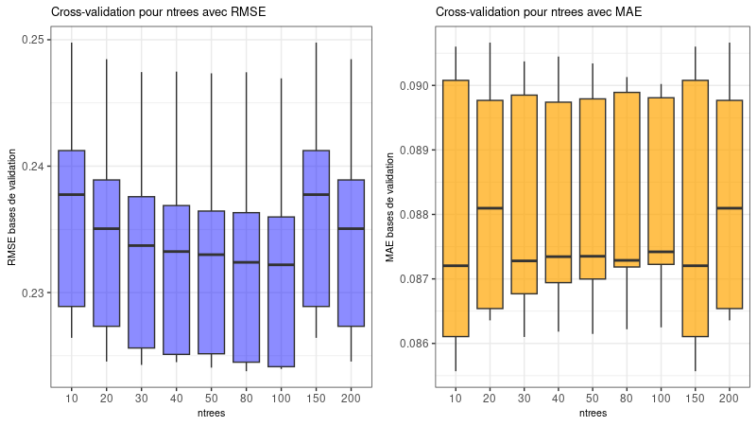


FIGURE 5.13 : *Etape 2 : Optimisation du paramètre ntrees dans le modèle RF avec variables de risque classiques et télématiques (pour le meilleur mtry retenu à l'étape 1, voir figure 5.12).*

MSE	MAE	RMSE	$RMSE_{mean}$
0.05128279	0.08603667	0.226457	4.635171

TABLE 5.8 : *Résultats de performance prédictive du RF fréquence avec les variables tarifaires traditionnelles et télématiques.*

5.3.1.2 Évaluation de la valeur ajoutée des données télématiques dans la modélisation de la sévérité

Dans cette sous-partie, il est question cette fois d'évaluer le gain potentiel de précision prédictive conséquente à la prise en compte des données télématiques dans la modélisation de la sévérité de sinistres.

Pour ce faire, nous ajustons puis évaluons les performances prédictives de deux modèles de sévérité : l'un avec uniquement des variables de risque classiques et l'autre avec en plus des variables télématiques. Nous nous limitons ici à l'ajustement des modèles de type GLM.

(a) Modèle GLM-Tweedie avec uniquement des variables tarifaires classiques

Débutons par la mise en oeuvre d'un modèle GLM avec loi de Tweedie, uniquement à partir des variables de risque classiques disponibles dans notre base de données.

Pour son implémentation nous avons utilisé le package *statmod* du logiciel R. Les sorties du modèle sont présentées dans le tableau 5.9.

	Estimate
(Intercept)	5.34^{***}
	(0.49)
TerritoryGzone_B	0.22^{**}
	(0.08)
TerritoryGzone_C	0.64^{***}
	(0.10)
Car.useCommute	-0.04
	(0.17)
Car.useFarmer	-1.28^{**}
	(0.42)
Car.usePrivate	-0.17
	(0.18)
Car.ageG[3, 5]	-0.13
	(0.08)
Car.ageG[6, 9]	-0.47^{***}
	(0.08)
Car.ageG[10, +[	-1.05^{***}
	(0.10)
Insured.ageG[21, 30]	0.02
	(0.42)
Insured.ageG[31, 40]	-0.46
	(0.42)
Insured.ageG[41, 50]	-0.32
	(0.41)
Insured.ageG[51, 60]	-0.39
	(0.42)
Insured.ageG[61, 70]	-0.94^{*}
	(0.42)
Insured.ageG[71, 80]	-0.82
	(0.43)
Insured.ageG[81, +]	-0.75
	(0.50)
Credit.scoreG]750, 800]	-0.51^{***}
	(0.09)
Credit.scoreG]800, 850]	-1.09^{***}
	(0.08)
Credit.scoreG]850, 900]	-1.47^{***}
	(0.09)
Annual.miles.driveG]6000, 7500]	0.76^{***}
	(0.19)
Annual.miles.driveG]7500, 12 500]	0.90^{***}
	(0.19)
Annual.miles.driveG]12 500, +[	0.89^{***}
	(0.21)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

TABLE 5.9 : *Résumé du GLM Tweedie (lien logarithmique) avec variables tarifaires classiques.*

Après avoir mis en place le modèle, nous nous en sommes servi pour réaliser des prédictions au niveau de la base de test. Les résultats des métriques de performance prédictive obtenus sont résumés dans le tableau 5.10.

Déviance	MSE	MAE	RMSE	$RMSE_{mean}$
5 716 786.74	1 301 920	250.6309	1 141.017	8.587901

TABLE 5.10 : *Résultats de performance prédictive du GLM-coût avec uniquement les variables tarifaires traditionnelles.*

(b) Modèle GLM-Tweedie avec variables tarifaires classiques et télématiques

Nous avons enfin ajusté la sévérité agrégée de sinistres avec un modèle GLM-Tweedie en utilisant comme prédicteurs les variables de risque classiques et les variables télématiques. Les performances prédictives du modèle obtenu sur la base de test sont présentées dans le tableau 5.11 ci-dessous.

Déviance	MSE	MAE	RMSE	$RMSE_{mean}$
5 016 936	1 256 309	237.5303	1 120.852	8.436126

TABLE 5.11 : *Résultats de performance prédictive du GLM-coût avec variables tarifaires traditionnelles et télématiques.*

5.3.1.3 Analyse synthèse et comparaison des modèles mis en oeuvre sans variables télématiques *vs* avec variables télématiques

Dans cette partie, nous reprenons les résultats de performances prédictives des modèles ajustés ci-dessus. L'objectif est d'analyser les performances des modèles de fréquence et de sévérité lorsqu'on met de côté les données télématiques dans un premier temps et lorsqu'on les prend en compte dans un second temps.

On observe dans le tableau 5.12 que le gain relatif de performance prédictive en cas de prise en compte des variables télématiques dans les modèles est non négligeable, que ce soit pour la fréquence ou pour la sévérité de sinistre.

De plus, l'amplitude de ce gain relatif de performance varie suivant la complexité du modèle utilisé pour modéliser la sinistralité.

Le gain relatif de performance prédictive en question est mesuré par la métrique suivante :

$$Gain\,relatif = \frac{metrique_{modele\,benchmark} - metrique_{modele\,cible}}{metrique_{modele\,benchmark}} \times 100$$

Rappelons néanmoins que pour les modèles de fréquence, le *modèle Benchmark* correspond au modèle GLM-Poisson ajusté uniquement sur les variables tarifaires classiques ; et pour la sévérité le modèle *Benchmark* correspond au modèle GLM-Tweedie ajusté uniquement sur les variables tarifaires classiques.

□ Comparaison globale de la performance prédictive des modèles

♣ Modèles de fréquence

Commençons par analyser les gains relatifs obtenus pour la fréquence de sinistres.

Lorsqu'on part d'un modèle de fréquence classique de type GLM, en prenant en compte uniquement les variables de risque classiques, et que par la suite on y augmente les variables télématiques, on réalise un gain relatif d'environ 5% sur la MSE et la MAE, et d'environ 2.5% sur la RMSE.

Lorsqu'on ajuste la fréquence de sinistres avec uniquement les variables de risque classiques et en utilisant cette fois un modèle plus sophistiqué tel qu'un *Random Forest*, on constate que les gains relatifs de performance réalisés sur la MSE et la RMSE sont infimes : de l'ordre de 0.77% et 0.40% respectivement. Ce gain relatif est 6 fois moins important que le gain relatif obtenu avec un modèle GLM lorsqu'on lui augmente les variables télématiques.

Ces résultats révèlent deux faits notables :

• Dans le cadre de notre étude, les données télématiques permettent indéniablement de réaliser une valeur ajoutée certaine et non négligeable dans l'amélioration de la performance prédictive des modèles de fréquence de sinistres.

• Au vu du gain relatif de performance infime réalisé par le *Random Forest* (utilisant uniquement les variables de risque classiques), on serait tenté de conclure qu'il n'est pas pertinent pour la modélisation de la fréquence dans notre étude et que l'on pourrait tout simplement simplement se limiter à la modélisation de la fréquence de sinistre par un simple modèle GLM.

Comme le soulignait déjà Delcaillau (2019) dans son mémoire d'actuariat, il est courant de voir que ces modèles complexes censés améliorer la performance prédictive sur les données étudiées ne le fassent pas toujours, sinon très marginalement, notamment en tarification automobile. Toutefois, ce gain négligeable de performance serait plus dû à la qualité des données utilisées (données d'ordre

publique, faiblement informatives sur le comportement de conduite de l'assuré) qu'à la structure même du modèle *Random Forest*.

Modèles	Métriques				Gain relatif (en %)		
	MSE	MAE	RMSE	$RMSE_{mean}$	MSE	MAE	RMSE
GLM-freq-classique (Benchmark)	0.0597765	0.09298055	0.2444923	5.004322	–	–	–
GLM-freq-complet	0.05687827	0.08876985	0.2384917	4.881498	4.85	4.53	2.45
RF-freq-classique	0.0593124	0.09186272	0.2435414	4.984857	0.78	1.20	0.40
RF-freq-complet	0.05128279	0.08603667	0.226457	4.635171	14.21	7.50	7.40
GLM-coût-classique (Benchmark)	1 301 920	250.6309	1 141.017	8.587901	–	–	–
GLM-coût-complet	1 256 309	237.5303	1 120.852	8.436126	3.50	5.30	1.76

Légende :
GLM-freq-classique : glm-poisson pour la fréquence, avec variables classiques uniquement (Benchmark fréq.) ;
GLM-freq-complet : glm-poisson pour la fréquence de sinistres, avec variables classiques et télématiques ;
RF-freq-classique : random forest pour la fréquence de sinistres, avec variables classiques uniquement ;
RF-freq-complet : random forest pour la fréquence de sinistres, avec variables classiques et télématiques ;
GLM-coût-classique : glm-Tweedie sévérité agrégée, avec variables classiques uniquement (Benchmark sév.) ;
GLM-coût-complet : glm-Tweedie pour la sévérité agrégée de sinistres, avec variables classiques et télématiques.

TABLE 5.12 : *Analyse de la performance des modèles avec et sans variables télématiques ; au milieu du tableau sont représentés les indicateurs de performance sur la base de test ; à droite sont représentés les gains relatifs par rapport au modèle GLM sans variables télématiques (appelé GLM classique).*

Étant donné que la prise en compte des variables télématiques amélioraient les performances du modèle GLM classique de départ, nous étions alors curieux de savoir comment évoluerait ce gain relatif si l'on remplaçait le modèle GLM par un modèle plus sophistiqué à même d'apprendre des relations plus complexes contenues dans les données. À cet effet, nous avons ajusté la fréquence par un modèle *Random Forest* sur l'ensemble des variables de risque classiques et variables télématiques. Les résultats sont plutôt impressionnants :

• Le gain relatif de performance réalisé avec le modèle *Random Forest* prenant en compte les données télématique est près de trois fois plus important que celui obtenu avec le *GLM* prenant en compte les variables télématiques, pour la MSE et la RMSE ;

La présence de variables tarifaires explicitement liées au comportement de conduite de l'assuré améliore donc la qualité des données disponibles pour la modélisation, et le modèle complexe *Random Forest* révèle clairement sa supériorité de précision prédictive par rapport au modèle GLM.

• Par rapport au modèle Benchmark (GLM avec variables classiques uniquement), le gain relatif de précision réalisé avec le modèle *Random Forest* prenant en compte l'ensemble des variables classiques et télématiques s'élève à plus de 14.21% pour la MSE et 7.40% pour la RMSE et la MAE.

Ces résultats montrent bien la plus value des données télématiques dans la modélisation de la fréquence de sinistre, dans notre cas d'études.

♣ Modèles de sévérité agrégée

En ce qui concerne l'utilité des données télématiques dans la modélisation de la sévérité agrégée de sinistres, les constats fait sont les mêmes que pour le cas de la fréquence :

• Dans notre étude, l'ajout des variables télématiques améliorent la précision prédictive des modèles de sévérité. On enregistre un gain relatif d'un peu plus de 5% sur la MAE.

• Il est important de rappeler que dans notre contexte, les modèles de sévérité agrégée correspondent directement aux modèles de prime pure. Ainsi, on pourrait dire que les données télématiques améliorent significativement la précision de la tarification automobile dans notre cas d'études.

• Comme dans le cadre de la modélisation de la fréquence, on peut penser qu'en utilisant un modèle plus sophistiqué dans le cadre de la modélisation de la sévérité, on améliorerait davantage le gain de performance prédictive.

□ Comparaison locale de la performance prédictive des modèles

Dans la partie précédente, nous avons pu constater que la prise en compte des données télématiques améliorent nettement la précision globale des modèles de sinistre (fréquence et sévérité).

Dans cette partie, nous nous intéressons au gain relatif de précision dû à l'utilisation des données télématiques, de manière beaucoup plus fine.

Par exemple, nous aimerons savoir : pour quels segments d'assurés la prise en compte des variables télématiques permet le mieux d'améliorer l'évaluation du risque ?

Pour répondre à cette question nous avons calculé le gain relatif de MSE et de MAE, et les prédictions moyennes issues de notre meilleur modèle de fréquence (respectivement de sévérité) sans variables télématiques à savoir le *RF-classique* (resp. le *GLM-classique*) et notre meilleur modèle de fréquence (resp. de sévérité) avec variables télématiques à savoir le *RF-complet* (resp. le *GLM-complet*), et ceci en faisant varier chaque fois les valeurs d'une seule variable explicative.

Les résultats obtenus sont présentés sur les figures 5.14 et 5.15. L'ensemble des calculs ont effectués sur les assurés de la base de test mise de côté dès le départ.

• Sur la figure 5.14, on constate clairement que l'amplitude de la plus-value des données télématiques dans la modélisation de la sinistralité varie suivant les segments de clientèle.

– Sur le segment des assurés âgés entre 16 et 20 ans, l'ajout des variables télématiques dans le meilleur modèle de fréquence sans variables télématiques, à savoir le *RF-classique*, réduit de 30% la valeur de la MSE, contre une réduction inférieure à un 20% pour les autres tranches d'âge.

– Dans le modèle de sévérité agrégé (prime pure), la prise en compte des données télématiques réduit de plus de 60% la MSE sur le segment des assurés les plus jeunes (16 à 20 ans).

– Pour ce même segment d'assurés (âgés entre 16 et 20 ans), on observe clairement qu'avec le modèle de sévérité complet (c'est-à-dire, celui prenant en compte les variables télématiques), on aboutit à des prédictions moyennes de fréquence et sévérité, relativement plus proche des valeurs réelles observées. Tandis que pour les autres tranches d'âge, les prédictions moyennes obtenues sont sensiblement les mêmes pour le modèle avec et sans variables télématiques et sont toutes les deux proches des valeurs moyennes réelles observées au niveau de la base de test.

– En ce qui concerne le score de crédit, on relève que la prise en compte des données télématique dans la modélisation de la fréquence de sinistre permet de réaliser un gain relatif de plus de 15% sur la MSE pour le segment d'assurés ayant un score de crédit inférieure à 800 points contre un gain relatif de MSE inférieure à 11% pour les autres segments.

• Sur la figure 5.15 on constate que :

– Pour la variable *Car.use*, pour la fréquence de sinistre, le gain relatif de précision (en termes de MSE) le plus important est observé chez les véhicules à usage commercial : Il est de l'ordre

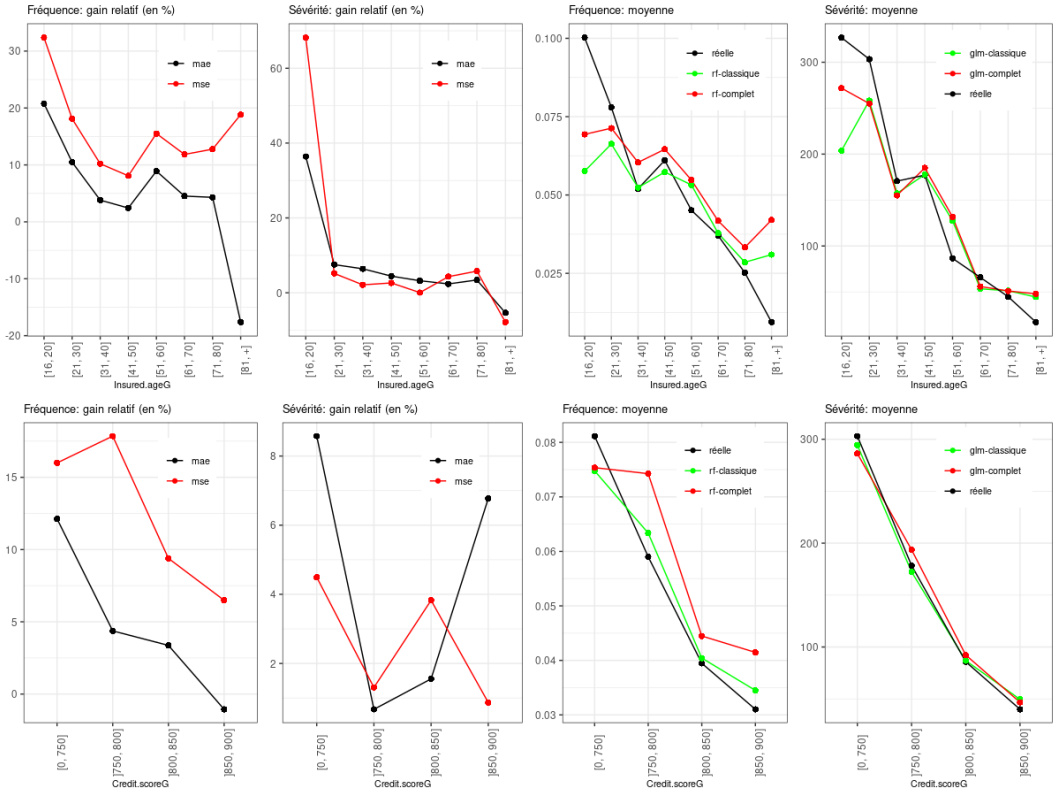


FIGURE 5.14 : *Gain de performance prédictive, fréquence moyenne prédite et sévérité agrégée (prime pure) moyenne prédite, selon les caractéristiques du conducteur assuré.*

de 20%. Les modèles avec et sans variables télématiques permettent d'aboutir à des prédictions moyennes de fréquence et de sévérité très voisines l'une de l'autre, quelque soit le type d'usage du véhicule, et toutes les deux proches des valeurs réelles observées au niveau de l'échantillon.

– Pour toutes nos classes d'âge de véhicule, prendre en compte les données télématiques permet d'améliorer les performances prédictives des modèles fréquence (à au moins 10%) et de sévérité en termes de MSE. On remarque néanmoins que l'amélioration est plus marquée chez les véhicules les plus récents (âgés de moins de 2 ans).

□ Une analyse concurrentiel entre les deux versions de modèles de sévérité agrégée : sans variables télématiques *vs* avec variables télématiques

Dans cette partie, nous nous plaçons dans un contexte de concurrence entre deux assureurs : l'un utilisant le modèle GLM-Tweedie[6] sans variables télématiques (*GLM-coût-classique*) pour fixer le niveau de prime pure de ses assurés et l'autre utilisant le modèle de sévérité avec variables télématiques (*GLM-coût-classique*).

Nous faisons les trois hypothèses simplificatrices suivantes :

(*i*) Les deux assureurs sont les seuls présents sur le marché : l'assureur A (celui utilisant le modèle *GLM-coût-classique*) ; l'assureur B (celui utilisant le modèle *GLM-coût-complet*).

(*ii*) On suppose que la prime commerciale se réduit à la prime pure. Autrement dit, on se place dans un contexte d'absence de chargement de sécurité, de chargement destiné à couvrir les frais de la compagnie et d'absence de taxes.

[6]Comme évoqué précédemment, les modèles GLM avec loi de Tweedie permettent d'ajuster directement la prime pure, sans recours à l'approche fréquence/sévérité.

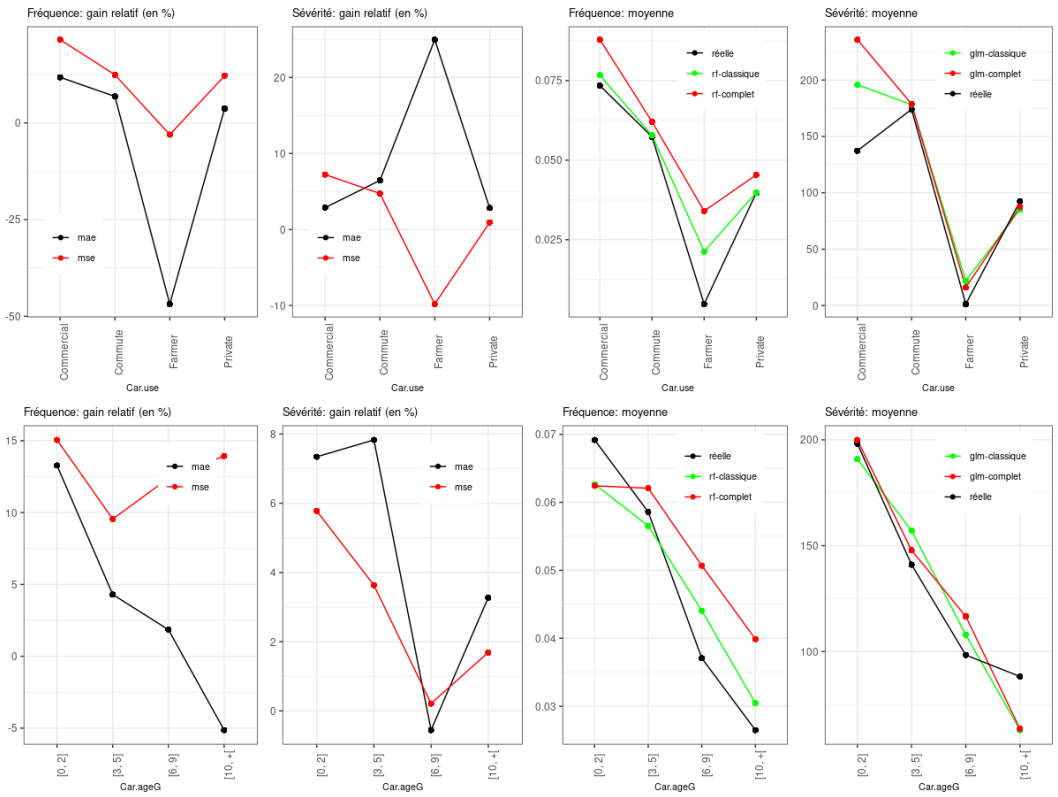


FIGURE 5.15 : *Gain de performance prédictive, fréquence et sévérité agrégé (prime pure) moyenne prédite, selon les caractéristiques du véhicule assuré.*

(*iii*) En outre, nous supposons que les assurés sont *homoéconomicus* (économiquement rationnels). Chacun d'entre eux se dirige vers l'assureur qui lui propose la prime la moins chère.

L'objectif ici est tout simplement de montrer que dans un contexte concurrentiel, l'usage de données télématiques pourrait améliorer considérablement le résultat de l'assureur (primes perçues moins les prestations versées) dans sa globalité, et même sur certains segments spécifiques de sa clientèle.

L'hypothèse (*iii*) nous permet de séparer notre base de test en deux sous populations : la première est constituée des assurés qui s'assure chez l'assureur A, c'est-à-dire ceux pour lesquels le tarif obtenu par le *GLM-coût-classique* est inférieur à celui obtenu par le *GLM-coût-classique* et la deuxième sous population est celle des assurés qui vont chez l'assureur B. On obtient alors la répartition décrite par le diagramme 5.16.

Analysons succinctement le compte de résultat des deux assureurs de manière globale et sur quelques segments de clientèle spécifiques.

D'entrée de jeu, on observe sur le diagramme 5.16 que l'assureur B attire la plus grosse part du marché (64%). Cela s'explique par le fait que la prise en compte des variables télématiques dans son modèle de tarification (*GLM-Tweedie-complet*) lui permet d'évaluer au plus juste le risque des assurés là où le modèle de l'assureur A (*GLM-Tweedie-classique*) sur-tarifie.

Cependant, le fait de détenir une clientèle plus nombreuse n'est pas toujours synonyme de rentabilité. Afin d'évaluer la rentabilité de son activité, l'assureur doit examiner son compte de résultat. Pour ce faire, il utilise plusieurs indicateurs de rentabilité, dont l'un des plus usuels est le *ratio de perte* (en anglais, *loss ratio*). Il est obtenu en effectuant le rapport entre le montant total des sinistres à indemniser (S) et le montant total des primes encaissées (P). Le *ratio de perte* doit

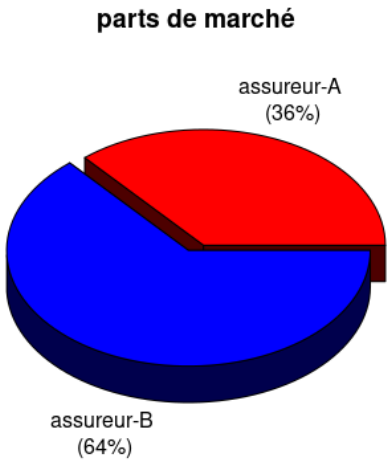


FIGURE 5.16 : *Part de marché des assureurs A et B dans la base de test sous l'hypothèse (iii) précédente. L'assureur-A tarifie avec un modèle GLM-Tweedie sans utiliser les variables télématiques (GLM-coût-classique) et l'assureur-B tarifie avec le même type de modèle mais utilise en plus les variables télématiques (GLM-coût-complet).*

être inférieur à 1 pour couvrir les coûts de gestion de l'assureur et assurer une bonne rentabilité. Les résultats de l'analyse des comptes de résultats des assureurs A et B sont répertoriés dans le tableau 5.13.

Segments	Primes (P)		Sinistres (S)		Résultats (P-S)		Loss Ratio (S/P)	
	A	B	A	B	A	B	A	B
global	**732 866**	**544 128**	**1 420 774**	**533 380**	**-687 908**	**+10 748**	**1.94**	**0.98**
age-client[16, 30]	140 924	122 905	407 573	80 186	-266 649	+42 719	2.89	0.65
age-véhicule[0, 5]	481 753	377 813	952 250	333 873	-470 497	+43 940	1.98	0.88
véhicule-privé	213 167	162 718	470 199	159 525	-257 032	+3 193	2.21	0.98
véhicule-commun	483 759	361 988	901 434	370 144	-417 675	-8 156	1.95	1.02
véhicule-agricole	1 095	1 489	276	0	+819	+1 489	0.25	0

TABLE 5.13 : *Analyse comparative de la rentabilité du portefeuille de l'assureur A et celui de l'assureur B.*

De manière globale,

• Pour l'assureur A le ratio de perte est supérieure à 1 (1.94), tandis qu'il est inférieur à 1 pour l'assureur B (0.98). L'assureur A a donc un résultat global négatif : il est déficitaire ; alors que l'assureur B tire des bénéfices de son activité.

• Lorsqu'on évalue la rentabilité de certains segments spécifiques de marchés, généralement considérés comme les plus à risque ou les plus coûteux, notamment celui des jeunes conducteurs ou des véhicules les plus récents, on se rend compte que l'inégalité de compétitivité entre les deux assureurs est davantage sévère. L'assureur B utilisant les données télématiques est plus de loin plus compétitif que l'assureur A n'utilisant pas

• En ce qui concerne la part de marché constituée par les véhicules à usage commun, les deux assureurs y sont contre-performants avec des loss ratio tous supérieure à 1. Toutefois, le déficit est moins accentué chez l'assureur B utilisant les données télématiques pour tarifer que chez l'assureur A ne les utilisant pas (1.95 contre 1.02).

On peut donc conclure que dans le jeu de la concurrence entre les deux assureurs, le modèle de tarification de l'assureur B, l'emporte sur celui de l'assureur A : grâce à son modèle plus pointu l'assureur B s'attire les "bons risques" et les "mauvais risques" vont chez l'assureur A.

Par cette représentation simplifié de la réalité, nous avons montré de manière très schématique que l'usage des données télématiques dans le processus de tarification en assurance automobile peut être d'une *plus-value* capitale pour la santé de l'activité de l'assureur. Nous avons également montré que les données télématiques peuvent permettre de mieux appréhender la sinistralité au niveau des segments de marché les plus à risque parfois difficile à cerner uniquement par des variables de risque classiques.

5.3.2 Prédiction des données télématiques

Dans la sous-section précédente nous avons montré que les variables télématiques permettent de tarifer au plus juste les contrats d'assurance automobile.

Cependant, en pratique, les assureurs n'ont pas toujours le luxe de disposer de données télématiques sur leur clientèle. Ceci pour diverses raisons que nous ne détaillons pas dans ce mémoire. Notons néanmoins que les principales raisons seraient d'une part liées à la réticence des automobilistes à partager leurs données personnelles de conduite, et d'autre part, le coût non négligeable pour l'assureur, de l'installation d'équipements nécessaires à collecte des données télématiques des assurés.

Dans cette section, il est question de prédire les valeurs des variables télématiques à partir des caractéristiques observables chez l'assuré à la souscription du contrat, notamment : l'âge de l'assuré, l'âge de son véhicule, l'usage fait de son véhicule, son score de crédit, sa durée de couverture, le nombre annuel de miles prévus à parcourir déclaré par l'assuré à la souscription.

Nous avons essayé de prédire ces variables télématiques à l'aide de plusieurs types de modèles : un modèle logistique, l'algorithme CART et un modèle de Forêt aléatoire. Il s'en est sorti que le modèle *Random Forest* avait de meilleurs performances sur notre échantillon de test. Nous l'avons donc retenu pour la prédiction de variables télématiques.

Étant donné que nous ne disposons que de très peu de prédicteurs pour l'ajustement des différentes variables télématiques (06 prédicteurs au total), dans nos différents modèles *Random Forest*, nous optimiserons uniquement l'hyperparamètre *ntrees* relatif au nombre d'arbres constituant la forêt aléatoire. L'hyperparamètre *mtry* est fixé à 5 et les autres hyperparamètres sont maintenus à leur valeur par défaut.

5.3.2.1 Total.miles.drivenB : distance totale annuelle parcourue en miles

Commençons par prédire la classe du conducteur en termes de distance totale annuelle parcourue (*Total.miles.drivenB*) en fonction de ses caractéristiques de risque classiques. *Total.miles.drivenB* est une variable binaire dont les deux modalités sont "[0, 4500]" et "]4500,+[correspondant respectivement à une distance totale annuelle parcourue de moins de 4500 miles et de plus de 4500 miles.

Les bases d'apprentissage et de test sont celles fixées dès le départ. Les sorties de l'optimisation de l'hyperparamètre *ntrees*, la courbe roc (sur l'échantillon de test) et l'importance des variables

du modèle optimal retenu sont présentés sur la figure 5.17. On retient $ntrees^* = 200$, ce qui nous donne un AUC autour de 81%.

Les trois variables les plus importantes dans ce modèle sont la durée de couverture de l'assuré, le nombre de miles annuels prévus à parcourir déclaré par le client à la souscription et l'âge du véhicule.

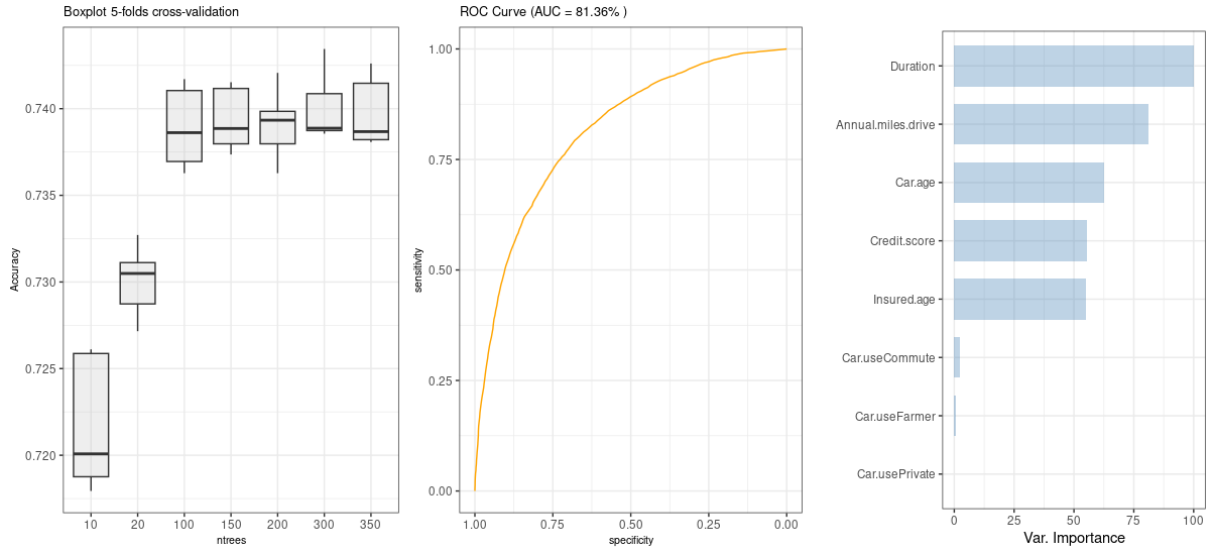


FIGURE 5.17 : *Prédiction de Total.miles.drivenB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntrees optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

5.3.2.2 Annual.pct.drivenB : pourcentage annualisée du temps passé sur la route

En ce qui concerne la variable binaire *Annual.pct.drivenB* correspondant à la classe du pourcentage de jours de conduite durant l'année avec pour modalités [0%, 50%] et]50%, 100%], après ajustement du modèle, les résultats obtenus sont présentés sur la figure 5.18.

Le nombre d'arbre optimal que nous retenons est $ntrees^* = 250$, ce qui nous donne un AUC d'environ 79%.

Les trois variables les plus importantes dans ce cas sont : en premier lieu, la durée de couverture de l'assuré, puis son score de crédit (généralement corrélée au niveau de revenus et la catégorie socioprofessionnelle, etc.) et l'âge du véhicule. À ce stade, nous n'interpréterons pas d'avantage le modèle, car il est avant tout à vocation prédictive plutôt qu'explicative.

5.3.2.3 Avgdays.weekB : nombre moyen de jours de conduite par semaine

Le nombre moyen de jours de conduite par semaine *Avgdays.weekB* est une variable binaire, dont les deux modalités sont : [0, 4] et]4, 7].

Sur le graphique 5.19 on peut lire l'AUC du modèle prédictif retenu qui est de 82.58%. Ce qui correspond à une bonne performance prédictive.

Les variables les plus influentes dans ce modèles sont :

– Tout d'abord, l'âge du véhicule : il est possible que les véhicules plus anciens soient moine tomber en panne ;

– Ensuite, vient le score de crédit de l'assuré ;

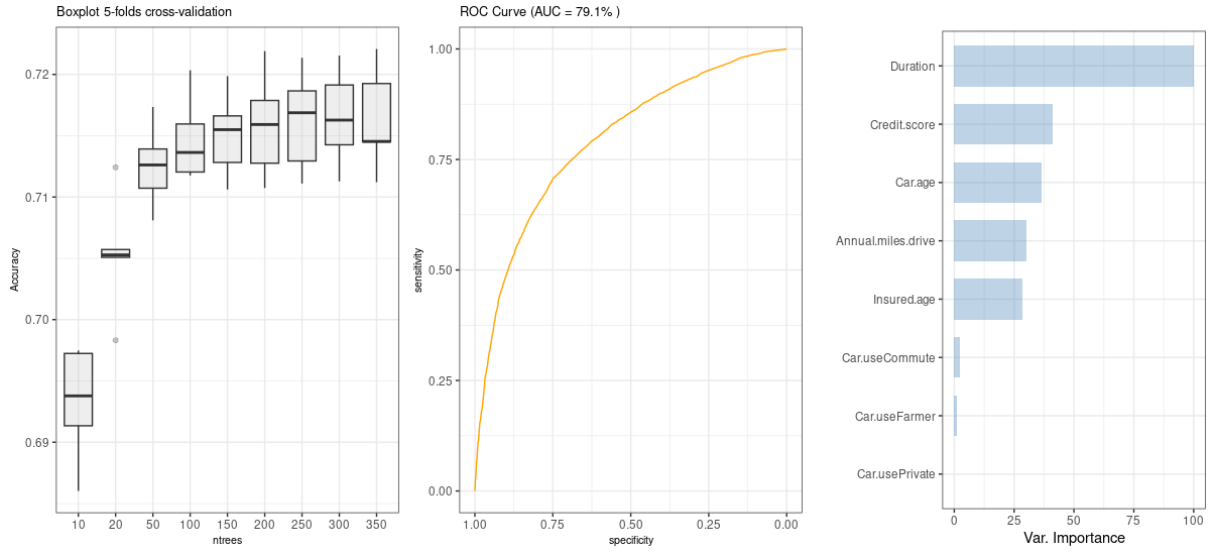


FIGURE 5.18 : *Prédiction de Annual.pct.drivenB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntrees optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

– Enfin, l'âge du conducteur assuré : les jeunes conducteurs sont généralement plus susceptibles à conduire le plus souvent que les conducteurs plus âgés. Les conducteurs plus âgés peuvent souvent avoir des emplois à temps chargés qui les empêchent de conduire autant que les conducteurs plus jeunes.

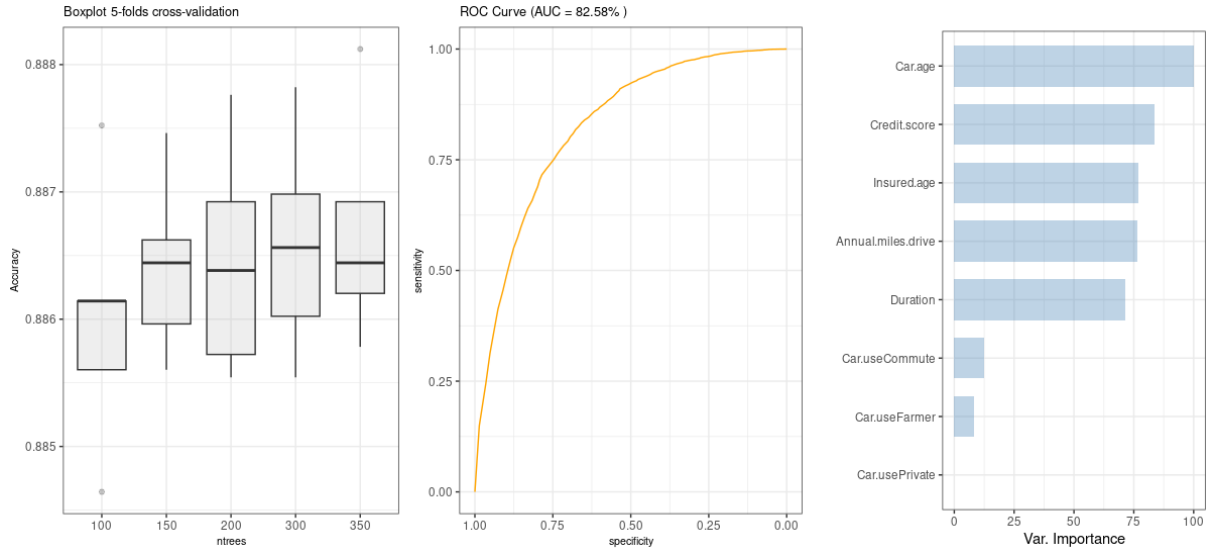


FIGURE 5.19 : *Prédiction de Avgdays.weekB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntrees optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

5.3.2.4 Accel.14milesB : nombre total annuel d'accélérations brusques d'intensité 14 mph/s

En ce qui concerne les variables télématiques relatives au comportement de l'assuré au volant, nous avons tout d'abord prédit les variables relatives au nombre d'accélérations brusques par 1000

miles d'intensité 14, 12, et 06 miles par heure par seconde (mph/s).

Comme le montre la figure 5.20 et les figures C.14, C.13 (en annexe), pour chacun des trois modèles optimaux ajustés pour ces trois variables, l'AUC varie entre 74% et 76%, ce qui est un indicateur d'une performance prédictive "acceptable".

Comme le montre les diagrammes d'importance de variables, on constate que l'âge du véhicule et l'âge de l'assuré sont parmi les caractéristiques les plus influentes dans l'ajustement du nombre d'accélération brusque d'intensité 14 et 12 mph/s.

Les jeunes conducteurs sont plus susceptibles à avoir des accélérations brusques que les conducteurs plus âgés, généralement plus prudents. L'âge du véhicule peut également avoir un impact indirect sur le nombre d'accélérations brusques. Par exemple, un véhicule plus ancien peut avoir des pièces usées qui peuvent occasionner des accélérations brusques involontaires. De plus, un véhicule plus ancien peut avoir une puissance moindre, ce qui peut pousser le conducteurs à multiplier des accélérations pour compenser la faible puissance de son véhicule.

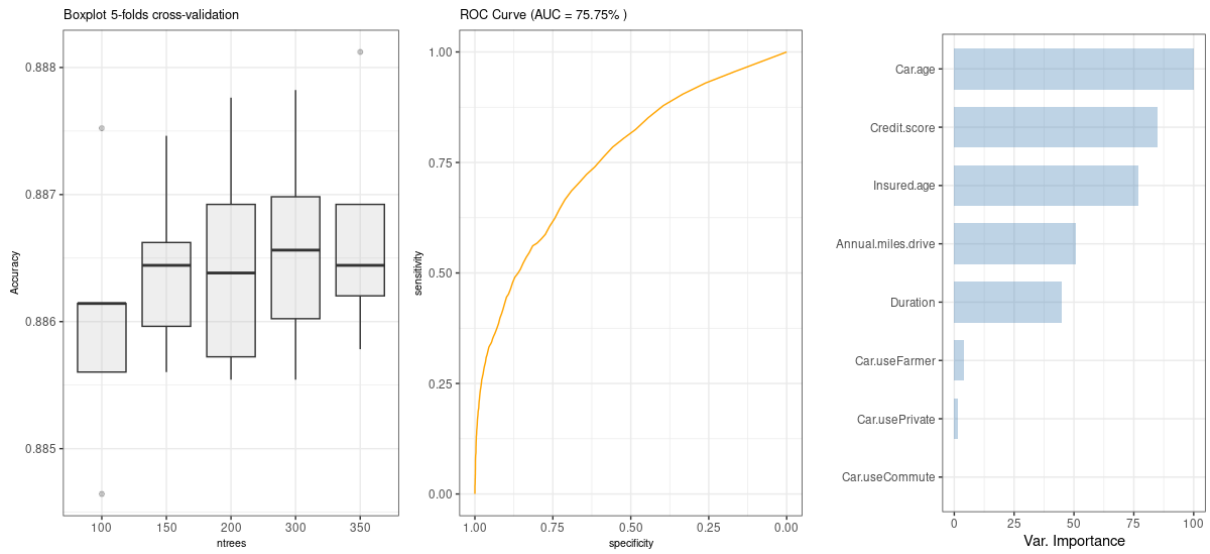


FIGURE 5.20 : *Prédiction de Accel.14milesB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntrees optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

5.3.2.5 Brake.09milesB : nombre total annuel de freinages brusques d'intensité 09 mph/s

Sur la figure 5.21 (et la figure C.15 en annexe), nous avons les sorties relatives aux modèles utilisés pour la prédiction du nombre de freinages brusques par 1000 miles d'intensité 09 mph/s et 08 mph/s respectivement.

Les niveaux de performance prédictives sont plutôt bons avec des AUC de 73%.

En ce qui concerne l'importance des variables dans le modèle, le constat est le même que dans le cas précédant des variables d'accélérations brusques. L'âge de l'assuré, l'âge du véhicule et le score de crédit de l'assuré sont les plus influentes à côté du nombre de miles prévu à parcourir durant l'année déclaré à la souscription par l'assuré.

En effet, un conducteur qui roule à une vitesse excessive aura plus de chances de devoir effectuer des freinages brusques pour éviter des obstacles.

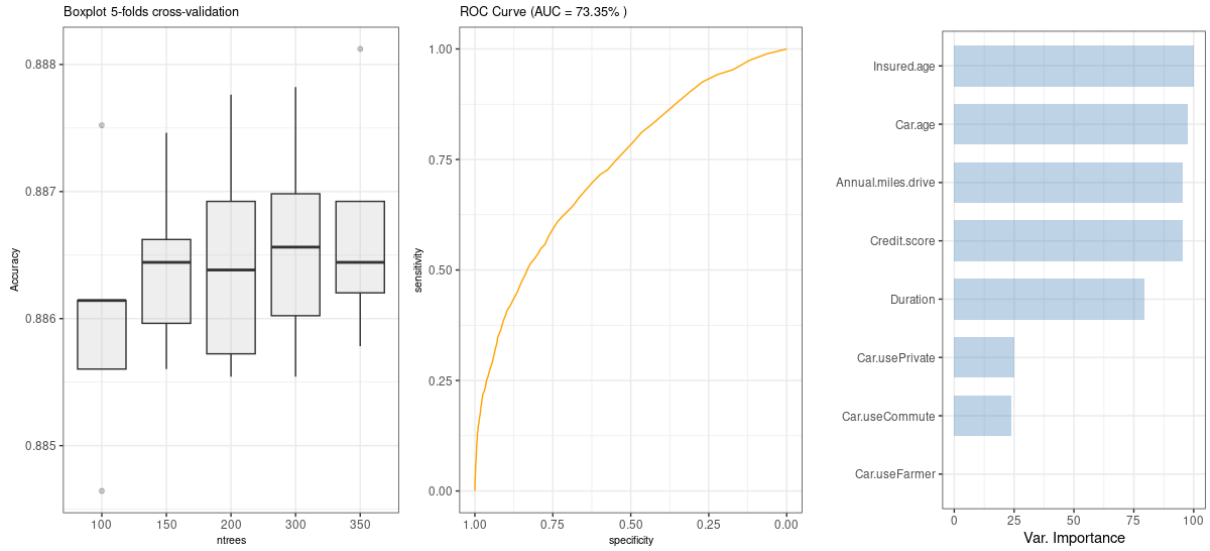


FIGURE 5.21 : *Prédiction de Brake.09milesB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntrees optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

5.3.2.6 Left.turn.intensity08B : nombre total annuel de virages à gauche par 1000 miles avec intensité 08 mph/s

En ce qui concerne la variable relative au nombre de virages à gauche par 1000 miles avec intensité 08 mph/s, comme le montre le graphique d'importance des variables sur la figure 5.22, les trois caractéristiques les plus importantes sont la durée de la couverture de l'assuré, l'âge de son véhicule et son propre âge.

Le modèle retenu pour la prédiction possède un AUC de 77%. Ce qui est caractéristique d'une bonne performance prédictive.

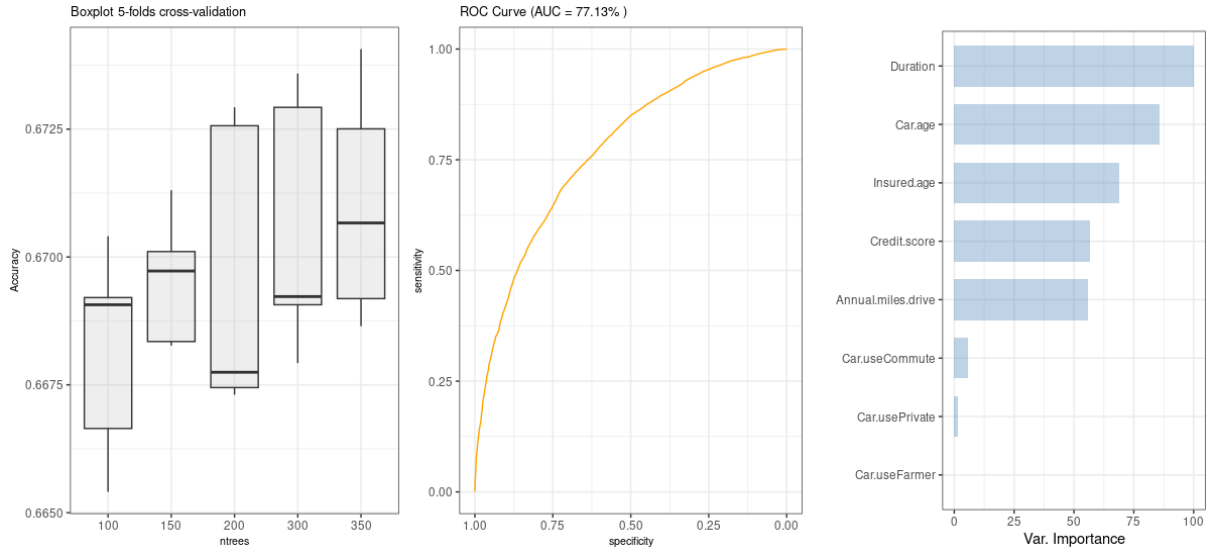


FIGURE 5.22 : *Prédiction de Left.turn.intensity08B : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntrees optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

5.3.2.7 Pct.drive.4hrsB : pourcentage annualisé de conduite d'une durée de 4 heures

La variable *Pct.drive.4hrsB* mesure le pourcentage de trajets d'une durée de 4 heures menés par conducteur assuré durant l'année, dans le total de ses trajets de conduite. Dans notre modèle, comme le montre la figure 5.23 les variables les plus importantes dans la prédiction de cette variable sont : l'âge du conducteur, l'âge de son engin, son score de crédit et d'autres variables liées au type d'usage du véhicule.

Notre modèle optimal (pour $ntrees^* = 100$ possède un AUC de l'ordre de 70%, ce qui est une performance prédictive "acceptable".

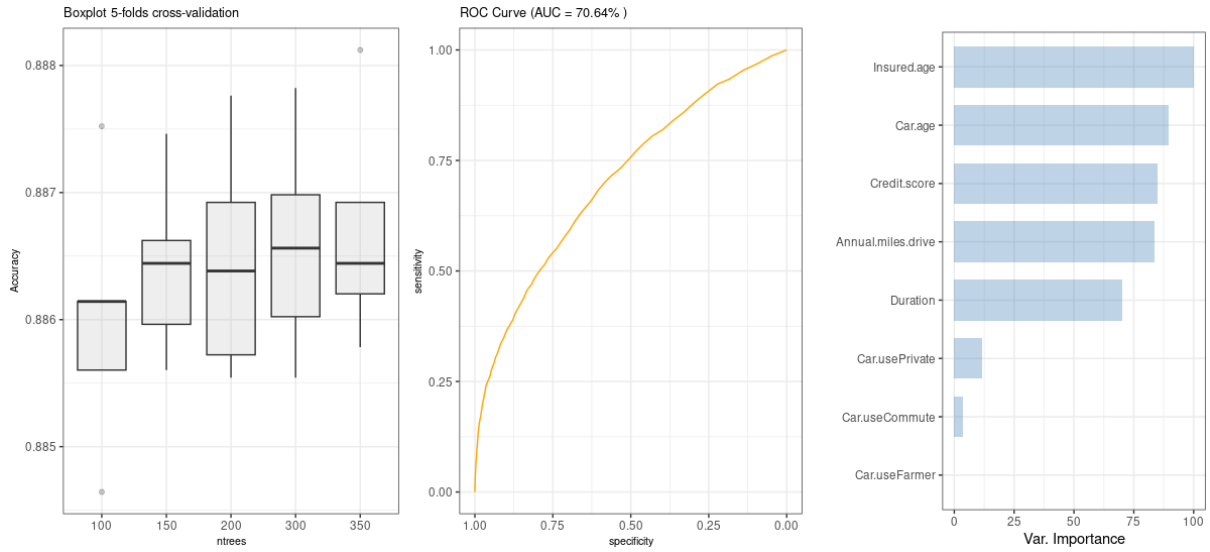


FIGURE 5.23 : *Prédiction de Pct.drive.4hrsB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntrees optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

5.3.2.8 Pct.drive.rush.amB : pourcentage annualisé de conduite pendant les heures de pointe en matinée

Le pourcentage annualisé de conduite d'un assuré pendant les heures de pointe en matinée est le pourcentage de kilomètre qu'un assuré a conduit entre 6h00 et 9h00 du matin durant une année. Il se mesure en rapportant le nombre total annuel de kilomètres parcourus entre 6h00 et 9h00 du matin à la distance totale parcourue durant l'année. Le pourcentage annualisé de conduite d'un assuré pendant les heures de pointe en soirée (*Pct.drive.rush.pmB*) est le pourcentage de kilomètres qu'un assuré passe à conduire entre 17h00 et 21h00 durant l'année.

Sur le graphique 5.24 (et le graphique C.17 en annexe), on peut bien voire que les modèles ajustés pour prédire ces variables ont de bonnes performances en terme de AUC (77,50% et 78.95%). De plus, nous notons que dans les deux cas, les variables les plus importantes dans les modèles de prédiction sont l'âge de l'assuré, l'âge de son véhicule et le score de crédit de l'assuré.

A présent nous avons ajusté l'ensemble des modèles de prédiction des 13 variables télématiques. Ces modèles ont été utilisés pour prédire les valeurs de variables télématiques pour l'ensemble des assurés de notre portefeuille de polices d'assurance initial. Les 13 variables prédites obtenues sont ensuite ajoutées aux variables explicatives de risque classiques pour la modélisation de la sinistralité.

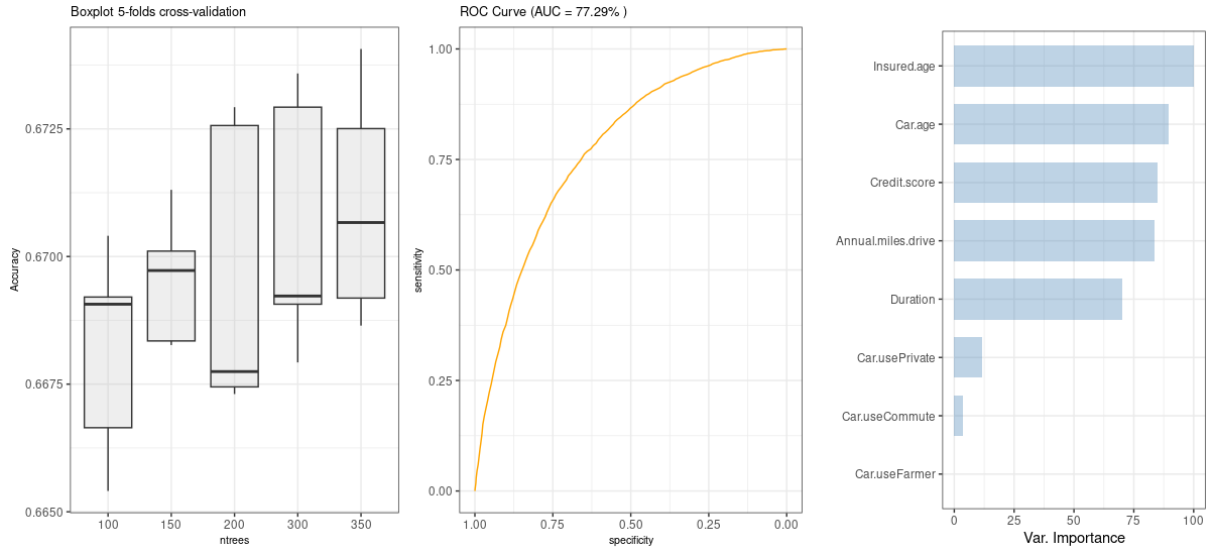


FIGURE 5.24 : *Prédiction de Pct.drive.rush.amB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntrees optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

Pour chacune des 13 variables télématiques binaires prédites, nous avons conservé comme valeur prédite, la probabilité prédite d'appartenance à une classe de référence (la classe supérieur de la variable), plutôt que la valeur de classe prédite par le modèle. Par exemple, pour la variable *Total.miles.drivenB*, à l'issue de sa prédiction, nous avons retenu comme nouvelle variable explicative la probabilité prédite d'appartenir à la classe]4500, +[pour chaque assuré, plutôt que la classe prédite par le modèle. Un tel choix permet d'introduire davantage de flexibilité dans l'ajustement des modèles de sinistralité. Le tableau 5.14 récapitule la description variables télématiques prédites qui seront augmentées aux différents modèles de sinistre dans la section suivante.

Variables télématiques prédites	Description
Accel.06miles_fit	Probabilité prédite d'appartenance de *Accel.06miles* à :]30, +[
Accel.11miles_fit	Probabilité prédite d'appartenance de *Accel.11miles* à :]0, +[
Accel.14miles_fit	Probabilité prédite d'appartenance de *Accel.14miles* à :]0, +[
Brake.08miles_fit	Probabilité prédite d'appartenance de *Brake.08miles* à :]10, +[
Brake.09miles_fit	Probabilité prédite d'appartenance de *Brake.09miles* à :]5, +[
Left.turn.int.08_fit	Probabilité prédite d'appartenance de *Left.turn.int.08* à :]150, +[
Total.miles.driven_fit	Probabilité prédite d'appartenance de *Total.miles.driven* à :]4500, +[
Avgdays.week_fit	Probabilité prédite d'appartenance de *Avgdays.week* à :]4, 7]
Annual.pct.driven_fit	Probabilité prédite d'appartenance de *Annual.pct.driven* à :]50, 100]
Pct.drive.2hrs_fit (%)	Probabilité prédite d'appartenance de *Pct.drive.2hrs* à :]0, 100]
Pct.drive.4hrs_fit (%)	Probabilité prédite d'appartenance de *Pct.drive.4hrs* à :]0, 100]
Pct.drive.rush.am_fit (%)	Probabilité prédite d'appartenance de *Pct.drive.rush.am* à :]5, 100]
Pct.drive.rush.pm_fit (%)	Probabilité prédite d'appartenance de *Pct.drive.rush.pm* à :]10, 100]

TABLE 5.14 : *Résumé des treize (13) variables télématiques prédites qui seront augmentées dans l'ajustement des modèles de sinistralité à la section suivante.*

5.4 Modélisation de la sinistralité

Dans cette section nous modélisons la fréquence et la sévérité agrégée des sinistres, à l'aide des modèles de régression classiques utilisés en tarification non-vie et des modèles plus complexes d'apprentissage statistique, et ce, en y ajoutant les variables télématiques prédites dans la partie précédente.

Pour la modélisation de la fréquence et de la sévérité, nous avons ajusté trois types de modèle, à savoir :

- un modèle linéaire généralisé (avec loi de Poisson pour la fréquence et avec loi de Tweedie pour la sévérité) ;
- un modèle LocalGLMnet ;
- un modèle de forêt aléatoire.

5.4.1 Modèles de fréquence mis en oeuvre

GLM Poisson

La procédure d'ajustement est la même que celle abordé dans la section 5.3.1 où nous mettions en exergue la valeur ajoutée des données télématiques dans la modélisation de la fréquence de sinistres. La seule différence est qu'ici, à la place des variables télématiques, nous utilisons les valeurs prédites des variables télématiques −plus précisément, les probabilités prédites d'appartenance aux classes spécifiques de ces variables détaillées dans le tableau 5.14. Compte tenu de l'équidispersion de la fréquence de sinistre dans notre jeu de données, nous ajustons la fréquence de sinistres par un modèle GLM avec une loi de Poisson.

Les résultats de l'estimation des coefficients du modèle mis en place sont répertoriés dans le tableau C.4 en annexe.

LocalGLMnet avec loi de Poisson

Ensuite, nous ajustons la fréquence par une architecture LocalGLMnet avec loi Poisson, ayant une profondeur $d = 4$. Les différentes couches ont $(q_0, q_1, q_2, q_3, q_4) = (24, 18, 12, 6, 24)$ neurones pour la première jusqu'à la dernière couche.

Dans le modèle LocalGLMnet, nous n'avons pas nécessairement besoin d'utiliser les versions catégorisées des variables explicatives numériques (à savoir l'âge des assurés, l'âge des véhicules, le score de crédit, le nombre de miles prévu à parcourir dans l'année déclaré par l'assuré à la souscription) comme dans les GLM. Nous avons introduit ces variables en l'état dans le modèle.

Pour l'implémentation, nous nous sommes inspiré du notebook du cours *Deep Learning with Actuarial Applications in R* de l'Association suisse des actuaires, et qui sert d'accompagnement à la mise en oeuvre des modèles LocalGLMnet. Ledit notebook est accessible via le lien github suivant : [mise en oeuvre du modèle LocalGLMnet avec R].

Random Forest

Nous terminons par l'ajustement d'un modèle complexe d'apprentissage statistique de type *Random Forest*. Après optimisation des hyperparamètres principaux, nous avons retenu comme nombre d'arbres optimal $ntrees = 200$ et comme nombre de variables échantillonnées pour la formation des différents noeuds, $mtry = 8$.

5.4.1.1 Analyse comparative de la qualité d'ajustement des modèles mis en place

Pour chacun des trois modèles mis en place, analysons le nuage de points des résidus de Pearson contre les valeurs prédites. Pour un modèle bien ajusté, ces deux quantités doivent être asymptotiquement décorrélées. Autrement dit : mieux le nuage est dispersé, mieux le modèle est bien ajusté.

Sur la figure 5.25, on observe que le modèle LocalGLMnet est celui pour lequel le nuage de points est le plus dispersé. On retient donc que parmis les tois modèles de fréquence mis en place, le modèle LocalGLMnet serait celui qui ajuste le mieux la fréquence de sinistres.

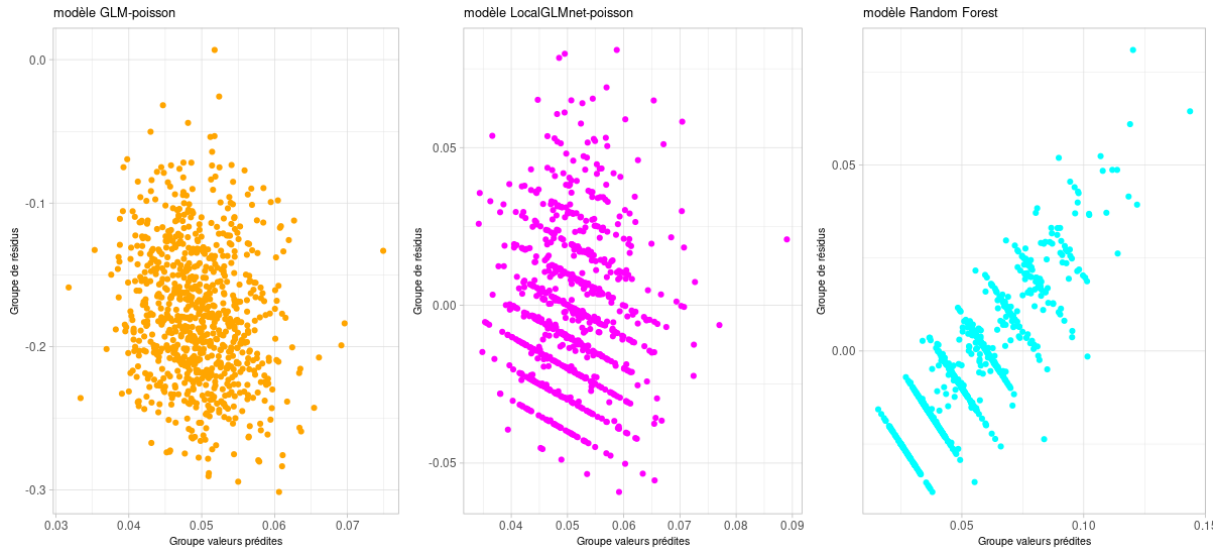


FIGURE 5.25 : *Nuages des résidus vs. prédictions pour chaque modèle de fréquence mis en place.*

5.4.1.1 Analyse comparative de la performance prédictive des modèles

Afin d'évaluer la performance prédictive des modèles mis en oeuvre, nous utilisons des métriques de précision présentées plus haut dans ce chapitre 1, à savoir la MSE, la MAE et la RMSE_mean, que nous calculons sur la base de test.

Nous retenons comme meilleur modèle celui ayant la plus petite RMSE. Dans le tableau 5.15, on constate que l'ensemble des modèles mis en place ont des performances prédictives voisines.

En outre, la figure 5.26 des *Boxplots* des fréquences prédites par les trois modèles montre que les distributions de ces fréquences prédites sont assez voisines.

Le modèle LocalGLMnet avec loi de Poisson se démarque de très peu des deux autres modèles avec une RMSE de 0.2407691, soit un gain relatif de +0.64% par rapport au modèle GLM-Poisson et de 0.08% par rapport au *Random Forest*.

Pour l'ensemble des trois modèles implémentés, le ratio de la fréquence moyenne prédite par la fréquence moyenne observée au niveau de la base test est inférieure à 1. Cela signifie que dans la globalité, les trois modèles de fréquence implémentés sous estime la fréquence effective de sinistres des assurés. Le *Random Forest* avec un ratio de 0.97, est celui qui permet le mieux de se rapprocher de la fréquence moyenne effective observée au niveau de la base test.

Pour aller plus loin, nous pouvons comparer les résultats des tableaux 5.12 et 5.15. On observe que le meilleur modèle de fréquence sans variables télématiques à savoir le modèle *RF-freq-classique* (confère tableau 5.12) performe moins bien en termes de RMSE, que le meilleur modèle de fréquence

avec variables télématiques prédites, à savoir le modèle *LocalGLMnet-Poisson* (confère tableau 5.15), avec un gain relatif de $+1,14\%$ en faveur de ce dernier.

Ce dernier résultat montre la plus-value, même si infime, des variables télématiques prédites dans la modélisation de fréquence.

Cependant, le modèle fréquence le plus performant utilisant les variables télématiques prédites, à savoir le modèle LocalGLMnet-Poisson du tableau 5.15 reste moins précis que le modèle fréquence le plus performant avec variables télématiques réelles, à savoir *RF-freq-complet* du tableau 5.12, avec un écart relatif de l'ordre $+6\%$ sur la RMSE. Ce résultat interroge sur la qualité des données télématiques prédites qui se seraient vues dénaturées suite au processus de prédiction.

Modèles	moy. pred./moy. réelle	MSE	MAE	RMSE	$RMSE_{mean}$
GLM-Poisson	0.8794369	0.05871528	0.08529553	0.2423124	4.959702
LocalGLMnet-Poisson	0.9115864	0.05796977	0.08565878	0.2407691	4.928114
Random Forest	0.9767824	0.05806044	0.08783279	0.2409573	4.931966

TABLE 5.15 : *Comparaison de la performance prédictive des modèles de fréquence.*

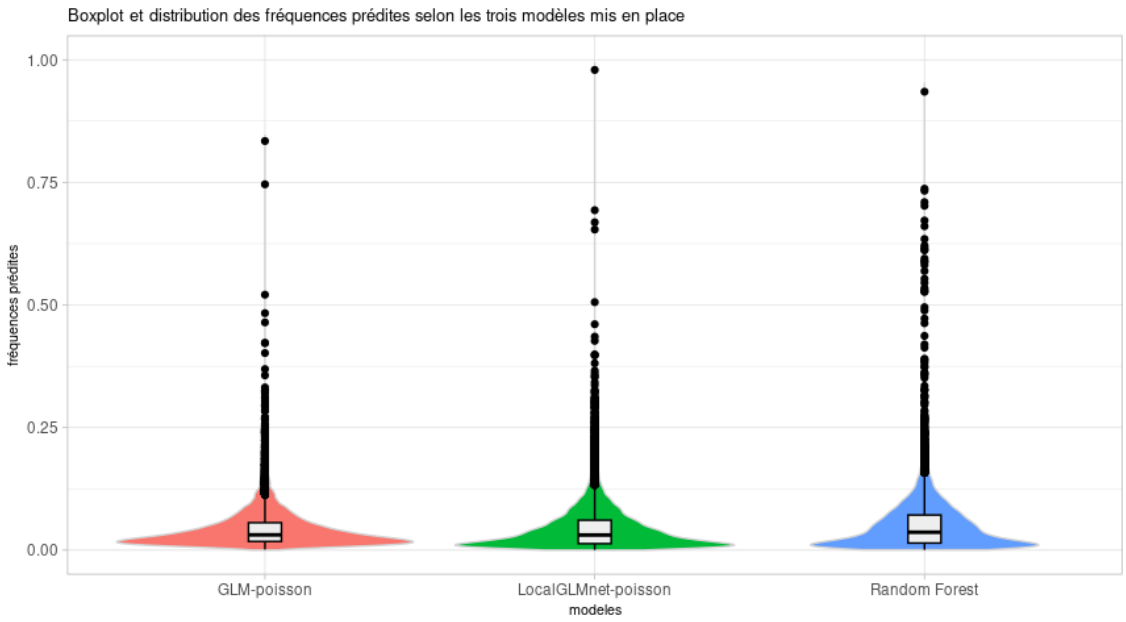


FIGURE 5.26 : *Boxplot et distributions des fréquences prédites sur la base de test, selon les trois modèles mis en place.*

5.4.2 Modèles de sévérité mis en place

Pour la modélisation de la sévérité agrégée (qui correspond ici à la prime pure annuelle), l'approche est la même que pour les modèles de fréquence.

GLM-Tweedie

Nous commençons par l'ajustement d'un modèle GLM avec loi de Tweedie. En effet, dans notre base de données nous ne disposons pas d'informations sur le montant des sinistres individuels, mais plutôt sur le montant agrégée des sinistres sur toute la période de couverture. Dans un tel contexte,

la loi de Tweedie est l'une des plus appropriées parmi les lois exponentielles pour modéliser le montant agrégée de sinistres.

Les résultats obtenus à l'issue de l'ajustement du modèle sont disponibles dans le tableau C.5 en annexe.

LocalGLMnet-Tweedie

Ensuite, nous ajustons le montant agrégé des sinistres par une architecture LocalGLMnet avec loi Tweedie, ayant une profondeur d = 4, avec un nombre de neurones $(q_0, q_1, q_2, q_3, q_4) = (24, 18, 12, 6, 24)$ par couche.

Random Forest

Comme dans le cadre de la modélisation de la fréquence, nous terminons par l'ajustement d'un modèle complexe d'apprentissage statistique de type *Random forest*. Après optimisation des hyperparamètres principaux, nous retenons comme nombre d'arbres optimal $ntrees = 200$ et comme nombre de variables échantillonnées pour la formation des différents noeuds, $mtry = 8$.

5.4.2.1 Analyse comparative de la qualité d'ajustement des modèles mis en place

Nous comparons les résidus standardisés contre les valeurs de sévérité prédites pour chaque modèle mis en place.

On remarque sans trop d'efforts sur la figure 5.27 que c'est le nuage du modèle GLM avec loi Tweedie, qui est le plus dispersé. Donc, il s'agirait du modèle qui s'ajuste le mieux les montants agrégés de sinistres au niveau de notre base de données.

Cependant, le modèle qui ajuste le mieux les données d'entraînement n'est pas nécessairement celui pour lequel on obtient de meilleures performances prédictives sur des données non vues par le modèle (sur la base de test). C'est pour cette raison que nous procédons ensuite, dans le paragraphe suivant, à l'analyse des performances prédictives de nos différents modèles sur la base de test.

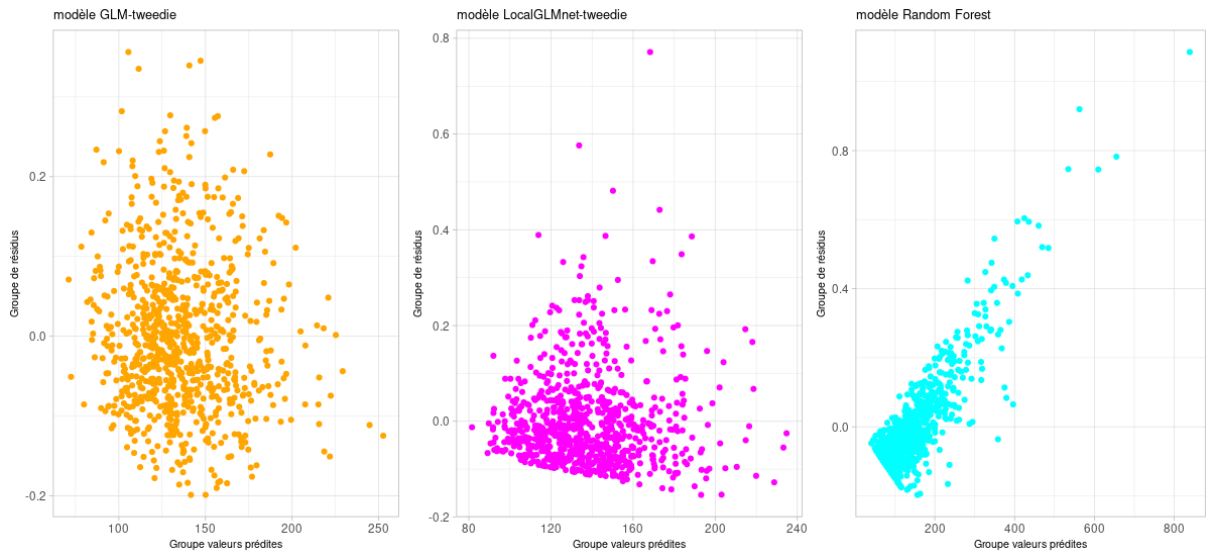


FIGURE 5.27 : *Nuages des résidus vs. prédictions pour chaque modèle de sévérité (agrégée) mis en place.*

5.4.2.2 Analyse comparative de la performance prédictive des modèles de sévérité

On fait deux constats frappant dès la première lecture du tableau 5.16 :

• premièrement, pour la modélisation de la sévérité agrégée du montant des sinistres (la prime pure), le modèle Random Forest est à la fois le plus prudent dans la globalité avec un ratio coût moyen prédit sur coût moyen observé égale à 1.045721 ;

• deuxièmement, le modèle Random Forest est le plus précis au sens de la RMSE ou MSE (mais le moins bon en termes de MAE), suivi par le modèle LocalGLMnet-Tweedie.

Une analyse plus approfondie du tableau 5.16 en comparaison avec les résultats obtenus dans le tableau 5.12 montre l'apport indéniables des données télématiques prédites dans la précision prédictive du calcul de la prime pure, quoique infime :

– On observe un gain relatif global d'environ 0.7% entre le modèle *GLM-coût-classique* du tableau 5.12 ne prenant pas en compte les variables télématiques et le modèle *GLM-Tweedie* du tableau 5.16 prenant en compte les valeurs prédites des variables télématiques.

– Cette valeur ajoutée des variables télématiques prédites dans la modélisation de la sévérité agrégée (prime pure) est d'autant plus visible lorsqu'on se restreint à des segments de marché spécifiques d'assurés. Par exemple sur le segment des assurés les plus jeunes (âgés entre 16 et 20 ans inclus), généralement réputés être les plus à risque, on s'aperçoit qu'en utilisant un modèle GLM-Tweedie avec variables télématiques prédites, on réduit de près de 40% la valeur de la RMSE par rapport à la situation où on omet les variables télématiques prédites et se sert juste des variables de risque classiques pour la tarification.

Modèles	moy. pred./moy. réelle	MSE	MAE	RMSE	$RMSE_{mean}$
GLM-Tweedie	0.8177954	1285130	223.1136	1133.636	8.532343
LocalGLMnet-Tweedie	0.8809933	1282644	229.391	1132.539	8.524087
Random Forest	1.045721	1276388	248.962	1129.774	8.503275

TABLE 5.16 : *Comparaison de la performance prédictive des modèles de sévérité (agrégée) mis en place.*

Sur la figure 5.28 on observe que les distributions des montants prédits de sinistres par les trois modèles sont assez proches. En observant de plus près, on remarque que les queues de distribution des montants prédits par le modèle *Random Forest* sont un peu plus épaisses que celles obtenues par les deux autres modèles : le modèle *Random Forest* prend mieux en compte les sinistres extrêmes dans son ajustement.

5.5 Interprétation des modèles de fréquence mis en place

Dans cette section, nous présentons les résultats de l'interprétation de deux des trois modèles de fréquence mis en place, à savoir le *GLM-poisson* et le *LocalGLMnet-poisson*. Les résultats de l'interprétation du modèle *Random Forest* étant quasiment similaires à ceux du modèle *LocalGLMnet-poisson*, nous avons choisir de ne pas les présenter dans ce mémoire, pour éviter de rallonger davantage le mémoire.

Pour l'interprétation des modèles *GLM-Poisson* et *LocalGLMnet-poisson* nous avons procédé suivant deux approches :

• Une interprétation basée sur le modèle (IBM) ;

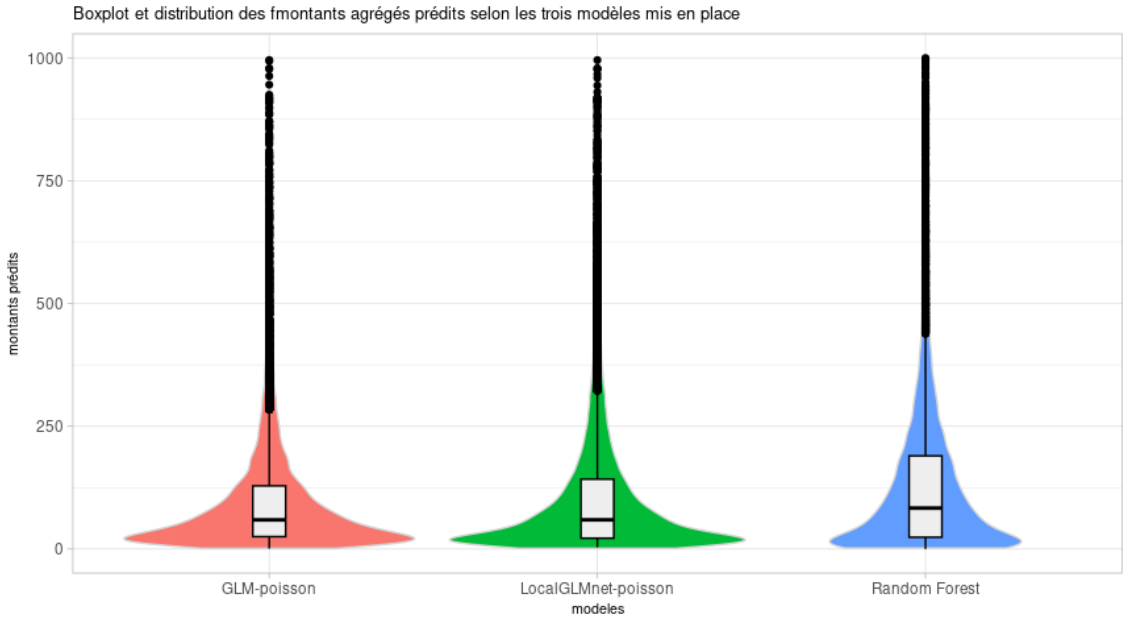


FIGURE 5.28 : *Boxplot et distributions des coûts prédits sur la base de test, selon les trois modèles mis en place.*

• Une interprétation agnostique au modèle, utilisant les outils d'interprétation présentés au chapitre 4.

Le fait d'interpréter ces modèles en utilisant ces deux approches d'interprétabilité, nous permettra de confronter les résultats de l'interprétation issus de chacune des deux approches et de s'assurer de la cohérence des interprétations obtenues à l'aide des outils d'interprétation présentés au chapitre 4.

Les modèles de sévérité pourraient également être interprété en suivant la même procédure. Nous ne présenterons pas les résultats de leur interprétation dans ce présent mémoire.

5.5.1 Interprétation du GLM fréquence

5.5.1.1 Interprétation basée sur le modèle (IBM)

Commençons par comprendre comment fonctionne notre modèle GLM fréquence. Comment le modèle aboutit-il aux prédictions ? Quelles caractéristiques ont-elles tendance à accroître ou à réduire la fréquence de sinistres chez les assurés ? De combien changerait la fréquence prédite d'un assuré si une ou plusieurs de ses caractéristiques venaient à être modifiée(s) ? Pour obtenir des réponses à ces différentes questions, il suffit de s'appuyer sur les sorties du modèle GLM fréquence récapitulés dans le tableau C.4 en annexe.

Tout d'abord, comme nous l'avons évoqué au chapitre 3, le modèle GLM est généralement simulable. Autrement dit, étant donné un assuré quelconque pour lequel nous prédisons la fréquence de sinistre à l'aide de notre modèle GLM, il est possible de reconstituer l'ensemble du processus ayant mené à la valeur prédite, et ce, dans un laps de temps raisonnable. Cela découle du fait que la valeur prédite par le GLM se décompose simplement comme une somme pondérée des caractéristiques de l'assuré, à laquelle on applique une fonction de lien (qui correspond à la fonction logarithme pour les différents modèles GLM mis en place dans ce mémoire).

Plus précisément, on a :

$$
\begin{aligned}
frequence \;=\;& \ln^{-1}\left[\hat{\beta}_{intercept} + \hat{\beta}_{TerritoryGzone_B}\,\mathbb{I}_{TerritoryGzone_B} + \hat{\beta}_{TerritoryGzone_C}\,\mathbb{I}_{TerritoryGzone_C}\right.\\
&+ \hat{\beta}_{Car.useCommute}\,\mathbb{I}_{Car.useCommute} + \hat{\beta}_{Car.useFarmer}\,\mathbb{I}_{Car.useFarmer}\\
&+ \hat{\beta}_{Car.usePrivate}\,\mathbb{I}_{Car.usePrivate}\\
&+ \hat{\beta}_{Car.ageG[0,2]}\,\mathbb{I}_{Car.ageG[0,2]} + \hat{\beta}_{Car.ageG[3,5]}\,\mathbb{I}_{Car.ageG[3,5]} + \hat{\beta}_{Car.ageG[6,9]}\,\mathbb{I}_{Car.ageG[6,9]}\\
&+ \hat{\beta}_{Car.ageG[10,+[}\,\mathbb{I}_{Car.ageG[10,+[} + \hat{\beta}_{Insured.ageG[16,20]}\,\mathbb{I}_{Insured.ageG[16,20]}\\
&+ \hat{\beta}_{Insured.ageG[21,30]}\,\mathbb{I}_{Insured.ageG[21,30]} + \hat{\beta}_{Insured.ageG[31,40]}\,\mathbb{I}_{Insured.ageG[31,40]}\\
&+ \hat{\beta}_{Insured.ageG[41,50]}\,\mathbb{I}_{Insured.ageG[41,50]} + \hat{\beta}_{Insured.ageG[51,60]}\,\mathbb{I}_{Insured.ageG[51,60]}\\
&+ \hat{\beta}_{Insured.ageG[61,70]}\,\mathbb{I}_{Insured.ageG[61,70]} + \hat{\beta}_{Insured.ageG[71,80]}\,\mathbb{I}_{Insured.ageG[71,80]}\\
&+ \hat{\beta}_{Insured.ageG[81,+[}\,\mathbb{I}_{Insured.ageG[81,+[} + \hat{\beta}_{Credit.score.G[0,750]}\,\mathbb{I}_{Credit.score.G[0,750]}\\
&+ \hat{\beta}_{Credit.score.G]750,800]}\,\mathbb{I}_{Credit.score.G]750,800]} + \hat{\beta}_{Credit.score.G]800,850]}\,\mathbb{I}_{Credit.score.G]800,850]}\\
&+ \hat{\beta}_{Credit.score.G]850,900]}\mathbb{I}_{Credit.score.G]850,900]}\\
&+ \hat{\beta}_{Annual.miles.driveG[0,6000]}\mathbb{I}_{Annual.miles.driveG[0,6000]}\\
&+ \hat{\beta}_{Annual.miles.driveG]6000,7500]}\,\mathbb{I}_{Annual.miles.driveG]6000,7500]}\\
&+ \hat{\beta}_{Annual.miles.driveG][7500,12500]}\,\mathbb{I}_{Annual.miles.driveG][7500,12500]}\\
&+ \hat{\beta}_{Total.miles.driven\_fit}\,X_{Total.miles.driven\_fit} + \hat{\beta}_{Annual.pct.driven\_fit}\,X_{Annual.pct.driven\_fit}\\
&+ \hat{\beta}_{Pct.drive.rush.am\_fit}\,X_{Pct.drive.rush.am\_fit} + \hat{\beta}_{Pct.drive.rush.pm\_fit}\,X_{Pct.drive.rush.pm\_fit}\\
&+ \hat{\beta}_{Left.turn.intensity08\_fit}\,X_{Left.turn.intensity08\_fit} + \hat{\beta}_{Brake.08miles\_fit}\,X_{Brake.08miles\_fit}\\
&+ \hat{\beta}_{Brake.09miles\_fit}\,X_{Brake.09miles\_fit} + \hat{\beta}_{Pct.drive.4hrs\_fit}\,X_{Pct.drive.4hrs\_fit}\\
&+ \hat{\beta}_{Pct.drive.2hrs\_fit}\,X_{Pct.drive.2hrs\_fit} + \hat{\beta}_{Accel.14miles\_fit}\,X_{Accel.14miles\_fit}\\
&+ \hat{\beta}_{Accel.11miles\_fit}\,X_{Accel.11miles\_fit} + \hat{\beta}_{Accel.06miles\_fit}\,X_{Accel.06miles\_fit}\\
&+ \left.\hat{\beta}_{Avgdays.week\_fit}\,X_{Avgdays.week\_fit}\right]
\end{aligned}
$$

où les coefficients $\hat{\beta}$ sont disponibles dans le tableau C.4 en annexe.

En guise d'illustration considérons un assuré *lambda* dans notre base de test, avec les caractéristiques suivantes :

Insured.age : **70 ans**	*Car.age* : **8 ans** *Credit.score* : **787**	*Annual.miles.drive* : **6213.71**
Car.use : **Private**	*TerritoryG* : **zone_B**	*Total.miles.driven_fit* : **0.11**
Annual.pct.driven_fit : **0.29**	*Pct.drive.rush.am_fit* : **0.15**	*Left.turn.intensity08_fit* : **0.72**
Pct.drive.rush.pm_fit : **0.22**	*Brake.08miles_fit* : **0.11**	*Brake.09miles_fit* : **0.01**
Pct.drive.4hrs_fit : **0.00**	*Pct.drive.2hrs_fit* : **0.34**	*Accel.14miles_fit* : **0.11**
Accel.11miles_fit : **0.21**	*Accel.06miles_fit* : **0.64**	*Avgdays.week_fit* : **0.99**

TABLE 5.17 : Caractéristiques de l'assuré *lambda*

$$
\begin{aligned}
frequence_{lambda} \;=\;& \exp\left[-5.27 + (-0.04) + (-0.15) + (-0.34) + 0.01 + (-0.20) + 0.77\right.\\
&+ 1.30 \times 0.11 + 0.73 \times 0.29 + (-0.09) \times 0.15 + 0.09 \times 0.22 + 0.36 \times 0.72\\
&+ 0.46 \times 0.11 + 1.00 \times 0.01 + 0.22 \times 0 + 0.64 \times 0.34 + 0.11 \times 0.11\\
&+ \left.0.10 \times 0.21 + (-0.02) \times 0.64 + 0.33 \times 0.99\right]\\
=\;& exp\left[-3.98\right]\\
=\;& 0.019
\end{aligned}
$$

Dans notre modèle GLM-poisson, étant donné que notre fonction de lien est le logarithme, son inverse qui correspond à la fonction exponentielle est croissante. Ainsi, pour cerner l'effet d'une caractéristique sur le niveau de risque des assurés, il suffit de se fier au signe du poids $\hat{\beta}$ de la caractéristique concernée dans le tableau C.4. Lorsque le poids est positif, cela voudrait dire que, la présence (ou

l'augmentation de la valeur) de cete caractéristique chez un assuré accroît relativement la fréquence de sinistre. Le signe positif pourrait alors nous permettre de mettre en évidence les segments d'assurés les plus à risque.

Inversement, une caractéristique avec un poids négatif stipule que la présence (ou l'augmentation de la valeur) de la caractéristique en question chez un assuré réduit relativement sa fréquence de sinistre. Un signe négatif pourrait nous permettre d'identifier les segments d'assurés, relativement, les moins à risque.

Cependant, il est rigoureux de se limiter à l'interprétation du signe des caractéristiques qui sont statistiquement significatives aux seuils habituels (10%, 5%, 1%), c'est à dire celles qui ont une p-valeur inférieure à 0.05 pour le test de nullité de leur poids.

Récapitulons quelques segments spécifiques d'assurés les plus à risque mis en évidence par notre modèle GLM :

– Véhicules à usage commercial ;

– Véhicules âgés entre [0, 2] ans ;

– Assurés âgés entre [16, 20] et [71, +] ans ;

– Assurés les moins solvables au sens du score de crédit ;

– Par ailleurs, on se rend compte qu'à mesure que la probabilité de rouler plus 4500 miles par an s'accroît chez les assurés, la fréquence de sinistre a également tendance à s'accroître ; de même, en ce qui concerne le nombre de freinage brusques : la fréquence de sinistre s'accroît à mesure que la probabilité d'effectuer plus de 5 freinages brusques d'intensité 09 mph/s durant l'année augmente.

Enfin, on peut aussi calculer la fréquence prédite pour le profil de risque moyen. Sur notre jeu de données de test, le profil de risque moyen possède les caractéristiques suivantes :

Insured.age : **51 ans**	*Car.age* : **5.78 ans** *Credit.score* : **801**	*Annual.miles.drive* : **9141**
Car.use : **Commute**	*TerritoryG* : **zone_B**	*Total.miles.driven_fit* : **0.41**
Annual.pct.driven_fit : **0.46**	*Pct.drive.rush.am_fit* : **0.66**	*Left.turn.intensity08_fit* : **0.39**
Pct.drive.rush.pm_fit : **0.68**	*Brake.08miles_fit* : **0.29**	*Brake.09miles_fit* : **0.13**
Pct.drive.4hrs_fit : **0.15**	*Pct.drive.2hrs_fit* : **0.70**	*Accel.14miles_fit* : **0.09**
Accel.11miles_fit : **0.25**	*Accel.06miles_fit* : **0.42**	*Avgdays.week_fit* : **0.86**

TABLE 5.18 : Caractéristiques du profil de risque moyen du jeu de données de test

$$
\begin{aligned}
freq_{profil\_moyen} &= \exp\left[-5.27 + (-0.04) + (-0.19) + (-0.34) + (-0.00) + (-0.58) + 0.67\right. \\
&+ 1.30 \times 0.41 + 0.73 \times 0.46 + (-0.09) \times 0.66 + 0.09 \times 0.68 + 0.36 \times 0.39 \\
&+ 0.46 \times 0.29 + 1.00 \times 0.13 + 0.22 \times 0.15 + 0.64 \times 0.70 + 0.11 \times 0.09 \\
&+ \left. 0.10 \times 0.25 + (-0.02) \times 0.42 + 0.33 \times 0.86\right] \\
&= exp\left[-3.6843\right] \\
&= 0.025
\end{aligned}
$$

Lorsqu'on compare la fréquence prédite de l'assuré *lambda* (dont les caractéristiques sont décrites dans le tableau 5.17) à la fréquence prédite du profil de risque moyen (dont les caractéristiques sont présentées dans le tableau 5.18), on s'aperçoit que la fréquence prédite du profil de risque moyen est plus élevée que celle de l'assuré *lambda*.

L'écart entre leur fréquence prédite de sinistre peut clairement s'expliquer en comparant leur profil de risque. On constate que les deux assurés ont des volumes et comportements de conduite assez disparates :

– alors que le profil de risque moyen a une probabilité de 41% de rouler plus de 4500 miles durant l'année, l'assuré *lambda* quant à lui n'a qu'une probabilité de 11% de rouler plus de 4500 miles durant l'année.

– De plus, lorsque le profil de risque moyen a une probabilité de près de 50% de rouler plus d'un jour sur deux durant l'année, l'assuré *lambda* a une probabilité inférieure à un tier de rouler plus d'un jour sur deux durant l'année.

– En ce qui concerne le comportement de conduite, lorsque l'assuré *lambda* a une probabilité de 1% d'excéder plus de 5 freinages brusques d'intensité 09 mph/s durant l'année, le profil de risque moyen a une probabilité de 13% d'excéder plus de 5 freinages brusques d'intensité 09 mph/s durant l'année.

En somme, l'assuré *lambda* aurait une fréquence de sinistre moins élevée que celle du profil de risque moyen parce qu'il a une propension à rouler beaucoup moins importante que celle du profil de risque moyen. En outre son comportement de conduite semble plus prudent que celui du profil de risque moyen.

5.5.1.2 Interprétation post hoc du modèle GLM-féquence : vérification de la cohérence des résultats avec l'approche IBM et l'approche post hoc

Avant de passer à l'interprétation du modèle LocalGLMnet, nous souhaitons implémenter quelques outils d'interprétation post hoc présentés au chapitre 4, afin de s'assurer de la cohérence des interprétations obtenues par les deux approches (post hoc et IBM) du modèle GLM-Poisson mis en oeuvre.

Nous commençons par une interprétation globale du modèle GLM-Poisson, puis nous essayons d'expliquer une prédiction ponctuelle obtenue par le modèle GLM-Poisson, notamment celle de l'assuré *lambda* (présenté dans le tableau 5.17).

(A) Analyse globale du modèle GLM-Poisson

L'interprétation au niveau global s'intéresse aux relations générales apprises par un modèle.

□ Importance globale des caractéristiques : t-statistic de Student, MR, SFIMP

Le score d'importance d'une variable au niveau d'un modèle tente de capturer la contribution de la variable, dans un ensemble de données, à la formation des prédictions.

On distingue un grand nombre de méthodes pour la mesure de l'importance des variables. Dans cette sous-partie notre objectif est de vérifier est-ce que la hiérarchisation de l'importance des variables basée sur la t-statistic présentées au chapitre 2 dans le cadre de la présentation du modèle GLM, est similaire à la hiérarchisation obtenue par les autres outils de mesure d'importance de variables agnostiques au modèle présentés au chapitre 4, à savoir la MR (basée sur la permutation des caractéristiques) et SFIMP (basée sur les valeurs de Shapley de la théorie des jeux).

Les résultats obtenus par ces différentes méthodes sont présentés par les trois diagrammes de la figure 5.29 : on observe de légères différences sur les diagrammes.

Tout d'abord, rappelons que lors de sa mise en place, le modèle GLM re-code automatiquement les caractéristiques catégorielles, en plusieurs nouvelles variables indicatrices par rapport à une catégorie de référence de la variable. C'est la raison pour laquelle on peut observer que sur le diagramme d'importance des variables basée sur la t-statistique de Student, certaines variables catégorielle apparaissent suivant les catégories. Cependant, cela n'empêche aucunement la comparaison des résultats obtenus avec ceux des deux autres méthodes : globalement, on observe une convergence de la hiérarchisation de l'importance des variables suivant les trois approches.

D'après les trois approches, les caractéristiques jugées les plus importantes dans la prédiction de la fréquence de sinistre par le GLM-Poisson sont : la distance totale parcourue durant l'année (*Total.miles.driven_fit*) ; le niveau de score de crédit de l'assuré (*Credit.scoreG*) ; la fréquence de conduite durant les jours l'année (*Annual.pct.driven_fit*) et le nombre annuel de freinages brusques d'intensité 09 mph/s (*Brake.09miles_fit*).

Les variables les moins importantes dans la prédiction de la fréquence de sinistre par le GLM-Poisson sont : le nombre annuel d'accélération d'intensité 06 mph/s, 11 mph/s, 14 mph/s (*Accel.06miles_fit*, *Accel.11miles_fit*, *Accel.14miles_fit*) et le territoire de conduite de l'assuré (*TerritoryG*) ;

Avec la méthode MR (PFI), on constate que l'ensemble des caractéristiques ont à peu près le même niveau d'importance dans le modèle : cela n'est pas surprenant, car aucune caractéristique prise individuellement, n'est fortement explicative de la variable cible (fréquence de sinistre). Raison pour laquelle leur permutation individuelle n'impacte que dans de très faibles proportions la valeur prédite du nombre de sinistre.

Pour la mise en oeuvre de ces différentes méthodes, nous avons utilisé le package *DALEX* de *R* pour le calcul de l'importance basée sur les permutations (MR) et le package *iml* de *R*, pour le calcul de l'importance basée sur les valeurs de Shapley.

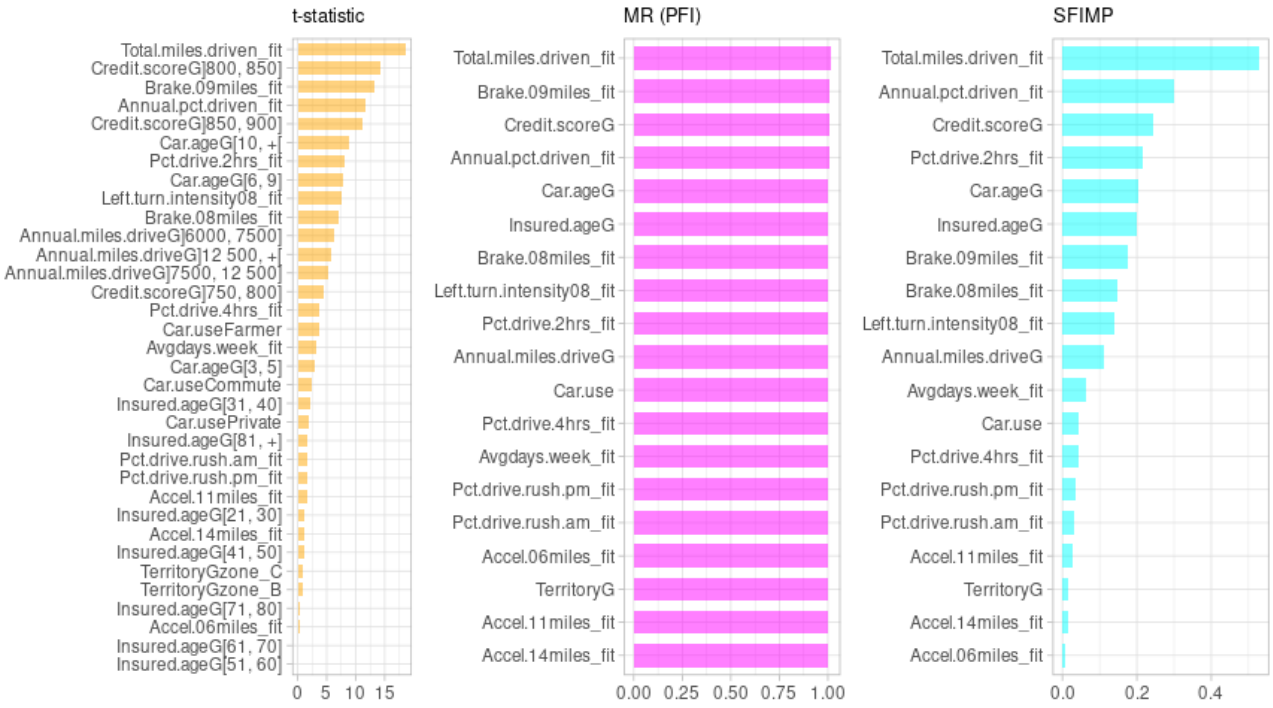


FIGURE 5.29 : *Importance des variables dans le modèle GLM fréquence, basée sur la t-statistique, sur la permutation des caractéristiques (MR), et sur les valeurs de Shapley (SFIMP).*

□ Analyse de l'effet global des caractéristiques : PDP, ALE-plot

• PDP

Les graphiques de dépendances partielles (en anglais Partial Dependences Plots) sont des outils particulièrement utiles à la compréhension relations apprises par un modèle. Le PDP trace le changement de la valeur moyenne prédite lorsque la caractéristique spécifiée varie marginalement.

Dans le cas où les caractéristiques ne sont pas fortement corrélées, le diagramme de PDP traduit l'effet marginal moyen de la caractéristique concernée sur la cible. Dans notre contexte, nous nous

attendons à observer une certaine cohérence entre les poids du GLM et les diagrammes PDP. Lorsqu'on compare les diagrammes de dépendances partielles représentées sur la figure 5.30 et les diagrammes des contributions des différentes caractéristiques dans le GLM (confère figure 5.31), on observe que les courbes obtenues sont quasiment identiques à un changement d'échelle près : ce qui est cohérent avec la théorie.

Les graphiques de PDP ont été implémentés à l'aide du package *DALEX* sous *R*.

FIGURE 5.30 : *Graphiques de dépendance partielles (PDP) des 19 caractéristiques pour le GLM-fréquence.*

- **ALE-Plot**

Les graphiques PDP présentent des limites dans la mesure de l'effet des différentes caractéristiques lorsque celles-çi sont fortement corrélées entre-elles.

C'est la raison pour laquelle ont été introduit les graphiques ALE qui permettent de mesurer l'effet des caractéristiques sur la variable cible en tenant compte des potentielles interactions existantes entre-elles.

Les tracés ALE sont centrés sur zéro. Cela rend leur interprétation agréable, car la valeur à chaque point de la courbe ALE est la différence avec la prédiction moyenne.

Dans notre contexte, étant donné que les caractéristiques sont mutuellement faiblement corrélées, et compte tenu du fait que le modèle GLM est structurellement additif, les courbes ALE et PDP seront quasiment identiques, à un changement d'échelle près. Les résultats de la mise en oeuvre des courbes

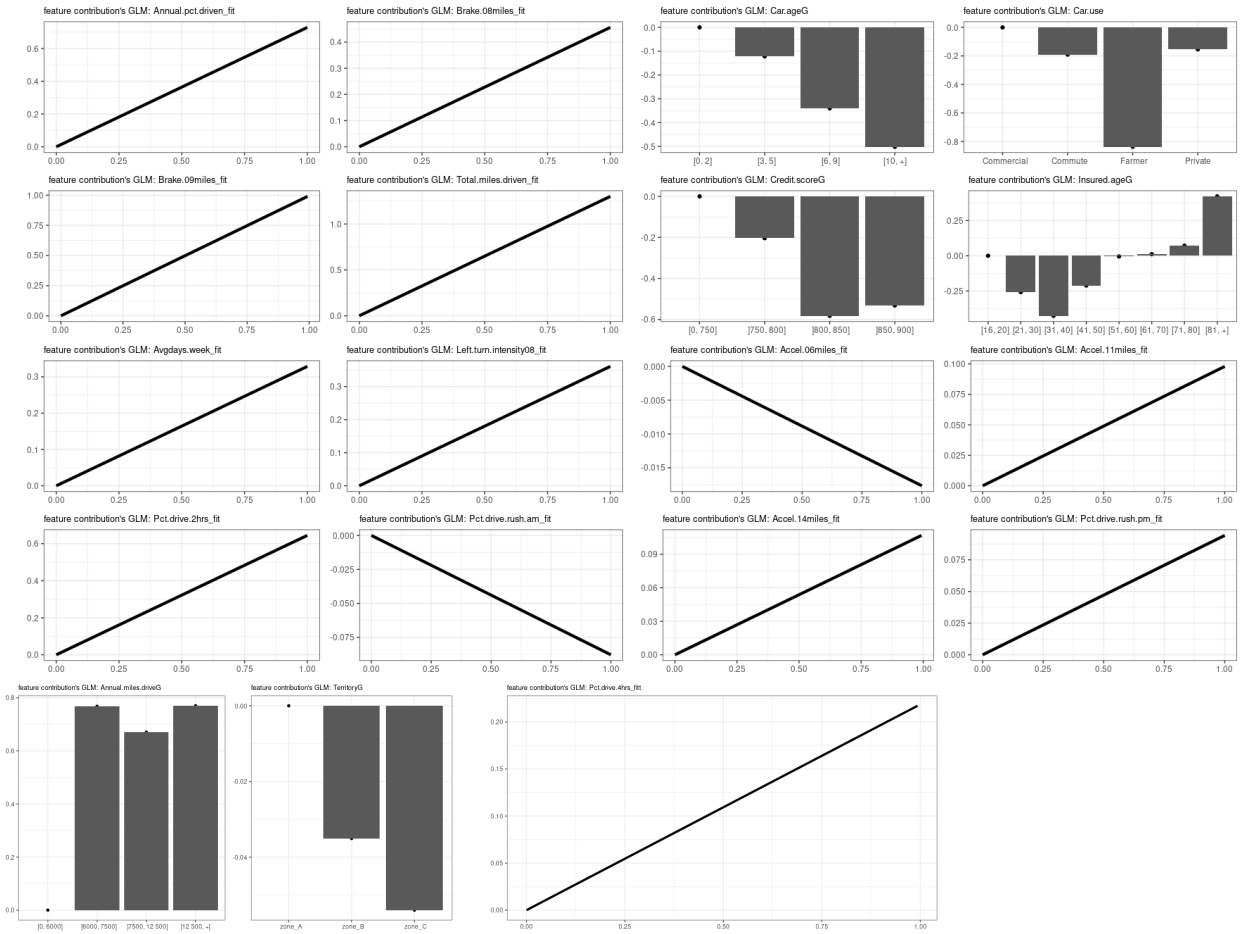


FIGURE 5.31 : *Graphiques des contributions individuelles ($\hat{\beta}_j\, x_j$) des 19 caractéristiques dans le GLM-fréquence.*

ALE sont représentés sur la figure 5.32 : on observe bien les faits attendus, ce qui est cohérent avec la théorie.

□ Analyse de l'interaction globale entre les caractéristiques : Indice de Sobol et H-statistique de Friedman

• H-statistique de Friedman

La H-statistique de Friedman est présenté dans le chapitre 4. Il s'agit d'un outil qui permet de quantifier d'une part, le niveau d'interaction entre les paires de caractéristiques dans un modèle, et d'autre part, il permet également de mesurer l'interaction totale d'une caractéristique dans un modèle, c'est-à-dire, dans quelle mesure la caractéristique interagit avec l'ensemble des autres caractéristiques du modèle.

Dans notre cas, nous mesurons les interactions totales de chaque caractéristique dans le modèle. Les résultats obtenus à l'aide du package *iml* de R sont donnés sur la figure 5.33. Comme attendu, les interactions entre les différentes caractéristique dans le modèle GLM sont quasiment nulles : ce qui est cohérent avec la théorie, car de par spécification du modèle linéaire généralisée, on contraint explicitement les covariables à ne pas interagir les unes avec les autres.

FIGURE 5.32 : *Graphiques des effets locaux accumulés (ALE) des 19 caractéristiques pour le GLM-fréquence.*

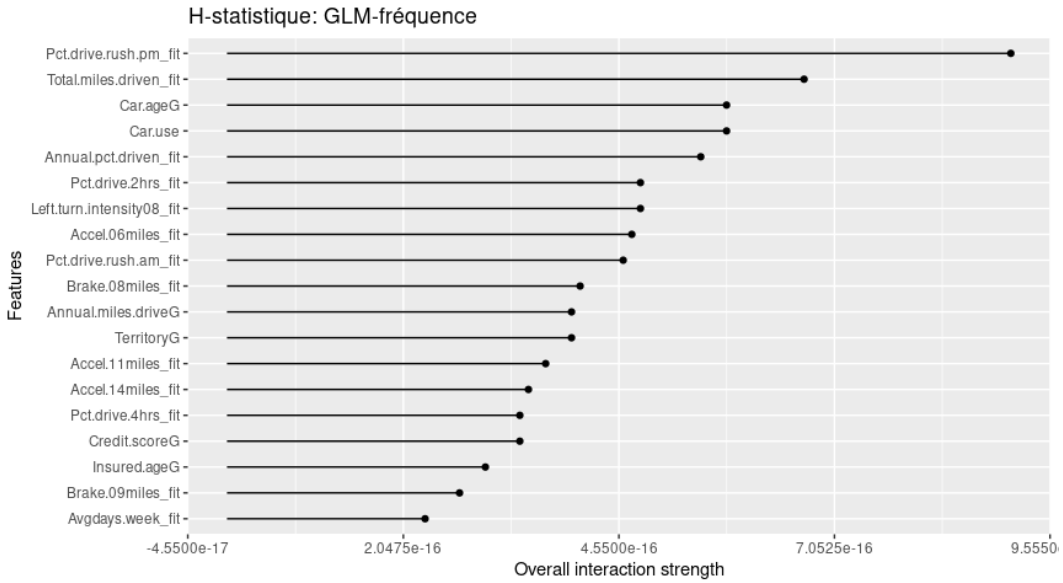


FIGURE 5.33 : *Résultats des H-statistique de l'interaction totale des 19 caractéristiques dans le GLM fréquence.*

- **Indice de Sobol**

Une autre approche pour mesurer l'interaction entre les caractéristiques est basée sur le calcul des indices de Sobol. La présentation des bases théoriques de cette méthode a été abordé dans le chapitre

4. Tout comme avec la H-statistique, les indices de Sobol peuvent permettre de quantifier d'une part, le niveau d'interaction entre les couples de caractéristiques dans un modèle, et d'autre part, mesurer l'interaction totale d'une caractéristique dans un modèle. Ici, nous nous limitons au calcul des indices de sensibilité de premier ordre.

Les résultats obtenus à l'aide du package *sensitivity* de R sont donnés sur la figure 5.34. Les estimations des indices de Sobol ont été obtenus par la méthode d'estimation de Sobol, version Saltelli 2002, présentée au chapitre 4.

Sur la figure 5.34, pour chaque caractéristique, la barre de couleur gris représente la valeur de l'indice de sensibilité de premier ordre de la variable. Elle exprime la proportion de la variance du modèle qui est expliquée par la caractéristique (hormis son interaction avec les autres caractéristique).

Sur la première barre de la figure, on retrouve la valeur totale des effets principaux de l'ensemble des caractéristique (en rouge), elle vaut 1.03 : autrement dit, la totalité de la variance du modèle est expliquée uniquement par les effets principaux (ou effets de premier ordre). Ce qui stipule que l'ensemble des effets d'interactions entre les caractéristiques sont négligeables, voir inexistantes dans notre modèle. Ce dernier résultat est tout à fait cohérent avec la théorie, et rejoint les résultats obtenus précédemment avec la méthode basée sur la H-statistique de Friedman, car de par sa spécification, le modèle GLM, capte uniquement les effets principaux des variables.

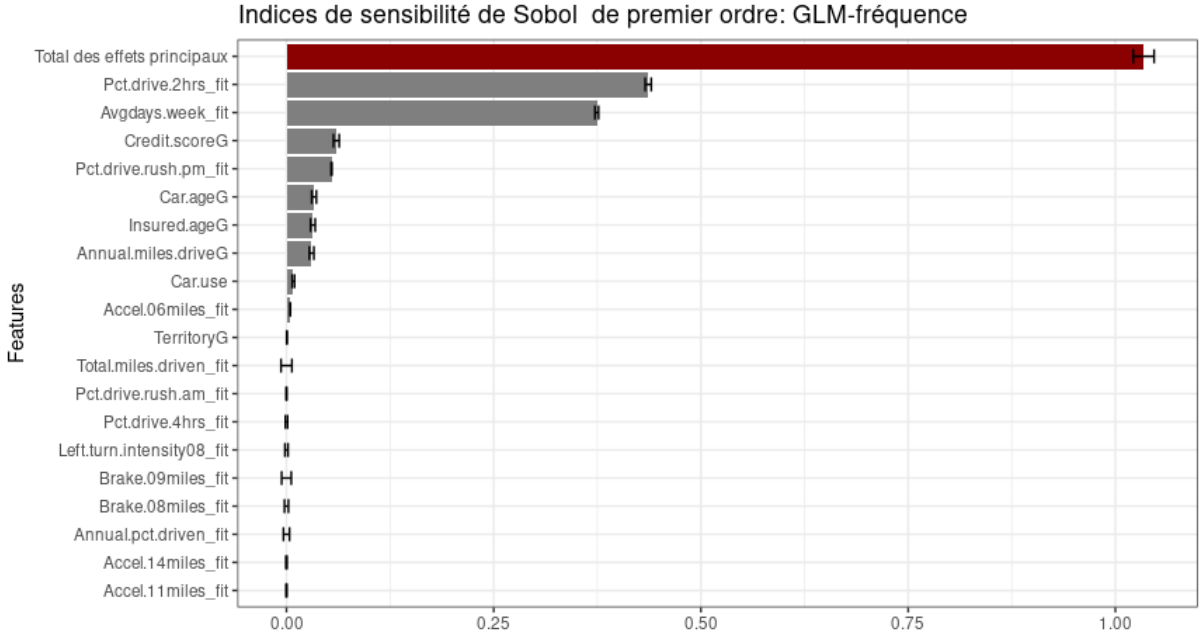


FIGURE 5.34 : *Résultats indices de sensibilité global et de premier ordre des 19 caractéristiques dans le GLM fréquence.*

(B) Analyse locale du modèle

Les explicateurs locaux permettent la deuxième manière de générer des explications pour notre modèles. Le principe est le suivant : étant donné une prédiction, l'explicateur local fournit des informations relatives aux processus de prédiction qui ne sont valables que pour l'instance particulière concernée et ne peuvent pas être généralisées à l'ensemble du modèle.

Dans notre contexte, nous nous focaliserons sur l'explication de la prédiction de l'instance *lambda* de la base de test, dont les caractéristiques ont été récapitulées plus haut dans le tableau 5.17. A cet effet, nous utiliserons les outils LIME et SHAP présentés dans le chapitre 4.

□ LIME

Rappelons que l'approche LIME (Local Interpretable Model-agnostic Explanations) est une méthode d'interprétabilité agnostique au modèle, généralement utilisée pour expliquer les prédictions individuelles des algorithmes complexes.

Dans ce paragraphe, nous souhaitons interpréter la prédiction de l'instance *lambda* faite par notre modèle GLM, à l'aide de la méthode LIME. Les résultats obtenus sont présentés sur la figure 5.35.

Il est important de remarquer que les valeurs "Actual Prediction" et "LocalModel prediction" inscrite au dessus de la figure 5.35 correspondent respectivement à la valeur prédite de l'instance d'intérêt par le vrai modèle (ici, notre GLM-fréquence) et la prédiction de l'instance issue du modèle de substitution locale.

La figure 5.35 montre bien que l'assuré *lambda* a une faible fréquence prédite de sinistre parce qu'il a une faible probabilité de rouler plus 4500 miles durant l'année ($Total.miles.driven\_fit$=0.11). De plus, son véhicule est relativement âgé, ce qui expliquerait le fait qu'il n'utilise pas fréquemment son véhicule, et donc, il a moins d'occasion d'être victime d'accidents.

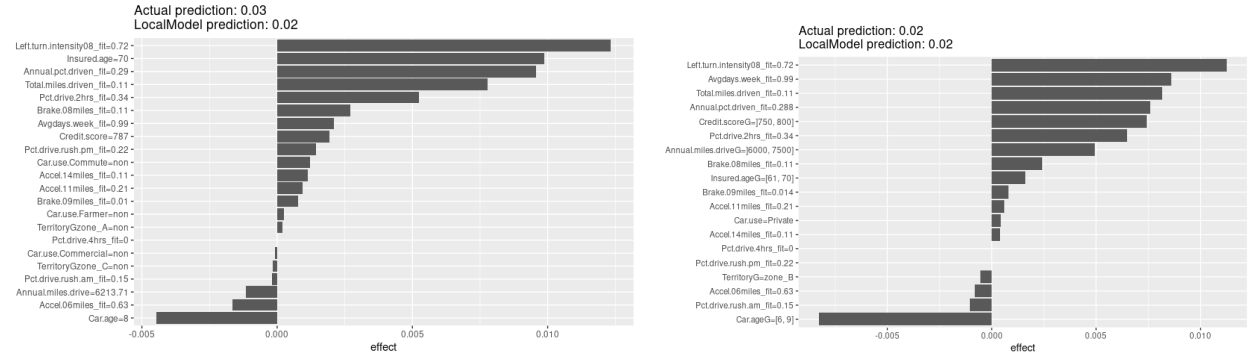


FIGURE 5.35 : *Résultats LIME pour notre assuré lambda dans le modèle GLM-fréquence : résultats avec package DALEX (à gauche), et avec package iml (à droite).*

□ SHAP

Passons à présent à l'interprétation de la prédiction de l'instance lambda par la méthode SHAP. Les résultats obtenus sont donnés sur la figure 5.36. On retrouve que : la faible probabilité de l'assuré *lambda* de rouler plus de 4500 miles durant l'année, contribue principalement à réduire sa fréquence prédite de sinistre durant l'année. Par ailleurs, son comportement de conduite sont également des facteurs qui amoindrissent sa fréquence prédite de sinistre : il a de faibles probabilités d'effectuer plus de 5 freinages brusques d'intensité 09 mph/s ou 08 mph/s durant l'année.

Cette similitude entre les interprétations basées sur ces différentes approches, confirme la cohérence entre l'interprétation basée sur le modèle et de celle agnostique au modèle pour la prédiction de l'assuré *lambda*.

En définitive, les interprétations globales et locales de notre GLM fréquence par les méthodes *post hoc* agnostiques au modèle présentées au chapitre 4 sont cohérentes avec la théorie et conformes aux interprétations intrinsèques au modèle. Ceci constitue un argument en faveur de la pertinence et de la légitimité de ces outils d'interprétation *post hoc* indépendants du modèle présentés au chapitre 4. Ils seront dorénavant utilisés sans craintes dans la suite de ce mémoire pour l'interprétation d'un modèle relativement plus complexe.

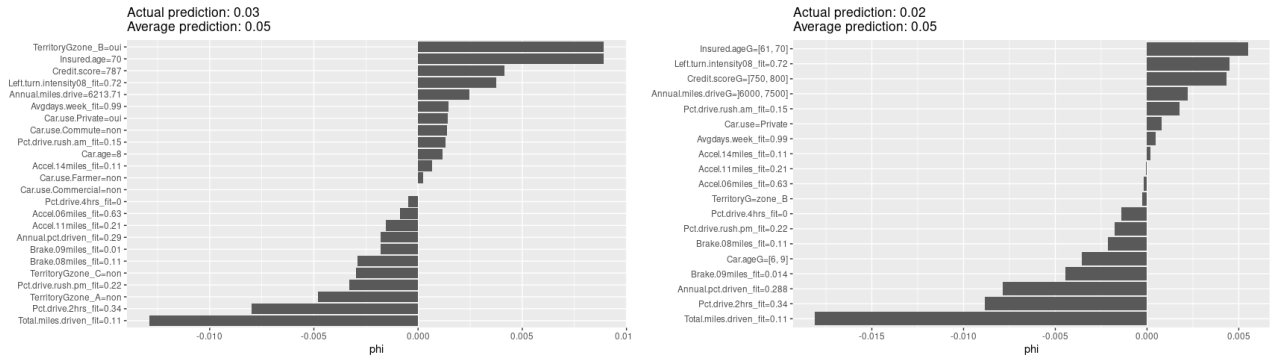


FIGURE 5.36 : *Résultats SHAP pour notre assuré lambda dans le modèle GLM-fréquence : résultats avec package DALEX (à gauche), et avec package iml (à droite).*

5.5.2 Interprétation du LocalGLMnet fréquence

5.5.2.1 Interprétation basée sur le modèle (IBM)

Comme nous l'avons indiqué plus haut, les modèles LocalGLMnet sont des modèles hybrides du point de vu de l'interprétabilité. Il combine à la fois une structure linéaire inspirée des modèles linéaires généralisés et une structure complexe basée sur des réseaux de neurones de type *Feed Forward Network*. Ils ont été présentés dans le chapitre 2.

□ Importance des variables basée sur les poids de régression estimés

Une fois le modèle ajusté, l'une des premières informations qu'on souhaite savoir est l'importance des variables, afin de comprendre quelles sont les variables les plus utilisées par le modèle, et celles qui influent le plus sur les valeurs des prédictions.

Pour ce faire, la mesure de l'importance des variables spécifique aux LocalGLMnet définie dans le chapitre 2 a été mis en oeuvre. Les résultats sont donnés sur la figure 5.37.

On observe un bouleversement notable de l'ordre d'importance des variables par rapport aux résultats obtenus précédemment dans le cadre du GLM fréquence (cf. figure 5.29) :

– Les variables *Total.miles.driven_fit*, *Annual.pct.driven_fit*, *Car.age* qui étaient situés tout en tête du classement des variables les importantes dans le GLM-fréquence se situent désormais parmis les variables les moins importantes dans le LocalGLMnet.

– *A contrario*, les variables *Avgday.week_fit* et *Accel11miles_fit* qui étaient parmi les variables les moins importantes dans le modèle GLM se trouvent classées parmi les caractéristiques les plus importantes dans le LocalGLMnet.

Ce bouleversement de l'ordre d'importance des variables pourrait provenir du fait que les modèles n'utilisent pas nécessairement les variables de la même manière, et donc leur attribue des niveaux d'importance différents.

□ Contribution des caractéristiques ($\hat{\beta}_j(x)x_j$)

La figure 5.38 nous donne la contribution des différentes variables $\hat{\beta}_j(x)x_j$ dans les prédictions de notre modèle LocalGLMnet-Poisson pour 10000 assurés prélevés aléatoirement dans notre base de test.

La courbe de couleur bleu sur ces différents graphiques correspond à l'ajustement spline décrite par la contribution de la caractéristique concernée. Les droites en pointillé de couleur orange correspondent

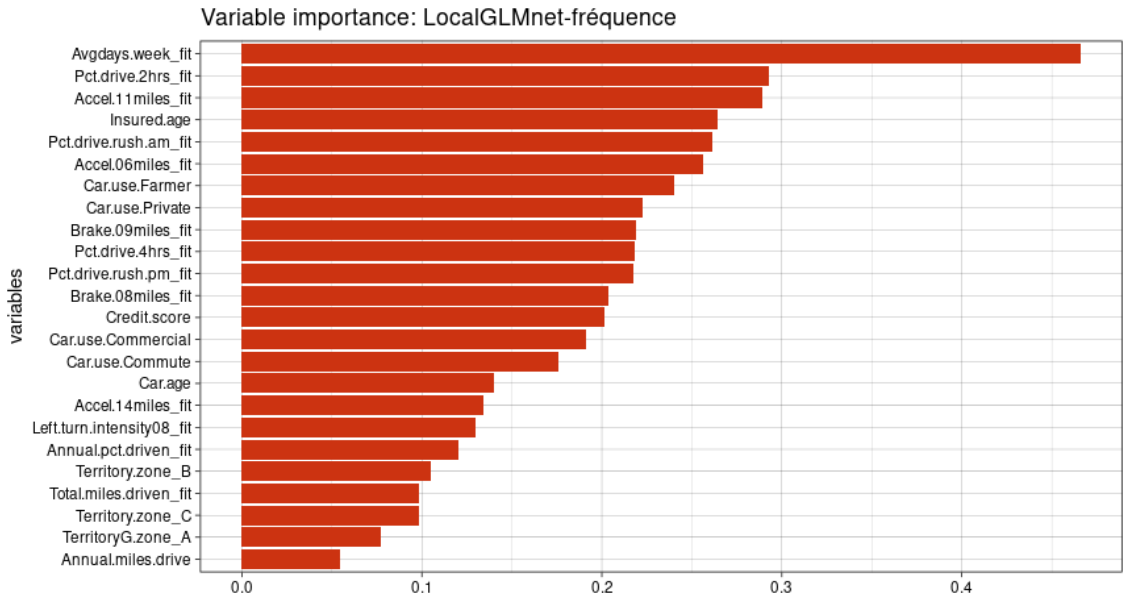


FIGURE 5.37 : *Importance des variables dans le modèle LocalGLMnet, suivant l'approche basée sur les poids de régression estimés présentée au chapitre 2. Il s'agit d'une approche analogue à celle des t-statistiques dans le cas des GLM.*

aux droites d'équations $y = 0.25$ et $y = -0.25$. On observe que les variables contribuent inéquitablement dans les prédictions :

– Les variables *Car.age*, *Avgday.week_fit*, *Credit.score*, *Car.use.Farmer* et *Annual.miles.drive* figurent parmi les variables ayant les plus grandes contributions ;

– Les variables *Annual.pct.driven_fit*, *Accel.11miles_fit*, *Accel.11miles_fit*, *Accel.11miles_fit* ont des contributions quasiment nulles ;

– Les variables *Avgday.week_fit* et *Annual.miles.drive* ont une contribution en moyenne croissante sur la fréquence sinistre, tandis que les variables *Car.age* et *Car.use.Farmer* ont une contribution en moyenne, décroissante sur la fréquence de sinistre ;

– La variable *Credit.score* a une contribution quadratique sur la fréquence de sinistre ;

– La contribution de la variable *Insured.age* est plus difficile à interpréter.

Lorsqu'on compare ces différents résultats à ceux obtenus dans le cadre du modèle GLM-Poisson représentés sur la figure 5.31, on constate que le sens de monotonie des contributions sont globalement identiques pour l'ensemble des variables.

Cette stabilité tend à faire confiance aux résultats obtenus. Rappelons une fois de plus qu'il s'agit ici d'une interprétation globale de la contribution des variables. Au niveau d'une instance particulière, il se peut que certaines variables jugées peu contributives à ce niveau, soient finalement essentielles. Ceci pourra être observé plus bas grâce aux méthodes d'interprétation locale telles que LIME et SHAP.

□ Analyse de l'interaction entre les caractéristiques ($\partial_{x_k}\hat{\beta}_j(x)$)

Un des plus gros avantages des modèles LocalGLMnet est que en plus de leur architecture linéaire facile à interpréter, comme nous l'avons vu précédemment avec la contribution des variables, ils tiennent compte de l'interaction entre les différentes caractéristiques dans l'ajustement des coefficients de régression.

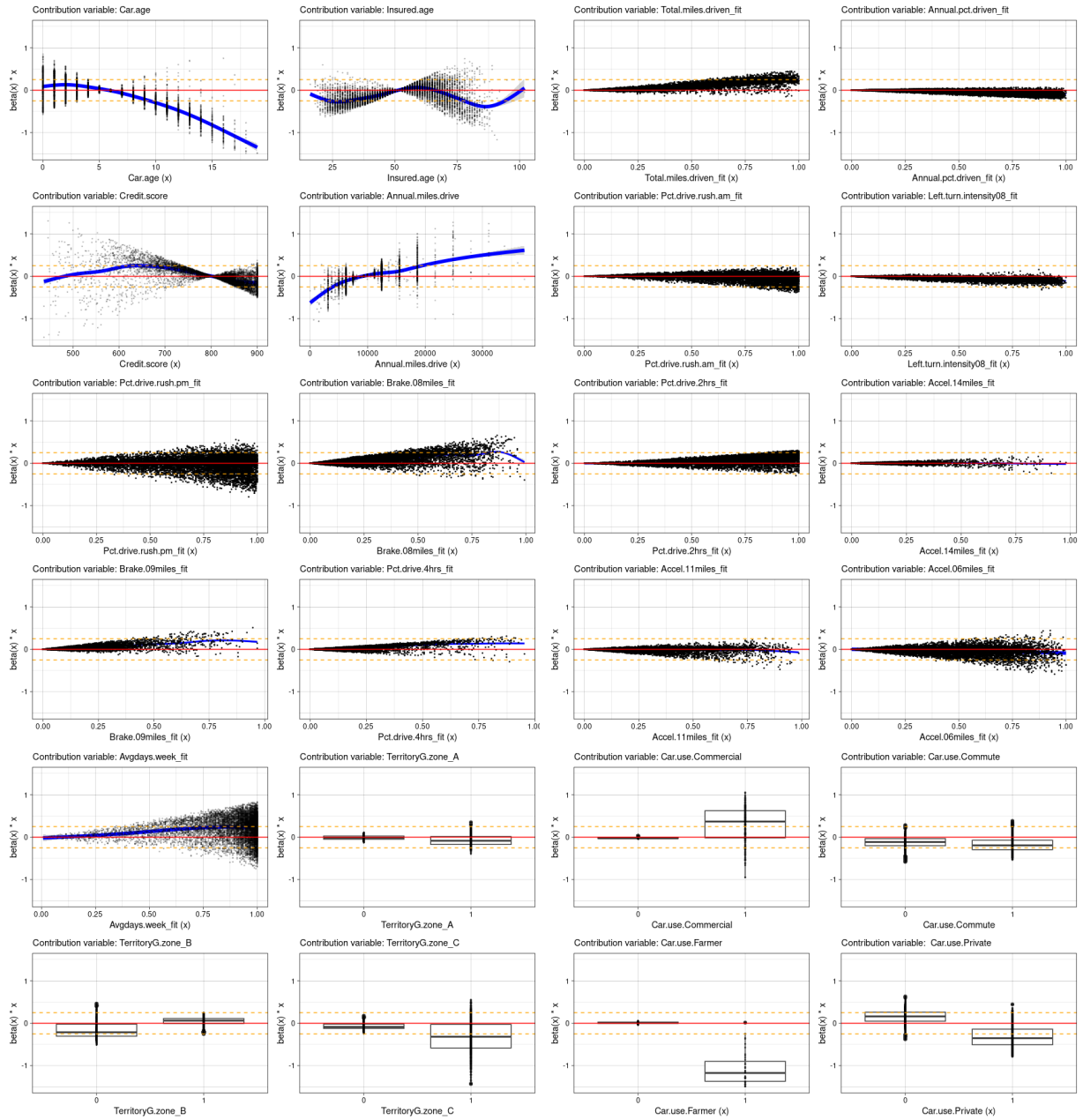


FIGURE 5.38 : *Nuages des contributions* $\hat{\beta}_j(x^{(i)}) \cdot x_j^{(i)}$, *des* j *caractéristiques* ($1 \leq j \leq 24$) *dans notre LocalGLMnet-fréquence, pour 10000 instances* $x^{(i)}$ *prélevées aléatoirement dans la base de test.*

Pour analyser ces interactions, nous suivons tout simplement la démarche présentée au chapitre 2 lors de la présentation des modèles LocalGLMnet : il s'agit tout simplement d'étudier la monotonie des gradients $(\partial_{x_k} \hat{\beta}_j(x))$ des coefficients des caractéristiques continues ($1 \leq k \leq 17$ et $1 \leq j \leq 17$).

La figure C.18 nous montre l'ajustement spline des gradients des caractéristiques continues dans notre modèle.

• Premièrement, on observe que :

– Les variables *Annual.miles.drive*, *Annual.pct.driven_fit*, *Pct.drive.rush.am_fit*, *Pct.drive.rush.pm_fit*, *Accel14miles_fit*, *Accel11miles_fit*, *Accel06miles_fit*, *Avgdays.week_fit*, *Pct.drive.4hrs_fit*, *Brake.08miles_fit* et *Left.turn.intensity08_fit* ont une contribution assez linéaire sur la fréquence de sinistre puisque leur terme $\partial_{x_j} \hat{\beta}_j(x) \approx 0$. Nous pouvons d'ailleurs

le constater sur les ajustements splines de leur contribution, représentés en couleur bleu, sur la figure 5.38.

– Les variables *Brake.09miles_fit*, *Pct.drive.2hrs_fit*, *Car.age*, *Credit.score* quant à elles ont une contribution approximativement quadratique sur la fréquence de sinistre, puisque leur terme $\partial_{x_j}\hat{\beta}_j(x) \approx constant \neq 0$.

– En ce qui concerne la variable *Total.miles.driven_fit*, son terme $\partial_{x_j}\hat{\beta}_j(x)$ est linéaire en x_j, ce qui signifie tout simplement qu'elle aurait une contribution cubique sur la fréquence de sinistre.

– Enfin, nous constatons une effet complexe de la variable *Insured.age* sur la fréquence de sinistre, qui a une expression quadratique de son gradient $\partial_{x_j}\hat{\beta}_j(x)$, suivant la composante x_j.

• Deuxièmement, nous nous concentrons sur l'analyse des interactions hétérogènes entre les différentes caractéristiques continues. Ceci nécessite l'étude des courbes des fonctions $\partial_{x_k}\hat{\beta}_j(x)$ pour les $k \neq j$ (confère figure C.18 en annexe). Les interactions les plus significatives peuvent être clairement observées :

– Entre *Insured.age* et *Left.turn.intensity08_fit*, et aussi entre *Insured.age* et *Brake.09miles_fit*.

La première interaction indique qu'une probabilité plus élevée d'effectuer plus de 30 virages à gauche par 1000 miles d'intensité 08 mph/s lorsqu'on est un conducteur "jeune" entraîne un accroissement de la prédiction de la fréquence de sinistre plus élevée que chez un assuré plus âgé. Ce résultat est intuitivement logique, car la réussite d'un virage à gauche dépend de plusieurs facteurs dont principalement l'expérience de conduite et la prudence au volant. Or, il est bien établi que les jeunes sont généralement moins expérimentés et moins prudents que les plus âgés, par conséquent les jeunes sont relativement plus enclin à échouer leur virage à gauche, donc provoquer relativement plus de sinistres.

La seconde paire d'interaction nous révèle qu'une probabilité plus élevée d'effectuer plus de 5 freinages brusques par 1000 miles d'intensité 09 mph/s pour un conducteur "âgé", entraîne un accroissement de la prédiction de la fréquence de sinistre plus élevée que chez un assuré moins âgé.

– Entre *Car.age* et *Total.miles.driven_fit*.

Cette interaction indique qu'une probabilité plus élevée de conduire plus de 4500 miles durant l'année avec un véhicule neuf entraîne un accroissement de la prédiction de la fréquence de sinistre relativement plus élevée. Cela peut provenir du fait qu'avec les véhicules neuf on a généralement des grandes capacités de vitesse et d'accélération, ce qui peut parfois favoriser plus de sinistres.

– Entre le *Credit.score* et *Total.miles.driven_fit*, entre le *Credit.score* et *Left.turn.intensity08_fit* et entre le *Credit.score* et *Insured.age*.

Dans le premier cas, l'interaction révèle qu'un accroissement de la probabilité de rouler plus de 4500 miles durant l'année pour un assuré ayant un score de crédit plus faible conduit à un accroissement de la prédiction de la fréquence de sinistre, relativement plus élevée que chez un assuré ayant un score de crédit élevé. Ce résultat est intuitivement logique, puisque les personnes ayant des scores de crédit plus faibles peuvent être plus à risque en matière de conduite, car elles sont plus susceptibles de souffrir de stress financier et de mauvaises habitudes de conduite.

Dans le deuxième cas, l'interaction révèle qu'un accroissement de la probabilité d'effectuer plus de 30 virages à gauche par 1000 miles d'intensité 08 mph/s pour un conducteur assuré ayant un faible score de crédit mène à un accroissement de la prédiction de la fréquence de sinistre relativement plus élevée que chez un assuré possédant un score de crédit plus élevé.

Dans le troisième cas, l'interaction révèle qu'un accroissement de l'âge pour un conducteur ayant un score de crédit faible conduit à un accroissement de la prédiction de la fréquence de sinistre relativement moins élevée que chez un assuré ayant un score de crédit élevé.

– Entre *Total.miles.driven_fit* et *Annual.pct.driven_fit*.

Cette interaction révèle qu'un accroissement de la probabilité de conduire plus d'un jour sur deux durant l'année pour un assuré ayant une faible probabilité de conduire plus de 4500 miles durant l'année, entraîne un accroissement de la prédiction de la fréquence de sinistre relativement moins élevée que chez un assuré ayant une forte probabilité de conduire plus de 4500 miles durant l'année. Ce résultat est intuitivement logique, car plus on roule, plus on est enclin à avoir des occasions d'accidents.

Par ailleurs, un accroissement de la probabilité de conduire plus de 4500 miles durant l'année pour un assuré ayant une faible probabilité de conduire plus d'un jour sur deux durant l'année, entraîne un accroissement plus important de la prédiction de la fréquence de sinistre que chez un assuré ayant une forte probabilité de rouler plus d'un jour sur deux durant l'année. Ce résultat pourrait s'expliquer par le fait que : un conducteur qui accumule un volume élevé de conduite sur peu de jour durant l'année, toutes chose égales par ailleurs, effectue des trajets plus long et est alors plus enclin à faire des accidents qu'un conducteur lissant son haut volume de conduite sur plusieurs jour de l'année, car étant spécialisé dans des courts trajets, et maîtrisant généralement leur trajet.

• En ce qui concerne les effets d'interactions linéaires (non constants), nous nous limitons ici à l'analyse des graphes d'interactions de quelques caractéristiques classées les plus importantes dans le modèle LocalGLMnet mis en place (confère figure 5.37) :

– Sur le graphe correspondant à la variable *Avgdays.week_fit*, on observe des interactions quasiment linéaires entre la variable *Avgdays.week_fit* et les variables *Total.miles.driven_fit*, *Annual.pct.driven_fit*, *Left.turn.intensity08_fit*, puisque $\partial_k \hat{\beta}_j(x) \approx constante \neq 0$.

– Sur le graphe correspondant à la variable *Pct.drive.rush.am_fit*, on observe des interactions linéaires entre la variable *Pct.drive.rush.am_fit* et les variables *Total.miles.driven_fit*, *Annual.pct.driven_fit*, *Brake.09miles_fit*, puisque $\partial_k \hat{\beta}_j(x) \approx constante \neq 0$.

– Sur le graphe correspondant à la variable *Brake.09miles_fit*, on observe des interactions quasiment linéaires entre la variable *Brake.09miles_fit* et les variables *Total.miles.driven_fit*, *Annual.pct.driven_fit*, *Left.turn.intensity08_fit*, puisque $\partial_k \hat{\beta}_j(x) \approx constante \neq 0$.

En somme, on constate que les variables *Total.miles.driven_fit*, *Annual.pct.driven_fit*, *Left.turn.intensity08_fit*, *Car.age* et *Brake.09miles_fit* interagissent linéairement avec quasiment l'ensemble des autres variables télématiques dans leur impact sur la fréquence de sinistre.

□ Simulation du processus de calcul d'une prédiction dans le modèle LocalGLMnet

Dans ce paragraphe nous essayons de simuler le processus de prédiction d'une instance donnée. Comme dans le cadre des modèles *GLM*, une prédiction se décompose simplement comme une somme pondérée des caractéristiques de l'assuré, à laquelle on applique l'inverse d'une fonction de lien.

Cependant, dans le cadre des *LocalGLMnet*, les poids affectés aux différentes caractéristiques ne sont pas constants, ils sont dépendants des caractéristiques de chaque assuré. Ces poids permettent de mieux comprendre les caractéristiques ayant le plus contribué à la formation de la valeur prédite.

En guise d'illustration, considérons deux assurés quelconques de notre jeu de données de test, dont l'un est l'assuré *lambda* présenté plus haut dans le tableau 5.17, et un autre que nous nommons assuré *beta*, dont les caractéristiques sont présentées dans le tableau 5.19.

La fréquence de sinistre prédite par notre modèle LocalGLMnet est de 0.03 pour pour l'assuré *lambda* et de 2.77 pour l'assuré *beta*.

Essayons à présent de comprendre comment le modèle aboutit à ces deux prédictions.

Le processus de prédiction se fait en trois étapes :

Insured.age : **20 ans**	*Car.age* : **0 an** ; *Credit.score* : **580**	*Annual.miles.drive* : **12427.2**
Car.use : **Commute**	*TerritoryG* : **zone_C**	*Total.miles.driven_fit* : **0.92**
Annual.pct.driven_fit : **0.63**	*Pct.drive.rush.am_fit* : **0.11**	*Left.turn.intensity08_fit* : **0.96**
Pct.drive.rush.pm_fit : **0.95**	*Brake.08miles_fit* : **0.95**	*Brake.09miles_fit* : **0.83**
Pct.drive.4hrs_fit : **0.90**	*Pct.drive.2hrs_fit* : **0.98**	*Accel.14miles_fit* : **0.85**
Accel.11miles_fit : **0.87**	*Accel.06miles_fit* : **0.96**	*Avgdays.week_fit* : **1**

TABLE 5.19 : Caractéristiques de l'assuré *beta*

– Initialement, on normalise toutes les caractéristiques continues et binaires de la base de test, afin de les ramener toutes à la même échelle. Pour les caractéristiques catégorielles non binaires, on les transforme par un encodage à chaud (one-hot encoding) en autant de variables factices (*dummy* variables) que de catégories, puis on normalise chacune des variables factices obtenues.

– Ensuite, pour un assuré donné, on se sert du modèle *Feed Forward Network* ajusté, pour calculer le poids des différentes caractéristiques spécifiques à cet assuré.

– Enfin, une fois les poids des caractéristiques obtenus pour l'assuré, on calcule la prédiction par une somme pondérée des caractéristiques, à laquelle on applique l'inverse de notre fonction de lien (qui ici est exponentielle).

Pour un individu i la fréquence de sinistre est alors calculée par la formule suivante :

$$\begin{aligned}
frequence_i &= \ln^{-1}[\hat{\beta}_{intercept} + \hat{\beta}^{(i)}_{TerritoryGzone_A}\tilde{X}^{(i)}_{TerritoryGzone_A} + \hat{\beta}^{(i)}_{TerritoryGzone_B}\tilde{X}^{(i)}_{TerritoryGzone_B} \\
&+ \hat{\beta}^{(i)}_{TerritoryGzone_C}\tilde{X}^{(i)}_{TerritoryGzone_C} + \hat{\beta}^{(i)}_{Car.useCommercial}\tilde{X}^{(i)}_{Car.useCommercial} \\
&+ \hat{\beta}^{(i)}_{Car.useCommute}\tilde{X}^{(i)}_{Car.useCommute} + \hat{\beta}^{(i)}_{Car.useFarmer}\tilde{X}^{(i)}_{Car.useFarmer} \\
&+ \hat{\beta}^{(i)}_{Car.usePrivate}\tilde{X}^{(i)}_{Car.usePrivate} + \hat{\beta}^{(i)}_{Car.age}\tilde{X}^{(i)}_{Car.age} + \hat{\beta}^{(i)}_{Insured.age}\tilde{X}^{(i)}_{Insured.age} \\
&+ \hat{\beta}^{(i)}_{Credit.score}\tilde{X}^{(i)}_{Credit.score} + \hat{\beta}^{(i)}_{Annual.miles.drive}\tilde{X}^{(i)}_{Annual.miles.drive} \\
&+ \hat{\beta}^{(i)}_{Total.miles.driven_fit}\tilde{X}^{(i)}_{Total.miles.driven_fit} + \hat{\beta}^{(i)}_{Annual.pct.driven_fit}\tilde{X}^{(i)}_{Annual.pct.driven_fit} \\
&+ \hat{\beta}_{Pct.drive.rush.am_fit}\tilde{X}^{(i)}_{Pct.drive.rush.am_fit} + \hat{\beta}^{(i)}_{Pct.drive.rush.pm_fit}\tilde{X}^{(i)}_{Pct.drive.rush.pm_fit} \\
&+ \hat{\beta}^{(i)}_{Left.turn.intensity08miles_fit}\tilde{X}^{(i)}_{Left.turn.intensity08miles_fit} + \hat{\beta}^{(i)}_{Brake.08miles_fit}\tilde{X}^{(i)}_{Brake.08miles_fit} \\
&+ \hat{\beta}^{(i)}_{Brake.09miles_fit}\tilde{X}^{(i)}_{Brake.09miles_fit} + \hat{\beta}^{(i)}_{Pct.drive.4hrs_fit}\tilde{X}^{(i)}_{Pct.drive.4hrs_fit} \\
&+ \hat{\beta}^{(i)}_{Pct.drive.2hrs_fit}\tilde{X}^{(i)}_{Pct.drive.2hrs_fit} + \hat{\beta}^{(i)}_{Accel.14miles_fit}\tilde{X}^{(i)}_{Accel.14miles_fit} \\
&+ \hat{\beta}^{(i)}_{Accel.11miles_fit}\tilde{X}^{(i)}_{Accel.11miles_fit} + \hat{\beta}^{(i)}_{Accel.06miles_fit}\tilde{X}^{(i)}_{Accel.06miles_fit} \\
&+ \hat{\beta}^{(i)}_{Avgdays.week_fit}\tilde{X}^{(i)}_{Avgdays.week_fit}]
\end{aligned}$$

où $\tilde{X}^{(i)}_j$ désigne la valeur normalisée de la variable j pour l'assuré i ; $\hat{\beta}_{intercept}$ correspond à l'intercept du modèle et est donc indépendant des assurés.

On peut à présent retrouver les fréquences prédites de nos assurés *lambda* et *beta* par notre modèle LocalGLMnet-Poisson en effectuant les opérations suivantes :

$$\begin{aligned}
frequence_{lambda} &= \exp[-3.28 + 0.24\times(-0.52) + (-0.005)\times 0.73 + 0.08\times(-0.40) \\
&+ (-0.21)\times(-0.16) + (-0.03)\times(-0.99) + 0.03\times(-0.12) + (-0.13)\times 1.08 \\
&+ (-0.24)\times 0.56 + 0.27\times 1.20 + (-0.13)\times(-0.17) + 0.01\times(-0.76) \\
&+ 0.03\times 0.11 + (-0.14)\times 0.29 + (-0.05)\times 0.15 + (-0.05)\times 0.22 \\
&+ (-0.20)\times 0.72 + 0.30\times 0.11 + 0.13\times 0.01 + 0.12\times 0 + 0.07\times 0.33 \\
&+ 0.07\times 0.11 + (-0.07)\times 0.21 + (-0.10)\times 0.63 + 0.012\times 0.99] \\
&= exp[-3.526747] \\
&= 0.030
\end{aligned}$$

$$
\begin{aligned}
frequence_{beta} &= \exp\left[-3.28 + (-0.16) \times (-0.52) + 0.14 \times (-1.37) + (0.07) \times 2.54\right. \\
&+ 0.06 \times (-0.16) + 0.04 \times 1 + 0.22 \times (-0.12) + (-0.28) \times (-0.93) \\
&+ 0.18 \times (-1.45) + (-0.096) \times (-2.03) + (-0.14) \times (-2.64) + 0.19 \times 0.85 \\
&+ 0.08 \times 0.92 + (-0.007) \times 0.63 + 0.21 \times 0.11 + 0.39 \times 0.95 + 0.05 \times 0.96 \\
&+ 0.71 \times 0.95 + 0.58 \times 0.83 + 0.40 \times 0.90 + 0.19 \times 0.98 + 0.31 \times 0.85 \\
&+ \left.(-0.01) \times 0.87 + 0.60 \times 0.96 + 0.49 \times 1\right] \\
&= exp\left[1.018379\right] \\
&= 2.77
\end{aligned}
$$

On constate clairement que la composante qui contribue le plus positivement à la fréquence prédite de sinistre de l'assuré *lambda* est *Insured.age* et la composante qui contribue le plus à réduire la fréquence prédite de sinistre chez cet assuré est *Left.turn.intensity08_fit*.

Par ailleurs, contrairement au modèle GLM-Poisson où la contribution de la variable *Total.miles_driven_fit* était la plus importante dans la fréquence prédite de sinistre de l'assuré *lambda*, avec le modèle LocalGLMnet, sa contribution directe n'est pas substantiellement significative. Ceci pourrait s'expliquer par le fait que, dans le modèle LocalGLMnet, la variable *Total.miles_driven_fit* agit sur la fréquence de sinistre à travers ses multiples interactions avec les autres caractéristiques, plutôt que directement.

En outre, on note un écart significatif entre les fréquences prédites de sinistre pour nos deux assurés : l'assuré *beta* a une fréquence prédite de sinistre de très loin supérieure à celle de l'assuré *lambda*. Ce qui reflète bien la réalité, puisque la valeur réelle observée du nombre de sinistre de l'assuré *lambda* est de 0, alors que celle de l'assuré *lambda* est de 2.

On peut identifier les facteurs explicatifs de cet écart entre les fréquences prédites des deux assurés, en comparant termes à termes les contributions des différentes caractéristiques dans le calcul des fréquences ci-dessus. On note des différences importantes de la contribution des variables *Brake.08miles_fit*, *Accel.06miles_fit*, *Avgdays.week_fit*, *Brake.09miles_fit* et *Credit.score* chez les deux assurés. Ils ont un comportement de conduite très différent : l'assuré *beta* a une propension à accélérer et à freiner brusquement beaucoup plus importante que celle de l'assuré *lambda*, ce qui explique principalement l'écart entre les fréquences prédites de sinistre des deux assurés.

Sur la figure 5.39, on relève un fait notable : les deux assurés ont une forte propension à rouler plus de quatre (04) jours par semaine ($Avgdays.week\_fit_{lambda} = 0.99$, $Avgdays.week\_fit_{beta} = 1$), cependant la contribution de cette variable dans la fréquence prédite de l'assuré *lambda* est largement inférieure à sa contribution dans le calcul de la fréquence prédite de l'assuré *beta*. Cela s'expliquerait tout simplement par l'existence des effets d'interaction entre les caractéristiques dans notre modèle LocalGLMnet.

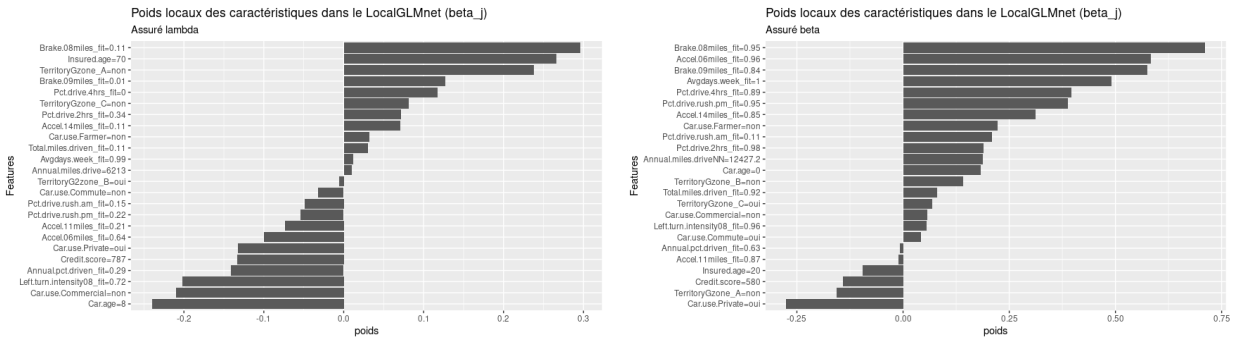


FIGURE 5.39 : *Poids locaux des caractéristiques dans la prédiction de la fréquence de sinistre des assurés lambda (à gauche) et beta (à droite), dans le modèle LocalGLMnet* ($\beta_j(x^{(i)})$).

5.5.2.2 Interprétation *post hoc* du modèle LocalGLMnet fréquence

À présent nous pouvons passer à l'interprétation *post hoc* de notre modèle LocalGLMnet à l'aide des outils présentés au chapitre 4. L'objectif est non seulement de vérifier la cohérence entre les interprétations ci-dessus basées sur le modèle et celles obtenues par le biais des outils d'interprétation agnostiques au modèle présentés au chapitre 4, mais aussi, d'obtenir éventuellement de nouveaux éclaircissements sur le fonctionnement de notre modèle, aussi bien au niveau global que local.

(A) Analyse globale du modèle

□ Importance des caractéristiques

Comme à l'accoutumée, nous commençons par étudier l'importance globale des caractéristiques dans le modèle, afin d'identifier les caractéristiques qui influent le plus, de manière globale, sur les prédictions.

Nous souhaitons vérifier est-ce que les hiérarchisations de l'importance des caractéristiques obtenus par la méthode basée sur les permutations (MR) et celle basée sur les valeurs de Shapley (SFIMP) présentées au chapitre 4 coïncident bien avec la hiérarchisation de l'importance des caractéristiques basée sur le modèle (confère figure 5.37).

Les résultats obtenus à l'aide de ces deux méthodes agnostiques au modèle (MR et SFIMP) sont donnés sur la figure 5.40. L'analyse comparative des figures 5.37 et 5.40 révèle trois faits notables :

– Premièrement les résultats obtenus par ces méthodes agnostiques au modèle, à savoir, celle basée sur les permutations (MR) et celle basée sur les valeurs de Sphaley (SFIMP) coïncident à quelques exceptions près.

De plus, elles sont conformes aux résultats obtenus dans le cadre du modèle GLM (confère figure 5.29). D'après ces résultats, globalement, les variables les plus utilisées par le modèle LocalGLMnet pour réaliser ses prédictions sont : *Total.miles_driven_fit*, *Car.use.Private*, *Car.age*, *Credit.score* et *Insured.age*.

La présence de ces variables améliore le mieux la performance prédictive du modèle. Réciproquement, leur absence dégraderait le plus la performance prédictive du modèle.

– Deuxièmement, lorsqu'on compare les résultats obtenus avec MR et SFIMP à ceux obtenus plus haut par la méthode d'évaluation de l'importance globale des caractéristiques spécifique aux LocalGLMnet (confère figure 5.37), on note une inversion flagrante de l'importance de certaines variables.

La probabilité de parcourir plus de 4500 miles durant l'année (*Total.miles.driven_fit*) qui se trouvait en queue du classement pour la méthode de mesure de l'importance globale des variables spécifique aux LocalGLMnet, est jugée comme caractéristique la plus importante avec les méthodes MR et SFIMP.

Cette inversion proviendrait du fait que les méthodes fonctionnent différemment et mesure des réalités différentes : la méthode SFIMP et MR jugent l'importance des variables sur la base de leur capacité à réduire l'erreur globale de prédiction du modèle, alors que la méthode spécifique au LocalGLMnet présentée au chapitre 2 évalue la contribution quantitative de chaque caractéristique à la formation des valeurs prédites.

Ainsi, les résultats de ces différentes approches sont certes différents, mais pas incohérents. En effet, certaines caractéristiques pourraient améliorer significativement la performance prédictive du modèle par le biais de leur interaction avec les autres caractéristiques, sans pourtant contribuer directement sur la variable cible. C'est d'ailleurs le cas de la variable *Total.miles.driven_fit* dans notre contexte, qui interagit substantiellement avec les autres caractéristiques (confère. figure C.18 en annexe) et donc améliore la précision prédictive du modèle, sans pourtant avoir une contribution directe significative sur la fréquence prédite de sinistre (confère figure 5.38).

– Pour finir, nous pouvons faire un troisième commentaire sur la forme du diagramme de l'importance des variables issue de la méthode MR (*Model reliance* en anglais).

On remarque qu'elle a une forme quasiment rectangulaire. Autrement dit, d'après ces résultats, toutes les caractéristiques auraient sensiblement le même niveau d'importance dans le modèle.

Pour mieux interpréter ce résultat, rappelons le principe général de la méthode MR : elle consiste à mesurer l'importance d'une caractéristique en calculant l'augmentation de l'erreur de prédiction du modèle conséquente à une permutation des valeurs de cette caractéristique.

Pour une caractéristique donnée, une grande valeur de MR (supérieure à 1) signifie que celle-ci améliore significativement la précision prédictive du modèle. Une valeur de MR égale à 1 signifie que le modèle ne dépend pas substantiellement de la caractéristique concernée en ce sens qu'un brouillage des valeurs de la caractéristique n'a aucune incidence sur la précision prédictive du modèle.

On souligne que toutes nos caractéristiques ont toute un MR légèrement au dessus de 1, donc chacune d'entre-elles a une contribution au moins marginale à la performance prédictive du modèle. Cependant, aucune d'entre elles ne démarque considérablement des autres.

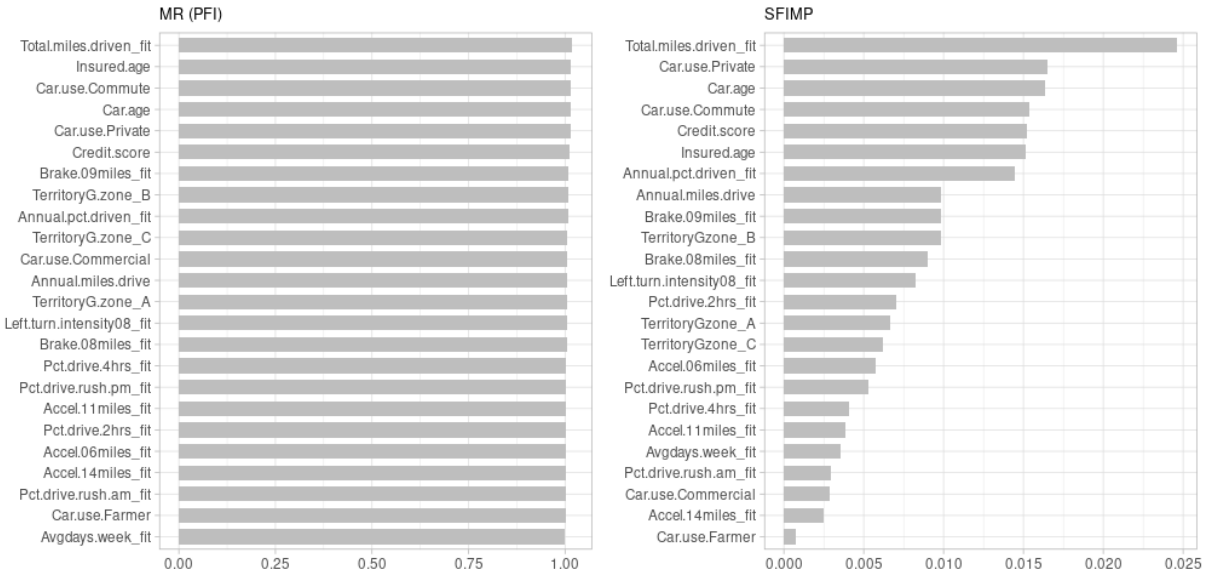


FIGURE 5.40 : *Importance des variables dans le modèle LocalGLMnet fréquence, basée sur la permutation des caractéristiques (MR), et sur les valeurs de Shapley (SFIMP).*

□ Effet marginal des caractéristiques : effets locaux accumulés (ALE)

Analysons à présent les graphiques des effets locaux accumulés (ALE). Les résultats obtenus à l'aide du package DALEX de R sont représentés sur la figure 5.41. On rappelle qu'en abscisse de chaque diagramme ALE, il s'agit non des valeurs des caractéristiques en l'état, mais de leur valeur normalisée. Cela n'a aucune incidence particulière sur la pertinence des interprétations qui seront faites.

En analysant ces différentes courbes ALE pour chacune des caractéristiques, on retrouve quasiment les mêmes courbes que celle obtenues précédemment lors du tracé des ajustements splines des contributions des différentes caractéristiques dans le modèle LocalGLMnet, à un changement d'échelle près (confère 5.38). Ce constat est plutôt cohérent avec la théorie.

De la lecture des diagrammes ALE, nous pouvons faire les remarques complémetaires suivantes, tout en gardant à l'esprit que corrélation n'est pas synonyme de causalité :

– *Credit.score :* comme nous l'avions observé au niveau de la contribution des variables, la sinistralité est minimale chez les assurés à très faible score de crédit et chez les assurés à fort score de crédit, et relativement élevée chez les conducteurs à score de crédit intermédiaire.

– *Car.age :* on observe une décroissance de l'impact de l'âge du véhicule sur la sinistralité. Cela peut provenir du fait que lorsqu'on détient une voiture depuis longtemps, on maîtrise mieux son fonctionnement et ses potentiels dysfonctionnements, ainsi nous sommes davantage prudent dans son usage, et par conséquent nous commettons moins de sinistres.

– *Insured.age :* comme observé en analyse préliminaire, l'âge a un effet quadratique sur la fréquence de sinistre : les plus jeunes et les plus âgés sont les plus risqués relativement aux assurés d'âge intermédiaire. L'effet de l'âge sur la sinistralité demeure nettement plus accentué chez les personnes âgées que chez toutes les autres classes d'âge.

– *Car.use :* les véhicules à usage agricole (*Car.use.Farmer*) semblent être les moins sinistrés. Ce résultat est conforme à l'analyse descriptive préliminaire que l'on avait réalisé sur la figure 5.9.

– *Annual.miles.drive* : on observe une croissance de l'impact du nombre annuel de miles à parcourir déclaré à la souscription par l'assuré, jusqu'à un certain seuil (assez élevé), à partir duquel l'effet devient décroissant, jusqu'à s'annuler. Cela peut provenir du fait que les assurés qui dès la souscription déclare un volume espéré de conduite très élevé sont généralement des conducteurs assez expérimentés, donc plus habiles et prudents.

– *Brake09miles_fit, Brake08miles_fit, Accel11miles_fit :* de manière générale, on observe une croissance de l'impact de la probabilité d'avoir un nombre élevé de freinages brusques d'intensité 08 mph/s et 09 mph/s par 1000 miles, et du nombre d'accélération d'intensité 11 mph/s par 1000 miles sur la sinistralité, ce qui est intuitivement logique, car ils sont révélateurs des comportements de conduite imprudente.

Jusqu'à ce stade, nous avons globalement obtenu des résultats d'interprétations similaires à ceux de l'interprétation basée sur le modèle. Ce qui est plutôt rassurant.

□ Analyse des interactions globales : Courbes ICE, H-statistiques et indices de Sobol

• Courbes ICE : c-ICE et d-ICE

Afin de peaufiner notre compréhension de l'effet des différentes caractéristiques sur la fréquence de sinistre, nous investiguons les potentiels effets hétérogènes des caractéristiques dans notre modèle de fréquence.

Nous analysons les courbes ICE (*Individual Conditionnal Expectation*, en anglais) des différentes caractéristiques.

Nous nous contentons de présenter les résultats uniquement pour les caractéristiques qui se sont précédemment démarquées à l'étape de l'analyse des interactions basée sur le modèle (confère figure C.18), à savoir : *Insured.age*, *Credit.score*, *Car.age*, *Total.mile.driven_fit*, *Left.intensity.08_fit*, *Brake09miles_fit*.

Nous avons utilisé les versions dérivées de la courbe ICE à savoir c-ICE (ICE centée) et d-ICE (ICE dérivée) pour leur clarté visuelle et leur caractère intuitif qui rendent faciles les interprétations.

Sur les graphiques c-ICE de la figure 5.42, l'axe vertical à droite affiche les changements de $\hat{f}$ par rapport à sa valeur médiane due à l'accroissement de la caractéristique concernée, en interaction avec les autres caractéristiques du modèle.

Les graphiques c-ICE de la figure 5.42 montrent clairement que l'effet cumulatif de l'âge du véhicule *Car.age* sur la fréquence prédite de sinistre augmente dans certains cas et diminue dans d'autres. Il

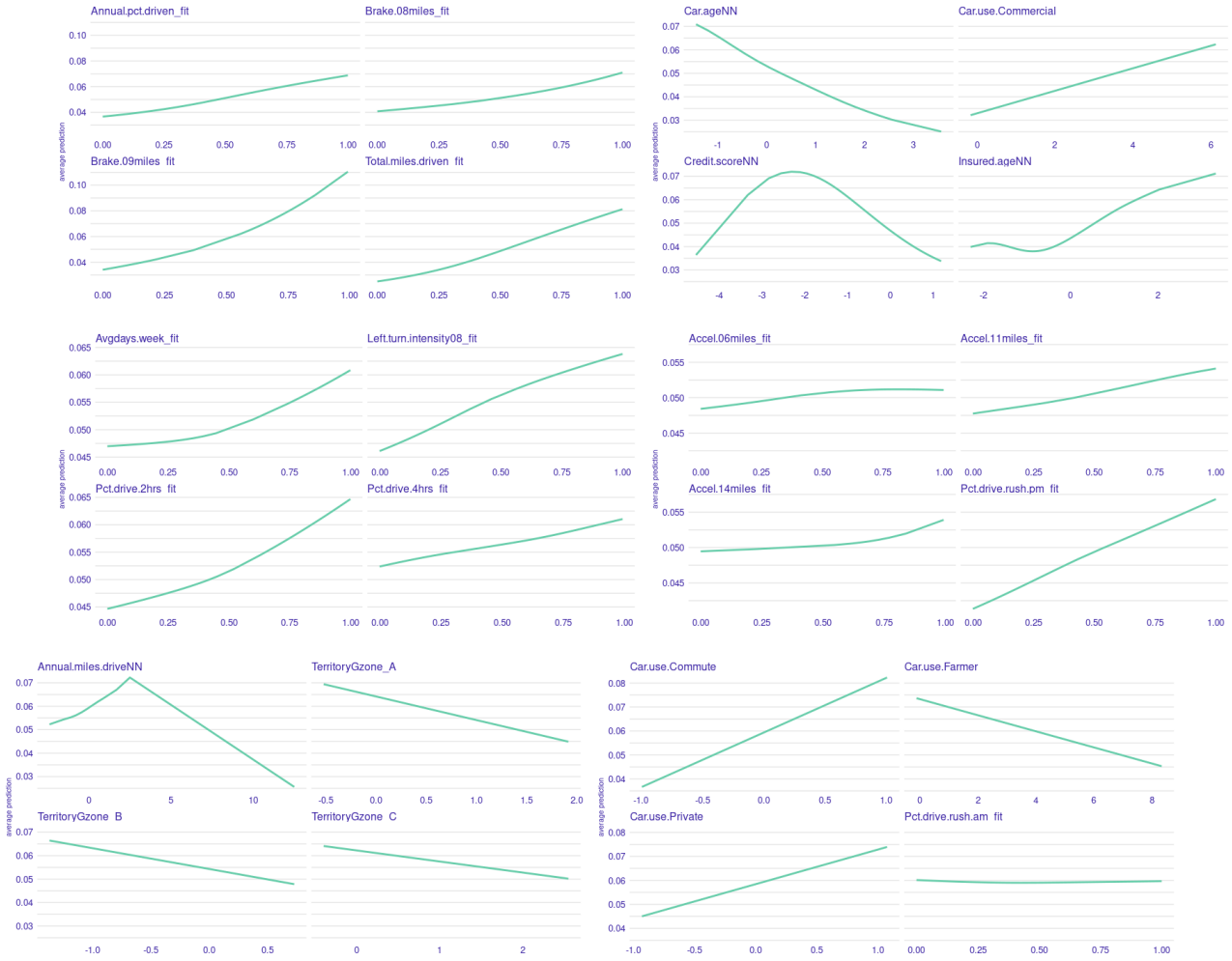


FIGURE 5.41 : *Graphiques des effets locaux accumulés (ALE) des 24 caractéristiques pour le LocalGLMnet-fréquence.*

en est de même pour les caractéristiques *Left.turn.intensity08_fit*, *Insured.age* et *Credit.score* qui font référence respectivement au nombre de virage à gauche d'intensité 08 mph/s par 1000 miles, à l'âge du conducteur assuré et au score de crédit du conducteur assuré. De telles divergences des courbes ICE centrées suggèrent l'existence d'interactions entre les caractéristiques concernées et le reste des caractéristiques dans le modèle.

En ce qui concerne les variables *Brake09.miles_fit* et *Total.mile.driven_fit* qui font respectivement référence au nombre d'accélération d'intensité 09 mph/s par 1000 miles et à la distance annuelle parcourue par un conducteur assuré, leurs graphiques c-ICE révèlent que leur effet cumulatif sur le nombre de sinistre est toujours croissant, mais avec effet plus accentué chez certains assurés que chez d'autres.

L'analyse des courbes ICE dérivées (d-ICE) nous apporte des informations supplémentaires. En effet, comme il peut être difficile d'évaluer visuellement les dérivées $\partial \hat{f}$ à partir des tracés c-ICE, il est pratique de tracer directement une estimation de la dérivée partielle directement : c'est ce que font les graphiques d-ICE. Les détails de cette procédure sont donnés dans le chapitre 4.

Pour une caractéristique donnée, lorsqu'elle n'interagit pas avec les autres caractéristiques dans le modèle, les courbes du tracé d-ICE sont toutes équivalentes et le tracé présente une seule ligne épaisse. Lorsque des interactions existent, les lignes dérivées sont hétérogènes.

Dans notre contexte, en se limitant aux zones contenant des points noirs (en effet, les zones ne contenant des points sont des zones d'extrapolation et donc peuvent mener aux interprétations fallacieuses), on observe bien que lorsque l'âge du véhicule est inférieur au quantile d'ordre 80% (qui correspond à 9 ans dans notre jeu de données), il y a des observations pour lesquelles la dérivée est négative et d'autres pour lesquelles la dérivée est positive, ce qui suggère l'existence d'interaction entre l'âge du véhicule *Car.age* et d'autres caractéristiques du modèle.

Il en est de même pour l'âge de l'assuré *Insured.age*, *Left.intensity.08miles_fit* et *Credit.score*.

Cependant, en ce qui concerne les variables *Total.mile.driven_fit* et *Brake09.miles_fit*, l'analyse de leurs courbes d-ICE confirme bien qu'elles ont toujours effet cumulatif positif sur la fréquence de sinistre. L'ampleur de cet effet cumulatif n'est pas linéaire, mais il s'intensifie au fur et à mesure que les valeurs de ces variables s'accroissent. Pour deux assurés donnés, l'un ayant une faible probabilité de rouler plus de 4500 miles durant l'année et l'autre ayant une forte probabilité de rouler plus de 4500 miles durant l'année, un accroissement supplémentaire d'une unité de la valeur leur probabilité de rouler plus de 4500 miles aura un effet croissant sur leur fréquence prédite de sinistre, plus accentué chez le second assuré que chez le premier.

Enfin, l'écart-type des dérivées partielles en chaque point, tracé dans le panneau inférieur des courbes d-ICE, sert de résumé utile pour mettre en évidence les régions d'hétérogénéité dans les dérivées estimées (c'est-à-dire les preuves potentielles d'interaction dans le modèle ajusté).

L'analyse de l'évolution de l'écart type de la dérivée ICE dans le volet inférieur des différents graphiques d-ICE de la figure 5.42 montre :

– Un effet hétérogène de l'âge du véhicule (*Car.age*) sur la fréquence prédite de sinistre –graphique en haut, à gauche–. La plus grande hétérogénéité s'observant lorsque l'âge du véhicule est faible. Il en est de même pour la caractéristique *Credit.Score*.

– Pour les caractéristiques *Brake09.miles_fit*, *Total.mile.driven_fit* et *Insured.age* la plus grande hétérogénéité de leur effet se produit pour des valeurs élevées.

– Aucune hétérogénéité significative n'est observée dans l'effet *Left.intensity.08miles_fit* sur le modèle. Ce résultat montre que cette caractéristique interagit faiblement avec les autres dans son influence sur la fréquence de sinistre.

- **H-statistiques**

Cet indicateur nous permet d'identifier et de classer les caractéristiques entre elles, suivant leur force d'interaction avec les autres caractéristiques dans le modèle. Dans notre cas, nous avons calculé les H-statistique totale des différentes caractéristiques, afin d'identifier celles qui ont des effets d'interaction les plus significatifs dans le modèle.

L'interaction basée sur la statistique H totale de Friedman peut s'interpréter comme la part de variance du modèle (ou des prédictions) expliquée par les interactions de la caractéristique concernée avec les autres. Elle est nulle en cas d'absence d'interaction entre la caractéristique concernée et les autres caractéristiques du modèle. Une valeur proportionnellement grande de la H-statistique de Friedman indique une forte présence d'interaction entre la caractéristique concernée et les autres caractéristiques dans le modèle.

Les résultats obtenus à l'aide du package *iml* de R sont donnés sur la figure 5.43.

On observe clairement que les caractéristiques *Car.age*, *Total.miles.driven_fit*, *Annual.miles.drive_fit* et *Brake08miles_fit* figurent parmi celles qui interagissent le plus avec les autres caractéristiques dans le modèle.

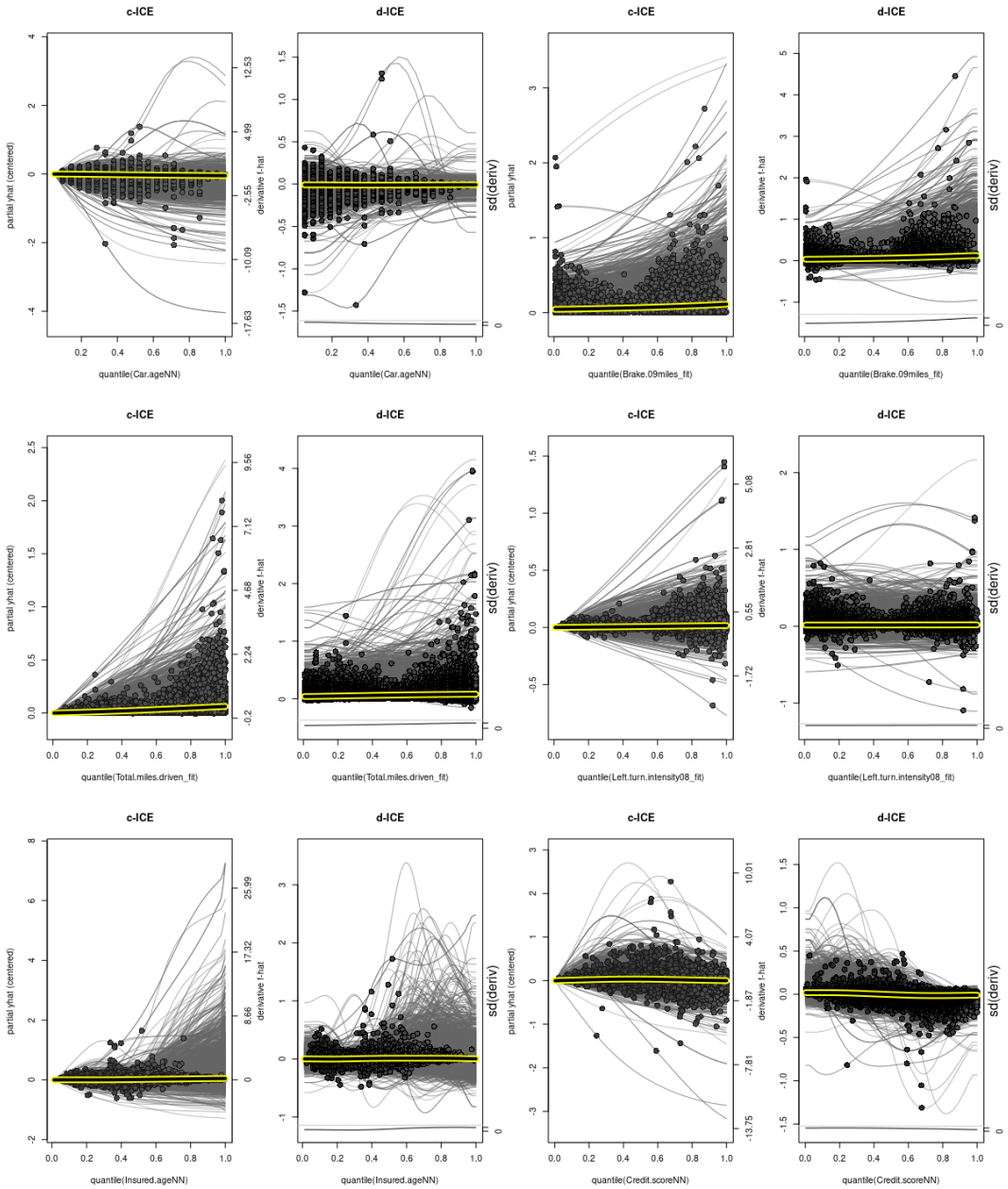


FIGURE 5.42 : *Courbes c-ICE et d-ICE de quelques caractéristiques de notre modèle LocalGLMnet-fréquence : la ligne en jaune sur les graphiques c-ICE correspond à la courbe PDP centrée, et sur les graphiques d-ICE elle correspond à la dérivée de la courbe PDP.*

Les caractéristiques *Avgday.week* et *Accel.14miles_fit* quant à elles figurent parmi les caractéristiques qui interagissent le moins avec les autres, au sens de la H-statistique de Friedman.

En tête du classement on observe les variables binaires *Car.use.Commute* et *Territory.Gzone_C*. Cependant, cela proviendrait en partie du fait que la H-statistique de Friedman a tendance à surestimer les effets d'interaction des variables catégorielles. Il faut donc prendre du recul dans l'analyse des Statistiques H des variables catégorielles.

L'une des principales limites des H-statistiques provient du fait qu'elle permet certes de repérer les caractéristiques qui interagissent potentiellement le plus avec les autres, mais sans toutefois fournir d'informations supplémentaires sur la nature ou la forme de l'interaction existante entre les caractéristiques.

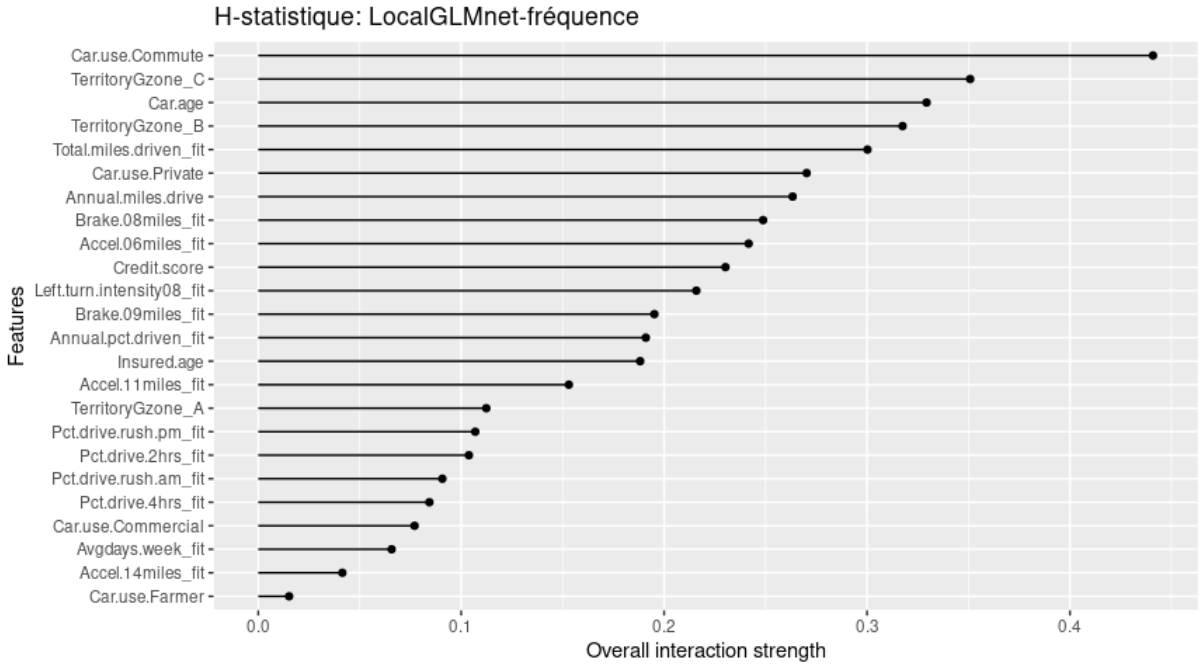


FIGURE 5.43 : *Résultats H-statistique de l'interaction totale des caractéristiques dans le modèle LocalGLMnet-fréquence.*

- **Indices de Sobol**

Un autre outil intéressant pour mesurer la force d'interaction entre les caractéristiques est l'indice de sensibilité de Sobol présenté au chapitre 4. Le principe de cette approche s'inspire essentiellement du principe de la décomposition réalisée en analyse de la variance, plus connue sous le nom de *ANOVA decomposition*.

L'indice de sensibilité de Sobol entre deux caractéristiques encore appelé indice de Sobol d'ordre 2 est mesurée en effectuant le ratio de [la différence entre la variance du modèle ajusté sur la paire de caractéristiques concernées et la somme des variances des modèles ajustés individuellement sur chacune des deux caractéristiques] par la [variance totale du modèle].

De la même manière, il est possible de calculer les indices de Sobol d'ordre $k > 2$ ou d'ordre $k = 1$ (indice de premier ordre).

Ainsi, l'interaction totale (indice de Sobol total) d'une caractéristique est obtenu en sommant de tous les indices de sensibilité (de différents ordres) de ladite caractéristique.

Les indices de Sobol d'ordres quelconques et total sont des quantités théoriquement toujours positives et comprises entre 0 et 1. Dans notre cas, nous nous sommes restreint à l'implémentation des indices de sensibilité total et de premier ordre.

Les résultats obtenus suite à l'implémentation à l'aide de la fonction *sobol2002* du package *sensibility* de R sont présentées sur la figure 5.44.

Pour une variable donnée, un écart significatif entre l'indice totale et l'indice de premier ordre indique une forte présence d'effets d'interaction de ladite variable avec les autres. On observe des estimations négatives de la valeur de certains indices de premier ordre et même de certains indices totaux, ce qui est théoriquement absurde. Il est également absurde d'obtenir un indice de premier ordre strictement supérieur à l'indice total. Ces imprécisions computationnelles sont généralement observée pour des caractéristiques peu influentes dans l'explication du phénomène étudié.

Seules les estimations des indices relatifs aux caractéristiques *Total.miles.driven_fit* et *Left.turn.intensity08_fit* sont conformes à la théorie. Pour cette première caractéristique, on observe un écart relativement important entre la valeur de l'indice total et celle de l'indice de premier ordre. Ceci signifie que la caractéristique *Total.miles.driven_fit* interagit considérablement avec les autres caractéristiques du modèle : ce qui est en parfaite cohérence avec l'ensemble des résultats d'analyse d'interaction obtenus plus haut.

En ce qui concerne la caractéristique *Left.turn.intensity08_fit* on constate que l'indice d'interaction totale est relativement faible (de l'ordre de 8%). Cela signifie que cette caractéristique contribue faiblement à la variabilité des prédictions. De plus, l'écart entre l'indice d'interaction totale et l'indice d'interaction de premier ordre étant assez faible, cela signifie que interagit *Left.turn.intensity08_fit* faiblement avec les autres caractéristiques du modèle.

Nous ne prendrons pas le risque d'interpréter les indices de Sobol des autres caractéristiques, car leurs valeurs étant non conformes à la théorie.

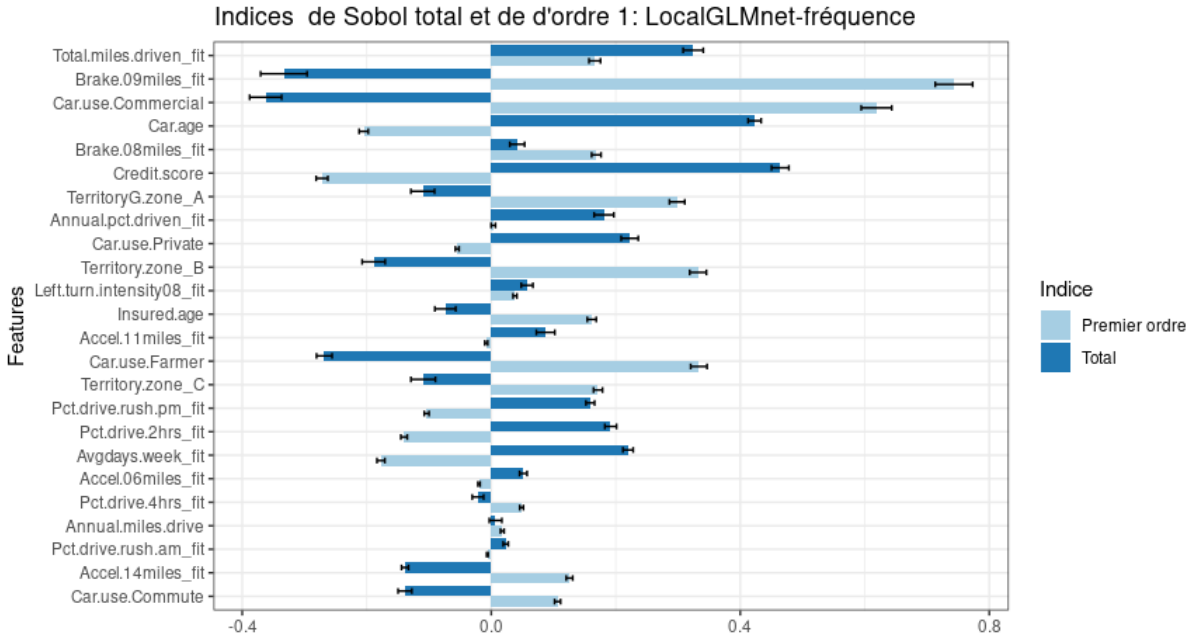


FIGURE 5.44 : *Résultats indices de Sobol de l'interaction totale et de premier ordre des caractéristiques dans le modèle LocalGLMnet-fréquence.*

• Focus sur les interactions bi-directionnelles entre les caractéristiques : ALE de second ordre

Dans les parties précédentes, à l'aide des courbes c-ICE et d-ICE, des H-statistiques totales et des indices de Sobol totaux et de premier ordre, nous avons pu identifier et hiérarchiser les caractéristiques ayant des effets hétérogènes les plus importants dans notre modèle LocalGLMnet-Poisson.

Il en est ressorti que l'âge du véhicule (*Car.age*), la distance distance annuelle parcourue (*Total.miles.driven_fit*) figurent parmi les caractéristiques qui interagissent le plus avec les autres caractéristiques dans leur effet sur la fréquence prédite de sinistre.

Pour l'instant, nous ignorons la nature de ces interactions. Plus précisément, pour une caractéristique donnée, par exemple l'âge du véhicule, nous ignorons les autres caractéristiques avec lesquelles elle interagit, et l'effet de leur(s) interaction(s) sur la fréquence prédite de sinistre.

Pour investiguer la nature et la forme de ces effets interactions entre les caractéristiques, nous mettons en place les courbes ALE de second ordre présentées au chapitre 4.

Étant donné que nous disposons de 24 caractéristiques dans notre modèle LocalGLMnet-Poisson, représenter les courbes ALE de second ordre de toutes les paires de caractéristiques reviendrait à représenter $C_{24}^2 = 276$ graphiques ALE : ce qui est énorme. Nous nous limitons ici, à la représentation des paires de caractéristiques issues du croisement entre la caractéristique *Car.age* −qui semble être celle ayant le plus d'interactions avec les autres caractéristiques dans notre modèle LocalGLMnet− et six autres variables ayant également des H-statistique totale relativement significative (confère figure 5.43).

Les résultats obtenus à l'aide du package *iml* de R sont présentés sur la figure 5.45. Nous interprétons les différents graphiques ALE de second ordre présentés sur cette figure tout en gardant à l'esprit que les corrélations ne sont pas synonymes de causalité.

– Commençons par analyser la paire de caractéristiques $\{Car.age, Total.miles.driven\_fit\}$.

Il y existe une "forte" interaction entre ces deux caractéristiques, puisque les valeurs de l'ALE d'ordre 2 issu du croisement entre ces deux caractéristiques varient entre -0.022 et +0.0134.

On constate une plus grande fréquence prédite de sinistre lorsque la voiture est récente (âgée de moins de deux ans) et que la probabilité prédite de parcourir plus de 4500 miles durant l'année est supérieure à 75%. *A contrario*, les voitures âgées de plus 16 ans qui ont une probabilité au dessus de 60% de rouler plus de 4500 miles de route durant l'année sont associées à une fréquence prédite de sinistre relativement moindre.

– L'analyse du graphique ALE associé à la paire $\{Car.age, Brake09miles\_fit\}$ montre également une forte interaction entre ces deux caractéristiques dans le modèle, avec un pic de sinistralité chez les conducteurs qui ont une voiture récente (âgée de moins de deux ans) et qui ont une forte probabilité d'effectuer plus de 5 freinages brusques d'intensité 09 mph/s par 1000 miles, ce qui n'est pas du tout étrange. Par contre, chez les assurés ayant une probabilité prédite supérieure à 60% d'effectuer plus 5 virages d'intensité 09 mph/s par 1000 miles, mais qui ont une ancienne voiture (âgée d'au moins 6 ans), la fréquence prédite de sinistre qui leur est associée est relativement moins élevée.

– Le graphique relatif à la paire $\{Car.age, Credit.score\}$ nous révèle d'intéressantes informations. Pour les assurés ayant score de crédit en deçà de 600 points, l'effet de l'âge du véhicule sur la sinistralité est différencié : ceux ayant une véhicule récent ont un niveau de sinistralité relativement plus faible que ceux qui ont un vieux engin, toutes choses égales par ailleurs. Cette seconde catégorie d'assuré correspond souvent à des jeunes qui ont hérité de l'ancienne voiture d'un parent, ou à des personnes d'âge intermédiaire au chômage, ou encore à des personnes âgés en situation de précarité financière.

Chez les assurés ayant un score de crédit intermédiaire (entre 650 points et 750 points), on observe un effet inverse, ceux ayant une voiture récente (de moins de 4 ans) sont cette fois associés à une fréquence prédite de sinistre relativement plus élevée que ceux ayant une voiture plus ancienne.

– Une autre paire de caractéristique intéressante est $\{Car.age, Insured.age\}$. On observe un phénomène classique en assurance automobile : les jeunes assurés disposant d'une voiture très âgée (de plus de seize ans) sont associés à un nombre important de sinistre, de même que les jeunes assurés titulaire d'une voiture neuve (de moins d'un an).

– Quant à la paire $\{Car.age, Left.turn.intensity08\_fit\}$, on observe un pic de sinistralité chez les assurés disposant d'un véhicule récent de moins de deux (02) ans et qui ont une probabilité prédite supérieure à 80% d'effectuer plus de trente virages à gauche d'intensité 08 mph/s par 1000 miles.

– En ce qui concerne la paire de caractéristiques $\{Car.age, Accel06mile\_fit\}$, leur interaction est quasiment inexistante.

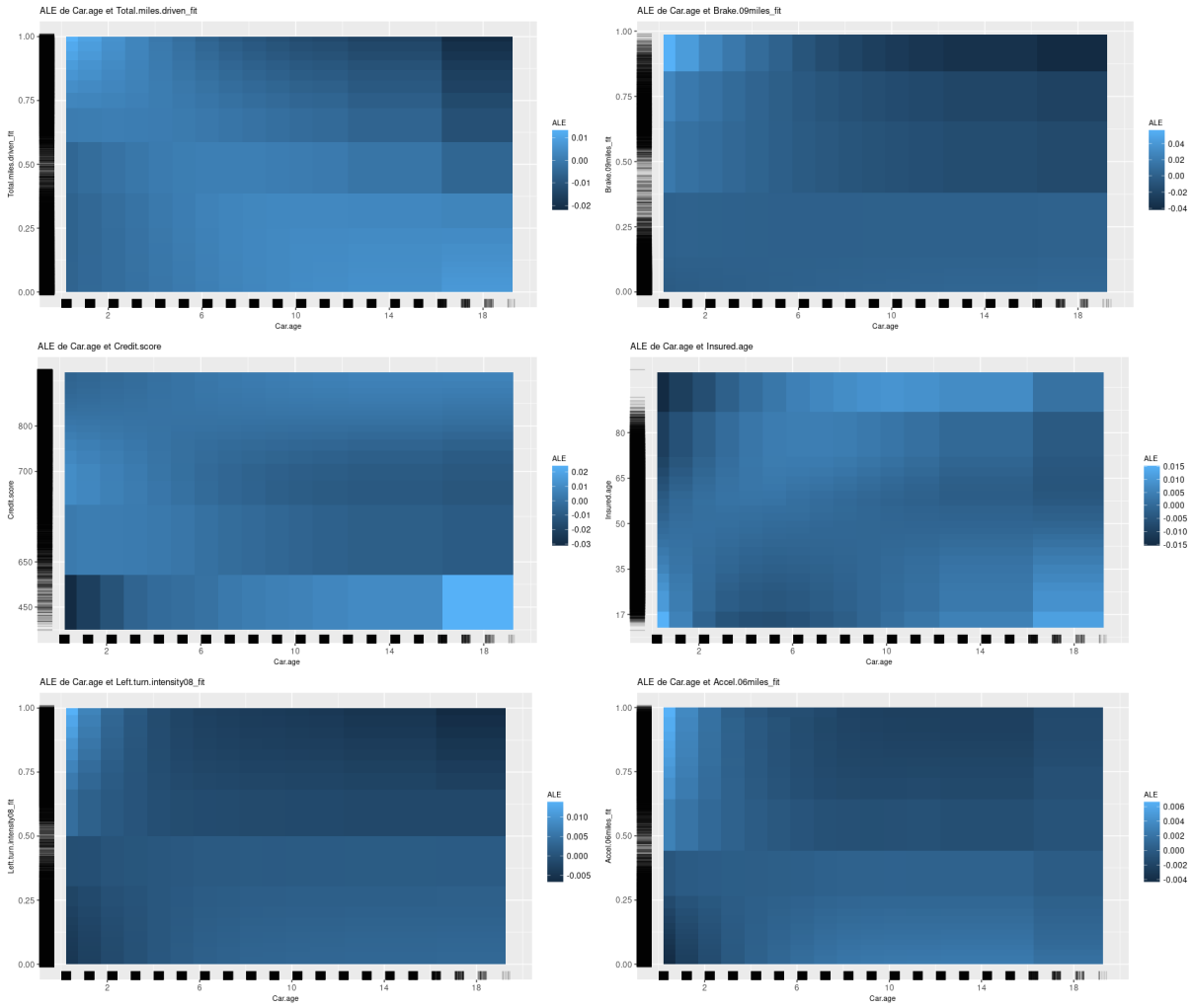


FIGURE 5.45 : *Graphiques ALE de second ordre issus du croisement entre la variable Car.age et six autres variables explicatives.*

Nous retrouvons ainsi les interprétations conformes à celles obtenues en s'appuyant sur l'analyse du graphique des splines de gradients du coefficient associé à la caractéristique *Car.age* dans notre modèle LocalGLMnet −noté $\partial_k \hat{\beta}_j$− et disponible sur la figure C.18 en annexe.

(B) Analyse locale du modèle LocalGLMnet

Dans cette partie, nous essayons de comprendre les prédictions des assurés *lambda* et *beta* présentés dans les tableaux 5.17 et 5.19 respectivement, en utilisant les outils agnostiques au modèle présenté au chapitre 4, à savoir LIME et SHAP. L'idée est d'étudier la conformité entre les interprétations basées sur modèle et celles issues de l'utilisation de ces outils indépendants du modèle.

□ LIME

Comme nous l'avons évoqué lors de la présentation de la méthode LIME au chapitre 4, son principe général consiste à approcher le modèle boite-noire par un modèle linéaire régularisée de type Lasso, au voisinage de l'instance dont on souhaite expliquer la prédiction.

Ainsi, les interprétations obtenues sont issue du modèle local construit, qui lui-même est fonction du voisinage de points échantillonnés autour de l'instance d'intérêt pour son ajustement. Pour deux voisinages distincts de l'instance d'intérêt, on peut aboutir à des explications totalement différentes.

Il est alors nécessaire de s'assurer de la stabilité des résultats avant de débuter toute quelconque interprétation.

Pour ce faire, nous avons réalisé 100 simulations LIME pour l'interprétation des fréquences prédites pour chacun de nos deux assurés d'intérêt (assurés *lambda* et *beta*). Pour chaque simulation, le voisinage échantillonné pour l'ajustement du modèle local varie aléatoirement. Les résultats LIME obtenus sont représentés par les graphiques de boîtes à moustaches de la figure 5.46 (en bas).

On note que les effets des caractéristiques sont assez stables pour chacun des deux assurés : car les boîtes à moustaches des poids locaux des différentes caractéristiques sont assez resserrées (confère figure 5.46 (en bas)).

On peut alors interpréter en toute quiétude, les résultats de LIME présentés sur les graphiques de la figure 5.46 (en haut).

• Commençons par l'assuré *lambda* :

– La quantité *"Actual prediction"* correspond à la fréquence prédite de sinistre par le modèle LocalGLMnet-Poisson. La quantité *"LocalModel prediction"* quant à elle correspond à la fréquence prédite de sinistre par le modèle de substitut local ajusté pour l'explication de la fréquence prédite de l'assuré *lambda* par le LocalGLMnet-Poisson. On observe que ces deux quantités sont assez proches l'une de l'autre : ce qui est plutôt rassurant. En effet, si *"LocalModel prediction"* était très éloignée de *"Actual prediction"*, on sous-entendrait d'office que le modèle local n'est pas assez fidèle au modèle boite-noire au voisinage de l'instance à expliquer. Par conséquent, on remettrait en question la pertinence des interprétations issues de ce modèle de substitut local.

– Les caractéristiques les plus influentes dans l'aboutissement à la fréquence prédite de sinistre de l'assuré *lambda* sont par ordre décroissant d'importance : *Lef.turn.intensity08miles_fit*, *Insured.age*, *Annual.pct.driven_fit*, *Total.miles.driven_fit* qui ont toutes une contribution positive sur la fréquence prédite de sinistre et la caractéristique *Car.age* qui a une contribution négative sur la fréquence prédite de sinistre (elle contribue à réduire la fréquence de sinistre de l'assuré *lambda*.

– Les poids respectifs des caractéristiques *Left.turn.intensity08miles_fit* et *Total.miles.driven_fit* dans le modèle de substitut local oscille autour de 0.013 et 0.0075 respectivement.

– Ainsi, si la probabilité prédite d'effectuer plus de 30 virages à gauche d'intensité 08 mph/s par 1000 miles de l'assuré *lambda* passait par exemple de 0.72 (valeur actuelle) à 0.20, autrement dit s'il adoptait promptement un comportement de conduite plus prudent en terme de nombre de virages à gauche de grande intensité, toutes choses égales par ailleurs, il réduirait fréquence prédite de sinistre de 0.03 à $0.03 + 0.013 \times (0.20 - 0.72) = 0.02$.

– Les interprétations issues de LIME sont conformes aux interprétations basées sur le modèle, en ce sens que les caractéristiques les plus importantes révélées par LIME, à savoir *Insured.age* (pour les effets positifs), et *Car.age* (pour les effets négatifs) sont également parmi celles qui ont les poids les plus importants dans la prédiction de la fréquence de sinistre de l'assuré *lambda* par le modèle LocalGLMnet mis en place (confère figure 5.39).

– On relève une ambiguïté : la caractéristique *Left.turn.intensity08miles_fit* a un poids local négatif dans le calcul de la fréquence prédite de sinistre de l'assuré *lambda* par le LocalGLMnet-Poisson (confère figure 5.39). Or, elle a un poids positif significatif dans LIME (confère 5.46 (en haut)). Ceci n'est pas très surprenant car les deux poids sont calculés sur des ensembles de données différents. Les poids de LIME sont calculer sur un "petit" voisinage de l'instance d'intérêt, alors que les poids issue du LocalGLMnet, bien que qualifiés de locaux pour une instance donnée, sont obtenus suite à l'ajustement du modèle sur l'ensemble de la base de données d'entraînement et donc incorpore implicitement un composante globale.

• En ce qui concerne l'assuré *beta*, il s'agit de l'assuré ayant la fréquence prédite de sinistre la plus élevé de notre jeu de données test.

– D'entrée de jeu, on observe une différence significative entre la fréquence prédite par le modèle LocalGLMnet-Poisson (2.77) et la fréquence prédite par le modèle de substitut local (0.25). Toutefois, cette dernière valeur reste substantiellement élevée par rapport à la prédiction moyenne du modèle LocalGLMnet-fréquence qui est d'environ 0.05. Ce modèle de substitut local rend donc bien compte du caractère hautement risqué de l'assuré *beta* relativement aux autres assurés de la base test (même si de manière absolue, il sous-estime la fréquence prédite de sinistre de l'assuré *beta*).

Commençons par analyser la stabilité des poids issus de la méthode LIME. Sur le graphique en bas et à droite de la figure 5.46, nous avons représenté les boites à moustaches des poids de caractéristiques issus de 100 simulations LIME, dans lesquelles nous avons fait varier la largeur du voisinage à chaque simulation. On observe que les poids obtenus sont assez stables car les boîtes à moustaches sont assez resserrées.

– Sur le graphique en haut à droite de la figure 5.46, on note que les variables qui influencent le plus la fréquence prédite de sinistre de l'assuré *beta* positivement sont *Total.miles.driven_fit*, *Brake09miles_fit* et *Credit.score*. Par ailleurs, seule la variable *Insured.age* contriburait à réduire la fréquence prédite de sinistre de l'assuré *beta*.

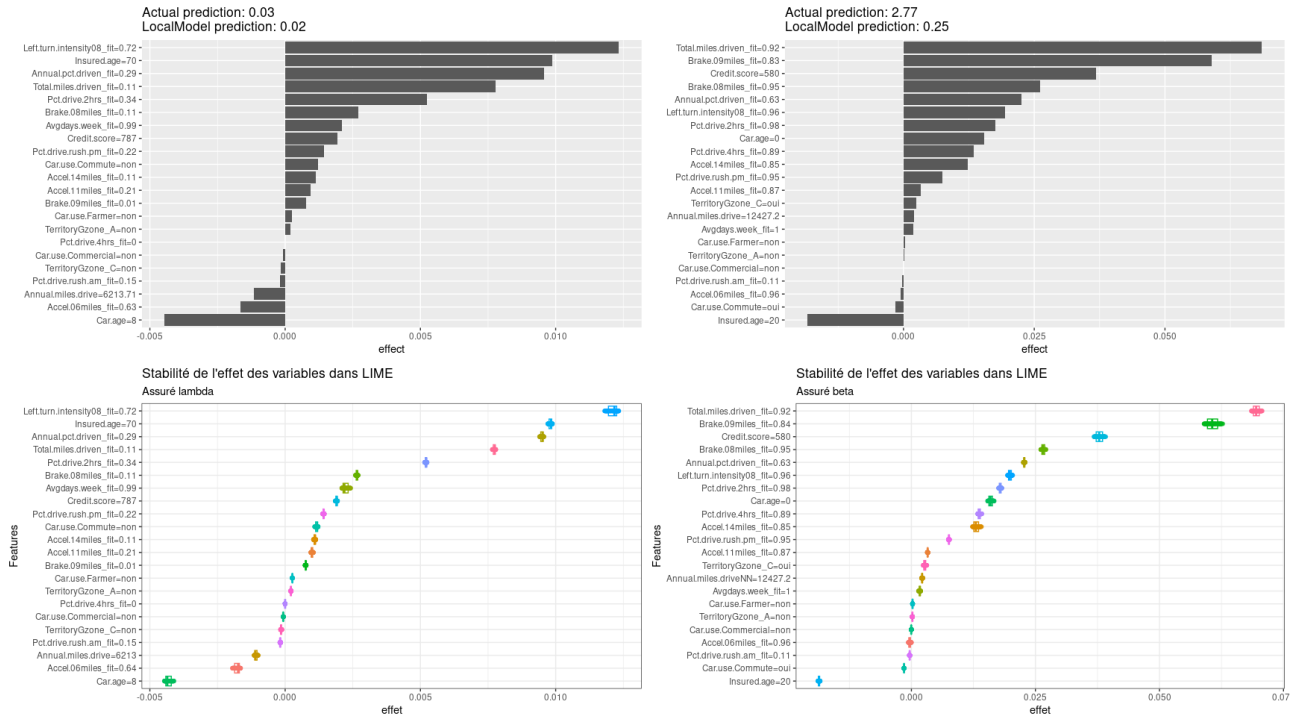


FIGURE 5.46 : *Explications LIME des prédictions de nos assurés lambda et beta dans le modèle LocalGLMnet-Poisson : assuré lambda (en haut, à gauche), et assuré beta (en haut, à droite) ; Analyse de la stabilité de l'effet des variables dans le modèle de substitut local pour 100 simulations LIME en faisant varier la largeur des voisinages à chaque itération, pour l'assuré lambda (en bas, à gauche) et l'assuré beta (en bas, à droite).*

□ SHAP

A présent, passons à l'explication des prédictions des assurés *lambda* et *beta* en utilisant cette fois la méthode SHAP.

La méthode SHAP a été présenté au chapitre 4. Contrairement à l'approche LIME, SHAP ne base pas ses explications sur un modèle de substitution locale, mais plutôt sur une répartition équitable de la valeur de la fréquence prédite de sinistre de l'assuré en termes de contribution des différentes caractéristiques à la formation de la fréquence prédite.

Tout comme la méthode LIME, la méthode SHAP utilise une perturbation du jeu de données initiale, par la méthode de Monte Carlo pour générer des explications locales.

Les résultats obtenus à l'aide du package *iml* de R sont donnés sur les figures 5.47. Les graphiques en dessous de cette figure correspondent à l'analyse de la stabilité des contributions obtenues à l'issue de 200 simulations d'explications de la même instance (*lambda* ou *beta*). Sur ces graphiques, on constate que les boîtes à moustaches des différents graphiques sont plus ou moins dispersés (relativement aux résultats de LIME). Cependant, si l'on utilise la *médiane* comme critère d'ordonnancement de l'importance des différentes caractéristiques, on cremarque que l'ordre d'importance des caractéristiques est assez bien établi, pour les deux assurés.

En outre, on constate un fait notable : les explications issues de SHAP ne sont pas conformes à celles issues de LIME pour l'assuré *lambda*. Ceci n'est pas nécessairement absurde, car les deux méthodes fonctionnent différemment et peuvent éventuellement aboutir à des résultats différents.

• Commençons par l'assuré *lambda*.

Pour celui ci, les explications issues de SHAP paraissent plus cohérentes que celles issues de LIME, quoi que légèrement plus instables. Ceci proviendrait du fait que LIME étant basée sur un modèle linéaire, il ne prend pas minutieusement en compte les potentiels effets d'interaction entre les caractéristiques. Or, comme nous l'avons analysé plus haut, notre modèle LocalGLMnet-Poisson contient plusieurs interactions entre les caractéristiques.

La supériorité de SHAP sur LIME dans ce cas précis n'est pas généralisable : il existe bien de situations dans lesquelles LIME fournir des explications plus cohérentes que SHAP.

D'après les résultats de la méthode SHAP (confère figure 5.47 en haut à gauche), la caractéristique qui contribue le plus à réduire la fréquence prédite de sinistre de l'assuré *lambda* est sa probabilité prédite de parcourir plus de 4500 miles de route durant l'année (*Total.miles.driven_fit*). Cette dernière est relativement faible pour cet assuré (elle vaut précisément 0.11), ce qui sous-entend que l'assuré *lambda* roule relativement peu durant l'année, ce qui limite logiquement sa fréquence prédite de sinistre.

En outre, toujours d'après SHAP, l'âge avancé de l'assuré *lambda* (70 ans) représentée par la variable *Insured.age* fait partie des caractéristiques ayant la plus grande contribution positive à sa fréquence prédite de sinistres. Ce qui est cohérent : car les plus âgés sont plus enclin à avoir un nombre élevé de sinistres dans dans notre base de données.

• Passons à présent à l'interprétation de la fréquence prédite pour l'assuré *beta*.

Pour cet assuré, nous observons une certaine conformité entre les interprétations issues de SHAP et LIME. D'après le graphique en haut et à droite de la figure 5.47, on note que :

Les caractéristiques qui sont les plus influentes positivement sur la fréquence prédite de sinistre de cet assuré sont : le score de crédit (*Credit.score*), qui est relativement bas pour l'assuré *beta* (580 sur 900 points) ; ensuite ses fortes probabilités prédites d'effectuer plusieurs accélérations et freinages brusques à grandes intensités par 1000 miles (*Brake.09miles_fit*, *Brake.09miles_fit*, *Accel.14miles_fit*) ; sans oublier sa forte probabilité prédite de parcourir plus de 4500 miles de route durant la période de couverture (0.92), ce qui est intuitevement logique, car cela correspond bien à des comportements de haut risque en conduite automobile.

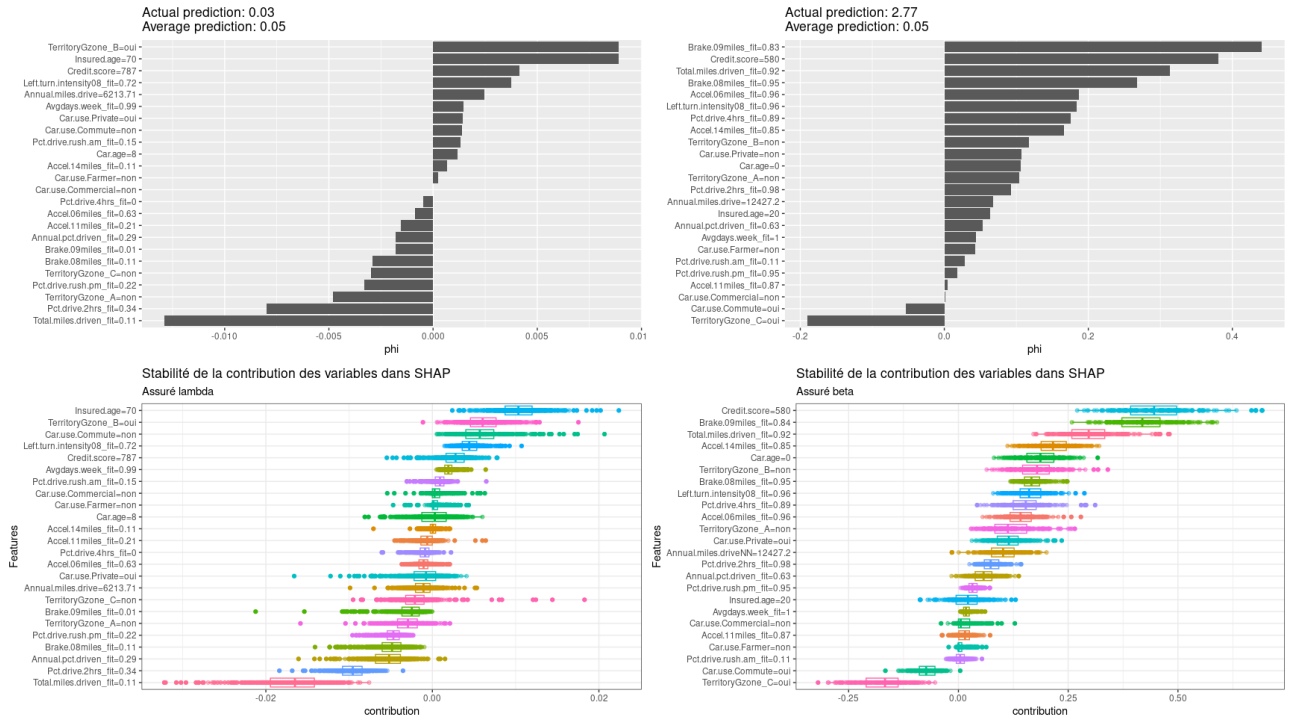


FIGURE 5.47 : *Explications SHAP des prédictions de nos assurés lambda et beta dans le modèle LocalGLMnet-Poisson : assuré lambda (en haut, à gauche), et assuré beta (en haut, à droite) ; Analyse de la stabilité de la contribution locale des variables pour 200 simulations SHAP, pour l'assuré lambda (en bas, à gauche) et l'assuré beta (en bas, à droite).*

5.5.3 Une attaque à la fiabilité éthique de la méthode LIME

Comme nous l'avons évoqué dans le chapitre 3, l'une des principales raisons de la recherche de l'interprétabilité est *la prise de décision juste et éthique*.

De nouvelles réglementations pour la protection des données ont vus le jour dans l'Union européenne (RGPD). Elles proposent que les individus affectés par des décisions algorithmiques aient un droit à l'explication des décisions prises (Goodman et Flaxman (2017)).

Dans le contexte spécifique de l'assurance en France, il existe d'autres contraintes réglementaires liées à l'utilisation des modèles d'apprentissage automatique pour la prise de décision juste et éthique. L'Autorité de Contrôle Prudentiel et de Résolution (ACPR) veille à la protection des consommateurs et exige que l'assureur puisse justifier de manière détaillée et exacte toutes les décisions prises pour le calcul d'une prime d'assurance. Les tarifs doivent être explicites, notamment pour éviter le risque de discrimination.

Comment pouvons-nous être sûr qu'une prédiction basée sur un modèle complexe d'apprentissage statistique n'est pas discriminatoire sur la base de la race, du sexe, de l'appartenance religieuse, etc ? Pour répondre à cette problématique, les outils d'interprétabilité présentés au chapitre 4 tels que LIME et SHAP peuvent être utilisées. Dans les sous-sections précédentes nous avons eu l'occasion de définir et d'illustrer longuement leur fonctionnement. Malheureusement, ces outils d'interprétabilité, quoique pertinents et précis, ils sont peu fiables, car étant falsifiables.

Plus précisément, étant donné un modèle biaisé, c'est-à-dire contenant des variables discriminatoires (au sens juridique du terme), nous pouvons masquer efficacement les biais discriminatoires de sortes que les explications LIME et SHAP ne puissent pas les détecter lors des interprétations locales.

Dans le cadre de la présente étude, une caractéristique potentiellement discriminatoire serait par

exemple le score de crédit (*Credit.score*).

Afin de mettre en place la falsification des interprétations issues des méthodes LIME et SHAP, nous procédons comme suit :

• Premièrement, nous énumérons quelques raisons pour lesquelles le score de crédit peut être perçu dans notre étude comme discriminatoire ;

• Puis, nous présenterons une procédure par laquelle l'effet du score de crédit peut être masqué dans les explications générées par LIME de la fréquence prédite de sinistre (par notre modèle LocalGLMnet-Poisson) d'un assuré quelconque de notre base de données test. Nous mettons en oeuvre cette procédure de falsification pour modifier les explications LIME de la fréquence prédite des assurés *lambda* et *beta*.

5.5.3.1 Discussion autour du caractère éthique de la variable *Credit.score* en tarification automobile

Depuis près de trois décennies, de nombreux travaux de recherche menés en assurance automobile ont permis d'établir scientifiquement une liaison entre le pointage de crédit d'un assuré et sa probabilité de subir un sinistre, donc de présenter des réclamations.

D'après les études de Brockett et Golden (2007) des attributs psychologiques et comportementaux −par exemple un type de personnalité en quête de sensations fortes− sont communes aux conducteurs automobile présentant à la fois des coûts de sinistres plus élevés et des cotes de crédit plus faibles.

Pour résumer, ces auteurs établissent un lien entre "être un mauvais conducteur" et "avoir un faible score de crédit". En 2016, ces mêmes chercheurs vont plus loin dans leurs analyses et montrent que la corrélation observée entre le score de crédit et le nombre de réclamation n'est pas dû au chevauchement avec les autres variables de souscription existantes, mais que les cotes de crédit contiennent des informations importantes qui ne sont pas déjà intégrées dans d'autres variables de notation traditionnelles (telles que l'âge, le sexe, les antécédents de conduite, etc.).

Leurs observations sont bien conformes à ceux que nous avons constaté dans notre étude, lors de l'analyse des modèles de fréquence de sinistres mis en place. Nous avons observé par diverses procédures que la variable score de crédit (*Credit.score*) figurait parmi les variables les plus importantes dans l'ajustement de la fréquence de sinistre, aussi bien à l'échelle globale qu'au niveau des prédictions individuelles dans les modèles GLM-Poisson et LocalGLMnet-Poisson (confère figures 5.29, 5.40 pour l'importance globale, 5.46 et les figures 5.47 pour l'importance locale au niveau des prédictions des assurés *lambda* et *beta*).

En pratique, la prise en compte du score de crédit dans le calcul de la prime automobile permet souvent de faire diminuer substantiellement la prime d'assurance automobile lorsque le dossier de crédit du souscripteur est bon ou excellent (pointage de crédit supérieure ou égale 700/900). L'inverse est aussi vrai : un mauvais dossier de crédit (ceux ayant une note de crédit inférieure à 600 points/900) peut parfois payer deux fois, même trois fois plus cher son automobile.

Si en pratique les assurés ayant un excellent dossier de crédit − c'est à dire ceux pour lesquels la prime n'est pas pénalisée par leur pointage de crédit− sont favorables à l'utilisation du score de crédit dans le processus de tarification, ce n'est pas le cas pour les assurés ayant un mauvais score de crédit, car ils se verraient doublement pénalisés.

En effet, dans certains contextes l'évaluation du risque de crédit est jugée au moins en partie discriminatoire sur certaines caractéristiques sensibles telle que la couleur de la peau (confère Rice et Swesnik (2013)). Ainsi, la prise en compte du score de crédit dans le calcul de la prime d'assurance pourrait être perçue comme un substitut euphémique de l'appartenance ethnique ou raciale.

Kiviat (2019) explique de manière beaucoup plus détaillée le caractère sensible de cette variable

score de crédit dans le contexte de la tarification en assurance automobile. Elle étudie l'effet causal de la variable score de crédit sur le comportement de conduite des assurés. Elle explique que le score de crédit bien qu'ayant un effet causal sur le comportement de conduite, il est également lié au statut socio-économique, lui-même lié à la couleur de peau, donc problématique.

L'assureur se trouve donc en face d'un dilemme : il dispose d'une variable assez significative dans la modélisation du risque de sinistralité de l'assuré, mais assez sensible d'un point de vue éthique. Il peut alors être tenté de maintenir cette variable dans son modèle pour affiner ses tarifs. Cependant, en cas de recours à l'explication du calcul de la prime d'un assuré par la méthode LIME, par exemple, il pourrait masquer l'influence de cette variable sensible dans le modèle de tarification.

5.5.3.2 Étapes de la mise en oeuvre de l'attaque des explications LIME

□ Intuition de la démarche adoptée

Afin de truquer les explications locales des fréquences prédites de sinistres obtenues par la méthode LIME, nous nous inspirons de la démarche méthodologique présentée dans l'article de Slack *et al.* (2020). Nous reprenons leurs notations.

La procédure peut se résumer en les cinq (05) étapes suivantes :

1− Premièrement, nous construisons le modèle biaisé f (modèle avec la variable sensible). Dans notre contexte f correspond à notre modèle LocalGLMnet-Poisson. Il est qualifié de "biaisé" parce qu'il contient une variable sensible (ici, le score de crédit $-$*Credit.score*), qui est potentiellement discriminatoire.

2− Ensuite, nous générons des perturbations pour chaque instance de données dans l'ensemble de données d'entraînement $\mathcal{X}$ de notre modèle biaisé f en ajoutant un bruit aléatoire échantillonné suivant une loi $\mathcal{N}(0,\, 0.05^2)$, et ce, pour chaque caractéristique.

Nous obtenons une nouvelle base de données perturbées, de même dimension que $\mathcal{X}$, que nous notons $\mathcal{X}_p$. Ce choix perturbation, est similaire à celui qu'utilise la méthode LIME du package *DALEX* de *R* pour se générer un voisinages de points de l'instance à expliquer (mais avec plutôt $\mathcal{N}(0,\, 1)$).

À partir de $\mathcal{X}$ et $\mathcal{X}_p$, nous constituons une nouvelle base de données $\mathcal{X} \bigcup \mathcal{X}_p$. Dans cette nouvelle base de données, nous créons une nouvelle variable nommée OOD (*out-of-distribution* , en anglais), qui prend la valeur 1 pour toutes les observations perturbées ($\mathcal{X}_p$) et 0 pour les observations originales ($\mathcal{X}$).

3− Nous ajustons ensuite un classifieur sur $\mathcal{X} \bigcup \mathcal{X}_p$, avec pour variable cible OOD et pour prédicteurs, l'ensemble des variables explicatives du modèle f. La finalité de ce prédicteur est de détecter les observations obtenues suite aux perturbations, des observations réelles ou originales.

Comme l'illustre la figure 5.48, à l'aide d'une analyse en composante principale, on observe que les données synthétiques obtenues après perturbation − en adoptant la méthodologie décrite à l'étape précédente− sont globalement significativement différents des données originales, de sorte qu'un classifieur même simple de type CART par exemple, puisse permettre de bien séparer les données synthétiques des originales. Dans notre cas, nous avons mis en place un classifieur de type modèle Random Forest pour effectuer cette tâche de détection des données synthétiques des données originales. Nous désignons ce classifieur par : is_ODD. Il nous renvoie la probabilité qu'une instance fixée de $\mathcal{X} \bigcup \mathcal{X}_p$ soit issue d'une pertubaliton aléatoire.

4− Nous calibrons enfin notre modèle non biaisé, noté ψ, sur la base de données d'entraînement originale $\mathcal{X}$, dans laquelle les valeurs de la variable sensible *Credit.score* sont tous remplacer par de nouvelles valeurs issues d'une loi uniforme sur un intervalle arbitraire (dans notre cas nous avons opté

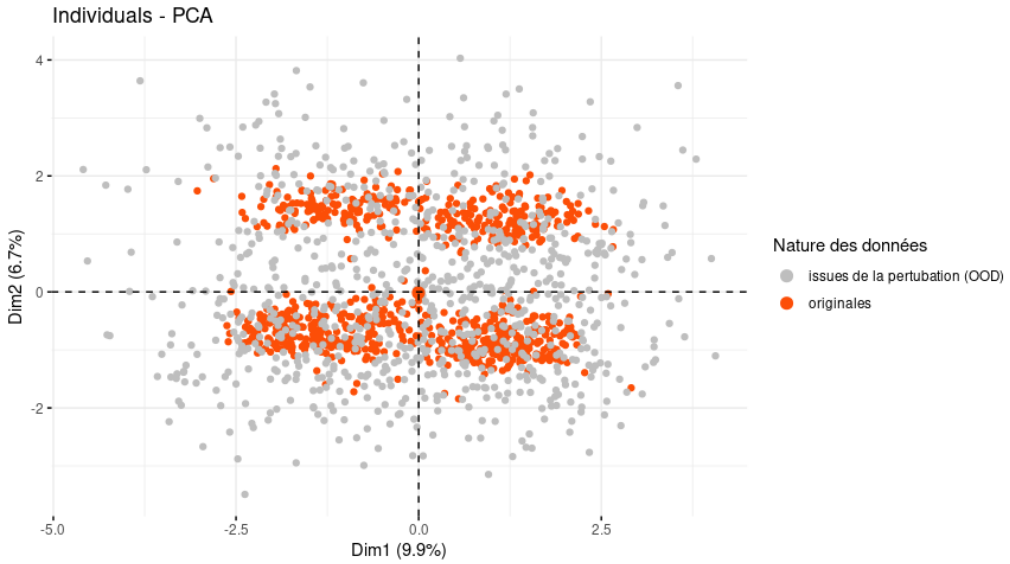


FIGURE 5.48 : *ACP appliquée à un échantillon réduit de taille 833 (réduit uniquement pour des raison de lisibilité du graphe) de notre jeu de données d'entraînement, $\mathcal{X}$ (en rouge) ainsi qu'à ses perturbations de type LIME, $\mathcal{X}_p$ (gris). Même dans cet espace à juste deux dimensions, nous pouvons bien voir que les points de données générés par les perturbations sont distribués très différemment des instances des données originales* .

pour une uniforme sur [50, 60]), de manière à être parfaitement décorrelée de la variable *Credit.score* originale.

5− Enfin, nous obtenons facilement notre modèle adverse (aussi surnommé *"cheval de Troie"*). Nous le notons e et il est défini par :

$$e(x) = \begin{cases} f(x), \text{ si } is\_OOD(x) \leq 0.8 \\ \psi(x), \text{ si } is\_OOD(x) > 0.8 \end{cases} \tag{5.1}$$

C'est le modèle e qui sera déployé en industrie pour le calcul des primes. e se comportera exactement comme f lorsqu'il fera des prédictions sur des instances échantillonnées suivant la distribution réelles (X), cependant, il ne révélera pas les biais sous-jacents de f lorsqu'il sera interrogé avec une méthode d'explication locale post hoc telles que LIME. En effet, lorsque e sera interrogé par LIME, il générera les explications en s'appuyant sur le modèle non biaisé ψ, plutôt que sur le modèle biaisé f.

□ Illustration de l'attaque des explications LIME pour les assurés *lambda* et *beta*

Dans la sous-section 5.5.2 nous avons eu à réaliser les explications LIME de la fréquence prédite de sinistre des assurés *lambda* et *beta* par notre modèle LocalGLMnet-Poisson (confère figure 5.46).

D'après ces explications, la caractéristique *Credit.score* figurait parmi les caractéristiques contribuant le plus à augmenter la fréquence prédite de sinistre chez les assurés *lambda* et *beta*, avec un effet plus accentué chez l'assuré *beta*, ayant un score de crédit faible (avec un pointage de seulement 580 points sur 900).

Cependant, observons de nouveau les explications LIME de ces assurés à l'issue de l'attaque des explications. Les résultats obtenus sont représentés par les graphiques de la figure 5.49. On relève deux faits notables :

• Sur l'entête des graphiques d'au-dessus correspondant aux nouvelles explications LIME des fré-

quences prédites de sinistre des assurés *lambda* (à gauche) et *beta* (à droite), on observe que le modèle adverse e fournit les mêmes prédictions de fréquence pour les deux assurés que celles du modèle biaisé f (0.03 pour le premier assuré et 2.77 pour le second), ce qui est cohérent avec la théorie ;

On relève une différence au niveau des prédictions fournies par les modèles de substitution locale (0.06 vs 0.02 pour l'assuré *lambda* et 0.24 vs 0.25 pour l'assuré *beta*). Cette différence est logique, car les explications locales sont générées à partir du modèle non biaisé ψ et non plus sur le modèle original, biaisé f.

• La remarque la plus importante est que pour l'assuré *beta*, celui pour lequel la caractéristique sensible *Credit.score* posait le plus de soucis −de par sa grande influence positive dans la prédiction de la fréquence de sinistre−, on observe clairement que dans les nouvelles explications LIME truquées, son poids affiché est quasiment nul.

Il s'agit en réalité du poids de cette caractéristique dans le modèle non biaisé ψ, où la caractéristique sensible n'intervient pas en réalité.

Ainsi, ni le régulateur, ni l'assuré *beta*, ne pourra être en mesure de détecter l'influence du score de crédit dans sa fréquence prédite de sinistre à partir des explications LIME, pourtant elle continue bien d'y intervenir.

Il en est de même pour l'assuré *lambda* et il en sera ainsi pour l'ensemble des nouveaux assurés pour lesquels on essayera d'avoir les explications des prédictions faites à partir du modèle adverse e : le poids attribué à la caractéristique *Credit.score* par la méthode LIME restera quasiment nul.

Afin de s'assurer de la stabilité des nouvelles explications LIME, nous avons réalisées 200 simulations LIME en modifiant à chaque itération la largeur du voisinage utilisé pour les explications locales. Les graphiques en dessous de la figure 5.49 représentent les boites à moustaches des poids des différentes caractéristiques obtenus après les 200 simulations. On observe que les boites à moustaches obtenues sont assez resserrées, ce qui indique la stabilité des explications LIME truquées, avec un poids attribué à la caractéristique *Credit.score* toujours quasiment nul, pour les assurés *beta* et *lambda*. Ces résultats démontrent que la technique d'explication LIME a été efficacement trompée par le classificateur adverse e.

En adoptant une procédure similaire, à quelques ajustement près, il est également possible de truquer les explications issues de la méthode SHAP. Globalement, il est possible de truquer les explications issues de toutes les méthodes d'interprétation locale *post hoc* basée sur une perturbation des caractéristiques de l'instance d'intérêt, en adoptant une démarche similaire à celle de la méthode LIME présentée ci-dessus.

5.6 Ingénierie des caractéristiques

Dans cette section, nous souhaitons terminer par une optimisation de notre modèle GLM-Poisson. Pour cela, nous faisons recours à de l'ingénierie des caractéristiques basée sur les résultats issues de l'interprétation du modèle "hybride" LocalGLMnet. Concrètement, il s'agit de s'inspirer des relations apprises dans le modèle sophistiqué LocalGLMnet-Poisson, pour enrichir le modèle GLM-Poisson et accroître par la même occasion sa performance prédictive.

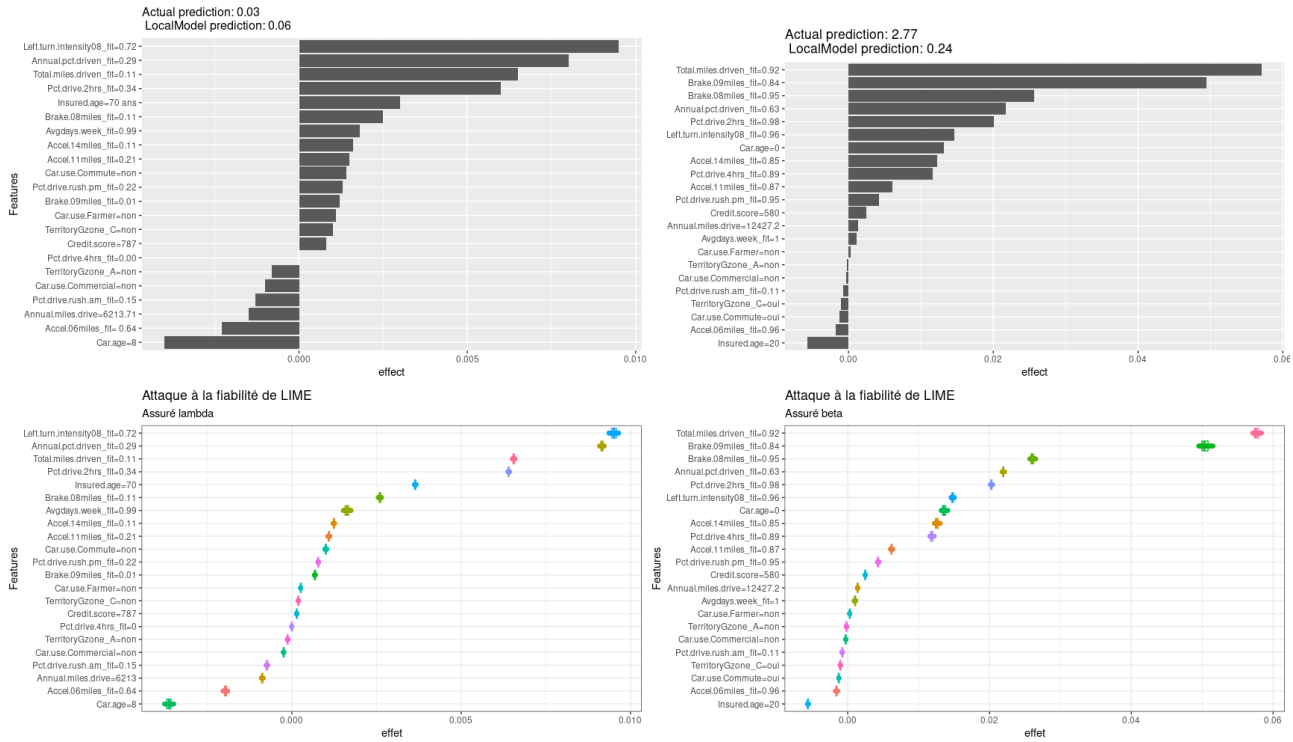


FIGURE 5.49 : *Attaque à la fiabilité des explications LIME pour les assurés lambda (à gauche) et beta (à droite). Les graphique en dessous correspondent aux boites à moustaches des poids des caractéristiques obtenues après 200 simulations LIME, pour l'assuré lambda (en bas, à gauche) et pour l'assuré beta (en bas, à droite).*

5.6.1 Intégration des relations non-linéaires et d'interaction détectées dans le modèle LocalGLMnet-Poisson au modèle GLM-Poisson initial

Les méthodes d'interprétation *post hoc* présentées au chapitre 4 permettent de détecter les régions d'interaction entre les différentes caractéristiques. Cependant, elles ne nous fournissent pas les détails sur la forme de la relation d'interaction : d'où l'un des principaux intérêts du modèle LocalGLMnet, qui en plus de prendre en compte les potentielles interactions entre les caractéristiques lors de son ajustement, nous révèle les formes de celles-ci, à l'issue des interprétations basées sur le modèle (confère figures 5.38 et C.18).

Toutefois, l'intérêt de faire recours aux interprétations *post hoc* du modèle LocalGLMnet est que cela nous a permis de vérifier la convergence des interprétations basées sur le modèle et des interprétations *post hoc*.

En ce qui concerne la détection des effets non-linéaires des caractéristiques sur la variable cible (effet polynomial, par exemple), nous nous inspirons également des résultats issus de l'interprétation *post hoc* et de l'interprétation intrinsèque du modèle LocalGLMnet-Poisson (confère figures 5.41, 5.42 et C.18).

Les interactions qui seront augmentées sont celles relatives aux variables *Car.age*, *Insured.age*, *Credit.score*, *Total.miles.driven_fit*, et qui ont été identifiées grâce aux tracés des graphiques de gradients des poids d'attentions dans le modèle LocalGLMnet (confère figure C.18) et synthétisé dans la sous-section 5.5.2.

Étant donné le nombre élevé de potentielles relations d'interaction significatives à prendre en

compte, nous ajusterons le nouveau modèle GLM à l'aide d'une régularisation de type LASSO présentée au chapitre 2. L'objectif est de ne retenir que les facteurs d'interactions les plus significatifs dans la prédiction de la fréquence de sinistre.

Dans le nouveau modèle GLM-Poisson avec interactions intégrées qui est mis en place, les variables numériques qui avaient initialement été regroupées en catégories sont utilisées en leur état numérique (étant donné que nous prenons en compte les termes polynomiaux dans le nouveau modèle).

Commençons par déterminer le paramètre de pénalisation λ optimal dans notre modèle de régularisation LASSO. Nous procédons par validation croisée. Les résultats obtenus à l'aide de la fonction *glmnet* du package *glmnet* de *R* sont présentés sur la figure 5.50. Le graphique en haut et à gauche nous donne l'évolution de la valeur des poids des différentes caractéristiques suivant les valeurs de λ. Les graduations de l'axe au dessus de ce graphique nous indique le nombre de caractéristique ayant un poids non nul, pour une valeur de $\ln(\lambda)$ donnée.

Pour $\ln(\lambda) = -10$, le modèle régularisé sélectionne 36 caractéristiques. Avec ce nombre relativement élevé de prédicteurs, on pourrait s'attendre à un meilleur ajustement du modèle sur la base d'entraînement, mais au détriment de la parcimonie (perte en interprétabilité) et d'un éventuel surapprentissage. Par contre, pour $\ln(\lambda) = -4$, le modèle sélectionne juste 3, il est donc parcimonieux, donc facilement interprétable, cependant, il contient trop peu de variables pour bien modéliser la fréquence de sinistre. Quelle valeur de λ faudrait-il donc choisie, de manière à optimiser simultanément la performance prédictive et descriptive de notre modèle ?

La réponse à cette question est donnée par le graphique en haut et à droite de la figure 5.50. Sur ce graphique, la bande blanche autour de la ligne rouge correspond aux intervalle de confiance de la déviance du modèle, suite à une procédure de validation croisée. On retient la valeur de $\ln(\lambda)$ à partir de laquelle la déviance du modèle explose pour la première fois : c'est à dire $\ln(\lambda) = -6$. Pour ce choix de $\ln(\lambda)$, on retient 18 caractéristiques dans notre modèle GLM-Poisson enrichi. Les caractéristiques sélectionnées ainsi que leur poids sont représentées sur le graphique en dessous de la figure 5.50.

Étant donné que dans le modèle LASSO, les variables sont au préalable remises à la même échelle par une opération standardisation, nous pouvons comparer les coefficients des variables entre-eux et identifier les variables ayant l'effet le plus important dans le nouveau modèle GLM-Poisson enrichi. On constate bien que ce sont les termes polynomiaux des caractéristiques $Isured.age$, $Total.miles.driven\_fit$ et $Brake.09miles\_fit$ qui ont les effets les plus significatifs ; ensuite viennent les termes d'interaction linéaire entre la variable $Avgdays.week\_fit$ et
$Total.miles.driven\_fit$, puis $Avgdays.week\_fit$ et $Annual.pct.driven\_fit$.

Nous pouvons enfin visualiser graphiquement l'effet de l'ajout des effets non linéaires et d'interaction des caractéristiques dans le modèle GLM-Poisson de départ.

La figure 5.51 illustre la différence de la forme d'interaction entre les caractéristiques
$Avgdays.week\_fit$ et $Total.miles.driven\_fit$ dans le modèle GLM-Poisson initial et le modèle GLM-Poisson avec interactions et non linéarité intégrées. On voit bien que dans le nouveau modèle GLM complexifié, l'interaction entre ces deux variables est mieux prise en compte, avec un effet amplifié pour les individus ayant à la fois une grande probabilité de rouler plus de 4500 miles de distance dans l'année et une grande probabilité de rouler en moyenne, plus de 4 jours sur 7 durant la semaine. Par contre, dans le modèle initial, l'interaction est inexistante entre ces deux variables, car le plan de réponse ne présente aucune courbure.

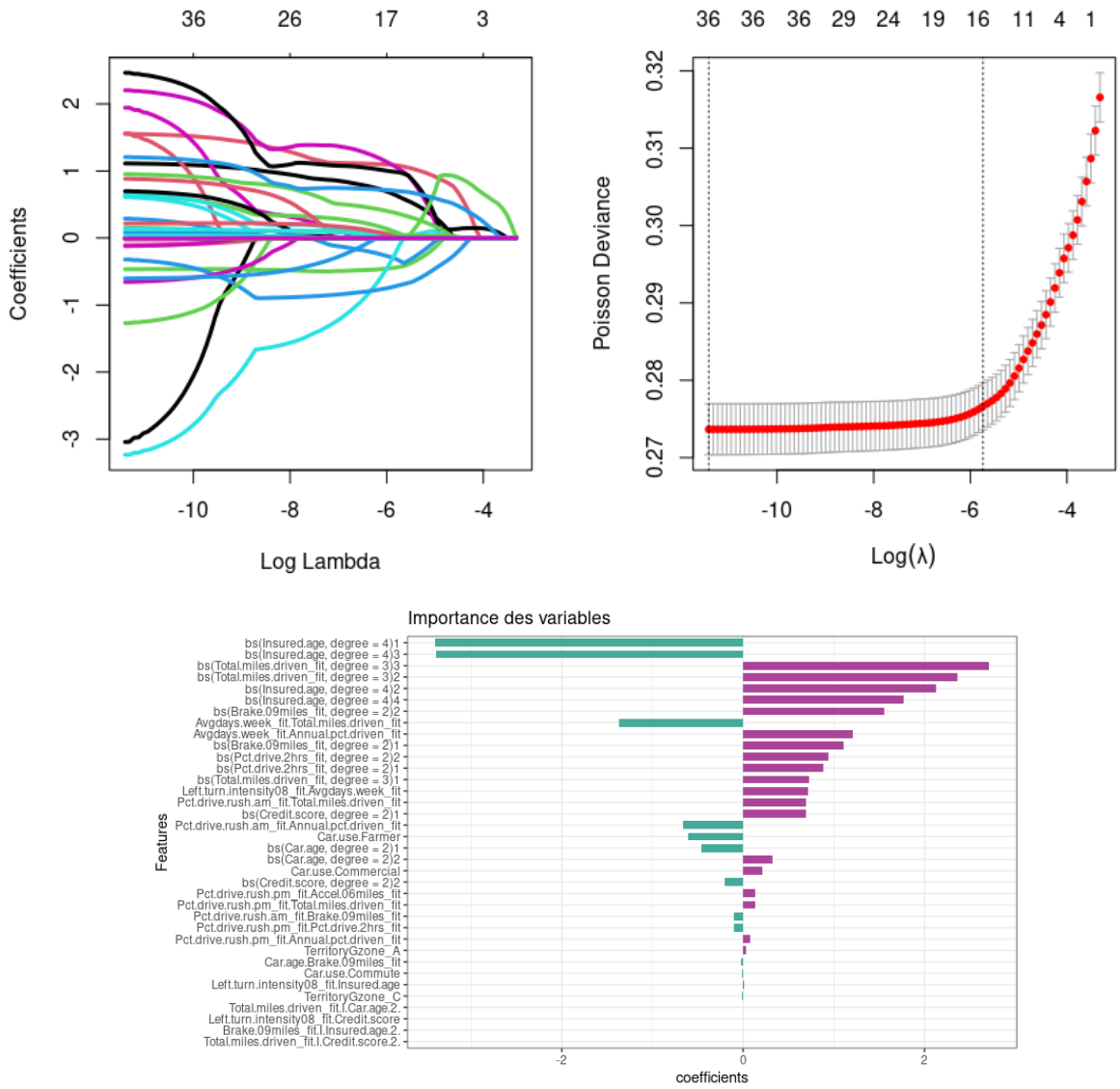


FIGURE 5.50 : *Choix optimal du paramètre de pénalisation λ dans la régularisation LASSO (graphiques du haut) ; Résultats des variables retenues par le modèle Lasso pour le* $\ln(\lambda)$ *optimal (en bas).*

5.6.2 Comparaison de la performance prédictive du GLM-Poisson initial et du GLM-Poisson avec effets non-linéaire et d'interaction ajoutés

5.6.2.1 Qualité d'ajustement du modèle GLM-Poisson avant et après ajout des effets d'interaction et de non linéarité

Avant d'effectuer la comparaison de la précision prédictive du modèle GLM-Poisson initial et du modèle GLM-Poisson complexifié, il est intéressant de comparer avant tout leur qualité d'ajustement. À cet effet, sur la figure 5.52 nous représentons le nuage des fréquences prédites de sinistres des assurés de notre base d'entraînement en fonction des résidus obtenus. Le modèle le mieux ajusté correspond à celui dont le nuage est le plus dispersé ; c'est-à-dire celui pour lequel il n'existe pas de tendance apparente entre les valeurs prédites et les résidus. Sous ce critère, c'est le modèle GLM complexifié qui

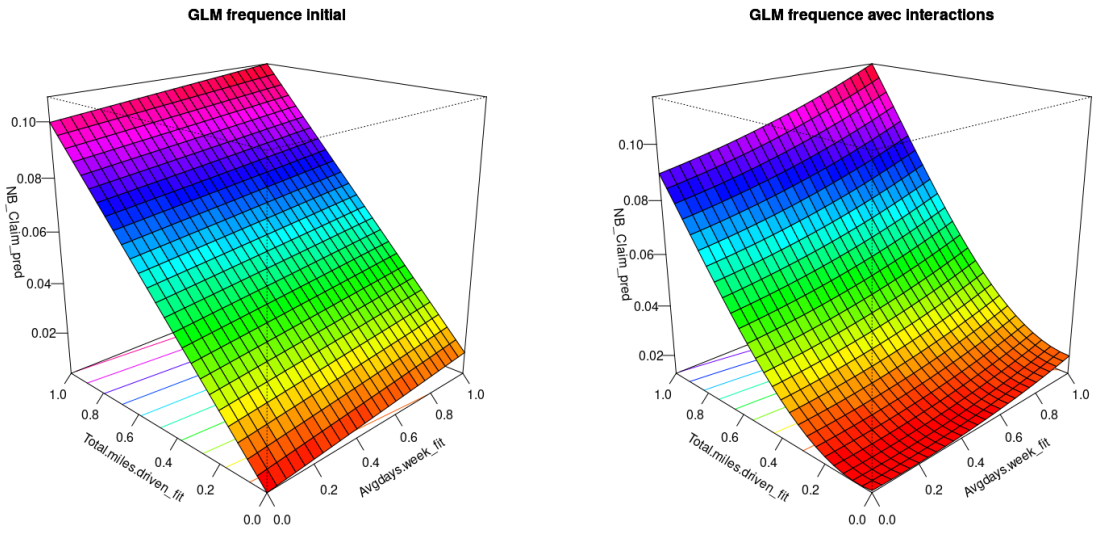


FIGURE 5.51 : *Illustration de la différence de l'effet d'interaction entre les variables sur la fréquence prédite, dans un modèle GLM fréquence initial (à gauche), et dans le modèle GLM fréquence complexifié (à droite).*

semble avoir le nuage le plus dispersé. Ce dernier est donc considéré comme celui qui s'ajuste le mieux aux données d'entraînement.

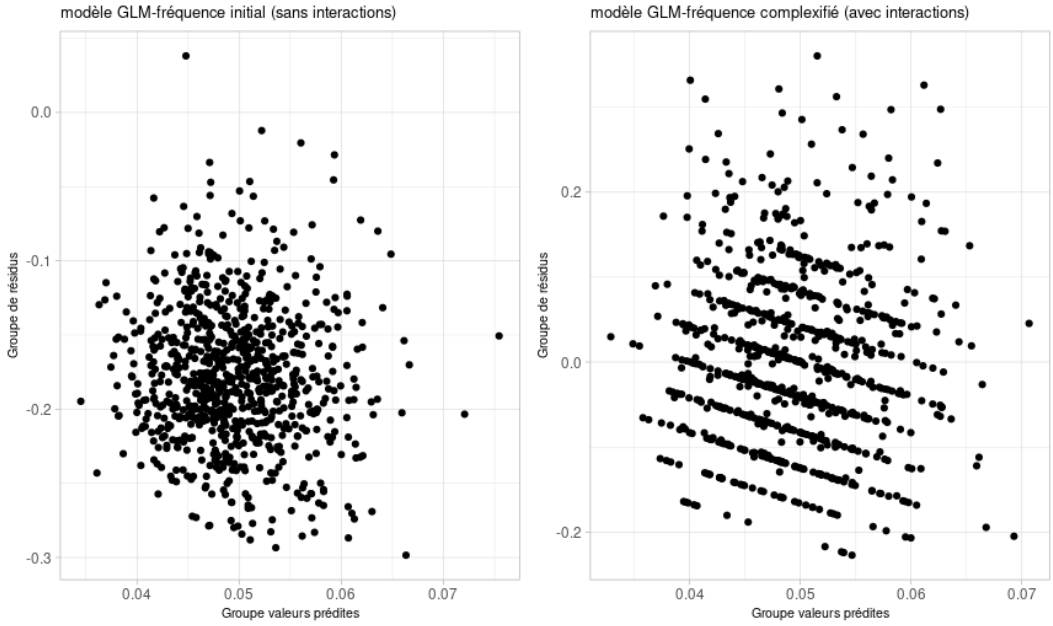


FIGURE 5.52 : *Nuages des résidus vs. prédictions pour le modèle GLM fréquence initial (à gauche) et pour le modèle GLM fréquence complexifié (à droite).*

Un autre critère évoqué plus haut, pour comparer la qualité d'ajustement de deux modèles linéaires généralisés est la *Déviance résiduelle* du modèle, plus elle est faible mieux c'est. Pour le modèle GLM-Poisson initial nous avons une *Déviance résiduelle* de l'ordre de 22 704 et pour le nouveau modèle GLM-Poisson enrichi, la *Déviance résiduelle* s'élève à 22 712.83. Selon ce critère c'est le modèle GLM initial qui serait très marginalement le mieux ajusté. La prise en compte des effets d'interaction et

des effets non-linéaires des caractéristiques n'a pas eu un impact substantiel dans l'amélioration de la déviance du modèle initial. Au contraire, elle a dégradé (de très peu) la déviance du modèle initial.

5.6.2.2 Comparaison de la performance prédictive du modèle GLM-Poisson avant et après ajout des effets d'interaction et de non-linéarité

Après avoir ajusté notre modèle GLM-Poisson avec interactions et non-linéarité augmentées, nous nous en sommes servi pour réaliser les prédictions de fréquence de sinistre des assurés de notre base test, mise de côté dès le départ.

Les résultats des métriques calculées sont disponibles dans le tableau 5.20. Dans ce tableau, dans une optique d'analyse de la plus value de l'intégration des interactions dans la modélisation de la fréquence de sinistre, nous avons également repris les résultats obtenus pour le modèle GLM-Poisson initial du tableau 5.15.

On observe très clairement que le gain apporté par la prise en compte des interactions dans le modèle GLM-Poisson est infime, voir inexistant. En effet, la métrique que l'on a retenu pour la comparaison des deux modèles est la RMSE, et le gain de RMSE en comparaison du modèle GLM-Poisson initial est de l'ordre de 0.15 %. En plus du gain négligeable sur la RMSE, le modèle GLM-Poisson complexifié réalise une perte d'environ 4.10 % sur le MAE. On relève néanmoins que la prédiction moyenne obtenue à l'aide du modèle GLM-Poisson complexifié se rapproche nettement mieux de la fréquence moyenne réelle de sinistres observée au niveau de la base test (avec un ratio de 0.9650) ; autrement dit, le biais du modèle GLM-Poisson complexifié est moindre que celui du modèle GLM-Poisson initial.

Cependant, bien que ces résultats conséquents à l'intégration des effets d'interaction entre les caractéristiques dans le modèle soient peu satisfaisants, voire décevants en terme de gain relatif de précision prédicitve, la méthodologie adoptée dans notre étude demeure réutilisable sur de nouvelles bases de données, ou pour de nouvelles expériences dans lesquelles les effets d'interaction entre les caractéristiques ont un effet significatif sur la variable cible.

Modèles fréquence	Métriques					Gain relatif (en %)		
	$\frac{\text{moy. pred.}}{\text{moy. réelle}}$	MSE	MAE	RMSE	$RMSE_{mean}$	MSE	MAE	RMSE
GLM initial (benchmark)	0.8794369	0.058715	0.085295	0.24231	4.959702	–	–	–
GLM complexifié	0.9650467	0.058534	0.088793	0.24193	4.952040	0.31	−4.10	0.15

TABLE 5.20 : *Comparaison de la performance prédictive du modèle GLM fréquence initial et du modèle GLM fréquence complexifié (après l'ajout des effets non linéaires et d'interaction).*

Conclusion

En France, 83%. C'est la part d'assureurs qui considèrent que l'IA va profondément modifier les processus internes et la relation client, selon une étude menée en 2022 par l'ACPR.

L'interprétabilité des modèles d'apprentissage statistique revêt une importance cruciale dans le domaine de l'actuariat. Deux principales raisons justifient cela : en premier lieu, l'interprétabilité de ces modèles permet de satisfaire aux exigences réglementaires de transparence du processus de tarification édictées par l'ACPR ; et en second lieu, l'interprétabilité permet d'éviter les problèmes d'éthique et de gouvernance liés à la non transparence de ces modèles.

Tout au long de ce mémoire, l'objectif était double. Premièrement, nous avons exploré les différentes méthodes et techniques visant à améliorer l'interprétabilité des modèles opaques d'apprentissage statistique. Deuxièmement, nous avons étudié les enjeux des données télématiques en assurance automobile en vue d'une tarification plus précise et plus équitable.

Les méthodes d'interprétabilité étudiées dans ce mémoire ont été présentées suivant deux grandes catégories, à savoir :

• Les méthodes d'explication globales, c'est-à-dire celles qui fournissent des interprétations valables pour l'ensemble du jeu de données. Dans cette catégorie nous avons présentés plusieurs méthodes suivant leur utilité. Les méthodes permettant d'évaluer l'importance des différentes caractéristiques dans le modèle *boîte noire* (MR et SFIMP). Les méthodes permettant d'évaluer l'effet marginale des différentes caractéristiques sur la variable cible, à savoir : les graphiques de dépendance partielle (PDP) et les graphiques d'effets locaux accumulés (ALE-plot). Enfin, les méthodes qui permettent d'étudier l'interaction entre les différentes caractéristiques dans le modèle (Indice de Sobol, H-statistique de Friedman). Les résultats obtenus à l'aide des indices de Sobol étaient assez imprécis, ce qui peut s'expliquer par le faible pouvoir prédictif des différentes caractéristiques sur la variable cible.

• Les méthodes d'explication locales, c'est-à-dire celles qui permettent d'interpréter la valeur prédite d'un seule observation de la base de données. Dans cette catégorie, nous avons présenté la boîte à outils ICE (*Individual conditionnal expectation*). Il s'agit d'outils de visualisation graphique qui permettent d'étudier l'effet local des caractéristiques sur les prédictions de la variable cible. Puis nous avons dévellopés les méthodes LIME et SHAP qui sont les plus populaires. Cependant, nous avons montré que ces dernières méthodes revêtaient une limite principale : elles sont peu fiables. En effet, les interprétations issues de ces méthodes peuvent être falsifiées à la guise du data-scientist (confère sous-section 5.5.3).

Dans notre cas d'application en assurance automobile, nous avons pu mettre en évidence quatre (04) faits notables :

• Nous avons pu montrer qu'un modèle d'apprentissage statistique supervisé aussi complexe soit-il, peut toujours s'interpréter par le biais des outils présentés au chapitre 4. L'illustration a été mené sur le cas des modèles LocalGLMnet dont les poids sont ajustés par des réseaux de neurones artificiels complexes.

• Par la suite, nous relevons que les variables télématiques permettent d'améliorer la précision prédictive et descriptive des modèles de tarification automobile. Dans notre étude, lorsqu'on part d'un modèle de fréquence classique de type GLM, en prenant en compte uniquement les variables de risque

classiques, et que par la suite on y augmente les variables télématiques, on réalise un gain relatif d'environ 5% sur la MSE et la MAE, et d'environ 2.5% sur la RMSE.

En outre, pour la modélisation de la fréquence de sinistre, lorsque l'on prend en compte les données télématiques, et qu'en plus, on utilise un modèle de fréquence complexe de type *Random Forest* capable de mieux extraire les informations contenues dans ces données télématiques, on réalise un gain relatif de précision encore plus significatif : 14.21% sur la MSE, de 7.50% sur la MAE et de 7.40% sur la RMSE, par rapport au modèle GLM-fréquence n'utilisant pas de données télématiques.

Cette amélioration est davantage accentuée lorsqu'on se focalise sur des segments spécifiques d'assurés. Par exemple, l'utilisation des variables télématiques réduit la MSE de la fréquence prédite de sinistres de plus 30% sur le segment des assurés de moins de 21 ans.

• Un autre fait marquant que l'on retient est que les modèles de type LocalGLMnet constitueraient une bonne alternative aux modèles classiques de type GLM et aux modèles complexes d'apprentissage statistique de type Random Forest en tarification automobile. Au niveau de notre cas d'application, en plus de challenger le modèle Random Forest en termes de performance prédictive (RMSE), les modèles LocalGLMnet offrent une grande flexibilité dans leur interprétation.

• Enfin, nous avons montré qu'à l'issue de l'interprétation du modèle complexe, les interprétations résultantes (effets non-linéaire et interactions détectées) pouvaient être utilisées pour optimiser les performances du modèle GLM-Poisson de départ. Cependant, dans notre étude, l'amélioration des performances de ce dernier était infime (0.15% de gain de RMSE).

Cependant, bien que ces résultats conséquents à l'intégration des effets d'interaction entre les caractéristiques dans le modèle soient peu satisfaisants, voire décevants en terme de gain relatif de précision prédicitve, la méthodologie adoptée dans notre étude demeure réutilisable sur de nouvelles bases de données, ou pour de nouvelles expériences dans lesquelles les effets d'interaction entre les caractéristiques ont un effet significatif sur la variable cible.

En somme, cette étude sur l'interprétabilité des modèles de tarification opaques en actuariat a souligné l'importance de rendre ces modèles plus transparents et a proposé des méthodes robustes, tout en illustrant rigoureusement leur mise en oeuvre sur un cas pratique de tarification automobile. Cet enjeu d'interprétabilité permet ainsi aux actuaires de répondre aux exigences réglementaires de transparence des processus de tarification (et bien d'autres encore), de veiller aux contraintes éthiques posées par l'avènement de l'intelligence artificielle, et aussi de tirer partie de la performance prédictive des modèles complexes d'apprentissage statistique, tout ceci afin de mieux comprendre les facteurs de risque émergents des portefeuilles et favoriser ainsi une gestion optimale et équitable du risque au sein de la société d'aujourd'hui et de demain. Les différentes méthodes explorées dans ce mémoire offrent donc des outils prometteurs pour investiguer les mécanismes sous jacents aux modèles d'apprentissage statistique complexes.

Pour l'avenir, des perspectives intéressantes seraient, d'une part de concevoir des métriques robustes qui permettront d'évaluer la qualité des interprétations issues des méthodes d'interprétation *post hoc* présentées dans ce mémoire. D'autre part, il serait judicieux de promouvoir (vulgariser) davantage l'usage de ces méthodes d'interprétation existantes à l'heure actuelle, au sein de la communauté actuarielle, afin de faciliter la tâche d'interprétation aux actuaires, et de contribuer ainsi à la démocratisation des algorithmes avancés d'apprentissage statistique, notamment dans le secteur de l'assurance.

Annexe A

Autres modèles couramment utilisés en actuariat

A.1 Régression linéaire

La régression linéaire est la version ou l'approche la plus simplifiée de l'apprentissage supervisé. Elle est utilisée dans le cadre de la prédiction d'une variable quantitative à partir de variables numériques ou catégorielles. La régression linéaire existe depuis longtemps et fait jusqu'aujourd'hui l'objet d'innombrables manuels. La paternité de l'expression "régression linéaire" est attribué à Galton (1886) qui constata chez les êtres humains, un phénomène de régression de la taille des fils en fonction de la taille de leur père. Bien qu'elle puisse sembler quelque peu simpliste par rapport aux approches modernes de l'apprentissage statistique, elle constitue un bon point de départ pour la bonne compréhension des modèles complexes d'apprentissage statistiques. En effet, de manière générale, ces nouvelles approches sont inspirées ou tout simplement basées sur la régression linéaire. Dans cette section, nous passons en revue les idées qui sous-tendent le modèle de régression linéaire, abordons l'approche d'estimation par les moindres carrées ordinaires et nous rappelons les principaux résultats relatifs au modèle linéaire.

A.1.1 Principe général de la régression linéaire

Dans la régression linéaire, il s'agit de prédire une variable numérique y généralement appelée *variable cible*, sur la base d'une ou de plusieurs variables explicatives, notées $(x_j)_{1\leq j\leq p}$. Elle suppose qu'il existe une relation approximativement affine entre la variable d'intérêt y et les variables x_j. Mathématiquement, le modèle de régression prend la forme suivante :

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \quad \text{avec} \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{A.1}$$

Comme l'indique l'équation A.1, dans le modèle linéaire, l'on suppose que la variable cible est une somme pondérée des p caractéristiques x_j. Les coefficients β_j représentent les pondérations des différentes caractéristiques. La composante ϵ représente l'erreur de prédiction, c'est-à-dire la différence entre la prédiction et le résultat réel. Ces erreurs sont supposées suivre une distribution normale, ce qui signifie d'une part que, l'erreur peut aussi bien être positive que négative, et d'autre part, le modèle commet de nombreuses petites erreurs et très rarement de grosses erreurs.

Cependant, dans la pratique, les coefficients $\beta_j,\ 1 \leq j \leq p$ ne sont pas connus d'avance. Par conséquent, pour pouvoir effectuer des prédictions, ou tout simplement expliquer l'effet des différentes caractéristiques sur la variable cible, il est nécessaire d'utiliser les données historiques pour estimer les coefficients $\beta_j,\ 1 \leq j \leq p$.

Soit $Y = (y^{(i)})_{1\leq i\leq n}$ et $X = (x_j^{(i)})_{1\leq i\leq n,\, 0\leq j\leq p}$ (avec $x_0^{(i)} = 1$, pour tout $i = 1, \cdots, n$) les données que nous disposons pour n individus. Sous forme matricielle, le modèle de régression linéaire s'écrit : $Y = X\beta + \epsilon$. L'objectif est d'obtenir les estimations $\hat{\beta}$ des coefficients de telle sorte que le modèle linéaire (A.1) s'ajuste le mieux aux données disponibles. Autrement dit, nous voulons trouver le plan qui soit le plus "proche" des n points de données disponibles (voir figure A.1, pour le cas tri-dimensionnelle). Afin

de déterminer ce plan, l'approche de loin la plus courante en pratique consiste à minimiser le critère des moindres carrés. Ce qui revient à résoudre le programme de minimisation suivant :

$$\hat{\beta} = \underset{(\beta_j)_{0\leq j\leq p}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y^{(i)} - \beta_0 - \sum_{j=1}^{p} \beta_j x_j^{(i)} \right)^2 \tag{A.2}$$

On trouve finalement :

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y \tag{A.3}$$

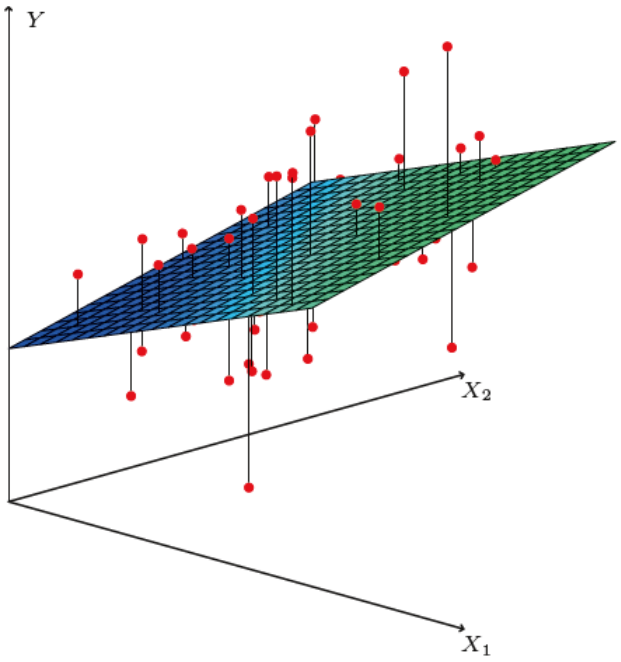


FIGURE A.1 : *Dans un cadre tridimensionnel, avec deux prédicteurs et une réponse, la ligne de régression des moindres carrés devient un plan. Le plan est choisi de manière à minimiser la somme des carrés des distances verticales entre chaque observation (en rouge) et le plan (issue du livre de Hastie* et al. *(2009)).*

A.1.2 Interprétation du modèle régression linéaire

• Effet marginal des variables explicatives

L'une des raisons pour lesquelles le modèle de régression linéaire est prisé en actuariat est la facilité de son interprétation. En effet, la linéarité de la relation apprise rend l'interprétation facile.

De prime abord, notons que l'interprétation du poids d'une caractéristiques dans le modèle dépend de la "nature" de la caractéristiques (catégorielle, binaire, ou numérique).

(*i*) Pour les *caractéristiques numériques* : l'augmentation de la caractéristique d'une unité supplémentaire, toutes choses étant égales par ailleurs, implique une variation de la variable cible d'un niveau égale au poids associé à cette caractéristique.

(*ii*) Pour une *caractéristique binaire* : la modification de la caractéristique de la catégorie de référence à l'autre catégorie modifie la valeur de la variable cible d'un niveau égale au poids associé à la catégorie concernée.

(*iii*) Pour une *caractéristique catégorielle avec L catégories* : nous n'avons besoin que de $L-1$ catégories. Comme dans le cas binaire précédent, une catégorie est retenue comme catégorie de référence et toutes les autres catégories s'interprètent par rapport à elle. L'interprétation de chacune des $L-1$ catégories est la même que l'interprétation pour le cas d'une caractéristique catégorielle binaire.

(*iv*) Enfin, *l'interception* noté β_0 correspond au poids associé à la "variable constante". Son interprétation est la suivante : Pour un individus avec toutes les variables numériques à zéro et les valeurs des variables catégorielles fixées aux catégories de référence, la prédiction du modèle est égale au poids de *l'interception*. L'interprétation de *l'interception* n'est généralement pas pertinente car les observations avec toutes les valeurs de variables quantitatives à zéro n'a souvent aucun sens réel. Son interprétation n'a généralement de sens que lorsque les variables explicatives numériques ont été au préalable standardisées. Ainsi, *l'interception* reflète alors la prédiction attendu pour "l'individu moyen" du jeu de données d'entraînement.

- **Évaluation de l'importance des variables**

Suite à l'estimation des poids $\beta_j,\ 1 \leq j \leq p$ par l'approche des moindres carrés ordinaires nous avons également accès aux écart-type estimés des différents coefficients notés $se(\hat{\beta}_j),\ 1 \leq j \leq p$. Ces écart-type peuvent être utilisés pour effectuer des tests d'hypothèses de nullité des coefficients. Ainsi, le test plus courant consiste à tester l'hypothèse nulle :

$$H0 : \text{Il n'y a pas de relation entre } x_j \text{ et } y, \text{ autrement dit } \beta_j = 0 \tag{A.4}$$

contre l'hypothèse alternative :

$$H1 : \text{Il existe une relation entre } x_j \text{ et } y, \text{ autrement dit } \beta_j \neq 0 \tag{A.5}$$

Pour une variable explicative x_j, étant donné que l'on ne dispose pas de la vraie valeur de son poids β_j, pour tester l'hypothèse nulle, nous déterminons tout simplement si son estimation $\hat{\beta}_j$ est généralement suffisamment éloigné de zéro pour que nous puissions être confiant que β_j est bien non nul. Pour ce faire, si la valeur de $se(\hat{\beta}_j)$ est petite −c'est-à-dire si $\hat{\beta}_j$ ne varie que très peu, c'est-à-dire, est assez précis−, alors même des valeurs relativement faibles mais non nulles de $\hat{\beta}_j$ (en valeur absolue) peuvent mettre à l'évidence qu'il existe une relation entre x_j et y. En revanche, si $se(\hat{\beta}_j)$ est élevé, alors, $\hat{\beta}_j$ devrait être grand (en valeur absolue) pour que nous puissions rejeter l'hypothèse nulle (confère chapitre 3 du livre de Hastie *et al.* (2009)). Ainsi, en pratique, l'importance d'une variable dans un modèle de régression linéaire peut être mesurée par la valeur absolue de la statistique t définie par :

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{A.6}$$

qui n'est rien d'autre que le poids estimé corrigé des effets d'échelle par son *erreur standard*.

Une interprétation typique de t serait : l'importance d'une variable augmente avec l'augmentation de son poids ; cependant, plus le poids estimé est imprécis (variance élevée), moins la caractéristique est importante.

A.1.3 Régularisation RIDGE et LASSO

Par définition, la régularisation traduit tout simplement l'action de régulariser, ou tout simplement, l'action de rendre régulier. Pourquoi parle-t-on donc de régularisation dans le cadre de la régression linéaire ? Qu'est ce qui est irrégulier et que l'on aimerait rendre régulier ? Et enfin, comment y parvenir ?

Il y a deux raisons principales pour lesquelles nous ne sommes souvent pas satisfaits des estimations des moindres carrés :

– La première est l'imprécision des prédictions. En effet, les estimateurs $\hat{\beta}$ des MCO des poids β ont généralement un faible biais, mais une forte variance : ce qui accroît mécaniquement l'imprécision des estimations. Or, la précision des prédictions peut parfois être améliorée en réduisant ou en annulant certains coefficients. En procédant ainsi, nous agissons sur le compromis biais-variance : on "sacrifie" un peu de biais dans l'optique d'améliorer la précision globale du modèle.

– Par ailleurs, la deuxième raison est relative à l'interprétabilité du modèle. En effet, avec un très grand nombre de prédicteurs l'interprétation du modèle devient complexe et difficile à comprendre, ce qui détériore et la précision descriptive et la précision prédictive du modèle (car sur-apprentissage). Ainsi, nous aimerions souvent déterminer un sous ensemble plus restreint de variables qui présente les effets les plus forts sur la variable cible.

Ainsi, en apprentissage statistique, lorsqu'on parle de régularisation, il est tout simplement question d'améliorer la précision des estimateurs des coefficients, pour *in fine*, améliorer la précision des prédictions.

• Régression Lasso

Cette approche de régularisation permet de palier aux deux raisons évoquées çi-dessus. L'estimation Lasso est définie par le problème d'optimisation suivant :

$$\hat{\beta}^{lasso} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{j=1}^{n} \left(y^{(j)} - \beta_0 - \sum_{i=1}^{p} x_j^{(i)} \beta_j \right)^2, \text{ sous contrainte } \sum_{i=1}^{p} |\beta_j| \leq t \tag{A.7}$$

Le Lasso est un moyen automatique et pratique d'introduire de la parcimonie dans le modèle de régression linéaire.

• Régression Ridge

Relativement aux moindres carrés ordinaires, la régression ridge (tout comme la régression lasso) permet de réduire la variance des estimateurs au détriment de l'introduction d'un biais.

L'estimation Ridge est définie par le problème d'optimisation suivant :

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \beta_0 - \sum_{j=1}^{p} x_j^{(i)} \beta_j \right)^2 + \lambda \left(\sum_{j=1}^{p} \beta_j^2 \right) \tag{A.8}$$

Ce problème admet une solution explicite :

$$\hat{\beta}^{ridge} = \left(\lambda Id_{p+1} + \frac{1}{n} \sum_{i=1}^{n} X^{(i)} X^{(i)^T} \right)^{-1} \left(\frac{1}{2} \sum_{j=1}^{n} X^{(i)} Y^{(i)} \right) \tag{A.9}$$

Le biais de l'estimateur $\hat{\beta}^{ridge}$ augmente avec la valeur de λ, tandis que la variance diminue avec λ, d'où un compromis biais-variance à faire dans le choix de λ.

Le choix optimal de λ est généralement obtenu en adoptant une approche par validation croisée.

A.1.4 Avantages et inconvénients du modèle de régression linéaire

• Avantages

Le modèle de régression linéaire présente plusieurs avantages. Nous nous contenterons ici, d'énumérer juste quelques uns :

– Premièrement, la modélisation des prédictions sous forme de somme pondérée rend transparente la façon dont les prédictions sont produites. Et avec la régularisation Lasso, nous pouvons nous assurer la parcimonie du modèle.

– En outre, Le modèle de régression linéaire de par son aspect simple et intuitif, est utilisé par un très grand nombre de personnes de par le monde entier : par conséquent, il existe une forte communauté scientifique développer autour de ce modèle et il est facilement implémentable sous plusieurs logiciels (R, Python, SAS, Excel, etc.).

– Enfin, d'un point de vue théorique, le modèle linéaire constitue le socle de plusieurs autres modèles populaires (GLM, GAM, etc.).

• **Inconvénients**

En ce qui concerne les inconvénients, notons que :

– Le modèle linéaire ne peut représenter que des relations linéaires entre la variable cible et les variables explicatives. Ainsi, chaque potentielle non-linéarité ou interaction entre variables doit être fabriquer à la main pour être ensuite explicitement donnée au modèle. Or, cette pratique est assez limitée car les effets de non-linéarité ou d'interaction à prendre (manuellement) en compte peuvent parfois être nombreuses.

– Par ailleurs, les coefficients du modèle de régression linéaire sont dans certaines circonstances difficiles à interpréter en termes de "causalité" ou "corrélation".

– Enfin, le modèle de régression linéaire repose sur des hypothèses parfois peu réalistes, notamment en assurance où les variables d'intérêt (fréquence ou du coût d'un sinistre, probabilité d'occurrence d'événements, etc.) sont généralement à valeurs dans $\mathbb{N}$, $\mathbb{R}^+$ ou $[0, 1]$ et donc ne suivraient pas une distribution gaussienne. Il s'avère donc crucial pour l'actuaire de se tourner vers des modèles tenant mieux en compte les réalités métiers que ne le fait le modèle de régression linéaire.

A.2 Arbres de décision : algorithme CART

A.2.1 Généralités sur l'algorithme CART

Les fondements théoriques et pratiques des arbres (méthode CART) ont été présentés pour la première fois par Breiman *et al.* (1984). L'algorithme CART regroupe deux méthodes analytiques : les arbres de classification (CT) lorsque la variable dépendante est catégorielle, et les arbres de régression (RT) lorsque la variable dépendante est numérique. Les algorithmes CT et RT partagent des principes statistiques communs et ne diffèrent que par quelques détails. Dans ce mémoire nous présentons uniquement les aspects statistiques communs à CT et RT, en utilisant CART comme expression générale.

L'algorithme CART est une méthode analytique qui décompose les relations entre une variable cible notée Y, et un groupe de prédicteurs, X.

A.2.2 Procédures statistiques de CART

Statistiquement, l'algorithme CART effectue des partitions binaires successives des observations de la base d'apprentissage, au fur et à mesure que l'arbre grandit.

Lorsqu'une partition est faite, elle génère deux noeuds (un à gauche, puis un autre à droite). Ces noeuds sont appelés *noeuds enfants*, tandis que le noeud à partir duquel ils ont été généré est appelé *noeud parent*. Les noeuds situés à la fin de l'arbre sont appelés les *noeuds terminaux*, tandis que le noeud à l'origine de l'arbre (le tout premier noeud) est appelé *noeud racine*.

Avec la méthode CART, le partitionnement des noeuds en sous-groupes à chaque niveau est guidé non pas par un test statistique mais par un critère statistique appelé *impureté*. L'impureté mesure le

degré auquel les éléments d'un noeud appartiennent à différentes catégories (ou valeurs) de la variable cible.

Typiquement, un groupe est dit "pur" lorsque l'ensemble de ses éléments appartiennent à une même catégorie (ou valeur) de la variable dépendante, tandis qu'un groupe est dit "impur" lorsque l'ensemble de ses éléments appartiennent tous à des catégories (ou valeurs) différentes de la variable dépendante.

Lorsqu'un noeud est pur, on stoppe sa division, sinon, il faut décider soit d'arrêter le partitionnement et d'accepter le groupe associé comme noeud terminal (une décision imparfaite, certes), soit on sélectionne un autre prédicteur puis on continue le développement de l'arbre jusqu'à ce que chacun des noeuds terminaux soient purs ou tout simplement jusqu'à un niveau de partitionnement jugé "convenable".

A.2.3 Conditions d'arrêt de l'arborescence CART

Une question directement liée à la discussion faite dans le paragraphe précédent est celle de savoir : quand est-ce qu'il est convenable d'arrêter le partitionnement d'un noeud ?

Cette question vue sous un autre angle est équivalente au compromis "biais-variance" présenté dans le chapitre 1. En effet, si l'on arrête le partitionnement trop tôt, l'arborescence sera plutôt stable (variance faible), mais trop petite pour refléter la véritable structure des données (biais élevé). *A contrario*, si l'on arrête le partitionnement trop tard, l'arbre résultant sera assez précis (biais faible) sur l'échantillon d'entraînement, mais également trop grand pour être stable ou significatif sur des données non vues, car certains noeuds terminaux risqueraient d'être quasiment vides (variance forte).

En pratique, plusieurs, règles d'arrêt existent :

• La plus populaire d'entre elles est basée sur la réduction des impuretés. L'idée est de fixer un seuil et de comparer la réduction des impuretés avec ce seuil. L'arbre CART continue de croître tant qu'il existe un prédicteur capable de réduire l'impureté en deçà du seuil fixé. Lorsque toutes les caractéristiques ne parviennent plus à réduire l'impureté en deçà de ce seuil le faire, on arrête le partitionnement. Toutefois, la difficulté de cette approche réside dans le choix du bon seuil à fixer, d'autant plus que l'impureté est un concept très abstrait (qui n'a pas un sens statistique concret).

• Une autre approche, consiste à fixer d'avance, l'effectif minimal des noeuds terminaux. Par exemple, on peut fixer cet effectif à 50 observations ou à 5% du nombre total d'individus dans l'échantillon. Ainsi, pour cet exemple lorsqu'un noeud contiendra plus de 50 observations, s'il est déjà pur, on arrête son partitionnement, sinon on continue à le partitionner en noeud fils jusqu'à obtenir soit des noeuds purs, ou alors des noeuds soit des noeud de 50 observations. Ce critère d'arrêt a pour avantage de limiter l'existence de noeuds terminaux non significatifs (de tailles très faibles).

• Une troisième méthode repose sur le test statistique d'adéquation du Chi-2. Sa procédure est clairement détaillée par Hart *et al.* (2000). Elle consiste à comparer si la partition obtenue à l'issue d'un découpage basé sur une caractéristique donnée est statistiquement significativement différente d'une partition aléatoire. Plus précisément, avec cette méthode, à chaque noeud, l'on teste pour l'ensemble des caractéristiques si le fait de partitionner le noeud sur la base de cette caractéristique est statistiquement équivalent à partitionner aléatoirement le noeud en deux noeuds fils. On arrête l'arborescence du noeud dès lors qu'aucune caractéristique ne parvient plus à améliorer le partitionnement du noeud par rapport à un partitionnement aléatoire.

• La toute première technique d'arrêt de l'arbre fut proposée par Breiman *et al.* (1984). Elle consiste à adopter la validation pour décider de quand arrêter l'arborescence. L'idée est d'utiliser un sous-ensemble des données (par exemple 90% du total) pour développer l'arbre et d'utiliser le reste des données pour valider l'arbre. Plus l'arbre est grand, plus l'erreur dans la structure diminue, jusqu'à ce qu'un sur-ajustement survienne (confère figure A.2).

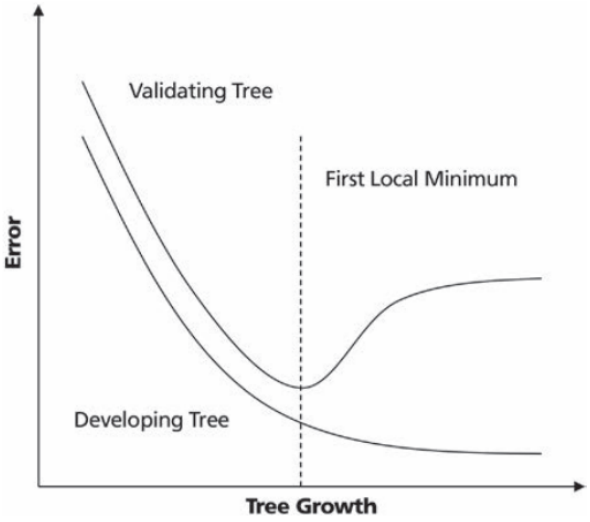


FIGURE A.2 : *Illustration de l'arrêt de l'arbre par validation (issue de l'article Ma (2018)).*

A.2.4 Élagage de l'arbre CART

En pratique, il arrive parfois de déclarer un noeud comme terminal assez prématurément, ce qui provoque des biais dans la structure de l'arbre. Pour palier cela, une stratégie consiste à laisser un arbre CART se développer complètement jusqu'à ce que la norme d'impureté minimale soit atteinte partout dans l'arbre. Ensuite, l'on examine toutes les paires de noeuds enfants descendant des mêmes noeuds parents un niveau au-dessus. Toute paire dont l'élimination ne conduit qu'à une "petite" amélioration de l'impureté est supprimée et son noeud parent devient un noeud terminal provisoire (provisoire dans le sens où ce noeud peut à son tour être supprimé à la prochaine étape) : cette méthode est appelée élagage de l'arbre CART. Il s'agit de la principale approche alternative pour arrêter l'arbre CART.

Dans la pratique, pour procéder à l'élagage, Breiman *et al.* (1984) ont proposé la mesure de la complexité des coûts. L'idée de base est d'attacher une pénalité à la tentative de faire pousser un arbre très grand. Plus l'arbre est grand, plus la pénalité est élevée.

Cela peut être observé à partir de la définition mathématique de la mesure de la complexité des coûts : pour un arbre T de taille $|T|$ (nombre de noeuds terminaux),

$$R_\alpha(T) = R(T) + \alpha|T| \tag{A.10}$$

où $R(T)$ est la mesure du risque de l'arbre T, c'est à son erreur d'apprentissage : par exemple le taux de mauvaise classification pour les problèmes de classification, et l'erreur quadratique moyenne pour les problèmes de régression ; α : le coefficient de pénalité non négatif. Comme on peut le constater plus l'arbre est grand (c'est-à-dire plus $|T|$ est élevé), plus la mesure de complexité de coûts $\alpha|T|$ est élevée.

L'objectif de l'élagage revient alors dans un premier temps à trouver, pour chaque niveau de α, l'arbre T_α qui minimise $R_\alpha(T)$. Par la suite, le choix du meilleur α à retenir s'obtient par une procédure validation-croisée (confère chapitre 1) : on choisir le meilleur $\hat{\alpha}$ qui minimise le risque $R(T_\alpha)$. Notre arbre final est noté $T_{\hat{\alpha}}$.

A.2.5 Formalisation mathématique de la procédure CART

Dans cette section nous empruntons les notations de James *et al.* (2013).

• **Mécanisme de construction des arbres de régression**

Donnons nous un ensemble de données $\left\{(x^{(i)},\, y^{(i)})\right\}_{i=1}^{n}$, avec $x^{(i)} = \left(x_1^{(i)},\, \cdots,\, x_d^{(i)}\right)$ le vecteur des caractéristiques à valeurs dans un ensemble $\mathcal{X}$ et $y^{(i)} \in \mathbb{R}$ la variable dépendante pour la $i-$ème observation. Supposons premièrement que nous disposons d'une partition de $\mathcal{X}$ en M régions $R_1,\, \cdots,\, R_M$, modélisons la réponse y par la constante $c_m \in \mathbb{R}$ dans chaque région R_m, $m \in \{1,\, \cdots,\, M\}$ respectivement, et désignons par f le modèle global. Alors :

$$f(x) = \sum_{m=1}^{M} c_m \mathbb{I}\{x \in R_m\},\ x \in \mathcal{X}. \tag{A.11}$$

L'estimation de f s'obtient en utilisant la méthode des moindres carrés ordinaires et l'on obtient, pour tout $m \in \{1,\, \cdots,\, M\}$,

$$\hat{c}_m = \frac{1}{|R_m|} \sum_{x^{(i)} \in R_m} y^{(i)}$$

Par l'expression de l'erreur totale $\sum_{i=1}^{n} \left(y^{(i)} - f(x^{(i)})\right)^2$, on voit bien que cette erreur totale dépend de f et donc des c_m ; or, l'estimation des c_m ci-dessus montre qu'ils sont eux-mêmes fonction des R_m. Ainsi, l'erreur totale dépend de la partition de $\mathcal{X}$ choisie. Comment donc choisir notre partition $\{R_m\}_{m=1}^{M}$ de manière à minimiser au mieux l'erreur totale ? En pratique, l'algorithme RT procède comme suit :

Étape 1− On choisit une caractéristique X_j quelconque que l'on segmente en une paire de demi-plans suivant un point seuil s (noter que si X_j est catégorielle, s n'est rien d'autre qu'un label ou un sous-groupe de labels de ses catégories) :

$$R_1(j,\, s) = \{X \mid X_j \leq s\} \text{ et } R_2(j,\, s) = \{X \mid X_j > s\}$$

Étape 2− Pour déterminer la meilleure variable X_j et le meilleur point seuil s à considérer, on résout le programme :

$$(j,\, s) = \underset{j,\, s}{\operatorname{argmin}} \left[\min_{c_1} \sum_{x^{(i)} \in R_1(j,\, s)} \left(y^{(i)} - c_1\right)^2 + \min_{c_2} \sum_{x^{(i)} \in R_2(j,\, s)} \left(y^{(i)} - c_2\right)^2 \right]$$

Étape 3− Une fois notre paire de partition de l'étape 2 obtenue, on repartit les données d'apprentissage dans les deux régions issue du partitionnement, et on répète le processus pour chacune des deux régions.

• **Mécanisme de construction des arbres de classification**

Lorsque la variable dépendante y est catégorielle et prend les valeurs $k \in \{1,\, \cdots,\, K\}$, on procède comme suit : Pour chaque région R_m, on calcule la proportion de la classe k dans R_m par la formule :

$$\hat{p}_{m,\,k} = \frac{1}{|R_m|} \sum_{x^{(i)} \in R_m} \mathbb{I}\{y^{(i)} = k\}$$

Ainsi étant donné une observation dans la région R_m, sa classe prédite sera $k(m) = \underset{k}{\operatorname{argmax}}\, \hat{p}_{m,\,k}$, c'est-à-dire la classe majoritaire de y dans R_m.

A présent, comment choisir optimalement la partition $\{R_m\}_{m=1}^{M}$ de manière à minimiser l'erreur totale de classification ? La procédure est similaire à celle des arbres de régression, juste l'étape 2 qui

se voir légèrement modifiée et devient :

$$(j,\, s) = \underset{j,\, s}{\operatorname{argmin}} \left[\min\, Q_1(j,\, s) + \min\, Q_2(j,\, s)\right]$$

avec Q pouvant correspondre à différente mesure d'impureté :

– *Erreur de mauvaise classification (Misclassification error)*

$$Q_1(j,\, s) = \frac{1}{|R_1(j,\, s)|} \sum_{x^{(i)} \in R_1(j,\, s)} \mathbb{I}\{y^{(i)} \neq k(j,\, s)\} = 1 - \hat{p}_{1,\, k(j,\, s)} \text{ et}$$

$$Q_2(j,\, s) = \frac{1}{|R_2(j,\, s)|} \sum_{x^{(i)} \in R_2(j,\, s)} \mathbb{I}\{y^{(i)} \neq k(j,\, s)\} = 1 - \hat{p}_{2,\, k(j,\, s)}$$

– *Indice de Gini*

$$Q_1(j,\, s) = \sum_{k \neq k'} \hat{p}_{1,\, k} \hat{p}_{1,\, k'} = \sum_{k=1}^{K} \hat{p}_{1,\, k}(1 - \hat{p}_{1,\, k}) \text{ et } Q_2(j,\, s) = \sum_{k \neq k'} \hat{p}_{2,\, k} \hat{p}_{2,\, k'} = \sum_{k=1}^{K} \hat{p}_{2,\, k}(1 - \hat{p}_{2,\, k})$$

– *Entropie croisée ou déviance*

$$Q_1(j,\, s) = \sum_{k=1}^{K} \hat{p}_{1,\, k} \log(\hat{p}_{1,\, k}) \text{ et } Q_2(j,\, s) = \sum_{k=1}^{K} \hat{p}_{2,\, k} \log(\hat{p}_{2,\, k})$$

Avantages et inconvénients de CART

• **Avantages**

La méthode CART possède plusieurs avantages, entre autres nous avons :

– Interprétabilité : le mécanisme de prédictions par la méthode CART est assez intuitif et facilement compréhensible même par un profane.

– Capture les interactions entre les caractéristiques et les potentielles effets non linéaires des prédicteurs sur la variable cible.

– Les données sont partitionnées en des groupes distincts parfois plus faciles à comprendre.

– Il n'est pas nécessaire d'effectuer des transformations préalables sur les variables (par exemple, pas besoin de standardiser au préalable les caractéristiques). Par exemple avec les modèles linéaire on est parfois emmené à passer au logarithme de certaines variables. En outre la méthode CART étant non paramétrique est plus robuste au valeurs aberrantes relativement aux modèles paramétriques.

• **Inconvénients**

– Instabilité : de légères modifications dans la base de données d'entraînement peut parfois créer une arborescence complètement différente. C'est la raison pour laquelle, au lieu de prendre ses décisions sur la base d'un seul arbre, il peut être intéressant d'utiliser plusieurs arbres et de les mettre en collaboration pour tirer une décision plus robuste. Ceci est par exemple possible à l'aide d'une architecture de Forêt aléatoire (en anglais, *Random Forest*), bien qu'on y perd en transparence.

– Les noeuds terminaux augmentent rapidement avec la profondeur de l'arbre, rendant l'arbre plus difficile à parcourir.

Annexe B

Autres méthodes d'interprétabilité post hoc

B.1 Modèles de substitution globaux

□ Principe général

L'approche d'interprétabilité par un modèle de substitution global consiste à former un modèle nativement interprétable pour approximer les prédictions d'un modèle boîte noire aussi précisément que possible. De sorte que l'on puisse tirer des conclusions sur le modèle boîte noire en interprétant le modèle de substitution. Autrement formulé, nous disposons d'une fonction de prédiction complexe f, et la tâche consiste à approcher f aussi étroitement que possible par une fonction de substitution g, sous la contrainte que g soit nativement interprétable −par exemple g peut être un modèle linéaire (simple ou généralisé) ou un arbre de décision. L'idée de modèles de substitution peut également être trouvée dans la littérature sous différents noms : Modèle d'approximation, métamodèle, modèle de surface de réponse, etc.

□ Étapes de mise en oeuvre d'un modèle de substitution global

La mise en oeuvre des modèles de substitution est assez intuitive. Elle ne nécessite pas beaucoup de théorie pour la comprendre. Elle requiert juste l'accès aux données et à la fonction de prédiction f pour que l'on souhaite substituer. Il s'agit bien d'une méthode agnostique au modèle, car aucune information préalable sur la structure interne du modèle n'est nécessaire. Pour former un modèle de substitution, Molnar (2020) propose la démarche suivante subdivisée en sept (07) étapes très simples :

1 : Sélectionner un jeu de données $\mathcal{D}$. Il peut s'agir du même jeu de données que celui utilisé pour l'apprentissage du modèle de boîte noire ou d'un nouveau jeu de données de la même distribution.

2 : Pour le jeu de données $\mathcal{D}$ sélectionné, effectuer les prédictions à l'aide du modèle de boîte noire. Récupérer toutes les prédictions obtenues.

3 : Sélectionner un type de modèle interprétable : modèle linéaire, arbre de décision, ou autres.

4 : Entraîner le modèle interprétable choisi sur le jeu de données $\mathcal{D}$ avec pour variable cible les prédictions obtenues à l'étape 2 à partir de la boîte noire.

5 : A l'issue de l'étape 4, nous avons maintenant notre modèle de substitution.

6 : Évaluer dans quelle mesure le modèle de substitution reproduit les prédictions du modèle de boîte noire.

7 : Interpréter le modèle de substitution.

A l'étape 6, pour évaluer dans quelle mesure le modèle de substitution reproduit les prédictions du modèle de boîte noire, nous pouvons utiliser les métriques usuelles : l'AUC pour les problèmes de classification (ou courbe ROC en cas de classification binaire), la MSE pour les problèmes de régression (ou R^2, voir pseudo-R^2 si le type de modèle de substitution choisi est le modèle linéaire). En outre, on peut comparer visuellement les courbes de distributions empiriques de $\left\{\hat{y}_*^{(i)}\right\}_{i=1}^{n}$ et $\left\{\hat{y}^{(i)}\right\}_{i=1}^{n}$,

valeurs prédites par le modèle de substitution et le modèle boîte noire respectivement, à partir du jeu d'entraînement initial $\mathcal{D}$.

A l'étape 7, il faut noter qu'en cas de bon ajustement entre le modèle substitut et le modèle sous-jacent, l'interprétation du modèle de substitution est toujours valable parce qu'elle fait des déclarations sur le modèle sous-jacent et non sur le monde réel. Cependant, l'interprétation du modèle de substitution devient hors de propos si le modèle de boîte noire sous-jacent est mauvais, car alors le modèle de boîte noire lui-même n'est pas pertinent.

□ **Avantages et limites**

• **Avantages**

– La flexibilité : cela se traduit par la possibilité de choisir le type de modèle de substitution avec lequel l'on se sent le plus à l'aise, soit le modèle linéaire, soit les arbres, ou autre type. De plus, il s'agit d'une méthode d'interprétation indépendante du modèle à interpréter.

– Avec les métriques usuelles, il est possible d'évaluer rigoureusement la capacité du modèle de substitution à reproduire les prédictions du modèle sous-jacent.

– L'approche de cette méthode est très intuitive : non seulement elle est facile à mettre en œuvre, mais elle est aussi facile à expliquer aux personnes qui ne sont pas familières avec la science des données ou l'apprentissage automatique.

• **Limites**

– Tout d'abord, en utilisant cette approche d'interprétabilité, les interprétations faites à partir du modèle de substitution ne sont valables que sur le modèle sous-jacent, mais pas sur le monde réel (du moins pas directement). Car le modèle de substitution n'a jamais accès aux vraies valeurs de la variable d'intérêt.

– Chaque type de modèle interprétable que l'on dispose pour la substitution présente lui-même ses propres avantages et ses inconvénients. Par exemple, les modèles linéaires ne prennent pas en compte les interactions entre variables de manière automatique. On voit donc qu'avec un simple modèle linéaire, il sera difficile de mettre en lumière les éventuels effets d'interactions dans le modèle.

□ **Quelques méthodes de substitution globales**

Dans la littérature, il existe plusieurs méthodes de substitution globales. Nous récapitulons les plus populaires d'entre-elles dans le tableau B.1 ci-dessous.

Méthodes	**Références**
SP-LIME	**?**
k-LIME	*Hall* et al. *(2017)*
Soft Decision Tree	*Frosst et Hinton (2017)*
Binary Decision Tree	*Yang* et al. *(2018)*

TABLE B.1 : *Quelques méthodes de substitution globales.*

B.2 Méthodes locales

B.2.1 Explications par les ancres

Les ancres constituent un nouveau système agnostique de modèle qui explique le comportement des modèles complexes avec des règles de haute précision appelées ancres.

Les ancres représentent des conditions locales "suffisantes" pour les prédictions. Cette méthode a été introduite par les mêmes chercheurs qui ont introduit la méthode LIME à savoir Ribeiro *et al.* (2018).

Comme la méthode LIME, les ancres sont construits à partir de perturbations pour générer des explications locales. Cependant, alors qu'avec la méthode LIME, l'on construit un modèle de substitution local, avec les ancres, les explications résultantes sont présentées sous la forme de règles de décisions du type : SI (conditions vérifiées) ALORS (Résultat).

La construction des ancres est basée sur des techniques d'apprentissage par renforcement. Pour un espace de perturbation donné, alors que LIME se contente d'apprendre une frontière de décision locale qui se rapproche le mieux de celle du modèle boîte noire, les ancres, de par leur conception (apprentissage par renforcement) construisent des explications dont la couverture est adaptée au comportement du modèle.

Avantages et inconvénients

• **Avantages**

– Facile à interpréter : de par la nature des sortie sous le format SI ... ALORS.
– Rapidité d'exécution : l'algorithme de mise en oeuvre peut être parallélisé.
– Explications plus précises que les explications issues de LIME.
– Les ancres de par leurs constructions fournissent des explications parcimonieuses.

• **Inconvénients**

– Comme tout méthode basée sur les permutations, il faut prêter une attention particulière à la manière dont les permutations sont faites afin de limiter les explications biaisées.

– À notre connaissance, il n'existe pas encore de packages *R* ni Python en accès libre permettant de mettre en place les ancres pour les problèmes de régression ou de classification multi-classes. Le package existant permet uniquement de construire les ancres pour les modèles de classification binaire.

B.2.2 Autres méthodes d'explications locales

Méthodes	Références
MUSE	*Lakkaraju* et al. *(2019)*
LORE	*Guidotti* et al. *(2018)*
MAPLE	*Plumb* et al. *(2018)*
MES	*Turner (2016)*

TABLE B.2 : *Quelques méthodes d'explication locales.*

Annexe C

Résultats complémentaires du cas d'application mené au chapitre 5

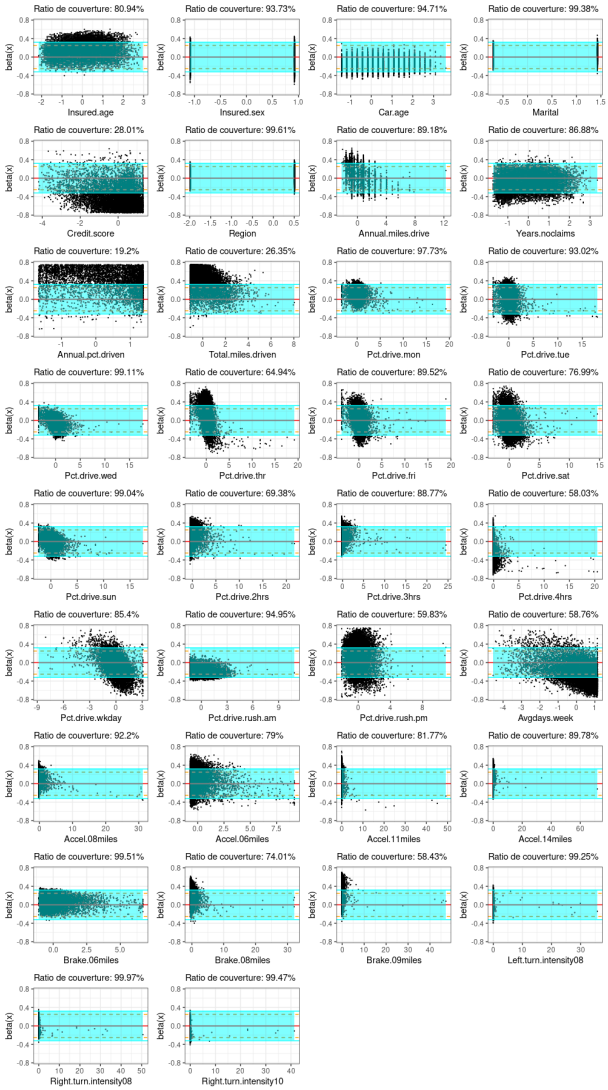


FIGURE C.1 : *Poids d'attention* $\hat{\beta}_j(x_+^{(i)})$ *sur la fréquence de sinistres, des caractéristiques continues et binaires, pour* 10 000 *instances* $x_+^{(i)}$ *de la base de test sélectionnées aléatoirement ; la bande bleu clair indique la zone de confiance* I_α *au niveau de signification* $\alpha = 0,1\%$ *du test d'hypothèse nulle* $H_0 : \hat{\beta}_j(x_+) = 0$.

Classes / Variables	iq1	iq2	iq3	iq4	iq5
Accel.06milesG	[0, 7]	]7, 17]	]17, 32]	]32, 63]	]63, +[
Brake.08milesG	[0, 2]	]2, 4]	]4, 7]	]7, 13]	]13, +[
Left.turn.int.08G	[0, 4]	]4, 29]	]29, 144]	]144, 485]	]485, +[
Total.miles.drivG	[0, 1215]	]1215, 2643]	]2643, 4520]	]4520, 7818]	]7818, +[
Avgdays.weekG	[0, 4.5]	]4.5, 5.5]	]5.5, 6]	]6, 6.5]	]6.5, 7]
Annual.pct.drivG (%)	[0, 20]	]20, 40]	]40, 50]	]50, 85]	]85, 100]
Pct.drive.monG (%)	[0, 11]	]11, 13]	]13, 15]	]15, 17]	]17, 100]
Pct.drive.thrG (%)	[0, 12]	]12, 14]	]14, 16]	]16, 18]	]18, 100]
Pct.drive.satG (%)	[0, 10]	]10, 13]	]13, 15]	]15, 17]	]17, 100]
Pct.drive.sunG	[0, 8]	]8, 10]	]10, 12]	]12, 14]	]14, 100]
Pct.drive.rush.amG	[0, 3]	]3, 6]	]6, 10]	]10, 16]	]16, 100]
Pct.drive.rush.pmG	[0, 8]	]8, 11]	]11, 15]	]15, 19]	]19, 100]
Credit.scoreG (%)	[0, 750]	]750, 800]	]800, 850]	]850, 900]	–

TABLE C.1 : *Résumé des classes retenues à l'issue du regroupement des variables explicatives continues ; les intervalles* iq_j, $j = 1, \cdots, 5$ *sont délimitées par les quintiles des variables correspondantes.*

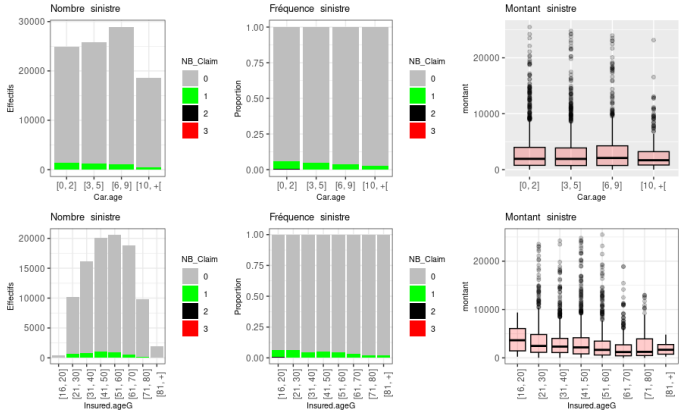


FIGURE C.2 : *Analyse de la sinistralité selon les variables Car.ageG et Insured.ageG.*

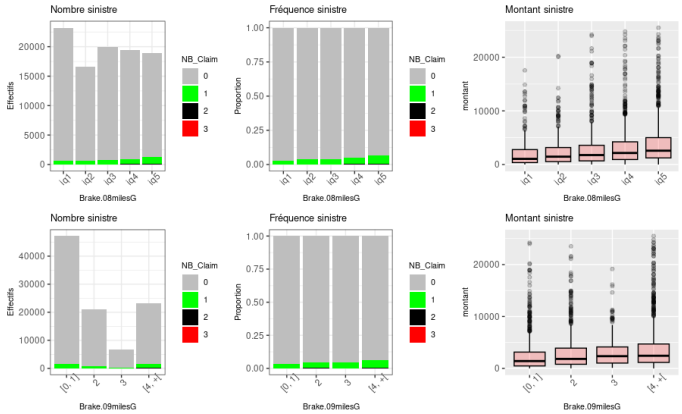


FIGURE C.3 : *Analyse de la sinistralité selon les variables Brake.08miles et Brake.09milesG.*

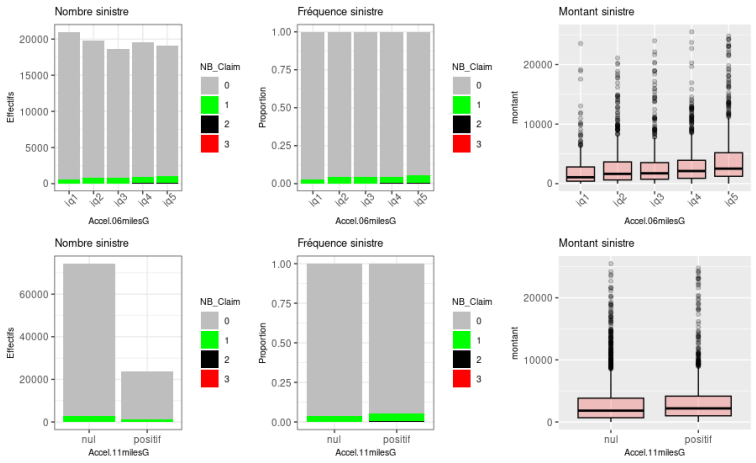


FIGURE C.4 : *Analyse de la sinistralité selon les variables Accel.06milesG et Accel.11milesG.*

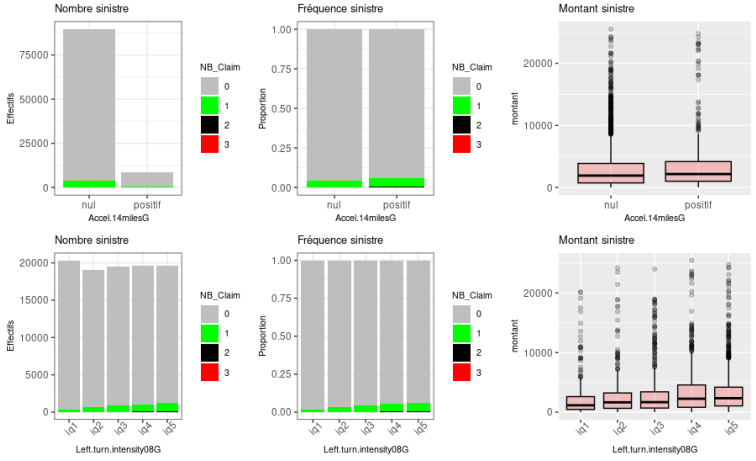


FIGURE C.5 : *Analyse de la sinistralité selon les variables Accel.14milesG et Left.turn.intensity08G.*

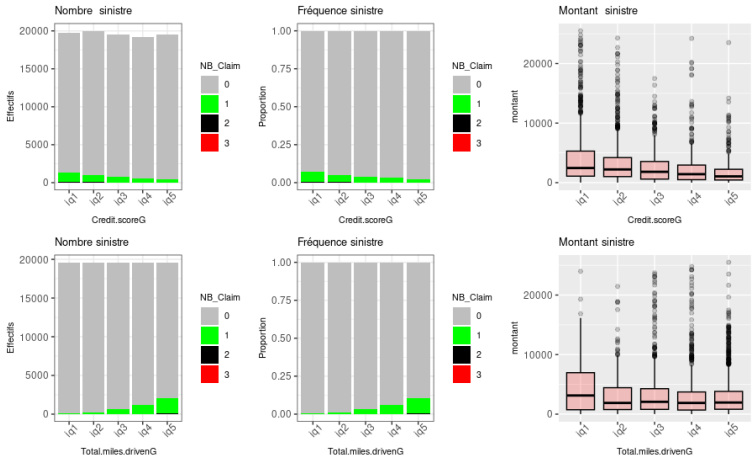


FIGURE C.6 : *Analyse de la sinistralité selon les variables Credit.scoreG et Totalmiles.drivenG.*

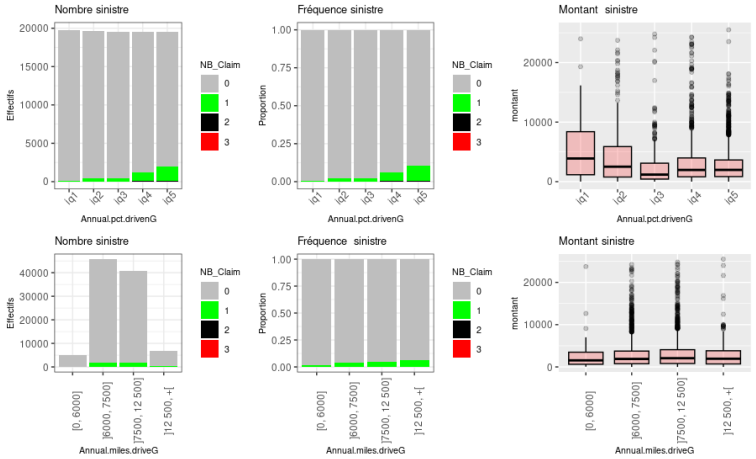


FIGURE C.7 : *Analyse de la sinistralité selon les variables Annual.pct.drivenG et Annual.miles.driveG.*

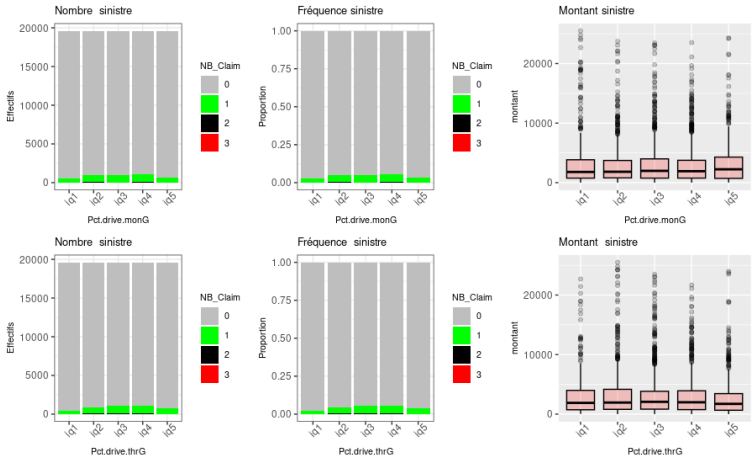


FIGURE C.8 : *Analyse de la sinistralité selon les variables Pct.drive.monG et Pct.drive.thrG.*

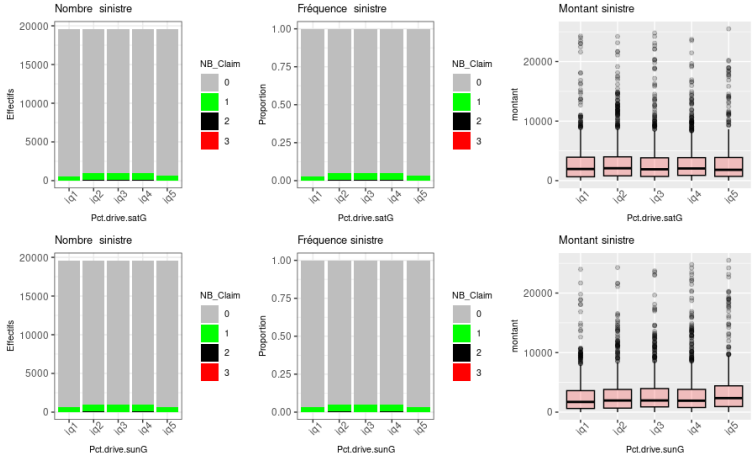


FIGURE C.9 : *Analyse de la sinistralité selon les variables Pct.drive.satG et Pct.drive.sunG.*

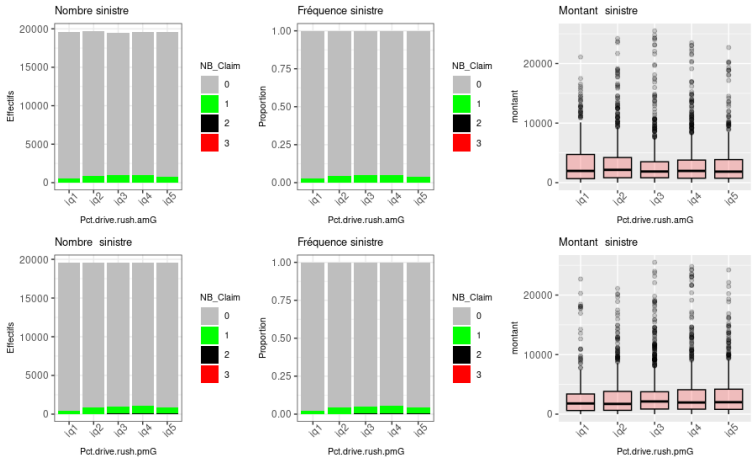


FIGURE C.10 : *Analyse de la sinistralité selon les variables Pct.drive.rush.amG et Pct.drive.rush.pmG.*

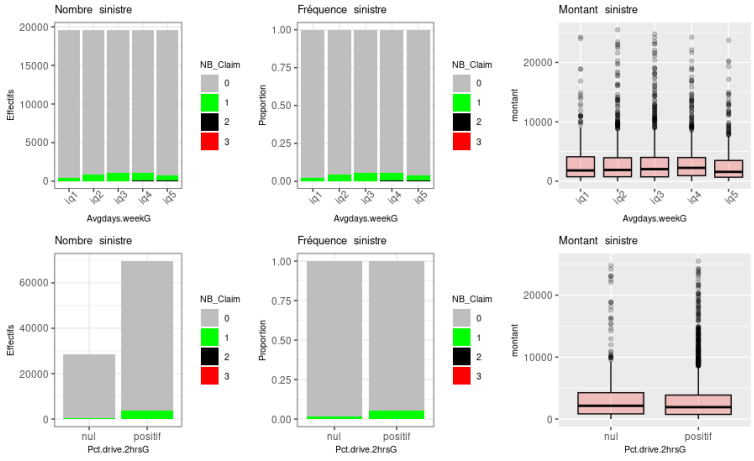


FIGURE C.11 : *Analyse de la sinistralité selon les variables Avgdays.weekG et Pct.drive.2hrsG.*

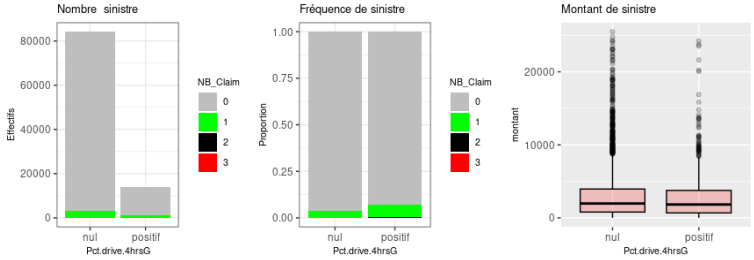


FIGURE C.12 : *Analyse de la sinistralité selon les variables Pct.drive.4hrsG.*

	Estimate
(Intercept)	-4.52^{***}
	(0.27)
TerritoryGzone_B	-0.03
	(0.04)
TerritoryGzone_C	-0.05
	(0.05)
Car.useCommute	-0.23^{**}
	(0.08)
Car.useFarmer	-0.84^{***}
	(0.22)
Car.usePrivate	-0.20^{*}
	(0.08)
Car.ageG[3, 5]	-0.12^{**}
	(0.04)
Car.ageG[6, 9]	-0.34^{***}
	(0.04)
Car.ageG[10, +[	-0.56^{***}
	(0.06)
Insured.ageG[21, 30]	-0.27
	(0.21)
Insured.ageG[31, 40]	-0.49^{*}
	(0.20)
Insured.ageG[41, 50]	-0.29
	(0.20)
Insured.ageG[51, 60]	-0.12
	(0.20)
Insured.ageG[61, 70]	-0.16
	(0.21)
Insured.ageG[71, 80]	-0.16
	(0.22)
Insured.ageG[81, +]	0.19
	(0.25)
Credit.scoreG]750, 800]	-0.22^{***}
	(0.05)
Credit.scoreG]800, 850]	-0.61^{***}
	(0.04)
Credit.scoreG]850, 900]	-0.58^{***}
	(0.05)
Annual.miles.driveG]6000, 7500]	0.74^{***}
	(0.12)
Annual.miles.driveG]7500, 12 500]	0.70^{***}
	(0.12)
Annual.miles.driveG]12 500, +[	0.86^{***}
	(0.13)
Brake.08milesB]10, +[	0.29^{***}
	(0.04)
Brake.09milesB]5, +[	0.68^{***}
	(0.05)
Accel.06milesB]30, +[	-0.01
	(0.04)
Accel.11milesB]0, +[	0.09^{*}
	(0.04)
Accel.14milesB]0, +[	0.08
	(0.06)
Left.turn.intensity08B]150, +[	0.22^{***}
	(0.03)
Total.miles.drivenB]4500, +[	0.95^{***}
	(0.05)
Annual.pct.drivenB]50%, 100%]	0.60^{***}
	(0.04)
Pct.drive.rush.amB]5%, 100%]	-0.07
	(0.04)
Pct.drive.rush.pmB]10%, 100%]	0.08^{*}
	(0.04)
Avgdays.weekB]4, 7]	0.23^{**}
	(0.07)
Pct.drive.2hrsB]0, 100%]	0.44^{***}
	(0.06)
Pct.drive.4hrsB]0, 100%]	0.15^{***}
	(0.04)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

TABLE C.2 : *Résumé GLM Poisson (lien logarithmique) avec variables tarifaires classiques completées par les variables télématiques.*

	Estimate
(Intercept)	3.09^{***}
	(0.60)
TerritoryGzone_B	0.20^{*}
	(0.09)
TerritoryGzone_C	0.43^{***}
	(0.12)
Car.useCommute	-0.23
	(0.19)
Car.useFarmer	-1.39^{**}
	(0.50)
Car.usePrivate	-0.30
	(0.20)
Car.ageG[3, 5]	-0.21^{*}
	(0.09)
Car.ageG[6, 9]	-0.40^{***}
	(0.09)
Car.ageG[10, +[	-0.91^{***}
	(0.12)
Insured.ageG[21, 30]	0.08
	(0.50)
Insured.ageG[31, 40]	-0.28
	(0.50)
Insured.ageG[41, 50]	-0.03
	(0.50)
Insured.ageG[51, 60]	0.02
	(0.50)
Insured.ageG[61, 70]	-0.38
	(0.51)
Insured.ageG[71, 80]	-0.13
	(0.52)
Insured.ageG[81, +]	0.22
	(0.59)
Credit.scoreG]750, 800]	-0.40^{***}
	(0.10)
Credit.scoreG]800, 850]	-1.01^{***}
	(0.09)
Credit.scoreG]850, 900]	-1.42^{***}
	(0.11)
Annual.miles.driveG]6000, 7500]	0.67^{**}
	(0.22)
Annual.miles.driveG]7500, 12 500]	0.85^{***}
	(0.22)
Annual.miles.driveG]12 500, +[	0.89^{***}
	(0.25)
Brake.08milesB]10, +[	0.44^{***}
	(0.10)
Brake.09milesB]5, +[	1.07^{***}
	(0.11)
Accel.06milesB]30, +[	0.00
	(0.08)
Accel.11milesB]0, +[	0.18^{*}
	(0.09)
Accel.14milesB]0, +[	-0.03
	(0.13)
Left.turn.intensity08B]150, +[	0.28^{***}
	(0.07)
Total.miles.drivenB]4500, +[	0.73^{***}
	(0.10)
Annual.pct.drivenB]50%, 100%]	0.75^{***}
	(0.09)
Pct.drive.rush.amB]5%, 100%]	-0.18^{*}
	(0.08)
Pct.drive.rush.pmB]10%, 100%]	0.18^{*}
	(0.09)
Avgdays.weekB]4, 7]	0.30^{*}
	(0.14)
Pct.drive.2hrsB]0, 100%]	0.43^{***}
	(0.10)
Pct.drive.4hrsB]0, 100%]	0.11
	(0.09)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

TABLE C.3 : *Résumé GLM Tweedie (lien logarithmique) avec variables tarifaires classiques completées par les variables télématiques.*

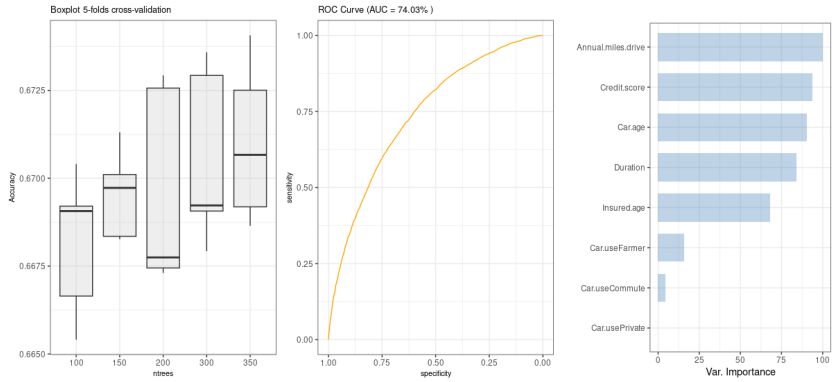


FIGURE C.13 : *Prédiction de Accel.06milesB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntree optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

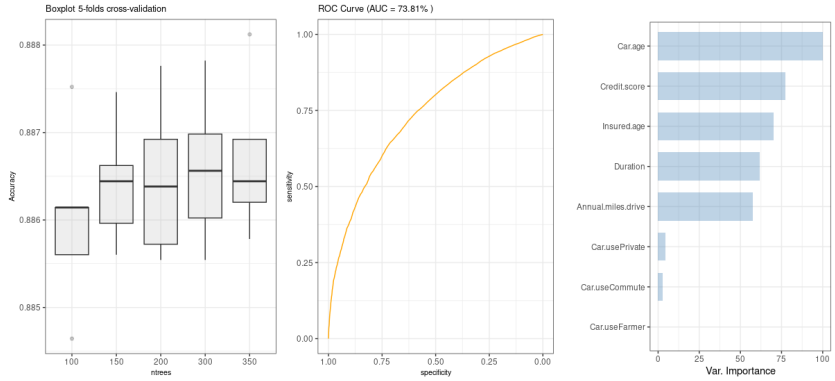


FIGURE C.14 : *Prédiction de Accel.11milesB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntree optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

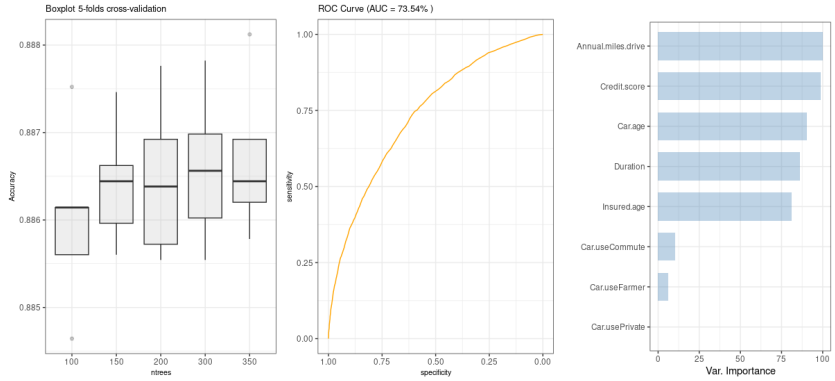


FIGURE C.15 : *Prédiction de Brake.08milesB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntree optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

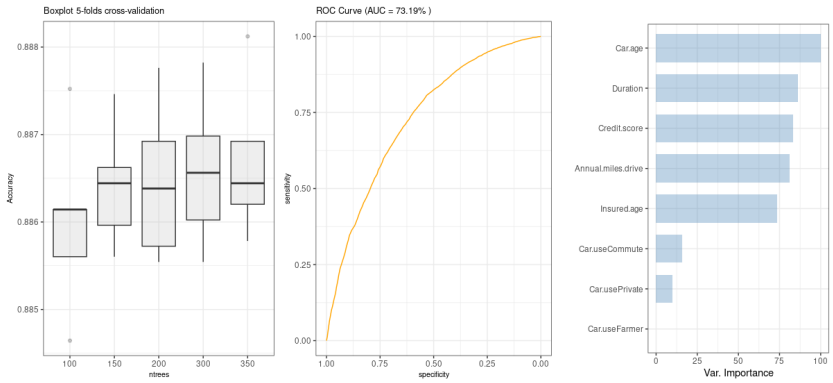


FIGURE C.16 : *Prédiction de Pct.drive.2hrsB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntree optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

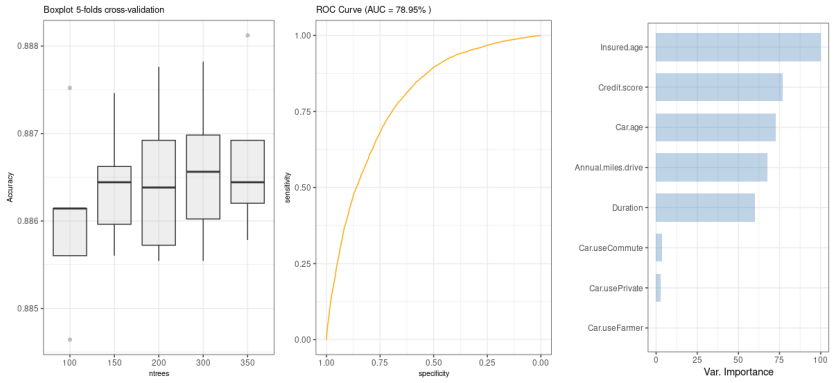


FIGURE C.17 : *Prédiction de Pct.drive.rush.pmB : Validation croisée sur ntrees (à gauche) ; courbe ROC sur la base test, pour le ntree optimal retenu (au milieu) ; importance des variables pour le modèle optimal (à droite).*

	Estimate
(Intercept)	-5.27^{***}
	(0.28)
TerritoryGzone_B	-0.04
	(0.04)
TerritoryGzone_C	-0.05
	(0.05)
Car.useCommute	-0.19^{*}
	(0.08)
Car.useFarmer	-0.84^{***}
	(0.22)
Car.usePrivate	-0.15
	(0.08)
Car.ageG[3, 5]	-0.12^{**}
	(0.04)
Car.ageG[6, 9]	-0.34^{***}
	(0.04)
Car.ageG[10, +[	-0.50^{***}
	(0.06)
Insured.ageG[21, 30]	-0.26
	(0.21)
Insured.ageG[31, 40]	-0.43^{*}
	(0.21)
Insured.ageG[41, 50]	-0.21
	(0.20)
Insured.ageG[51, 60]	-0.00
	(0.20)
Insured.ageG[61, 70]	0.01
	(0.21)
Insured.ageG[71, 80]	0.07
	(0.22)
Insured.ageG[81, +]	0.42
	(0.25)
Credit.scoreG]750, 800]	-0.20^{***}
	(0.05)
Credit.scoreG]800, 850]	-0.58^{***}
	(0.04)
Credit.scoreG]850, 900]	-0.53^{***}
	(0.05)
Annual.miles.driveG]6000, 7500]	0.77^{***}
	(0.12)
Annual.miles.driveG]7500, 12 500]	0.67^{***}
	(0.13)
Annual.miles.driveG]12 500, +[	0.77^{***}
	(0.14)
Total.miles.driven_fit	1.30^{***}
	(0.07)
Annual.pct.driven_fit	0.73^{***}
	(0.06)
Pct.drive.rush.am_fit	-0.09
	(0.05)
Left.turn.intensity08_fit	0.36^{***}
	(0.05)
Pct.drive.rush.pm_fit	0.09
	(0.06)
Brake.08miles_fit	0.46^{***}
	(0.06)
Brake.09miles_fit	1.00^{***}
	(0.07)
Pct.drive.4hrs_fit	0.22^{***}
	(0.06)
Pct.drive.2hrs_fit	0.64^{***}
	(0.08)
Accel.14miles_fit	0.11
	(0.09)
Accel.11miles_fit	0.10
	(0.06)
Accel.06miles_fit	-0.02
	(0.05)
Avgdays.week_fit	0.33^{**}
	(0.11)
Deviance	22703.86
Num. obs.	83350

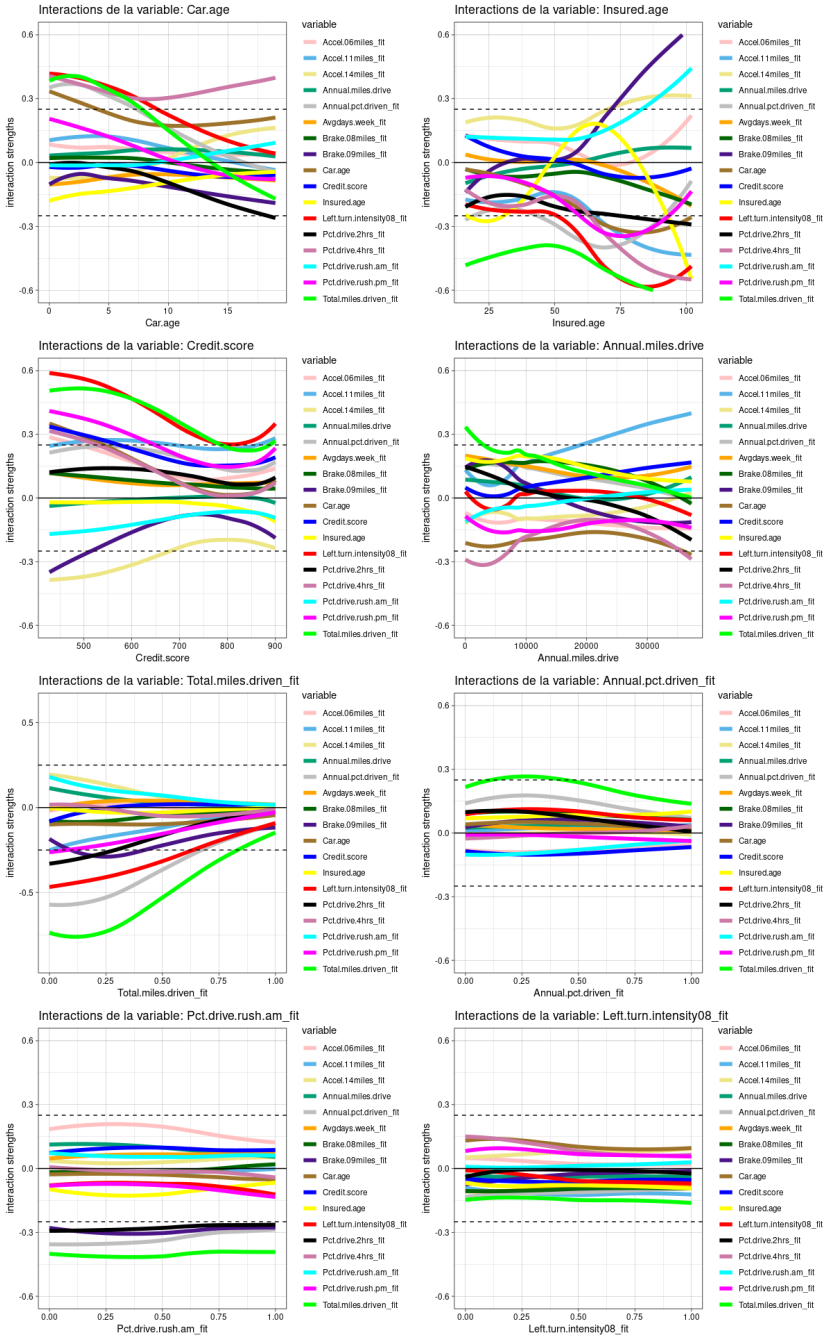
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

TABLE C.4 : Résumé du modèle de fréquence GLM Poisson (lien logarithmique).

	Estimate
(Intercept)	2.17^{***}
	(0.62)
TerritoryGzone_B	0.17
	(0.09)
TerritoryGzone_C	0.40^{***}
	(0.12)
Car.useCommute	−0.19
	(0.19)
Car.useFarmer	-1.42^{**}
	(0.51)
Car.usePrivate	−0.26
	(0.20)
Car.ageG[3, 5]	-0.19^{*}
	(0.09)
Car.ageG[6, 9]	-0.40^{***}
	(0.09)
Car.ageG[10, +[	-0.84^{***}
	(0.12)
Insured.ageG[21, 30]	0.09
	(0.50)
Insured.ageG[31, 40]	−0.26
	(0.49)
Insured.ageG[41, 50]	0.02
	(0.49)
Insured.ageG[51, 60]	0.09
	(0.49)
Insured.ageG[61, 70]	−0.24
	(0.50)
Insured.ageG[71, 80]	0.01
	(0.52)
Insured.ageG[81, +]	0.38
	(0.59)
Credit.scoreG]750, 800]	-0.39^{***}
	(0.10)
Credit.scoreG]800, 850]	-0.96^{***}
	(0.09)
Credit.scoreG]850, 900]	-1.33^{***}
	(0.11)
Annual.miles.driveG]6000, 7500]	0.71^{**}
	(0.22)
Annual.miles.driveG]7500, 12 500]	0.82^{***}
	(0.22)
Annual.miles.driveG]12 500, +[	0.83^{**}
	(0.25)
Total.miles.driven_fit	1.12^{***}
	(0.14)
Annual.pct.driven_fit	1.00^{***}
	(0.13)
Pct.drive.rush.am_fit	-0.31^{**}
	(0.12)
Left.turn.intensity08_fit	0.42^{***}
	(0.10)
Pct.drive.rush.pm_fit	0.17
	(0.13)
Brake.08miles_fit	0.63^{***}
	(0.14)
Brake.09miles_fit	1.55^{***}
	(0.17)
Pct.drive.4hrs_fit	0.07
	(0.14)
Pct.drive.2hrs_fit	0.68^{***}
	(0.15)
Accel.14miles_fit	−0.03
	(0.20)
Accel.11miles_fit	0.21
	(0.14)
Accel.06miles_fit	0.05
	(0.12)
Avgdays.week_fit	0.50^{*}
	(0.21)
Deviance	4981424.82
Num. obs.	83350

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

TABLE C.5 : Résumé du modèle de coût GLM Tweedie (lien logarithmique).


Interactions de la variable: Car.age
Car.age
Interactions de la variable: Insured.age
Insured.age
Interactions de la variable: Credit.score
Credit.score
Interactions de la variable: Annual.miles.drive
Annual.miles.drive
Interactions de la variable: Total.miles.driven_fit
Total.miles.driven_fit
Interactions de la variable: Annual.pct.driven_fit
Annual.pct.driven_fit
Interactions de la variable: Pct.drive.rush.am_fit
Pct.drive.rush.am_fit
Interactions de la variable: Left.turn.intensity08_fit
Left.turn.intensity08_fit
interaction strengths
variable
Accel.06miles_fit
Accel.11miles_fit
Accel.14miles_fit
Annual.miles.drive
Annual.pct.driven_fit
Avgdays.week_fit
Brake.08miles_fit
Brake.09miles_fit
Car.age
Credit.score
Insured.age
Left.turn.intensity08_fit
Pct.drive.2hrs_fit
Pct.drive.4hrs_fit
Pct.drive.rush.am_fit
Pct.drive.rush.pm_fit
Total.miles.driven_fit

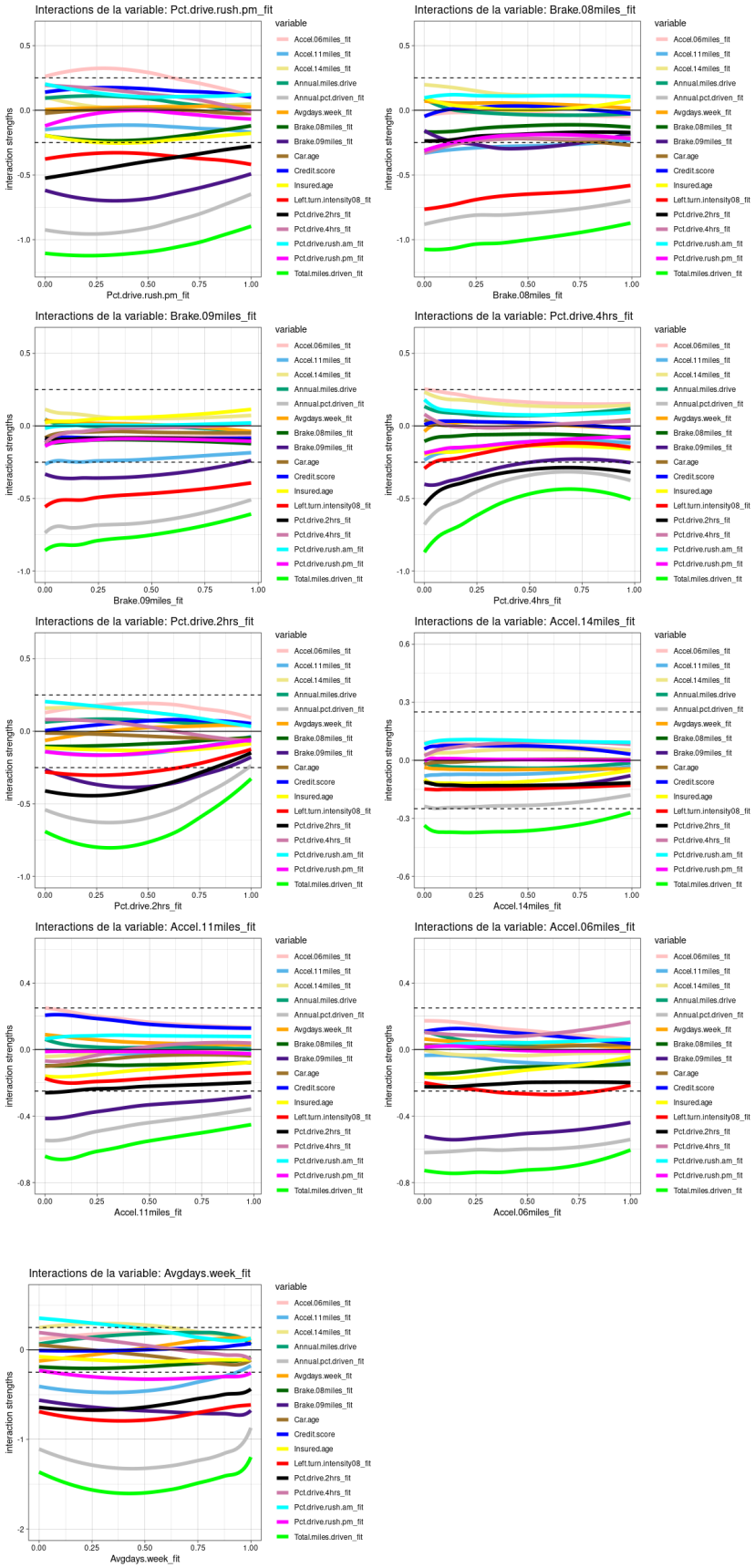


FIGURE C.18 : *Analyse des interactions : ajustements splines des gradients* $\partial_{x_k}\hat{\beta}_j(x^{(i)})$ *des caractéristiques continues* ($1 \leq j, k \leq 17$) *dans notre modèle LocalGLMnet-Poisson, pour l'ensemble des instances* $x^{(i)}$ *de la base de test.*

Bibliographie

Abbasi-Asl, R. et Yu, B. (2017). Interpreting convolutional neural networks through compression. *arXiv*, 1711.02329.

Alvarez-Melis, D. et Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv*, 1806.08049.

Apley, D. W. et Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 82(4):1059–1086.

Barry, L. et Charpentier, A. (2022). L'équité de l'apprentissage machine en assurance.

Basu, S., Kumbier, K., Brown, J. B. et Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948.

Bertsimas, D. et Dunning, I. (2016). Multistage robust mixed-integer optimization with adaptive partitions. *Operations Research*, 64(4):980–998.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. et Stone, C. J. (1984). *Classification and regression trees*. Wadsworth and Brooks.

Brockett, P. L. et Golden, L. L. (2007). Biological and psychobehavioral correlates of credit scores and automobile insurance losses : Toward an explication of why credit scoring works. *Journal of Risk and Insurance*, 74(1):23–63.

Burkart, N. et Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. et Elhadad, N. (2015). Intelligible models for healthcare : Predicting pneumonia risk and hospital 30-day readmission. *In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730.

Casalicchio, G., Molnar, C. et Bischl, B. (2019). Visualizing the feature importance for black box models. *In Joint European conference on machine learning and knowledge discovery in databases*, pages 655–670. Springer.

Chastaing, G., Gamboa, F. et Prieur, C. (2012). Generalized hoeffding-sobol decomposition for dependent variables-application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448.

Chawla, N. V., Bowyer, K. W., Hall, L. O. et Kegelmeyer, W. P. (2002). Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chouldechova, A. (2017). Fair prediction with disparate impact : A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Cohen, S., Dror, G. et Ruppin, E. (2007). Feature selection via coalitional game theory. *Neural Computation*, 19(7):1939–1961.

Dalalyan, A. (2018). Apprentissage statistique, cours de troisème année-voie data science statistique et apprentissage. *ENSAE Paris, 3A*.

Delcaillau, D. (2019). Contrôle et transparence des modèles complexes en actuariat. *Mémoire de l'Institut des Actuaires*.

Denuit, M. et Charpentier, A. (2005). *Mathématiques de l'assurance non-vie : Tome II Tarification et provisionnement*. Economica.

D'Haultfoeuille, X. (2021). Econométrie 1. *ENSAE Paris, 2A*.

Efron, B. et Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596.

Escofier, B. et Pagès, J. (1998). Analyses factorielles simples et multiples. *Dunod, Paris*, page 284.

Farrar, D. E. et Glauber, R. R. (1967). Multicollinearity in regression analysis : the problem revisited. *The Review of Economic and Statistics*, pages 92–107.

Fisher, A., Rudin, C. et Dominici, F. (2019). All models are wrong, but many are useful : Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81.

Fliche, O. et Yang, S. (2018). Artificial intelligence : Challenges for the financial sector. *Banque de France Discussion Paper*.

Freedman, D. A. (1991). Statistical models and shoe leather. *Sociological methodology*, pages 291–313.

Friedman, J., Hastie, T., Tibshirani, R. *et al.* (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friedman, J. H. (2001). Greedy function approximation : a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Friedman, J. H. et Popescu, B. E. (2008). Predictive learning via rule ensembles. *The annals of applied statistics*, pages 916–954.

Frosst, N. et Hinton, G. (2017). Distilling a neural network into a soft decision tree. *arXiv preprint arXiv :1711.09784*.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Goldstein, A., Kapelner, A., Bleich, J. et Pitkin, E. (2015). Peeking inside the black box : Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65.

Goodfellow, I., Bengio, Y. et Courville, A. (2016). *Deep learning*. MIT press.

Goodman, B. et Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Greenwell, B. M., Boehmke, B. C. et McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv preprint arXiv :1805.04755*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. et Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Hall, P., Gill, N., Kurka, M. et Phan, W. (2017). Machine learning interpretability with h2o driverless ai. *H2O. ai*.

Hart, P. E., Stork, D. G. et Duda, R. O. (2000). *Pattern classification*. Wiley Hoboken.

Hastie, T., Tibshirani, R., Friedman, J. H. et Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*, volume 2. Springer.

Henelius, A., Puolamäki, K., Boström, H., Asker, L. et Papapetrou, P. (2014). A peek into the black box : exploring classifiers by randomization. *Data mining and knowledge discovery*, 28(5):1503–1529.

Hoeffding, W. et Robbins, H. (1948). The central limit theorem for dependent random variables. *Duke Mathematical Journal*, 15(3):773–780.

Homma, T. et Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J. et Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154.

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kim, B., Glassman, E., Johnson, B. et Shah, J. (2015). ibcm : Interactive bayesian case model empowering humans via intuitive interaction.

Kiviat, B. (2019). The moral limits of predictive practices : The case of credit-based insurance scores. *American Sociological Review*, 84(6):1134–1158.

Kumbier, K., Basu, S., Brown, J. B., Celniker, S. et Yu, B. (2018). Refining interaction search through signed iterative random forests. *arXiv preprint arXiv :1810.07287*.

Lakkaraju, H., Kamar, E., Caruana, R. et Leskovec, J. (2019). Faithful and customizable explanations of black box models. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 131–138, New York, NY, USA. Association for Computing Machinery.

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X. et Detyniecki, M. (2017). Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv :1712.08443*.

Laugel, T., Renard, X., Lesot, M.-J., Marsala, C. et Detyniecki, M. (2018). Defining locality for surrogates in post-hoc interpretablity. *arXiv preprint arXiv :1806.07498*.

Lipton, Z. C. (2018). The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Lou, Y., Caruana, R. et Gehrke, J. (2012). Intelligible models for classification and regression. *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158.

Lundberg, S. M. et Lee, S.-I. (2017). A unified approach to interpreting model predictions. *In Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

Ma, X. (2018). *Using classification and regression trees : A practical primer*. IAP.

Malot-Tuleau, C. (2006). *Introduction à la sélection des variables*. Université Nice Sophia-Antipolis.

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. et Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Nelder, J. A. et Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society : Series A (General)*, 135(3):370–384.

Olden, J. D., Joy, M. K. et Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological modelling*, 178(3-4):389–397.

Perla, F., Richman, R., Scognamiglio, S. et Wüthrich, M. V. (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, 2021(7):572–598.

Plumb, G., Molitor, D. et Talwalkar, A. S. (2018). Model agnostic supervised local explanations. *Advances in neural information processing systems*, 31.

Ribeiro, M. T., Singh, S. et Guestrin, C. (2016). "why should i trust you ?" : Explaining the predictions of any classifier. *arXiv*, 1602.04938.

Ribeiro, M. T., Singh, S. et Guestrin, C. (2018). Anchors : High-precision model-agnostic explanations. *In Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Rice, L. et Swesnik, D. (2013). Discriminatory effects of credit scoring on communities of color. *Suffolk UL Rev.*, 46:935.

Richman, R. et Wüthrich, M. V. (2022). Localglmnet : interpretable deep learning for tabular data. *Scandinavian Actuarial Journal*, pages 1–25.

Robert, C. Y. (2018). Extrem value theory, cours de troisème année-voie actuariat. *ENSAE Paris, 3A*.

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions technip.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D. et Batra, D. (2016). Grad-cam : Why did you say that ? *arXiv preprint arXiv :1611.07450*.

Slack, D., Hilgard, S., Jia, E., Singh, S. et Lakkaraju, H. (2020). Fooling lime and shap : Adversarial attacks on post hoc explanation methods. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

So, B., Boucher, J.-P. et Valdez, E. A. (2021). Synthetic dataset generation of driver telematics. *Risks*, 9(4):58.

Sobol', I. M. (1990). On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280.

Tillé, Y. (2010). Balanced sampling by means of the cube method. *In Presentation to the International Statistical Seminar of EUSTAT, Bilbao, Basque Country*.

Tissot, J.-Y. et Prieur, C. (2012). Variance-based sensitivity analysis using harmonic analysis. *HAL*, 00680725.

Tsang, M., Cheng, D. et Liu, Y. (2017). Detecting statistical interactions from neural network weights. *arXiv preprint arXiv :1705.04977*.

Turner, R. (2016). A model explanation system. *In 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE.

Yang, C., Rangarajan, A. et Ranka, S. (2018). Global model interpretation via recursive partitioning. *In 2018 IEEE 20th International Conference on High Performance Computing and Communications ; IEEE 16th International Conference on Smart City ; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1563–1570. IEEE.

Zhang, Y., Song, K., Sun, Y., Tan, S. et Udell, M. (2019). " why should you trust my explanation ?" understanding uncertainty in lime explanations. *arXiv preprint arXiv :1904.12991*.