Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Elias BOUTAHAR

Titre : Application à la tarification automobile de méthodes de partitionnement récursif de modèles linéaires généralisés.

Confidentialité : ☐ OUI ☒ NON (Durée : ☐ 1 an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires *Signature*

. . . . . .

. . . . . .

. . . . . .

Membres présents du jury de l'ISFA

. . . . . .

. . . . . .

. . . . . .

Entreprise :
Nom : Prim'Act

Signature :

Directeur de mémoire en entreprise :
Nom : Quentin GUIBERT

Signature :

Directeur de mémoire en école :
Nom : Denys POMMERET

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

Signature du candidat

Résumé

Dans un contexte hyperconcurrentiel du marché de l'assurance automobile, les sociétés d'assurances recherchent l'optimisation de leur ratio combiné. Cette rentabilité s'agence autour des produits proposés par l'assureur. Ainsi, les assurés étant naturellement attirés par les tarifs les plus bas obligent les sociétés d'assurances à revoir leur tarification. Dernièrement, cet effet a été accentué par la loi Hamon (2015) facilitant la procédure de résiliation de contrat pour l'assuré. Par conséquent, les sociétés d'assurances se doivent d'être performantes sur l'estimation et la compréhension du tarif d'assurance automobile.

Afin d'estimer la prime pure, les assureurs utilisent classiquement les modèles *GLM* pour leur facilité d'exécution et leur interprétabilité permettant une bonne gestion du risque. Cependant, l'essor des modèles de *machine learning* plus performants que les modèles *GLM* étoffent le panel de modèles. La performance des modèles *machine learning* a un prix, une moins bonne interprétabilité empêchant une bonne gestion du risque essentiel aux assureurs. Dans l'optique de combiner performance et interprétabilité, les modèles *MOB* (*MOdel-Based recursive partitioning*) segmentent les données afin d'effectuer un modèle paramétrique sur chaque segments homogènes pouvant permettre de combiner ces deux aspects. Dans ce mémoire, le modèle paramétrique utilisé est le modèle *GLM*. Lors de la segmentation, la problématique de l'équilibre entre mutualisation et segmentation est un point primordial. En effet, une segmentation extrême n'est pas envisageable car la loi forte des grands nombres et le théorème central limite essentiels à la prédiction de la prime pure ne sont plus applicables. Ces contraintes opérationnelles obligent les assureurs à développer une multitude de modèles adaptables aux spécificités de leur portefeuille.

L'objectif du mémoire est d'étudier les modèles *GLM*, *CART*, *XGBoost* et *MOB* puis d'analyser la prime pure sur un portefeuille d'assurance automobile. Cette étude a permis de comparer ces différents modèles en termes de performance et d'interprétabilité. Les modèles *MOB* ont montré une efficacité sur des données très hétérogènes tout en conservant les aspects d'interprétabilités des modèles *GLM* et *CART*.

Mots-clés : *Tarification automobile*, *Segmentation*, *GLM*, *CART*, *MOB*, *GLM trees*, *Machine learning*, *XGBoost*, *SHAP*, *Interprétabilité*, *Explicabilité*.

Abstract

In a hyper-competitive context of the automobile insurance market, insurance companies are looking to optimize their combined ratio. This profitability revolves around the products offered by the insurer. Thus, policyholders being naturally attracted by the lowest rates, insurance companies are forced to review their pricing. Recently, this effect has been accentuated by the Hamon law (2015), which has facilitated the procedure of contract cancellation for policyholders. Therefore, insurance companies must be highly efficient when it comes to estimating and understanding the car insurance rate.

In order to estimate the pure premium, insurers conventionally use *GLM* models because they are easy to apply and to interpret, allowing good risk management. However, the rise of machine learning models, which are more efficient than *GLM* models, is expanding the range of models. The performance of machine learning models comes with a price: a lesser interpretability prevents a good risk management, essential to insurers. In order to combine performance and interpretability, the *MOB* (MOdel-Based recursive partitioning) models, segment the data in order to perform a parametric model on each homogeneous segment that could combine these two aspects. In this paper, the parametric model used is the *GLM* model. During the segmentation, the problem of the balance between mutualization and segmentation is a crucial point. Indeed, an extreme segmentation is not possible because the strong law of large numbers and the central limit theorem, both essential to predict the pure premium, are no longer applicable. These operational constraints force insurers to develop a multitude of models that can be adapted to the specificities of their portfolio.

The aim of this paper is to study the *GLM*, *CART*, *XGBoost* and *MOB* models and then evaluate the pure premium obtained with each of them on a car insurance portfolio. This study compared the models mentioned above in terms of performance and interpretability. The *MOB* models showed an efficiency on very heterogeneous data while keeping the interpretability aspect of the *GLM* and *CART* models.

Keywords: *Pricing*, *Segmentation*, *GLM*, *CART*, *MOB*, *GLM trees*, *Machine learning*, *XGBoost*, *SHAP*, *Interpretabily*, *Explicability*.

Note de Synthèse

Dans un contexte hyperconcurrentiel, les assureurs ont pour objectif d'optimiser la modélisation de la prime pure et de maîtriser l'interprétabilité des modèles pour une gestion du risque plus efficace. Une multitude de modèles sont testés afin de trouver le modèle adapté à leur portefeuille. Récemment, les modèles de *machine learning* ont démontré sur diverses applications une très bonne performance mais leur interprétabilité reste compliquée et utilise des méthodes indépendantes de la modélisation de ces modèles tel que l'algorithme *SHAP* ou *LIME*. Le modèle GLM reste le modèle classique car son interprétabilité à l'aide des coefficients permet de comprendre les variations des primes pures en fonction des modalités. Par ailleurs, le modèle *CART* est basé sur des arbres de décision regroupant des classes homogènes. L'idée de rassembler des individus avec le même comportement semble intuitivement être une bonne méthode, attention toutefois à ne pas effectuer une forte segmentation car les différents groupes doivent avoir des effectifs suffisamment grands afin d'appliquer le théorème central limite et la loi forte des grands nombres. Dans ce mémoire, nous introduisons les modèles *MOB* qui effectuent une segmentation des données et associent chaque segment à un modèle paramètrique. Dans le but d'optimiser la prédiction tout en conservant la qualité d'interprétabilité des modèles GLM, nous segmentons le jeu de données à l'aide des arbres de décision et au sein de chaque feuille nous estimons un modèle GLM. Ce modèle est un modèle GLM *trees* faisant parti de la classe des modèles *MOB*. Afin de comparer les performances de ce modèle avec celles des modèles de *machine learning*, nous allons étudier les modèles GLM, *CART*, *XGBoost* et *MOB*.

Cadre de l'étude

Les données utilisées sont des données automobile provenant d'un partenaire de Prim'Act. Avec ces données, nous allons modéliser la fréquence et le coût des sinistres. La base de données contient 108 729 observations et 27 variables. Par exemple, nous avons des variables telles que la formule auto étant un chiffre associé à un assuré, la zone étant la région du foyer de l'assuré ou encore son ancienneté de contrat. La mise en place de la modélisation nécessite trois échantillons. Le premier est utilisé dans la construction du modèle, le second dans l'optimisation des paramètres et le dernier pour la mesure de performance. Afin de conserver une cohérence dans la comparaison des performances, ces trois échantillons sont inchangés à travers les différents modèles. Pour chaque modèle, nous considérons un modèle fréquence et un modèle sévérité. La prime pure est obtenue en multipliant la fréquence et le coût estimés. La qualité de prédiction est mesurée par l'erreur moyenne quadratique (MSE) et de l'erreur moyenne absolue (MAE).

Présentation des différents modèles

Modèle GLM

Le modèle GLM est un grand classique de la modélisation s'expliquant par sa simplicité d'interprétation. L'effet des variables explicatives est quantifié par un coefficient permettant d'identifier les mauvais risques. Afin d'estimer la prime pure, nous modélisons deux modèles GLM, un modèle fréquence et un modèle coût sans sinistres extrêmes biaisant l'estimation. Le calcul de la prime pure s'effectue ainsi

$$Primepure = \mathbb{E}(Y_{freq}) \times \mathbb{E}(Y_{cout}) = \exp(\sum_{i=1}^{p} \beta_i X_i + \beta_0) \times \exp(\sum_{i=1}^{p} \beta_i' X_i + \beta_0'),$$

avec Y_{freq} la variable fréquence à expliquer, Y_{cout} la variable coût à expliquer, $X_i, 1 \leq i \leq p$ les variables explicative, $(\beta_1, ..., \beta_p)$ les coefficients du GLM fréquence, $(\beta_1', ..., \beta_p')$ les coefficients du GLM coût, β_0 l'*intercept* du modèle fréquence et β_0' l'*intercept* du modèle coût. Dans l'équation ci-dessus, les différents effets sont interprétables par les β_i et β_i' permettant de quantifier et contrôler le risque. L'optimisation du modèle GLM a été effectuée en sélectionnant les variables pertinentes par la méthode *forward* sur l'échantillon d'apprentissage.

Modèle *CART*

Le modèle *CART* est un algorithme d'arbre de décisions. Les arbres binaires segmentent le jeu de données afin de créer des feuilles terminales correspondant à un type de population. Dans le cas d'un arbre de régression (la variable Y est quantitative), nous affectons à chaque feuille une valeur étant la moyenne des observations associées à cette feuille pour l'échantillon d'apprentissage.
Concernant l'optimisation du modèle *CART*, nous avons modélisé avec l'échantillon d'apprentissage l'arbre maximal (le plus complexe). Cet arbre n'est pas le meilleur car une forte segmentation des données empêche l'application des lois statistiques. Par la suite, nous effectuons un élagage de l'arbre afin de sélectionner l'arbre optimal.

Modèle *XGBoost*

L'algorithme *XGBoost* est un algorithme ensembliste agrégeant des arbres de décision. A chaque itération, l'arbre construit apprend de l'erreur de son prédécesseur et la corrige dans le sens du gradient. Le modèle *XGBoost* s'optimise à travers ses différents hyperparamètres. Parmi eux se trouve, la profondeur de l'arbre maximal, le pourcentage d'observations utilisées pour construire un arbre, le pourcentage des variables utilisées pour construire un arbre, le taux d'apprentissage, le paramètre de lissage du modèle, le nombre d'observation minimal par nœud et le nombre d'arbres à implémenter. Les paramètres sont indépendants les uns envers les autres. Par conséquent, l'optimisation du modèle *XGBoost* a été effectué en testant 2700 combinaisons des paramètres. La combinaison minimisant le MSE par validation croisée sur l'échantillon d'apprentissage est sélectionnée.

Modèle *MOB*

L'algorithme *MOB* est un algorithme de partitionnement récursif qui permet d'estimer un modèle paramétrique dans chaque segment homogène. Dans notre cas, nous utilisons le modèle *glm trees* faisant parti de la famille des modèles *MOB*. Ces modèles utilisent un modèle GLM pour chaque feuille terminale. Par conséquent, parmi les variables explicatives, certaines sont sélectionnées en variables de classification et d'autres en variables de régression. Les variables de classification sont les variables décisionnelles sur la séparation du nœud et les variables de régression sont les variables utilisées par le modèle GLM dans chaque feuille. Dans l'algorithme *MOB*, un test statistique d'instabilité des paramètres du modèle GLM est effectué afin de décider la nécessité d'un découpage supplémentaire du nœud par rapport à une variable de classification. Cette propriété est intéressante car elle permet de fixer un critère d'arrêt dans le découpage et pourrait évité une segmentation trop fine. Afin d'optimiser le modèle *MOB* nous avons mis en place trois approches de sélection de variables :

- La première approche consiste à sélectionner les variables significatives dans *CART* (nous avons choisi d'en prendre deux) et à effectuer une régression sur les autres variables.
- La seconde approche consiste à prendre toutes les variables explicatives et de sélectionner une variable de classification parmi les variables, les autres variables seront utilisées en variables de régression. La variable minimisant notre erreur est sélectionnée en variable de classification idéale. Nous réitérons en testant les variables non utilisées dans la régression en variables de classification (avec maximum une variable en classification). L'algorithme s'arrête lorsque l'erreur d'une itération à l'autre ne baisse pas. Par la suite, nous conservons définitivement ces variables de régression et testons l'ajout de variables de classification non utilisées en régression.
- La troisième approche consiste à sélectionner des variables optimales et d'effectuer toutes les combinaisons possibles en variables de régression et de classification.

Pour les deux dernières approches, nous sélectionnons la combinaison minimisant le MSE sur l'échantillon de validation. Sur notre jeu de données, la deuxième approche donne la combinaison de variables de régression et classification minimisant le MSE.

Performance et interprétabilité des modèles

Comparaison des performances

Les différentes mesures évoquées par la suite permettent de comparer les performances de nos modèles.

Modèle	**MSE**	**RMSE**	**MAE**	**Temps de calcul**
GLM1	0,489	0,700	0,453	28s
GLM2	0,486	0,697	0,452	Instantané
CART	0,600	0,774	0,490	Instantané
XGBoost	0,603	0,777	0,551	1h 43mn
***MOB* approche 2**	**0,478**	**0,691**	**0,441**	**7h 29mn**

TABLE 1 – MSE, RMSE et MAE des différents modèles fréquence sur l'échantillon test

Sur le tableau 1, nous avons les différentes erreurs associées à chaque modèle que nous avons étudié précedemment. Le modèle GLM1 est un modèle GLM optimisé par la méthode classique *forward*

selection sélectionnant les meilleurs variables sur le critère AIC. Le modèle GLM2 est modèle GLM avec les mêmes variables que le modèle *MOB* (y compris les variables de classifications). Ce modèle nous permet de savoir si la meilleure performance du modèle *MOB* était dûe aux variables de régression ou à la segmentation. Nous observons que les modèles *CART* et *XGBoost* sont moins performants que le modèle GLM sur notre jeu de données. La mauvaise performance du modèle *XGBoost* peut s'expliquer par l'échantillonnage établi en amont, par le faible nombre de variables ou par le choix des hyperparamètres. Par ailleurs, le modèle *MOB* est le meilleur modèle fréquence sur notre jeu de données.

Modèle	**MSE**	**RMSE**	**MAE**	**Temps de calcul**
GLM1	1 760 525	1 327	779	35s
GLM2	1 758 095	1 326	779	Instantané
CART	1 766 018	1 329	781	Instantané
XGBoost	1 756 596	1 325	775	38mn
***MOB* approche 2**	**1 754 889**	**1 325**	**778**	**23h 12mn**

Table 2 – MSE, RMSE et MAE des différents modèles coût

Sur le tableau 2, nous avons les différentes erreurs associées à nos modèles coût. Le modèle *MOB* obtient encore une fois la meilleure performance en terme de RMSE. Le temps de calcul très important du modèle *MOB* est dû à notre algorithme. En effet, un modèle *MOB* peut prendre 30 minutes à être modélisé (voir plus en fonction du jeu de données) et notre algorithme teste énormément de modèles. Par conséquent, nous suggerons d'ajouter un argument à notre algorithme afin de s'arrêter au bout d'un certain temps et de prendre le meilleur modèle parmi toutes les combinaisons testées. Le modèle *XGBoost* a la meilleure performance en MAE pour le modèle coût.

Interprétabilité des modèles

Dans cette note, nous nous focalisons sur l'interprétabilité des modèles *MOB*.
Le modèle *MOB* fréquence possède 4 feuilles et le modèle *MOB* coût possède 2 feuilles avec respectivement un modèle GLM par feuille dont le détail est :

- La première feuille du modèle fréquence contient 24 009 observations. Ces données ont une ancienneté de contrat inférieure ou égale à 0,041.
- La seconde feuille du modèle fréquence contient 11 879 observations. Ces données ont une ancienneté de contrat comprise entre 0,041 et 1,777 avec 1,777 inclus.
- La troisième feuille du modèle fréquence contient 22 064 observations. Ces données ont une ancienneté de contrat comprise entre 1,777 et 7,943 avec 7,943 inclus.
- La dernière feuille du modèle fréquence contient 3 457 observations. Ces données ont une ancienneté de contrat strictement supérieur à 7,943.
- La première feuille du modèle coût contient 11 883 observations. Ces données ont une formule auto égale à 1, 3 ou 4.
- La seconde feuille du modèle coût contient 4 790 observations. Ces données ont une formule auto égale à 2 ou 5.

Ainsi, chaque feuille à son propre modèle GLM.

Modalité ou variable	Modèle fréquence				Modèle coût	
	Feuille 1	Feuille 2	Feuille 3	Feuille 4	Feuille 1	Feuille 2
Formule0	Référence	Référence	Référence	Référence	Non présent	Non présent
Formule1	0,5366	0,7156	0,3861	- 2,5680	Non présent	Non présent
Formule2	1,4679	1,6008	1,7294	- 1,0886	Non présent	Non présent
Formule3	1,6704	1,6810	1,6505	- 0,9369	Non présent	Non présent
Formule4	2,0436	1,9797	1,9858	- 0,1884	Non présent	Non présent
Formule5	1,4134	1,5596	1,6372	- 1,0127	Non présent	Non présent
auto_zone2	Référence	Référence	Référence	Référence	Référence	Référence
auto_zoneAUTRES	0,1353	0,1262	0,1288	- 1,5553	0,0823	0,2117
anc_contrat	Non présent	Non présent	Non présent	Non présent	- 0,0134	- 0,0012

TABLE 3 – Coefficients des modèles GLM en fonction de la feuille terminale

Le tableau 3 correspond aux différents coefficients proposés par le GLM pour chaque feuille. Ces coefficients permettent d'identifier les mauvais risques et d'avoir une relation linéaire entre les modalités d'une même feuille.

auto_formule	anc_contrat		anc_contrat		anc_contrat		anc_contrat		anc_contrat	
	0		1		2		7		8	
	auto_zone		auto_zone		auto_zone		auto_zone		auto_zone	
	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES
1	65	80	70	86	102	126	95	118	0	0
2	68	96	71	99	119	167	118	166	0	0
3	201	249	183	226	360	445	337	416	0	0
4	291	362	247	304	504	622	471	582	0	0
5	64	91	68	95	108	152	108	151	0	0

TABLE 4 – Grille de tarification pour le modèle *MOB*

Le tableau 4 correspond aux différents tarifs payés par les assurés en fonction de leur formule auto, leur zone et l'ancienneté de contrat. Avec les coefficients du tableau 3, nous pouvons comprendre l'impact des modalités. Nous allons procéder à un exemple afin de comprendre la différence de comportement créée par la segmentation. D'une part, nous prenons un assuré avec une ancienneté de contrat égale à 0, une formule auto égale à 1 et habitant dans la région associée à 2. Nous voulons connaitre le prix ainsi que l'effet engendré par une personne avec une ancienneté de contrat égale à 0, une formule auto égale à 1 et habitant dans la région associée à "AUTRES". Nous rappellons que ces individus appartiennent à la feuille 1 pour la fréquence et le coût. D'autre part, nous prenons un assuré avec une ancienneté de contrat égale à 0, une formule auto égale à 2 et habitant dans la région associée à 2. Nous voulons connaitre le prix et l'effet d'un assuré avec les mêmes caractéristiques hormis une région associée à "AUTRES". Ces individus appartiennent à la feuille 1 pour la fréquence et la feuille 2 pour le coût. Notons $PrimePure1$, $PrimePure1^{'}$ les primes pures que devraient payer les 2 individus associés à la première feuille du modèle *MOB* fréquence et coût et $PrimePure2$, $PrimePure2^{'}$ les primes pures que devraient payer les 2 individus associés à la première feuille du modèle *MOB* fréquence et la deuxième feuille du modèle *MOB* coût.

$$\begin{aligned} Primepure1^{'} &= Primepure1 \times \exp(0,1353) \times \exp(0,0823) \\ &= Primepure1 \times \exp(0,2176) \\ &= 65 \times 1,24309 \approx 80 \end{aligned}$$

$$\begin{aligned} Primepure2^{'} &= Primepure2 \times \exp(0,1353) \times \exp(0,2117) \\ &= Primepure2 \times \exp(0,3470) \\ &= 68 \times 1,414817 \approx 96 \end{aligned}$$

Cet exemple illustre la différence de comportement du changement de zone en fonction des feuilles. Nous précisons que dans un modèle GLM classique, les coefficients de passage entre ces primes pures ne sont pas identiques car un changement de formule (passage d'une auto formule 1 à 2) est constaté ajoutant un effet. Cet effet est linéaire car le passage d'une modalité à une autre dans un modèle GLM se fait en multipliant par le même coefficient. Dans le modèle *MOB*, le fait de changer de formule auto peut entrainer un changement de feuille et une modification de tous les coefficients créant un effet non linéaire. De plus, nous avons modifié notre base de données de sorte que :

- Le cout moyen d'un sinistre pour les personnes âgées de moins de 40 ans strictement est augmenté de 4 000.
- Le coût moyen d'un sinistre pour les personnes âgées de moins de 40 ans strictement et possédant une voiture avec 7 chevaux fiscaux est augmenté de 7 000.

Une application du modèle *MOB* sur ce jeu de données a permis de capter la segmentation car parmi les segmentations effectuées, une feuille est assimilée à chaque groupe modifié. Ainsi, le modèle *MOB* a permis une baisse de 22 % du MSE sur le deuxième meilleur modèle.

Conclusion

Le modèle *MOB* a montré une forte efficacité sur des données hétérogènes tout en conservant la puissance d'interprétabilité des modèles GLM et *CART* afin d'avoir une bonne connaissance des différents risques. Sur nos données étant homogènes, le modèle *MOB* apporte tout de même une amélioration. Les contraintes liées à ce modèle restent l'optimisation des variables de régression et de classification ainsi que l'optimisation des différents paramètres similaires à ceux du modèle *CART*. Naturellement, nous pouvons penser qu'une segmentation très fine entraine une meilleure performance. Cet instinct est faux car une segmentation forte de ces données diminue le nombre d'observations par segment empêchant l'application de la loi forte des grands nombres et du théorème central limite essentiels à l'inférence statistique.

Synthesis note

In a hyper-competitive context, insurers aim to optimize the modeling of their pure premium and to master the interpretability of models for a more efficient risk management.
A multitude of models are tested in order to find the right model for their portfolio. Recently, machine learning models have demonstrated great performances on various applications, but their interpretability remains complicated and uses methods independent of the modeling of these models such as the *SHAP* or *LIME* algorithm.
Conventionally, the GLM model remains the classic model because its interpretability using coefficients allows us to understand the variations in pure premiums. On the other hand, the CART model is based on decision trees regrouping homogeneous classes. The idea of grouping individuals with the same behavior intuitively seems to be a good and valid method, but caution is advised not to perform a strong segmentation because the different groups must be large enough to apply the central limit theorem and the strong law of large numbers. In this paper we will introduce MOB models. These models perform a segmentation and associate each segment to a parametric model. In order to optimize the prediction while keeping the quality of interpretability of GLM models, we will perform the segmentation by using decision trees and by estimating a GLM model at each terminal leaf. This model is a GLM trees model belonging to the class of MOB models. In order to compare the performances of this model with those of machine learning models we will study the GLM, CART, XGBoost and MOB models.

Scope of the study

The data used is automotive data from a Prim'Act partner. With this data, we will model the frequency and cost of claims. The database contains 108,729 observations and 27 variables. For example, we have variables such as the car formula being a number associated with a policyholder, the zone being the region of the policyholder's home or the policyholder's seniority
The implementation of the model requires three samples. The first one is used in the construction of the model, the second in the optimization of the parameters and the last one for the performance quality evaluation. In order to maintain the performances comparison coherent, these three samples are unchanged across the different models. For each model, we consider a frequency model and a severity model. The pure premium is obtained by multiplying the estimated frequency and cost. The prediction quality is given by the mean squared error (MSE) and the mean absolute error (MAE).

Presentation of the different models

GLM model

The GLM model is a classic in modeling because of its simplicity of interpretation. The effect of the explanatory variables is quantified by a coefficient allowing to identify bad risks. In order to estimate the pure premium, we model two GLM models, a frequency model and a cost model without extreme

claims biasing the estimation. The calculation of the pure premium is done as follows:

$$Purepremiums = \mathbb{E}(Y_{freq}) \times \mathbb{E}(Y_{cout}) = \exp(\sum_{i=1}^{p} \beta_i X_i + \beta_0) \times \exp(\sum_{i=1}^{p} \beta_i' X_i + \beta_0'),$$

With Y_{freq} the frequency variable to be explained, Y_{cost} the cost variable to be explained, $X_i, 1 \leq i \leq p$ are te explanatory variables, $(\beta_1, ..., \beta_p)$ are the coefficients of the GLM frequency,$(\beta_1', ..., \beta_p')$ are the coefficients of the GLM cost, β_0 the intercept of the frequency model and β_0' the intercept of the cost model. In the above equation, the different effects can be interpreted with the coefficients β_i and β_i', allowing to quantify and control the risk. The optimization of the GLM model was performed selecting the variables via forward stepwise selection on the training sample.

CART model

The CART model is a decision tree algorithm. These binary trees segment the dataset to create terminal leaves corresponding to a population type. In the case of a regression tree (the variable Y is quantitative), we assign to each leaf a value being the average of the observations associated with that leaf for the training sample.
Regarding the optimization of the CART model, we have modeled using the training sample the maximal tree (the most complex one). This tree is not the best one because a strong segmentation of the data prevents the application of statistical laws. Then, we prune the tree to select the optimal tree. The errors of the model on the test sample are the following:

XGBoost model

The XGBoost algorithm is a set algorithm aggregating decision trees. At each iteration, the constructed tree learns from the error of its predecessor and corrects it in the direction of the gradient.
The XGBoost model is optimized through these different hyperparameters. Among them we find the depth of the maximum tree, the percentage of observations used to build a tree, the percentage of variables used to build a tree, a learning rate, a model smoothing parameter, a minimum number of observations per node and the number of trees to implement.
The parameters are independent of each other. Therefore, the optimization of the XGBoost model was performed by testing 2700 parameters. The combination minimizing the MSE by cross validation on the apprentissage sample is selected. The *MOB* algorithm is a recursive partitioning algorithm which allows to estimate a parametric model in each homogeneous segment. In our case, we use the *glm trees* model which is part of the *MOB* family of models. These models use a GLM model for each terminal leaf. Among the explanatory variables, some are selected as classification variables and others as regression variables. The classifications variables are the decision variables on the node separation and the regression variables are the variables used by the GLM model in each terminal leaf. In the MOB algorithm, a statistical test of instability of the parameters of the GLM model is carried out in order to decide the need for an additional division of the node with respect to a classification variable. This property is interesting because it allows to set a stopping criterion in the cutting and avoid a too fine segmentation. In order to optimize the MOB model, we implemented 3 methods of variable selection:

— The first method consists in selecting the significant variables in CART (we chose to take two) and then performing a regression on the other variables.
— The second method consists in taking all the explanatory variables and selecting a classification variable among the variables, the other variables will be used as regression variables. The variable minimizing our error is selected as the ideal classification variable. We reiterate by testing the variables not used in the regression as classification variables (with maximum one

variable in classification). The algorithm stops when the error from one iteration to the next does not decrease. Afterwards, we definitively keep these regression variables and test the addition of classification variables not used in the regressions.

— The third method consists in selecting optimal variables and performing all possible combinations of regression and classification variables.

For this last two methods, We select the combination that minimizes the MSE on validation sample. On this dataset, the 2nd method gives the combination of regression and classification variables minimizing the MSE.

Performance and interpretability of the models

Performance of the models

The different measures previously mentioned allow us to compare the performance of our models.

Model	MSE	RMSE	MAE	Calculation time
GLM1	0,489	0,700	0,453	28s
GLM2	0,486	0,697	0,452	Instant
CART	0,600	0,774	0,490	Instant
XGBoost	0,603	0,777	0,551	1h 43mn
***MOB* method 2**	**0,478**	**0,691**	**0,441**	**7h 29mn**

Table 5 – MSE, RMSE and MAE of the different frequency models

In the table 5, we have the different errors associated to each model we have studied previously. The GLM1 model is a GLM model optimized by the classical method forward selection, selecting the best variables on the AIC criteria. The GLM2 model is a GLM model with the same variables as the MOB model (including the classification variables). This model allows us to know if the better performance of the MOB model was due to the regression variables or to the segmentation. We observe that the CART and XGBoost models perform worse than the GLM model on our dataset. The poor performance of the XGBoost model can be explained by the upstream sampling, by the small number of variables or by the choice of hyperparameters. On the other hand, the MOB model is the best frequency model on our dataset.

Model	MSE	RMSE	MAE	Calculation time
GLM1	1 760 525	1 327	779	35s
GLM2	1 758 095	1 326	779	Instant
CART	1 766 018	1 329	781	Instant
XGBoost	1 756 596	1 325	775	38mn
***MOB* method 2**	**1 754 889**	**1 325**	**778**	**23h 12mn**

Table 6 – MSE, RMSE and MAE of the different cost models

In table 2, we have the different errors associated to our cost models. The MOB model obtains once again the best performance in terms of RMSE. The very important computation time of the MOB model is due to our algorithm. Indeed, a MOB model can take 30 minutes to be modeled (or more depending on the dataset) and our algorithm tests a lot of models. Therefore, we suggest to add an argument to our algorithm in order to stop after a certain time and to take the best model among all tested combinations. The XGBoost model has the best performance in MAE for the cost model.

Interpretability of the models

The study of the interpretability of the GLM, *CART* and *XGBoost* models is performed in the chapter 5. The MOB frequency model has 3 sheets and the MOB cost model has 2 sheets with respectively one GLM model per sheet with the following details:

- The first sheet of the frequency model corresponds to the insureds with a contract length under 0.041.
- The second sheet of the frequency model corresponds to the insureds with a contract length between 0.042 and 1.777.
- The third sheet of the frequency model corresponds to policyholders with a contract length over 1.778 and 7.943.
- The last sheet of the frequency model corresponds to policyholders with a contract length over 7.944.
- The first sheet of the cost model corresponds to the policyholders with a car formula equal to 1, 3 and 4.
- The second sheet of the cost model corresponds to the policyholders with a car formula equal to 2 and 5.

Variable	Frenquency models				Cost models	
	Leaf 1	Leaf 2	Leaf 3	Leaf 4	Leaf 1	Leaf 2
Formule0	Reference	Reference	Reference	Reference	Not present	Not present
Formule1	- 7,1317	- 7,5369	- 6,7300	4,7061	Not present	Not present
Formule2	0,5366	0,7156	0,3861	- 2,5680	Not present	Not present
Formule3	1,4679	1,6009	1,4000	- 0,9369	Not present	Not present
Formule4	1,6704	1,6810	1,6505	- 0,1884	Not present	Not present
Formule5	2,0436	1,9797	1,9858	- 0,1987	Not present	Not present
auto_zone2	Reference	Reference	Reference	Reference	Reference	Reference
auto_zoneAUTRES	- 0,0676	- 0,0631	- 0,0644	0,0603	0,0823	0,2117
anc_contrat	Non présent	Non présent	Non présent	Non présent	- 0,0134	- 0,0012

Table 7 – GLM coefficient as a function of the end leaf

Thus, each sheet has its own GLM model.

auto_formule	anc_contrat		anc_contrat		anc_contrat		anc_contrat		anc_contrat	
	0		1		2		7		8	
	auto_zone		auto_zone		auto_zone		auto_zone		auto_zone	
	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES
1	65	80	70	86	102	126	95	118	0	0
2	68	96	71	99	119	167	118	166	0	0
3	201	249	183	226	360	445	337	416	0	0
4	291	362	247	304	504	622	471	582	0	0
5	64	91	68	95	108	152	108	151	0	0

Table 8 – Pricing grid for the MOB model

Table 7 corresponds to the different coefficients proposed by the GLM for each leaf. These coefficients allow us to identify bad risks and to have a linear relationship between the modalities of the same leaf. Table 8 shows the different rates paid by policyholders according to their car formula, their zone and the length of time they have been insured. With the coefficients in table 7, we can understand the impact of the modalities. We will proceed with an example in order to understand the difference in behavior created by segmentation. On the one hand, we take a policyholder with a contract length equal to 0, an car formule equal to 1 and living in the region associated with 2. We want to know the price as well as the effect generated by a person with a contract length equal to 0, an car formule

equal to 1 and living in the region associated with "OTHER". We recall that these individuals belong to sheet 1 for frequency and cost. On the other hand, we take an insured with a contract length equal to 0, a car formule equal to 2 and living in the region associated to 2. We want to know the price and effect of an insured with the same characteristics except for a region associated with "OTHER". These individuals belong to sheet 1 for the frequency and sheet 2 for the cost. Let us note $Purepremium1$, $Purepremium1^{'}$ the pure premiums that should be paid by the 2 individuals associated to the first sheet of the model *MOB* frequency and cost and $Purepremium2$, $Purepremium2^{'}$ the pure premiums that should be paid by the 2 individuals associated to the first sheet of the model *MOB* frequency and the second sheet of the model *MOB* cost.

$$\begin{aligned} Purepremium1^{'} &= Purepremium1 \times \exp(0.1353) \times \exp(0.0823) \\ &= Purepremium1 \times \exp(0.2176) \\ &= 65 \times 1.24309 \approx 80 \end{aligned}$$

$$\begin{aligned} Purepremium2^{'} &= Purepremium2 \times \exp(0.1353) \times \exp(0.2117) \\ &= Purepremium2 \times \exp(0.3470) \\ &= 68 \times 1.414817 \approx 96 \end{aligned}$$

This example illustrates the difference in the behavior of the zone change according to the leaves. We specify that in a classical GLM model, the coefficients of passage between these pure premiums are not identical because a change of car formula (passage from a car formule 1 to 2) is seen adding an effect. This effect is linear because the passage from one modality to another in a GLM model is done by multiplying by the same coefficient. In the model *MOB*, the fact of changing the car formula can lead to a change of sheet and a modification of all the coefficients creating a non linear effect. In addition, we modified our database so that:

- The average cost of a claim for people strictly under 40 is increased by 4,000.
- The average cost of a claim for people strictly under 40 and owning a car with 7 horsepower is increased by 7,000.

An application of the MOB model on this dataset captured the segmentation because among the segmentations performed, one leaf is assimilated to each modified group. Thus, the MOB model resulted in a 22% decrease of the MSE on the second best model.

Conclusion

The MOB model showed a strong efficiency on heterogeneous data while keeping the interpretability power of the GLM and CART models in order to have a good knowledge of the different risks. On our data being homogeneous, the MOB model still brings an improvement. The constraints related to this model remain the optimization of the regression and classification variables as well as the optimization of the various parameters similar to those of the CART model. Naturally, we may think that a very fine segmentation leads to a better performance. This instinct is false because a strong segmentation of the data decreases the number of observations per segment preventing the application of the strong law of large numbers and the central limit theorem essential to the prediction.

Remerciements

Je tiens tout d'abord à remercier Prim'Act pour m'avoir accueilli dans leur équipe.

Un grand merci à Quentin Guibert, consultant chez Prim'Act et professeur associé à l'université Paris Dauphine, pour son investissement incroyable et ses multiples conseils très pertinents ainsi qu'à Frédéric Planchet, directeur associé chez Prim'Act, professeur à l'Institut de Science Financière et d'Assurances et membre agrégé de l'institut des actuaires, pour la confiance et l'intérêt qu'il a su porter à mon travail.

Je tiens à remercier Denys Pommeret professeur à l'Institut de Science Financière et d'Assurances et à l'université Aix-Marseille pour ses conseils, son soutien et la confiance qu'il m'a accordé durant ces quatre dernières années.

Je remercie Pierre Bousseau, mon tuteur chez Prim'Act, pour son soutien et ses précieux conseils. Une attention particulière à Taline Tosbath et Maxime Ben Brik pour leurs aides ainsi que tous les membres de Prim'Act pour leurs conseils et leur convivialité.

Merci à mon camarade de classe et ami, David Delrio actuaire chez AXA pôle pilotage et actuariat en IARD pour son aide et ses multiples conseils.

Enfin, j'ai une pensée particulière pour mon père Mohamed Boutahar maître des conférences à l'Institut de Mathematiques de Marseille ainsi que ma mère Latifa Boutahar pour leur présence, aide et soutien tout le long de ma scolarité et bien plus.

Introduction

En 2019, l'assurance automobile représente un chiffre d'affaires de $22,8$ milliards d'euros en France. Cette importance s'explique par l'obligation pour chaque individu possédant un véhicule de souscrire à un contrat d'assurance automobile au tiers. Le grand nombre d'assurés potentiels fait de ce marché un marché hyperconcurrentiel où conserver ses assurés et en attirer de nouveaux sont un véritable challenge. Cette concurrence est accentuée par la loi Hamon de 2015 facilitant les résiliations de contrat d'assurance automobile par les assurés. Par conséquent, les assureurs redoublent d'effort afin de mettre en place des tarifs avantageux tout en développant de nouvelles garanties afin de satisfaire leurs clients. Cette concurrence omniprésente nécessite pour l'assureur d'effectuer une tarification précise et adaptée à son portefeuille. Cette tarification repose sur une analyse de son portefeuille et d'une segmentation plus ou moins fine de ce dernier. La segmentation permet de créer des sous-populations porteuses d'un risque commun et de leur affecter un tarif en fonction du risque. La segmentation ne doit pas pencher vers la discrimination car le principe de base de l'assurance est la mutualisation des risques. Cette analyse doit être effectuée en adéquation avec les objectifs de l'entreprise et un suivi continu doit être établi. La prime pure d'assurance automobile est modélisée classiquement par des modèles de régression linéaire. Ces modèles ont l'avantage d'être interprétables permettant une gestion du risque adaptée. Parallèlement, l'essor des méthodes machine learning en actuariat fait envisager de nouvelles possibilités de modélisation. Ces modèles ont démontré sur une multitude d'applications de meilleurs performances que les modèles de régression linéaire mais leur interprétabilité reste compliquée. Par conséquent, les modèles *GLM* restent le modèle phare car la non interprétabilité des modèles est rédhibitoire. Le but de ce mémoire est de concurrencer les modèles *machine learning* tout en conservant la puissance d'interprétabilité des modèles *GLM*. Pour ce faire, nous introduisons le modèle *MOdel-Based recursive partitioning*. Ce modèle segmente les données et effectue un modèle statistique sur chaque segment homogène. Dans ce mémoire, nous considérons le modèle *GLM tree* faisant parti de la famille des modèles *MOB* et utilisant des modèles *GLM* pour chaque feuille terminale. Un modèle *GLM* pour chaque segment homogène peut s'avérer plus performant qu'un modèle *GLM* sur l'ensemble des données.

Pour ce faire, nous commençons par présenter les différentes notions de tarification en insistant sur la segmentation essentielle à la modélisation. Par la suite, nous présentons et analysons statistiquement la base de données. La troisième partie consiste à étudier le modèle *GLM* référence. Par la suite, nous étudions les modèles *CART*, *XGBoost* et *MOB* tout en comparant les performances de ces différents modèles. La dernière partie est une analyse de l'interprétabilité des modèles.

Table des matières

Chapitre 1

Contexte de l'étude

1.1 Le marché de l'assurance automobile

L'assurance se décompose en deux catégories en fonction de la nature du risque : l'assurance de personnes d'une part et l'assurance de biens et de responsabilité de l'autre. En 2019[35], les chiffres d'affaires de ces deux catégories sont respectivement 169, 3 milliards d'euros et 58, 6 milliards d'euros soit un total de 227, 9 milliards d'euros de cotisations pour les assurances françaises.

Dans ce mémoire, nous allons nous intéresser aux assurances de biens et de responsabilité incluant l'assurance automobile. Les assurances de bien et de responsabilité regroupent les branches d'assurances n'étant pas directement liées à la vie humaine. Appelé par l'abréviation IARD, (Incendie Accident et Risques Divers), ces assurances ont pour but de protéger et indemniser les entreprises et particuliers contre des sinistres non liés à la vie humaine. Dans cette partie, les différents chiffres sont issus de la fédération française des assurances[35]. Les cotisations dédiées aux assurances de biens et de responsabilités constituées des branches automobiles, biens des particuliers, biens des professionnels et agricoles, responsable civile générale, construction, catastrophes naturelles, transports et d'autres branches non-vie représentent 58, 6 milliards d'euros en 2019 dont 40% propre à la branche automobile soit 22, 8 milliards d'euros. Les cotisations dans le secteur de l'automobile représentent un dixième des cotisations totales des assurances françaises et la moitié des cotisations assurances de biens et de responsabilité. Dans ce secteur, les cotisations ont augmenté de 3,2 % en 2019. Ces chiffres peuvent s'expliquer par l'inflation.

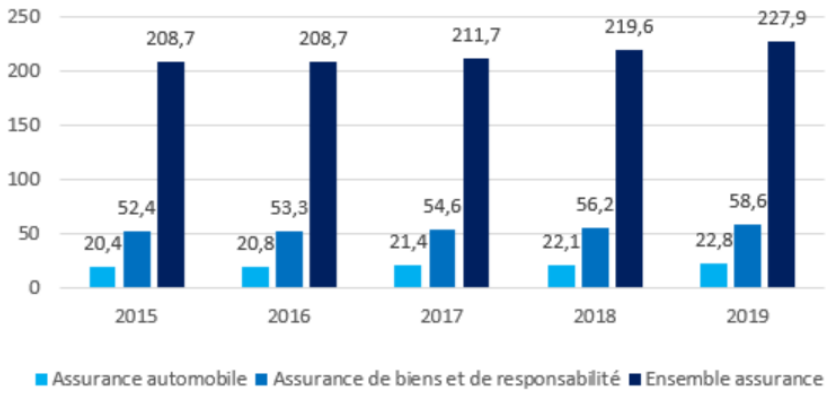

FIGURE 1.1 – Cotisations des assurances françaises en milliards d'euros

Sur le graphique 1.1, nous observons que de 2015 à 2019 les cotisations des assurances françaises, de biens et de responsabilité et automobile, augmentent d'une année à l'autre. Cette augmentation ne dépasse pas les 5%. Chaque année, la proportion des cotisations d'assurance automobile correspond à 10% des cotisations totales.
En 2019, les prestations sont de $183,7$ milliards d'euros pour l'ensemble des assurances françaises. Les prestations en assurance de biens et de responsabilité représentent $42,1$ milliards d'euros dont $17,5$ milliards d'euros pour l'assurance automobile. Dans ce dernier secteur, les prestations ont augmenté de $6,25\%$.

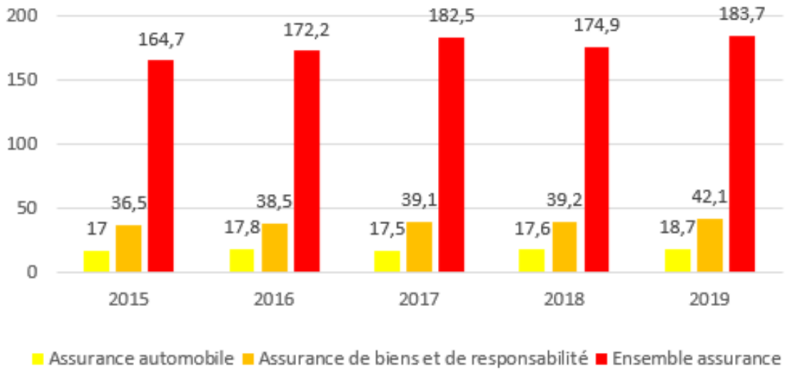

FIGURE 1.2 – Prestations des assurances françaises en milliards d'euros

Sur le graphique 1.2, nous observons que de 2015 à 2019, les prestations des assurances automobile augmentent d'une année à l'autre excepté en 2018 où les prestations totales des assurances françaises stagnent par rapport à celles de 2017. Ces chiffres peuvent s'expliquer par l'inflation combinée à la sensibilisation des automobilistes. En 2018, on observe une baisse de la mortalité de 5,3 % et une baisse des accidents corporels de 4,9 % et le nombre de blessés de 4,8 %. Ces baisses s'expliquent par des campagnes de sensibilisation des conducteurs et surtout par la baisse à 80 km/h de la vitesse maximale autorisée permettant d'épargner 127[30] vies en six mois. Tout comme les cotisations, les prestations en assurance automobile correspondent à un dixième des prestations totales des assurances françaises et la moitié des prestations de biens et de responsabilité. Ces ordres de grandeur montrent l'importance économique du marché de l'assurance automobile ainsi que les constantes évolutions entrainant diverses études à ce sujet.
Comme vu précedemment, l'assurance automobile est un marché très important car chaque usagé a l'obligation de souscrire à un contrat d'assurance automobile au tiers mais aussi par la volonté de certains assurés de souscrire à certaines garanties apportant une meilleure couverture. On compte notamment parmi celles-ci :

— La responsabilité civile appelée assurance au tiers. Il s'agit de la garantie minimale et obligatoire qui couvre les dommages corporels et matériels. Le responsable de l'accident ne sera pas indemnisé pour les dommages subis. La responsabilité civile ne couvre pas les dommages subis par le véhicule assuré, les dommages subis sur un circuit privé, les dommages concernant des biens et animaux appartenant au conducteur et les dommages causés par les professionnels de la réparation et du contrôle technique.
— Les garanties dommages prennent en charge l'endommagement du véhicule suite à un accident ou un acte de vandalisme. Parmi elles, nous avons :

1. La garantie bris de glace couvre le remplacement des parties vitrées du véhicule. Cette garantie s'applique si les dommages sont la conséquence d'actes de vandalisme ou de chocs avec un piéton ou un autre véhicule.

2. La garantie vol rembourse le véhicule en cas de vol. Certaines compagnies installent des mesures de sécurité (garage, traçage du véhicule,..). Cette garantie exclue les vols lorsque le conducteur a laissé les clés à l'intérieur du véhicule. L'assuré peut assurer le contenu du véhicule dans une limite imposée par le contrat. Des mesures strictes sont prises contre les vols frauduleux.

3. La garantie incendie couvre le véhicule contre les destructions causées par les flammes. En cas de catastrophes naturelles, cette garantie peut s'activer si un arrêté ministériel est fait. Cette garantie peut s'annuler si les flammes ont été causé à cause d'une cigarette.

4. La garantie collision couvre contre les dommages causés au véhicule à la suite d'une collision avec un autre véhicule ou un objet mobile. Elle ne couvre pas les accidents causés par le conducteur seul ou les collisions avec un véhicule faisant un délit de fuite.

Après avoir explicité la responsabilité civile obligatoire et les garanties dommages, d'autres garanties complémentaires existent telle que la garantie protection juridique couvrant les frais de justice, de défense ou d'expertise si l'assuré est convoqué devant les tribunaux ou encore la garantie assistance permettant en cas de panne de prendre en charge le remorqueur. Le conducteur peut aussi souscrire à des garanties personnelles couvrant le décès ou l'invalidité. Nous pouvons voir sur le graphique 1.3 que les garanties facultatives (dommages) sont non négligeables car elles représentent 64 % des cotisations en assurance automobile.

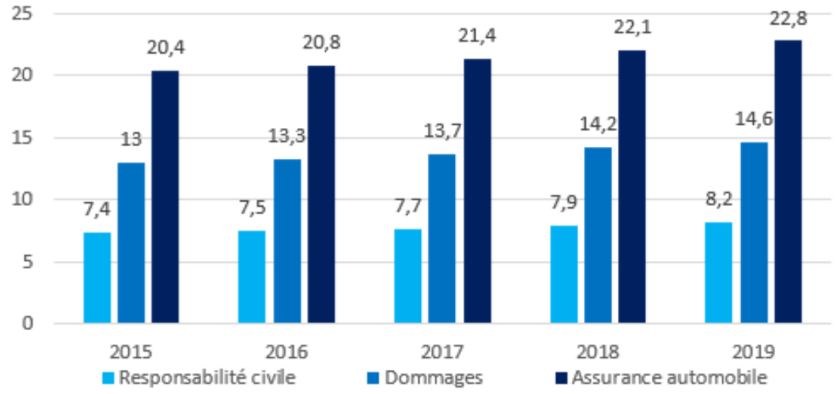

FIGURE 1.3 – Cotisations des assurances françaises en milliards d'euros

La compétitivité des assurances nous amène à optimiser et créer de nouveaux modèles afin d'attirer le plus d'assurés possibles et de redresser les résultats techniques. Cette compétitivité est accentuée par l'entrée le 1er janvier 2015 de la loi Hamon permettant à l'assuré de résilier son contrat à tout instant. Les assurances doivent redoubler d'efforts afin de proposer des prix concurrentiels et rentables. Pour ce faire, les assureurs mettent à jour régulièrement leur produit. La conception du produit se fait par des échanges entre les différentes directions de la compagnie d'assurance. La première étape consiste à examiner les besoins des clients et de les cibler. Par la suite, les actuaires effectuent la tarification du produit, il s'agit de l'objet de ce mémoire. Enfin, le pôle juridique rédige des documents contractuels du produit créé en s'assurant que les clauses soient claires et conformes à la législation. Le produit peut être proposé aux assurés via un réseau de distribution (courtiers, agents généraux, internet, ...). Ces modifications peuvent entrainer un ajustement du tarif géré par l'actuaire. Pour finir, les produits d'assurances ont un suivi constant. Un reporting est fait afin d'analyser la rentabilité du produit, la satisfaction des clients et d'identifier les nouveaux besoins.

1.2 Mutualisation et segmentation

Contexte assurentiel

La mutualisation des risques est le fondement du système d'assurance. Historiquement, l'assurance s'est construit sur la solidarité d'un groupe d'individu soumis au même risque. Ces individus payent des cotisations à l'assurance afin de régler des sinistres qui surviennent à quelques-un seulement. Sans la mutualisation, un individu devrait régler seul un sinistre pouvant attendre des centaines voir des milliers d'euros ce qui est impossible.

Actuellement, les sociétés d'assurance disposent d'un volume de données colossales et d'un large panel de méthodes de modélisation. Dans un contexte de forte concurrence, ces récentes évolutions obligent les assurances à segmenter de plus en plus leur tarif afin de proposer le tarif le plus adéquat et concurrentiel au risque. Rappelons que la problématique en assurance est la méconnaissance du coût des sinistres futurs. Ce procédé se nomme inversion du cycle de production. Par conséquent, la prime pure doit être modélisée en fonction de l'estimation du coût des sinistres futurs.

Principes de mutualisation versus segmentation

La segmentation est une technique permettant à un assureur de classer les risques selon certains critères pour établir son tarif et/ou déterminer les modalités des garanties offertes. La segmentation se fait en adéquation avec le profil de son portefeuille d'assurés et des différents risques. La segmentation engendre différentes catégories de risques en fonction des caractéristiques de l'assuré et des garanties couvertes par l'assureur. Chaque catégorie subira une mutualisation afin d'obtenir un tarif en adéquation avec le risque. Donnons un exemple concret afin de comprendre l'importance de la segmentation.

Prenons un assureur A n'effectuant aucune segmentation. Tous les assurés de l'assureur A paieront le même tarif avec une prime égale à l'espérance mathématique de la charge annuelle. Un profit est effectué sur les assurés à faible probabilité de sinistre et une perte est effectuée sur les assurés à forte probabilité de sinistre.

Prenons un assureur B effectuant une segmentation. Un assuré ayant un profil de risque à faible probabilité de sinistre paiera moins cher qu'un assuré à profil de risque à forte probabilité de sinistre. Un assuré à faible probabilité de sinistre choisira d'aller chez l'assureur B car il paiera moins cher et un assuré à forte probabilité de sinistre choisira d'aller chez l'assureur A.

Finalement, l'assureur A qui n'aura effectué aucune segmentation attirera les mauvais risques alors que l'assureur B ayant effectué une segmentation attirera les bons risques ce qui entrainera des pertes pour l'assureur A et du profit pour l'assureur B. Avec cet exemple, nous comprenons que les "bons risques" sont attirés par des tarifs segmentés car plus avantageux pour eux. À contrario, les "mauvais risques" sont attirés par des tarifs mutualisés car plus avantageux pour eux.

Les limites de la segmentation

Une segmentation excessive n'est pas enviseable car des limites sont identifiées.

— La limite principale de la segmentation est la connaissance des risques du portefeuille pour que la loi des grands nombre et le théorème central limite puissent s'appliquer. Par conséquent, une classe de risque homogène doit être assez grande afin de pouvoir utiliser ces théorèmes essentiels au calcul de la prime pure. En pratique, la connaissance du portefeuille n'est pas parfaite. Plus nous segmentons, plus la connaissance du risque diminue et augmente l'incertitude de l'estimation. Statistiquement, la clé essentielle à l'estimation est le volume de données afin d'appliquer la loi des grands nombres et de contôler la volatilité de l'estimateur.

— Une segmentation fine entraine des modèles complexes : la compléxité d'un modèle augmente les risques opérationnels d'estimation entrainant une augmentation des tarifs ce qui va à l'encontre du point fort de la segmentation : avoir un tarif plus attractif.
— Une segmentation avec un modèle inapté peut être dangereux car la déformation du portefeuille dans le temps biaiserait l'estimation.
— Une segmentation fine apporte théoriquement un tarif plus juste mais l'instabilité est forte et la gestion du risque moins efficace.
— La segmentation reste compliquée dans certains cas, le cas des jeunes conducteurs pose souvent problème car ils n'ont pas d'ancienneté d'assurance et sont donc considérés comme une catégorie à mauvais risque. D'autres mauvais risques existent tels que les conducteurs faisant énormément de sinistres. Ces catégories de mauvais conducteurs font souvent l'objet de refus auprès des assurances classiques et doivent se tourner vers des assurances spécialisées dans les mauvais risques. Si aucun accord n'est trouvé, ces assurés peuvent se tourner vers le Bureau Central de Tarification qui obligera un assureur à couvrir la responsabilité civile étant obligatoire.

Ces différents éléments convergent vers le fait qu'une segmentation fine n'offre aucune garantie de tarif plus concurrentiel.

Un équilibre entre la mutualisation et la segmentation

Théoriquement, la mutualisation est efficace si l'application du théorème central limite est faisable. Cette condition dépend du volume de données et est facilement respectée. La segmentation permet d'avoir des tarifs concurrentiels car plus nous segmentons plus la somme des primes pures diminuent. Une forte segmentation exige une maîtrise des risques techniques et opérationnels. Un équilibre doit être trouvé entre la mutualisation et la segmentation afin de combiner tarif concurrentiel et contrôle du risque. Au delà de l'aspect économique, un aspect éthique subsiste. La question sur la segmentation d'un tarif provient de la pression concurrentielle. En effet, les assureurs qui appliquent une segmentation forte des tarifs voient leur somme des primes pures baisser. Ainsi, les assureurs concurrents sont incités à segmenter de manière plus fine afin de proposer des tarifs plus concurrentiels et de conserver les bons risques. La logique de la segmentation très forte va à l'encontre du principe de base d'un contrat d'assurance étant « d'assurer une mutualisation des risques par la mise en commun des risques de personnes soumises à des risques équivalents mais indépendants »[24]. Une segmentation forte entrainerait une prime trop élévée pour les conducteurs à risques et même une exclusion de ces assurés. Pour ne pas franchir le cap entre segmentation et discrimination, les classes de risques doivent être[14] :

— Légitime : devoir de satisfaire l'intérêt général.
— Objective et pertinente : une mise en oeuvre fiable et justifiée scientifiquement.
— Nécessaire et efficace : justification d'aucune alternative plus rentable.
— Proportionnelle : équilibre entre intérêts servis et conséquences préjudiciables.

Un équilibre doit être trouvé entre la segmentation et la mutualisation lors de la modélisation de la prime pure. Une analyse rigoureuse de l'hétérogénéité du portefeuille doit être faite en amont afin d'appliquer une segmentation et obtenir des classes homogènes. L'intérêt final est d'expliquer le tarif par segment afin de cibler les segments contenant les mauvais risques pour une gestion du risque adaptée.

1.3 Modélisation de la prime pure

Une police d'assurance est un contrat passé entre l'assuré et l'assureur pour une durée déterminée. Cette durée est appelée période de couverture et est souvent reconductible d'un an en assurance automobile.

L'assuré et l'assureur sont engagés l'un envers l'autre pendant cette période. L'assuré doit payer la prime d'assurance pour une durée établie au préalable et l'assureur s'engage à payer tout sinistre garanti dans le contrat pendant la période de couverture. À la souscription, l'assureur reçoit une prime d'un montant connu et s'engage à couvrir un risque d'un montant inconnu. Ce procédé amène l'assureur à estimer les futures pertes afin de rester rentable. L'estimation de ces pertes se fait par la création de modèles et de nouveaux produits. Le produit modélisé est la prime pure étant le montant attendu des sinistres. Mathématiquement, la prime pure correspond à l'espérance mathématique des pertes. Classiquement, deux modèles sont utilisés : le modèle individuel et le modèle collectif[7].

Le modèle individuel

Soit n le nombre d'assurés fini et déterministe

$$S^{ind} = \sum_{i=1}^{n} Y_i,$$

où S^{ind} et Y_i sont la charge sinistres globale et la charge de sinistres totale pour l'assuré i.

Le modèle individuel calcule la charge sinistre individuelle de chaque assuré. Ces charges sont sommées entre elles afin d'obtenir la charge sinistre globale du portefeuille. L'hypothèse étant que les charges individuelles de sinistre Y_i sont des variables indépendantes pouvant avoir des lois différentes. En réalité, cette hypothèse n'est pas forcément vraie.

Le modèle collectif

Soit N un nombre de sinistres aléatoire sur la période.

$$S^{col} = \begin{cases} \sum_{i=1}^{N} X_i & \text{si } N > 0 \\ 0 & \text{sinon} \end{cases}$$

X_i : montant du i^e sinistre

S^{col} : Charge sinistres globale.

Cette décomposition permet d'observer l'influence des différentes variables et la décomposition coût - fréquence. Afin d'utiliser ce modèle, deux hypothèses sont faites : les variables aléatoires N et X_i sont indépendantes, c'est-à-dire indépendance entre la fréquence et le coût des sinistres. Les variables aléatoires (X_i) sont indépendantes, stationnaires et de même loi. De ces hypothèses en découlent 2 propriétés.

Propriété n°1

Supposons que N et X_i admettent un moment d'ordre un. Alors

$$E[S] = E[N]E[X].$$

On observe dans cette formule que la prime moyenne est le produit entre le nombre de sinistre moyen et le coût moyen d'un sinistre.

Propriété n°2

Supposons que N et X_i admettent un moment d'ordre deux. Alors

$$Var[S] = E[N]Var[X] + Var[N]E^2[X].$$

La variance permet d'étudier la volatilité des sinistres afin d'ajuster les chargements de sécurité. La modélisation permet de trouver une relation entre la variable à expliquer et les variables explicatives. Dans le modèle collectif, la variable à expliquer est la variable fréquence ou coût et les variables explicatives sont les caractéristiques du conducteur.

Modélisation de la fréquence

La variable aléatoire fréquence de sinistres prend des valeurs entières positives. Il s'agit d'une variable de comptage et doit être modélisée par une loi de probabilité discrète. La loi la plus utilisée est la loi de Poisson.

Soit N une variable aléatoire suivant une loi de Poisson de paramètre λ réel positif, alors pour tout n entier naturel on a

$$P(N = n) = \frac{\lambda^n}{n!}e^{-\lambda}.$$

où $P(N = n)$ désigne la probabilité d'avoir n sinistres.

L'espérance et la variance de la variable aléatoire N s'écrivent

$$E(N) = Var(N) = \lambda.$$

La loi de Poisson se caractérise par une équidispersion demandant une homogénéité du portefeuille. Des tests sont effectués au préalable sur l'échantillon afin de savoir si on est dans un cas d'équidispersion ou de surdispersion. Dans le cas d'une hétérogénéité du portefeuille une deuxième loi peut être utilisée pour la modélisation du nombre de sinistres. Il s'agit de la loi Binomiale Négative tenant compte de la surdispersion.

Soit N une variable aléatoire suivant une loi Binomiale Négative de paramètres r entier naturel et p $\in$ [0,1], alors pour tout n entier naturel on a

$$P(N = n) = C^n_{n+r-1}p^r(1-p)^n.$$

L'espérance et la variance de la variable aléatoire N s'écrivent

$$E(N) = \frac{r(1-p)}{p},$$
$$Var(N) = \frac{r(1-p)}{p^2}.$$

Modélisation du coût des sinistres

La variable aléatoire coût des sinistres prend des valeurs réelles positives. Le choix de la loi dépend de la queue de distribution des montants de sinistres. Si la queue de distribution est à queue lègère de bons candidats peuvent être la loi Gamma, loi log-normale ou encore la loi inverse gaussienne.

Soit X une variable aléatoire suivant une loi Gamma de paramètres α et θ, pour $\alpha > 0$ et $\theta > 0$, c'est à dire que X est une variable aléatoire continue ayant comme densité de probabilité f

$$f(x) = \begin{cases} \frac{\theta^\alpha}{\Gamma(\alpha)}e^{-\theta x}x^{\alpha-1} & \text{si } x > 0, \\ 0 & \text{sinon,} \end{cases}$$

avec $\Gamma(\alpha) = \int_0^{+\infty} e^{-x} x^{\alpha-1} dx$.

L'espérance et la variance de la variable aléatoire X s'écrivent

$$E(X) = \frac{\alpha}{\theta},$$
$$Var(X) = \frac{\alpha}{\theta^2}.$$

En pratique, nous modélisons séparément les sinistres communs dits attritionnels et les sinistres graves. Les sinistres attritionnels seront généralement à queue légère et les sinistres graves à queue lourde. Les sinistres graves peuvent être modélisés par une loi de Pareto généralisée. Le modèle collectif est utilisé dans la suite du mémoire. Nous étudions par conséquent les modèles fréquences et coûts des sinistres en faisant l'hypothèse classique d'indépendance.

1.4 Des nouvelles approches de tarification

Classiquement, la modélisation de la prime pure se fait par le modèle linéaire généralisé[29]. Ces modèles de références sont souvent utilisés pour leur simplicité d'application, leur temps de calcul rapide, leur bons résultats et surtout leur interprétabilité. Le modèle linéaire généralisé [1] permet à partir des caractéristiques du conducteur et du véhicule de calculer une prime pure. De plus, l'effet des caractéristiques du conducteur et du véhicule sont directement quantifiables par des coefficients constituant une véritable force d'interprétabilité. Toujours dans un contexte concurrentiel, les assureurs aimeraient trouver de nouveaux modèles afin d'optimiser le calcul de la prime pure. Récemment, dans la littérature académique actuarielle et dans des récents mémoires d'actuariat, les méthodes de *machine learning* sont mises en avant. Pour diverses applications, les modèles de *machine learning* ont montré un gain de performances par rapport aux modèles de régression linéaires généralisés notamment dans des situations de modélisation où l'aspect non linéaire apparaît. Ces modèles sont capables d'apprendre des relations très complexes entre les données et d'en extraire un maximum d'informations possibles. L'inconvénient des méthodes de *machine learning* est la difficulté d'interprétabilité des modèles. La non-interprétabilité d'un modèle engendre une non-connaissance réelle du risque et par conséquent une difficulté supplémentaire de pilotage du risque.

Modèle linéaire généralisé versus *machine learning*

Les modèles GLM et de *machine learning* sont les principaux modèles étudiés dans la tarification automobile. D'un côté, nous avons le modèle GLM étant une référence par l'explication très simple des prédictions à l'aide de coefficients. En effet, ces coefficients permettent de quantifier un risque et de le maitriser. De l'autre, nous avons l'émergence des modèles *machine learning* pouvant être plus performants que les modèles classiques mais dont l'interprétabilité reste compliquée. Sans interprétabilité, un modèle n'a aucune valeur car ce modèle ne permet pas de cibler les mauvais risques. Réglementairement, l'ACPR [2] et le RGPD [3] restent vigilants sur l'utilisant des modèles *machine learning* appelées *blackbox*. Cette métaphore est dûe au fait que les données sont utilisées par l'algorithme pour créer le modèle sans avoir une réelle idée du processus de modélisation. Par conséquent, l'interprétabilité de ces modèles est un défi d'actualité pour les data-scientists et les actuaires. Récemment, de nombreuses techniques telles que les méthodes LIME ou SHAP[12] sont de plus en plus développées et utilisées sur les modèles de *machine learning* afin d'interpréter ces modèles. Ces méthodes permettent d'avoir l'importance

1. Appelé plus communément GLM, *Generalized Linear Model*.
2. Autorité de contrôle prudentiel et de résolution.
3. Le règlement général sur la protection des données.

des variables engendrant une connaissance du risque. Néanmoins, ces méthodes ne sont pas encore assez maitrisées par les actuaires et donc très peu utilisées. Afin d'illuster les points forts et faibles du modèle GLM et des modèles *machine learning*, nous reprenons un exemple issu d'un mémoire[27] d'actuariat comparant le modèle GLM et le modèle *gradient boosting machine* faisant parti de la famille des méthodes de *machine learning*. Cet exemple n'utilise pas les méthodes SHAP et LIME étant très récentes. Pour commencer, nous allons présenter très rapidement un exemple afin de comprendre l'importance des modèles GLM et leur force d'interprétation. Dans cette étude, le modèle de coût est étudié. Par conséquent, la variable à expliquer est le coût et les variables explicatives sont les caractéristiques du conducteur. Le modèle linéaire généralisé avec un lien log (ce lien est souvent choisi pour conserver l'aspect multiplicatif entre la fréquence et le coût) se présente de la forme suivante

$$\mathbb{E}(Y) = \exp(\sum_{i=1}^{p} \beta_i X_i + \beta_0),$$

où, Y est la variable à expliquer (le coût), X_i est la variable explicative i (caractéristique du conducteur ou véhicule), β_i est le coefficient pour la variable explicative i (effet), p est le nombre de variables explicatives et β_0 est l'intercept (effet commun à toutes les variables).

A titre d'exemple, le GLM suivant a été obtenu :

X_i	β_i
marqueAUDI	0,092
marqueBMW	0,150
marqueCITROEN	-0,168
marqueFIAT	-0,176
marqueFORD	-0,183
marqueMERCEDES	0,168

TABLE 1.1 – Coefficients du modèle GLM dans l'exemple de mémoire [27]

Nous ne donnons pas toutes les variables explicatives et coefficients du modèle GLM coût, le but de cet exemple est de comprendre la puissance d'interprétabilité de ces coefficients. En effet, nous pouvons lire directement sur ces coefficients, les effets de chaque modalités de variables. Par exemple, parmi ces différentes marques, la marque Mercedes est la plus couteuse, suivi de la marque BMW puis AUDI. Les marques les moins couteuses sont les marques FORD, FIAT et CITROEN. De plus, ces coefficients quantifient le passage d'une modalité à une autre. Cet aspect est developpé dans la partie interprétabilité du mémoire 5. En résumé, le modèle GLM est simple, efficace et quantifiable. Le modèle de *gradient boosting machine*[4] fait parti de la famille des modèles *machine learning*. Après création de ce modèle et optimisation, l'erreur MSE sur l'échantillon test est égal à 2 411 535 ce qui fait une baisse de 34 % par rapport à l'erreur du modèle GLM.

GLM	**GBM**	**Écart**
3 639 942	2 411 535	-34%

TABLE 1.2 – Comparaison des erreurs du modèle du GLM et *GBM* dans l'exemple de mémoire [27]

4. Appelé GBM.

Le gain en performance est non négligeable expliquant l'intérêt des modèles de *machine learning* par la communauté actuarielle. L'inconvénient de ces modèles est l'interprétabilité des variables. En effet, dans les modèles GLM, nous avons directement l'effet des variables par les coefficients alors que dans le modèle *GBM* nous n'avons rien qui permettent d'expliquer l'interprétabilité du modèle. La seule information que nous avons en sortie du *GBM* est l'importance des variables explicatives dans la modélisation du modèle coût.

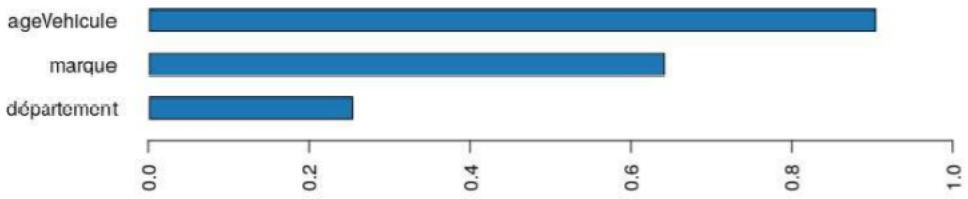

FIGURE 1.4 – Importance de certaines variables explicatives du modèle de *gradient boosting machine* dans l'exemple de mémoire [27]

Sur le graphique 1.4, nous avons l'importance de certaines variables explicatives. Plus le coefficient associé à une variable explicative est fort, plus la variable est importante dans la conception du modèle. Mais cette information ne donne aucune information sur l'effet des variables explicatives sur la variable à expliquer. Ce phénomène de *black box* est récurrent dans les modèles de *machine learning*. En résumé, ces modèles n'ont pas de sortie permettant d'étudier l'interprétabilité du modèle contrairement aux modèles GLM donnant en sortie des coefficients permettant de quantifier l'impact des variables sur la valeur prédite. Un autre inconvénient que nous pouvons mettre en avant est que certains modèles de *machine learning* ont un temps de calcul très élevé. Cet exemple permet d'appréhender la perte d'interprétabilité engendrée par certains modèles de *machine learning*. L'interprétabilité des différents modèles est étudiée dans la partie 5.

Objectif du mémoire

Nous avons vu précédemment que le modèle GLM pouvait avoir une performance plus faible que celle des modèles de *machine learning* mais que l'interprétabilité des variables explicatives étaient plus compliquées pour les modèles de *machine learning*. Ce dernier point est très important car il permet réellement de comprendre les résultats des modèles et de mieux piloter le risque. L'idéal serait de combiner la performance des méthodes de *machine learning* avec l'interprétabilité du modèle GLM. Dans ce contexte, nous introduisons les *Modele-Based Recursive Partionning*[43] [5]. Ces modèles segmentent le jeu de données en segments homogènes et créent des modèles GLM pour chaque segment. Intuitivement, ce différents modèles linéaires généralisés devraient avoir une meilleure performance que le modèle GLM global (c'est à dire un modèle GLM classique sur tout le jeu de données) tout en conservant leur propriété d'interprétabilité. L'objectif du mémoire est d'avoir une idée de la possibilité d'utilisation des modèles *MOB* dans la tarification automobile. Dans ce but, nous allons comparer les performances du modèle GLM, modèles de *machine learning* (*CART* et *XGBoost*) et des modèles *MOB* en les appliquant à la tarification automobile et de mener une réflexion sur l'interprétabilité des différents modèles.

5. Dont l'abbréviation est *MOB*.

Chapitre 2

Base de données et analyses descriptives

2.1 Base de données

Dans cette section, nous allons présenter la base de données et les différentes variables et modalités.

Présentation des données

Notre base de données est un portefeuille d'assurance automobile. Une ligne correspond à un contrat pour un assuré. La durée du contrat est quantifiable par des dates d'entrée et de fin de contrat. Un même assuré peut être présent sur plusieurs lignes mais les périodes d'exposition ne se chevauchent pas. Si un assuré est présent plusieurs fois sur plusieurs contrats différents, cet assuré aura des numéros d'avenant différents. Cette base de données contient 108 729 observations dont l'information est répartie à travers 27 variables. Les différentes variables sont présentées dans le tableau 2.1.

Variable	**Contenu**	**Nature**
nopol	numéro de police	quantitif
no_avenant	numéro d'avenant	quantitatif
auto_vehicule_marque	marque du véhicule	qualitatif
auto_vehicule_cv	nombre de chevaux fiscaux du véhicule	qualitatif
auto_groupe_gta	groupe GTA	qualitatif
auto_classe_gta	classe GTA	qualitatif
auto_code_tarif	code tarif	qualitatif
auto_dpt_garage	département garage	qualitatif
auto_zone	zone	qualitatif
auto_conducteur_datenais	date de naissance du conducteur	date
auto_conducteur_datepermis	date de permis du conducteur	date
auto_conducteur_sexe	sexe du conducteur	qualitatif
auto_conducteur_situfamil	situation familiale du conducteur	qualitatif
auto_date_premiere_entree	date entrée du premier contrat	date
auto_date_effet	date d'effet du contrat	date
auto_date_fin	date de fin du contrat	date
auto_elite	code de rabais technique	qualitatif
auto_franchise_bg	franchise brise de glace	quantitatif
auto_toit_panoramique	toit panoramique	qualitatif
auto_retro_exterieurs	possibilité de rachat de franchise	qualitatif
auto_type_assistance	type d'assistance	quantitatif
auto_fractionnement	fractionnement	qualitative
tarif_5_ans	code réduction tarifaire suivant ancienneté auto	qualitative
code_profession	code sociaux professionnel	qualitative
sra_carrosserie	type de carrosserie	qualitative
nombreSinistres	nombre de sinistres	quantitatif
coutMoyenBrut	montant de sinistres	quantitatif

TABLE 2.1 – Différentes variables de la base de données

Sur la table 2.1, nous avons un résumé des différentes variables, de leur contenu et de leur nature lors de la modélisation. Un contrat est défini par le numéro de police et d'avenant. La clé numéro de police et l'avenant sont unique. Les variables nombre de sinistres et cout moyen brut (plus précisément coût moyen brut divisé par nombre de sinistres) seront nos variables à expliquer. Les autres variables sont les variables explicatives.

Nous allons expliciter chacune des variables :

- Le numéro de police correspond à un numéro d'assuré individuel.
- Le numéro d'avenant correspond à un contrat. Par exemple, si un assuré change de contrat tout en restant dans le portefeuille, son numéro d'avenant augmentera d'un.
- Les marques de véhicule présents dans notre portefeuille sont : Audi, Autres, BMW, Citroën, Fiat, Ford, Mercedes, Nissan, Opel, Peugeot, Renault, Seat, Toyota et Volkswagen.
- Le nombre de cheveux fiscaux du véhicule est compris entre moins 4 et plus 16.
- Le groupe SRA contient des nombres compris entre 1 et 16. Plus le code est grand plus le véhicule a une puissance fiscale élevée et est dangereux.
- La classe SRA contient des lettres comprises entre A et Z avec une catégorie non notée propre à notre base de données. Plus la classe a une lettre proche de Z plus le véhicule est coûteux.
- Le code tarif est un code interne dépendant du kilométrage.
- Le code de rabais technique correspond à une notation interne à la société en fonction des caractéristiques de l'individu.
- Le département garage correspond au département du foyer de l'assuré. Les départements présents dans notre base sont 02 (Aisne), 60 (Oise), 62 (Pas-De-Calais), 80 (Somme) et AUTRES.
- La zone correspond à la region du foyer de l'assuré 02 (Haut de France) ou AUTRES.

— La date de naissance du conducteur donne des âges d'assurés compris entre 18 et 115 ans.
— La date de permis du conducteur donne des anciennetés de permis entre 0 et 115 ans.
— La date d'effet de contrat correspond à la date d'entrée du contrat. Ces dates sont comprises entre le 01/01/2006 et le 31/12/2015.
— Les dates de fin de contrat sont comprises entre le 01/01/2006 et le 31/12/2015.
— Les dates d'entrée du premier contrat correspond aux dates d'entrée de contrat avec un avenant égal à 1. Ces dates correspondent à l'entrée de l'assuré dans le portefeuille. Elles sont comprises également entre le 01/01/2006 et le 31/12/2015.
— La situation familiale du conducteur correspond au statut de l'assuré. Les différentes possibilités sont conjoint, divorcé, marié, union libre et veuve.
— La franchise brise de glace est une variable binaire, 1 signifie que l'assuré a une franchise brise de glace, 0 sinon.
— Le type d'assistance prend ses valeurs parmis 0, 1 et 2. 0 est une assistance dépannage kilométrique minimale, 2 maximale.
— La variable fractionnement correspond à la cadence de réglement de la prime pure : annuelle, semestrielle ou mensuelle.
— Le code profession regroupe des individus avec des profils métiers similaires. Les différents codes de profession sont compris entre 0 et 9.

Dans cette section, les variables que nous allons calculer et énumérer sont utilisées lors de la modelisation des modèles fréquence et coût. La variable exposition contient l'exposition au risque du contrat en année. Elle correspond à la différence entre la date de fin et la date de début pour un numéro d'assuré et un avenant donné. La variable ancienneté du permis correspond à la différence entre la date d'effet du contrat et la date d'acquisition du permis du conducteur. La variable âge du conducteur correspond à la différence entre la date d'effet du contrat et la date de naissance du conducteur. La variable ancienneté de contrat correspond à la différence entre la date d'effet du contrat et la date d'arrivée de l'assuré dans le portefeuille. Ces différentes variables sont divisées par 365,25 afin d'avoir une information annuelle. La variable coût moyen unit correspond à la division du coût moyen brut par le nombre de sinistres. Explicitement, cette variable est le coût moyen pour un sinistre d'un contrat. Cette variable sera la variable à expliquer dans notre modèle sévérité.

Présentation des variables quantitatives

Le tableau 2.2 présente les variables quantatives de notre jeu de données. La présentation consiste à donner pour chaque variable : le minimum, le 1^{er} quantile, la médiane, la moyenne, le 3^{me} quantile et le maximum. Une présentation des variables qualitatives est présente en annexe A.1.

	age_conducteur	**anc_contrat**	**anc_permis_effet**	**coutMoyenUnit**	**nbsinistres**
Minimum	17,95	0	0	0	0
1er Qu.	31,29	0	10,34	0	0
Médiane	43,69	1,06	22,58	0	0
Moyenne	45,46	2,22	24,68	339,01	0,38
3e Qu.	55,68	3,66	33,90	103,92	1,00
Maximum	114,98	12,38	114,98	1 935 313,57	12,00

TABLE 2.2 – Statistiques descriptives des variables quantitatives

L'âge du conducteur, son ancienneté de permis et l'ancienneté de son contrat ont des valeurs cohérentes et sont des variables explicatives. Le nombre de sinistres est un nombre entier compris entre 0 et 12 et le coût moyen unitaire est compris en 0 et 1 935 513 (sinistre très extrême représentant 3,4 % de la

somme des montants de sinistres). Une analyse plus détaillée des variables explicatives a été effectué dans la partie 2.2.

Retraitement des données

Nous conservons les observations dont l'âge est compris entre 18 et 80 ans car nous avons remarqué des données aberrantes après 80 ans. Nous passons de 108 729 lignes à 104 872 lignes. Nous conservons les observations dont l'exposition est non nulle (date de début de contrat identique à la date de fin de contrat), nous passons de 104 872 lignes à 102 349 lignes. Cette base sera notre base fréquence. Concernant la base sévérité, nous gardons les observations de la base fréquence dont le coût moyen unitaire est positif. Nous passons de 102 349 lignes à 27 919 lignes. Après ces quelques retraitements, nos bases fréquence et coût sont propres et ne présentent aucune valeur manquante ou manifestement aberrante.

2.2 Analyse statistiques des variables et regroupement

Une analyse de l'hétérogénéité du portefeuille par variable doit être faite en amont afin de créer des classes homogènes. L'objectif est dans un premier temps de regrouper les modalités pour chaque variable afin d'optimiser les algorithmes de prédiction que nous verrons par la suite et dans un second temps d'appréhender l'effet des différentes variables explicatives sur la fréquence ou le coût. Dans notre base fréquence, la probabilité de faire un accident pour un individu sur une année est de 0,26. Concernant la base sévérité, le coût moyen d'un sinistre est de 1 220 €.

Âge du conducteur

Nous allons étudier l'évolution de la fréquence et du coût moyen d'un sinistre en fonction de l'âge.

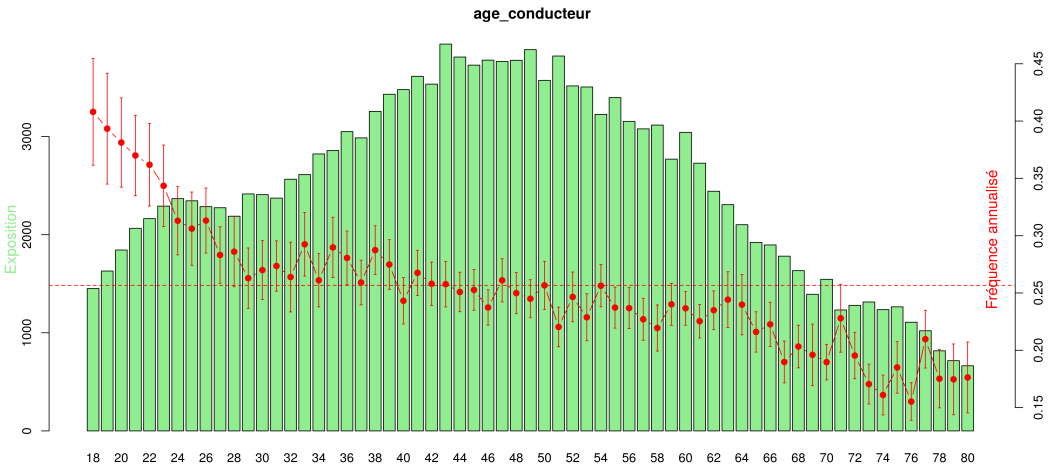

FIGURE 2.1 – Fréquence annuelle des sinistres en fonction de l'âge conducteur

Sur la figure 2.1, nous observons une baisse de la sinistralité en fonction de l'âge ce qui est cohérent. Pour la suite, un regroupement des âges n'est pas envisagé. Nous observons une pente qui décroît énormément entre 18 et 26 ans puis une décroissance moins forte à partir de 26 ans. Pour la première pente, la fréquence se situe entre 0,30 et 0,41. Pour la seconde, la fréquence se situe entre 0,15 et 0,30. Aucun âge n'est sous representé.

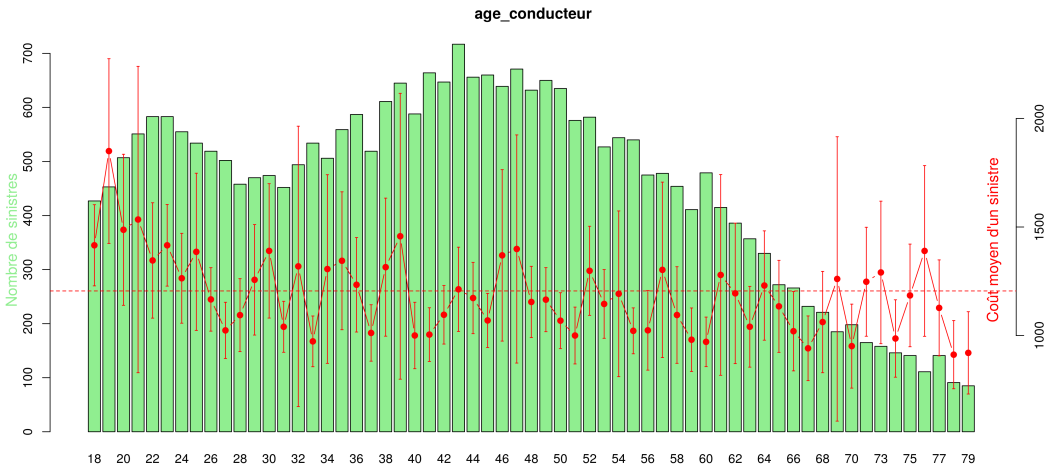

FIGURE 2.2 – Coût moyen d'un sinistre en fonction de l'âge conducteur

Sur la figure 2.2, nous avons enlevé du graphique le coût moyen des sinistres pour les 71 ans et 80 ans car ils étaient très élevés (1 993 € pour 71 ans et 3 435 € pour 80 ans). De 19 ans à 27 ans, nous observons une décroissance du coût moyen d'un sinistre. Le coût moyen d'un sinistre de cette catégorie d'âge est compris entre 1 850 € et 1 024 €. A partir de 27 ans le coût moyen d'un sinistre oscille entre 920 € et 1 457 €.

Marque du véhicule

Notre portefeuille est constitué de 14 marques. Intuitivement, nous pensons que les mêmes types de véhicule auront le même comportement. Par exemple, les citadines devraient avoir des fréquences et des coûts moyens d'un sinistre similaire.

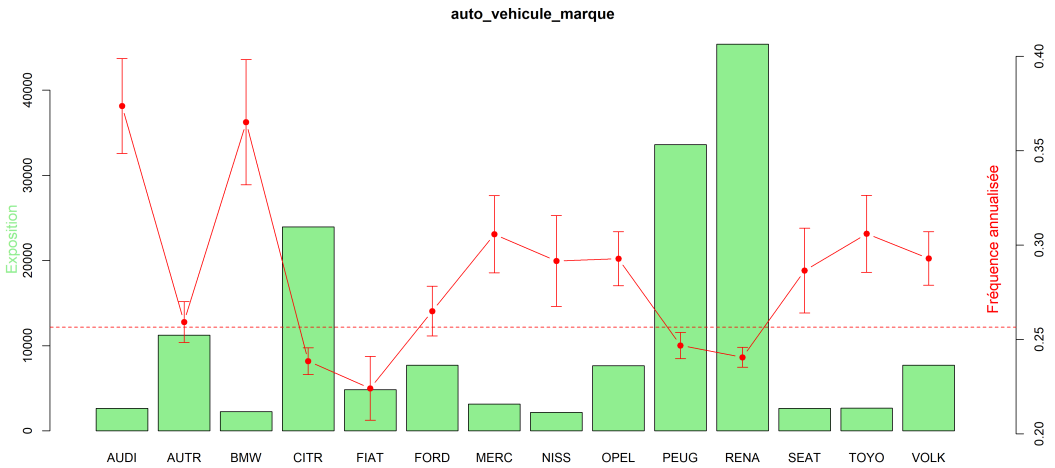

FIGURE 2.3 – Fréquence annuelle des sinistres en fonction de la marque du véhicule

Sur la figure 2.3, nous observons que nous pouvons effectuer des regroupements de marques de voiture ayant une fréquence de sinistres annuels très proche. Nous pouvons regrouper les marques :

- BMW et Audi
- Mercedes, Nissan, Opel, Seat, Toyota et Volkswagen
- Citroën, Fiat, Ford, Peugeot, Renault et Autres

La fréquence de sinistres oscille entre 0,22 et 0,38. Aucune marque n'est sous représentée.

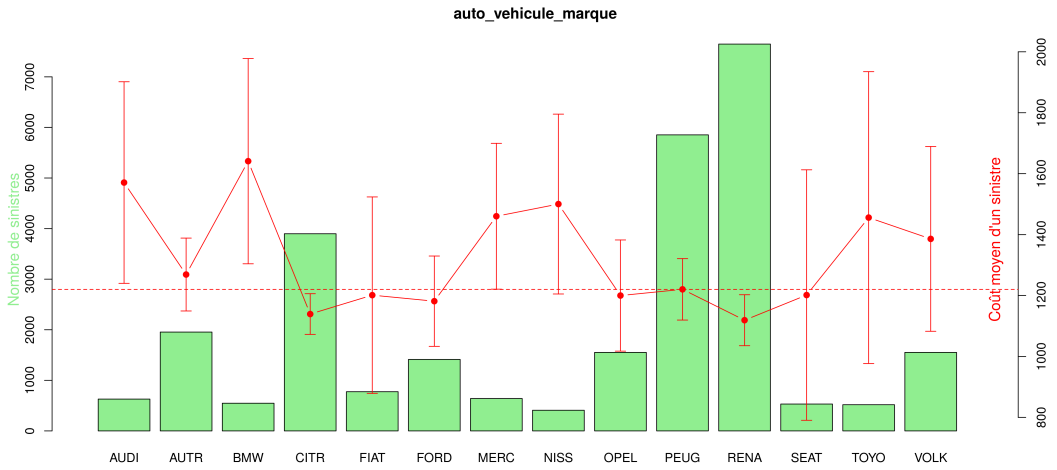

FIGURE 2.4 – Coût moyen d'un sinistre en fonction de la marque du véhicule

Sur la figure 2.4, les coûts moyens d'un sinistre donnent des regroupements similaires à ceux de la fréquence de sinistres. Les coûts moyens sont compris entre 1 641 € et 1 118 €.

Carrosserie

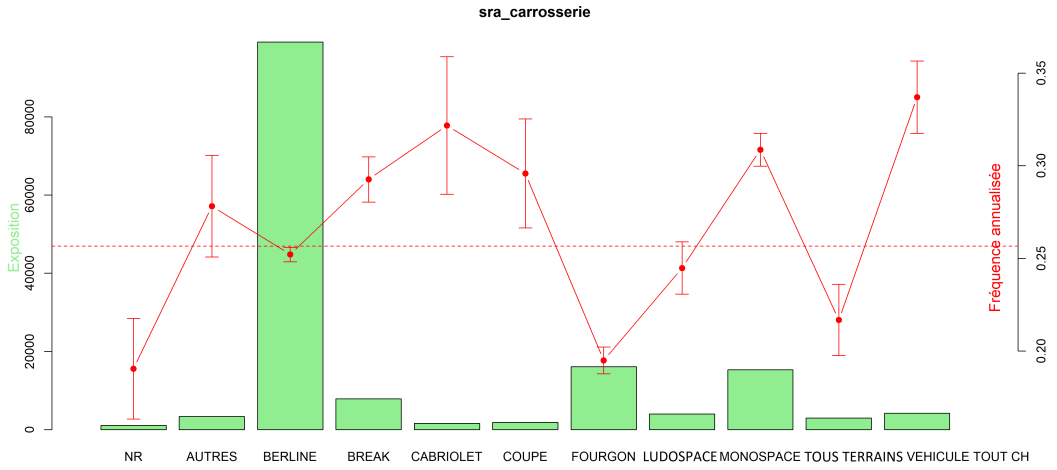

FIGURE 2.5 – Fréquence annuelle des sinistres en fonction de la carrosserie du véhicule

Sur la figure 2.5, nous observons que certaines catégories sont sous représentées telles que la catégorie NR et Cabriolet. Les regroupements de classes que nous allons effectuer vont permettre de pallier cette contrainte. Les regroupements des catégories sont :

- Fourgon, NR et Tous terrains
- Berline, Ludospace
- Break, Cabriolet, Coupe, Monospace, Vehicule tout ch et Autres.

La fréquence annuelle de sinistres oscille entre 0,19 et 0,34.

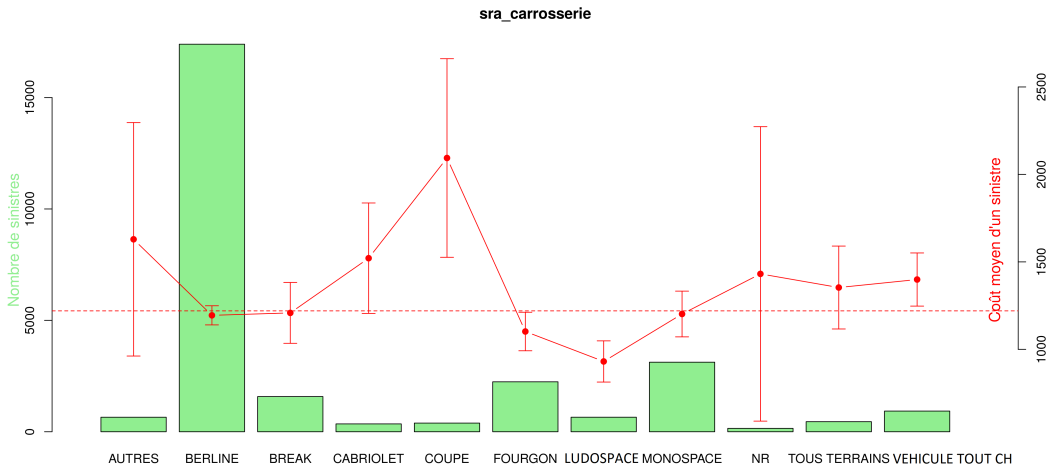

FIGURE 2.6 – Coût moyen d'un sinistre en fonction de la carrosserie du véhicule

Sur la figure 2.6, les coûts moyens de sinistres se situent entre 830 € et 1 210 €. Les regroupements faits précédemment sur la fréquence ne coïncident pas avec ceux que nous aurions pu faire avec les coûts moyens. Nous les conservons par choix arbitraire.

Classe SRA

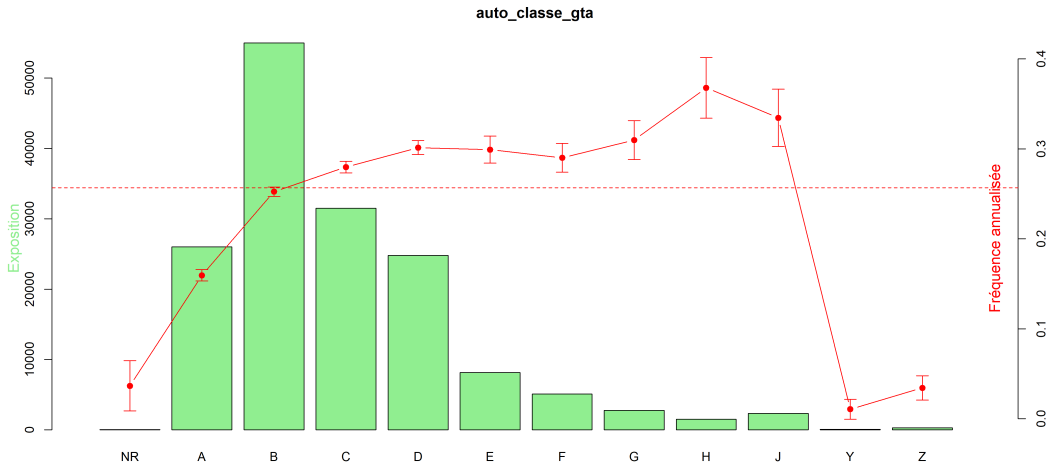

FIGURE 2.7 – Fréquence annuelle des sinistres en fonction de la classe auto

Sur la figure 2.7, la fréquence annuelle de sinistres des différentes modalités prenne des valeurs entre 0 et 0,37. L'exposition de certaines modalités est faible. Nous pouvons créer des classes ayant des fréquences de sinistres similaires :

— B, C, D, E, F et G
— H et J
— NR, Y et Z

La modalité "A" est bien représentée et n'a pas le même comportement que les autres modalités. Les modalités "NR", "Y" et "Z" sont peu représentées (respectivement,54,88 ; 94,61 ; 291,59 d'exposition annuelle) sont regroupées malgré un comportement en coûts moyens différent (voir figure 2.8).

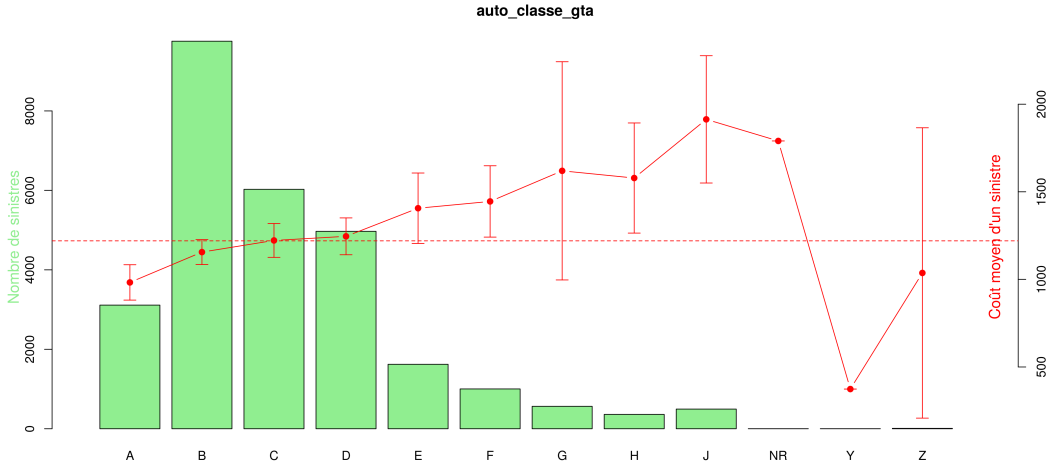

FIGURE 2.8 – Coût moyen d'un sinistre en fonction de la classe auto

Sur la figure 2.8, les coûts moyens d'un sinistre sont compris entre 983 € et 1 913 € sauf pour la modalité "Y" pour laquelle le cout moyen d'un sinistre est de 374 €. Les regroupements faits précedemment sont en adéquation avec la distribution des coûts moyens d'un sinistre pour une modalité donnée sauf pour le regroupement NR-Y-Z.

Groupe SRA

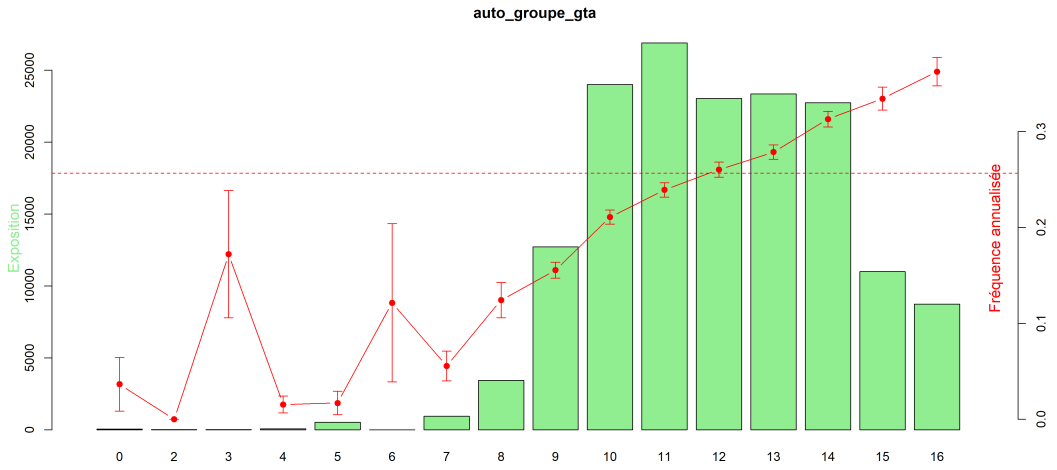

FIGURE 2.9 – Fréquence annuelle des sinistres en fonction du groupe auto

Sur la figure 2.9, la fréquence annuelle des sinistres des différentes modalités prenne des valeurs entre 0 et 0,38. L'exposition annuelle de certaines modalités est faible. Nous pouvons créer des classes ayant des fréquences de sinistres similaires :

— 0, 2, 4, 5 et 7
— 3, 6 et 8

Les modalités 0, 2, 4 et 5 sont sous représentées. Nous décidons de former une modalité pour les modalités 0, 2, 4, 5 et 7. Les modalités 3 et 6 sous representées forment une même modalité avec la modalité 8 car elles ont le même comportement.

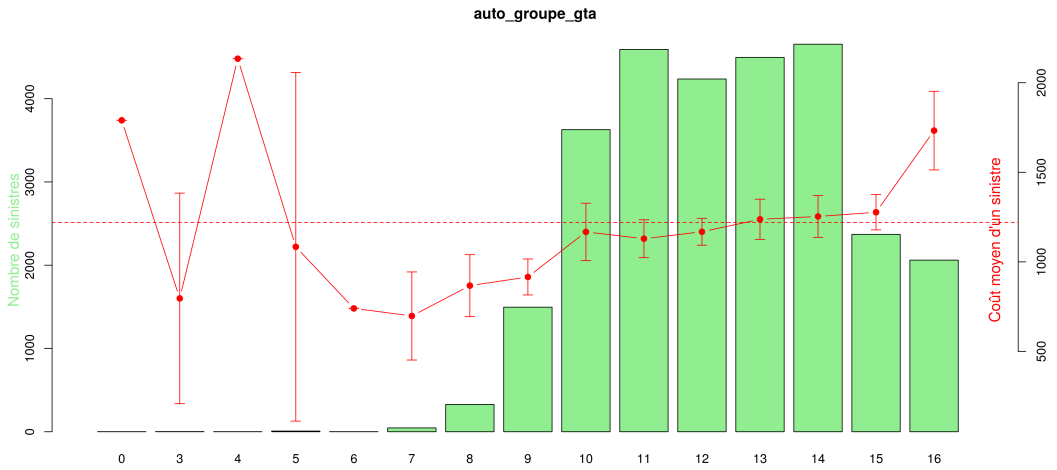

FIGURE 2.10 – Coût moyen annuel d'un sinistre en fonction du groupe auto

Sur la figure 2.10, les coûts moyens d'un sinistre sont compris entre 698 € et 2 134 €. Les regroupements faits précedemment sont en adéquation avec la distribution des coûts moyens d'un sinistre pour une modalité donnée sauf pour les modalités 0 et 4. Nous n'avons pas fait de groupe entre ces modalités car elles sont sous représentées dans la base sévérité.

Code profession

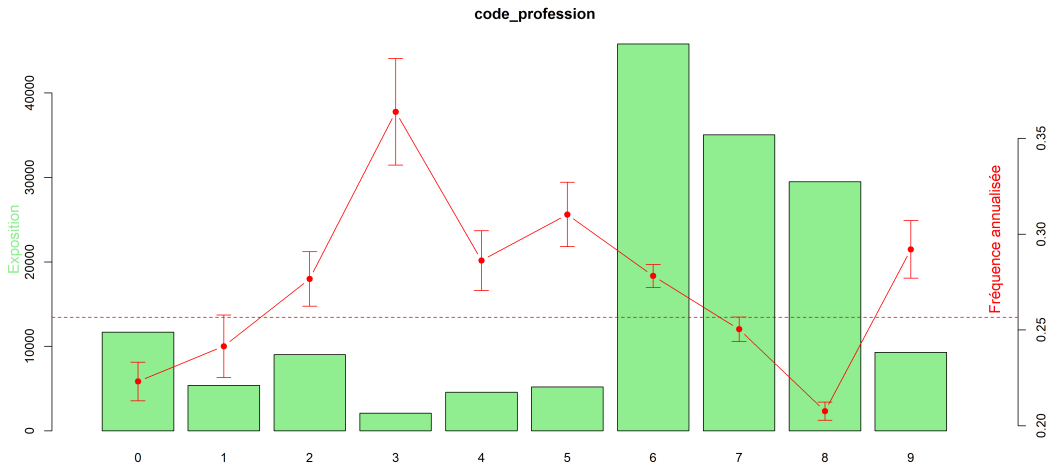

FIGURE 2.11 – Fréquence annuelle des sinistres en fonction du code de profession

Sur la figure 2.11, la fréquence annuelle des sinistres des différentes modalités prenne des valeurs entre 0,20 et 0,36. Nous créeons des classes ayant des fréquences de sinistres similaires :

- 0 et 8
- 1 et 7
- 2, 4, 5, 6 et 9

Aucune modalité n'est sous représentée. La modalité 3 est non regroupée car sa fréquence annuelle de sinistre est haute.

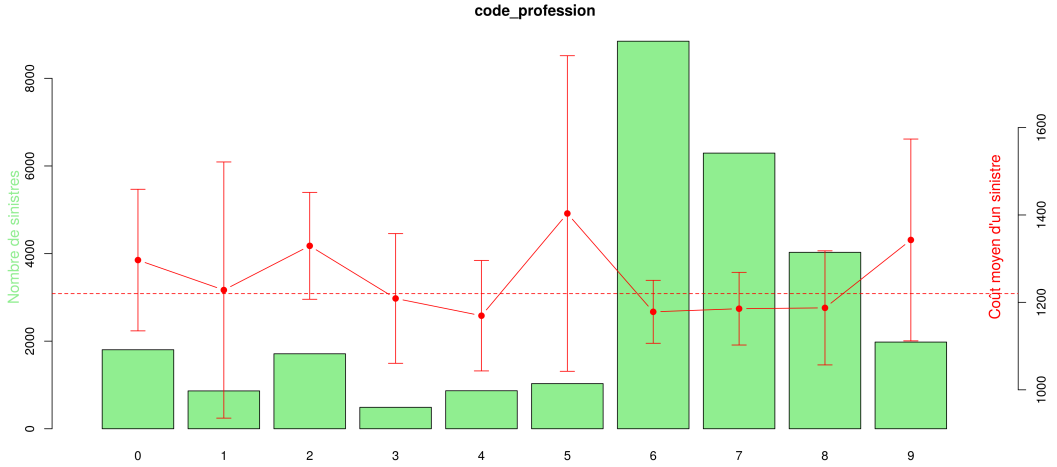

FIGURE 2.12 – Coût moyen d'un sinistre en fonction du code de profession

Sur la figure 2.12, les coûts moyens sont très homogènes par modalité. Les regroupements faits précedemment sont conservés.

Auto formule

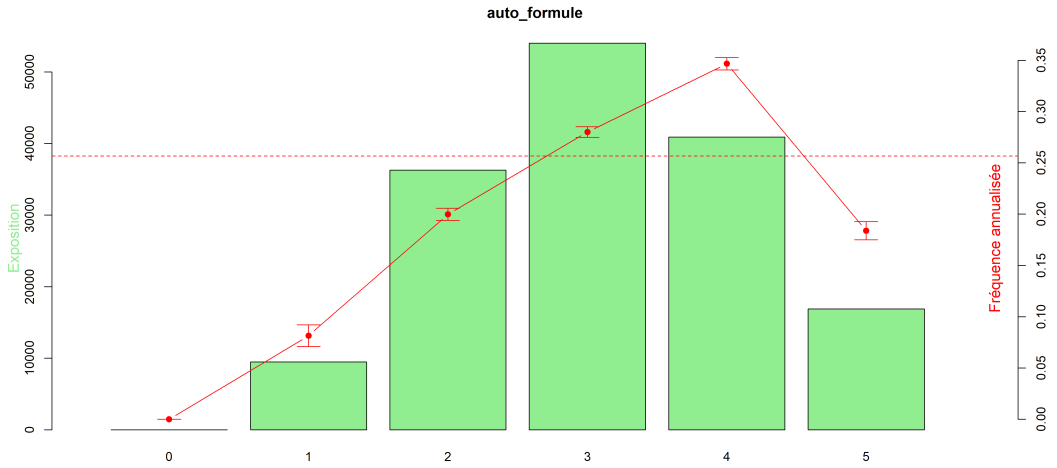

FIGURE 2.13 – Fréquence annuelle des sinistres en fonction des modalités de la formule auto

Sur la figure 2.13, la fréquence annuelle des sinistres des différentes modalités prenne des valeurs entre 0,08 et 0,35 hormis la modalité 0 ne contenant aucun sinistre. Aucun regroupement n'est effectué (nous aurions pu regrouper la modalité 0 peu représentée avec une autre modalité)

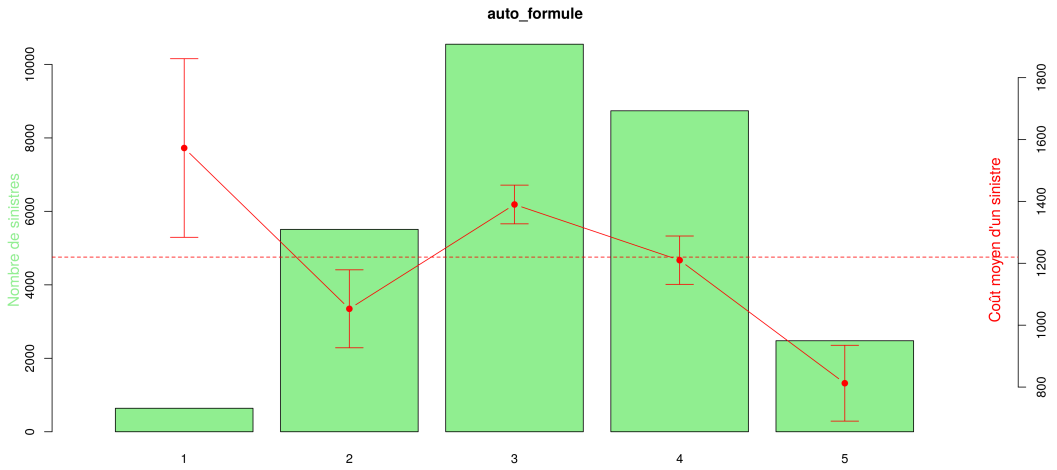

FIGURE 2.14 – Coût moyen annuel d'un sinistre attritionnel en fonction des modalités de la formule auto

Sur la figure 2.14, les coûts moyens d'un sinistre sont compris entre 812 € et 1572 €.
Les statistiques descriptives des autres modalités sont présentes en annexe A.2. Les regroupements entre modalités peuvent permettre de gagner en performance et en temps de calcul.

2.3 Analyse des corrélations entre variables

Dans cette section, l'objectif est d'étudier la dépendance des variables deux à deux. Pour analyser la corrélation entre les variables, nous allons utiliser le test d'indépendance du χ^2, l'analyse en composantes principales et l'analyse en composantes multiples.

Outils d'analyse des données

Nous allons présenter le test d'indépendance du χ^2, le calcul du V de Cramer et une explication rapide des analyses en composantes principales et multiples. Commençons par le test d'indépendance du χ^2,
Soit P une loi de probabilité d'un couple aléatoire (X, Y) à valeurs dans $\mathbb{R}^2$, on veut tester l'indépendance de X et Y en se basant sur un échantillon $(X_i, Y_i)_{1\leq i\leq n}$.
Ceci revient à tester $H_0 : P(dx, dy) = P_1(dx)P_2(dy)$ contre $H_1 : P(dx, dy) \neq P_1(dx)P_2(dy)$; où P_1 et P_2 sont respectivement les lois marginales de X et Y.
On construit alors une partition uniforme de l'ensemble des valeurs prises par X : $A_1, ..., A_r$, et de l'ensemble des valeurs prises par Y : $B_1, ..., B_s$. Dans le cas de variables aléatoires finies, r et s représentent le nombre de modalités de X et Y. Nous pouvons noter

$$N_{i,j} = \sum_{k=1}^{n} \mathbf{1}_{\{(X_k,Y_k)\in A_i\times B_j\}},$$

et

$$d_n = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{i,j} - np_{i,j})^2}{np_{i,j}}.$$

Nous montrons que

$$d_n \xrightarrow{\mathcal{L}} \mathcal{X}^2(rs - 1),$$

sous H_0, $p_{i,j} = p_{i,\bullet}p_{\bullet,j}$.
En pratique, les probabilités marginales sont inconnues et donc nous les estimons par les estimateurs empiriques

$$\widehat{p}_{i,\bullet} = \frac{1}{n}\sum_{j=1}^{s} N_{i,j}, \widehat{p}_{\bullet,j} = \frac{1}{n}\sum_{i=1}^{r} N_{i,j}.$$

Nous montrons alors que

$$\mathcal{X}^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{i,j} - n\widehat{p}_{i,\bullet}\ \widehat{p}_{\bullet,j})^2}{n\ \widehat{p}_{i,\bullet}\ \widehat{p}_{\bullet,j}},$$

suit asymptotiquement une loi de χ^2((r-1)(s-1)).

Le V de Cramer
En s'appuyant sur le test du χ^2, le V de Cramer permet de normaliser et de comparer l'intensité du lien entre deux variables. Cette mesure s'écrit

$$V = \sqrt{\frac{\chi^2}{n(\min(r-1, s-1))}}.$$

Ce coefficient est compris entre 0 et 1. Plus V est proche de 1, plus les variables sont corrélées.

Analyse en composante principale
L'analyse en composante principale dite ACP est une méthode d'analyse de données permettant d'analyser les informations quantitatives et de comprendre les différents effets d'interactions entre les données.
Théoriquement, on applique une ACP sur un échantillon structuré dans une matrice, chaque colonne représentant une variable. Avec cette matrice, nous calculons des coefficients de corrélations quantifiant les interactions entre les colonnes.
L'équivalent d'une ACP pour les variables qualitatives est l'analyse des correspondances multiples dite ACM.

Applications aux données de fréquence et coût

Notre base de données est composée de variables qualitatives et quantitatives. Un test du χ^2 est effectué pour les variables qualitatives et quantitatives. Une ACP est effectuée sur les variables quantitatives, une ACM est effectuée sur les variables qualitatives.

Test du χ^2 et calcul du V de Cramer

Nous avons calculé le V de Cramer à l'aide du logiciel `R`[32]. À présent, nous allons analyser les différentes interactions entre les variables explicatives. La représentation graphique se fait avec le package `corrplot`[39].

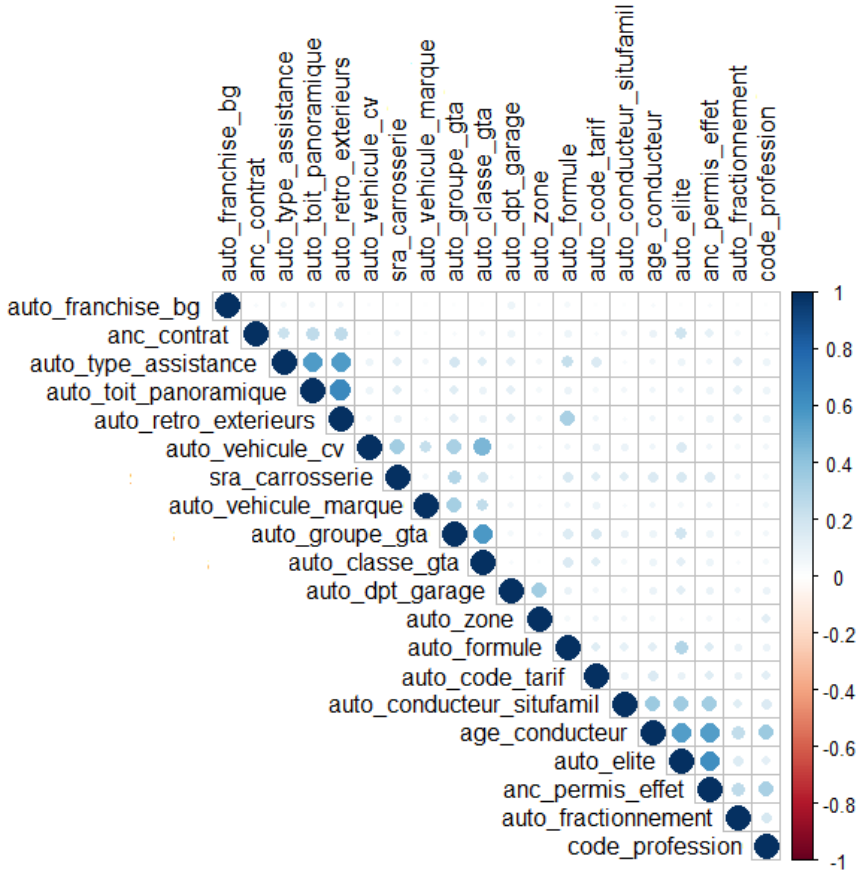

FIGURE 2.15 – Coefficient V de Cramer entre les variables explicatives

Sur la figure 2.15, nous pouvons lire les différentes corrélations entre les variables explicatives à l'aide de la couleur et de la taille du cercle. Nous pouvons observer que certaines variables sont corrélées :

- Le groupe et classe SRA du véhicule sont corrélées ce qui est cohérent car le groupe SRA donne une information sur la puissance fiscale du véhicule et le code SRA donne une information sur le prix du véhicule. Ces éléments sont liés.
- Les chevaux fiscaux du véhicule et la classe SRA sont corrélés. Cette corrélation est cohérente car le prix d'un véhicule est lié aux chevaus fiscaux du véhicule.
- L'âge du conducteur et l'ancienneté de permis du conducteur sont corrélés. Cette information naturellement cohérente.
- La région et le département du foyer sont corrélés. Nous nous attendions à une plus forte corrélation mais ceci est dû aux modalités de ces 2 variables. En effet, la plupart des observations ont une modalité égale à 2 pour la région correspondant à 4 départements (parmis les 5 départements possibles). Cette information a du mal à être capté par des outils statistiques mais la vérification des correspondances de la région avec les départements permet de conclure que cette variable est cohérente.

Ces différentes analyses permettent d'examiner la colinéarité de certaines variables et de s'assurer de la fiabilité des données. En effet, si nous avions pas eu de corrélation entre l'âge du conducteur et l'ancienneté de permis du conducteur, nous aurions détecté une erreur dans le jeu de données. Les corrélations entre les variables sont cohérentes et nous permettent de continuer notre étude.

Analyse en composantes principales

En utilisant le logiciel `R`[32], du package `FactoShiny`[37], nous avons effectué une ACP sur les variables quantitatives étant âge du conducteur, ancienneté du permis et ancienneté du contrat.

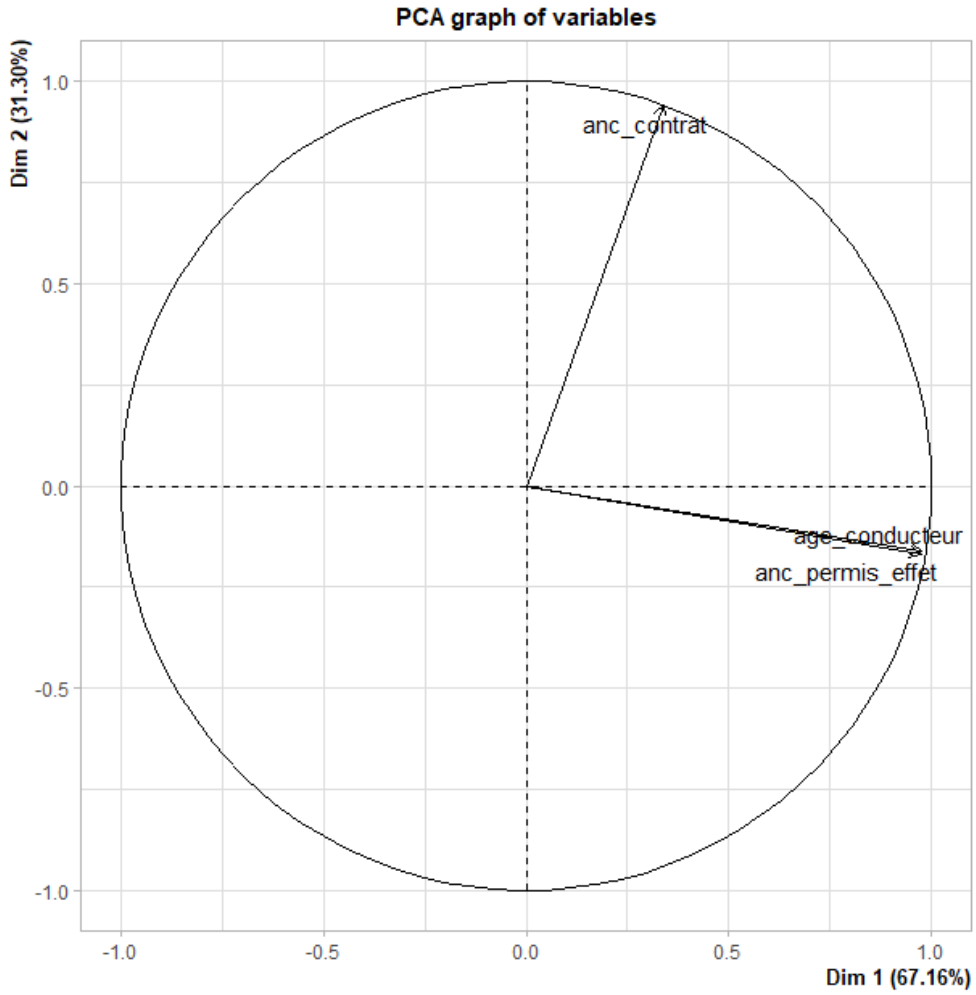

FIGURE 2.16 – ACP des variables âge du conducteur, ancienneté du permis et ancienneté du contrat.

Sur le graphique 2.17, plus les flèches vont dans la même direction plus les variables sont corrélées. On observe une extrême corrélation entre l'âge du conducteur et l'ancienneté de son permis ce qui est cohérent.

Analyse en composantes multiples

Tout comme pour l'ACP, le package utilisé est `FactoShiny`[37]. L'objectif est toujours de quantifier la corrélation entre les variables. Nous utilisons une ACM pour les variables qualitatives.

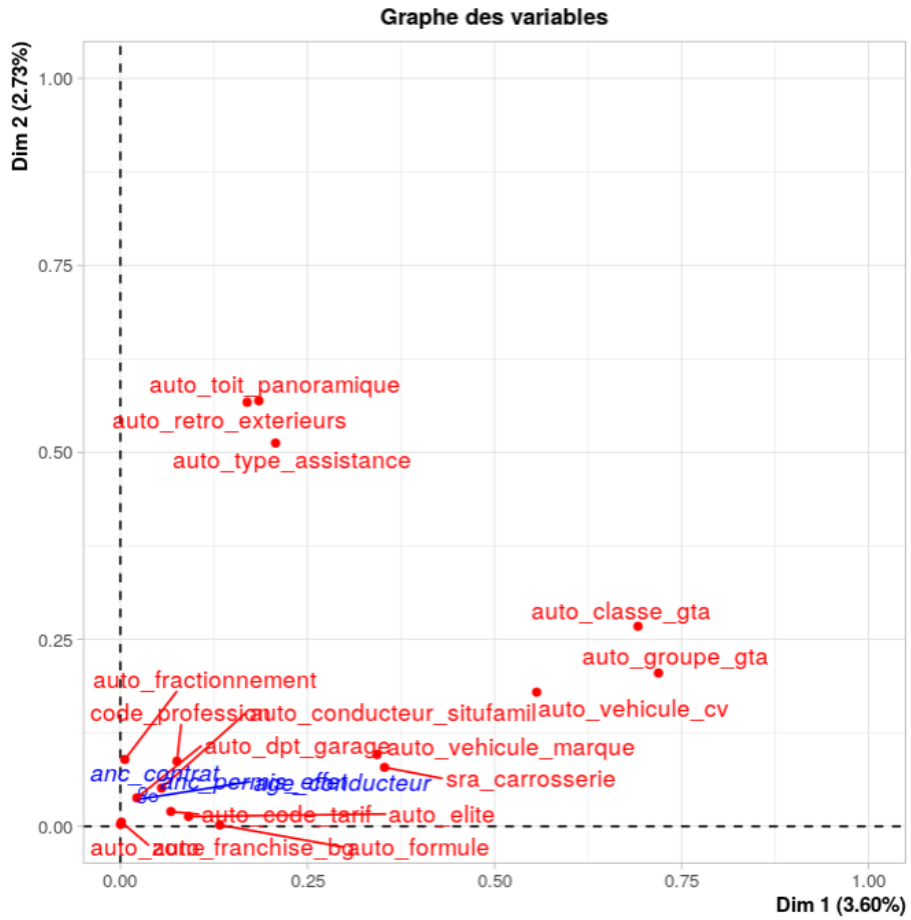

FIGURE 2.17 – ACM des variables qualitatives

Sur ce graphique 2.17, plus les variables sont proches, plus elles sont corrélées. On observe une forte corrélation entre les variables :

— Le groupe SRA, la classe SRA et les chevaux fiscaux du véhicule sont corrélés.

— La marque et la carrosserie du véhicule sont corrélées.

Intuitivement, ces différentes corrélations sont correctes. Nous ne pouvons pas interprêter la corrélation des autres variables sur ce graphique.

Ces différentes analyses ont permis de s'assurer de la fiabilité des données. En effet, les corrélations entre les variables sont cohérentes sur ces différents tests de corrélation. De plus, nous avons une idée de la colinéarité entre certaines variables.

Chapitre 3

Modèle linéaire généralisé

Dans ce chapitre, nous allons présenter le modèles linéaires généralisés. Ce modèle joue un rôle essentiel pour la détermination de la prime pure en assurance non vie.
Les modèles GLM permettent d'exprimer l'espérance d'une variable à expliquer en fonction des variables explicatives. Dans notre cas, la variable Y correspond à la fréquence ou au coût et les variables explicatives correspondent aux informations de l'assuré dont dispose l'assureur.

3.1 Présentation du modèle

Nous effectuons une présentation synthétique du modèle linéaire généralisé. Nous nous plaçons dans un contexte de régression où nous cherchons à expliquer une variable Y par p variables explicatives $X_1, ..., X_p$. Nous disposons d'un n-échantillon $(X_1, Y_1), ..., (X_n, Y_n)$ où les $x_i = (X_{i,1}, ..., X_{i,p})$ sont supposées fixés et les Y_i sont des variables aléatoires réelles indépendantes avec i nombre d'observations.

Distribution exponentielle

Les variables aléatoires $Y_1, ..., Y_n$ ont une densité de probabilité exponentielle de la forme

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right),$$

θ est appelé le paramètre naturel et ϕ le paramètre de dispersion (ou paramètre de nuisance), $b(.)$ et $c(.)$ sont des fonctions réelles qui dépendent de la loi des Y_i. On montre que l'espérance et la variance des Y_i sont données par

$$E(Y_1) = \frac{db(\theta)}{d\theta}, V(Y_1) = \phi\frac{d^2b(\theta)}{d\theta^2}.$$

Dans le cas où la variable est discrète, f est la vraisemblance c'est à dire $f(y, \theta, \phi) = P(Y = y)$.

Loi	vraisemblance	θ	ϕ
Binomiale $B(n,p)$	$C_n^y p^y \left((1-p)^{n-y}\right)$	$\log\left(\frac{p}{1-p}\right)$	1
Binomiale négative $BN(r,p)$	$C_{y+r-1}^y p^r (1-p)^y$	$\log(1-p)$	1
Poisson $P(\lambda)$	$e^{-\lambda}\frac{\lambda^y}{y!}$	$\log(\lambda)$	1
Gamma $\Gamma(\alpha,\beta)$	$\frac{1}{\Gamma(\alpha)\beta^k} y^{\alpha-1} e^{-\frac{y}{\beta}}$	$-\frac{1}{\alpha\beta}$	$\frac{1}{\alpha}$
Gaussienne $N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$	μ	σ^2

TABLE 3.1 – Vraisemblance, θ et ϕ en fonction des différentes lois

Loi	$b(\theta)$	$c(y,\phi)$
Binomiale $B(n,p)$	$n\log\left(1+e^{\theta}\right)$	$\log C_n^y$
Binomiale négative $BN(r,p)$	$-r\log\left(1-e^{\theta}\right)$	$\log C_{y+r-1}^y$
Poisson $P(\lambda)$	e^{θ}	$-\log(y!)$
Gamma $\Gamma(\alpha,\beta)$	$-\log(-\theta)$	$-\log(y)+\alpha\log(\alpha y)-\log(\Gamma(\alpha))$
Gaussienne $N(\mu,\sigma^2)$	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$

TABLE 3.2 – $b(\theta)$ et $c(y,\phi)$ en fonction des différentes lois

Les tableaux 3.1 et 3.2 précisent $\theta, \phi, b(.)$ et $c(.)$ pour quelques lois usuelles.

Modélisation par un modèle linéaire généralisé

Pour expliquer l'espérance de la variable Y_i nous devons spécifier :

- les variables explicatives $X_{i,j}, 1 \leq j \leq p$.
- la famille exponentielle de lois.
- la fonction de lien $g(.)$.

Pour tout i, la densité de Y_i est donnée par

$$f(y_i,\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi)\right),$$

$$E(Y_i) = b'(\theta_i),$$

, où b' est la dérivée de b par rapport à x. Un lien naturel, dit aussi canonique consiste à choisir g telle que :

$$g(E(Y_i)) = \theta_i.$$

Si nous souhaitons expliquer $E(Y_i)$ par les $X_{i,j}, 1 \leq j \leq p$, alors θ_i va dépendre de celles-ci via la formule

$$\theta_i = b'^{-1}\left(g^{-1}(X_i'\beta)\right),$$

car

$$E(Y_i \mid X_i) = g^{-1}(X_i'\beta) = b'(\theta_i).$$

Pour notre modèle linéaire généralisé les paramètres à estimer sont : $\beta = (\beta_0, ..., \beta_p)'$ et ϕ.

3.2 Estimateur et choix du modèle

Estimateur du modèle

Notons $\boldsymbol{\beta} = (\beta_0, ..., \beta_p, \phi)'$.
En supposant que les variables $Y_1, ..., Y_n$ sont indépendantes, la log-vraisemblance de $Y = (Y_1, ..., Y_n)'$ est donnée par

$$\mathcal{L}(Y, \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi), \tag{3.1}$$

avec $\theta_i = b'^{-1}\left(g^{-1}(X_i'\beta)\right)$.
L'estimateur du maximun de vraisemblance de $\boldsymbol{\beta}$ est solution de l'équation du premier ordre suivante :

$$\frac{\partial \mathcal{L}(Y, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0. \tag{3.2}$$

Choix des variables

L'objectif est de sélectionner le meilleur modèle c'est à dire sélectionner les variables optimisant le modèle. Pour ce faire, nous comparons les modèles entre eux en fonction de deux critères. Le critère d'information d'Akaike et le critère d'information bayésien.

Critère AIC

Le critère d'information d'Akaike appelé AIC permet de comparer les différents modèles. Ce critère s'appuie sur le maximum de vraisemblance et le nombre de paramètres du modèle. Il s'écrit :

$$AIC = 2 \times (p - \mathcal{L}(\hat{\mu} \mid Y)),$$

où p et $\mathcal{L}$ sont respectivement le nombre de variables du modèle et la log-vraisemblance du modèle.
Le meilleur modèle est celui présentant l'AIC le plus faible.

Critère BIC

Le critère d'information bayésien appelé BIC permet de comparer les différentes modèles. Il s'écrit :

$$BIC = -2 \times \mathcal{L}(\hat{\mu} \mid Y) + p \times \ln(n)$$

où

- p est le nombre de variables du modèle.
- $\mathcal{L}$ est la log-vraisemblance du modèle.
- n est le nombre d'observations.

Le meilleur modèle est celui présentant le BIC le plus faible. Ce critère pénalise de manière plus importante le nombre de variables du modèle.

Sélection des variables

Dans notre démarche, certaines variables peuvent ne pas être significatives. Une sélection individuelle des variables ne peut pas être envisagée du fait de l'interaction entre les variables. En effet, dans un modèle une variable peut ne pas être significative avec certaines variables et peut le devenir en ajoutant ou supprimant une variable du modèle. Ces contraintes techniques nous amènent à considérer trois méthodes afin de sélectionner une combinaison de variables intéressantes.

Forward Selection
Cette méthode consiste à utiliser un modèle avec une seule variable explicative. Cette variable peut être la variable la plus significative du modèle. Nous ajoutons la variable améliorant le plus le modèle. Ainsi de suite, jusqu'à ce que le modèle ne s'améliore plus. Dans notre cas, l'amélioration du modèle signifie une baisse de l'AIC ou BIC.

Backward Selection
Cette méthode consiste à utiliser un modèle avec toutes les variables explicatives. Nous enlèvons la variable diminuant le plus l'AIC ou BIC. Ainsi de suite, jusqu'à ce que l'AIC augmente lors de la supression d'une variable.

Stepwise Selection
Cette méthode consiste à utiliser un modèle avec les variables jugées les plus significatives du modèle. À chaque étape, nous enlèverons la variable la moins significative et nous rajouterons la variable la plus significative. Une variable peut être significative avec une combinaison de variables et ne plus l'être avec une autre combinaison de variables. Cette méthode est une combinaison des deux méthodes précédentes.

Qualité d'ajustement du modèle

L'objectif de cette partie est de présenter des outils de validation des différents modèles. Posons $\mu_i = E(Y_i \mid X_i)$, $\sigma_i^2 = var(Y_i \mid X_i)$ et soient $\hat{\mu}_i, \hat{\sigma}_i^2$ les estimateurs du maximum de vraisemblance de μ_i et σ_i^2. Pour la validation des modèles, nous pouvons analyser les résidus.

En général nous considèrons trois types de résidus :

— Les résidus bruts :
$$\varepsilon_i = Y_i - \hat{\mu}_i.$$

— Les résidus de déviance :
$$\varepsilon_{i,D} = signe(d_i) * \sqrt{|d_i|},$$
avec la déviance résiduelle
$$D = \sum_{i=1}^{n} d_i,$$
l'expression de d_i dépend de la loi choisie.

— Les résidus normalisés de *Pearson* :
$$\varepsilon_{i,P} = \frac{Y_i - \hat{\mu}_i}{\hat{\sigma}_i}.$$

Les tests d'adéquation du modèle sont basés sur ces deux derniers résidus.

Déviance résiduelle
Il s'agit d'évaluer la qualité du modèle en considérant toutes les variables explicatives et en se basant sur les écarts entre observations et estimations. Le modèle estimé est comparé au modèle saturé (c'est à dire au modèle possédant autant de paramètres que d'observations. Cette comparaison est basée sur la différence des log-vraisemblances, qu'on appelle déviance résiduelle et est donnée par
$$D = 2\left(\mathcal{L}(Y,Y) - \mathcal{L}(Y,\hat{\beta})\right),$$

où $\mathcal{L}(Y,Y)$ et $\mathcal{L}(Y,\hat{\beta})$ sont respectivement les log-vraisemblances du modèle saturé et estimé. Nous montrons qu'asymptotiquement D suit une loi de *Khi-deux* à $n-(p+1)$ degrés de liberté ($(p+1)$ est le nombre de paramètres inconnus du modèle). On peut alors effectuer un test de rejet ou d'acceptation du modèle selon la valeur de la déviance. Nous acceptons le modèle avec un risque α si $D \leq x_\alpha$, avec $P(\mathcal{X}^2(n-(p+1)) > x_\alpha) = \alpha$.

Test de *Pearson*
Le test de *Pearson* est un test de type *Khi-deux* dont le but est de comparer les valeurs observées Y_i aux valeurs estimées par le modèle. Posons $\mu_i = E(Y_i \mid X_i)$, $\sigma_i^2 = var(Y_i \mid X_i)$ et soient $\hat{\mu}_i, \hat{\sigma}_i^2$ les estimateurs du maximum de vraisemblance de μ_i et σ_i^2. La statistique de test est donnée par :

$$\mathcal{X}^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2},$$

On montre qu'asymptotiquement $\mathcal{X}^2$ suit une loi de *Khi-deux* à $n-(p+1)$ degrés de liberté.

En résumé, en appliquant le test de déviance et/ou le test de *Pearson*, le modèle est généralement accepté si $D \leq n-(p+1)$ et/ou $\mathcal{X}^2 \leq n-(p+1)$. Les résidus de deviance et de *Pearson* peuvent aussi être validés par des analyses graphiques.

3.3 Echantillonnage et mesure d'erreur

Echantillonnage

La validation croisée est une méthode d'estimation de la fiabilité des modèles se basant sur l'échantillonnage de la base. Pour ce faire, la base de données est découpée en 3 sous-échantillons. Ce découpage est effectué pour tester des données indépendantes du modèle afin d'observer l'efficacité réelle du modèle. La validation croisée permet d'éviter le surapprentissage (modèle ayant trop appris de ces données estimant mal de nouvelles entrées). Le découpage se fait généralement ainsi :

- L'échantillon d'apprentissage (*train*) contient 60 % de la base de données sélectionnée aléatoirement afin de construire le modèle.
- L'échantillon de validation contient 20 % de la base de données sélectionnée aléatoirement afin d'optimiser les paramètres du modèle.
- L'échantillon test contient 20 % de la base de données sélectionnée aléatoirement afin d'évaluer les prédictions.

Mesure d'erreur

L'évaluation des prédictions se fait par des mesures d'erreur que nous allons expliciter. L'erreur quadratique moyenne permet d'évaluer la précision des valeurs prédites. L'erreur quadratique moyenne est définie par

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 ,$$

où y_i est la valeur réelle de la i^{me} observation du jeu de données et $\hat{y}_i$ est la valeur prédite par le modèle de la i^{me} observation.
Il existe d'autres mesures d'erreur équivalentes au MSE telles que le RMSE et le MAE

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \qquad \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| .$$

Les mesures d'erreur MSE et RMSE sont équivalentes. Le MSE et RMSE donne un poids plus élévé aux grosses erreurs que le MAE. Aucun de ces indicateurs n'est meilleur que les autres.

3.4 Application

Notre modèle GLM est construit sur un échantillon d'apprentissage contenant 60 % des données sélectionnées aléatoirement et est testé sur 20 % des données (échantillon test). Le modèle GLM n'a pas de paramètres d'optimisation d'où l'utilisation de seulement deux échantillons. L'échantillon d'apprentissage contient 61 409 observations, l'échantillon test contient 20 470 observations. Pour la suite de nos travaux, ces échantillons seront inchangés en fonction des modèles afin de conserver une cohérence dans nos comparaisons de mesure d'erreur. Dans cette partie, la modélisation des modèles GLM se fait avec la fonction `glm` du package `MASS`[38] sur `R`[32].

Modélisation de la fréquence

La variable à expliquer Y est le nombre de sinistres et les variables explicatives X_i, sont les 20 variables caractéristiques du conducteur vues précédemment. L'exposition correspond à la durée annuelle du contrat d'un assuré. Par conséquent, ce n'est pas une variable explicative mais plutôt un poids. En effet, si nous n'utilisons pas l'exposition, une personne exposée un jour avec un sinistre et une personne exposée un an avec un sinistre auront le même comportement dans le modèle ce qui n'est pas le cas. Par conséquent, l'exposition est ajoutée en *offset*.

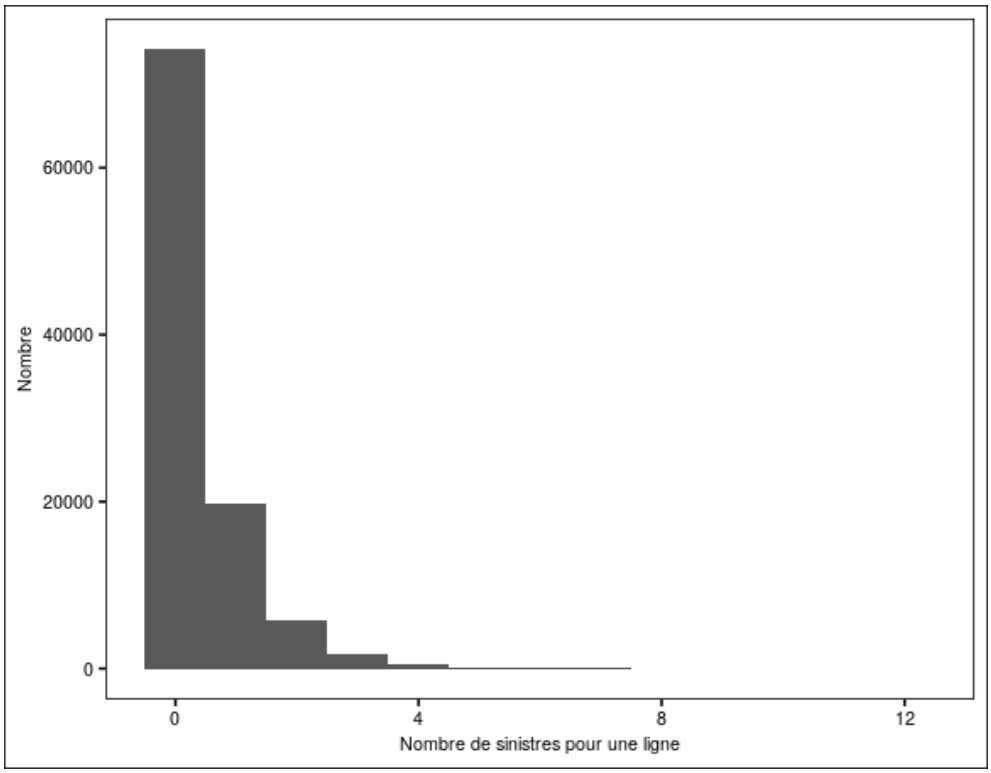

FIGURE 3.1 – Distribution du nombre de sinistres

Dans le graphique 3.1, nous avons en abscisse le nombre de sinistres et en ordonnée le nombre d'observations total associé au nombre de sinistres. Par exemple en 0, nous avons 74 180 individus avec aucun sinistre parmi nos 102 349 observations. Le nombre de sinistre maximum pour une observation est de 12. Nous observons une forte masse en 0.

Choix de la loi

Tout d'abord, nous devons choisir la loi de notre modèle. Pour la fréquence, les lois utilisées sont la loi de Poisson et la loi Binomiale Négative. Le lien choisi est le lien log pour conserver l'aspect multiplicatif.

Choix des variables

Dans cette partie, la sélection des variables est effectuée pour la loi de Poisson et la loi Binomiale-Négative sur le jeu de données sans regroupement des modalités et avec regroupement des modalités (voir la partie 2.2 pour plus de détails sur les regroupements). Le regroupement a été effectué en fonction des similitudes de fréquences annuelles de sinistres des différentes modalités pour chaque variable. Le choix de notre modèle GLM optimisé est celui dont les variables minimisent l'AIC. Afin de sélectionner les variables optimales dans la conception du modèle, nous utilisons les démarches *backward selection* et *forward selection*. Ces méthodes sont appliquées par la fonction `STEPAIC` du package `MASS`[38]. Dans un premier temps, nous avons essayé d'utiliser la méthode *backward selection* mais elle conserve énormément de variables sans pour autant que l'AIC baisse significativement. Par conséquent, nous décidons d'utiliser la méthode *forward selection* qui nous donnera un bon AIC avec peu de variables. Comme expliqué dans la partie 3.2, la méthode *forward* permet de tester différents modèles GLM en fonction de différentes variables. L'avantage est de pouvoir interpreter la fréquence en fonction de moins de variables.

Modalités non regroupées		**Modalités regroupées**	
Loi de Poisson	**Loi Binomiale-Négative**	**Loi de Poisson**	**Loi Binomiale-Négative**
auto_formule	auto_formule	auto_formule	auto_formule
auto_code_tarif	auto_anc_permis	auto_code_tarifGroupe	auto_anc_permi
auto_anc_permis	auto_code_tarif	auto_anc_permis	auto_code_tarifGroupe
auto_groupe_gta	auto_elite	auto_groupe_gtaGroupe	auto_elite
auto_elite	auto_groupe_gta	auto_elite	auto_groupe_gtaGroupe
auto_fractionnement	auto_fractionnement	auto_fractionnement	auto_fractionnement
auto_zone	auto_zone	auto_zone	auto_zone
auto_dpt_garage	auto_dpt_garage	auto_dpt_garage	auto_dpt_garage
auto_type_assistance	auto_type_assistance	auto_type_assistance	auto_type_assistance
exposition	exposition	exposition	exposition

TABLE 3.3 – Variables sélectionnées par la méthode *forward*

L'ordre des variables dans le tableau 3.3 correspond à l'ordre d'ajout des variables par la méthode *forward*. La limite de cette méthode est d'avoir tendance à ajouter des variables dont la significativité est discutable. Afin de s'assurer de n'avoir que des variables significatives dans notre modèle, nous allons étudier les p-values de nos différentes variables et modalités en sortie de R. Nous considérons qu'une variable ou modalité n'est pas significative si sa p-value est supérieure à 0,05. Par exemple, nous nous plaçons dans le cas de la loi de Poisson avec regroupement des modalités.

	Estimate	Std Error	z value	Pr(>|z|)
(Intercept)	-2,66	9,78	-0,27	0,79
auto_formule1	-6,96	48,90	-0,14	0,89
auto_formule2	0,54	9,78	0,06	0,96
auto_formule3	1,39	9,78	0,14	0,89
auto_formule4	1,67	9,78	0,17	0,86
auto_formule5	1,99	9,78	0,20	0,84
auto_code_tarifGroupe1	-0,36	0,02	-23,81	0,00
auto_code_tarifGroupe2	-0,12	0,01	-9,74	0,00
auto_code_tarifGroupe3	0,14	0,02	8,66	0,00
anc_permis_effet	-0,01	0,00	-21,57	0,00
auto_groupe_gtaGroupe1	-0,96	0,11	-8,53	0,00
auto_groupe_gtaGroupe2	0,03	0,02	1,78	0,08
auto_groupe_gtaGroupe3	0,10	0,02	5,43	0,00
auto_groupe_gtaGroupe4	0,12	0,02	6,70	0,00
auto_groupe_gtaGroupe5	0,19	0,02	10,50	0,00
auto_groupe_gtaGroupe6	0,23	0,02	12,85	0,00
auto_groupe_gtaGroupe7	0,27	0,02	12,94	0,00
auto_groupe_gtaGroupe8	0,36	0,02	16,38	0,00
auto_groupe_gtaGroupe9	-0,20	0,05	-4,31	0,00
auto_elite1	-0,15	0,01	-22,66	0,00
auto_fractionnement1	-0,06	0,01	-6,15	0,00
auto_fractionnement2	0,07	0,01	9,48	0,00
auto_zone1	-0,07	0,01	-9,04	0,00
auto_dpt_garage1	-0,04	0,02	-2,23	0,03
auto_dpt_garage2	0,13	0,02	8,14	0,00
auto_dpt_garage3	0,01	0,02	0,39	0,70
auto_dpt_garage4	-0,03	0,01	-2,76	0,01
auto_type_assistance1	-0,08	0,02	-3,62	0,00
auto_type_assistance2	0,01	0,01	1,11	0,27

TABLE 3.4 – Sortie de R du modèle GLM Poisson sans regroupement de modalité

Sur le tableau 3.4, la p-value correspond à la colonne Pr(>|z|). Nous observons que la formule auto pour le GLM Poisson sans regroupement de modalité n'est pas une variable significative. Par conséquent, nous la supprimons des variables explicatives de notre modèle. Nous effectuons ce procédé pour les 4 modèles du tableau 3.3. En résumé, l'auto formule est supprimé des variables explicatives de tous les modèles et le code tarif est supprimé pour les modèles sans regroupement de variables. Par ailleurs, les autres colonnes du tableau 3.4 en particulier la première colonne donne l'impact des modalités sur la fréquence que nous étudions par la suite sur le modèle GLM fréquence final dans la partie 3.4 de ce chapitre. De plus, d'après l'étude sur la corrélation des variables explicatives effectuée dans le paragraphe 2.3 ces variables n'ont pas de corrélation forte. Cette sélection de variables nous convient car des variables corrélées contiennent une information presque similaire. Pour la suite, nous considérons les 4 modèles GLM avec des variables explicatives significatives.

Jeu de données	Modèle GLM	AIC	BIC
Modalités non regroupées	**Loi de Poisson**	155 000	155 200
	Loi Binomiale-Négative	153 417	153 627
Modalités regroupées	**Loi de Poisson**	153 755	153 984
	Loi Binomiale-Négative	152 431	152 670

TABLE 3.5 – Comparaison de l'AIC et BIC en fonction de la loi et du jeu de données

Dans le tableau 3.5, nous observons que les modèles GLM avec regroupement des modalités a un meilleur AIC et BIC. Concernant le choix de la loi, le modèle GLM Binomiale Négative a un meilleur AIC et BIC que le modèle GLM Poisson. Mais ces différents écarts restent négligeables. Pour la suite, nous sélectionnons le jeu de données avec modalités regroupées. Au delà d'une meilleure qualité de modèle par rapport aux critères AIC et BIC, le regroupement de modalités permet d'avoir une interprétabilité plus aisée du modèle car nous avons moins de modalités et donc moins coefficients dans le modèle GLM mais aussi de raccourcir les temps de calcul des prochains modèles.

Validation du modèle

La validation d'un modèle GLM se fait par l'analyse des résidus du modèle GLM. Les résidus permettant de connaitre la qualité d'ajustement d'un modèle sont les résidus de déviance et de *Pearson* présentés dans la partie 3.2. En programmation sur `R` s'effectue avec la fonction `residuals`.
Les résidus présentés sont ceux du modèle GLM avec une loi de Poisson mais les résidus avec un modèle Binomiale Négative sont identiques. Nous commencçons avec les résidus de deviance permettant d'avoir une information sur la qualité du modèle.

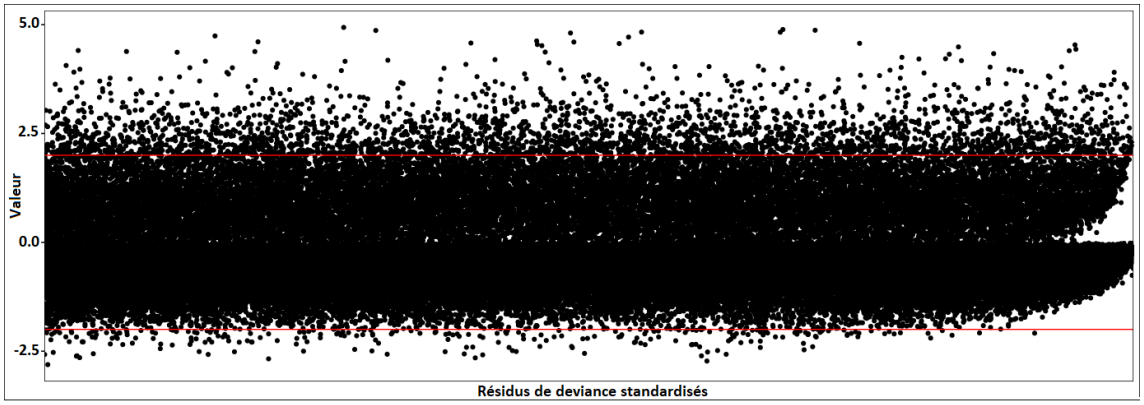

FIGURE 3.2 – Résidus de deviance standardisés

Une bonne qualité d'ajustement est traduite par l'appartenance des résidus de deviance standardisés à l'intervalle $[-2 : 2]$. Sur le graphique 3.2, nous observons que la plupart des points appartiennent à cet intervalle (environ 97 %). Les résidus de déviance valident le modèle GLM. De plus, nous avons effectué un test du $\mathcal{X}^2$ sur nos résidus de deviance et obtenu une p-value égale à 1. Le modèle est validé par le test.

Les résidus de *Pearson* vont permettre d'avoir une information complémentaire sur la qualité du modèle.

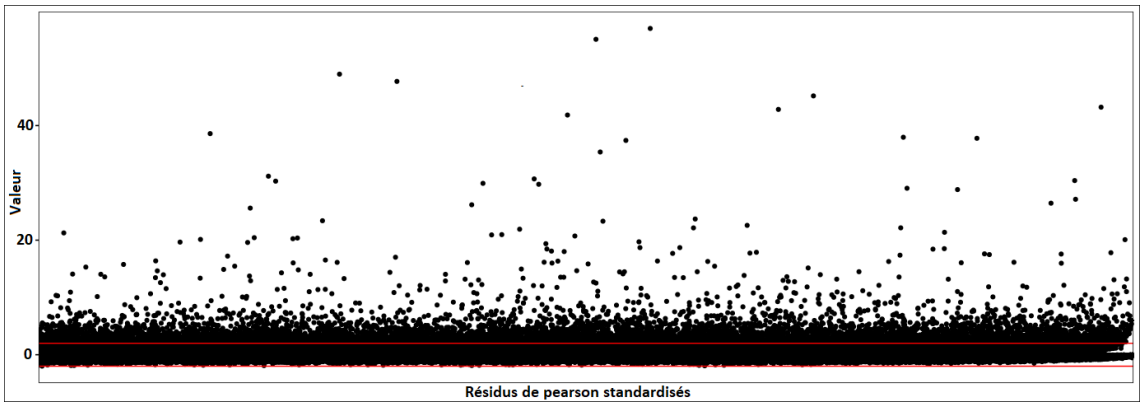

FIGURE 3.3 – Résidus de *Pearson* standardisés

Tout comme les résidus de deviance, une bonne qualité d'ajustement est synonyme d'appartenance à l'intervalle $[-2 : 2]$. Sur le graphique figure 3.3, nous observons que les résidus de *Pearson* ne sont pas dans cet intervalle. Nous pouvons souligner que nous avons 93 % des résidus de *Pearson* compris entre $[-2 : 2]$. Les résidus de *Pearson* ne valident pas graphiquement le modèle GLM fréquence mais la plupart des points appartiennent à cet intervalle. De plus, nous avons effectué un test du $\mathcal{X}^2$ sur nos résidus de *Pearson* et obtenu une p-value égale à 0. Le modèle est rejeté par le test.

Au vu des différents éléments vus précédemment, nous pouvons choisir d'utiliser une loi de Poisson ou Binomiale Négative car l'AIC, le BIC et les résidus sont très proches. Nous choissons d'utiliser une loi de Poisson car l'objet de notre mémoire porte sur les modèles *MOB*. Malheureusement, ces modèles n'ont pas d'implémentation de la loi Binomiale Négative sur le logiciel R. Afin d'avoir une cohérence de comparaisons entre les modèles, nous préférons comparer des modèles avec les mêmes lois.

Modèle final

Encore une fois, nous sélectionnons les variables optimales par méthode *forward*, puis nous supprimons les variables non significatives c'est à dire les variables dont la p-value est supérieure à 0,05.
Dans le tableau 3.6 ci-dessous, nous observons que toutes les variables sélectionnées sont significatives. Concernant la première colonne du tableau 3.6, les valeurs correspondent aux cofficients GLM donnant à l'impact de la modalité sur la fréquence prédite. Plus la valeur est élevée, plus la fréquence prédite l'est. Par exemple, si nous nous focalisons sur la variable groupe SRA, nous observons que le groupe SRA 16 est la modalité augmentant le plus la fréquence prédite. Ce qui est normal car plus la valeur du groupe SRA est elevée plus le véhicule est considéré comme dangereux. Ces coefficients donnent aussi l'impact sur la fréquence engendrée par un changement de modalité. Une explication plus approfondie de ces différents coefficients est étudiée dans la partie 5 portant sur l'interprétabilité des différents modèles. Afin de comparer notre modèle GLM fréquence avec les autres modèles, nous calculons le MSE, RMSE et MAE sur l'échantillon test.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3,46318	0,16135	-21,46373	0,00000
auto_code_tarifGroupe-20-REF	0,35223	0,01776	19,83508	0,00000
auto_code_tarifGroupe-30 et -40	0,64255	0,02516	25,53909	0,00000
auto_code_tarifGroupe40	0,84882	0,05780	14,68459	0,00000
anc_permis_effet	-0,00527	0,00061	-8,63283	0,00000
auto_groupe_gtaGroupe10	1,17097	0,15591	7,51065	0,00000
auto_groupe_gtaGroupe11	1,29044	0,15575	8,28547	0,00000
auto_groupe_gtaGroupe12	1,33890	0,15593	8,58669	0,00000
auto_groupe_gtaGroupe13	1,38865	0,15590	8,90753	0,00000
auto_groupe_gtaGroupe14	1,49437	0,15591	9,58493	0,00000
auto_groupe_gtaGroupe15	1,55647	0,15665	9,93608	0,00000
auto_groupe_gtaGroupe16	1,62397	0,15697	10,34544	0,00000
auto_groupe_gtaGroupe3-6-8	0,78104	0,16743	4,66485	0,00000
auto_groupe_gtaGroupe9	0,99153	0,15730	6,30329	0,00000
auto_eliteNR	0,17021	0,01692	10,06226	0,00000
auto_fractionnementM	0,12611	0,01760	7,16629	0,00000
auto_fractionnementS	0,02652	0,02292	1,15715	0,24721
auto_zoneAUTRES	0,14166	0,02107	6,72173	0,00000
auto_dpt_garage60	0,23000	0,03740	6,14904	0,00000
auto_dpt_garage62	0,12800	0,04439	2,88325	0,00394
auto_dpt_garage80	0,15685	0,03058	5,12915	0,00000
auto_dpt_garageAUTRES	0,02966	0,03687	0,80452	0,42110
auto_type_assistance1	0,30311	0,03922	7,72898	0,00000
auto_type_assistance2	0,39807	0,03898	10,21283	0,00000
code_professionGroupe1-7	-0,10274	0,02106	-4,87853	0,00000
code_professionGroupe2-4-5-6-9	-0,00232	0,01888	-0,12284	0,90223
code_professionGroupe3	0,05724	0,05025	1,13900	0,25470

Table 3.6 – Sortie R du modèle GLM fréquence final

MSE	**RMSE**	**MAE**
0,489	0,700	0,453

Table 3.7 – MSE, RMSE et MAE du modèle GLM fréquence final sur l'échantillon test

Le tableau 3.7 présente les différentes mesures qui seront comparées avec les mesures des autres modèles que nous introduirons dans ce chapitre partie 4.4.

Modélisation de la sévérité

Dans cette partie, la variable à expliquer Y est le coût moyen d'un sinistre et les variables explicatives X_i, sont les 20 variables caractéristiques du conducteur et du véhicule (nous avons enlevé l'exposition car l'exposition du contrat n'est pas à prendre en compte) .

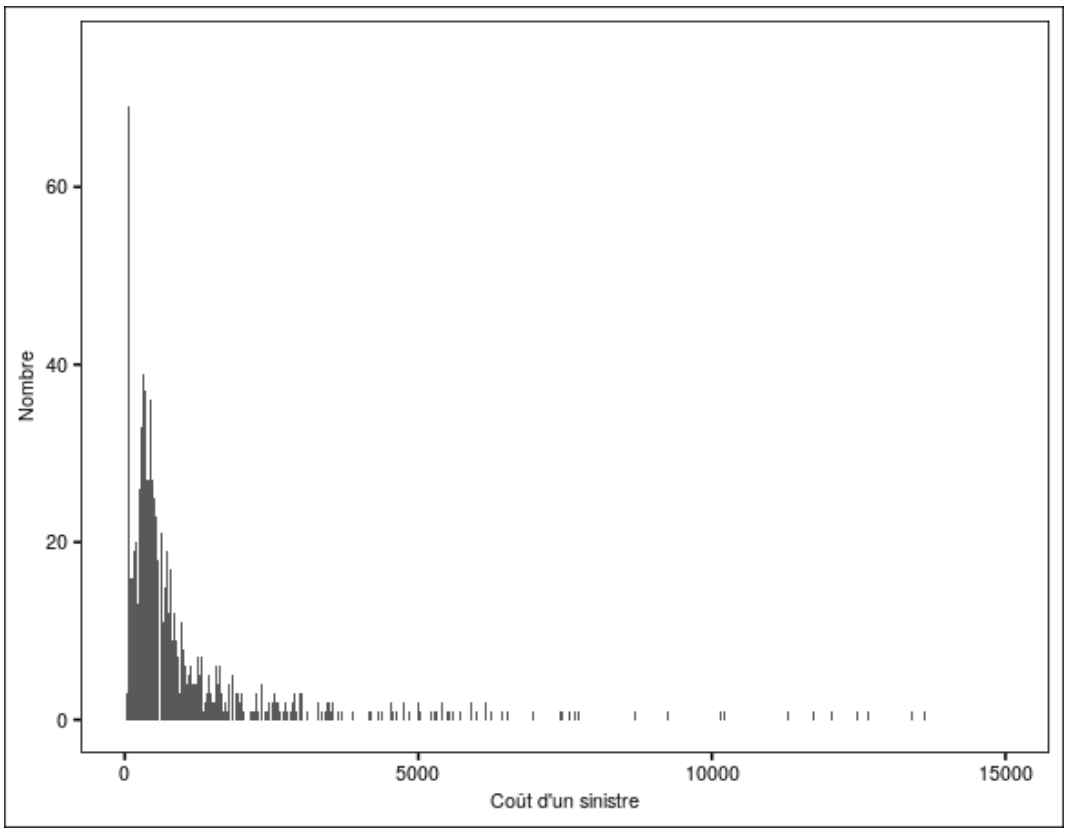

FIGURE 3.4 – Distribution du coût moyen

Dans le graphique 3.4, nous avons en abscisse le coût moyen d'un sinistre et ordonnée le nombre d'observations total associé à ce coût moyen. N'oublions pas que certains sinistres extrêmes sont présents dans la base et non représentés dans le graphique 3.4 comme par exemple le sinistre coûtant 1 935 314. Ces sinistres risquent de poser problème dans la modélisation du modèle coût.

Détermination du seuil

Une analyse sur le modèle avec sinistres extrêmes a été effectuée et n'a pas été concluante car les sinistres extrêmes biaisent le modèle. Par conséquent, nous décidons de supprimer les sinistres extrêmes de notre base. La détermination d'un seuil de sinistres extrêmes peut se faire par diverses méthodes (mean excess plot, théorie des valeurs extrêmes ou encore avis d'expert en prenant par exemple le quantile à 99,5 %). Dans notre étude, le choix du seuil dépendra de la qualité d'ajustement du modèle mesuré par les résidus de deviance du modèle GLM avec une loi Gamma. Nous allons partir d'un seuil de valeur extrême égal à 19 500 € (choix arbitraire) et à chaque itération baisser le seuil de 500 € puis faire un test du $\mathcal{X}^2$ pour calculer la p-value. Nous rappelons que le test du $\mathcal{X}^2$ est validé lorsque la p-value est supérieur à 5 %.

Sup du montant	p-value du $\mathcal{X}^2$	Quantile sur la base totale
19 500	0,156429	99,71
19 000	0,196982	99,70
18 500	0,243504	99,70
18 000	0,377578	99,67
17 500	0,466603	99,66
17 000	0,573581	99,65
16 500	0,726754	99,62
16 000	0,855255	99,60
15 500	0,962331	99,56
15 000	0,988571	99,53
14 500	0,995265	99,51
14 000	0,998423	99,49
13 500	0,999955	99,44
13 000	0,999996	99,41
12 500	**1,000000**	**99,36**

TABLE 3.8 – Choix du seuil de sinistres extrêmes

Sur le tableau 3.8 nous avons fourni une borne supérieure à nos montants de sinistres et sélectionnés les données dont les montants sont inférieurs à cette borne. Nous avons crée un modèle GLM avec l'échantillon total et effectué un test du $\mathcal{X}^2$. La p-value correspond à la valeur de ce test. Le quantile sur la base total correspond au pourcentage de données présent dans notre nouvelle base de données par rapport à la base de données initiale. Nous avons décidé de choisir un seuil égal à 12 500 car la p-value est maximale et 99,30 % des données ont un coût inférieur à ce seuil (seulement 180 sinistres ne font pas parti de notre base attritionnelle). Nous précisons que peu importe le seuil de sinistres extrêmes choisi, les modèles les plus performants restent identiques. Notre base sinistres attritionnels a 27 739 observations, 16 643 sont dédiées à l'échantillon d'apprentissage et 5 548 à l'échantillon test.

Choix de la loi

Nous choissons d'utiliser le jeu de données avec regroupement de modalités pour les mêmes raisons que la section précédente 3.4. Cette fois, la variable à expliquer est le cout d'un sinistre et les variables explicatives sont les caractéristiques du conducteur et du véhicule. Les lois sélectionnées pour le modèle GLM coût sont la loi Gamma et la loi inverse gaussienne.

Choix des variables

Comme fait précédemment dans la partie 3.1. La méthode *backward* n'est toujours pas efficace. Par conséquent, nous décidons d'utiliser la méthode *forward* qui nous donnera un bon AIC avec moins de variables. Après la sélection des variables par méthode *forward* nous supprimons les variables dont la p-value est supérieure à 0,05. Nous obtenons pour les modèles GLM coût Gamma et inverse gaussienne

Modèle GLM	AIC	BIC
Loi Gamma	438 763	526 218
Loi inverse gaussienne	477 727	526 679

TABLE 3.9 – AIC et BIC des modèles GLM coût en fonction de la loi

Le tableau 3.9 nous montre que sur notre modèle GLM coût, la loi Gamma est plus adaptée que la loi inverse gaussienne car l'AIC et le BIC de la loi Gamma sont inférieurs à ceux de la loi inverse gaussienne. Par conséquent, nous choisissons d'utiliser pour la suite de nos travaux la loi Gamma.

Validation du modèle

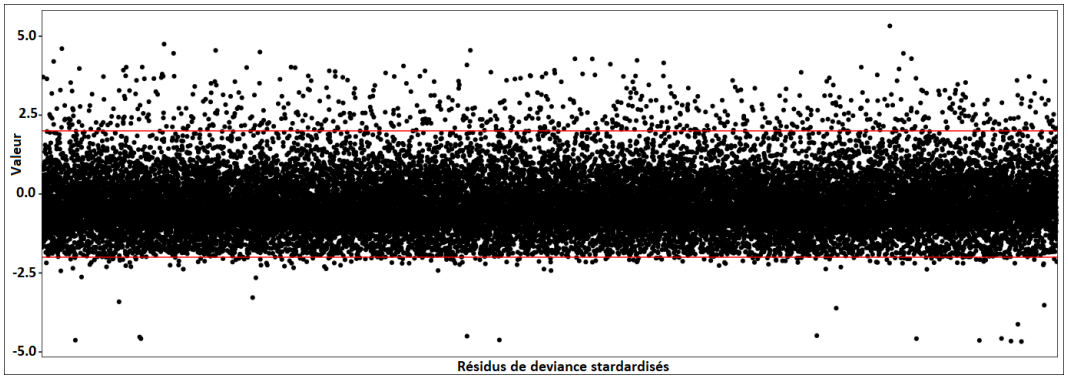

FIGURE 3.5 – Résidus de deviance standardisés

Sur le graphique 3.5, nous observons que la plupart des résidus de deviance appartiennent à l'intervalle $[-2, 2]$ (environ 98,55 %). Les résidus de déviance valident le modèle lGLM coût.
De plus, nous avons effectué un test du $\mathcal{X}^2$ sur nos résidus deviance et nous obtenons une p-value égale à 1 ce qui valide notre modèle.

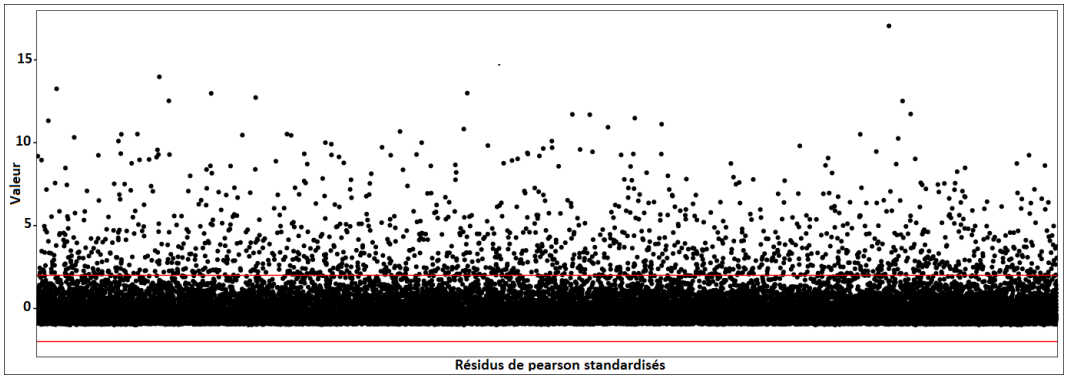

FIGURE 3.6 – Résidus de *Pearson* standardisés

Sur le graphique 3.6, nous observons que les résidus de *Pearson* sont globalement dans l'intervalle $[-2 : 2]$ (environ 95,47 %). Les résidus de *Pearson* valident graphiquement le modèle GLM.
Nous avons effectué un test du $\mathcal{X}^2$ sur nos résidus *Pearson* et nous obtenons une p-value égale à 0 ce qui rejete notre modèle.

Modèle Final

Nous sélectionnons les variables optimales par méthode *forward* et parmi les variables sélectionnées, nous supprimons les variables non significatives.
Dans le tableau 3.10 ci-dessous, toutes les variables sélectionnées sont significatives. Par la suite, nous calculons le MSE, RMSE et MAE sur l'échantillon test de notre modèle GLM coût afin de le comparer avec les autres modèles. Encore une fois, si nous nous focalisons sur le groupe SRA, la modalité augmentant le plus le coût prédit est le groupe 16. Ce qui est cohérent car plus le groupe SRA est élevé, plus le véhicule est couteux. Une explication plus approfondie des différents coefficients est effectuée dans la partie 5

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.17614	0.24516	29.27106	0.00000
auto_eliteNR	0.21557	0.02421	8.90516	0.00000
auto_formule2	-0.50402	0.06796	-7.41647	0.00000
auto_formule3	-0.11925	0.06708	-1.77770	0.07547
auto_formule4	-0.20818	0.06893	-3.02029	0.00253
auto_formule5	-0.63788	0.07172	-8.89405	0.00000
auto_groupe_gtaGroupe10	-0.15601	0.22535	-0.69231	0.48875
auto_groupe_gtaGroupe11	-0.05723	0.22523	-0.25411	0.79942
auto_groupe_gtaGroupe12	0.01636	0.22605	0.07238	0.94230
auto_groupe_gtaGroupe13	0.06790	0.22634	0.30000	0.76418
auto_groupe_gtaGroupe14	0.06841	0.22715	0.30119	0.76328
auto_groupe_gtaGroupe15	0.06413	0.23048	0.27825	0.78083
auto_groupe_gtaGroupe16	0.20466	0.23325	0.87742	0.38027
auto_groupe_gtaGroupe3-6-8	-0.20791	0.24036	-0.86499	0.38706
auto_groupe_gtaGroupe9	-0.18625	0.22743	-0.81891	0.41285
auto_dpt_garage60	0.01169	0.05655	0.20667	0.83627
auto_dpt_garage62	-0.01855	0.06759	-0.27438	0.78379
auto_dpt_garage80	-0.07678	0.04638	-1.65541	0.09786
auto_dpt_garageAUTRES	0.06499	0.05491	1.18364	0.23657
auto_toit_panoramiqueN	0.11300	0.02103	5.37243	0.00000
auto_toit_panoramiqueO	-0.02651	0.07080	-0.37441	0.70810
auto_vehiculecvGroupe12-15 et plus	-0.06740	0.09028	-0.74660	0.45531
auto_vehiculecvGroupe13-14-8-9	-0.09514	0.06049	-1.57293	0.11576
auto_vehiculecvGroupe4 et moins	0.02262	0.07657	0.29539	0.76770
auto_vehiculecvGroupe5	-0.03776	0.07066	-0.53429	0.59315
auto_vehiculecvGroupe6	-0.13666	0.06667	-2.04969	0.04041
auto_vehiculecvGroupe7	-0.12542	0.06395	-1.96113	0.04988
auto_zoneAUTRES	0.09218	0.03169	2.90880	0.00363
sra_carrosserieGroupeBREAK	0.07580	0.02557	2.96400	0.00304
sra_carrosserieGroupeFOURGON-NR	0.01276	0.03901	0.32700	0.74367
anc_contrat	-0.00867	0.00394	-2.20327	0.02759

TABLE 3.10 – Sortie R du modèle GLM coût optimal

MSE	**RMSE**	**MAE**
1 760 525	1 327	779

TABLE 3.11 – MSE, RMSE et MAE du modèle GLM coût optimal

Le tableau 3.11 présente les différentes mesures qui seront comparées avec les métriques des autres modèles que nous introduirons dans ce chapitre partie 4.27.

Chapitre 4

Modélisation de la sinistralité à l'aide d'arbres de décision

4.1 Algorithme *CART*

Dans ce chapitre, nous allons nous intéresser aux modèles *CART*, *XGBoost* et *MOB* utilisant des arbres de régression ou de classification.

4.1.1 Présentation du modèle

L'algorithme *CART*[5] fait parti des algorithmes se basant sur les arbres de décision.
Les arbres de régression ou classification sont des outils d'exploration des données et d'aide à la décision qui permettent d'expliquer et de prédire une variable quantitative ou qualitative à partir de plusieurs variables explicatives. Cet algorithme permet d'obtenir des classes d'individus construites à partir des variables explicatives de manière à ce que les individus d'une même classe soient le plus homogènes possibles du point de vue de la variable d'intérêt. Les principaux avantages de cette méthode sont sa simplicité, sa performance et la représentation du modèle en forme d'arbres permettant de comprendre facilement la structure du modèle. La structure de l'arbre commence par un regroupement de l'ensemble des observations à la racine de l'arbre appelée *root*. Chaque division sépare chaque nœud en deux nœuds fils plus homogènes que le nœud père au sens d'un critère à préciser et dépendant du type de la variable Y (qualitative ou quantitative). Chaque nœud terminal de l'arbre est appelé feuille et se voit attribuer une valeur pour la méthode de régression et une classe pour celle de classification. L'avantage du modèle *CART* par rapport au modèle GLM est son approche non paramètrique n'imposant aucune loi de probabilité pour la variable à expliquer Y. Néanmoins, l'algorithme *CART* présente certaines faiblesses. En effet, il est très facile de tomber dans un sur-ajustement, les modèles sont particulièrement instables et très sensibles aux fluctuations de l'échantillon ayant servi à construire l'arbre. La sensibilité des variables explicatives est moins interprétable que pour les modèles GLM.

Construction d'un arbre binaire

Soient $X_1, ..., X_p$ p variables explicatives et la Y variable à expliquer. Pour commencer la construction d'un arbre, nous partons de la racine contenant l'ensemble de données. Par la suite, la construction de cet arbre binaire consiste à déterminer une séquence de nœuds. Un nœud est défini par le choix d'une variable explicative et d'une division qui induit une partition en deux classes. Autrement dit, à chaque nœud, correspond un sous-ensemble des données auquel est appliquée une dichotomie.
Une division est définie par une valeur seuil de la variable explicative sélectionnée si celle-ci est quantitative ou par un partage en deux groupes de modalités si cette variable est qualitative.

La procédure est ensuite itérée sur chacun des deux nœuds obtenus.

L'algorithme considéré nécessite trois éléments importants :
- Définir un critère qui permette de sélectionner la meilleure division parmi toutes celles admissibles pour les différentes variables.
- Définir une règle permettant de dire qu'un nœud est terminal. Si c'est le cas, le nœud devient donc une feuille.
- Définir une règle pour affecter une valeur ou une classe de la variable d'intérêt à chaque feuille obtenue.

Critère de division

Une division est dite admissible si aucun des deux nœuds fils qui en découlent n'est vide. Le critère de division repose sur la définition d'une fonction d'hétérogénéité. L'objectif étant de partager les individus en deux groupes les plus homogènes possible au sens de la variable à expliquer. L'hétérogénéité d'un nœud se mesure alors par une fonction positive qui doit être :
- Nulle si et seulement si le nœud est homogène : cela revient à dire que tous les individus appartiennent à la même modalité de Y ou prennent la même valeur de Y.
- Maximale lorsque les valeurs de Y sont très dispersées.

La division du nœud t crée deux fils, gauche et droit que nous nommerons tg et td. Parmi toutes les divisions admissibles du nœud t, l'algorithme retient celle qui rend la somme des hétérogénéités des nœuds fils minimale. Autrement dit si on appelle R_{tg} l'hétérogénéité du fils gauche et R_{td} l'hétérogénéité du fils droit, la somme $R_{tg} + R_{td}$ doit être minimale (puisque ces hétérogénéités doivent être le plus faible possible).
Cela revient aussi à résoudre à chaque étape (ou nœud) t de construction de l'arbre

$$\max_{\text{division de } X_j, 1 \leq j \leq p} \left(R_t - (R_{tg} + R_{td})\right).$$

Règle d'arrêt et affectation

L'arbre s'arrête à un nœud donné, qui devient donc feuille, lorsqu'il est homogène ou lorsqu'il n'existe plus de répartition admissible ou encore si le nombre d'observation qu'il contient est inférieur à une valeur seuil. Dans le cas d'un arbre de régression (Y est quantitative), on affecte à chaque feuille une valeur qui est généralement la moyenne des observations associées à cette feuille. Dans celui d'un arbre de classification (Y est qualitative), on affecte à chaque feuille une classe de Y en considérant la règle suivante : la feuille prend la classe de celle qui est la mieux représentée dans le nœud.

Critère d'homogénéité

Nous devons distinguer les deux cas, si Y est quantitative (régression) ou qualitative (classification). Dans le cas de la régression, l'hétérogénéité du nœud t est définie par la variance que nous noterons

$$R_t = \frac{1}{|t|} \sum_{i=1}^{|t|} (Y_i - \overline{Y}_t)^2,$$

avec $|t|$ l'effectif du nœud t et $\overline{Y}_t$ la moyenne empirique des observations se trouvant dans le nœud t. Pour un nœud donné, l'erreur globale que l'on commet en séparant les individus suivant un certain critère : $X_j, j \in \{1, ..., p\}, c \in \mathbb{R}$ est donnée par

$$E_t(X_j, c) = \frac{1}{|tg|} \sum_{i=1}^{|tg|} (Y_i - \overline{Y}_{tg})^2 + \frac{1}{|td|} \sum_{i=1}^{|tg|} (Y_i - \overline{Y}_{kd})^2,$$

avec $|tg|$ le nombre d'individus dans le nœud fils gauche de t, et $|td|$ le nombre d'individus dans le nœud fils droit de t
L'objectif est de trouver pour chaque nœud, la variable et la règle de division qui contribuera à la plus forte diminution de l'hétérogénéité de chacun des 2 nœuds fils revenant à minimiser l'erreur $E_t(X_j, c)$. Ce qui revient aussi à dire que l'on conserve la division qui rend le plus significatif possible le test de Fisher comparant les moyennes des deux nœuds fils. La condition de coupure adoptée est donc donnée par $X_{j^*} > c^*$, est telle que

$$(j^*, c^*) = arg \min_{j,c} E_t(X_j, c),$$

X_{j^*} est donc la variable de coupure et c^* est le seuil de coupure. Un biais de sélection apparait.

Soit Y une variable qualitative à m modalités, plusieurs fonctions d'hétérogénéité peuvent être définies pour un nœud : un critère défini à partir de la notion d'entropie ou à partir de la concentration de Gini. En pratique c'est souvent l'indice moyen de Gini qui est choisi par défaut. Ainsi l'hétérogénéité d'un nœud est définie par :

$$R_t = 1 - \sum_{j=1}^{m} p_{j,t}^2,$$

où $p_{j,t}$ est la fréquence de la j-ième classe de Y dans le nœud t.
Comme dans le cas quantitatif, pour un nœud t donné, une perte globale d'information que l'on obtient en séparant les individus suivant $X_j \in \mathbf{c}$ ou non est donné par :

$$G_t(j, \mathbf{c}) = R_{tg} + R_{td}.$$

Dans ce cas $\mathbf{c}$ est sous ensemble des modalités de X_j. Le but est de trouver la division admissible $X_j \in \mathbf{c}$ qui minimise G_t.

Elagage

Nous pouvons construire des arbres excessivement raffinés conduisant à des modèles de prévisions très instables car fortement dépendant de l'échantillon permettant son estimation. Cette situation de sur-ajustement est à éviter. Il est préférable de considérer des modèles plus parcimonieux et donc plus robustes au moment de la prévision. Tous les sous arbres sont admissibles, mais, comme leur nombre est de croissance exponentielle, il n'est pas envisageable de tous les considérer. Pour contourner ce problème, une démarche consistent à construire une suite emboitée de sous-arbres de l'arbre maximal puis de choisir parmi cette suite l'arbre optimal qui minimise un risque.
Soit T un arbre et t un nœud non terminal de T. Elaguer T à partir de t consiste à créer un nouvel arbre T' qui n'est autre que T privé de tous les descendants de t. Tout arbre T' obtenu par élagage de T est un sous-arbre de T, ce que l'on note $T' \subset T$.

Pour un arbre T, le critère d'élagage est donné par

$$R(T) = \sum_{t=1}^{|T|} R_t,$$

et

$$crit_\alpha(T) = R(T) + \alpha|T|,$$

avec $|T|$ le nombre total de ses feuilles et α paramètre de pénalisation qui a pour but de réduire le nombre de feuilles de l'arbre. Pour chaque valeur de α, il existe un unique sous-arbre de T_{max} , noté T_α, tel que

$$T_\alpha = arg \min_{T \subset T_{max}} crit_\alpha(T),$$

si $crit_\alpha(T_\alpha) = crit_\alpha(T)$, alors $T_\alpha \subset T$.

Détermination du premier élément de la suite des arbres
Soit T_0 le premier élément de la suite d'arbres associé à $\alpha = 0$.
T_0 est le sous-arbre de T_{max} obtenu en élagant tous les nœuds t pour lesquels

$$R(t) = R(tg) + R(td).$$

Ainsi T_0 satisfait :
— $crit_0(T_0) = 0$.

— Pour tout nœud t de $T_0, R(t) > R(td) + R(tg)$.

Détermination du deuxième élément de la suite des arbres
Par définition de T_0, on a pour tout nœud t de T_0, $crit_0(t) > crit_0(T_0^t)$, soit $R(t) > R(T_0^t)$.
De ce fait, tant que α demeurera suffisamment petit, on aura

$$R(t) + \alpha > R(T_0^t) + \alpha|T_0^t|,$$

α suffisamment petit signifie

$$\alpha < \frac{R(t)R(T_0^t)}{|T_0^t| - 1} = s(t, T_0^t),$$

pour tout nœud t de T_0.
Lorsque α atteint un seuil, cela signifie que $crit_\alpha(t) = crit_\alpha(T_0^t)$ et que le nœud t devient préférable à la branche issue de t. Posons

$$\alpha_1 = \min_{t \text{ nœud de } T_0^t} s(t, T_0^t),$$

et définissons $T_1 = T_{\alpha_1}$ comme étant le sous-arbre de T_0 obtenu en élagant toutes les branches issues de nœuds minimisant $s(t, T_0^t)$. T_1 est le deuxième élément de la sous-suite d'arbres. Il existe une suite finie strictement croissante de réels positifs notée $(\alpha_k), k \in \{0, ..., K\}$ telle que : $\forall k \in \{0, ..., K-1\}$, si $\alpha \in [\alpha_k, \alpha_{k+1}[$ alors $T_\alpha = T_{\alpha_k} = T_k$ À la fin de la phase d'élagage, nous disposons de plusieurs sous-arbres, $T_0, ..., T_{K-1}$ et donc de plusieurs estimateurs.

4.1.2 Application

À l'aide de la library `rpart`[36] du logiciel `R`[32], nous allons modéliser des arbres de régression *CART*.

Modélisation de la fréquence

Dans le modèle fréquence, la variable à expliquer est le nombre de sinistres et les variables explicatives sont les caractéristiques du véhicule et du conducteur. Pour les mêmes raisons que pour le modèle GLM construit dans la partie 3.4 l'exposition est utilisée comme poids.

Arbre maximal
La première étape consiste à modéliser un arbre maximal. Pour ce faire, le paramètre de complexité dit cp est ajusté (cp pour *complexity parameter* correspond au paramètre α en paragraphe 4.1.1). Un cp égal à 0 ne donne aucune pénalisation à la complexité, à contrario un cp égal à 1 donne une pénalisation maximale à la complexité. Par conséquent, l'arbre maximal est obtenu avec un cp égal à 0.

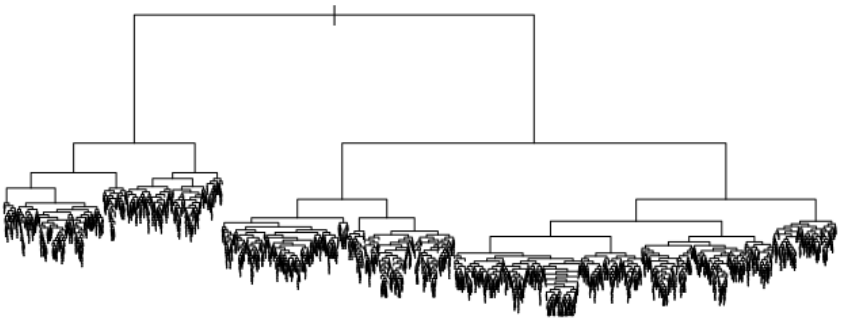

FIGURE 4.1 – Arbre maximal

L'arbre maximal présent sur le graphique 4.1 segmente trop le jeu de données. Par conséquent, un élagage est effectué afin de s'arrêter à la segmentation optimale.

Elagage de l'arbre

Cette partie consiste à élaguer l'arbre maximal. Nous partons de l'arbre minimal avec 0 division puis à chaque division nous calculons le cp, l'erreur de validation croisée et l'écart type de l'estimation de l'erreur de validation croisée. Dans la pratique sur le logiciel R[32], ces sorties sont présentes dans une variable de l'arbre maximal nommé *cptable*. Plutôt que de choisir directement une valeur de cp telle que l'erreur de validation croisée soit minimal nous appliquons la régle de *One Standard Error* [5] appelé aussi *"1-SE"*, règle de l'écart type. Cette régle consiste à créer un seuil égal à la somme du minimum de l'erreur de validation croisée avec l'écart type de l'estimation de l'erreur de validation croisée associée à ce minimum. Le cp optimal est le maximum des cp (arbre le moins complexe) dont l'erreur de validation croisée est inférieur à ce seuil.

	CP	**nsplit**	**rel error**	**xerror**	**xstd**
9	0,00139	10	0,93406	0,93615	0,00609
10	0,00107	12	0,93128	0,93431	0,00607
11	0,00095	13	0,93021	0,93363	0,00607
12	**0,00084**	14	0,92925	0,93281	0,00607
...	...	...	...	...	...
22	0,00032	43	0,91485	0,92739	0,00604
23	0,00032	44	0,91453	0,92695	0,00604
24	0,00032	45	0,91420	**0,92690**	**0,00604**
25	0,00031	46	0,91388	0,92749	0,00605
26	0,00029	48	0,91326	0,92780	0,00606
27	0,00029	49	0,91297	0,92871	0,00608

TABLE 4.1 – Étape d'élagage de l'arbre fréquence

Le tableau correspond à la sortie *cptable* de l'arbre maximal. Pour chaque ligne, nous avons le nombre

de divisions nsplit, le cp et les mesures associées (*xerror*=erreur de validation croisée, *xstd*=ecart type de l'estimation de l'erreur de validation croisée). Concrètement, le seuil est déterminé en ajoutant 0,92690 et 0,00604 ce qui donne 0,93294. Le premier cp avec un *xerror* inférieur au seuil est atteint lors de la treizième itération.

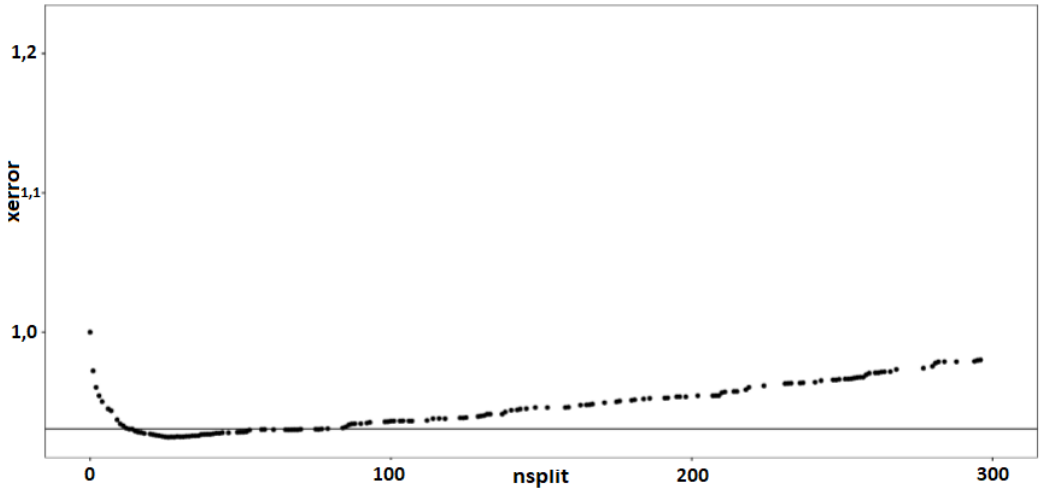

FIGURE 4.2 – Optimisation du cp

Le graphique 4.2 correspond à l'erreur de validation croisée en fonction du nombre de segmentation. Nous avons également tracé une droite correspondant au seuil. Le cp optimal est associé à la première erreur en dessous de ce seuil. L'arbre final est déterminé en utilisant la commande prune, l'arbre maximal et le cp optimal. Nous pouvons faire un commentaire concernant la sur-segmentation des données vu dans la partie 1.2. L'erreur n'est pas une fonction décroissante en fonction du nombre de segmentation. Cet exemple illustre le fait que une forte segmentation n'optimise pas l'erreur.

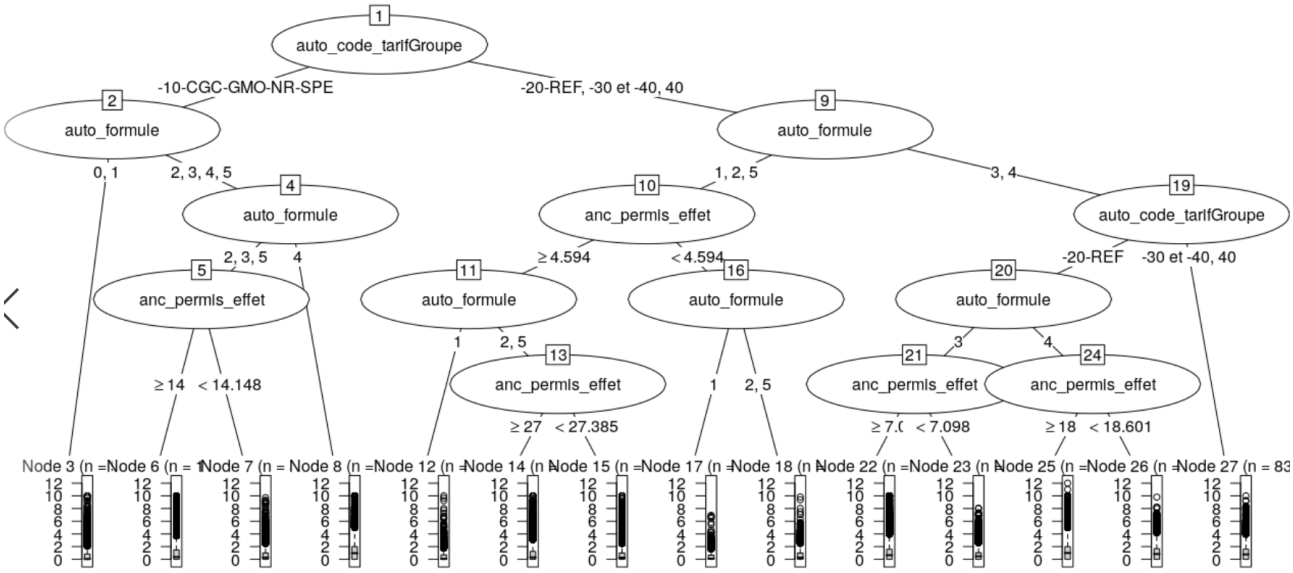

FIGURE 4.3 – Arbre optimal fréquence

L'arbre optimal sur le graphique 4.3 a 14 feuilles terminales (la variable à expliquer Y a 14 possibilités de fréquence de sinistralité). Chaque observation appartient à une feuille en fonction de ses variables explicatives et est associée à une valeur de fréquence de sinistralité.

Nous avons le modèle *CART* optimal sur notre jeu de données d'apprentissage. Comme pour les modèles GLM, les prédictions ont été effectuées sur un échantillon test indépendant de l'échantillon d'apprentissage.

MSE	**RMSE**	**MAE**
0.600	0.774	0.490

TABLE 4.2 – MSE, RMSE et MAE sur l'échantillon test

Les erreurs sur le tableau 4.2 sont celles du modèle final CART fréquence sur l'échantillon test. Ces erreurs seront comparées avec les erreurs des autres modèles présentés dans ce chapitre partie 4.4.

Modélisation de la sévérité

Arbre maximal

De même que pour la modélisation de la fréquence pour le modèle *CART*, nous construisons l'arbre maximal en ne pénalisant pas la complexité de l'arbre (cp=0).

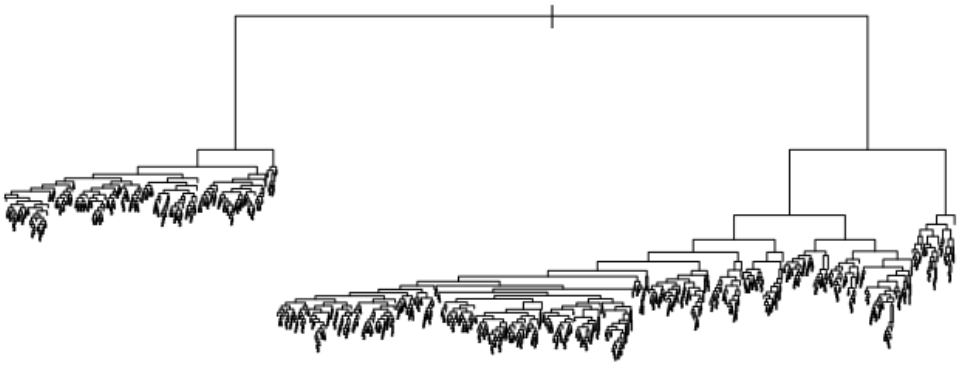

FIGURE 4.4 – Arbre maximal coût

L'arbre maximal sur le graphique 4.4 doit subir un élagage afin d'obtenir la segmentation optimale.

Elagage de l'arbre

	CP	nsplit	rel error	xerror	xstd
1	0,016320	0	1	1,000126	0,034793
2	0,007990	1	0,983680	0,983896	0,034267
3	0,002992	2	0,975690	0,979292	0,034087
4	0,002068	3	0,976788	0,976998	0,034053
5	**0,001531**	4	0,970629	**0,972698**	0,034011
6	0,001417	5	0,969099	0,976976	0,033949
7	0,001143	6	0,967682	0,979835	0,034040
8	0,001139	7	0,966538	0,984919	0,034052

TABLE 4.3 – Étape d'élagage de l'arbre

Le tableau 4.3 présente la sortie *cptable* de l'arbre maximal. Nous voulions utiliser la méthode *One Standard Error* comme pour le modèle fréquence mais cette méthode nous donne un cp sans division car l'erreur de validation croisée en fonction du nombre de division est croissante à partir du seuil. Le modèle avec aucune division est celui dont toutes les observations ont un montant de sinistres prédit égal au coût moyen. Par conséquent, nous décidons de choisir pour l'élagage de l'arbre optimal le cp tel que l'erreur de validation croisée soit minimine c'est à dire 0,001531.

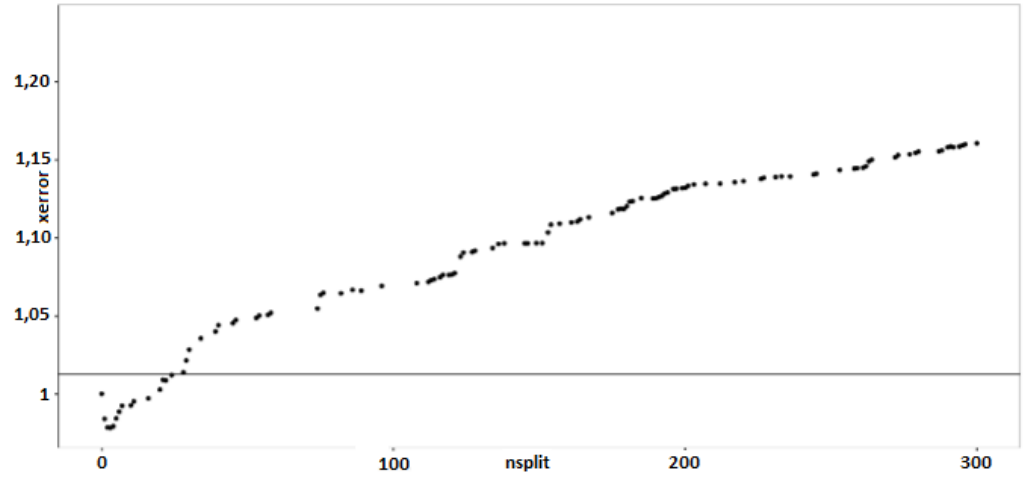

FIGURE 4.5 – Optimisation du cp

Le graphique 4.5 explique le choix de prendre un paramètre de compléxité minimisant l'erreur de validation croisée car choisir le cp minimisant le nsplit tout en étant en dessous du seuil revient à prendre un modèle avec une feuille terminale c'est à dire un modèle avec une même prédiction en coût moyen pour toutes les observations étant la moyenne des coûts moyens.

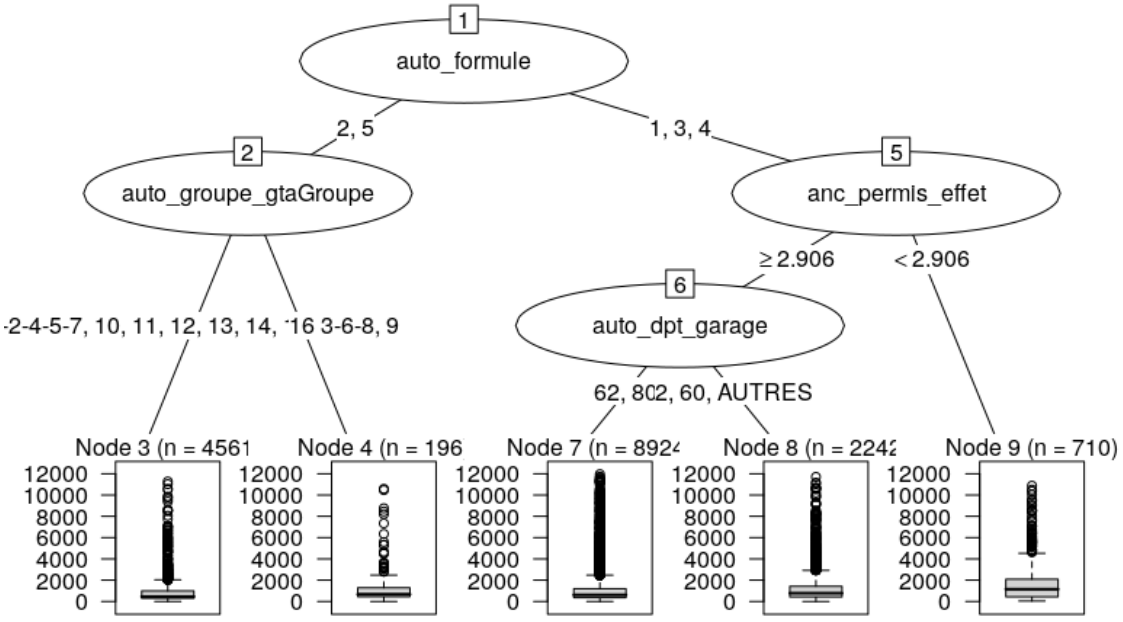

FIGURE 4.6 – Arbre optimal coût

L'arbre optimal sur le graphique 4.6 ne contient que 4 feuilles terminales. Le faible nombre de feuilles terminales traduit que les prédictions des coûts moyens ne peuvent prendre que 4 valeurs. Ce nombre de feuilles peut s'expliquer par l'homogénéité de nos données par rapport au coût moyen.

MSE	**RMSE**	**MAE**
1 766 018	1 329	781

TABLE 4.4 – MSE, RMSE et MAE sur l'échantillon test

Les résultats présentés dans le tableau 4.4 seront comparés avec les résultats des autres modèles que nous introduirons dans ce chapitre partie 4.27.

4.2 *Extreme gradient boosting*

4.2.1 Présentation du modèle

Les algorithmes basés sur l'aggrégation des modèles ont été développés afin d'améliorer l'ajustement tout en évitant ou en contrôlant le sur-ajustement. Dans cette section, nous allons présenter le principe de ces algorithmes dans le cas où la variable à expliquer est quantitative.

Bagging

L'algorithme *Bagging* pour *bootstrap* aggregation a été introduit par Breiman [3]. Soit Y une variable à expliquer $X = (X_1, ..., X_p)$ les variables explicatives et $f(X)$ un modèle fonction de X. Soit $E = ((X_1, Y_1), ..., (X_n, Y_n))$ un échantillon de loi F.
Considérant B échantillons indépendants notés $\{E_b\}_{b=1,...,B}$, une prévision par agrégation de modèles, dans le cas où la variable à expliquer Y est quantitative, est définie par

$$\hat{f}_B(.) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{E_b}(.).$$

Il s'agit d'une simple moyenne empirique des résultats obtenus pour les modèles associés à chaque échantillon.
Cependant, il n'est pas toujours commode de supposer les échantillons indépendants, notamment lorsque le nombre d'observations n'est pas très grand, d'où l'idée d'utiliser le bootstrap. Les B échantillons indépendants, sont donc remplacés par B réplications d'échantillons bootstrap obtenus chacun par n tirages avec remise selon la loi empirique $\hat{F}$.

: Algorithme du Bagging

Soit X_0 à prévoir et $E = ((X_1, Y_1), ..., (X_n, Y_n))$ un échantillon
pour *b=1 à B* **faire**
- Tirer un échantillon bootstrap E_b
- Estimer $\hat{f}_{E_b}(X_0)$ sur l'échantillon bootstrap

fin
Calculer l'estimation moyenne $\hat{f}_B(X_0) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{E_b}(X_0)$.

Boosting

Le boosting adopte le même principe général que le bagging : construction d'une famille de modèles qui sont ensuite agrégés par une moyenne pondérée des estimations. Dans le boosting chaque modèle est une version adaptative du précédent en donnant plus de poids, lors de l'estimation suivante, aux observations mal ajustées ou mal prédites. Il donne donc plus d'importance aux observations les plus difficiles à ajuster tandis que l'agrégation de l'ensemble des modèles réduit le risque de sur-ajustement.
Les algorithmes de boosting sont caractérisés par :

- La façon de pondérer l'importance des observations mal estimées lors de l'itération précédente.
- Leur objectif selon le type de la variable à prédire Y : binaire, qualitative à k classes ou quantitative.
- La fonction perte pour mesurer l'erreur d'ajustement.
- La façon d'agréger les modèles de base successifs.

La littérature sur le sujet présente donc de très nombreuses versions de cet algorithme ; qui vont du premier algorithme *AdaBoost* introduit par Freund et Schapire [17] au dernier *XGBoost* proposé par Chen et Guestrin [8]. L'algorithme suivant est une version du boosting, proposé par Drucker [13], adaptée à la régression.

: Algorithme du Boosting pour la régression

Soit X_0 à prévoir et
$E = ((X_1, Y_1), ..., (X_n, Y_n))$ un échantillon
Initialiser p par la distribution uniforme $\mathbf{p} = \{p_i = 1/n, i = 1, ..., n\}$
pour *m=1 à M* **faire**
- Tirer avec remise dans E un échantillon E_m^* suivant $\mathbf{p}$
- Estimer $\hat{f}_m$ sur l'échantillon E_m^*
- Calculer à partir de l'échantillon initial E :
  $l_m(i) = \mathbf{l}(Y_i, \hat{f}_m(X_i)), i = 1, ..., n,$ (l : fonction perte)
  $\hat{e}_m = \sum_{i=1}^n p_i l_m(i)$;
  $w_i = g(l_m(i))p_i$, (g continue non décroissante)
- Calculer les nouvelles probabilités : $p_i = \frac{w_i}{\sum_{i=1}^n w_i}$

fin
Calculer $\hat{f}(X_0)$ moyenne ou médiane des prévisions $\hat{f}_m(X_0), 1 \leq m \leq M$, pondérées par des coefficients $log(\frac{1}{\beta_m})$

La fonction perte $\mathbf{l}$ peut être par exemple quadratique, $\mathbf{l}(y, f(x)) = (y - f(x))^2/2$, et β_m est donné par :

$$\beta_m = \frac{\hat{e}_m}{L - \hat{e}_m}, \; L_m = \sup_{i=1,...n} l_m(i).$$

L_m étant le maximum d'erreur observée par le modèle $\hat{f}_m$ sur le modèle initial.

Boosting et gradient adaptatif

Dans le même esprit d'approximation adaptative, Friedman [16] a proposé (*gradient boosting machine)*[1] une famille d'algorithmes basés sur une fonction perte supposée convexe et différentiable notée $\mathbf{l}$. Le principe de base est le même que pour *AdaBoost*, construire une séquence de modèles de sorte que dans chaque étape, chaque modèle ajouté à la combinaison apparaisse comme un pas vers une meilleure solution. La principale innovation est que ce pas est franchi dans la direction du gradient de la fonction perte, afin d'améliorer les propriétés de convergence. On présente ci-dessous l'algorithme du Gradient Tree Boosting qui une version du gradient boosting, proposée par Friedman [16], adaptée à la régression.

1. dont l'acronyme est GBM

: Algorithme du *Gradient Boosting Machine* pour la régression

Soit X_0 à prévoir
Initialiser $\hat{f}_0 = argmin_\gamma \sum_{i=1}^{n} \mathbf{l}(Y_i, \gamma)$
pour *m=1 à M* **faire**
- Calculer $r_{mi} = -\left[\frac{\partial \mathbf{l}(Y_i, f(X_i))}{\partial f(X_i)}\right]_{f=f_{m-1}}, i = 1, ..., m$
- Ajuster un arbre de régression A_m aux couples $(r_{mi}, X_i), i = 1, ..., n$
- Calculer à partir de l'échantillon initial E :
  $l_m(i) = \mathbf{l}(Y_i, \hat{f}_m(X_i)), i = 1, ..., n$, (l : fonction perte)
- Calculer γ_m en résolvant : $\min_\gamma \sum_{i=1}^{n} \mathbf{l}(Y_i, f_{m-1}(X_i) + \gamma A_m(X_i))$,
- Mettre à jour : $\hat{f}_m(x) = \hat{f}_{m-1}(x) + \gamma_m A_m(x)$

fin
Calculer la prévision finale :$\hat{f}_M(X_0)$.
.

Extreme gradient boosting
Plus récemment, Chen et Guestrin [8] ont proposé une dernière version du boosting avec *L'eXtreme Gradient Boosting* (*XGBoost*). Dans cet algorithme, le développement de *Taylor* à l'ordre 2 de la fonction objectif permet d'optimiser la suite d'arbres de régressions. Le nombre de paramètres qu'il est nécessaire de prendre en compte est assez conséquent, ce qui rend l'algorithme très coûteux en temps de calcul. Cependant, l'implementation de *XGBoost* permet de parallèliser les calculs engendrant un gain en temps de calcul considérable. L'algorithme *XGBoost* inclut plusieurs fonctionnalités chacune associée avec un ou des paramètres supplémentaires à optimiser. Pour plus de détails, nous renvoyons le lecteur à l'article original de Chen et Guestrin [8].

4.2.2 Application

Pour les modèles *XGBoost*, les variables qualitatives sont transformées en variables binaires car *XGBoost* n'accepte que des variables quantitatives. À l'aide des library `xgboost`[9] et `caret`[22] de `R`[32], nous allons créer des modèles *XGBoost*. Les paramètres de tunning des modèles *XGBoost* que nous optimisons sont :

- *max_ depth* correspondant à la profondeur d'arbre maximal.
- *colsample_ bytree* est le pourcentage des variables utilisées pour construire un modèle.
- *subsample* est le pourcentage des observations utilisées pour construire un arbre.
- eta est le taux d'apprentissage.
- *gamma* correspond à la régularité du modèle, plus le paramètre est grand plus le modèle sera lisse.
- *min_ child_ weight* est le nombre d'observations minimum dans chaque nœud pour poursuivre le développement de l'arbre.
- *nround* est le nombre d'arbres à implémenter.

Nous allons créer plusieurs modèles à l'aide de l'échantillon d'apprentissage en utilisant plusieurs combinaisons de paramètres et retenir le modèle ayant minimisé l'erreur. Au départ, nous avions mesuré l'erreur sur l'échantillon de validation et sélectionné les paramètres minimisant l'erreur. Malheureusement, ce modèle n'a pas donné de bonne performance sur l'échantillon test. Par conséquent, nous avons décidé d'effectuer une *k-fold* validation croisée à 80 % sur l'échantillon d'apprentissage (avec k=5). Nous illustrons cette méthode par ce schéma :

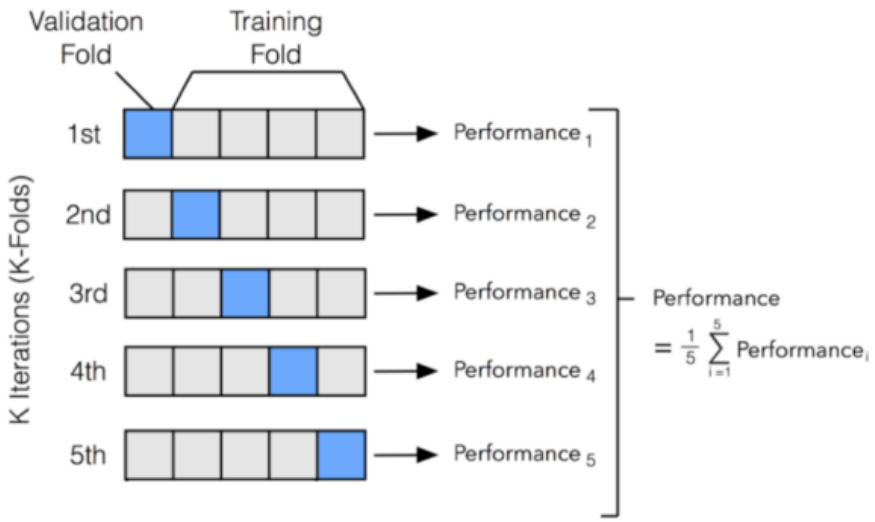

FIGURE 4.7 – Exemple de validation croisée avec k=5

Sur la figure 4.7, nous observons que notre échantillon d'apprentissage est divisé en 5. Les 20 premiers pourcents de l'échantillon d'apprentissage sont l'échantillon de validation. Les 80 autres pourcents sont l'échantillon d'apprentissage. Le modèle est modélisé sur l'échantillon d'apprentissage et l'erreur est calculée sur l'échantillon de validation. Nous réitérons ce processus sur nos 4 autres échantillons de validation distincts en prenant en échantillon d'apprentissage le complémentaire. L'erreur finale est donnée par la moyenne des erreurs. Ce procédé est effectué à chaque combinaison de paramètres modélisant un modèle *XGBoost*. Les paramètres choisis sont ceux dont le modèle associé minimise l'erreur.
Après avoir effectué des essais, nous avons remarqué que lorsque le paramètre *nround* est supérieur à 50, le modèle ne s'améliore pas et fait du surapprentissage. De ce fait, nous choisissons de fixer *nround* à 50 et essayons les différentes combinaisons possibles de ces paramètres de *tuning* :

Paramètres					
eta	0,01	0,1	0,5		
max_depth	1	3	5		
colsample_bytree	0,1	0,3	0,5	0,8	1
subsample	0,3	0,5	0,8	1	
gamma	0	0,2	0,5	0,8	1
min_child_weight	0,2	0,5	1		

TABLE 4.5 – Les différents paramètres utilisés pour la modélisation du modèle *XGBoost*
La grille de paramètres 4.5 permet de créer 2 700 combinaisons de paramètres créeant 2 700 modèles *XGBoost*.

Modélisation de la fréquence

Pour les modèles *XGBoost*, l'utilisation des variables est identique à celle de la partie 4.1.2.

Modèle *XGBoost* naïf

Nous effectuons un modèle *XGBoost* naïf, c'est à dire un modèle sans remplir les paramètres de tunning (des valeurs par défaut sont associées à ces paramètres dans la fonction `xgboost`[9] de `R`[32] mais ne sont pas optimaux sur tous les jeux de données). La prédiction et l'erreur sont calculées sur l'échantillon test.

MSE	RMSE	MAE
0,731	0,855	0,614

TABLE 4.6 – MSE, RMSE et MAE sur l'échantillon test fréquence du modèle naïf

Le tableau 4.6 présente les différentes métriques associées au modèle naïf. Ces mesures d'erreur vont permettre de comparer le gain en performance avec le modèle *XGBoost* optimisé.

Optimisation des paramètres

Les différentes combinaisons des paramètres du tableau 4.5 permettent de modéliser plusieurs modèles *XGBoost*. Nous choisissons les paramètres du modèle *XGBoost* minimisant le RMSE.

Combi	eta	maxdep	gamma	colsample	minchild	subsample	nround	RMSE	MAE
1 771	0,1	5	1	0,5	0,5	0,8	50	0,751	0,537
1 711	0,1	5	0,8	0,5	0,5	0,8	50	0,751	0,537
1 547	0,1	5	0	0,8	1	0,8	50	0,751	0,537
1 715	0,1	5	0,8	0,5	1	0,8	50	0,751	0,537
1 528	0,1	5	0	0,5	0,2	1	50	0,751	0,537
1 772	0,1	5	1	0,5	0,5	1	50	0,751	0,537
1 591	0,1	5	0,2	0,5	0,5	0,8	50	0,751	0,537
1 604	0,1	5	0,2	0,8	0,5	1	50	0,751	0,537
1 651	0,1	5	0,5	0,5	0,5	0,8	50	0,751	0,538
1 531	0,1	5	0	0,5	0,5	0,8	50	0,751	0,537

TABLE 4.7 – MSE, RMSE et MAE sur l'échantillon d'apprentissage fréquence par validation croisée

Le tableau 4.7 correspond à un tri des meilleurs RMSE par validation croisée sur l'échantillon d'apprentissage fréquence. Notre choix de paramètres pour le modèle optimisé est la combinaison numéro 1 771 minimisant le RMSE parmi les 2 700 combinaisons.

Paramètres	Optimal
eta	0,1
max_depth	5
colsample_bytree	0,5
subsample	0,8
gamma	1
min_child_weight	0,5
nrounds	50

TABLE 4.8 – Les paramètres optimaux du modèle *XGBoost* fréquence

Le tableau 4.8 correspond aux paramètres du modèle *XGBoost* fréquence optimal. Le temps de calcul est de 1 heure et 48 minutes en ayant paralléliser les calculs sur un serveur de calcul contenant 30 cœurs.

Modèle	MSE	RMSE	MAE
Sans optimisation	0,731	0,855	0,614
Avec optimisation	0,603	0,777	0,551
Pourcentage d'écart	-18 %	-10%	-10%

TABLE 4.9 – MSE, RMSE et MAE sur l'échantillon test fréquence

Les différentes erreurs du modèle *XGBoost* fréquence optimal sur l'échantillon test ainsi que leurs écarts avec le modèle *XGBoost* naïf sont donnés par le tableau 4.9. Plus précisément, le pourcentage d'écart correspond à l'évolution entre l'erreur des modèles sans optimisation et avec optimisation. L'optimisation des paramètres du modèle *XGBoost* a permis un gain de performance. Ces erreurs seront également comparées avec les erreurs des autres modèles dans la partie 4.4.

Modélisation de la sévérité

La variable à expliquer est le coût moyen d'un sinistre et les variables explicatives sont les caractéristiques du conducteur et du véhicule.

Modèle *XGBoost* naïf

Nous effectuons un modèle *XGBoost* à l'aide de notre échantillon d'apprentissage avec les paramètres par défaut. La prédiction et l'erreur sont calculées sur l'échantillon test.

MSE	RMSE	MAE
2 619 251	1 618	1 054

TABLE 4.10 – MSE, RMSE et MAE sur l'échantillon test coût du modèle naïf

Le tableau 4.10 permet de comparer les métriques avec le modèle optimisé afin de s'assurer du gain de performance.

Optimisation des paramètres

Les différentes combinaisons des paramètres du tableau 4.5 permettent de modéliser plusieurs modèles *XGBoost*.

Combi	eta	maxdep	gamma	colsample	minchild	subsample	nround	RMSE	MAE
1 459	0,1	3	1	0,3	0,5	0,8	50	1 327,77	783,11
2 051	0,5	1	1	0,1	1	0,8	50	1 327,89	786,34
1 403	0,1	3	0,8	0,3	1	0,8	50	1 327,97	783,56
1 467	0,1	3	1	0,5	0,2	0,8	50	1 327,98	782,76
1 336	0,1	3	0,5	0,3	0,2	1	50	1 328,13	783,64
1 283	0,1	3	0,2	0,3	1	0,8	50	1 328,14	783,26
1 338	0,1	3	0,5	0,3	0,5	0,5	50	1 328,17	783,24
1 984	0,5	1	0,8	0,1	0,2	1	50	1 328,19	786,31
1 395	0,1	3	0,8	0,3	0,2	0,8	50	1 328,21	783,40
1 948	0,5	1	0,5	0,5	0,2	1	50	1 328,23	786,05

TABLE 4.11 – MSE, RMSE et MAE sur l'échantillon d'apprentissage coût par validation croisée

Comme pour le modèle *XGBoost* fréquence, le tableau 4.11 correspond à un tri des meilleurs RMSE par validation croisée. Notre choix de paramètres pour le modèle optimisé est la combinaison numéro 1 459 minimisant le RMSE parmi les 2 700 combinaisons.

Paramètres	**Optimal**
eta	0,1
max_depth	3
colsample_bytree	0,3
subsample	0,8
gamma	1
min_child_weight	0,5
nrounds	50

TABLE 4.12 – Les paramètres optimaux du modèle *XGBoost* coût

Le tableau 4.12 correspond aux paramètres du modèle *XGBoost* coût optimal. Le temps de calcul est de 38 minutes en ayant paralléliser les calculs sur un serveur de calcul contenant 30 cœurs.

Modèle	**MSE**	**RMSE**	**MAE**
Sans optimisation	2 619 251	1 618	1 054
Avec optimisation	1 756 596	1 325	775
Pourcentage d'écart	-33 %	-18%	-26 %

TABLE 4.13 – MSE, RMSE et MAE sur l'échantillon test coût

Le tableau 4.13 donne les erreurs du modèle *XGBoost* final ainsi que les écarts de performance avec le modèle naïf. L'optimisation des paramètres du modèle *XGBoost* a permis un gain de performance. Ce modèle sera comparé avec les autres modèles coût de la partie 4.27.

4.3 *Model-based Recursive Partitioning*

4.3.1 Présentation du modèle

Le *MOdel-Based recursive partionning* (MOB), introduit par Zeileis et al. [43], Rusch et al. [33], a pour but d'élaborer un modèle segmenté $M(Y, X, Z, \beta_b), b = 1, ..., B$ afin d'avoir un meilleur ajustement des données que le modèle global $M(Y, X, Z, \beta)$. Le principe de la modélisation est décrit comme suit. La ségmentation est faite en appliquant une méthode de classification sur des variables de partitionnements $(Z_1, ..., Z_r)$. Ainsi la partition $I_b, b = 1, ..., B$ de l'espace $\mathcal{Z} = \mathcal{Z}_1 \times ... \times \mathcal{Z}_r$ engendré par les variables Z_j partage les données initiales en B segments distincts $\mathcal{B}_b$. Dans chaque segment $\mathcal{B}_b$ un modèle $M(Y, X, \beta_b)$ est estimé. Dans la suite nous allons illustrer le modèle MOB en utilisant une méthode de classification similaire à celle de CART et un modèle GLM dans chaque segment. La méthode CART fournit un arbre binaire, par conséquent le nombre de partitions noté B pour notre modèle *MOB* est égal à 2.

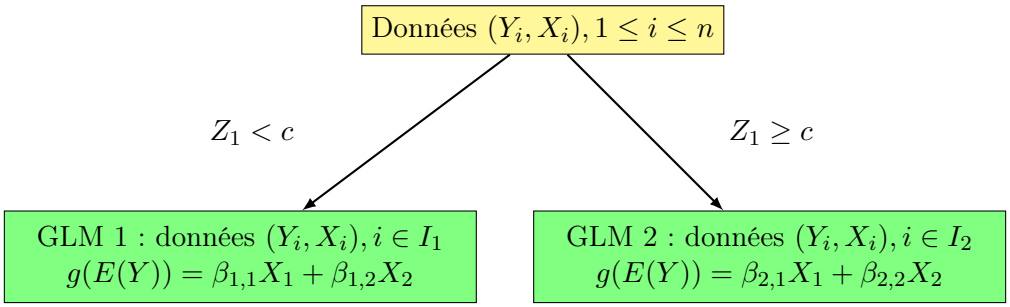

FIGURE 4.8 – Modèle MOB avec deux feuilles, deux variables explicatives (X_1, X_2) et une variable de partitionnement quantitative $Z_1, I_1 \bigcup I_2 = \{1, ..., n\}$.

La figure 4.8 présente un modèle MOB à deux feuilles. Nous observons que chaque feuille à son propre modèle GLM. Nous rappelons que le point faible du modèle GLM est de traiter seulement des relations linéaires. Dans le modèle MOB, l'introduction d'un arbre de décisions permet de traiter en amont de la non-linéarité des données tout en se ramenant à des modèles qui sont localement linéaires.

Modèle linéaire généralisé segmenté

Considérons le modèle GLM, $M(Y, Z, X, \beta)$, où Y est la variable à expliquer, X sont les variables explicatives (prédicteurs), Z sont les variables de partitionnement et β le paramètre inconnu.
Etant donné n observations $(Y_i, X_i, Z_i, i = 1, ..., n)$, le modèle peut être ajusté en minimisant une fonction objectif $\Phi(Y, X, \beta)$ qui s'exprime souvent sous la forme

$$\Phi(Y, X, \beta) = \sum_{i=1}^{n} \Phi(Y_i, X_i, \beta). \tag{4.1}$$

L'estimateur du paramètre β est donné par

$$\hat{\beta} = arg \min_{\beta} \Phi(Y, X, \beta).$$

Le modèle GLM segmenté, $M(Y, X, \beta_b), b = 1, ..., B$, nécessite l'estimation du paramètre $\beta = (\beta_1, ..., \beta_B)$, et donc la minimisation de la fonction objectif globale s'écrit

$$\sum_{b=1}^{B} \sum_{i \in I_b} \Phi(Y_i, X_i, \beta_b), \tag{4.2}$$

où I_b est l'ensemble des indices des observations appartenant au segment $\mathcal{B}_b$.
Cependant, la minimisation de la fonction (4.2) peut être facilement obtenue en minimisant localement $\sum_{i \in I_b} \Phi(Y_i, X_i, \beta_b)$. Une fois les estimateurs $\hat{\beta}_b$ dans chaque segment $\mathcal{B}_b$ obtenus, l'estimateur optimal de β est donné par $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_B)$. Plus précisement, pour notre modèle GLM, $\hat{\beta}_b$ est l'estimateur du maximum de vraisemblance de β_b et est déterminé en prenant les estimateurs des GLM (voir équations 3.1-3.2)

$$\Phi(Y_i, X_i, \beta_b) = -\mathcal{L}(Y_i, \beta_b) \tag{4.3}$$

$$= -\left(\sum_{i \in I_b} \frac{Y_i \theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi) \right), \tag{4.4}$$

avec $\mathcal{L}(Y_i, \beta_b)$ la log-vraisemblance de la variable Y_i appartenant au segment $\mathcal{B}_b$.

Algorithme de partitionement récursif

L'intérêt de l'algorithme est que si le modèle GLM global pour toutes les n observations n'est pas valide et si d'autres variables explicatives $Z_1, ..., Z_r$ sont disponibles, il pourrait être possible de partitionner les n observations par rapport à ces variables et trouver un modèle GLM approprié dans chaque segment de la partition. L'algorithme décrit ici tente de trouver une telle partition de manière adaptative.
L'idée de base est qu'à chaque nœud terminal correspond un seul modèle. Pour évaluer si la division d'un nœud est nécessaire, un test de fluctuation pour l'instabilité des paramètres est effectué. S'il y a une instabilité par rapport à l'une des variables de partitionnement Z_j on divise le nœud en deux fils droit et gauche et on répète la procédure. Si aucune instabilité significative n'est détectée l'algorithme de partitionement récursif s'arrête et retourne un arbre où chaque nœud terminal (ou feuille) est associé à un modèle GLM. Plus précisément, les étapes de l'algorithme sont :

1. Ajuster un modèle GLM en utilisant toutes les n observations $(Y_i, X_i, i = 1, ..., n)$, dans le nœud racine en estimant β via la minimisation de la fonction objectif du modèle GLM, Φ donnée par (4.1)-(4.4).
2. Évaluer si les estimations des paramètres sont stables par rapport à chacune des variables $Z_1, ..., Z_r$. S'il y a une instabilité globale, sélectionner la variable Z_j associée à l'instabilité la plus élevée. S'il y a stabilité globale arrêter l'algorithme (aucune segmentation n'est effectuée).
3. S'il y a une instabilité, diviser le nœud en deux nœuds fils et répéter la procédure sur chacun des nœuds fils.

Pour un nœud donné b, soient n_b le nombre d'observations lui appartenant, I_b l'indice de ces observations et soit β_b le paramètre associé au modèle GLM estimé sur ce nœud.

Estimation des paramètres

L'estimateur $\hat{\beta}_b$ de β_b est calculé en résolvant l'équation

$$\sum_{i \in I_b} S_b(Y_i, \beta_b) = 0, \tag{4.5}$$

où

$$S_b(Y_i, \beta_b) = \frac{\partial \mathcal{L}(Y_i, \beta_b)}{\partial \beta_b}$$

est le vecteur score et $\mathcal{L}(Y_i, \beta_t)$ est la log-vraisemblance des observations présentes dans le nœud b donnée par (4.4).

Test d'instabilité des paramètres

La deuxième étape permet de savoir si la segmentation de l'échantillon par rapport à Z_j peut capter des instabilités dans les paramètres et améliorer le modèle. Pour évaluer l'instabilité des paramètres, l'idée est de vérifier si les vecteurs scores $S_b(Y_i, \hat{\beta}_b)$ présentent des écarts systématiques de 0 sur Z_j. Ces écarts peuvent être capturés par le processsus de fluctuation empirique (voir [41], [42])

$$W_j(t) = \hat{J}^{-1/2} n_b^{-1/2} \sum_{i=1}^{\lfloor n_b t \rfloor} S_b(Y_i, \hat{\beta}_b)_{\sigma(Z_{ij})} \quad (0 \leq t \leq 1),$$

où $\sigma(Z_{ij})$ est la permutation d'ordre qui donne l'antirang de l'observation Z_{ij} dans le vecteur $Z_j = (Z_{1j}, \ldots, Z_{nj})^\top$. W_j est le processus de la somme partielle des scores classé par la variable Z_j convenablement normalisé. $\hat{J}$ est la matrice de covariance empirique du vecteur score $S_b(Y_i, \hat{\beta}_b)$ c'est à dire, $\hat{J} = n_b^{-1} \sum_{i=1}^{n_b} S_b(Y_i, \hat{\beta}_b) \left(S_b(Y_i, \hat{\beta}_b)\right)^\top$.
Pour évaluer les instabilités sur une variable quantitative Z_j, nous utilisons la statistique suivante :

$$\lambda_{\text{sup } LM}(W_j) = \max_{i=\underline{i},\ldots,\bar{i}} \left(\frac{i}{n_b} \cdot \frac{n_b - i}{n_b}\right)^{-1} \left\| W_j\left(\frac{i}{n_b}\right) \right\|_2^2,$$

$[\underline{i}, \ldots, \bar{i}]$ est l'intervalle associé aux points de ruptures potentiels de la variable Z_j.
Pour évaluer les instabilités sur une variable qualitative Z_j avec C modalités de Z_j, nous utilisons la statistique suivante

$$\lambda_{\chi^2}(W_j) = \sum_{c=1}^{C} \frac{|I_c|^{-1}}{n_b} \left\| \Delta_{I_c} W_j\left(\frac{i}{n_b}\right) \right\|_2^2,$$

où $\Delta_{I_c} W_j$ est l'accroissement du processus de fluctuation empirique sur les observations de la modalité c=1,..,C (associé à I_c), c'est à dire la somme des scores de la catégorie c.

Segmentation d'une variable

Dans la deuxième étape, le modèle est divisé par rapport à une variable Z_j. L'objectif est de connaitre le seuil de division pour les variables quantitatives et les deux groupes de modalités choisis pour les variables qualitatives. Nous utilisons encore la fonction objectif $\sum_{b=1}^{B} \sum_{i \in I_b} \Phi(Y_i, X_i, \beta_b)$ avec $B = 2$. Pour les variables quantitatives, la partition optimale est basée sur les méthodes de détection de points de ruptures et de changement structurel [2]. Pour les variables qualitatives, l'algorithme consiste à tester toutes les combinaisons de modalités possibles afin de créer deux groupes de modalités optimisant la fonction objectif.

Optimisation du modèle

Dans le modèle *MOB*, nous devons choisir des variables de classification et des variables de régression. Les variables de classification sont les variables segmentant le jeu de données. Les variables de régression sont les variables modélisant les différents modèles de GLM associés à chaque sous échantillon. Nous précisons que le modèle *MOB* a des paramètres similaires à ceux du modèle *CART* comme par exemple le paramètre de complexité pénalisant la complexité de l'arbre. Nous avons fait varier les différents paramètres (complexité de l'arbre, profondeur de l'arbre et nombre d'observations minimales par feuille terminale) mais les résultats sur notre jeu de données n'ont pas été amélioré.
Dans cette partie, aucune documentation extérieure n'est présente. L'objectif de cette partie est de choisir la combinaison optimale de variables de régression et de classification du modèle *MOB*. Nous

avons utilisé 3 approches. La première approche se base sur la classification effectuée par le modèle *CART* car ces deux modèles se ressemblent. La seconde approche est une variante de l'approche *backward* vu pour les approches de régression linéiaire. La dernière approche est une approche utilisant toutes les combinaisons de variables de classification et régression possibles.

Première approche
La première approche est une approche intuitive. En effet, les modèles MOB sont une combinaison des modèles GLM et des modèles *CART*. Par conséquent, nous choissons d'utiliser en variables de classification les variables les plus importantes du modèle *CART*. Les autres variables (ou une sélection des autres variables) seront utilisées en variables de régression.

Deuxième approche
Dans cette partie, l'approche utilisée est similaire aux méthodes *backward-forward* lors de la sélection des variables optimales dans le modèle GLM. Nous partons de toutes les variables et nous prenons une variable de classification parmi les variables, les autres variables sont utilisées en variables de régression. La variable minimisant notre erreur est sélectionnée en variable de classification. Par la suite, nous enlevons une variable parmi les variables de régression et nous essayons cette variable et la variable de classification précédemment choisi. Les modèles ne contiennent qu'une variable de classification. Nous réitérons cette étape (une étape est une suppression d'une variable de régression afin de la tester comme variable de classification). Si l'erreur augmente d'une étape à l'autre, nous nous arrêtons. Une fois le modèle final choisit (le modèle minimisant l'erreur parmi tous les modèles), nous essayons d'ajouter les variables non utilisées en variable de classification afin d'optimiser le modèle. Nous explicitons l'algorithme utilisé

: Algorithme de la deuxième approche d'optimisation des modèles *MOB*

Entrée : $E_{app} = (Y_i, X_{i,j}, 1 \leq i \leq p, 1 \leq j \leq m)$ échantillon d'apprentissage
Entrée : $E_{val} = (Y_i^*, X_{i,j}^*, 1 \leq i \leq p, 1 \leq j \leq m)$ échantillon de validation
Entrée : La variable à expliquer Y
Entrée : Les variables explicatives $X_j, j \in \{1, \ldots, m\}$, m nombre de variables explicatives
tant que $E_k = min(E)$ **faire**
  pour *j=1 à m* **faire**
    — *Sélection de* $X_{(1,\ldots,m)\setminus j}$ *en variables de régression*
    — *Sélection d'une variable de classification parmi les variables non utilisées dans la régression (toutes les combinaisons sont effectuées).*
    — *Construire le modèle MOB avec les variables de régression et de classification.*
    — *Construire* $\hat{Y}_j^*$ *à partir du modèle MOB et des variables explicatives de l'échantillon de validation.*
    — $\text{MSE}_\text{j} = \frac{1}{n} \sum_{i=1}^{p} \left(\hat{Y}_{i,j}^* - Y_{i,j}^* \right)^2$
  fin
  — $E_k = \min(MSE)$
  — $X'' = X_{argmin(MSE)}$
  — $X = X \setminus X''$
  — $E = (E; MSE)$
  — $m = m - 1$
fin

A cette étape, X contient les variables de régression, X" est la variable de classification et X' privé de X'' contient les variables non utilisées.

Afin d'optimiser au mieux, nous avons ajouté les variables non utilisées une à une à la variable de classification. Si l'erreur baisse nous conservons ce modèle et nous continuons à ajouter les variables non utilisées une à une... Si la mesure d'erreur ne baisse pas lors de l'ajout des variables une à une, nous nous arrêtons et conservons le modèle minimisant l'erreur parmi tous les modèles.

Troisième approche
La troisième approche utilisée est en théorie l'approche la plus optimale. Cette approche consiste à utiliser toutes les combinaisons possibles et de sélectionner le modèle minimisant l'erreur. Par exemple : si on a 3 variables explicatives qu'on nommera A, B et C. Nous avons 6 modèles. Un modèle en prenant A en variable de régression (respectivement classification) et B, C en variable de classification (respectivement régression). Un modèle en prenant B en variable de régression (respectivement classification) et A, C en variable de classification (respectivement régression). Un modèle en prenant C en variable de régression (respectivement classification) et A, B en variable de classification (respectivement régression). Ce qui fait 6 modèles. En pratique, cette approche a été mis en oeuvre mais pour peu de variables car elle est inutilisable pour plus de 10 variables (nous avons 1022 modèles avec 10 variables).

Application sur un jeu de données fictif

Dans cette partie, nous avons modifié notre jeu de données. Les modifications sont les suivantes :
- Le cout moyen d'un sinistre pour les personnes âgées de moins de 40 ans strictement est augmenté de 4 000.
- Le coût moyen d'un sinistre pour les personnes âgées de moins de 40 ans strictement et possédant une voiture avec 7 chevaux fiscaux est augmenté de 7 000.

Avec ces hypothèses fortes, nous allons observer si les modèles *MOB* arrivent à capter ces informations. Par la suite, une comparaison des modèles coût GLM, *CART*, *XGBoost* et *MOB* sur ce jeu de données est effectuée.

Résultats sur le jeu de données fictif
Nous optimisons analogiquement aux parties précédentes pour tous les modèles sauf le modèle *GLM* et *MOB*. Pour tous les modèles effectués sur l'échantillon d'apprentissage, nous utilisons toutes les variables explicatives (hormis l'exposition car notre étude est sur les modèles coûts). Pour le modèle *MOB*, nous utilisons toutes les variables hormis les variables *age_conducteur* et *auto_vehicule_cv* étant des variables de classification. Par conséquent, aucune optimisation n'est effectuée sur le modèle *MOB* car généralement les ordres de grandeur des erreurs sont les mêmes au vu de l'homogénéité des variables de régression utilisées.

Modèle	MSE	RMSE	MAE
GLM	2 913 194	1 707	1 247
CART	10 496 826	3 240	2 401
XGBoost	2 850 562	1 688	1 237
MOB	**2 223 220**	**1 491**	**925**

TABLE 4.14 – MSE, RMSE et MAE des différents modèles coûts sur l'échantillon test

Sur le tableau 4.14, nous avons les différentes erreurs sur l'échantillon test fictif. Nous observons que les modèles *MOB* sont plus performants que les autres modèles. Les modèles *MOB* détectent l'hétérogénéité du jeu de données fictif.

Le modèle *MOB* est constitué de 6 feuilles :

- La première feuille contient 1 500 observations. Ces observations ont un âge inférieur à 40 strictement et un véhicule avec un nombre de chevaux fiscaux égale à "4 et moins", "10-11" et "12-15 et plus".
- La deuxième feuille contient 4 281 observations. Ces observations ont un âge inférieur à 40 strictement et un véhicule avec un nombre de chevaux fiscaux égale à 5, 6 et "13-14-8-9".
- La troisième feuille contient 1 036 observations. Ces observations ont un âge inférieur à 40 strictement et un véhicule avec un nombre de chevaux fiscaux égale à 7.
- La quatrième feuille contient 2 885 observations. Ces observations ont un âge compris entre 40 et 47,40.
- La cinquième feuille contient 6 128 observations. Ces observations ont un âge compris entre 47,41 et 71,01.
- La sixième feuille contient 803 observations. Ces observations ont un âge compris supérieur à 71,02.

Au début de la partie 4.3.1, nous avions décidé d'augmenter le coût moyen des individus de moins de 40 ans possédant un véhicule avec un nombre de chevaux fiscaux différent de 7 de 3 000, ces individus ont été regroupés par le modèle *MOB* dans les feuilles 1 et 2. De même pour l'augmentation de 7 000 du coût moyen pour les individus de moins de 40 ans possédant un véhicule avec un nombre de chevaux égal à 7 regroupés dans la feuille 3. Par conséquent, les différentes feuilles sont en adéquation avec les modifications apportées au jeu de données.

Erreur locale sur le jeu de données fictif

Dans l'optique de comparer l'efficacité des modèles *MOB* sur des jeux de données hétérogènes. Nous allons comparer les erreurs pour chaque feuille. Pour ce faire, nous considérons un modèle GLM global et un modèle GLM local pour chaque feuille (équivalent au modèle *MOB*). La performance de chaque feuille est mesurée sur l'échantillon test.

Erreur	**Feuille 1**	**Feuille 2**	**Feuille 3**	**Feuille 4**	**Feuille 5**	**Feuille 6**
MSE	15 759 158	15 103 935	14 573 162	3 760 664	1 935 635	1 700 260
RMSE	3 970	3 886	3 817	1 939	1 391	1 304
MAE	3 688	3 594	3 472	1 645	969	900

TABLE 4.15 – MSE, RMSE et MAE des différentes feuilles sur l'échantillon test avec un modèle *GLM* global coût

Erreur	**Feuille 1**	**Feuille 2**	**Feuille 3**	**Feuille 4**	**Feuille 5**	**Feuille 6**
MSE	2 139 856	1 815 996	2 416 695	1 925 158	1 439 652	1 208 116
RMSE	1 463	1 348	1 555	1 388	1 200	1 099
MAE	928	822	820	809	715	662

TABLE 4.16 – MSE, RMSE et MAE des différentes feuilles sur l'échantillon test avec un modèle *MOB*

Nous observons que les différentes métriques par feuille sont nettement meilleures dans le modèle *MOB* (voir tableau 4.16 que dans le modèle *GLM* global (voir tableau 4.15). Par exemple, pour la feuille 2, le MSE baisse de 86 %, le RMSE baisse de 65 % et le MAE de 26 %. Cet exemple nous a permi de montrer que les modèles *MOB* peuvent capter l'hétérogéinité d'un portefeuille.

4.3.2 Application

Nous estimons les modèles *MOB* avec le package `partykit`[21] et la fonction `glmtree`.

Modélisation de la fréquence

L'objectif de cette partie est d'obtenir le meilleur modèle *MOB* fréquence en utilisant les 3 approches d'optimisation que nous avons mis en place dans la partie 4.3.1. Le modèle optimal est celui dont le RMSE est minimisé.

Approche 1
Cette approche a été explicitée dans la partie 4.3.1. Nous allons commencer par quantifier l'importance des variables dans le modèle CART fréquence. Ces données sont présents dans la sortie du modèle *CART*.

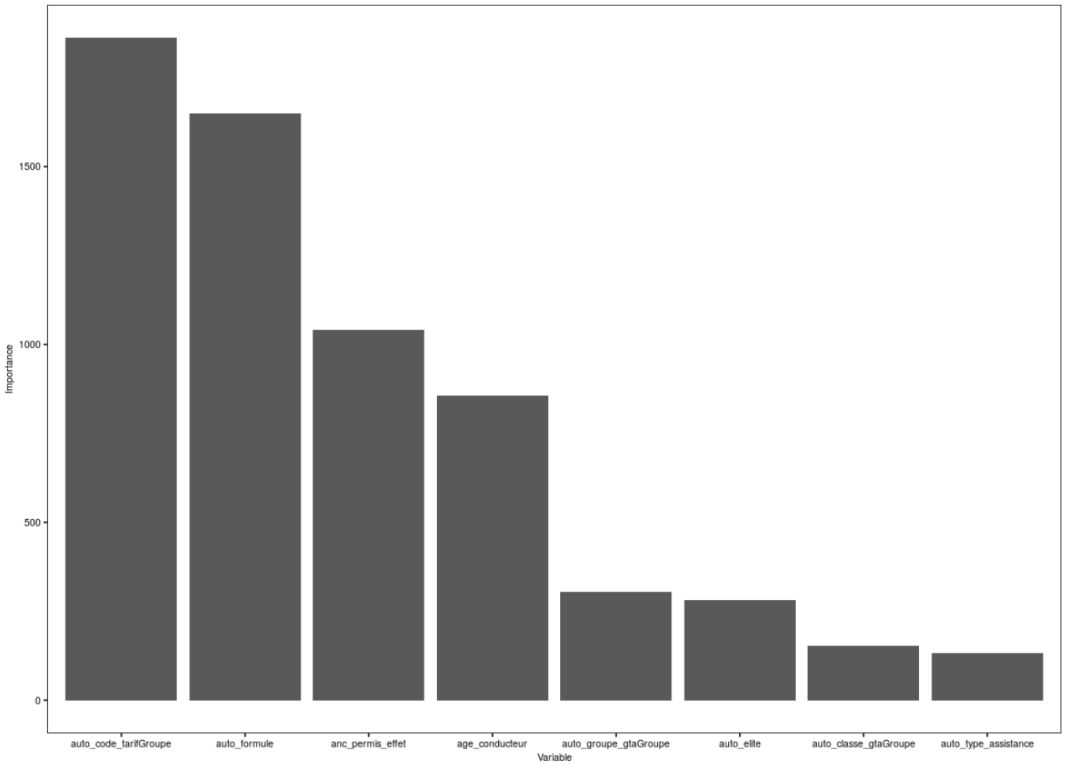

FIGURE 4.9 – Importance des variables dans le modèle *CART* fréquence

Au vu du graphique 4.9, nous choisissons comme variables de classification pour le modèle *MOB* le code de tarif, la formule auto et l'âge du conducteur. Les variables de régression choisies sont celles données par l'approche *forward selection* du modèle de régression linéaire vu en partie 3.1 :

- L'ancienneté de permis.
- Le groupe SRA.
- L'auto elite.
- La cadence de réglement.
- Le département du garage de l'assuré.
- Le code de profession de l'assuré.

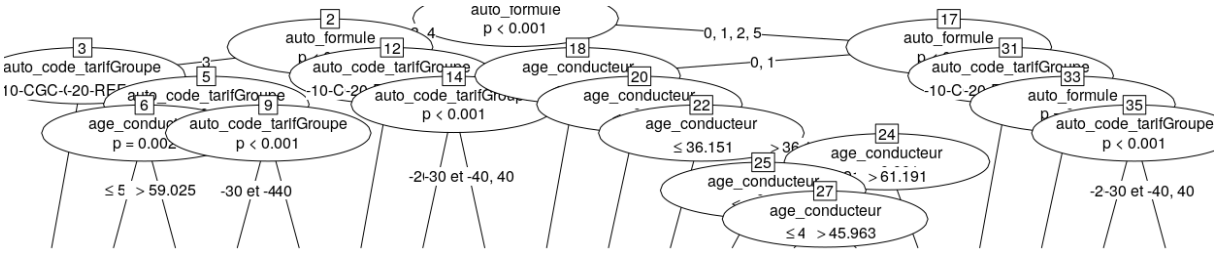

FIGURE 4.10 – Arbre *MOB* fréquence de l'approche 1

Sur le graphique 4.10, nous avons représenté l'arbre *MOB* découlant du choix de variables de régression et classification par la première approche. Ce modèle a 19 feuilles terminales. Chaque feuille correspond à une classe d'individu et à un modèle GLM attitré.

MSE	RMSE	MAE
0,579	0,761	0,536

TABLE 4.17 – MSE, RMSE et MAE sur l'échantillon test fréquence avec l'approche 1

Les métriques présentées dans le tableau 4.17 sont comparées aux métriques des autres approches d'optimisation *MOB* dans la partie 4.19.

Approche 2

Pour la deuxième approche expliquée dans la partie 4.3.1, on obtient :

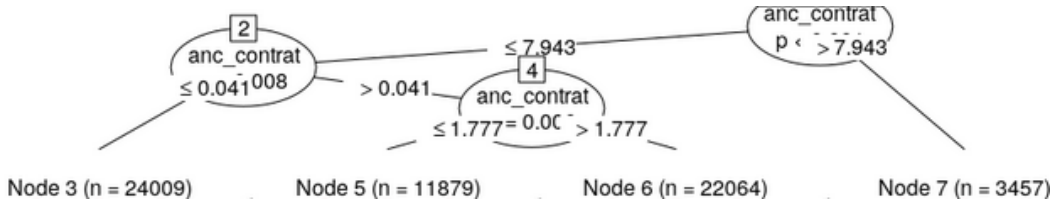

FIGURE 4.11 – Arbre *MOB* fréquence de l'approche 2

Sur le modèle *MOB* représenté par le graphique 4.11, notre algorithme a choisi de retenir la combinaison contenant la variable de classification ancienneté de contrat et les variables de régression sont :

- La marque du véhicule.
- Le nombre de chevaux fiscaux du véhicule.
- Le groupe SRA du véhicule.
- La formule auto.
- La région du garage de l'assuré.
- Le code tarif.
- La situation familiale du conducteur.
- Le type d'assistance dépannage.
- Le type de toit panoramique du véhicule.
- L'auto elite.
- La cadence de réglement de l'assuré.
- L'ancienneté de permis de l'assuré.

Ce modèle a seulement 4 feuilles terminales et donc 4 modèles de régression linéaire distinct pour chaque groupe d'individu.

MSE	RMSE	MAE
0,478	0,691	0,441

TABLE 4.18 – MSE, RMSE et MAE sur l'échantillon test fréquence avec l'approche 2

Les erreurs présentes sur le tableau 4.18 sont comparées aux autres erreurs des approches du modèle *MOB* de la partie 4.19.

Approche 3

La dernière approche d'optimisation est explicitée dans la partie 4.3.1. Nous choisissons d'effectuer toutes les combinaisons possibles en variables de classification et de régression choisies par l'approche *forward selection* du modèle de régression linéaire vu en partie 3.1 et en ajoutant la variable *auto_type_assistance* car nous pouvions encore choisir une variable. Plus le nombre de combinaisons augmente, plus nous avons de chance de tomber sur la combinaison optimale.

FIGURE 4.12 – Arbre *MOB* fréquence de l'approche 3

L'arbre représenté par le graphique 4.12 possède 9 feuilles terminales. La combinaison optimale choisie par la troisième approche a pour variables de classification : l'auto elite et pour variables de régression :

- Le groupe SRA du véhicule.
- Le code tarif.
- Le type d'assistance dépannage.
- L'ancienneté de permis de l'assuré.
- La cadence de réglement de l'assuré.
- Le département du garage de l'assuré.
- Le code de profession de l'assuré.

MSE	RMSE	MAE
0,491	0,700	0,454

TABLE 4.19 – MSE, RMSE et MAE sur l'échantillon test fréquence avec l'approche 3

Les erreurs du tableau 4.20 sont comparées avec celles des autres modèles *MOB* dans la partie 4.19.

Comparaison des différentes approches

Après avoir effectué les 3 approches d'optimisation des modèles *MOB*, nous allons expliciter et comparer les différentes métriques.

Approche	MSE	RMSE	MAE
Approche 1	0,487	0,698	0,446
Approche 2	**0,478**	**0,691**	**0,441**
Approche 3	0,491	0,700	0,454

TABLE 4.20 – MSE, RMSE et MAE sur l'échantillon test fréquence

Sur le tableau 4.20, la deuxième approche est choisie car elle minimise le RMSE. Nous observons que l'algorithme de l'approche 2 permet de trouver un équilibre entre la finesse de découpage et la robustesse

des estimations tout en testant toutes les variables. La première approche n'est pas controlée par le RMSE et effectue trop de segmentations rendant inefficace la loi forte des grands nombres et le théorème central limite. La troisième approche limite le nombre de variables et la sélection des variables reste délicate. Les segmentations effectuées par l'approche 2 sont :

- La première feuille contient 24 009 observations. Ces données ont une ancienneté de contrat inférieure ou égale à 0,041.
- La seconde feuille contient 11 879 observations. Ces données ont une ancienneté de contrat comprise entre 0,041 et 1,777 avec 1,777 inclus.
- La troisième feuille contient 22 064 observations. Ces données ont une ancienneté de contrat comprise entre 1,777 et avec 7,943 inclus.
- La dernière feuille contient 3 457 observations. Ces données ont une ancienneté de contrat strictement supérieur à 7,943.

Notre arbre a bien été constitué avec toutes les données de l'échantillon d'apprentissage fréquence car $24\,009 + 11\,879 + 22\,064 + 3\,457 = 61\,409$ correspondant à la taille de cet échantillon. Nous constatons que la segmentation par ancienneté de contrat A.6 n'est pas très intuitive car les fréquences annuelles de sinistres par ancienneté de contrat ne sont pas homogènes. Par exemple, intuitivement nous aurions regroupé les personnes avec une ancienneté de contrat égale à 8 avec ceux de la troisième feuille. De plus, nous précisons les différents coefficients des modèles GLM de chaque feuille en annexe A.4 car nous avons 4 feuilles et beaucoup de modalités. Pour la première feuille, les variables significatives du modèle GLM fréquence sont (nous rappellons qu'une variable est significative si $\Pr(>|t|)<0{,}05$) :

- La marque du véhicule.
- Le nombre de chevaux fiscaux du véhicule.
- Le groupe SRA du véhicule.
- La région du garage de l'assuré.
- Le code tarif.
- Le type d'assistance dépannage.
- L'auto elite.
- Le type de toit panoramique du véhicule.
- La cadence de réglement de l'assuré.
- L'ancienneté de permis de l'assuré.

Les variables significatives de la deuxième feuille du modèle GLM fréquence sont :

- La marque du véhicule.
- Le nombre de chevaux fiscaux du véhicule.
- Le groupe SRA du véhicule.
- La région du garage de l'assuré.
- Le code tarif.
- Le type d'assistance dépannage.
- L'auto elite.
- La cadence de réglement de l'assuré.
- L'ancienneté de permis de l'assuré.

Les variables significatives de la troisième feuille du modèle GLM fréquence sont :

- Le nombre de chevaux fiscaux du véhicule.
- Le groupe SRA du véhicule.
- La région du garage de l'assuré.
- Le code tarif.
- Le type d'assistance dépannage.
- L'auto elite.
- Le type de toit panoramique du véhicule.
- La cadence de réglement de l'assuré.

— L'ancienneté de permis de l'assuré.

Les variables significatives de la quatrième feuille du modèle GLM fréquence sont :

— La marque du véhicule.
— La formule auto.
— Le code tarif.
— La situation familiale de l'assuré.
— L'auto elite.
— L'ancienneté de permis de l'assuré.

Les variables de régression ont un comportement différents en fonction des modèles GLM fréquence associés à chaque feuille. L'exemple le plus flagrant est la significativité du modèle GLM de la dernière feuille. Nous observons que la formule auto est une variable significative seulement pour la dernière feuille ou encore que le groupe SRA du véhicule et la région de l'assuré sont des variables significatives pour toutes les modèles GLM sauf celui de la dernière feuille. Ces éléments peuvent justifier le fait de créer plusieurs modèles GLM sur des segments différents. Nous reviendrons sur les impacts des modalités sur la fréquence prédite pour chaque feuille dans le chapitre 5 portant sur l'interprétabilité des modèles.

Modélisation de la sévérité

Nous procédons de manière analogue à la partie 4.3.2.

Approche 1

Cette fois, nous quantifions l'importance des variables dans le modèle CART coût.

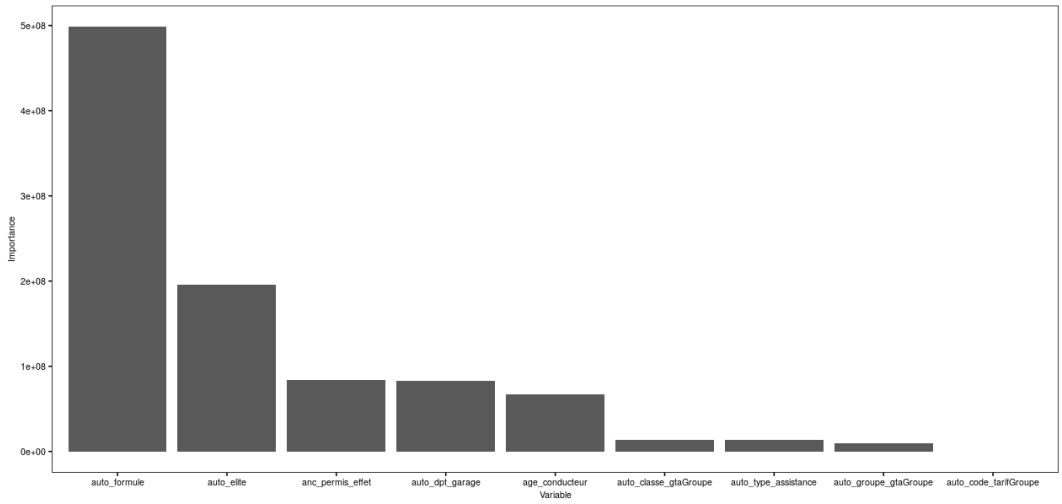

FIGURE 4.13 – Importance des variables dans le modèle *CART*

Au vu du graphique 4.13, nous choisissons comme variables de classification pour le modèle *MOB* la formule auto, l'auto elite et l'ancienneté de permis. Les variables de régression choisies sont celles données par l'approche *forward selection* du modèle GLM vu en partie 3.10 :

— Le groupe SRA du véhicule.
— Le département du garage de l'assuré.
— Le type de toit panoramique du véhicule.
— Le nombre de chevaux fiscaux du véhicule.
— La région du garage de l'assuré.
— La carrosserie du véhicule.
— L'ancienneté de contrat de l'assuré.

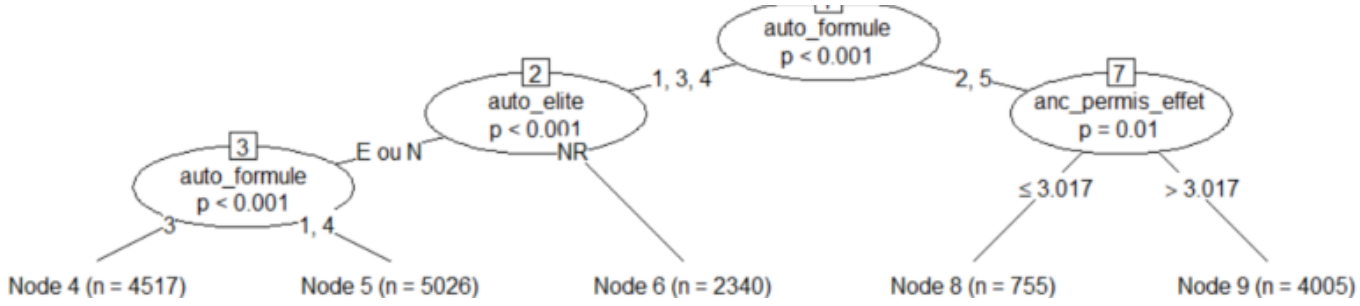

FIGURE 4.14 – Arbre *MOB* coût de l'approche 1

Sur le graphique 4.14, nous avons représenté l'arbre *MOB* avec les variables optimales de régression et classification selon la première approche. Ce modèle a 13 feuilles terminales.

MSE	RMSE	MAE
1 768 404	1 330	780

TABLE 4.21 – MSE, RMSE et MAE sur l'échantillon test coût avec l'approche 1

Les métriques présentées dans le tableau 4.21 sont comparées aux métriques des autres approches d'optimisation *MOB* coût dans la partie 4.23.

Approche 2

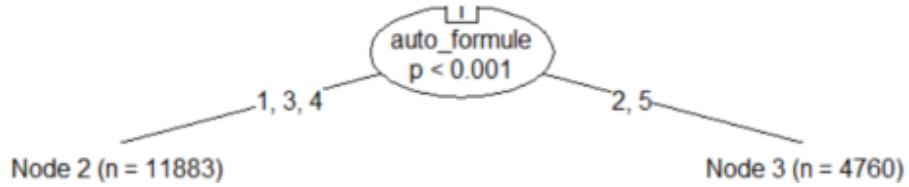

FIGURE 4.15 – Arbre *MOB* coût de l'approche 2

Sur le modèle *MOB* représenté par le graphique 4.16, notre algorithme a choisi de retenir la combinaison contenant la variable de classification formule auto et les variables de régression :

- Le groupe SRA du véhicule.
- La région de l'assuré.
- La situation familiale de l'assuré.
- L'auto elite.
- L'ancienneté de permis de l'assuré.
- L'âge du conducteur de l'assuré.
- L'ancienneté du contrat de l'assuré.

MSE	RMSE	MAE
1 754 889	1 324,72	778

TABLE 4.22 – MSE, RMSE et MAE sur l'échantillon test coût avec l'approche 2

Les erreurs présentes sur le tableau 4.22 sont comparées aux autres erreurs des approches du modèle *MOB* de la partie 4.23.

Approche 3

Nous choisissons d'effectuer toutes les combinaisons possibles en variables de classification et de régression choisies par l'approche *forward selection* du modèle de régression linéaire vu en partie 3.1. Nous rappelons que plus le nombre de combinaisons de variables augmente, plus nous avons de chance de tomber sur la combinaison optimale.

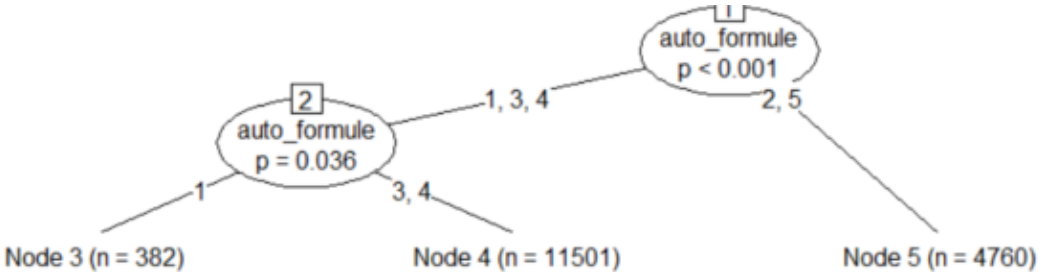

FIGURE 4.16 – Arbre *MOB* coût de l'approche 3

L'arbre représenté par le graphique 4.16 possède 2 feuilles terminales. La combinaison optimale choisie par la troisième approche a pour variable de classification la formule auto et pour variables de régression :

- L'auto elite.
- Le groupe SRA du véhicule.
- Le département de l'assuré.
- Le nombre de chevaux fiscaux du véhicule.
- L'ancienneté de contrat.

MSE	RMSE	MAE
1 755 745	1 325,05	779

TABLE 4.23 – MSE, RMSE et MAE sur l'échantillon test coût avec l'approche 3

Les erreurs du tableau 4.23 sont également comparées avec celles des autres modèles *MOB* dans la prochaine partie 4.23

Comparaison des différentes approches

Après avoir effectué les 3 approches d'optimisation des modèles *MOB*, nous allons expliciter et comparer les différentes métriques associées à chaque approche.

Approche	MSE	RMSE	MAE
Approche 1	1 768 404	1 330	780
Approche 2	**1 754 889**	**1 324,72**	**778**
Approche 3	1 755 745	1 325,05	779

TABLE 4.24 – MSE, RMSE et MAE sur l'échantillon test coût

Sur le tableau 4.24 représentant les erreurs en fonction des différentes approches *MOB*, nous observons que la deuxième approche minimise le RMSE. Cette fois, l'approche segmentant le moins est la plus optimale. Dans ce modèle, le segmentation se fait ainsi :

- La première feuille contient 11 883 observations. Ces données ont une formule auto égale à 1, 3 ou 4.
- La seconde feuille contient 4 790 observations. Ces données ont une formule auto égale à 2 ou 5.

Nous précisons encore une fois que notre arbre contient toutes les observations de l'échantillon d'apprentissage coût car 11 883 + 4 760 = 16 643 correspondant à la taille de cet échantillon. En comparant avec les statistiques descriptives de la variable *auto_formule* du graphique 2.14, la segmentation est justifiée car un comportement similaire en coût moyen entre les modalités de feuilles identiques est identifié (les sinistres extrêmes ne changent pas l'allure des coûts moyens car les expositions de ces différentes modalités sont non négligeables).
Nous présentons les coefficients GLM des deux feuilles terminales.

	Estimate	Std, Error	t value	Pr(>|t|)
auto_groupe_gtaGroupe10	-0,48628	0,36702	-1,32496	0,18521
auto_groupe_gtaGroupe11	-0,39705	0,36650	-1,08335	0,27867
auto_groupe_gtaGroupe12	-0,36099	0,36658	-0,98476	0,32476
auto_groupe_gtaGroupe13	-0,28531	0,36655	-0,77839	0,43635
auto_groupe_gtaGroupe14	-0,32587	0,36645	-0,88925	0,37388
auto_groupe_gtaGroupe15	-0,32648	0,36740	-0,88864	0,37421
auto_groupe_gtaGroupe16	-0,16989	0,36771	-0,46202	0,64408
auto_groupe_gtaGroupe3-6-8	-0,25441	0,40247	-0,63212	0,52732
auto_groupe_gtaGroupe9	-0,50065	0,37010	-1,35273	0,17617
auto_zoneAUTRES	0,08234	0,03593	2,29139	0,02196
auto_conducteur_situfamilD	0,00588	0,06323	0,09301	0,92590
auto_conducteur_situfamilM	-0,02312	0,03701	-0,62459	0,53225
auto_conducteur_situfamilU	-0,01227	0,03848	-0,31890	0,74981
auto_conducteur_situfamilV	0,09391	0,08335	1,12664	0,25992
auto_eliteNR	0,24084	0,03295	7,30860	0,00000
anc_permis_effet	-0,00662	0,00271	-2,44163	0,01464
age_conducteur	0,00365	0,00256	1,42328	0,15468
anc_contrat	-0,01339	0,00456	-2,94019	0,00329

TABLE 4.25 – Sortie R du modèle GLM coût sur la première feuille

Le tableau 4.25 présente les différents coefficients GLM de la première feuille. Nous observons sur le tableau 4.25 que le modèle GLM de la première feuille a pour variables significatives :

- La région de l'assuré.
- L'auto elite.
- L'ancienneté de permis.

	Estimate	Std, Error	t value	Pr(>|t|)
(Intercept)	6.29345	0.28227	22.29593	0.00000
auto_groupe_gtaGroupe10	0.04964	0.27313	0.18173	0.85580
auto_groupe_gtaGroupe11	0.12133	0.27300	0.44442	0.65676
auto_groupe_gtaGroupe12	0.23107	0.27387	0.84373	0.39887
auto_groupe_gtaGroupe13	0.17404	0.27363	0.63605	0.52478
auto_groupe_gtaGroupe14	0.23306	0.27557	0.84574	0.39774
auto_groupe_gtaGroupe15	0.33772	0.28195	1.19778	0.23106
auto_groupe_gtaGroupe16	0.67640	0.28232	2.39586	0.01662
auto_groupe_gtaGroupe3-6-8	-0.09282	0.28736	-0.32303	0.74669
auto_groupe_gtaGroupe9	0.02423	0.27622	0.08774	0.93009
auto_zoneAUTRES	0.21171	0.05614	3.77075	0.00016
auto_conducteur_situfamilD	-0.03689	0.08708	-0.42363	0.67185
auto_conducteur_situfamilM	-0.12795	0.05196	-2.46263	0.01383
auto_conducteur_situfamilU	-0.14465	0.04920	-2.94000	0.00330
auto_conducteur_situfamilV	-0.18196	0.13910	-1.30813	0.19089
auto_eliteNR	0.16150	0.04088	3.95086	0.00008
anc_permis_effet	-0.00414	0.00355	-1.16685	0.24333
age_conducteur	0.00495	0.00335	1.47780	0.13953
anc_contrat	-0.00117	0.00778	-0.15052	0.88036

TABLE 4.26 – Sortie R du modèle GLM coût sur la deuxième feuille

Sur le tableau 4.26, nous avons le modèle GLM de la deuxième feuille. Ses variables significatives sont :

- Le groupe SRA du véhicule.
- La région de l'assuré.
- L'auto elite.
- La situation familiale de l'assuré.

Les variables significatives ne sont pas les mêmes entre les deux modèles GLM. Par conséquent, la segmentation est justifiée car ces deux modèles GLM ont des comportements différents. Comme pour le modèle *MOB* fréquence, l'impact des modalités en coût moyen par feuille est présenté en dernière partie du mémoire 5 portant sur l'interprétabilité des modèles.

Conclusion sur les différentes approches d'optimisation

Au vu de notre étude, nous constatons que la deuxième approche donne de meilleurs performances que les autres approches. Ces résultats peuvent s'expliquer par le fait que la deuxième approche balaye l'ensemble de nos variables, le problème est que nous ne testons pas toutes les combinaions et qu'il peut exister une autre combinaison optimale. Par exemple, si le fait de supprimer une variable augmente le RMSE mais qu'une autre suppression baisse de manière conséquente le RMSE, l'algorithme ne captera pas cette information. Par ailleurs, la faiblesse de la première approche est que la segmentation à partir des variables les plus importantes selon le modèle *CART* nous enlève la possibilité de les utiliser en variables de régression. Ces variables peuvent être très importantes dans la régression et ne seront donc pas utilisées. La faiblesse de la troisième approche provient du temps de calcul. À partir de 11 variables, le temps de calcul est important (nous avons 2 046 combinaisons pour 11 variables, 4 094 combinaisons pour 12 variables, ...). Le choix des variables optimales utilisées dans la troisième approche reste à discuter. Le temps de calcul pour les modèles *MOB* dépend du nombre d'observations et du nombre de modalités pour une variable donnée. Nous précisons que le temps de calcul sans regroupement de modalité est mutiplié par 2 (les regroupements de modalité ont été effectués dans la partie 2.2).

4.4 Performance des différents modèles

Au cours de notre étude, nous avons étudié les modèles *GLM, CART, XGBoost* et *MOB*. Nous résumons les différentes métriques associées à chacun de ces modèles en fréquence et en coût.

Modèle fréquence

Modèle	MSE	RMSE	MAE	Temps de calcul
GLM1	0,489	0,700	0,453	28s
GLM2	0,486	0,697	0,452	Instantané
CART	0,600	0,774	0,490	Instantané
XGBoost	0,603	0,777	0,551	1h 43mn
***MOB* approche 2**	**0,478**	**0,691**	**0,441**	**7h 29mn**

TABLE 4.27 – MSE, RMSE et MAE des différents modèles sur l'échantillon test

Sur le tableau 4.27, nous avons les différentes erreurs associées à chaque modèle que nous avons étudié précedemment. Le modèle GLM2 est modèle GLM avec les mêmes variables que le modèle *MOB* (y compris les variables de classifications). Ce modèle nous permet de savoir si la meilleure performance du modèle *MOB* était dûe aux variables de régression ou à la segmentation. Nous observons que les modèles *CART* et *XGBoost* sont moins performants que le modèle GLM sur notre jeu de données. La mauvaise performance du modèle *XGBoost* peut s'expliquer par l'échantillonnage établi en amont, le faible nombre de variables et la combinaison de paramètres choisie n'étant pas forcément optimal pour cet échantillon. Par ailleurs, le modèle *MOB* est le meilleur modèle fréquence sur notre jeu de données.

Modèle coût

Modèle	MSE	RMSE	MAE	Temps de calcul
GLM1	1 760 525	1 327	779	35s
GLM2	1 758 095	1 326	779	Instantané
CART	1 766 018	1 329	781	Instantané
XGBoost	1 756 596	1 325	775	38mn
***MOB* approche 2**	**1 754 889**	**1 325**	**778**	**23h 12mn**

TABLE 4.28 – MSE, RMSE et MAE des différents modèles sur l'échantillon test

Sur le tableau 4.28, nous avons les différentes erreurs associées à nos modèles coût. Le modèle *MOB* obtient encore une fois la meilleure performance en terme de RMSE. Le temps de calcul très important du modèle *MOB* est dû à notre algorithme. En effet, un modèle *MOB* peut prendre 30 minutes à être modélisé (voir plus en fonction du jeu de données) et notre algorithme teste énormément de modèles. Par conséquent, nous suggerons d'ajouter un argument à notre algorithme afin de s'arrêter au bout d'un certain temps et de prendre le meilleur modèle parmi toutes les combinaisons testées. Le modèle *XGBoost* a la meilleure performance en MAE pour le modèle coût.

Conclusion

Les modèles *MOB* améliorent la prédiction de la fréquence de sinistres et du coût moyen d'un sinistre. Les limites des modèles *MOB* sont d'une part, que la performance sur des jeux de données homogènes reste très proche de celle des modèles GLM classiques même si nous pensons que les modèles *MOB* vont régulièrement avoir de meilleurs performances et d'autre part les temps de calcul élevés de nos algorithmes. Explicitement, les différentes modalités des variables ont une fréquence de sinistres proche et un coût moyen proche comme nous pouvons le voir lors de l'analyse statistiques des variables 2.2

(attention, les analyses sur le coût moyen comprennent les sinistres extrêmes, les coûts sont encore plus homogènes sur la base sinistre attritionnelle). Si le jeu de données est très hétérogène, le modèle *MOB* donne de très bonne performance par rapport aux autres modèles étudiés car ce modèle arrive à capter la variable responsable de cette hétérogénéité comme nous le montre l'étude faite dans la partie 4.3.1. Un test des modèles *MOB* sur un jeu de données avec plus d'observations peut s'avérer intéressant car les modèles *MOB* pourraient peut être créer plusieurs groupes respectant les conditions requises pour appliquer les différentes lois statistiques. La limite est le temps de calcul qui pourrait être trop grand. Encore une fois, la problèmatique de la segmentation est un thème central de l'étude.

Chapitre 5

Interprétabilité des différents modèles

Dans cette partie, l'objectif est de pouvoir interpréter pour chaque modèle le comportement de nos variables réponses en fonction des variables explicatives. Nous rappelons que dans notre cas, nous avons modélisé la fréquence et le coût afin d'obtenir la prime pure. La prime pure s'obtient en multipliant la fréquence prédite par le coût prédit. L'explication des variables réponses (fréquence et coût) est essentielle dans la compréhension du tarif. Par exemple, si nous ne comprenons pas la raison de l'augmentation forte d'un tarif, la gestion du risque sera défectueuse car nous n'avons pas la cause de ce phénomène. Dans ce but, nous allons étudier l'impact des variables explicatives sur le tarif pour nos modèles GLM, *CART*, *XGBoost* et *MOB*.

5.1 Concept d'interprétabilité

Les modèles GLM, *CART* et *MOB* offrent une interprétation claire de l'acheminemant ayant mené à une prédiction. Le modèle GLM permet d'expliquer les prédictions avec ses coefficients, le modèle *CART* explique les prédictions avec l'arbre et ses différentes segmentations et le modèle *MOB* explique les prédictions avec les segmentations de l'arbre et les coefficients GLM associés à chaque feuille terminale. Cependant, certains modèles de machine learning tel que le modèle *XGBoost* n'a aucune sortie permettant d'expliquer les prédictions. Les décisions prisent par ces modèles complexes sont difficilement interprétables d'où leur nom de modèles *blackbox*. Nous rappelons que ces modèles ont montré une performance supérieure aux autres modèles sur certaines applications. Par exemple, nous avons vu en introduction 1.4 que le modèle *GBM* (faisant parti de la famille des modèles *machine learning* non interprétable directement) était plus performant que le modèle GLM. Ainsi, un équilibre entre gain de performance et interprétabilité doit être établi. Par conséquent, lors de la conception d'un modèle d'apprentissage statistique, nous devons nous demander si l'on souhaite une meilleure performance ou une meilleure compréhension des prédictions. Certains scientifiques préconisent la performance au détriment de l'interprétabilité. Par exemple, Kuhn et Jonhson dans leur ouvrage [23] affirment que "tant que les modèles complexes sont correctements validés, l'utilisation d'un modèle construit pour l'interprétation plutôt que pour la performance prédictive peut s'avérer inappropriée". Cette phrase s'appuie sur une étude concernant la détection de spams dans les mails et l'évaluation du prix de maisons démontrant la non nécessité de l'interprétabilité des prédictions. Cependant, la plupart des études nécessitent à la fois une bonne performance et une interprétabilité des prédictions. Par exemple, si un diagnostic médical est établi mais que nous n'avons pas l'explication de ce diagnostique, cette étude aura été inutile car nous n'avons pas les causes de ce diagnostique. En résumé, l'interprétabilité d'un modèle permet de comprendre les modèles, prendre des décisions envers les mauvais risques, respecter la réglementation mais aussi réduire les biais éthiques et moraux [11].
Dans l'actuariat, le gain d'efficacité obtenu par les modèles de *machine learning* rendent ces modèles indispensables pour les assureurs car une forte concurrence existe entre ces différents acteurs. Face à

cette forte concurrence, un assureur ne suivant pas la concurrence s'expose au risque de récupérer les mauvais risques. Ces différents enjeux expliquent l'importance pour les actuaires de pouvoir expliquer les modèles de *machine learning*. De plus, un problème subiste, aucune définition mathématique rigoureuse n'existe concernant l'interprétabilité des modèles de *machine learning*. De nombreuses méthodes permettent d'expliquer un modèle mais aucune préconisation n'est donnée sur la méthode d'interprétabilité adéquate à un modèle donné. Une définition de l'interprétabilité est l'explication des différents effets en fonction des variables explicatives. Concrètement, il s'agit de la connaissance des modalités ou valeurs des différentes variables favorisant ou défavorisant une valeur prédite. Nous précisons qu'une interprétabilité globale est l'explication des effets globaux des variables sur la prédiction permettant d'avoir une compréhension globale du modèle et une interprétabilité locale (appelé aussi explicabilité) est la contribution de chaque modalité sur une prédiction. Par exemple, pour les modèles GLM, d'une part l'interprétabilité globale est donnée par les coefficients GLM. Un coefficient haut explique que la modalité en question augmente énormément la prédiction. D'autre part, l'interprétabilité locale est aussi donnée par les coefficients GLM permettant d'expliquer et quantifier le passage d'une modalité à une autre. Afin d'expliquer les modèles de *machine learning*, des méthodes d'interprétabilité à postériori sont utilisées c'est à dire des méthodes d'interprétabilité indépendantes de la modélisation du modèle et s'effectuant sur le modèle. Ces méthodes donnent une interprétabilité locale entrainant par la suite une interprétabilité globale.
Dans la suite de ce chapitre, nous commençons par expliquer les modèles GLM à partir des coefficients. Dans un deuxième temps, nous expliquons les modèles CART par l'arbre de decisions. Dans un troisième temps, nous introduisons l'algorithme SHAP[26] et l'utilisons sur le modèle *XGBoost* afin de l'expliquer. Pour finir, nous interprétons les différents effets des modèles *MOB* à l'aide des arbres des décisions et des coefficients GLM associés à chaque feuille terminale.

5.2 Modèle linéaire généralisé

Les modèles GLM possèdent une réelle force d'interprétabilité. Dans ces modèles, les coefficients permettent de quantifier l'effet d'une variable. Dans ces modèles, la prime pure se calcule de cette manière :

$$Primepure = \mathbb{E}(Y_{freq}) \times \mathbb{E}(Y_{cout}) = \exp(\sum_{i=1}^{p} \beta_i X_i + \beta_0) \times \exp(\sum_{i=1}^{p} \beta_i' X_i + \beta_0')$$

Avec Y_{freq} la variable fréquence à expliquer, Y_{cout} la variable coût à expliquer, $X_i, 1 \leq i \leq p$ les variables explicative, $(\beta_1, ..., \beta_p)$ les coefficients du GLM fréquence, $(\beta_1', ..., \beta_p')$ les coefficients du GLM coût, β_0 l'*intercept* du modèle fréquence et β_0' l'*intercept* du modèle coût.
Nous allons maintenant présenter et interpréter une grille de tarification. Cette grille a une base de variables explicatives communes. Les seules variables explicatives qui diffèrent sont la formule auto, la région de l'assuré et son ancienneté du contrat. L'objectif est d'interpréter les effets de ces 3 dernières sur la prime pure. Nous avons choisi ces variables car ce sont des variables pertinentes pour comprendre l'évolution du prix en fonction des différents modèles (par exemple, pour le modèle *MOB*, la segmentation se fait par l'ancienneté du contrat).

Les modalités des variables explicatives communes sont :

- *auto_vehicule_marque*="CITR-FIAT-FORD-PEUG-RENAULT-AUTRES"
- *auto_groupe_gta*="0-2-4-5-7"
- *auto_classe_gta*="B-C-D-E-F-G"
- *auto_dpt_garage*="AUTRES"
- *auto_conducteur_situfamil*="C"
- *auto_elite*="NR"
- *auto_franchise_bg*="0"
- *auto_toit_panoramique*="N"
- *auto_type_assistance*="0"
- *code_professionGroupe*="2-3-5-6-9"
- *sra_carrosserieGROUPE*="BERLINE-LUDOSPACE"
- *age_conducteur*=60
- *anc_permis_effet*=13
- *auto_vehiculecvGroupe*="4 et moins"
- *auto_code_tarifGroupe*="-20-REF"

auto_formule	anc_contrat 0		anc_contrat 1		anc_contrat 2		anc_contrat 7		anc_contrat 8	
	auto_zone		auto_zone		auto_zone		auto_zone		auto_zone	
	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES
1	138	174	137	173	136	171	130	164	129	163
2	83	105	83	104	82	103	78	99	78	98
3	122	155	121	153	120	152	115	146	114	144
4	112	142	111	140	110	139	105	133	104	132
5	73	92	72	91	72	90	69	87	68	86

TABLE 5.1 – Grille de tarification pour le modèle GLM

La grille 5.1 correspond au tarif en fonction de l'ancienneté de contrat, région de l'assuré et sa formule auto. Par exemple, un individu avec une ancienneté de contrat d'un an avec une formule auto 1 et habitant dans la région associée à 2 aura une prime pure égale à 137.

Modalité ou variable	Coefficient modèle fréquence	Coefficient modèle coût
auto_formule0	Non présent	Non présent
auto_formule1	Non présent	Référence
auto_formule2	Non présent	-0.50402
auto_formule3	Non présent	-0.11925
auto_formule4	Non présent	-0.20818
auto_formule5	Non présent	-0.63788
auto_zone2	Référence	Référence
auto_zoneAUTRES	0,14166	0,09218
anc_contrat	Non présent	-0,00867

TABLE 5.2 – Coefficient des modèles GLM fréquence et coût

Sur le tableau 5.2, nous avons les différents coefficients en sortie de R. Pour les variables quantitatives, ces coefficients correspondent aux β_i. Pour les variables qualitatives, chaque variable est associée à une modalité référence. Les coefficients des modalités non référence correspondent à l'écart par rapport à la modalité référence. Par exemple, grâce au lien *log*, pour passer du prix d'un assuré avec une modalité 0 à 1 pour la variable *auto_formule*, il suffit de multiplier par un coefficient. Afin d'expliquer le calcul de ce coefficient, notons $Primepure1$ le montant payé par un assuré avec une ancienneté de contrat

égale à 0, une localisation dans la région assimilée à 2 et une formule auto égale à 2 et $Primepure2$ le montant payé par un assuré avec les mêmes caractéristiques que précedemment hormis la formule auto égale à 3. À partir de la table 5.2 les prix payés par ces assurés sont respectivement 83 et 122.

$$\begin{aligned} Primepure2 &= Primepure1 \times \exp(0) \times \exp(-0,11925 - (-0,50402)) \\ &= Primepure1 \times \exp(0,38477) \\ &= Primepure1 \times 1,469276 = 83 \times 1,469276 \approx 122. \end{aligned}$$

Le $exp(0)$ provient de la non présence de la formule auto dans les variables explicatives du modèle GLM fréquence. Concernant les variables quantitatives, il suffit de multiplier la valeur de cette variable par le coefficient $\exp(\beta + \beta^{'})$.
Ces exemples illustrent la puissance d'interprétabilité des modèles GLM car pour les variables qualitatives nous pouvons calculer les différents coefficients de passage d'une modalité à l'autre afin d'identifier les modalités les plus couteuses. Pour les variables quantitatives, dans un modèle GLM avec un lien *log*, il suffit de multiplier la valeur de la variable par $\exp(\beta)$. Si $\beta > 0$, une augmentation de la variable entraine une augmentation du prix, si $\beta < 0$, une augmentation de la variable entraine une diminution du prix et si $\beta = 0$ la variable n'a aucun incidence sur le prix. Ces élements expliquent l'attirance de ce modèle par les assureurs.
Concernant la grille de tarification 5.1 des GLM, nous observons que les modalités les plus couteuses par variable dans l'ordre croissant sont :

— 5, 2, 4, 3, 1 pour la formule auto.
— 2 et AUTRES pour la région de l'assuré.

La prime pure est une fonction décroissante en fonction de l'ancienneté de contrat ce qui est cohérent. En résumé, les modèles GLM possèdent une forte capacité d'interprétabilité par des coefficients. Ces derniers nous permettent de connaitre les caractéristiques des conducteurs les plus coûteux afin de piloter ces mauvais risques.

5.3 Modèle *CART*

Le modèle *CART* est également simple à interpréter du moment que le nombre de feuilles terminales n'est pas trop important. Le principe d'arbres de décision permet de comprendre le tarif associé aux caractéristiques d'un assuré. Comme prédédemment, nous allons présenter une grille de tarification sous la même maille afin d'expliquer les tarifs proposés par le modèle *CART*. Rappelons que la tarification dans le modèle *CART* en régression associe chaque feuille à une valeur prédite obtenue en effectuant la moyenne des réponses. Dans notre cas, nous avons 15 feuilles terminales pour le modèle fréquence et 5 feuilles terminales pour le modèle coût. Ce qui fait 75 feuilles terminales pour le modèle fréquence $\times$ coût. Nous allons donc analyser les différentes possibilités de prime pure afin de cibler les mauvais risques. Dans un premier temps, nous faisons un rappel sur nos modèles *CART* fréquence et coût.

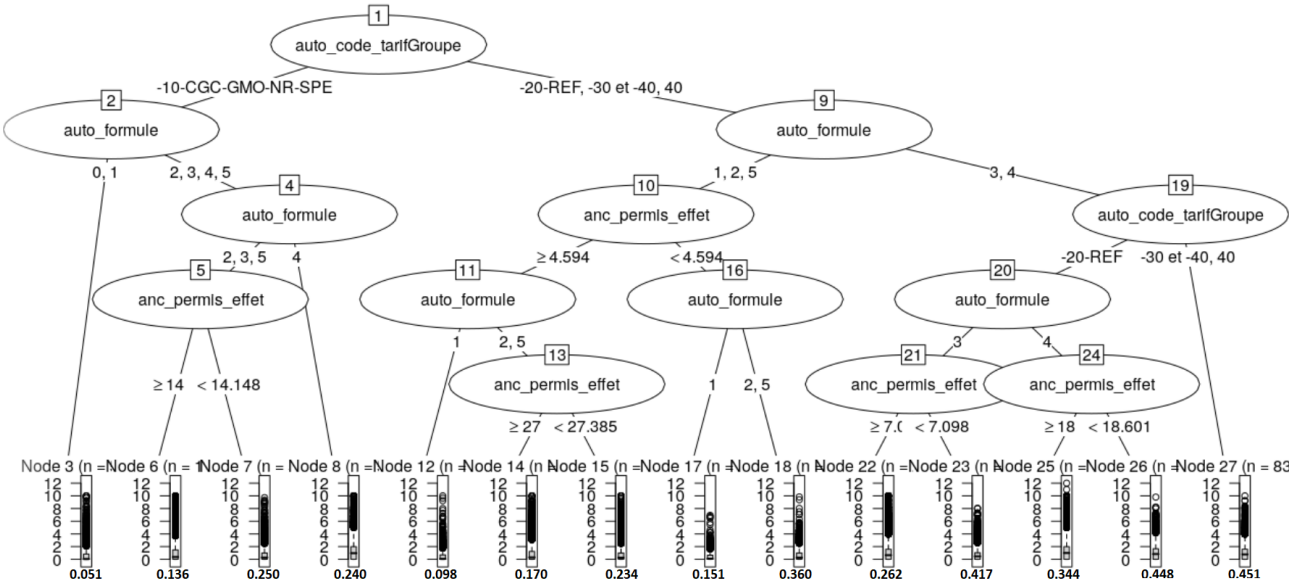

FIGURE 5.1 – Arbre optimal modèle *CART* pour la fréquence

La figure 5.1 présente l'arbre *CART* fréquence ainsi que les fréquences de sinistres annuelles associée à chaque feuille. Cet arbre permet de classifier les risques en fonction des caractéristiques du conducteur. Nous avons 14 groupes d'individus associés à 14 fréquences de sinistres. Chaque groupe a des caractéristiques différentes. Par exemple, le groupe d'assuré avec les plus mauvais risque en fréquence est celui avec un code tarif égal à "-40, 30 ou 40" et une formule auto égale à "3 ou 4". Sa probabilité annuelle de sinistres est égale à 0,451. Le groupe d'assuré avec le moins de risques en fréquence est celui dont le code tarif est "-10, CGC, GMO, NR ou SPE" et la formule auto est "0 ou 1". Sa probabilité annuelle de sinistres est de 0,0051. Cette analyse doit être complétée par celle de l'arbre *CART* coût.

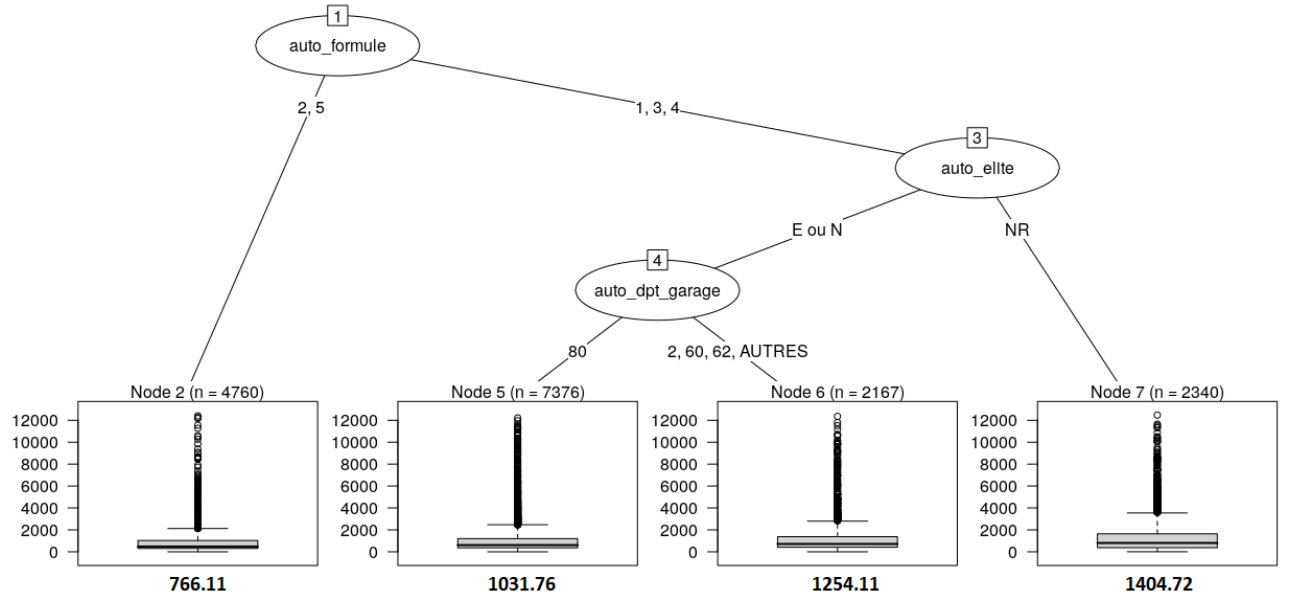

FIGURE 5.2 – Arbre optimal modèle *CART* pour le coût

La figure 5.2 présente l'arbre *CART* coût ainsi que le coût moyen d'un sinistre associé à chaque feuille. Cette fois, nous n'avons que 4 groupes d'individus associés à 4 coûts moyens pour un sinistre. Cette fois, le groupe d'individus avec les plus mauvais risque a pour formule auto "1, 3 ou 4" et un auto elite

égal à "NR". Son coût moyen d'un sinistre est égal à 1 404,72. Le groupe d'individus avec les meilleurs risques est celui avec une formule auto égale à "2 ou 5". Son coût moyen d'un sinistre est égal à 766,11.

auto_formule	anc_contrat		anc_contrat		anc_contrat		anc_contrat		anc_contrat	
	0		1		2		7		8	
	auto_zone		auto_zone		auto_zone		auto_zone		auto_zone	
	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES
1	137	137	137	137	137	137	137	137	137	137
2	179	179	179	179	179	179	179	179	179	179
3	368	368	368	368	368	368	368	368	368	368
4	629	629	629	629	629	629	629	629	629	629
5	179	179	179	179	179	179	179	179	179	179

TABLE 5.3 – Grille de tarifation pour le modèle *CART*

Le tableau 5.3 correspond aux différents tarifs pour les assurés en fonction de la formule auto, l'ancienneté du contrat et la région.
Nous pouvons comprendre les primes pures données par le tableau 5.3 en suivant les décisions des arbres fréquence 5.1 et coût 5.2 en fonction des variables explicatives :

- 137 = 0,098 × 1 404,72.
- 179 = 0,234 × 766,11.
- 368 = 0,262 × 1 404,72.
- 629 = 0,448 × 1 404,72.

Les arbres de décisions permettent de cibler les plus gros risques étant les conducteurs avec les plus grandes primes pures (fréquence × coût). Dans nos modèles *CART*, nous avons 37 possibilités de prime pure (ce n'est pas exactement 14 (nombre de possibilités pour la fréquence) × 4 (nombre de possibilités pour le coût) car certaines combinaisons sont impossibles). Parmi ces possibilités, les conducteurs les plus coûteux sont (pas forcément présent dans la grille) :

- Les conducteurs avec un auto elite égal à "NR", une formule auto égale à "3 ou 4" et un code tarif égal à "-40, -30 ou 40" correspondant à une prime pure égale à 634 (0,451 × 1 404,72)
- Les conducteurs avec un auto elite égal à "NR", une formule auto égale à "4", un code tarif égal à "-20 ou REF" et une ancienneté de permis inférieure à 18,601 correspondant à une prime pure égale à 629 (0,448 × 1 404,72)
- Les conducteurs avec un auto elite égal à "NR", une formule auto égale à "3", un code tarif égal à "-20 ou REF" et une ancienneté de permis inférieure à 7,098 correspondant à une prime pure égale à 586 (0,417 × 1 404,72)

Les résultats sont cohérents car les assurés avec les anciennetés de permis les moins élevés sont associés aux primes pures les plus élevées. En résumé, les modèles CART permettent d'avoir une interprétabilité visuelle, simple, rapide et de comparer la classification des risques avec celle des modèles GLM. L'arbre binaire permet de classer tous les individus du portefeuille en fonction de leur caractéristiques afin de connaitre les mauvais risques et d'agir en conséquence.

5.4 Modèle *XGBoost*

Le modèle *XGBoost* fait parti de la famille des modèles *machine learning* difficilement interprétable. Ce modèle n'a pas de sortie interprétable comme les modèles GLM ou le modèle *CART*. Par conséquent, l'un des objectifs des *data scientists* et des actuaires est de réussir à avoir une bonne interprétabilité de ce modèle. Dans ce but, nous introduisons l'algorithme de SHAP[26] [1] permettant d'expliquer des résultats issus des classifieurs ou régresseurs d'un modèle d'apprentissage automatique en construisant

1. SHapley Additive exPlanation

un modèle linéaire autour de la prédiction. Cet algorithme permet de donner l'importance de chaque variable explicative. Cette approche s'appuie sur des outils de la théorie des jeux avec la valeur de Shapley[34] et d'outils d'explications locales avec les algorithmes LIME[31]. Nous utilisons cet algorithme à l'aide du package `SHAPforxgboost`[25].

Algorithme *SHAP*

L'idée de l'algorithme de *SHAP* est de calculer la valeur de Shapley pour toutes les variables explicatives. Cette valeur est calculée à partir des écarts entre la valeur prédite pour toutes les combinaisons de modalités ou valeurs des variables et la valeur moyenne. Par exemple pour la fréquence, avec une combinaison, nous expliquons l'écart entre la fréquence annuelle des sinistres prédits et la fréquence moyenne annuelle des sinistres prédits. L'idée globale est d'effectuer toutes les combinaisons afin de moyenner l'impact de chaque variable. L'expression générale de la valeur de Shapley φ_i est

$$\varphi_i = \sum_{S \in S(\{1,\dots,p\}) \setminus \{i\}} \frac{|S|!(p-|S|-1)!}{p!} \left(f_x(S \cup i) - f_x(S)\right),$$

avec i la ième variable, p le nombre de variables, $S(\{1,...,p\}))$ est l'ensemble des sous-ensembles de $\{1,...,p\}$, f_x la fonction de prédiction en x, $f_x(S) = E\left[f(x) \mid S\right]$.
La prédiction peut être écrite comme la somme des différents effets des variables ajoutée à la valeur de base φ_0 étant la moyenne de toutes les prédictions

$$f(x) = y_{\text{pred}} \ = \varphi_0 + \sum_{i=1}^{p} \varphi_i z'_i$$

avec, y_{pred} la valeur prédite du modèle, φ_0 la valeur de base du modèle, $z' \in \{0,1\}^p$ prenant les valeurs $z'_i = 1$ lorsque la prédiction contient la variable i et $z'_i = 0$ sinon.
Cette expression permet de comprendre l'effet des variables sur une observation. En effet, si nous prenons les i tel que $z'_i = 1$, nous avons l'effet de chaque variable i par la valeur φ_i. Si φ_i est strictement positif, la variable i augmente la valeur de la prédiction si φ_i est strictement négatif, la variable i baisse la valeur de la prédiction sinon φ_i n'influe pas sur la valeur de prédiction. Nous pouvons aussi connaitre l'intensité de la variable i sur la valeur prédite car plus φ_i est grand plus la valeur de la prédiction augmentera. En moyennant les valeurs absolues des valeurs de *Shapley* pour chaque variable, nous avons l'importance globale des variables dans la modélisation du modèle. Pour plus de précisions concernant l'algorithme *SHAP*, nous vous renvoyons au mémoire[12] dédié à l'interprétabilité des modèles dont ceux de *machine learning* et de *XGBoost*.

Interprétabilité de la fréquence

Dans cette partie, nous allons donner et expliquer les fréquences en utilisant la même maille que précédemment. Concernant l'interprétabilité, nous utilisons l'algorithme *SHAP*. Cette fois nous décomposons l'interprétabilité de la fréquence et du coût.

auto_formule	anc_contrat		anc_contrat		anc_contrat		anc_contrat		anc_contrat	
	0		1		2		7		8	
	auto_zone		auto_zone		auto_zone		auto_zone		auto_zone	
	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES	AUTRES	2
1	0,39	0,39	0,39	0,39	0,39	0,39	0,39	0,39	0,39	0,39
2	0,52	0,52	0,53	0,53	0,53	0,53	0,53	0,53	0,50	0,50
3	0,74	0,74	0,74	0,74	0,74	0,74	0,74	0,74	0,73	0,73
4	1,10	1,10	1,13	1,13	1,14	1,14	1,16	1,16	1,15	1,15
5	0,51	0,51	0,51	0,51	0,51	0,51	0,50	0,50	0,48	0,48

TABLE 5.4 – Grille des fréquences annuelles de sinistres pour le modèle *XGBoost*

Le tableau 5.3 donne les fréquences annuelles de sinistres. Nous observons que les formules auto 3 et 4 sont les plus risqués. Le changement de région n'influence pas sur la prime pure et l'ancienneté de contrat influe peu sur la fréquence de sinistres.

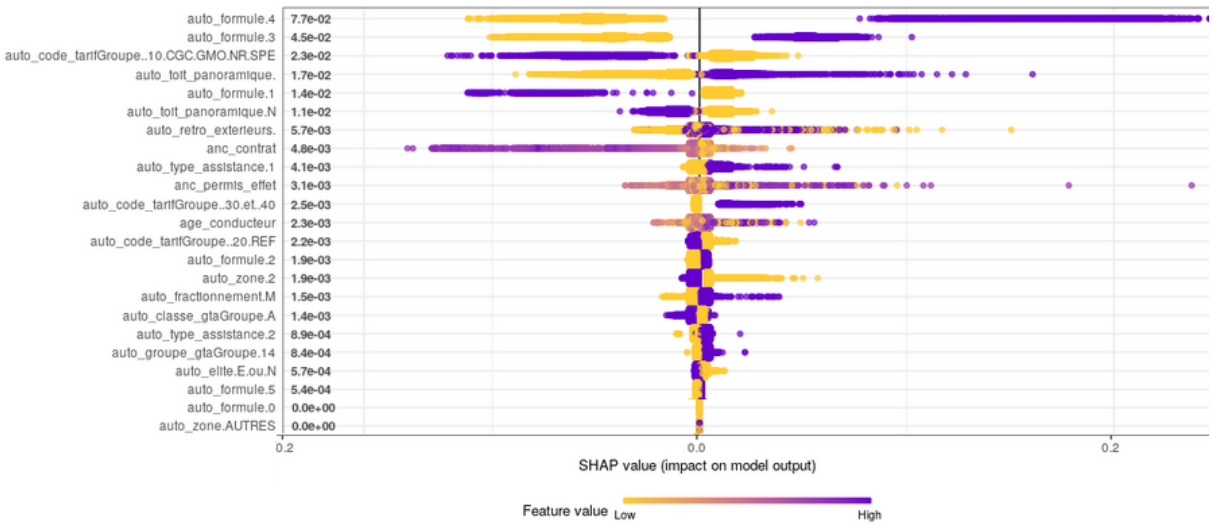

FIGURE 5.3 – Importance des variables par méthode *SHAP* pour le modèle fréquence *XGBoost*

En complément de la grille tarifaire 5.4, nous donnons les valeurs de *Shap* associées à la fréquence pour chaque modalité sur le graphique 5.3. La moyenne de nos prédictions des fréquence est 0,58 étant associé à la valeur 0 sur le graphique 5.3. Une *Feature value* élévée (bleu foncé) correspond à un nombre élévé d'observations, une *Feature value* faible correspond à un nombre faible d'observations. La valeur à côté du nom de la variable est la valeur de shap associée à la variable correspondant à la somme des valeurs absolues de toutes les valeurs de *Shapley* donnant l'importance de la variable sur le modèle. La position d'un point par rapport à l'origine donne l'écart entre la valeur prédite et la valeur moyenne. De la figure 5.3, nous pouvons déduire ces éléments :

- Une formule auto égale à 4 a de grande chance d'augmenter la fréquence prédite de minimum 0,08 allant jusqu'à ajouter 0,35, peu de chance de baisser la fréquence prédite de 0,02 à 0,11 et aucune chance d'avoir aucun impact sur la fréquence.
- Une formule auto égale à 3 a de grande chance d'augmenter la fréquence prédite de 0,03 à 0,09, peu de chance de baisser la fréquence prédite de 0,02 à 0,10 et aucune chance de n'avoir aucun impact sur la fréquence
- Une formule auto égale à 1 a de grande chance de baisser la fréquence prédite de maximum 0,12 et peu de chance d'augmenter la fréquence prédite de maximum 0,02.
- L'ancienneté de contrat aura de grande chance de baisser la fréquence prédite de maximum 0,14 et peu de chance de l'augmenter de maximum 0,03
- Une formule auto égale à 0,2 ou 5, une région associée à 2 a très peu d'influence sur la fréquence prédite.

Ces différentes interprétations se retrouvent sur la grille de tarification 5.3 car nous avons bien une stabilisation des formules auto 2 et 5, une bonne sinistralité pour la formule auto 1 et une mauvaise sinistralité pour la formule auto 4 et 3. De plus, un changement de région n'influence pas le tarif et l'impact de l'ancienneté de contrat n'est pas aberrant au vu des valeurs prises pas le spectre de cette variable sur la figure 5.3. De plus, avec la figure 5.3 nous pouvons cibler les modalités augmentant significativement la fréquence prédite permettant une gestion du risque adaptée. Par exemple, un code tarif égal à 30 ou 40 et une assistance égale à 1 ont une grande probabilité d'augmenter la fréquence prédite. À contrario, un code tarif égal à 10, CGC, GMO, NR ou SPE a de grande chance de baisser la fréquence prédite. Ces différents éléments montrent que l'algorithme *SHAP* permet une interprétation

globale. Par ailleurs, une interprétation locale est possible car l'attribution d'une modalité augmentant ou baissant la valeur de la fréquence prédite est quantifiée par sa valeur de *Shapley*. Par exemple, la valeur de *Shapley* de chaque modalité pour un assuré avec une fréquence prédite égale à 1,14 sur le tableau 5.4 (un assuré avec une formule auto égale à 4, une région associée à 2 et une ancienneté de contrat égal à 2 est :

Modalité	**Valeur de Shapley**
auto_formule4	0,30
auto_ancienneté de permis	0,23
auto_toit_panoramique	0,02
Le reste	0,01

TABLE 5.5 – Valeur de *Valeur de Shapley* des différentes modalités

Le tableau 5.5 donne les valeurs de *Shapley* associées aux différentes modalités. Ainsi, avec la moyenne des prédictions (nous rappellons qu'elle est égale à 0,58) et la valeur de *Shapley*, nous pouvons retrouver la fréquence prédite en ajoutant les valeurs de *Shapley* :

$$1,14 = 0,58 + 0,30 + 0,23 + 0,02 + 0,01.$$

Dans notre cas, les valeurs de *Shapley* sont positives ou négligeables mais elles peuvent bien évidemment être négatives et traduire une baisse de la fréquence prédite. Ce processus permet d'avoir une explicabilité du modèle en ciblant et quantifiant l'effet de chaque modalité sur la fréquence prédite. Contrairement aux modèles GLM, nous n'avons pas une explication direct d'une modalité à une autre. En effet, pour les modèles GLM le changement d'une modalité se traduit par la multiplication d'un coefficient. Dans l'algorithme de *Shap*, le changement d'une modalité modifie toutes les valeurs de *Shapley* donnant une moins bonne explication que les modèles GLM et une gestion du risque plus compliquée.

Interprétabilité du coût moyen unitaire

Dans cette partie, nous étudions les coûts moyens unitaires de manière analogue à la fréquence vu précedemment.

auto_formule	anc_contrat		anc_contrat		anc_contrat		anc_contrat		anc_contrat	
	0		1		2		7		8	
	auto_zone		auto_zone		auto_zone		auto_zone		auto_zone	
	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES	AUTRES	2
1	1 266	1 416	1 279	1 381	1 279	1 429	1 342	1 444	1 342	1 450
2	939	1 041	951	1 054	961	1 064	1 034	1 134	1 034	1 136
3	1 290	1 392	1 308	1 405	1 302	1 411	1 366	1 468	1 366	1 468
4	1 193	1 296	1 158	1 308	1 206	1 308	1 222	1 324	1 222	1 324
5	916	1 018	916	1 018	916	1 018	979	1 082	979	1 082

TABLE 5.6 – Grille des coûts unitaires de sinistres pour le modèle *XGBoost*

Le tableau 5.6 donne les coûts moyens unitaires de sinistres. Nous observons que les formules auto 1,3 et 4 sont les plus risquées. Un changement de région de 2 à AUTRES entraine une augmentation du tarif et l'ancienneté de contrat a un faible impact.

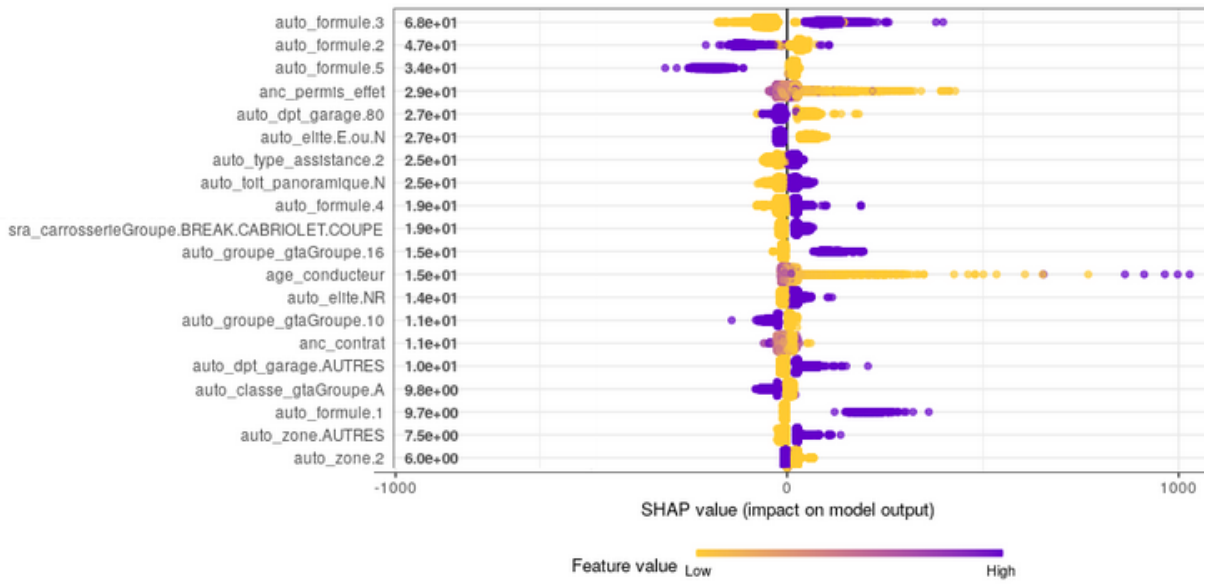

FIGURE 5.4 – Importance des variables par méthode *SHAP* pour le modèle coût *XGBoost*

Comme pour le modèle *XGBoost* fréquence, nous présentons les valeurs de *Shapley* associés au coût moyen unitaire sur la figure 5.4. Nous précisons que la valeur moyenne est de 1002, nous pouvons déduire de la figure 5.4 ces éléments :

- Une formule auto égale à 3 a de grande chance d'augmenter le coût moyen prédit de maximum 350, peu de chance de baisser le coût moyen prédit de maximum 250.
- Une formule auto égale à 2 a de grande chance de baisser le coût moyen prédit de maximum 250, peu de chance d'augmenter le coût moyen prédit de maximum 100 et de grande chance de l'augmenter de 100 à 175.
- Une formule auto égale à 5 a de grande chance de baisser le coût moyen prédit de maximum 400 et peu de chance d'augmenter le coût moyen prédit de maximum 50.
- Une formule auto égale à 4 a de grande chance d'augmenter ou baisser le coût moyen prédit de maximum de 175.
- Une formule auto égale à 1 a de grande chance d'augmenter le coût moyen prédit de 100 à 400.
- L'ancienneté de contrat n'a pas de grand influence sur le coût moyen prédit.
- Une région AUTRES a de grande chance d'augmenter le coût moyen prédit de maximum 250.

Ces différentes interprétations se retrouvent sur la grille de tarification 5.6. Par exemple, les formules auto 1 et 3 ont de grande chance d'augmenter le coût moyen prédit étant observable sur la grille 5.4 car les différents coûts moyens unitaires associés à la formule auto 1 et 3 sont les plus élevées. Les autres modalités ont aussi une cohérence entre le graphique 5.4 en sortie de l'algorithme *SHAP* et la grille de tarification 5.6.
De plus, avec la figure 5.4 de la méthode *SHAP* nous pouvons aussi identifier les mauvais risques comme un groupe SRA égal à 16 ce qui est cohérent et les bons risques comme une formule auto égale à 5. Comme pour le modèle fréquence, une interprétation globale et locale est désormais possible par l'application de l'algorithme *SHAP* sur le modèle *XGBoost*.

Grille de tarification du modèle *XGBoost*

Comme pour les modèles GLM et *CART*, nous donnons une grille de tarification étant le produit entre les fréquences et les coûts unitaires prédits par le modèle *XGBoost* donnés par les tableaux 5.4 et 5.6.

auto_formule	anc_contrat		anc_contrat		anc_contrat		anc_contrat		anc_contrat	
	0		1		2		7		8	
	auto_zone		auto_zone		auto_zone		auto_zone		auto_zone	
	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES	AUTRES	2
1	489	547	494	533	494	551	518	558	518	560
2	492	545	500	553	505	559	543	596	518	569
3	954	1 030	968	1 040	964	1 044	1 010	1 086	997	1 072
4	1 307	1 419	1 314	1 484	1 376	1 493	1 412	1 531	1 404	1 522
5	465	517	466	519	466	519	490	542	469	518

TABLE 5.7 – Grille de tarification le modèle *XGBoost*

Le tableau 5.7 donne les différents tarifs en fonction de la formule auto, l'ancienneté du contrat et la zone. Nous observons que ces tarifs sont très élévés s'expliquant par une prédiction élevée de la fréquence et du coût. Nous observons que les modalités les plus couteuses par ordre croissant sur cet exemple sont :

— Une formule auto égale à 5,2,1,3 et 4.
— 2 et AUTRES pour la zone.

La prime pure est une fonction légèrement décroissante en fonction de l'ancienneté de contrat. Nous observons que ces primes pures sont très élevées s'expliquant par la prédiction très élevée des fréquences annuelles de sinistres. La prédiction élevée de la fréquence est dûe au calibrage du modèle *XGBoost* non adapté à l'échantillon test car nous observons que la fréquence de sinistres sur cet échantillon est de 0,39 contre 0,56 pour les prédictions entrainant une surestimation de la fréquence. De plus, nous observons que les formules auto 4 et 3 sont très volatiles et donnent des fréquences prédites largement au dessus de la moyenne expliquant une prime pure très élevée. L'interprétabilité globale du modèle reste moins claire que pour les modèles GLM. En effet, les modèles GLM donnent un coefficient de passage d'une modalité à une autre alors que l'algorithme de *SHAP* donne le passage d'une modalité à une autre sachant toutes les autres modalités ce qui nous amène à devoir considérer toutes les combinaisons de modalités rendant le problème plus complexe.

5.5 Modèle *MOB*

Les modèles *MOB* en particulier le modèle *GLM trees* (faisant parti de la classe des modèles *MOB*) segmentent le portefeuille afin de créer des classes homogènes et d'estimer un modèle GLM pour chaque segment. Comme précédemment, nous donnerons une grille de tarification donnée par les modèles *MOB*. Rappelons les modèles *MOB* fréquence et coût.

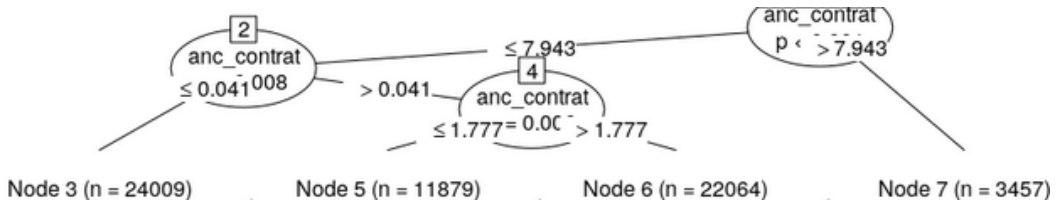

FIGURE 5.5 – Arbre optimal modèle *MOB* fréquence

L'arbre 5.5 correspond à l'arbre du modèle *MOB* fréquence. Nous avons 4 segmentations par rapport à l'ancienneté de contrat et donc 4 modèles GLM associés à chaque feuille.

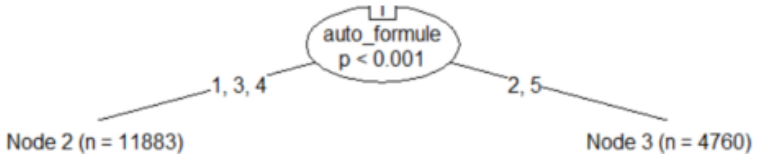

FIGURE 5.6 – Arbre optimal modèle *MOB* coût

La figure 5.6 présente l'arbre du modèle *GLM* coût. Nous avons 2 segmentations par rapport à la formule auto et donc respectivement 2 modèles GLM.

auto_formule	anc_contrat		anc_contrat		anc_contrat		anc_contrat		anc_contrat	
	0		1		2		7		8	
	auto_zone		auto_zone		auto_zone		auto_zone		auto_zone	
	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES	2	AUTRES
1	65	80	70	86	102	126	95	118	0	0
2	68	96	71	99	119	167	118	166	0	0
3	201	249	183	226	360	445	337	416	0	0
4	291	362	247	304	504	622	471	582	0	0
5	64	91	68	95	108	152	108	151	0	0

TABLE 5.8 – Grille de tarification pour le modèle *MOB*

À l'aide des arbres 5.5, 5.6 et des coefficients des GLM du tableau 5.9, nous allons interpréter l'effet des variables sur le tarif présenté dans le tableau 5.8.

Modalité ou variable	Modèle fréquence				Modèle coût	
	Feuille 1	Feuille 2	Feuille 3	Feuille 4	Feuille 1	Feuille 2
Formule0	Référence	Référence	Référence	Référence	Non présent	Non présent
Formule1	0,5366	0,7156	0,3861	- 2,5680	Non présent	Non présent
Formule2	1,4679	1,6008	1,7294	- 1,0886	Non présent	Non présent
Formule3	1,6704	1,6810	1,6505	- 0,9369	Non présent	Non présent
Formule4	2,0436	1,9797	1,9858	- 0,1884	Non présent	Non présent
Formule5	1,4134	1,5596	1,6372	- 1,0127	Non présent	Non présent
auto_zone2	Référence	Référence	Référence	Référence	Référence	Référence
auto_zoneAUTRES	0,1353	0,1262	0,1288	- 1,5553	0,0823	0,2117
anc_contrat	Non présent	Non présent	Non présent	Non présent	- 0,0134	- 0,0012

TABLE 5.9 – Coefficients du GLM en fonction de la feuille terminale

La compréhension de la tarification se fait en deux étapes :

— Application d'arbres de décision avec les variables explicatives de classification afin d'associer une observation à un modèle GLM.

— Application du modèle GLM sur les variables explicatives de régression.

Dans un premier temps, nous observons qu'étonnamment la prime pure des assurés avec une ancienneté de contrat égale à 8 est nulle. En regardant la base de données, nous avons remarqué que les assurés avec un groupe SRA égal à "0-2-4-5-5-7" et une ancienneté de contrat strictement supérieure à 7.943 n'ont aucun sinistre observé. Nous rappelons que le modèle MOB fréquence sur la figure 5.5 a décidé d'effectuer une segmentation pour les anciennetés de contrat strictement supérieures 7.943 engendrant un modèle GLM sur ces observations. Par conséquent, la fréquence de sinistres estimée de ces observations est très proche de 0. Nous précisons que ces observations sont au nombre de 21 sur l'ensemble de la base de données amenant le problème classique de l'utilisation des lois statistiques sur peu de données. Nous allons procéder à un exemple afin de comprendre la différence de comportement créée par la segmentation. D'une part, nous prenons un assuré avec une ancienneté de contrat égale à 0, une formule auto égale à 1 et habitant dans la région associée à 2. Nous voulons connaitre le prix ainsi que l'effet engendré par une personne avec une ancienneté de contrat égale à 0, une formule auto égale à 1 et habitant dans la région associée à "AUTRES". Nous rappellons que ces individus appartiennent à la feuille 1 pour la fréquence et le coût. D'autre part, nous prenons un assuré avec une ancienneté de contrat égale à 0, une formule auto égale à 2 et habitant dans la région associée à 2. Nous voulons

connaitre le prix et l'effet d'un assuré avec les mêmes caractéristiques hormis une région associée à "AUTRES". Ces individus appartiennent à la feuille 1 pour la fréquence et la feuille 2 pour le coût. Notons $PrimePure1$, $PrimePure1^{'}$ les primes pures que devraient payer les 2 individus associés à la première feuille du modèle *MOB* fréquence et coût et $PrimePure2$, $PrimePure2^{'}$ les primes pures que devraient payer les 2 individus associés à la première feuille du modèle *MOB* fréquence et la deuxième feuille du modèle *MOB* coût.

$$\begin{aligned} Primepure1^{'} &= Primepure1 \times \exp(0,1353) \times \exp(0,0823) \\ &= Primepure1 \times \exp(0.2176) \\ &= 65 \times 1,24309 \approx 80 \end{aligned}$$

$$\begin{aligned} Primepure2^{'} &= Primepure2 \times \exp(0,1353) \times \exp(0,2117) \\ &= Primepure2 \times \exp(0,3470) \\ &= 68 \times 1,414817 \approx 96 \end{aligned}$$

Cet exemple illustre la différence de comportement du changement de zone en fonction des feuilles. Nous précisons que dans un modèle GLM classique, les coefficients de passage entre ces primes pures ne sont pas identiques car un changement de formule (passage d'une auto formule 1 à 2) est constaté ajoutant un effet. Cet effet est linéaire car le passage d'une modalité à une autre dans un modèle GLM se fait en multipliant par le même coefficient. Dans le modèle *MOB*, le fait de changer de formule auto peut entrainer un changement de feuille et une modification de tous les coefficients créant un effet non linéaire plus difficilement interprétable. Par exemple, si nous voulons connaitre l'effet d'un passage d'une formule auto 1 à 2, ce changement ne peut pas être quantifié par la table 5.9 car les coefficients GLM de toutes les modalités sont différents (dans un modèle GLM classique seul les coefficients formule auto sont différents). Il est intéressant de préciser que la prime pure en fonction de l'ancienneté de contrat étant une variable quantitative n'est pas monotone pour le modèle *MOB*. Modéliser un tel effet est impossible avec les modèles GLM car la linéarité entraine une monotonie. De plus, nous pouvons nous questionner à propos de l'intérêt de la segmentation au vu de la similitude des coefficients du modèle fréquence sur la table 5.9. Certes ces coefficients sont très proches mais les autres coefficients des différents modèles GLM peuvent être différents comme par exemple les coefficients du nombre de chevaux fiscaux en annexe A.4. Concrétement, avec la table de tarification 5.8, nous observons que les modalités les plus couteuses par ordre croissant sont :

— 5, 1, 2, 3 et 4 pour la formule auto.
— 2 et AUTRES pour la région de l'assuré.
— 8, 0, 1, 7 et 2 pour l'ancienneté de contrat.

En résumé, les modèles *MOB* ont une interprétabilité similaire aux modèles GLM et *CART*. Les combinaisons des arbres de décision et des coefficients de régression permettent d'avoir une autre vision sur l'exposition de nos risques.

5.6 Synthèse des primes pures

Dans cette dernière partie, nous allons nous comparer les tarifs en fonction des modèles GLM, *CART*, *XGBoost* et *MOB*.

Modèle GLM

Concernant les modèles GLM, nous avons observé sur le tableau des tarifs 5.1 que les primes pures vont de 68 pour les assurés avec une ancienneté de contrat égale à 8, une formule auto égale à 5 et une

région associée à 2 jusqu'à 174 pour les assurés avec une ancienneté de contrat égale à 0, une formule auto égale à 1 et une région associée à AUTRES. Globalement, les différents tarifs modélisés par le modèle GLM sur les modalités choisies sont assez homogènes.
La table des tarifs 5.1 ne permet pas d'avoir une idée des primes pures car elle est restreinte sur certaines modalités. Afin d'avoir une idée des primes pures modélisées par le modèle GLM, nous calculons les primes pures sur l'échantillon de test.

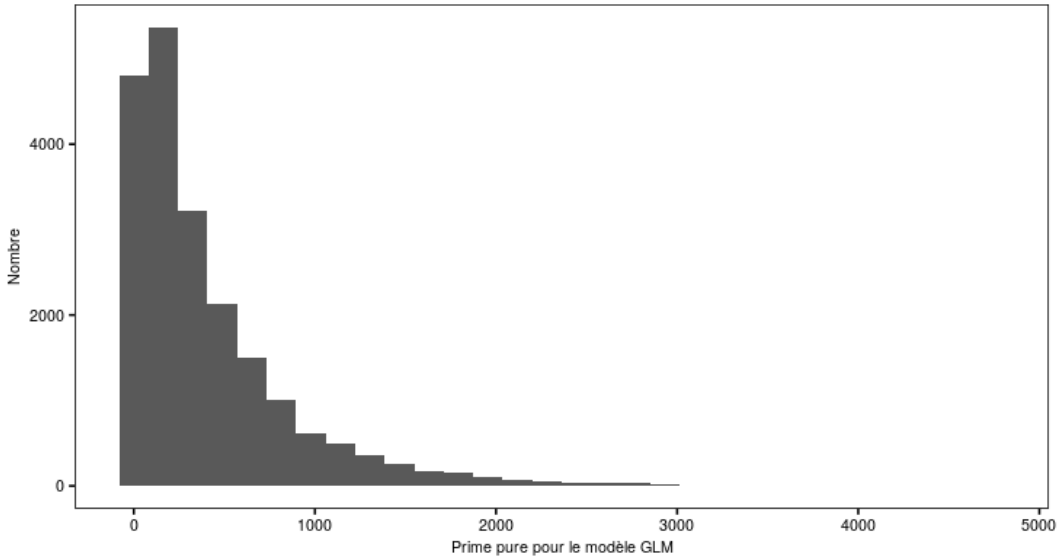

FIGURE 5.7 – Distribution des primes pures pour le modèle GLM

Sur la figure 5.7, nous avons en abscisse les primes pures modélisées par le modèle GLM sur l'échantillon test et en ordonnée le nombre d'observations total associé à la prime pure. Par exemple, nous avons environ 5500 observations avec une prime pure comprise entre 0 et 100. Ce graphique a l'allure d'une loi Gamma étant la loi de modalisation du modèle GLM coût.

Minimum	**1er Quantile**	**Médiane**	**Moyenne**	**3e Quantile**	**Maximum**	**Ecart type**
0	88	247	407	555	4 726	472

TABLE 5.10 – Statistiques descriptives des préditions de primes pures sur l'échantillon test par le modèle GLM
Le tableau 5.10 donne les paramétres statistiques des primes pures sur l'échantillon test modélisées par le modèle GLM. Ces paramètres seront comparés dans la dernière partie de ce chapitre 5.13.

Modèle *CART*

Pour le modèle *CART*, nous avons observé sur le tableau des tarifs 5.3 que les primes pures vont de 137 pour les assurés avec une formule auto égale à 1 jusqu'à 629 pour les assurés avec une formule auto égale à 4.
Comme pour les modèles GLM, la table des tarifs 5.3 des modèles *CART* ne permet pas d'avoir une idée des primes pures car elle est restreinte sur certaines modalités. Afin d'avoir une idée des primes pures modélisées par le modèle GLM, nous calculons les primes pures sur l'échantillon test.

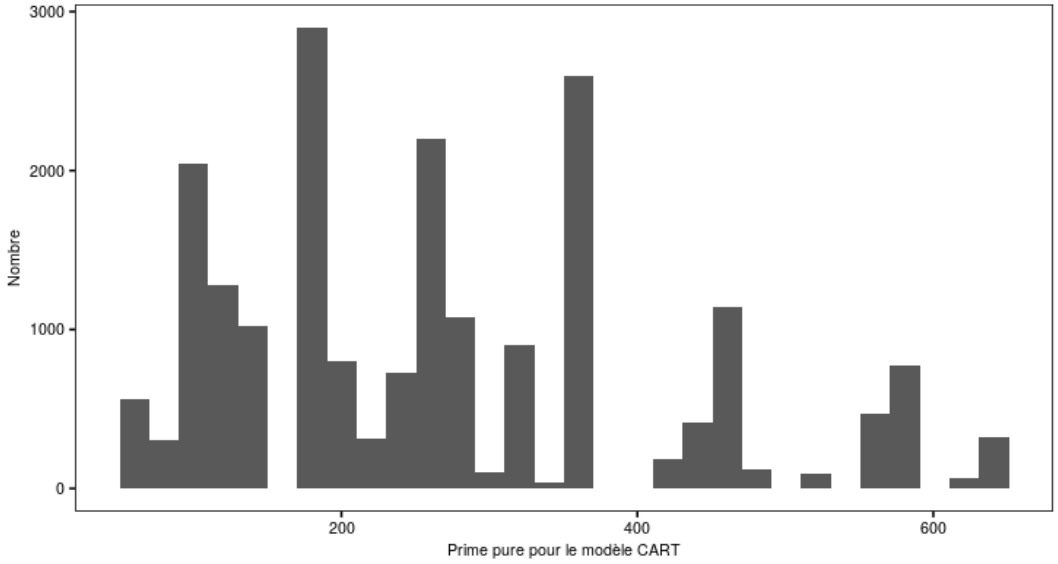

FIGURE 5.8 – Distribution de la prime pure pour le modèle *CART*

Sur la figure 5.8, nous avons en abscisse les primes pures modélisées par le modèle *CART* et en ordonnée le nombre d'observations total associé à la prime pure. Par exemple, nous avons environ 300 observations avec une prime pure égale à 634 correspondant aux feuilles terminales associées aux prédictions de fréquences et coûts unitaires les plus élevées vu dans la partie 5.3.

Minimum	**1er Quantile**	**Médiane**	**Moyenne**	**3e Quantile**	**Maximum**	**Ecart type**
53	141	270	269	356	634	145

TABLE 5.11 – Statistiques descriptives des préditions de primes pures sur l'échantillon test par le modèle *CART*

Le tableau 5.11 donne les paramétres statistiques des primes pures sur l'échantillon test modélisées par le modèle *CART*. Ces paramètres seront comparés dans la dernière partie de ce chapitre 5.13.

Modèle *XGBoost*

Sur le modèle GLM, nous avons observé sur le tableau des tarifs 5.7 que les primes pures vont de 465 pour les assurés avec une ancienneté de contrat égale à 0, une formule auto égale à 5 et une région associée à 2 jusqu'à 1531 pour les assurés avec une ancienneté de contrat égale à 7, une formule auto égale à 4 et une région associée à AUTRES.

Afin d'avoir une idée des primes pures modélisées par le modèle *XGBoost*, nous calculons les primes pures sur l'échantillon test.

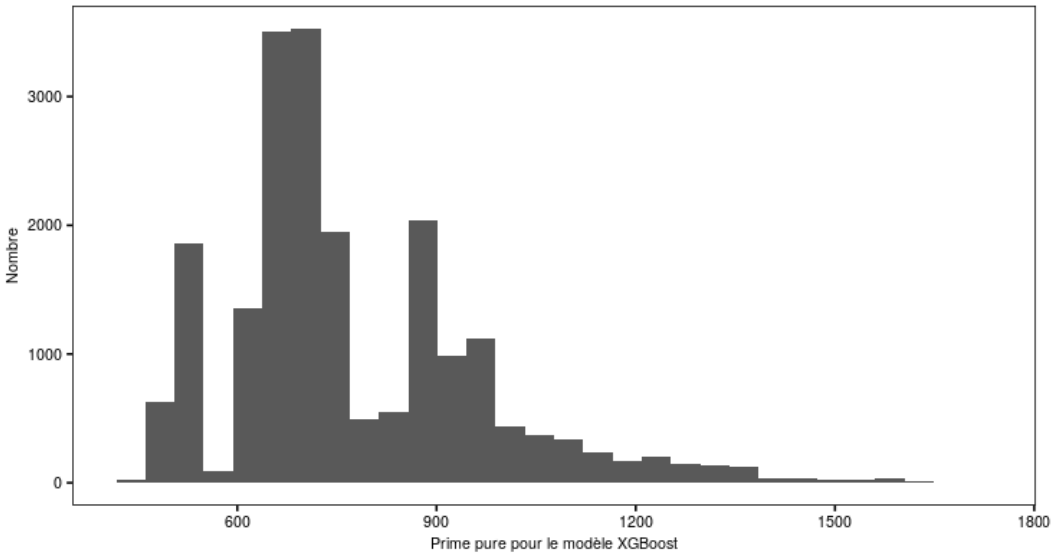

FIGURE 5.9 – Distribution de la prime pure pour le modèle *XGBoost*

Sur la figure 5.9, nous avons en abscisse la prime pure modélisée par le modèle *XGBoost* et en ordonnée le nombre d'observations total associé à la prime pure. Nous observons que nous avons aucune prime pure nulle et une forte masse en 750.

Minimum	**1er Quantile**	**Médiane**	**Moyenne**	**3e Quantile**	**Maximum**	**Ecart type**
442	649	714	771	891	1 717	193

TABLE 5.12 – Statistiques descriptives des préditions de primes pures sur l'échantillon test par le modèle *XGBoost*

Le tableau 5.12 donne les paramétres statistiques des primes pures sur l'échantillon test modélisées par le modèle *XGBoost*. Ces paramètres seront comparés dans la dernière partie de ce chapitre 5.13

Modèle *MOB*

Pour terminer, sur le modèle *MOB*, nous avons observé sur le tableau des tarifs 5.8 que les primes pures vont de 0 pour les assurés avec une ancienneté de contrat égale à 8 jusqu'à 622 pour les assurés avec une ancienneté de contrat égale à 2, une formule auto égale à 4 et une région associée à AUTRES. Comme pour les modèles GLM, *CART*, *XGBoost* et *MOB*, la table des tarifs 5.8 des modèles *MOB* ne permet pas d'avoir une idée des primes pures car elle est restreinte sur certaines modalités. Afin d'avoir une idée des primes pures modélisées par le modèle GLM, nous calculons las primes pures sur l'échantillon test.

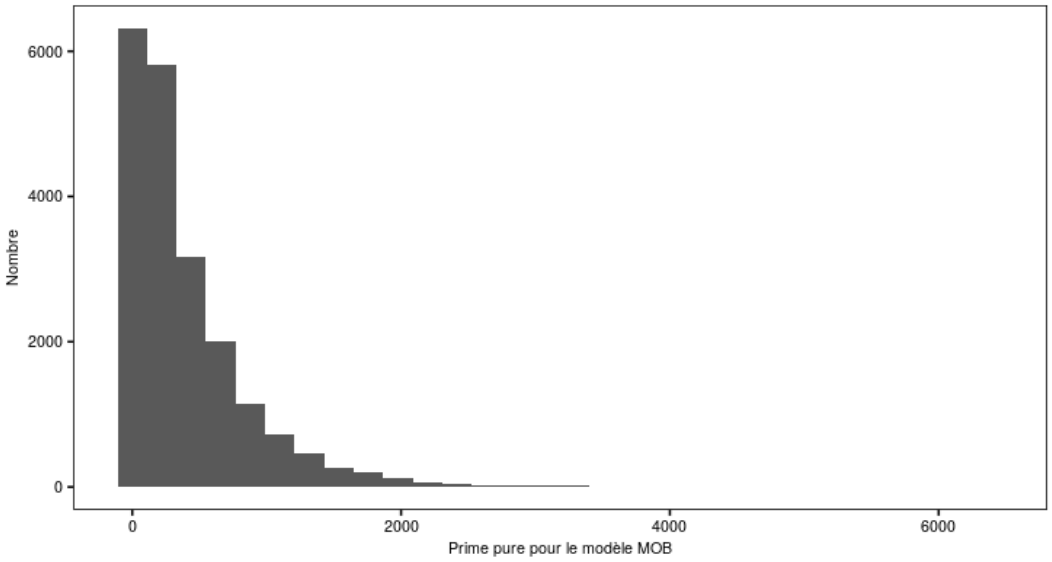

FIGURE 5.10 – Distribution de la prime pure pour le modèle *MOB*

Sur la figure 5.10, nous avons en abscisse les primes pures modélisées par le modèle *MOB* et en ordonnée le nombre d'observations total associé à la prime pure. Nous observons que ce graphique a l'allure d'une loi Gamma caractéristique de la modalisation du modèle *MOB* coût.

Minimum	**1er Quantile**	**Médiane**	**Moyenne**	**3e Quantile**	**Maximum**	**Ecart type**
0	80	243	400	553	6 369	472

TABLE 5.13 – Moyenne, écart type, minimum et maximum des préditions de primes pures sur l'échantillon test par le modèle *MOB*
Le tableau 5.13 donne les paramétres statistiques des primes pures sur l'échantillon test modélisées par le modèle *MOB*. Ces paramètres seront comparés dans la dernière partie de ce chapitre 5.13

Comparaison des primes pures

Nous commençons cette partie par une comparaison des primes pures des grilles de tarification vue précédemment et associés aux caractéristiques d'assuré décrites dans la partie 5.2.
Le modèle GLM a globalement les primes pures les plus faibles suivi du modèle *MOB* puis du modèle *CART* puis du modèle *XGBoost* s'expliquant par les prédictions de fréquences élevées. Les différences de tarifs se font globalement sur le changement de formule auto sauf pour le modèle *MOB* car l'ancienneté de contrat joue un rôle déterminant dans la tarification. En effet, nous rappelons que le modèle *MOB* fréquence effectue ses segmentations par rapport à l'ancienneté de contrat comme nous pouvons le voir sur la figure 5.5. Cette segmentation créée une non linéarité et des disparités sur chaque segment.
En prenant comme référence le modèle GLM, le modèle *CART* a des coûts globalement équivalents pour une formule auto égale à 1, des coûts globalement 5 fois plus élevés pour la formule auto 4 et des coûts globalement 2 fois plus élevé pour les formules auto égale à 2,3 ou 5.
Par rapport au modèle GLM, le modèle *XGBoost* a des coûts globalement 10 fois plus élevés pour une formule auto égale à 3, des coûts globalement 11 fois plus élevés pour la formule auto 4 et des coûts globalement 4 fois plus élevé pour les formules auto égale à 1,4 ou 5. Ces écarts s'expliquent par les prédictions de fréquences très élevées mais aussi par une modélisation donnant une classification des risques en fonction des modalités différentes pour ces modèles.
Pour le modèle *MOB* les primes pures sont très variables. Par exemple, pour une ancienneté de contrat égale à 8, le modèle *MOB* donne des primes pures égales à 0 contre environ 100 pour le modèle GLM.

À contrario, le modèle *MOB* donne des primes pures 5 fois plus élevées que celle du modèle GLM pour une formule auto égale à 4 et une ancienneté de contrat égale à 2.
Après avoir effectué une comparaison sur la grille de tarification propre à certaines modalités de variables. Nous allons nous intéresser à la comparaison des différentes primes pures de l'échantillon test en fonction du modèle.

Modèle	Min.	1er Qu.	Médiane	Moyenne	3e Qua.	Max.	Ecart type
GLM	0	88	247	407	555	4 726	472
CART	53	141	270	269	356	634	145
XGBoost	442	649	714	771	891	1 717	193
MOB	0	80	243	400	553	6 369	472

TABLE 5.14 – Moyenne, écart type, minimum et maximum des préditions de primes pures sur l'échantillon test en fonction du modèle

Le tableau 5.14 synthétise les paramètres statistiques des primes pures sur l'échantillon test modélisées par le modèle GLM, *CART*, *XGBoost* et *MOB*. Dans un premier temps, nous observons que le modèle *XGBoost* a des primes pures très élevées par rapport aux autres modèles. Cet écart s'explique par des prédictions élevées de la fréquence annuelle de sinistres. Le modèle *CART* a des montants plutôt homogènes au vu de l'écart type s'expliquant par la forte mutualisation. En effet, les modèles *CART* n'ont seulement 14 feuilles pour le modèle fréquence et 4 feuilles pour le modèle coût. Le peu nombre de feuilles entrainent une forte mutualisation et des primes pures proches. Pour finir, les modèles GLM et *MOB* ont des distributions de primes pures très proches. Ces différents éléments permettent de s'apercevoir que la tarification sur une grille ne peut pas être généralisée à toutes les modalités car chaque modèle donne des mauvais et bons risques différents.

Conclusion

À travers ce chapitre, nous avons observé que les modèles GLM, *CART* et *MOB* permettent d'expliquer ou quantifier les variations de prime pure à partir des sorties des modèles. L'explication du modèle *XGBoost* se fait par l'utilisation de l'algorithme de *SHAP* sur le modèle. De plus, nous observons que des disparités entre les niveaux de risque existent. Par exemple, en ayant fixé certaines modalités, pour les modèles *CART*, *XGBoost* et *MOB*, les plus mauvais risques sont les assurés avec une formule auto égale à 3 ou 4 et une formule auto égale à 1 est un très bon risque. À contrario, pour le modèle GLM une formule auto égale à 1 est le pire risque. Par ailleurs, la région de l'assuré n'a aucun impact dans le modèle *CART* mais dans les modèles GLM, *XGBoost* et *MOB*, une localisation dans la région associée à 2 est globalement moins risquée qu'une autre région. Cet exemple permet de comprendre le mécanisme d'interprétabilité de ces différents modèles et peut être généralisé afin d'expliquer tous les montants de primes pures.
Une combinaison de plusieurs modèles reste toujours intéressant afin d'avoir plusieurs points de vue sur les différents risques auquels nous sommes soumis et de pouvoir prendre les décisions adéquates.

Conclusion

Dans le cadre de l'amélioration des prédictions et de l'interprétabilité de la prime pure d'un assuré en assurance automobile, ce mémoire a présenté le modèle *MOB*. Ce modèle segmente le jeu de données et utilise un modèle paramétrique sur chaque segment. Pour le modèle paramétrique, nous avons choisi d'utiliser un modèle GLM car son interprétabilité est très simple et engendre une très bonne connaissance des différents risques. La segmentation en amont du jeu de données permet de prendre en compte de la non-linéarité des données pour ensuite utiliser des modèles GLM locaux sur chacun des segments. Ce processus permet d'améliorer la prédiction par rapport au modèle GLM classique car la non linéarité d'un jeu de données n'est pas capté par le modèle GLM. Naturellement, nous pouvons penser qu'une segmentation très fine entraine une meilleure performance. Cet instinct est faux car trop segmenter son jeu de données diminue le nombre d'observations par segmentation et empêche l'application des lois statistiques essentielles à la modélisation. De plus, sur des données hétérogènes, ce mémoire a montré que le modèle *MOB* est très performant. Les enjeux d'amélioration de ce modèle sont l'optimisation du choix des variables de régression et de classification et l'utilisation des différents paramètres similaires à ceux du modèle *CART*. Nous avons proposé une approche d'optimisation du choix des variables de régression et de classification mais les temps de calculs sont très élevés.
Une perspective d'évolution de ce mémoire serait d'utiliser ce modèle sur une base de données exhaustive afin d'observer si les différentes segmentations apportent une meilleure performance que les méthodes classiques et *machine learning*. Des contraintes opérationnelles liées au temps de calcul peuvent être rencontrées.

Bibliographie

[1] A.C. ACOCK et G.R. STAVIG. "A measure of association of Nonparametric Statistics". In : *Social Force* (1979), p. 1381-1386.

[2] J. BAI et P. PERRON. "Computation and Analysis of Multiple Structural Change Models". In : *Journal of Applied Econometrics* 2 (2003), p. 1-22. DOI : 10.1002/jae.659.

[3] L. BREIMAN. "Bagging predictors". In : *Machine Learning* 2 (1996), p. 123-140.

[4] L. BREIMAN. "Random forests". In : *Machine Learning* (2001).

[5] L. BREIMAN et al. *Classification And Regression Trees*. ISBN0412048418. Londres : Chapman et Hall/CRC, 1984.

[6] K.P. BURNHAM et D.R. ANDERSON. "Multimodel inference : understanding AIC and BIC in Model Selection". In : *Sociological Methods and Research* (1999), p. 261-304.

[7] A. CHARPENTIER et M. DENUIT. *Mathématiques de l'assurance non vie*. 2005.

[8] T. CHEN et C. GUESTRIN. "XGBoost : A Scalable Tree Boosting System". In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA : ACM, 2016, p. 785-794. ISBN : 978-1-4503-4232-2. DOI : 10.1145/2939672.2939785. URL : http://doi.acm.org/10.1145/2939672.2939785.

[9] T. CHEN et al. *xgboost : Extreme Gradient Boosting*. R package version 1.3.2.1. 2021. URL : https://CRAN.R-project.org/package=xgboost.

[10] *Code référencé SRA*. URL : https ://www.sra.asso.fr code de référencé SRAl.

[11] J. CUPE. "L'interprétabilité de l'IA". In : *Le nouveau défi des data scientists ActuIA* (2018).

[12] D. DELCAILLAU. "Contrôle et Transparence des modèles complexes en actuariat. Mémoire d'actuariat. EURIA". In : (2019).

[13] H. DRUCKER. "Improving regressors using boosting techniques". In : *Machine Learning : proceedings of the 14th International Conference*. San Francisco, California, USA : Morgan Kaufman, 1996, p. 107-15.

[14] B. DUBUISSON. *Solidarité, segmentation et discrimination en assurances, nouveau débat, nouvelles questions*. URL http ://hdl.handle.net/2078.1/83781. 2008.

[15] Y. FREUND et R.E. SCHAPIRE. "A Short Introduction to Boosting". In : *Journal of Japanese Society and Artificial Intelligence* 14 (1999), p. 771-780.

[16] J.H FRIEDMAN. "Stochastic gradient boosting". In : *Computational Statisrics and Data Analysis* 38 (2002), p. 367-378.

[17] Y. FRUEND et R.E. SCHAPIRE. "Experiments with a new boosting algorithm". In : *Machine Learning : proceedings of the Thirteenth International Conference*. San Francisco, California, USA : Morgan Kaufman, 1996, p. 1485-156.

[18] R. HASTIE, R. TIBSHIRANI et J. H. FRIEDMAN. *The Elements of Statistical Learning*. Springer, 2003.

[19] R.R. HOSKING. "The Analysis and Selection of Variables in Linear Regression". In : *Biometric* (1976), p. 261-304.

[20] T. HOTHORN, K. HORNIK et A. ZEILEIS. "Unbiased Recursive Partitioning : A Conditional Inference Framework". In : *Journal of Computational and Graphical Statistics* 15.3 (2006), p. 651-674. DOI : 10.1198/106186006X133933.

[21] T. HOTHORN et A. ZEILEIS. "partykit : A Modular Toolkit for Recursive Partytioning in R". In : *Journal of Machine Learning Research* 16 (2015), p. 3905-3909. URL : https://jmlr.org/papers/v16/hothorn15a.html.

[22] M. KUHN. *caret : Classification and Regression Training*. R package version 6.0-86. 2020. URL : https://CRAN.R-project.org/package=caret.

[23] M. KUHN et K. JOHNSON. *Applied Predictive Modeling*. Springer, 2013.

[24] G. LEROY et F. PLANCHET. "Un regard actuariel sur les évolutions de l'assurance automobile." In : *Risques* 105 (2016).

[25] Y. LIU et A. JUST. *SHAPforxgboost : SHAP Plots for 'XGBoost'*. R package version 0.1.0. 2020. URL : https://github.com/liuyanguu/SHAPforxgboost/.

[26] S. LUNDBERG et S.I. LEE. "A unified approach to interpreting model predictions". In : *31st Conference on Neural Information Processing Systems*. Long Beach, CA, USA, 2017, p. 4765-4774.

[27] M. DE LUSSAC. "Analyse du coût de sinistre matériels en assurance automobile. Mémoire d'actuariat. Université Paris-Dauphine". In : (2018).

[28] T. MACK. "Distribution-free calculation of the standard error of chain-ladder reserve estimates". In : *ASTIN Bulletin* 23 (1993), p. 213-225.

[29] J.A. NELDER et W.R.M. WEDDERBRUN. "Generalized Linear Models." In : *Journal of the Royal Statistical Society. Series A (General)* 135 (1972), p. 370-384.

[30] *Références statistiques sur les 127 vies sauvées en 2018*. URL : http//www.lci.fr/population/securite-routiere-la-mortalite-au-plus-bas-en-2018-la-limitation-a-80-km-h-aurait-sauve-127-vies-2122569.html.

[31] M. T. RIBEIRO, S. SINGH et C. GUESTRIN. "Explaining the Predictions of Any Classifier". In : *arXiv* (2016).

[32] RSTUDIO TEAM. *RStudio : Integrated Development Environment for R*. RStudio, PBC. Boston, MA, 2020. URL : http://www.rstudio.com/.

[33] T. RUSCH et A. ZEILEIS. "Gaining insight with recursive partitioning of generalized linear models". In : *Journal of Statistical Computation and Simulation* 83.7 (2013), p. 1301-1315.

[34] L. S. SHAPLEY. "A Value for n-Person Games". In : *Contribution to the Theory of Games* (1953), p. 303-317.

[35] *Statistiques de l'ensemble des assurances par la fédération française des assurances*. URL : http//www.ffa-assurance.fr.

[36] T. THERNEAU et B. ATKINSON. *rpart : Recursive Partitioning and Regression Trees*. R package version 4.1-15. 2019. URL : https://CRAN.R-project.org/package=rpart.

[37] P. VAISSIE, A. MONGE et F. HUSSON. *Factoshiny : Perform Factorial Analysis from 'FactoMineR' with a Shiny Application*. R package version 2.4. 2021. URL : https://CRAN.R-project.org/package=Factoshiny.

[38] W. N. VENABLES et B. D. RIPLEY. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York : Springer, 2002. URL : https://www.stats.ox.ac.uk/pub/MASS4/.

[39] T. Wei et V. Simko. *R package "corrplot" : Visualization of a Correlation Matrix.* (Version 0.84). 2017. URL : https://github.com/taiyun/corrplot.

[40] L. WILKINSON et G.E. DALLAL. "Tests of significance in forward selection regression with an F-to enter stopping rule." In : *Technometric* 23 (1981), p. 377-80.

[41] A. Zeileis. "A Unified Approach to Structural Change Tests Based on ML Scores, F statistics and OLS Residuals". In : *Econometric Reviews* 24.4 (2005), p. 445-466.

[42] A. Zeileis et K. Hornik. "Generalized M-Fluctuation Tests for Parameter Instability". In : *Statistica Neerlandica* 61.4 (2007), p. 488-508. DOI : 10.1111/j.1467-9574.2007.00371.x.

[43] A. Zeileis, T. Hothorn et K. Hornik. "Model-Based Recursive Partitioning". In : *Journal of Computational and Graphical Statistics* 17.2 (2008), p. 492-514. DOI : 10.1198/106186008X319331.

Annexe A

Annexes

A.1 Volumétrie des variables

Variable	Modalité	Fréquence
auto_classe_gta	B	37 277
	C	21 806
	D	17 670
auto_code_tarif	-10	29 048
	-20	67 024
	-30	7 657
auto_conducteur_situfami	C	25 117
	M	50 790
	U	22 304
auto_dpt_garage	60	8 052
	80	77 662
	AUTRES	1 2731
auto_elite	NR	32 041
	E ou N	76 888
auto_formule	2	26 384
	3	35 590
	4	22 973
auto_fractionnement	A	23 354
	M	71 568
	S	13 807
auto_franchise_bg	0	108 053
	1	676
auto_groupe_gta	10	15 989
	11	18 597
	13	16 573
auto_retro_exterieurs		50 939
	N	48 904
	O	88 86
code_profession	6	34 090
	7	25 813
	8	15 103

TABLE A.1 – Table détaillée des variables qualitatives 1/2

Variable	Modalité	Fréquence
auto_toit_panoramique		47 883
	N	5 9376
	O	1 670
auto_type_assistance	0	8 227
	1	38 223
	2	62 279
auto_vehicule_cv	4 et moins	18 510
	5	25 260
	6	29 068
auto_vehicule_marque	CITR	15 163
	PEUG	23 169
	RENA	31 517
auto_zone	2	96 194
	AUTRES	12 535

TABLE A.2 – Table détaillée des variables qualitatives 2/2

Les variables en gras sur les tableaux A.1 et A.2 correspondent aux variables dont toutes les modalités ainsi que leur volumétrie sont renseignées. Concernant les autres variables, nous avons renseigné les 3 modalités les plus représentatives. Les autres modalités des variables et leur volumétrie sont présentées en intégralité dans la partie 2.2.

Nombre de sinistres	Nombre d'observations
0	74180
1	19759
2	5860
3	1707
4	566
5	172
6	56
7	29
8	11
9	3
10	2
11	3
12	1

TABLE A.3 – Nombre d'observations en fonction du nombre de sinistres

A.2 Analyse statistiques des variables

Code tarif

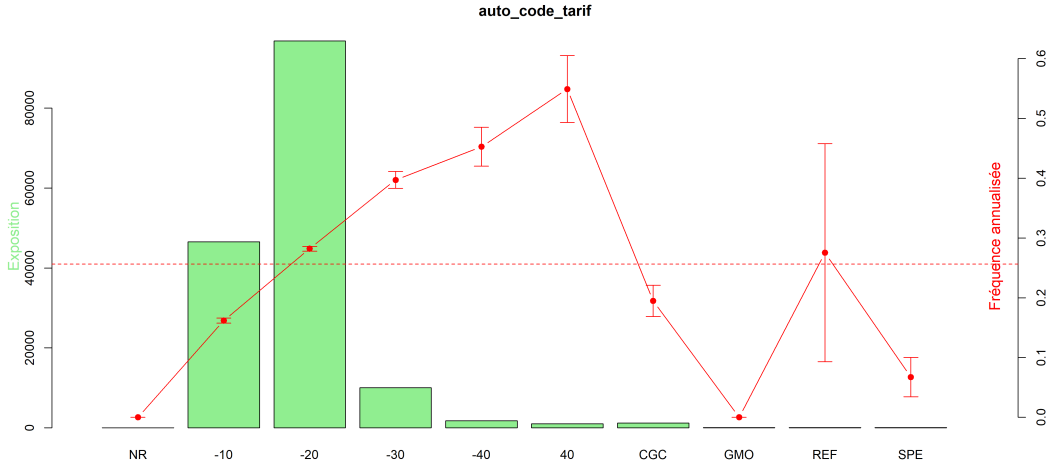

FIGURE A.1 – Fréquence annuelle des sinistres en fonction du code tarif.

Dans ce graphique, on observe qu'hormis les modalités -30, -20 et -10 les autres modalités sont sous représentés. Les regroupements vont permettre de créer des classes plus homogènes. Les regroupements effectués sont :

— -40 et -30
— -20 et REF
— -10, CGC, GMO, NR et SPE

La modalité "40" présente une fréquence de sinistre forte qu'on décide de ne pas regrouper. Les modalités GMO et NR ne représente que 76,26 d'exposition annuelle et aucun sinistre. Ces modalités sont regroupées par défaut avec la modalité possédant une fréquence de sinistre faible tout en étant bien représenté.
Les fréquences de sinistres oscillent entre 0 et 0,6 pour les différentes modalités de cette variable.

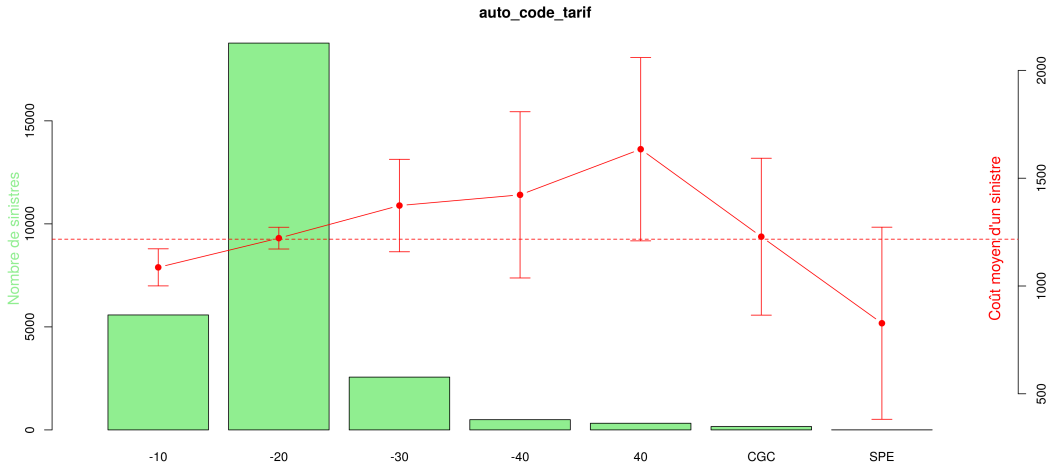

FIGURE A.2 – Coût moyen d'un sinistre en fonction du code tarif.

La modalité REF n'est pas représenté car son coût moyen pour un sinistre est de 6156 €. Ceci est dû à un sinistre extrême coutant 80 000 €. Si ce sinistre n'est pas comptabilisé le coût moyen d'un sinistre pour cette modalité est de 1233 €. Les regroupements effectués précédemment sont en adéquation avec le profil du coût moyen des montants d'un sinistre. Ces montants sont compris entre 827 € et 1635 € hormis la modalité "REF" avec le sinistre extrême.

Ancienneté permis effet

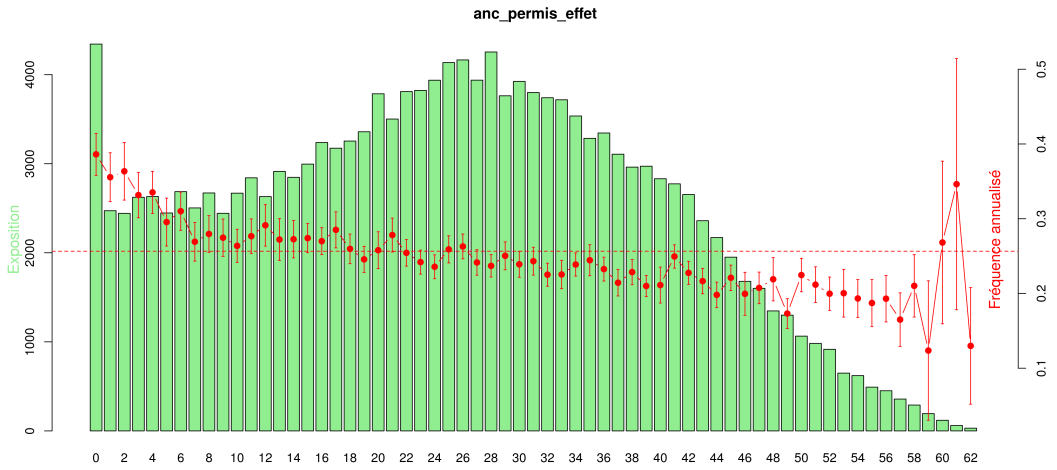

FIGURE A.3 – Fréquence annuelle des sinistres en fonction des modalités de la variable ancienneté permis effet.

Globalement, la fréquence annualisée des sinistres décroit de 0 à 57 ans pour osciller après 57 ans. Intuitivement, ce résultat semble cohérent car une personne plus expérimentée fera moins d'accident. La fréquence est comprise entre 0,19 et 0,39. Aucun regroupement n'est effectué.

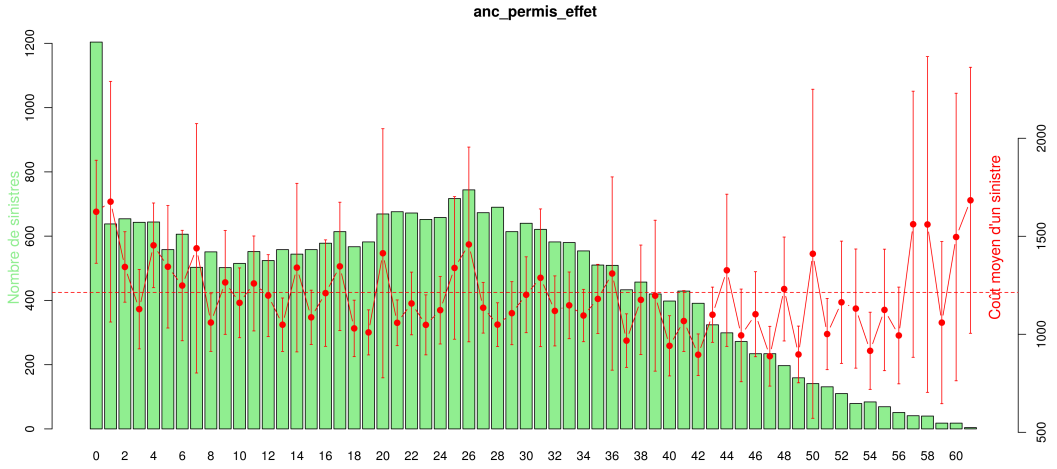

FIGURE A.4 – Coût moyen d'un sinistre attritionnel en fonction des modalités de la variable ancienneté permis effet.

Les coûts moyens d'un sinistre en fonction de l'ancienneté du permis ne sont pas monotone. Les coûts moyens sont compris entre 888 € et 12 401 € qui n'est pas répresenté sur le graphique.

Ancienneté contrat

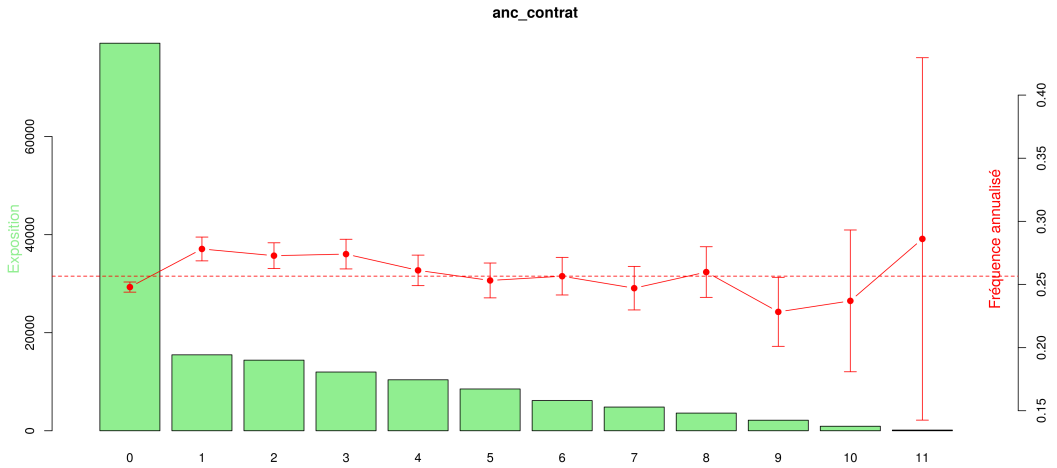

FIGURE A.5 – Fréquence annuelle des sinistres en fonction des modalités de la variable ancienneté contrat.

La fréquence annuelles des sinistres en fonction de l'ancienneté du contrat n'est pas monotone. Elle oscille entre 0,23 et 1,03 non représenté sur le graphique.

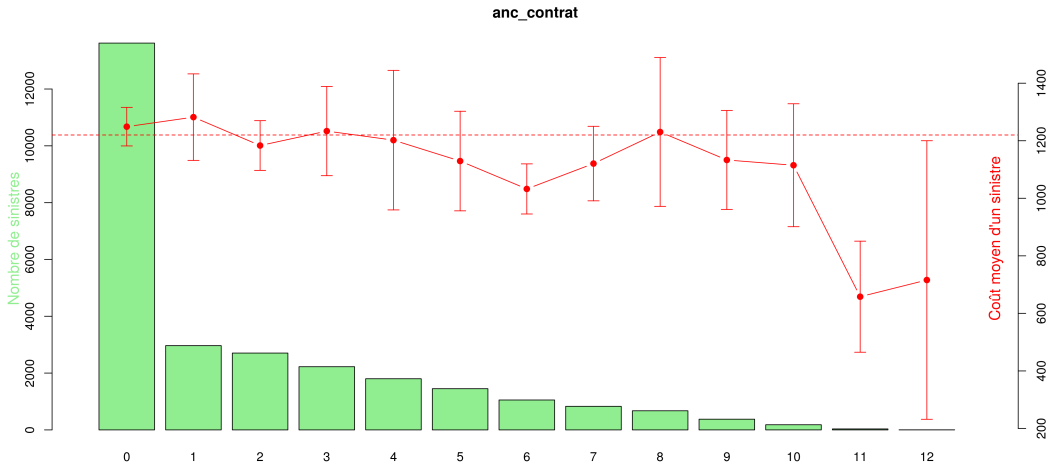

FIGURE A.6 – Coût moyen d'un sinistre en fonction des modalités de la variable ancienneté contrat.

Le coût moyen d'un sinistre en fonction de l'ancienneté du contrat a une légère décroissance. Le coût moyen d'un sinistre est compris entre 658 € et 1249 € .

Auto département garage

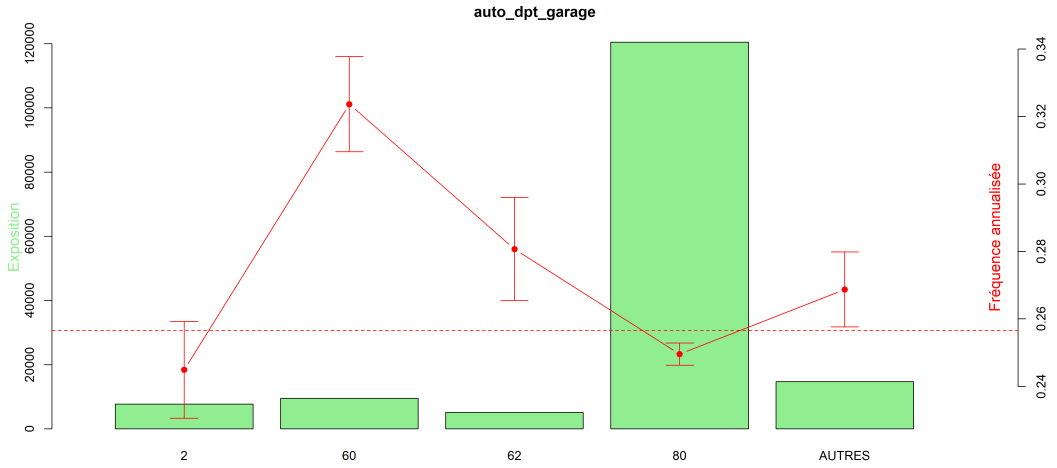

FIGURE A.7 – Fréquence annuelle des sinistres en fonction des modalités de la variable auto_dpt_garage

La fréquence annuelle des sinistres des différentes modalités prenne des valeurs entre 0,25 et 0,33. Aucun regroupement n'est effectué car aucune modalité n'est réellement sous représentée et le nombre de modalité est faible.

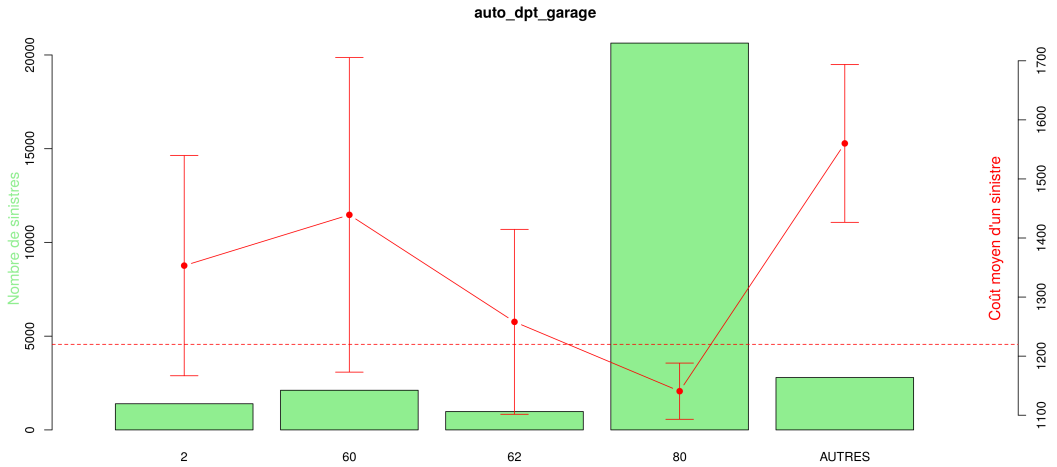

FIGURE A.8 – Coût moyen annuel d'un sinistre attritionnel en fonction des modalités de la variable auto_dpt_garage

Les coûts moyens d'un sinistre sont compris entre 1141 € et 1560 €.

Auto vehicule cv

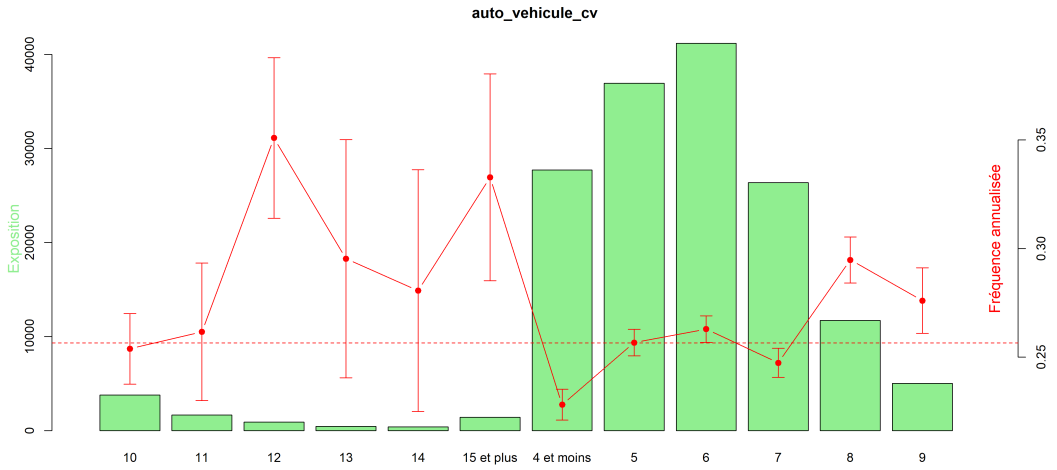

FIGURE A.9 – Fréquence annuelle des sinistres en fonction du nombre de chevaux fiscaux

La fréquence annuelle des sinistres des différentes modalités prenne des valeurs entre 0,23 et 0,34. Certaines modalités sont sous représentées. Pour pallier cette contrainte, nous créeons des modalités regroupement des modalités ayant des fréquences de sinistres similaires. Ces regroupements sont :

- 10 et 11
- 12 et 15 et plus
- 13, 14, 8 et 9

Avec ces regroupements, aucune nouvelle modalité n'est sous representée.

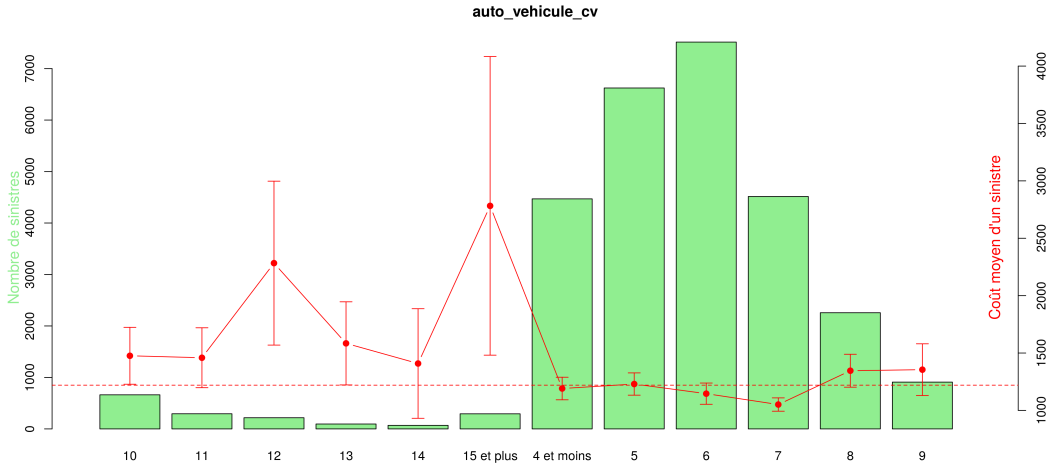

FIGURE A.10 – Coût moyen annuel d'un sinistre en fonction du nombre de chevaux fiscaux

Les coûts moyens d'un sinistre sont compris entre 1 051 € et 1 584 €. Les regroupements faits précedemment sont cohérents avec la distribution des coûts moyens.

auto zone

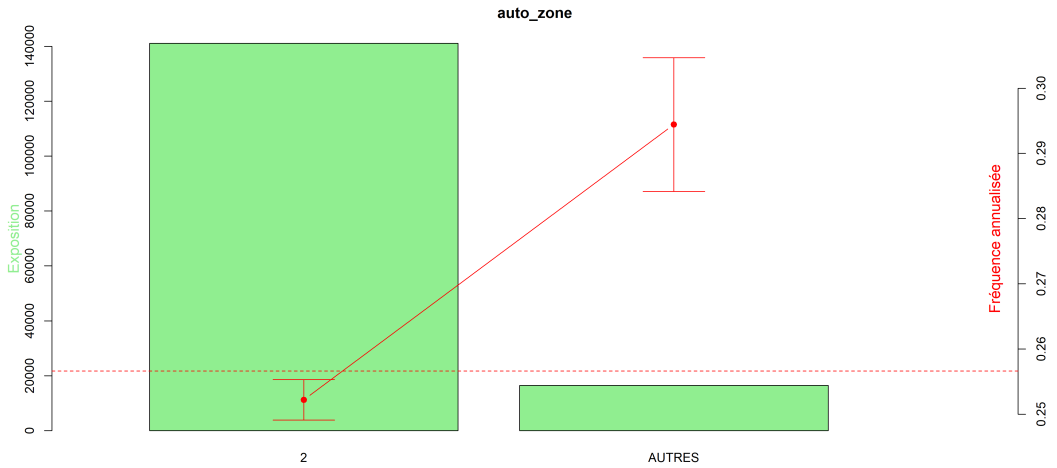

FIGURE A.11 – Fréquence annuelle des sinistres en fonction des modalités de la variable auto zone

La fréquence annuelle des sinistres des différentes modalités prenne les valeurs 0,25 et 0,29.

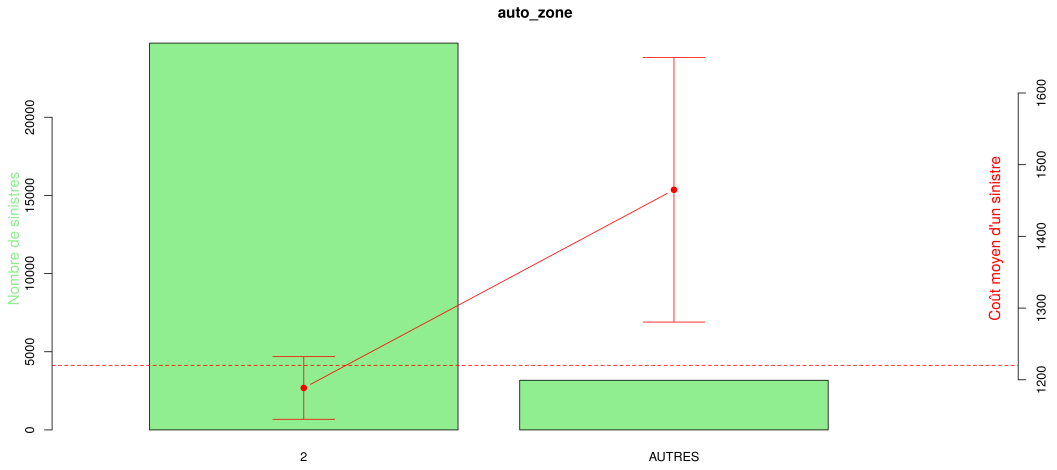

FIGURE A.12 – Figure 16 - Coût moyen annuel d'un sinistre attritionnel en fonction des modalités de la variable auto zone

Les coûts moyens d'un sinistre sont égaux à 1189 € et 1465 €.

Auto conducteur situation famiale

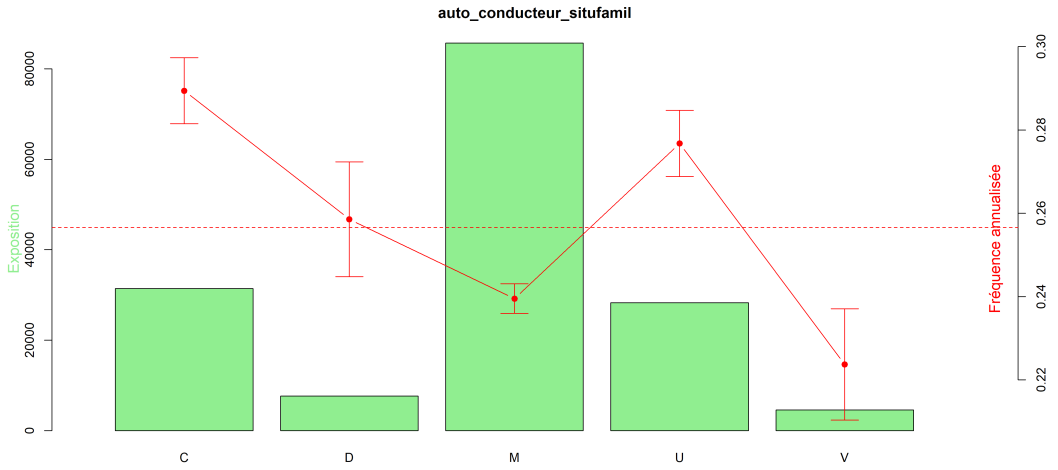

FIGURE A.13 – Fréquence annuelle des sinistres en fonction des modalités de la variable auto conducteur situation famiale

La fréquence annuelle des sinistres des différentes modalités prenne des valeurs entre 0,22 et 0,29. Aucun regroupement n'est effectué car aucune modalité n'est réellement sous représentée et le nombre de modalité est faible.

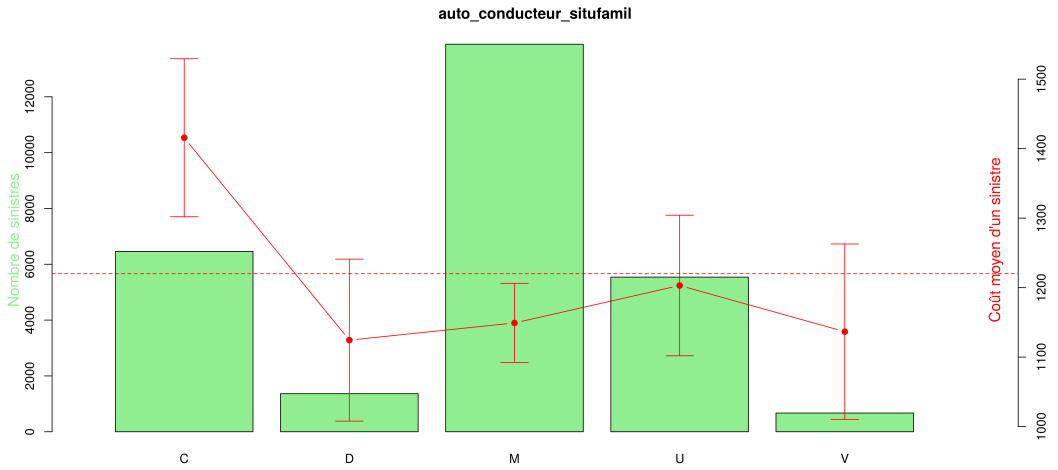

FIGURE A.14 – Coût moyen annuel d'un sinistre attritionnel en fonction des modalités de la variable auto conducteur situation famiale

Les coûts moyens d'un sinistre sont compris entre 1124 € et 1415 €.

Auto elite

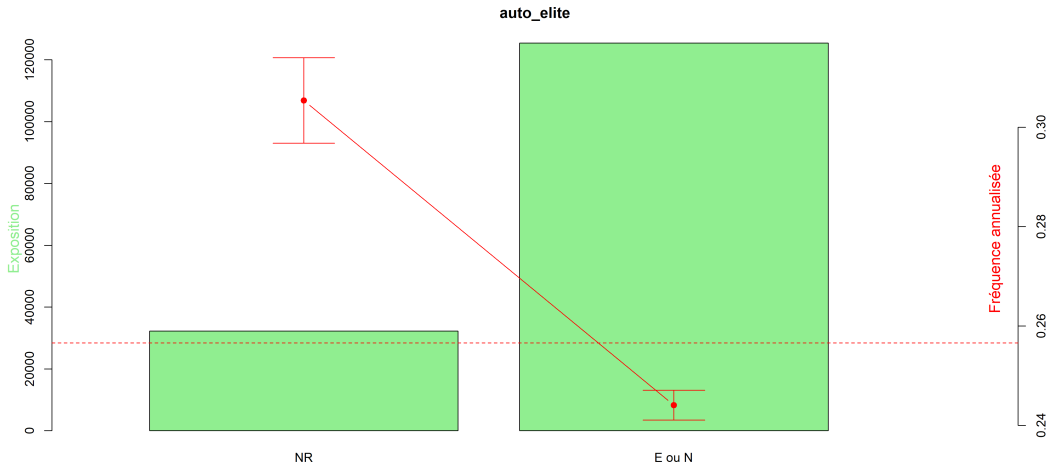

FIGURE A.15 – Fréquence annuelle des sinistres en fonction des modalités de la variable auto elite

La fréquence annuelle des sinistres des différentes modalités prenne les valeurs 0,24 et 0,30.

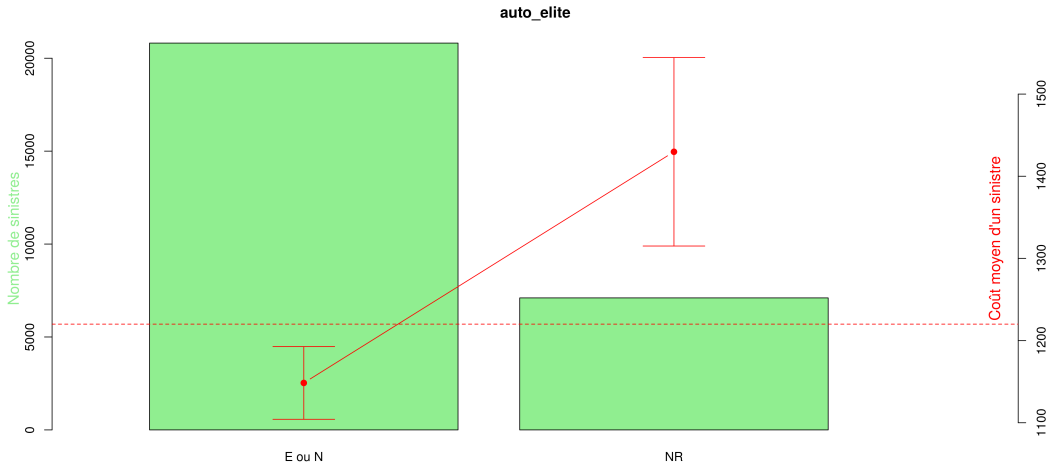

FIGURE A.16 – Coût moyen annuel d'un sinistre attritionnel en fonction des modalités de la variable auto elite

Les coûts moyens d'un sinistre sont égaux à 1148 € et 1429 €.

Auto franchise

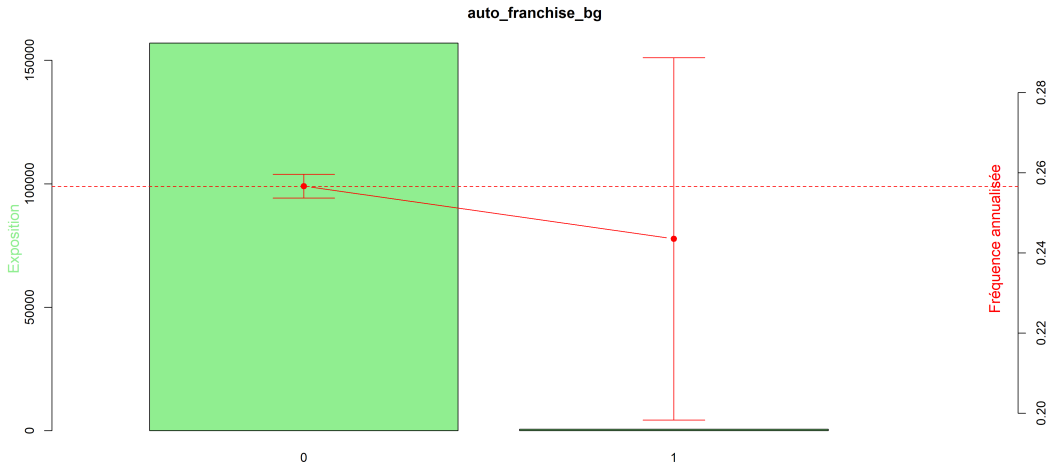

FIGURE A.17 – Fréquence annuelle des sinistres en fonction des modalités de la variable auto franchise

La fréquence annuelle des sinistres des différentes modalités prenne les valeurs 0,24 et 0,26. La modalité 1 est sous représentée avec une exposition annuelle égale à 657.

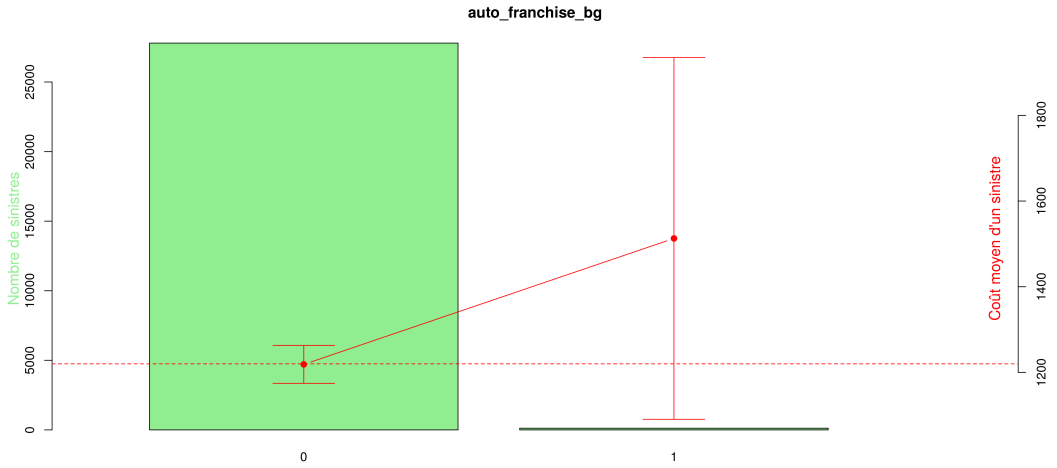

FIGURE A.18 – Coût moyen annuel d'un sinistre attritionnel en fonction des modalités de la variable auto franchise

Les coûts moyens d'un sinistre sont égaux à 1219 € et 1513 €.

Auto toit panoramique

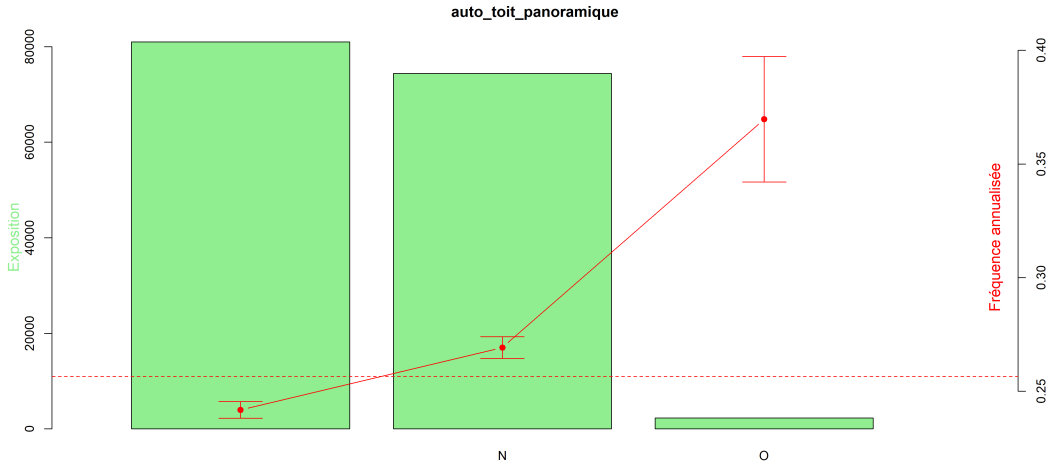

FIGURE A.19 – Fréquence annuelle des sinistres en fonction des modalités de la variable auto toit panoramique

La fréquence annuelle des sinistres des différentes modalités prenne les valeurs 0,24, 0,27 et 0,36.

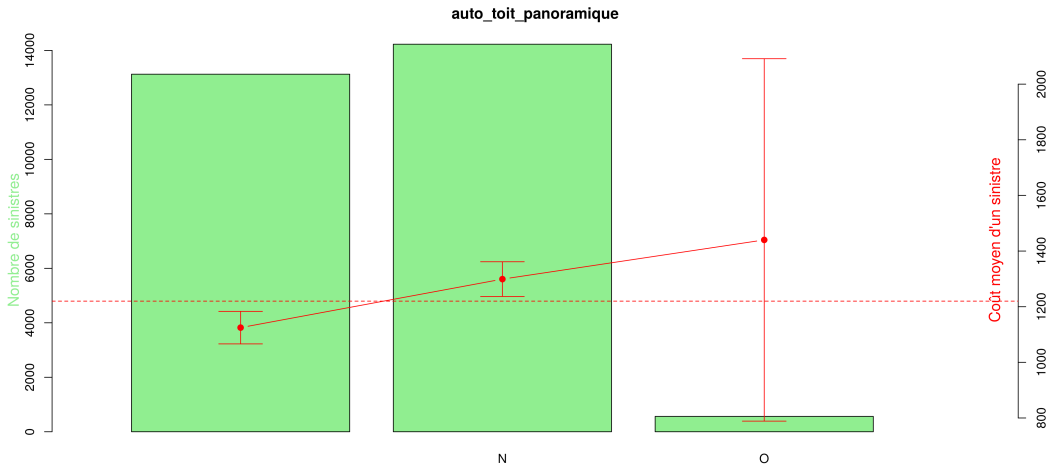

FIGURE A.20 – Coût moyen annuel d'un sinistre attritionnel en fonction des modalités de la variable auto toit panoramique

Les coûts moyens d'un sinistre sont égaux à 1124 €, 1299 € et 1440 €.

Auto retro exterieurs

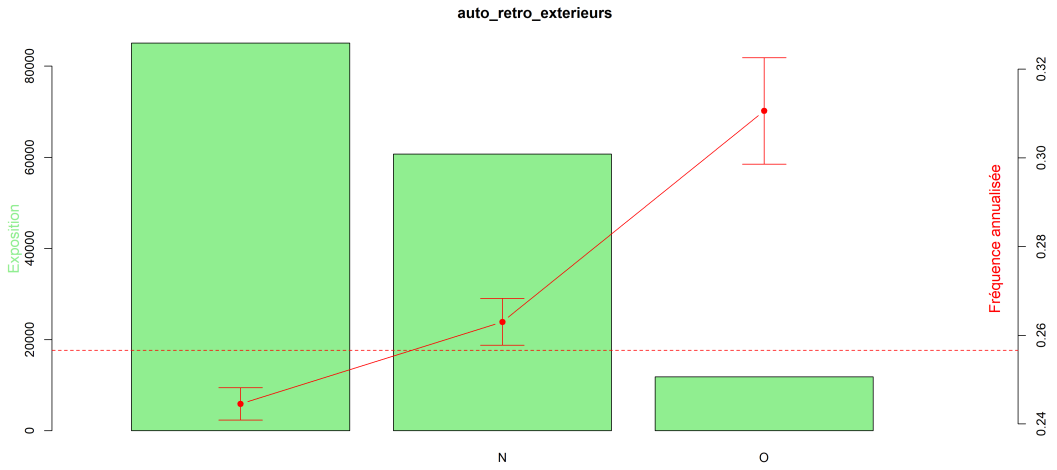

FIGURE A.21 – Fréquence annuelle des sinistres en fonction des modalités de la variable auto retro exterieurs

La fréquence annuelle des sinistres des différentes modalités prenne les valeurs 0,24, 0,27 et 0,31.

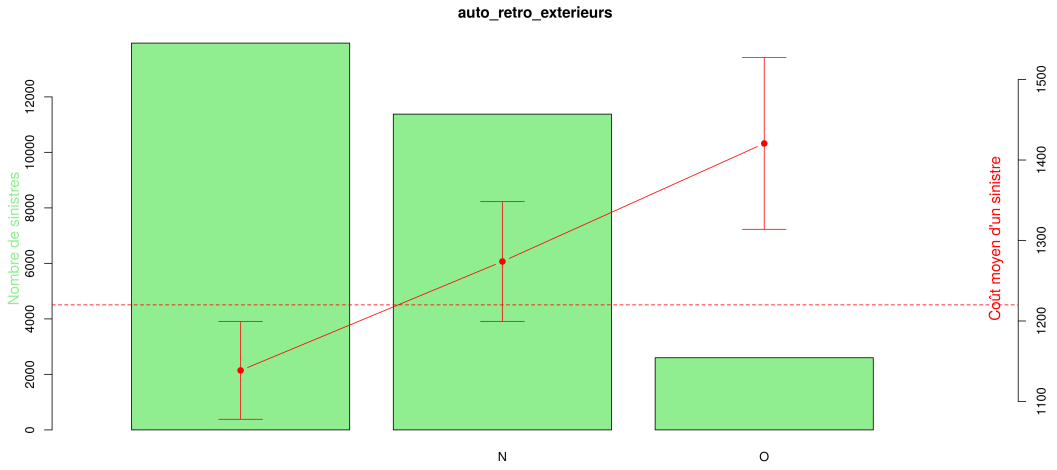

FIGURE A.22 – Coût moyen annuel d'un sinistre attritionnel en fonction des modalités de la variable auto retro exterieurs.

Les coûts moyens d'un sinistre sont égaux à 1139 €, 1274 € et 1420 €.

Auto type assistance

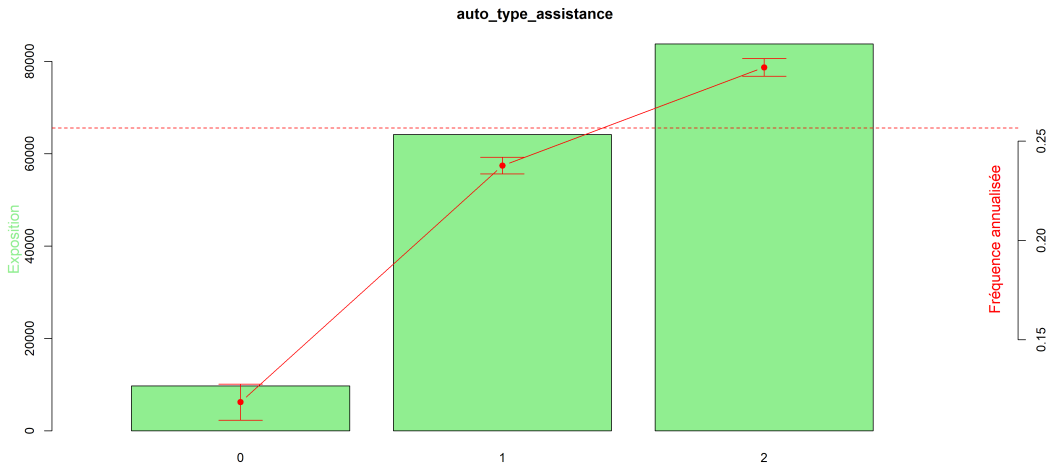

FIGURE A.23 – Fréquence annuelle des sinistres en fonction des modalités de la variable auto type assistance

La fréquence annuelle des sinistres des différentes modalités prenne les valeurs 0,12, 0,24 et 0,29.

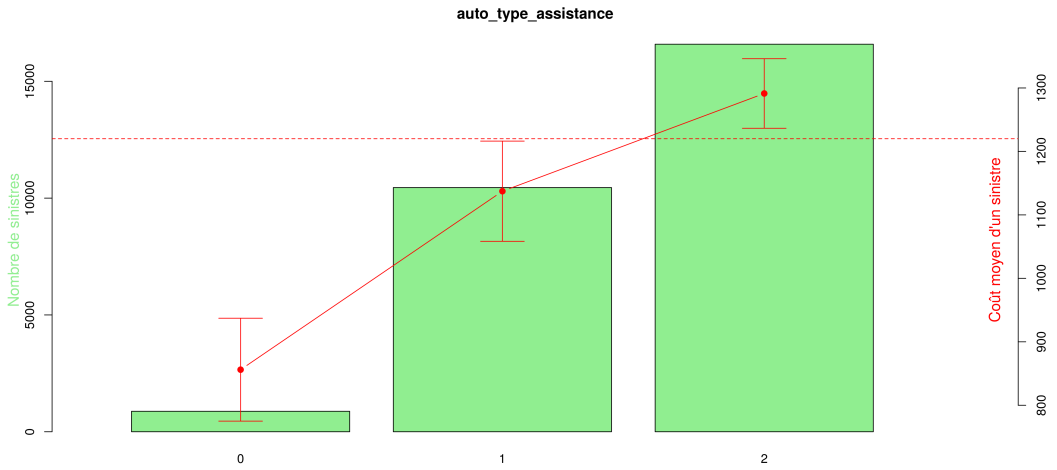

FIGURE A.24 – Coût moyen annuel d'un sinistre attritionnel en fonction des modalités de la variable auto type assistance

Auto fractionnement

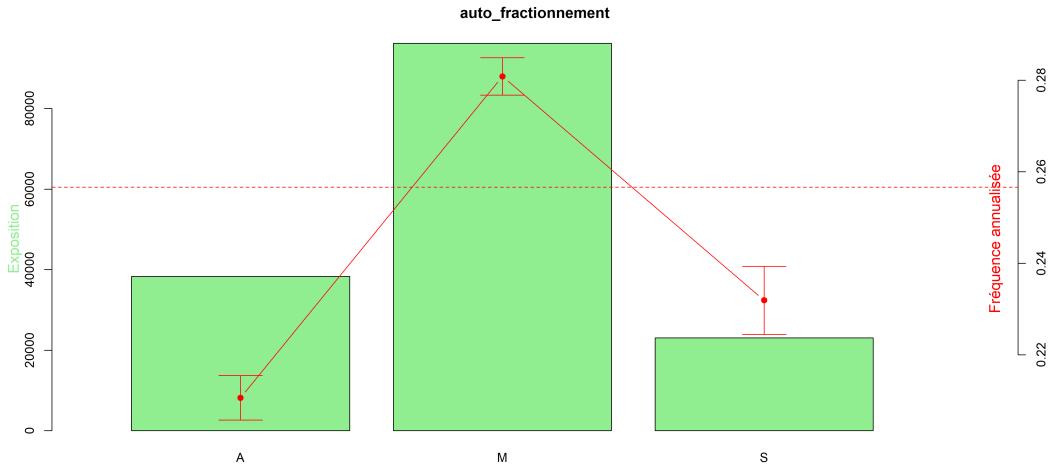

FIGURE A.25 – Fréquence annuelle des sinistres en fonction des modalités de la variable auto fractionnement

La fréquence annuelle des sinistres des différentes modalités prenne les valeurs 0.24 , 0.27 et 0.31.

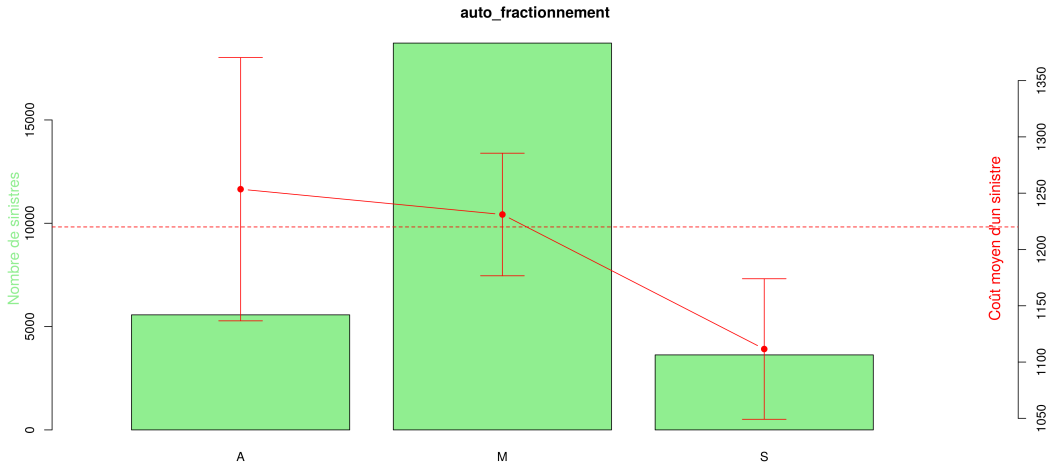

FIGURE A.26 – Coût moyen annuel d'un sinistre attritionnel en fonction des modalités de la variable auto fractionnement

Les coûts moyens d'un sinistre sont égaux à 1254 € , 1231 € et 1112 €.

A.3 Sortie des feuilles terminales du modèle *MOB* fréquence

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2,6343	17,4535	-0,1509	0,8800
auto_vehicule_marqueGroupe1	0,0640	0,0388	1,6527	0,0984
auto_vehicule_marqueGroupe2	-0,0513	0,0228	-2,2443	0,0248
auto_vehiculecvGroupe1	-0,0656	0,0528	-1,2421	0,2142
auto_vehiculecvGroupe2	-0,1750	0,0724	-2,4160	0,0157
auto_vehiculecvGroupe3	-0,0554	0,0317	-1,7502	0,0801
auto_vehiculecvGroupe4	0,1529	0,0366	4,1734	0,0000
auto_groupe_gtaGroupe1	-1,0562	0,1949	-5,4196	0,0000
auto_groupe_gtaGroupe2	-0,0298	0,0369	-0,8092	0,4184
auto_groupe_gtaGroupe3	0,1110	0,0347	3,1997	0,0014
auto_groupe_gtaGroupe4	0,1509	0,0366	4,1205	0,0000
auto_groupe_gtaGroupe5	0,2184	0,0366	5,9632	0,0000
auto_groupe_gtaGroupe6	0,3015	0,0384	7,8538	0,0000
auto_groupe_gtaGroupe7	0,3888	0,0481	8,0889	0,0000
auto_formule1	0,5366	17,4535	0,0307	0,9755
auto_formule2	1,4679	17,4534	0,0841	0,9330
auto_formule3	1,6704	17,4534	0,0957	0,9238
auto_formule4	2,0436	17,4534	0,1171	0,9068
auto_formule5	1,4134	17.4534	-0.812	0,9349
auto_zone1	-0,0676	0,0177	-3,8291	0,0001
auto_code_tarifGroupe1	-0,3539	0,0309	-11,4394	0,0000
auto_code_tarifGroupe2	-0,1491	0,0256	-5,8230	0,0000
auto_code_tarifGroupe3	0,1182	0,0328	3,6028	0,0003
auto_conducteur_situfamil1	0,0211	0,0274	0,7712	0,4406
auto_conducteur_situfamil2	0,0359	0,0411	0,8739	0,3822
auto_conducteur_situfamil3	-0,0343	0,0218	-1,5766	0,1149
auto_conducteur_situfamil4	-0,0451	0,0276	-1,6343	0,1022
auto_elite1	-0,1372	0,0140	-9,8072	0,0000
auto_toit_panoramique1	-0,0895	0,0441	-2,0305	0,0423
auto_toit_panoramique2	-0,0232	0,0345	-0,6716	0,5018
auto_type_assistance1	-0,0599	0,0454	-1,3190	0,1872
auto_type_assistance2	0,0792	0,0294	2,6979	0,0070
auto_fractionnement1	-0,0748	0,0188	-3,9792	0,0001
auto_fractionnement2	0,1099	0,0163	6,7414	0,0000
anc_permis_effet	-0,0100	0,0011	-8,8122	0,0000

TABLE A.4 – Coefficients de régression linéaire de la première feuille (à gauche) du modèle MOB fréquence

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2,6725	28,7759	-0,0929	0,9260
auto_vehicule_marqueGroupe1	0,0748	0,0510	1,4661	0,1426
auto_vehicule_marqueGroupe2	-0,0615	0,0306	-2,0114	0,0443
auto_vehiculecvGroupe1	0,0135	0,0754	0,1785	0,8583
auto_vehiculecvGroupe2	0,0074	0,0966	0,0767	0,9388
auto_vehiculecvGroupe3	-0,1121	0,0458	-2,4493	0,0143
auto_vehiculecvGroupe4	-0,0006	0,0533	-0,0113	0,9909
auto_groupe_gtaGroupe1	-1,0458	0,3705	-2,8228	0,0048
auto_groupe_gtaGroupe2	-0,0784	0,0620	-1,2638	0,2063
auto_groupe_gtaGroupe3	0,1636	0,0564	2,8999	0,0037
auto_groupe_gtaGroupe4	0,1167	0,0596	1,9589	0,0501
auto_groupe_gtaGroupe5	0,1535	0,0586	2,6218	0,0087
auto_groupe_gtaGroupe6	0,2975	0,0598	4,9733	0,0000
auto_groupe_gtaGroupe7	0,3161	0,0722	4,3783	0,0000
auto_formule1	0,7156	28,7758	0,0249	0,9802
auto_formule2	1,6008	28,7758	0,0556	0,9556
auto_formule3	1,6810	28,7758	0,0584	0,9534
auto_formule4	1,9797	28,7758	0,0688	0,9452
auto_formule5	1.5596	28,7758	0,0545	0,9582
auto_zone1	-0,0631	0,0247	-2,5508	0,0107
auto_code_tarifGroupe1	-0,3555	0,0432	-8,2269	0,0000
auto_code_tarifGroupe2	-0,1361	0,0357	-3,8147	0,0001
auto_code_tarifGroupe3	0,0897	0,0446	2,0135	0,0441
auto_conducteur_situfamil1	0,0200	0,0405	0,4936	0,6216
auto_conducteur_situfamil2	0,0416	0,0596	0,6979	0,4853
auto_conducteur_situfamil3	-0,0156	0,0335	-0,4658	0,6414
auto_conducteur_situfamil4	-0,0212	0,0381	-0,5561	0,5782
auto_elite1	-0,1633	0,0193	-8,4769	0,0000
auto_toit_panoramique1	0,0284	0,0552	0,5150	0,6065
auto_toit_panoramique2	0,0295	0,0450	0,6564	0,5116
auto_type_assistance2	-0,0875	0,0352	-2,4867	0,0129
auto_fractionnement1	-0,0529	0,0310	-1,7099	0,0873
auto_fractionnement2	0,0695	0,0246	2,8269	0,0047
anc_permis_effet	-0,0053	0,0017	-3,1262	0,0018

TABLE A.5 – Coefficients de régression linéaire de la deuxième feuille du modèle MOB fréquence

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2,6336	20,2102	-0,1303	0,8963
auto_vehicule_marqueGroupe1	0,0134	0,0384	0,3487	0,7273
auto_vehicule_marqueGroupe2	-0,0259	0,0225	-1,1491	0,2505
auto_vehiculecvGroupe1	-0,1309	0,0557	-2,3489	0,0188
auto_vehiculecvGroupe2	-0,0413	0,0694	-0,5956	0,5514
auto_vehiculecvGroupe3	-0,0586	0,0314	-1,8647	0,0622
auto_vehiculecvGroupe4	0,1009	0,0387	2,6042	0,0092
auto_groupe_gtaGroupe1	-0,4760	0,2425	-1,9626	0,0497
auto_groupe_gtaGroupe2	-0,0382	0,0432	-0,8847	0,3763
auto_groupe_gtaGroupe3	-0,0275	0,0409	-0,6722	0,5015
auto_groupe_gtaGroupe4	0,0513	0,0418	1,2277	0,2196
auto_groupe_gtaGroupe5	0,1203	0,0414	2,9037	0,0037
auto_groupe_gtaGroupe6	0,1949	0,0423	4,6113	0,0000
auto_groupe_gtaGroupe7	0,3206	0,0504	6,3554	0,0000
auto_formule1	0,3861	20,2103	0,0191	0,9848
auto_formule2	1,7294	20,2101	0,0693	0,9448
auto_formule3	1,6505	20,2101	0,0817	0,9349
auto_formule4	1,9858	20,2101	0,0983	0,9217
auto_formule5	1,6372	20,2101	0.0723	0,9469
auto_zone1	-0,0644	0,0162	-3,9766	0,0001
auto_code_tarifGroupe1	-0,3781	0,0331	-11,4093	0,0000
auto_code_tarifGroupe2	-0,0968	0,0280	-3,4616	0,0005
auto_code_tarifGroupe3	0,1647	0,0335	4,9114	0,0000
auto_conducteur_situfamil1	-0,0027	0,0272	-0,0992	0,9210
auto_conducteur_situfamil2	0,0092	0,0410	0,2238	0,8229
auto_conducteur_situfamil3	-0,0519	0,0213	-2,4421	0,0146
auto_conducteur_situfamil4	0,0240	0,0256	0,9364	0,3491
auto_elite1	-0,1302	0,0149	-8,7147	0,0000
auto_toit_panoramique1	-0,0043	0,0338	-0,1285	0,8978
auto_toit_panoramique2	-0,0608	0,0272	-2,2362	0,0253
auto_type_assistance1	-0,0130	0,0424	-0,3078	0,7582
auto_type_assistance2	-0,0059	0,0264	-0,2226	0,8238
auto_fractionnement1	-0,0506	0,0191	-2,6531	0,0080
auto_fractionnement2	0,0603	0,0156	3,8617	0,0001
anc_permis_effet	-0,0066	0,0012	-5,6781	0,0000

TABLE A.6 – Coefficients de régression linéaire de la troixième feuille du modèle MOB fréquence

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1,3610	30,9118	-0,0440	0,9649
auto_vehicule_marqueGroupe1	0,2440	0,1016	2,4014	0,0163
auto_vehicule_marqueGroupe2	-0,1719	0,0626	-2,7458	0,0060
auto_vehiculecvGroupe1	-0,0097	0,1507	-0,0645	0,9486
auto_vehiculecvGroupe2	-0,2411	0,2228	-1,0819	0,2793
auto_vehiculecvGroupe3	-0,0478	0,0985	-0,4853	0,6275
auto_vehiculecvGroupe4	0,0692	0,1257	0,5505	0,5820
auto_groupe_gtaGroupe1	-10,9060	278,1983	-0,0392	0,9687
auto_groupe_gtaGroupe2	1,1258	30,9112	0,0364	0,9709
auto_groupe_gtaGroupe3	1,1782	30,9111	0,0381	0,9696
auto_groupe_gtaGroupe4	1,2972	30,9111	0,0420	0,9665
auto_groupe_gtaGroupe5	1,4215	30,9111	0,0460	0,9633
auto_groupe_gtaGroupe6	1,3365	30,9111	0,0432	0,9655
auto_groupe_gtaGroupe7	1,5607	30,9113	0,0505	0,9597
auto_formule1	-2,5680	0,7165	-3,5841	0,0003
auto_formule2	-1,0886	0,3748	-5,3243	0,0002
auto_formule3	-0,9369	0,1146	-8,1721	0,0000
auto_formule4	-0,1884	0,0968	-1,9462	0,0516
auto_formule5	-1,0127	0,4343	-5,9605	0,0032
auto_zone1	0,0603	0,0640	0,9425	0,3459
auto_code_tarifGroupe1	-0,5520	0,1155	-4,7782	0,0000
auto_code_tarifGroupe2	-0,2280	0,0997	-2,2868	0,0222
auto_code_tarifGroupe3	-0,0311	0,1206	-0,2580	0,7964
auto_conducteur_situfamil1	-0,1500	0,0916	-1,6374	0,1015
auto_conducteur_situfamil2	0,1088	0,1279	0,8510	0,3948
auto_conducteur_situfamil3	0,1111	0,0626	1,7731	0,0762
auto_conducteur_situfamil4	-0,1821	0,0839	-2,1692	0,0301
auto_elite1	-0,2209	0,0541	-4,0861	0,0000
auto_toit_panoramique1	0,3346	0,1310	2,5549	0,0106
auto_type_assistance1	0,0601	0,1157	0,5193	0,6036
auto_fractionnement1	0,0318	0,0559	0,5687	0,5696
auto_fractionnement2	-0,0275	0,0478	-0,5757	0,5648
anc_permis_effet	-0,0152	0,0037	-4,1283	0,0000

TABLE A.7 – Coefficients de régression linéaire de la quatrième feuille du modèle MOB fréquence