Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 16/03/2022

Par : **Leslie GNANSOUNOU**

Titre : **Construction d'un véhiculier en assurance automobile à partir de méthodes de Machine Learning**

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

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

Membres présents du jury de la filière

Nom :

Membres présents du jury de l'Institut des Actuaires

Entreprise : Sia Partners

Signature :

Directeur du mémoire en entreprise :

Noms : Claire NICOLLE et Fabien CHERANCE
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

Secrétariat :

Signature du candidat

Bibliothèque :

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

Résumé

La segmentation tarifaire est aujourd'hui au cœur des enjeux de tarification en assurance. Secteur hyperconcurrentiel et en constante évolution, l'assurance IARD pousse ses acteurs à se démarquer sur tous les aspects du métier. Du côté de la tarification, les assureurs, afin de rester compétitifs, proposent des tarifs de plus en plus précis en différenciant leurs assurés par groupes de risques homogènes. Différents outils de mesures du risque existent et permettent à l'assureur d'individualiser au maximum le tarif. Par exemple, les zoniers utilisés en assurance multirisques habitation ainsi qu'en assurance automobile et les véhiculiers utilisés spécialement en assurance automobile.

Le véhiculier est une variable tarifaire regroupant de façon homogène, les caractéristiques de véhicules ayant les mêmes profils pour un risque donné. Plusieurs approches ont été développées au cours des dernières années pour construire des véhiculiers.

Ce mémoire propose une approche utilisant les méthodes de Machine Learning (ML), sur un portefeuille d'assurance couvrant la Responsabilité Civile Matérielle. Cette approche est basée sur la modélisation des résidus d'un modèle de tarification GLM, calibré sans les variables liées au véhicule. L'information non expliquée par le modèle GLM classique et contenue dans les résidus, sera modélisée grâce aux méthodes de Random Forest et de Gradient Boosting. Ces résidus modélisés seront par la suite clusterisés, afin de former des classes de véhicules, donc le véhiculier. Nous évaluerons l'apport et la pertinence de cette nouvelle variable à travers des analyses et des comparaisons de modèles.

Mots clés : Assurance non-vie, tarification, segmentation, classification de véhicules, Random Forest, Gradient Boosting.

Abstract

Pricing segmentation is at the heart of insurance pricing issues today. As a hyper-competitive and constantly evolving sector, property and casualty insurance pushes its players to stand out in all aspects of the business. In order to remain competitive, insurers, are offering increasingly precise rates by differentiating their policyholders by group of homogeneous risks. Different risk measurement tools exist and allow the insurer to individualize the rate as much as possible. For example, zoning variables are used in home insurance and in motor insurance, and vehicles classification are used in motor insurance.

The vehicle classification is a rating variable that groups together in a homogeneous way the characteristics of vehicles with the same profiles for a given risk. Several approaches have been developed over the last few years to construct these classifications.

This paper proposes an approach using Machine Learning (ML) methods, on an insurance portfolio covering Third Party Damage (TPD). This approach is based on the modeling of the residuals of a GLM pricing model, calibrated without the vehicle variables. The information not explained by the classical GLM model and contained in the residuals will be modeled using Random Forest and Gradient Boosting methods. These modeled residuals will then be clustered in order to form vehicle classes. We will evaluate the contribution and relevance of this new variable through analyses and model comparisons.

Keywords : Non-life insurance, pricing, segmentation,vehicle scoring, vehicle classification, Random Forest, Gradient Boosting.

Note de Synthèse

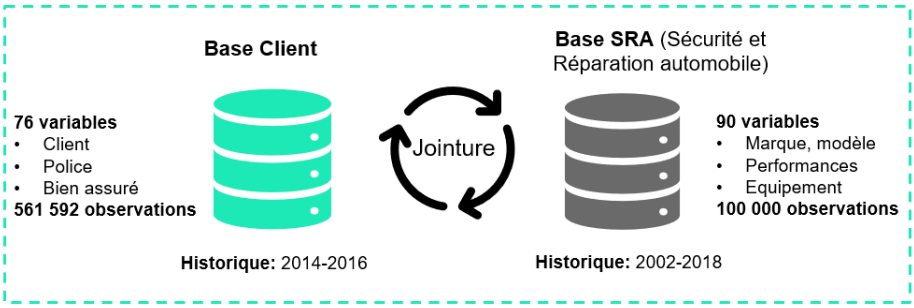
La segmentation tarifaire constitue un enjeu majeur en assurance IARD. Non seulement elle permet à l'assureur d'éviter l'antisélection en distinguant les bons risques des mauvais risques, mais elle aide également l'assureur à fidéliser sa clientèle. La segmentation peut être réalisée à travers plusieurs outils tels que les zoniers (classification géographique) ou encore les véhiculiers (classification de véhicules).

Ce mémoire est consacré à la construction d'un véhiculier, sur un portefeuille couvrant la Responsabilité Civile Matérielle (RCM), en s'appuyant sur les méthodes de Machine Learning. Les objectifs de l'étude sont multiples : affiner la segmentation et simplifier le modèle de tarification grâce au véhiculier, évaluer la pertinence du véhiculier pour la garantie RCM, évaluer l'apport du véhiculier en comparant les modèles incluant le véhiculier aux modèles l'excluant.

Présentation des bases de données

La base de donnée ayant servi à l'étude, correspond à la jointure d'un portefeuille client et d'une base SRA. Le portefeuille regroupe les informations liées aux polices, aux clients assurés et aux véhicules assurés, entre 2014 et 2017. Chaque ligne de la base représente une vision de contrat sinistré ou non, sur une période donnée que nous appellerons image. À chaque modification de contrat (survenance de sinistre, suspension, résiliation, avenant, etc.), une nouvelle image est créée. À cette base s'ajoute la base SRA (Sécurité et Réparation Automobile), une base générée par l'organisme du même nom et mise à disposition de tous les assureurs automobiles de la France. Cette base recense toutes les caractéristiques techniques et commerciales des véhicules, dans le but de maîtriser le mieux possible le coût des indemnisations des sinistres. Dans cette étude, cette base a été exploitée dans le but de récupérer toutes les informations disponibles sur les caractéristiques des véhicules assurés dans le portefeuille.

Lors de la construction de la base finale, des opérations de jointures ont été effectuées entre les bases. Les principales difficultés rencontrées étaient l'absence d'une clé commune aux deux bases, une écriture différente des modèles de voiture et la multiplicité des véhicules SRA. Ces difficultés ont été contournées en corrigeant les différentes erreurs d'écritures, puis en créant des clés de jointures à partir des variables communes aux deux bases.



Bases de l'étude

Traitements préliminaires

Avant l'application de la démarche de l'étude, des traitements préliminaires ont été réalisés. Dans un premier temps, les sinistres forfaitaires ont été identifiés. En effet, dans le but d'accélérer les procédures d'indemnisation pour la garantie RCM, les assureurs ont mis en place une convention permettant aux assurés d'être directement dédommagés par leur propre assureur. Il s'agit de la convention d'Indemnisation directe de l'assuré et de Recours entre Sociétés d'assurance Automobile (IRSA). Elle fonctionne de la manière suivante : lorsque le montant d'un sinistre est inférieur à 6 500 €, l'assureur non responsable indemnise directement son assuré et exerce un recours forfaitaire auprès de l'assureur responsable. Dans le cas où le montant du sinistre est supérieur à 6 500€, le recours est dit réel. Les recours forfaitaires sont des montants fixes et réévalués chaque année, ce qui entraîne une sur-représentation de ces montants dans la distribution des coûts moyens. Ce constat est effectué dans la distribution des coûts moyens du portefeuille étudié où l'on a pu observer une sur-représentation des montants entre 1 200 € et 1 400 €. Ces montants ont donc été retirés de la modélisation de la sévérité et seront rajoutés plus tard lors du calcul de la prime pure.

Dans un second temps un écrêtement des sinistres graves a été réalisé grâce à la méthode des dépassements moyens et au graphique Quantile-Quantile. Le seuil de sinistralité grave déduit de cette analyse est de 30 000 €.

Pour finir, les variables continues ont été discrétisées grâce à des algorithmes d'arbre de régression.

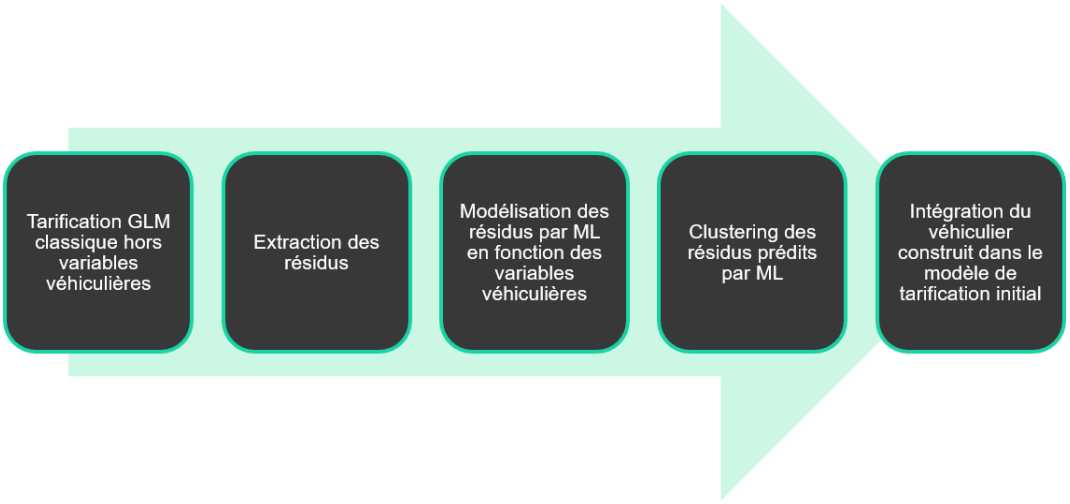
Description de l'approche mise en oeuvre

Le principe de la construction du véhiculier est de partir d'un modèle de tarification GLM classique, d'en extraire les résidus et de modéliser ces résidus à partir de méthodes d'apprentissage non supervisés (notamment les méthodes de Machine Learning). Ces résidus ainsi prédits seront ensuite classifiés/clusterisés formant ainsi le véhiculier.

L'objectif d'un véhiculier est de regrouper toute l'information liée aux véhicules en une seule variable qui sera ajoutée a posteriori au modèle de tarification initial. Deux véhiculiers seront construits, l'un pour le modèle de fréquence et l'autre pour le modèle de coût. Pour cela, les modèles GLM de fréquence et de coût initiaux sont construits hors variables véhiculières (afin d'isoler toute l'information liée au véhicule). Ensuite, les résidus de ces modèles seront modélisés en fonction des variables véhiculières. On entend par variables non véhiculières, toutes les variables liées à la police, au client et à sa zone géographique. Les variables véhiculières regroupent toutes les caractéristiques liées au véhicule.

Afin d'évaluer l'apport du véhiculier, un autre modèle GLM contenant les variables véhiculières et les variables non-véhiculières a été construit. Ce modèle servira de modèle comparatif avec les

modèles incluant le véhiculier. Cette comparaison mettra en évidence l'intérêt de réunir tout l'effet véhicule en une seule variable, plutôt que d'intégrer directement les variables véhiculières dans la modélisation. Les différentes étapes de l'approche mise en œuvre ont été résumées dans le schéma ci-dessous.



Étapes de construction du véhiculier

Mise en place des modèles de tarification GLM

En assurance IARD, les tarifs sont généralement construits à partir des modèles linéaires généralisés (GLM). On réalise une modélisation séparée de la fréquence et du coût afin de capter au mieux les effets de chaque risque. La construction des modèles se fait en trois étapes :

- Choix des lois paramétriques,
- Sélection des variables tarifaires,
- Analyse des résultats et validation des modèles.

Pour la première étape, les lois paramétriques ont été choisies à partir d'une analyse graphique et du test de Kolmogorov-Smirnov. La fréquence de sinistre a été modélisée avec une loi de Poisson et la sévérité avec une loi Gamma.

La seconde étape consiste en la sélection des variables tarifaires. Rappelons que les variables qui interviendront dans cette partie de l'étude sont les variables non véhiculières. Une pré-sélection de ces variables avait été effectuée grâce aux analyses univariées et à l'étude de la corrélation entre ces variables. Les variables qui ont été pré-sélectionnées sont celles consignées dans le schéma suivant.

Profession	Âge de l'assuré	Statut Marital
Formule de la police	Coefficient Bonus Malus (CRM)	Usage du véhicule
Mode de stationnement du véhicule	Mode d'apprentissage de la conduite	Ancienneté Bonus Malus

Sélection de variables non véhiculières

Ensuite, une étape de sélection automatique des variables a été mise en œuvre afin d'identifier celles qui influencent réellement la fréquence et le coût. Deux méthodes de sélection automatique de variables ont été comparées : la méthode backward du stepAIC et la régression pénalisée de Lasso. La méthode dont la sélection fournit le modèle optimal au sens des indicateurs de performance (RMSE et déviance) est la méthode Lasso. La fréquence de sinistre a donc été modélisée en fonction de huit variables et la sévérité en fonction de cinq variables.

L'analyse des résidus ainsi que la comparaison des valeurs prédites aux valeurs observées a permis de valider les modèles de fréquence et de coût.

Le modèle comparatif contenant l'ensemble des variables (véhiculières et non véhiculières) a été construit en suivant les étapes citées précédemment.

Modélisation des résidus par Machine Learning (ML)

L'étape suivante consiste à modéliser les résidus issus des modèles précédents en fonction des variables véhiculières (marque, carrosserie, carburant, etc.). Ces résidus contiennent en plus d'un bruit, l'information liée au véhicule qui n'a pas totalement été expliquée par les modèles GLM.

Dans le cadre de cette étude, nous avons privilégié l'utilisation des modèles facilement interprétables. À cet effet, les résidus ont été modélisés par Random Forest et Gradient Boosting qui sont des algorithmes d'apprentissage non paramétriques basés sur les arbres de régression. Contrairement aux GLM, ces méthodes ne font pas d'hypothèses sur la structure de la variable à expliquer. Cependant, il existe d'autres méthodes de Machine Learning tels que les réseaux neurones, qui pourraient être explorées.

Les méthodes de ML mises en place ont été calibrées à partir d'hyperparamètres renseignés par l'utilisateur. Afin d'obtenir les modèles les plus optimaux possibles, une phase supplémentaire d'optimisation des hyperparamètres a été implémentée par validation croisée. Les deux méthodes ont été comparées suivant le critère du RMSE (Erreur quadratique moyenne). Le RMSE était très proche pour les deux algorithmes, cependant le Gradient Boosting réalisait de meilleures performances.

Les résidus prédits par Gradient Boosting ont ensuite été agrégés suivant le code SRA (identifiant des véhicules) afin d'attribuer à chaque véhicule, un score unique. Le résidu agrégé correspond à la moyenne des résidus par code SRA. Ces résidus seront ensuite clusterisés/classifiés aboutissant ainsi au véhiculier.

Construction du véhiculier

Cette dernière étape consiste à classifier les résidus prédits afin de créer des groupes de risques homogènes. Le clustering est une méthode d'apprentissage automatique, qui consiste à regrouper des individus selon leurs similarités ou selon la distance qui les sépare. Nous avons implémenté deux des méthodes les plus couramment utilisées. Ces méthodes ont été comparées afin de choisir celle qui fournit la classification la plus optimale des données. La première méthode est celle des k-means, une méthode de partitionnement et la seconde est la méthode de Ward, une méthode de Classification Ascendante Hiérarchique (CAH).

La méthode de Ward fournit un partitionnement en 3 classes tandis qu'avec la méthode des k-means, des partitionnements de 9, 10 et 11 classes ont été réalisés. En effet, l'algorithme des k-means n'est pas capable de fournir le nombre de classes optimal k. Il a donc fallu identifier les valeurs de k, qui minimisent la variance intra-classe, grâce à la "méthode du coude". Nous disposons ainsi de 4 véhiculiers distincts pour chacun des modèles de fréquence et de sévérité. Ces nouvelles variables

tarifaires ont été intégrées dans les modèles GLM construits précédemment, et elles représentent la part explicative du risque liée au véhicule.

La classification optimale est celle qui minimise les critères d'AIC et de déviance, pour les modèles GLM. Le véhiculier optimal obtenu est donc celui qui segmente les données en 11 classes de risques. Intuitivement, le véhiculier de trois classes issu de la méthode de Ward n'aurait pas été retenu, car une segmentation en trois classes n'aurait pas été suffisamment fine pour permettre de se démarquer sur le marché. Il faut noter que les classes ont été numérotées de la classe la moins risquée à la classe la plus risquée.

Toute l'information liée au véhicule a donc été regroupée en une seule variable discriminant le risque en 11 classes. Il convient maintenant d'évaluer l'apport de cette nouvelle variable aux modèles initiaux en comparant les modèles avec véhiculier et les modèles sans véhiculier.

Analyse des résultats

Les modèles ci-dessous ont été comparés suivant les critères de l'AIC, de la déviance et de l'indice de Gini.

- Modèle A : GLM contenant les variables non véhiculières uniquement.
- Modèle B : GLM contenant les variables non véhiculières plus le véhiculier construit.
- Modèle C : GLM comparatif contenant les variables non véhiculières et les variables véhiculières.

Les performances des modèles ont été calculées sur les bases d'apprentissage, de validation et de test. Dans le tableau ci-dessous, nous avons comparé les modèles deux à deux, en calculant la variation des indicateurs lorsqu'on passe d'un modèle à l'autre. Pour la fréquence, le passage du modèle A au modèle B, améliore (diminue) l'AIC de 30 %, sur l'échantillon de validation. Autrement dit, l'intégration du véhiculier dans le modèle A, améliore l'AIC de 30 %. Le même constat est réalisé pour la déviance. De même, le passage du modèle C au modèle B montre que remplacer les variables véhiculières par le véhiculier améliore l'AIC de 11 % (sur l'échantillon de validation). Cependant, on n'observe pas une amélioration de la déviance, car le modèle C est plus complexe que le modèle B, donc l'écart entre la log-vraisemblance du modèle saturé et celle du modèle étudié, est plus faible.

En outre, la comparaison des indices de Gini des modèles permet d'aboutir aux mêmes conclusions. En effet, pour le modèle de fréquence, lorsqu'on passe du modèle C au modèle B, on constate une amélioration (augmentation) de l'indice de Gini de 17% sur l'échantillon d'apprentissage et de 16% sur l'échantillon de validation. Le modèle B est donc celui qui discrimine le mieux le risque. Les mêmes observations ont été faites pour le modèle de sévérité.

	Pourcentage d'amélioration des indicateurs				
	Indicateurs	Echantillon	A-B	C-B	A-C
Fréquence	AIC	Apprentissage	25%	4%	22%
		Validation	30%	11%	21%
		Test	30%	9%	23%
	Déviance	Apprentissage	22%	-7%	27%
		Validation	21%	-20%	34%
		Test	32%	-27%	46%

Figure 1 – Comparaison des modèles de fréquence

Modèles		Echantillon d'apprentissage	Echantillon de validation
Fréquence	Modèle C	19.83%	18.42%
	Modèle B	23.27%	21.34%
	Amélioration	17%	16%
Coût	Modèle C	43.12%	42.75%
	Modèle B	55.62%	53.09%
	Amélioration	29%	24%

FIGURE 2 – Comparaison des coefficients de Gini

Il résulte de la comparaison des trois modèles, que le modèle B réalise les meilleures performances sur chacun des échantillons. Par ailleurs, l'analyse de la qualité de prédiction (cf. figures 3 et 4) a remis en cause, la pertinence de la construction d'un véhiculier pour la garantie Responsabilité Civile Matérielle, notamment pour le modèle de fréquence. En effet, pour ce modèle, les prédictions des variables véhiculières sont moins précises que celles des variables non véhiculières. Le risque étudié étant les dommages causés à un tiers, la fréquence de la garantie RCM serait beaucoup plus influencée par les caractéristiques de l'assuré et de la police, plutôt que par les caractéristiques du véhicule. Le véhiculier serait donc plus pertinent pour le modèle de sévérité, pour lequel les prédictions des variables véhiculières sont meilleures. Le montant des dommages dépendrait de la puissance du choc, donc des caractéristiques du véhicule (performances et masse). En outre, une analyse plus approfondie des classes de risques formées, a révélé que les véhicules figurants dans les classes les plus risquées du modèle de sévérité, sont principalement des véhicules utilitaires ou des véhicules de luxe du type sportif. Les véhicules utilitaires ont la particularité d'être lourds et puissants. Ils peuvent donc facilement générer des montants de dommages élevés lors d'un accident.

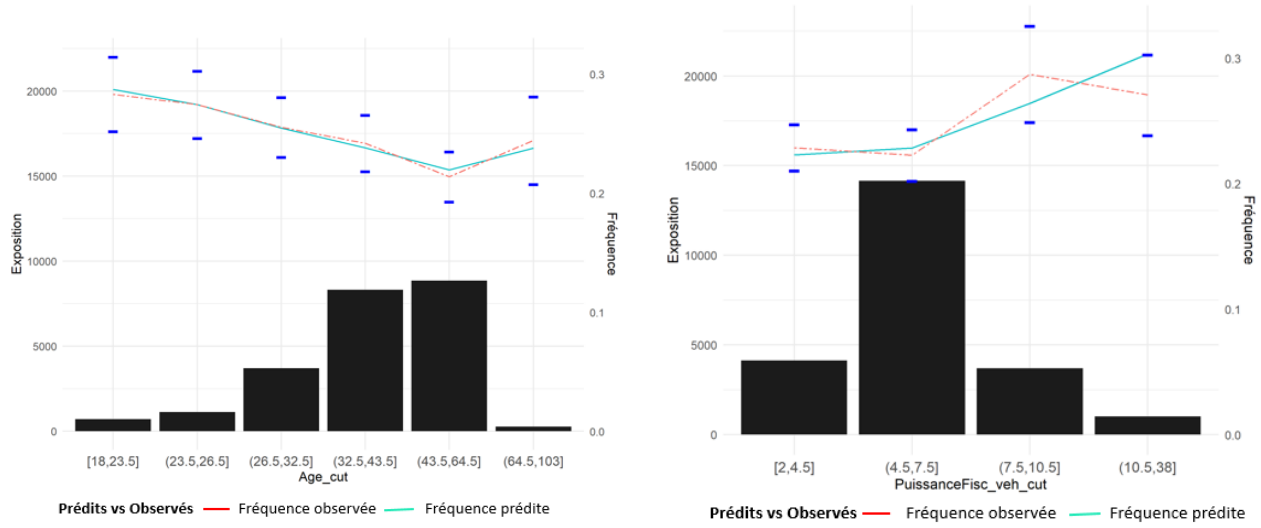


FIGURE 3 – Analyse des prédictions de la fréquence

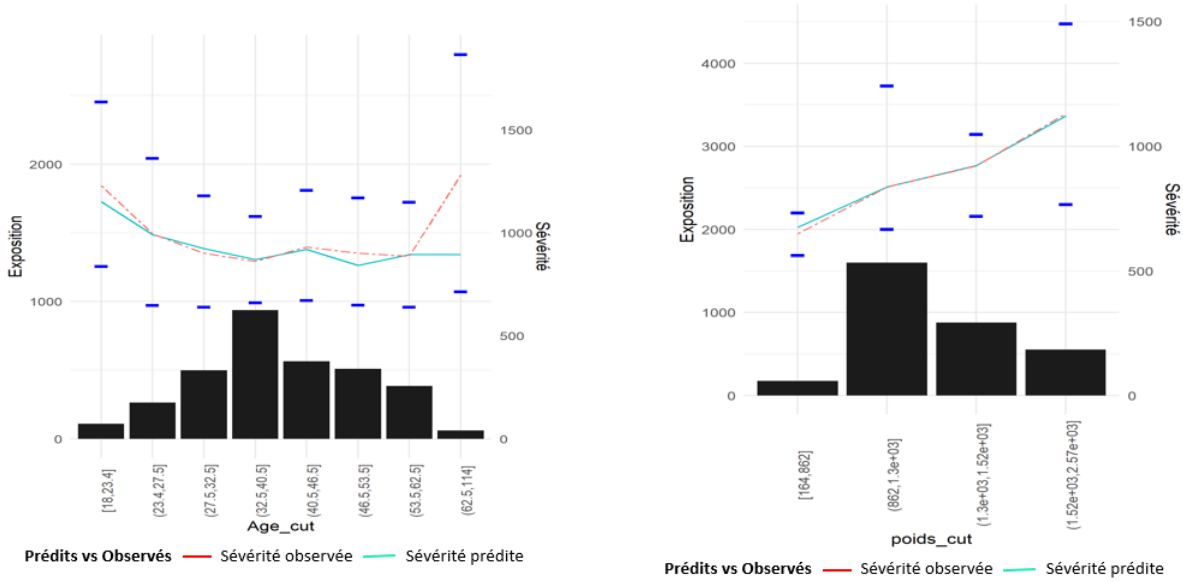


Figure 4 – Analyse des prédictions de la sévérité

Limites et points d'amélioration

L'étude a démontré que l'intégration d'un véhiculier dans les modèles de tarification GLM améliore non seulement la performance des modèles, mais permet également une bonne compréhension de la sinistralité par segments de risques. Cet outil permettra à l'assureur d'adapter son tarif suivant les différents profils de risques contenus dans son portefeuille. En effet, le tarif d'un assuré dont le véhicule est considéré comme hautement risqué par le véhiculier, pourrait être révisé à la hausse tandis que le tarif d'un assuré possédant un véhicule à faible risque, pourrait être révisé à la baisse. Par ailleurs, la construction d'un véhiculier est un processus long et couteux en temps. Pour une garantie donnée, le véhiculier est avantageux pour l'assureur lorsque sa valeur ajoutée est conséquente d'un point de vue commercial (en termes de gains).

Du point de vue des indicateurs de performance, les modèles contenant le véhiculier se sont révélés plus performants que les modèles contenant les variables véhiculières. Ainsi, synthétiser l'effet véhicule dans une seule variable permet d'affiner la segmentation, de simplifier les modèles et d'avoir de meilleures performances. Cependant, cette synthétisation de l'information ne se fait pas sans perte d'information. Une solution serait peut-être de créer un modèle alternatif contenant les variables véhiculières les plus pertinentes en plus d'un véhiculier qui regrouperait les variables les moins importantes. Ainsi, la perte de l'information serait minimisée.

En outre, le nombre de classes obtenues pour le véhiculier paraît insuffisant par rapport à ce qui s'obtient en général sur le marché (une trentaine de classes en moyenne). Cependant, compte tenu du volume de données à disposition de cette étude, réaliser une segmentation avec un nombre de classes trop élevé n'aurait pas été approprié, car certaines classes n'auraient pas été suffisamment représentées.

Il pourrait être utile de comparer cette approche à d'autres approches telles que celle basée sur le lissage spatial des résidus prédits par Machine Learning. Cette approche consisterait à créer une carte de véhicules en s'inspirant des méthodes de classifications géographiques afin d'effectuer un lissage par voisinage. Cette approche n'a pas été mise en œuvre par manque de temps, mais pourrait faire l'objet de travaux ultérieurs.

Enfin, la mise en place d'un système optimal de mise à jour du véhiculier, pourrait être utile. En

effet, il faudrait intégrer au fur et à mesure les nouveaux véhicules commercialisés ainsi que ceux qui n'étaient pas présents dans la base de donnée au moment de l'étude.

Executive summary

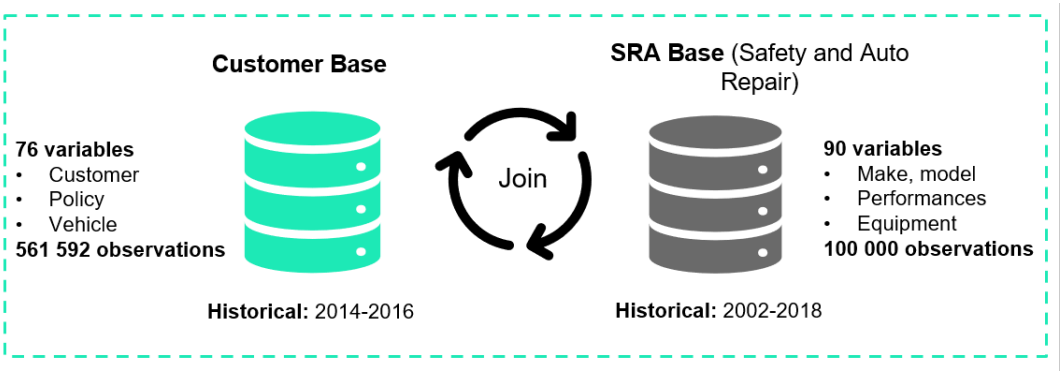
Rate segmentation is a major issue in property and casualty insurance. Not only does it allow the insurer to avoid anti-selection by distinguishing good risks from bad risks, but it also helps the insurer to build customer loyalty. Segmentation can be achieved through several tools, such as zoning variables (geographical classification) or vehicles classification.

This paper is devoted to the construction of a vehicle classification tool, on a portfolio covering Third Party Damage (TPD), based on Machine Learning methods. The objectives of the study are multiple : to refine the segmentation and simplify the pricing model thanks to the vehicle classification, to evaluate the relevance of the vehicle classification for the TPD coverage, to evaluate the contribution of the vehicle classification by comparing the models including the vehicle classification to the models excluding it.

Presentation of the databases

The database used for the study corresponds to the join of a client portfolio and a SRA database. The portfolio gathers information related to policies, insured clients and insured vehicles, between 2014 and 2017. Each line in the database represents a view of a contract that has or has not been subject to a claim, over a given period of time, which we will call an image. Each time a contract changes (claim occurrence, suspension, termination, endorsement, etc.), a new image is created. In addition to this database, the SRA (Automotive Safety and Repair) database is generated by the organization of the same name and is available to all car insurers in France. This database lists all the technical and commercial characteristics of the vehicles, with the aim of controlling the cost of compensation for claims as much as possible. In this study, this database was used to retrieve all the available information on the characteristics of the vehicles insured in the portfolio.

During the construction of the final database, join operations were performed between the databases. The main difficulties encountered were the absence of a key common to both databases, a different spelling of the car models and the multiplicity of SRA vehicles. These difficulties were circumvented by correcting the different writing errors, then by creating join keys from the variables common to both databases.



Databases presentation

Preliminary treatments

Prior to the application of the study approach, preliminary treatments were performed. First, lump sum claims were identified. Indeed, in order to accelerate the compensation procedures for the TPD coverage, the insurers have put in place an agreement allowing the insured to be directly compensated by their own insurer. This is the Direct Insured Compensation and Recourse Agreement between Automobile Insurance Companies (IRSA). It works as follows : when the amount of a claim is less than 6 500€, the insurer not at fault compensates its insured directly and exercises a lump-sum recourse against the insurer at fault. If the amount of the claim is higher than 6 500€, the recourse is called real. The lump-sum recoveries are fixed amounts that are revalued each year, which leads to an over-representation of these amounts in the distribution of average costs. This observation is made in the distribution of the average costs of the portfolio studied, where we observed an over-representation of amounts between 1 200€ and 1 400€. These amounts have therefore been removed from the severity model and will be added later when calculating the pure premium.

In a second step, a capping of the serious claims was carried out thanks to the method of average overruns and the Quantile-Quantile graph. The threshold of serious claims was set at 30 000€.

Finally, the continuous variables were discretized using regression tree algorithms.

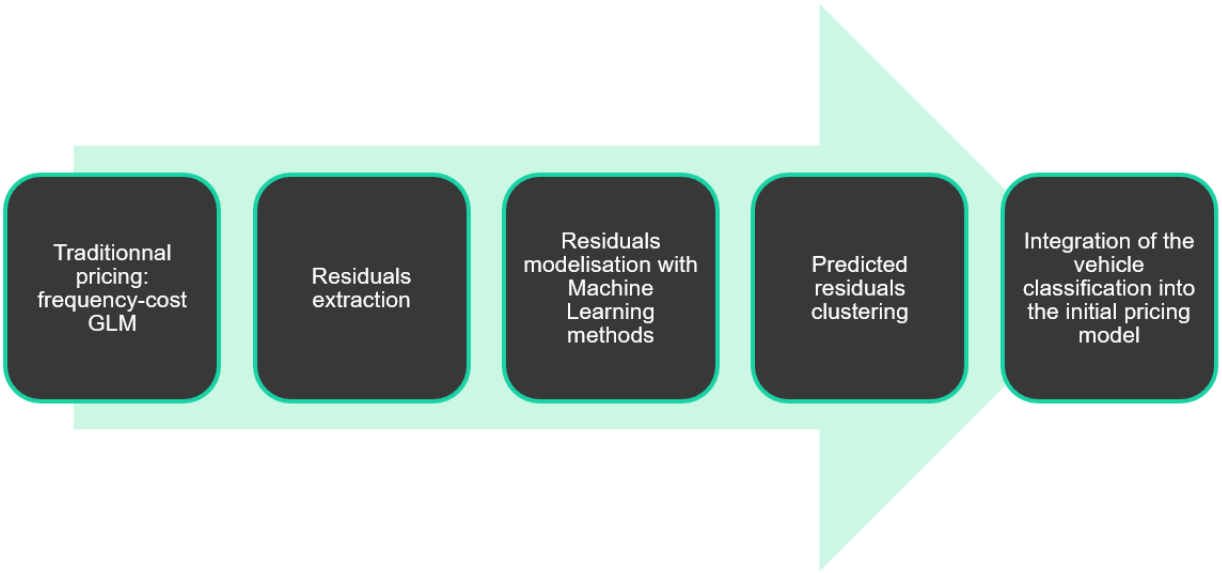
Description of the implemented approach

The principle of the construction of the vehicle classification is to start from a classical GLM pricing model, to extract the residuals and to model these residuals using unsupervised learning methods (especially Machine Learning methods). These predicted residuals will then be classified/clustered to form the vehicle classification.

The objective of a vehicle classification is to aggregate all vehicle-related information into a single variable that will be added to the initial pricing model. Two vehicles will be constructed, one for the frequency model and one for the cost model. For this purpose, the initial frequency and cost GLM models are constructed without vehicle variables (in order to isolate all vehicle-related information). Then, the predicted residuals are modeled as a function of the vehicle variables. Non-vehicle variables are defined as all variables related to the policy, the client and the geographical area. Vehicle variables include all vehicle-related characteristics.

In order to evaluate the contribution of the vehicle classification, another GLM model was constructed, this time containing the vehicle and non-vehicle variables. This model will serve as a comparative model with the models, including the vehicle classification. This comparison will highlight the interest of combining the entire vehicle effect into a single variable, rather than integrating the

vehicle variables directly into the model. The different steps of the approach have been summarized in the figure below.



Steps of the implemented approach

Implementation of the GLM pricing models

In property and casualty insurance, rates are generally constructed using generalized linear models (GLM). Frequency and cost are modeled separately to best capture the effects of each risk. The models are built in three stages :

- Choice of parametric laws,
- Selection of pricing variables,
- Analysis of results and validation of models.

For the first step, the parametric laws were chosen from a graphical analysis and the Kolmogorov-Smirnov test. The frequency of loss was modeled with a Poisson distribution and the severity with a Gamma distribution.

The second step consists in the selection of the rate variables. Recall that the variables that will be used in this part of the study are the non-vehicle variables. A pre-selection of these variables had been carried out thanks to the univariate analyses and the study of the correlation between these variables. The variables that were pre-selected are those shown in the following diagram.

Profession	Age	Marital status
Policy package	Bonus Malus	Vehicule usage
Parking mode	Driving learning method	Bonus Malus age

Selection of non-vehicle variables

Next, an automatic variable selection step was implemented in order to identify the variables that really influence frequency and cost. Two automatic variable selection methods were compared : the stepAIC backward method and the Lasso penalized regression. The method whose selection provides the optimal model in the sense of the performance indicators is the Lasso method. Loss frequency was therefore modeled as a function of eight variables and severity as a function of five variables.

The analysis of the residuals and the comparison of the predicted values to the observed values allowed the validation of the frequency and cost models.

The comparative model containing all the variables was built following the steps mentioned above.

Residuals modeling by Machine Learning methods (ML)

The next step is to model the residuals from the previous models as a function of the vehicle variables (make, body, fuel, etc.). These residuals contain, in addition to noise, information related to the vehicle that has not been fully explained by the GLM models.

In this study, we favored the use of easily interpretable models. To this end, the residuals were modeled by Random Forest and Gradient Boosting which are non-parametric learning algorithms based on regression trees. Unlike GLM, these methods do not make assumptions about the structure of the variable to be explained. However, there are other Machine Learning methods, such as neural networks, which could be explored.

The ML methods implemented were calibrated using hyperparameters provided by the user. In order to obtain the most optimal models possible, an additional phase of hyperparameter optimization was implemented by cross-validation. The two methods were compared according to the RMSE (root mean square error) criterion. The RMSE was very close for both algorithms, however the Gradient Boosting performed better.

The residuals predicted by Gradient Boosting were then aggregated according to the SRA code (vehicle identifier) in order to assign a unique score to each vehicle. The aggregated residual is the average of the residuals per SRA code. These residuals will then be clustered/classified, leading to the vehicle.

Construction of the vehicle classification

This last step consists in classifying the predicted residues in order to create homogeneous risk groups. Clustering is a machine learning method, which consists in grouping individuals according to their similarities or according to the distance that separates them. We have implemented two of the most commonly used methods. These methods have been compared in order to choose the one that provides the most optimal classification of the data. The first method is the k-means, a partitioning method and the second is the Ward's method, a Hierarchical Ascending Classification (HAC) method.

Ward's method provides a partitioning in 3 classes while with the k-means method, partitioning of 9, 10 and 11 classes have been achieved. Indeed, the k-means algorithm is not able to provide the optimal number of classes k. It was therefore necessary to identify the values of k, which minimize the intra-class variance, using the "elbow method". We thus have 4 distinct vehicles for each of the frequency and severity models. These new rating variables have been integrated into the GLM models constructed previously, and they represent the explanatory part of the risk related to the vehicle.

The optimal classification is the one that minimizes the AIC and deviance criteria for the GLM models. The optimal vehicle obtained is therefore the one that segments the data into 11 risk classes. Intuitively, the 3-class vehicle derived from Ward's method would not have been retained, because we need more finesse in the segmentation in order to stand out in the market. Note that the classes were numbered from the least risky to the most risky.

All the information related to the vehicle was therefore grouped into a single variable that discriminates the risk into 11 classes. It is now necessary to evaluate the contribution of this new variable to the initial models by comparing the models with and without vehicles.

Analysis of the results

The models below were compared according to the AIC, deviance and Gini index criteria.

- Model A : GLM containing the non-vehicle variables only,
- Model B : GLM containing the non-vehicle variables plus the constructed vehicle classification variable.
- Model C : Comparative GLM containing non-vehicle and vehicle variables.

The performances of the models were computed on the learning sample and then tested on a test sample to ensure the robustness of the models.

The performance of the models was calculated on the learning, validation and test bases. In the table below, we have compared the models two by two, by calculating the variation of the indicators when moving from one model to the other. For the frequency, the switch from model A to model B, improves (decreases) the AIC by 30 %, on the validation sample. In other words, the inclusion of the vehicle in model A improves the AIC by 30 %. The same observation is made for the deviance. Similarly, moving from Model C to Model B shows that replacing the vehicle variables with the vehicle improves the AIC by 11%(on the frequency validation sample). However, we do not observe an improvement in deviance, because model C is more complex than model B, so the difference between the log-likelihood of the saturated model and the one of the model under study is smaller.

Furthermore, the comparison of the Gini indices of the models leads to the same conclusions. Indeed, for the frequency model, when we switch from model C to model B, we see an increase in the Gini index of 17% on the training sample and 16% on the validation sample. Model B is therefore the one that best discriminates risk. The same observations were made for the severity model.

	Percentage of indicators improvement				
	Indicators	Sample	A-B	C-B	A-C
Frequency	AIC	Training	25%	4%	22%
		Validation	30%	11%	21%
		Testing	30%	9%	23%
	Déviance	Training	22%	-7%	27%
		Validation	21%	-20%	34%
		Testing	32%	-27%	46%

Figure 5 – Comparison of frequency models

Models		Training sample	Validation sample
Frequency	Model C	19.83%	18.42%
	Model B	23.27%	21.34%
	Improvement	17%	16%
Severity	Model C	43.12%	42.75%
	Model B	55.62%	53.09%
	Improvement	29%	24%

FIGURE 6 – Comparison of the Gini index

It results from the comparison of the three models, that the model B realizes the best performances on each sample. Moreover, the analysis of the prediction quality (cf. figures 7 and 8) has called into question the relevance of the construction of a vehicle classification for the TPD coverage, in particular for the frequency model. Indeed, for this model, the predictions of the vehicle variables are less precise than those of the non-vehicle variables. Since the risk under study is damage caused to a third party, the frequency of TPD coverage would be much more influenced by the characteristics of the insured and the policy, rather than by the characteristics of the vehicle. The vehicle classification would therefore be more relevant to the severity model, for which the predictions of the vehicle variables are better. The amount of damage would depend on the power of the impact, and thus on the characteristics of the vehicle (performance and mass). In addition, a more detailed analysis of the risk classes formed revealed that the vehicles in the highest risk classes of the severity model are mainly commercial vehicles or luxury vehicles of the sport type. Commercial vehicles have the particularity of being heavy and powerful. They can therefore easily generate high damage amounts in an accident.

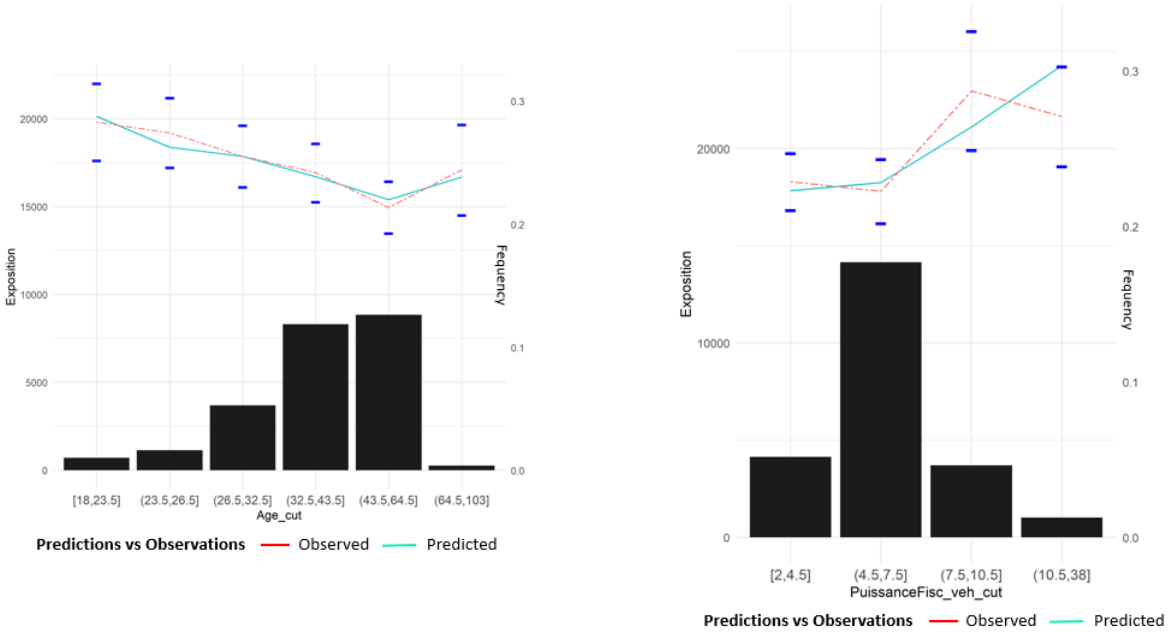


FIGURE 7 – Analysis of frequency predictions

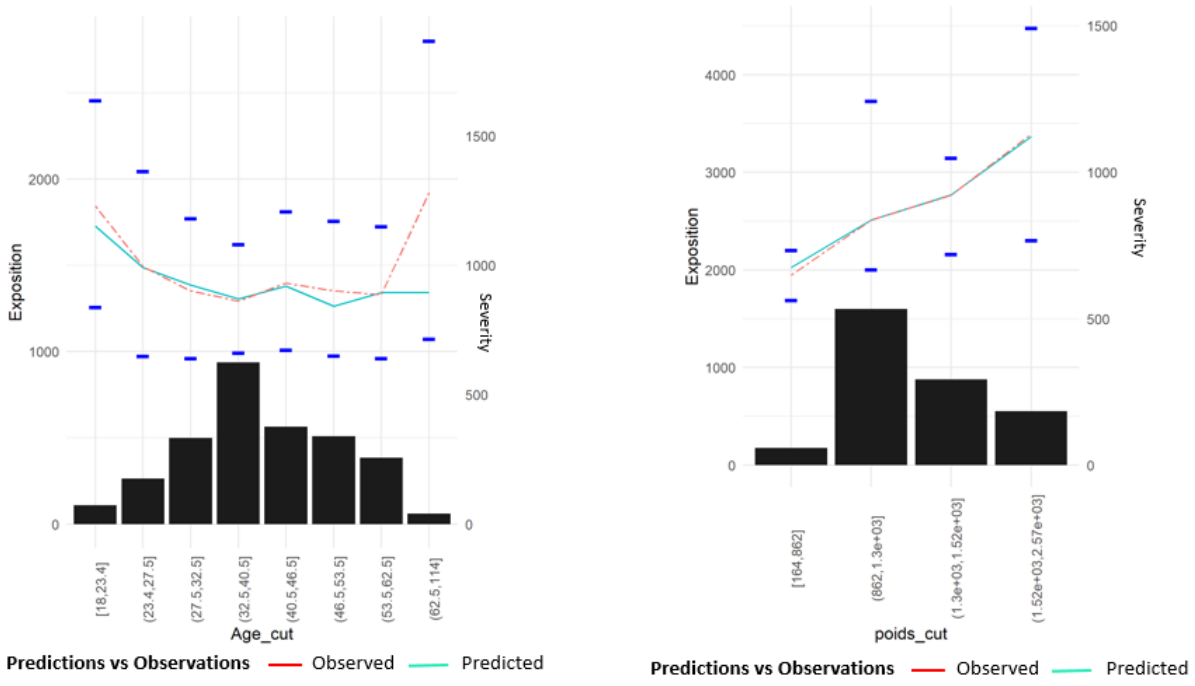


Figure 8 – Analysis of severity predictions

Limitations and areas for improvement

The study showed that incorporating a vehicle classification into GLM pricing models not only improves the performance of the models, but also provides a good understanding of the claims experience by risk segment. From the point of view of performance indicators, the models containing the vehicle classification were found to perform better than the models containing the vehicle variables. Thus, synthesizing the vehicle effect in a single variable makes it possible to refine the segmentation, simplify the models and achieve better performance. However, this synthesis of information is not done without loss of information. One solution might be to create an alternative model containing the most relevant vehicle variables in addition to a vehicle that groups together the least important variables. In this way, the loss of information will be minimized.

Moreover, the number of classes obtained for the vehicle seems insufficient compared to what is generally obtained on the market (about thirty classes on average). However, given the volume of data available for this study, a segmentation with too many classes would not have been appropriate, as some classes would not have been sufficiently represented.

It could be useful to compare this approach with other approaches, such as the one based on spatial smoothing of the residuals predicted by Machine Learning. This approach would consist in creating a map of vehicles inspired by geographical classification methods in order to perform a smoothing by neighborhood. This approach has not been implemented due to lack of time, but could be the subject of future work.

Finally, the implementation of an optimal system for updating the vehicle could be the subject of future studies. Indeed, it would be necessary to integrate new vehicles that are marketed as well as those that were not present in the database at the time of the study.

Remerciements

Je remercie tout particulièrement mes deux encadrants, Fabien CHERANCE actuaire manager et Claire NICOLLE, consultante senior tous deux au sein de l'Unité de Compétences (UC) Actuariat de Sia Partners. Je les remercie pour leur accompagnement et leurs conseils tout au long de la rédaction de ce mémoire.

Mes remerciements s'adressent également à Michaël DONIO et Ronan DAVIT, les deux directeurs de l'UC Actuariat qui m'ont offert l'opportunité d'effectuer mon stage au sein de leur service.

Aussi je remercie tous les membres de l'UC Actuariat pour leur chalereux accueil et leur sympathie et tout particulièrement Babacar BADJI manager senior au sein de l'UC pour tous ses conseils dans le cadre du bon déroulement de mon stage.

Enfin, je remercie ma tutrice académique Caroline HILLAIRET sans oublier ma famille pour tous leurs encouragements et leur soutien indéfectible.

Table des matières

Introduction

L'assurance automobile est un secteur saturé qui ne cesse d'évoluer. Face à la rude concurrence qui y règne, les acteurs du secteur s'attellent à identifier les différents profils de risque qui composent leur portefeuille, afin d'adapter au mieux leurs tarifs et ainsi fidéliser leur clientèle. Cette distinction des profils de bons risques et de mauvais risques est un processus appelé la segmentation qui est un regroupement des individus par groupe de risque homogène.

Le processus de tarification en assurance IARD est effectué par les actuaires. Ces derniers sont chargés de déterminer les différents paramètres à prendre en considération pour tarifer un risque donné, en partant d'une analyse statistique globale du portefeuille qu'ils ont à disposition. Ils sont donc chargés de réaliser la segmentation pour le risque étudié.

La segmentation peut être réalisée suivant différents paramètres et en assurance automobile, la segmentation suivant le véhicule est très courante. Il s'agit de créer une variable appelée véhiculier qui représentera une classification des différents types de véhicules selon le risque étudié.

L'objectif de ce mémoire est donc la construction d'un véhiculier sur un portefeuille d'assurance automobile couvrant la garantie Responsabilité Civile Matérielle. Pour cela, la composante du risque liée au véhicule sera modélisée à l'aide des techniques de Machine Learning, qui présentent l'avantage de permettre une optimisation de la performance prédictive des algorithmes. Nous partirons d'un modèle de tarification construit en excluant les variables liées au véhicule. Ensuite, nous modéliserons les résidus de ce modèle en fonction des variables exclues afin de capturer toute l'information liée au véhicule qui n'a pas été expliquée par le modèle initial. Le véhiculier sera construit grâce à des méthodes de clustering puis intégré dans le modèle initial en tant que variable tarifaire. Afin de mesurer son apport à l'optimisation des modèles, les modèles incluant le véhiculier seront comparés aux modèles l'excluant.

Ce mémoire est décomposé en plusieurs axes. Dans un premier temps, le contexte de l'étude sera présenté ainsi que les études préliminaires qui ont été effectuées avant la mise en œuvre de l'approche. Ensuite, chacune des étapes de l'approche sera mise en œuvre en passant par une description théorique des différentes notions évoquées. Enfin, les résultats de l'étude seront analysés, puis une vision critique de la démarche mise en œuvre sera donnée, à travers l'étude de ses limites.

Chapitre 1

Contexte et objectifs de l'étude

Dans ce premier chapitre, nous allons présenter le contexte de l'étude, l'assurance automobile et la garantie responsabilité civile matérielle. Ensuite, nous décrirons les objectifs de l'étude qui découlent notamment des besoins de segmentation dans le domaine de la tarification.

1.1 Enjeux de la tarification en assurance automobile

1.1.1 La segmentation et la mutualisation

En assurance, deux principes fondamentaux apparaissent antagonistes : la segmentation et la mutualisation. Pour faire face à une grande variabilité de la réalisation des risques et réduire l'exposition globale au risque, il y a besoin de les mutualiser, c'est-à-dire en considérer un grand nombre afin de tendre vers le risque moyen.

Basée sur la loi des grands nombres, la mutualisation suppose que les dommages subis par un ensemble d'individus sont des variables aléatoires indépendantes et identiquement distribuées. Autrement dit, la mutualisation est appliquée au sein d'une population dans laquelle les individus ont les mêmes profils de risque. Elle peut être vue comme une mise en commun, un partage des risques individuels.

Le second principe est celui de la segmentation en vue d'avoir des ensembles de risques homogènes. Ce besoin de segmentation vient du fait qu'en général les risques ne sont pas homogènes et il s'avère nécessaire de les regrouper afin de pouvoir appliquer une prime différenciée à chacun des groupes ayant un risque homogène.

La segmentation peut donc être définie comme une technique par laquelle, l'assureur découpe son portefeuille d'assurés en sous-portefeuilles, encore appelés classes de risques, par l'étude de caractères distinctifs, et ce, dans le but d'appliquer un tarif différencié selon la classe. Elle permet de différencier les « bons risques » des « mauvais risques ».

Ainsi, dans un contexte de marché très concurrentiel, la segmentation est aujourd'hui incontournable pour les assureurs, car elle leur permet non seulement de fidéliser leur portefeuille, mais aussi d'attirer de nouveaux clients représentant les « bons risques » afin d'éviter l'antisélection. Cependant, plusieurs facteurs sont à prendre en compte dans la réalisation d'une segmentation.

Les limites de la segmentation tarifaire

L'explosion exponentielle des données (Big Data) et des moyens de collecte de plus en plus complexes permet aujourd'hui des segmentations beaucoup plus fines en groupes plus restreints. Ce fait conduit

potentiellement à l'idée de tarification individuelle, entraînant ainsi un paradoxe avec le principe de base de l'assurance : la mutualisation.

De plus, dans un contexte dépourvu de contraintes réglementaires fortes, cette technique constitue une stratégie pour les assureurs afin de conquérir les segments de marché les plus rentables et de mieux appréhender ses risques en portefeuille. Cette voie peut conduire l'assureur à vouloir prendre en compte des informations dont l'utilisation n'est pas compatible avec le respect de la personne et de sa vie privée.

Le niveau de segmentation doit donc tenir compte des contraintes réglementaires telles que l'interdiction de tarifer selon le sexe de l'assuré, quand bien même la variable sexe s'est révélé discriminante en assurance auto/moto dans diverses études.

En outre, il faut noter que les segmentations fines entraînent des modèles complexes, ce qui augmente les risques de modèle, et d'estimation pouvant impacter les tarifs.

Enfin, la pression de la concurrence entraînant une forte segmentation permet de favoriser les risques systémiques et la discrimination. En effet, les assureurs qui segmentent leurs tarifs incitent à court terme leurs concurrents à segmenter aussi, sans quoi ils perdraient à la longue les « bons risques » et ne conserveraient que les « mauvais risques ». Dans le même temps, tandis que les tarifs individualisés récompensent les « bons risques », les individus les plus exposés aux risques ou les moins chanceux, se retrouvent, eux, obligés de subir des primes élevées, voire des exclusions.

Ainsi, il est primordial que les assureurs choisissent, de façon optimale, le degré de segmentation et de mutualisation adéquat dans l'élaboration de leurs tarifs.

1.1.2 Principes de tarification en assurance

En assurance Non-Vie, l'aléa du risque couvert réside non seulement sur la date de versement des flux, mais également sur le montant du sinistre. Il est donc impératif pour l'assureur de bien évaluer le risque auquel il est confronté afin d'adapter au mieux ses tarifs.

La tarification est une étape primordiale pour l'assureur dans le but de proposer des primes cohérentes aux assurés. Elle constitue une opération centrale et nécessite une phase rétrospective pour prendre en compte les données (qui sont nécessairement du passé) et une analyse prospective pour que la tarification soit en adéquation avec l'évolution des risques. Il est également nécessaire de prendre en compte les objectifs de rentabilité et les réactions de la concurrence.

La construction d'un tarif en assurance IARD passe par la construction de la prime pure. En effet, la prime d'assurance peut être décomposée de la manière suivante :

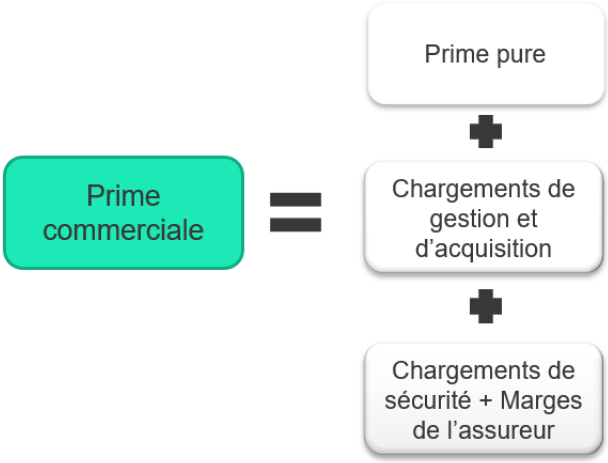


FIGURE 1.1 – Décomposition de la prime d'assurance

La prime pure correspond à l'espérance du montant des sinistres. Elle couvre le montant probable du sinistre dépendant des garanties souscrites. En supposant que la fréquence de survenance et les coûts des sinistres sont indépendants et identiquement distribués, la prime pure peut se calculer de la manière suivante en estimant la charge totale des sinistres :

$$E[S] = E\left[E\left[\sum_{i=1}^{N} X_i | N\right]\right] = E[N] \times E[X]$$

Avec
S : la charge totale des sinistres
N : le nombre total de sinistres
X_i : le montant du sinistre i.

La modélisation de la prime pure passe par une modélisation séparée du coût et de la fréquence de sinistre à l'aide de modèles statistiques adaptés. En effet, en assurance Non-Vie, le coût du sinistre est rarement connu. Il est donc important de simuler d'une part, la fréquence de sinistres et d'autre part, le coût afin de mieux cerner les différents facteurs influençant la sinistralité.

La prime commerciale correspond à la prime réellement versée par l'assuré. Elle est déterminée en ajoutant à la prime pure les chargements de gestion et d'acquisition permettant de financer les coûts d'acquisition et d'administration supportés par l'assureur. On peut également y inclure un chargement de sécurité ainsi que les marges de l'assureur. Enfin, à la prime commerciale peut être ajoutée la politique commerciale et stratégique de l'assureur. En effet, les agents ou les courtiers peuvent être amenés à faire des remises afin d'attirer un type d'assuré spécifique, par exemple, pour les bons risques.

La tarification constitue une étape très sensible dans le métier de l'actuaire. Une mauvaise estimation de la prime pure d'une police d'assurance entraînerait une sous-estimation ou une sur-estimation de la charge de sinistre, mettant ainsi en péril l'activité de l'assureur. Plus précisément, une mauvaise tarification met en péril l'activité commerciale de l'assureur (départ vers la concurrence en cas de sur-tarification) et une sous-provision de fonds propres (en cas de sous-tarification) qui pourrait entraîner la ruine de l'assureur.

Ainsi, face à un marché hyper-concurrentiel tel que l'assurance Non-Vie, les actuaires s'évertuent à développer des structures tarifaires innovantes à travers l'introduction de nouvelles variables tari-

faires.

1.1.3 L'assurance automobile en France

L'assurance Non-Vie

L'assurance Non-Vie regroupe l'ensemble des contrats d'assurance qui n'ont pas pour objet la vie de l'assuré. Il s'agit des contrats de type IARD (Incendie, Accidents et Risques Divers) qui couvrent les dommages et la protection des biens. C'est une opération par laquelle l'assuré contracte, moyennant un paiement (la prime ou cotisation), une prestation par l'assureur en cas de réalisation d'un risque. Contrairement à l'assurance vie où la survenance du sinistre est le plus souvent certaine, en assurance Non-Vie, elle est juste probable, c'est-à-dire incertaine.

Les deux principaux secteurs de l'assurance IARD sont l'Automobile et la Multi-Risques Habitation (MRH).

L'assurance automobile

Le contrat d'assurance automobile est obligatoire en France depuis 1958 selon la loi n°58-208 du 27 février 1958. L'objectif est de garantir le conducteur d'un véhicule automobile contre les conséquences des dommages matériels ou corporels causés par son véhicule à des tiers. En fonction du type de contrat souscrit, l'assurance automobile peut également couvrir les dommages matériels pour le véhicule assuré et les dommages corporels du conducteur.

L'assurance automobile se décompose en plusieurs garanties : les garanties obligatoires, les garanties de dommage au véhicule, les garanties complémentaires, la garantie personnelle du conducteur et les garanties facultatives de service.

Les garanties obligatoires : Les garanties obligatoires sont les garanties de la responsabilité civile matérielle et corporelle (également appelés « assurance au tiers »). Ce type de garantie couvre exclusivement les dommages causés à un tiers (passager, autre conducteur, piéton,…) ou au bien d'un tiers. Cependant, la responsabilité civile ne couvre pas les dommages relatifs à l'assuré, qu'ils soient corporels ou matériels.

Les garanties de dommages au véhicule : Ces garanties sont souscrites en plus de l'assurance au tiers et permettent de bénéficier d'une indemnisation lorsque le véhicule assuré est endommagé à la suite d'un sinistre. On distingue :

- **La garantie tous accidents tous risques** : couvre les dommages subis par le véhicule quelles que soient les circonstances de l'accident ou la responsabilité du conducteur.
- **La garantie dommages collision** : rembourse les dégâts causés au véhicule en cas de collision avec un tiers identifié tel qu'une autre voiture, un piéton ou un animal.
- **La garantie vol** : indemnise le propriétaire du véhicule en cas de vol.
- **La garantie incendie** : couvre tous les dommages causés au véhicule en cas d'incendie.
- **La garantie bris de glace** : couvre les dommages causés aux parties vitrées du véhicule tels que le pare-brise, les vitres latérales ou encore les rétroviseurs extérieurs.

Les garanties complémentaires : Ces garanties sont proposées en complément des garanties de dommages au véhicule. Peuvent s'y trouver :

- La garantie panne mécanique

- La garantie contenu du véhicule
- La garantie prêt de volant ou prêt du véhicule

La garantie personnelle du conducteur : Elle couvre les dommages corporels subis par le conducteur assuré, indépendamment de sa responsabilité dans l'accident. Cette garantie est souvent proposée dans les formules d'assurance tous risques.

Les garanties facultatives de services : L'assistance et la protection juridique sont des garanties supplémentaires qui peuvent être proposées par l'assureur. Les services fournis par l'assistance sont : le rapatriement du véhicule et de l'assuré en cas de panne ; le remplacement de pièces ; le prêt d'un véhicule de location et tout autre service prévu au contrat. La protection juridique quant à elle assure la couverture des frais de justice dans le cas d'une procédure de justice qui oppose l'assuré à un tiers.

1.1.4 La garantie Responsabilité Civile Matérielle

La responsabilité civile est l'obligation de réparer les dommages causés à autrui. Comme expliqué précédemment, la responsabilité civile est une garantie obligatoire pour tout véhicule terrestre motorisé. C'est une loi inscrite dans les articles "L211-1" du Code des Assurances et "L324-1" du Code de la route dont le non-respect constitue un délit passible d'une amende de 3 750 €.

La responsabilité civile se découpe en deux garanties : la responsabilité civile « matérielle » et la responsabilité civile « corporelle ».

- La responsabilité civile matérielle (RCM) est la garantie qui couvre les dégâts causés aux biens d'un tiers. Les sinistres encourus pour cette garantie sont fréquents, mais relativement peu coûteux.
- La responsabilité civile corporelle (RCC) est la garantie qui couvre les dégâts physiques causés à un tiers. Pour cette garantie, les sinistres sont moins fréquents, mais bien souvent plus coûteux.

Le portefeuille automobile étudié dans ce mémoire porte sur des contrats couvrant la responsabilité matérielle. Afin de simplifier et d'accélérer les procédures d'indemnisation en matière de sinistre automobile, les assureurs ont mis en place la convention IRSA [5]. Cette convention permet aux assurés, dans la plupart des cas, d'être dédommagés plus rapidement et directement par leur propre assureur, et ce, quel que soit leur degré de responsabilité.

La convention IRSA-IDA

La convention IRSA est définie comme la convention d'Indemnisation directe de l'assuré et de Recours entre Sociétés d'Assurance Automobile. Créée en 1968 sous l'appellation d'Indemnisation Directe des Assurés (convention IDA), elle change de nom en 1974. Cette convention est applicable aux accidents de circulation survenus en France et à Monaco, impliquant au moins deux véhicules terrestres à moteur (sous réserve que les deux organismes assureurs soient adhérents).

Il revient à l'assureur d'établir lui-même la responsabilité de son assuré, afin de l'indemniser des dommages et préjudices subis, avant de présenter un recours à l'assureur inverse. On distingue deux modes de recours : le recours forfaitaire et le recours réel.

- Le recours est forfaitaire lorsque le montant des préjudices matériels subis est inférieur à 6 500 €. Il est toujours proportionnel au niveau de responsabilité de l'auteur des dégâts. Le montant forfaitaire actuel en 2021 est de 1 678 € [1].
- Le recours est réel lorsque la valeur des dommages matériels dépasse 6 500 €. Dans ce cas, le niveau d'indemnisation correspond au montant réel des dommages subis.

Par exemple, considérons un sinistre non responsable d'un montant de 2 000 € dans le cadre de la convention IRSA en 2021. L'assureur aura un sinistre s'élevant à 1 678 € (montant forfaitaire en 2021) et l'assuré recevra une indemnité totale de son assureur à qui il restera à charge 322 € après recours.

Les contraintes réglementaires

L'assurance automobile est un secteur réglementé par plusieurs règles définies dans le code des assurances. L'une de ces contraintes est l'interdiction de différencier les tarifs en fonction des variables telles que le sexe de l'assuré. Il est donc interdit de distinguer les hommes des femmes lors des processus de tarification, même si la variable liée au genre de l'assuré s'est révélée discriminante. Cette décision a été prise par la Cour de Justice des Communautés Européennes (CJCE) dans un arrêt du 2 décembre 2012.

Il existe également des variables légales permettant de récompenser les bons risques et de pénaliser les mauvais risques. Le système de bonus-malus ou CRM (Coefficient de Réduction-Majoration) est un principe réglementé dans l'annexe de l'article A. 121-1 du Code des Assurances et qui consiste à récompenser le conducteur ou l'automobiliste qui se comporte bien sur la route en ne causant pas d'accident responsable. En effet, lors de chaque échéance annuelle du contrat, la prime due par l'assuré est déterminée en multipliant le montant de la prime de référence par le coefficient CRM, sur la base des sinistres impliquant la part de responsabilité. Le principe de calcul est le suivant :

- Le coefficient d'origine est 1 (pas de bonus-malus).
- Application d'une réduction de 5% par an sans accident responsable.
- Application d'une majoration de 25% par accident responsable.
- Application d'une majoration de 12, 5% par accident partiellement responsable.
- La réduction maximale est de 50%.
- La majoration maximale de 350%.
- La mise à jour est effective à la date anniversaire du contrat.

Tout assureur peut vérifier l'exactitude de la déclaration d'un nouveau client à partir du relevé d'information du contrat des assurances (RI), fourni par l'AGIRA (Association pour la Gestion des Informations sur le Risque Automobile).

1.2 Le véhiculier en assurance automobile

1.2.1 Utilité d'un véhiculier et objectifs de l'étude

En tarification IARD, une meilleure segmentation des risques passe principalement par l'introduction de nouvelles variables tarifaires telles que les zoniers (classification géographique) et les véhiculiers (classification de véhicules). Ces nouvelles variables permettent d'affiner la segmentation et de simplifier la structure tarifaire. En effet, elles permettent de regrouper toute l'information liée à la géographie ou à la catégorie de véhicule, sous forme de classes.

1. Les montants forfaitaires depuis 2014 sont présentés dans le tableau A.1 de l'annexe A.

En assurance automobile, la segmentation des risques passe notamment par la construction d'un véhiculier [2]. Le véhiculier est un regroupement de modèles ou versions de véhicules en groupes de risques homogènes permettant de capter plus d'informations afin de mieux discriminer le risque. En effet, le véhicule apparaît aujourd'hui comme un élément primordial de l'explication du risque pour la tarification a priori.

Plusieurs mémoires ont traité du sujet [3], notamment pour les garanties Dommages, Bris de Glace et Responsabilité Civile. De plus, il existe une classification de véhicules commune à tous les assureurs présentée dans la section suivante. Cependant, dans ce mémoire, nous construirons un véhiculier adapté à un portefeuille couvrant la Responsabilité Civile Matérielle (RCM). Cette étude nous permettra de mettre en évidence la pertinence d'un véhiculier pour la RCM, étant donné que ici le risque étudié porte sur les dommages causés à un tiers, et ne dépend donc pas uniquement de l'assuré.

Le véhiculier permet non seulement de simplifier les modèles de tarification, mais aussi de mieux expliquer le risque à travers une classification des véhicules par groupes de risques homogènes. Ainsi, les objectifs de ce mémoire sont multiples :

- Affiner la segmentation et simplifier la structure tarifaire en regroupant toute l'information liée au véhicule en une seule variable.
- Évaluer la pertinence d'un véhiculier pour la garantie Responsabilité civile Matérielle.
- Évaluer l'apport du véhiculier, à travers la comparaison de modèles avec véhiculier et de modèles sans véhiculier.

Dans la section suivante, nous présentons la classification de véhicule la plus utilisée en France.

1.2.2 Présentation du véhiculier SRA

En France, la classification de véhicules la plus utilisée en assurance automobile est celle fournie par l'association **Sécurité et Réparation Automobile** (SRA).

Créée en 1997, cette association professionnelle est dédiée à la constitution d'une base de données comportant toutes les informations existantes sur les véhicules de France et notamment un classement par caractéristiques techniques et commerciales. Cette base est mise à la disposition de tous les assureurs du secteur automobile.

Le classement SRA est un système de classement étoilé grâce auquel les professionnels utilisant la classification SRA peuvent très vite statuer sur la qualité des véhicules concernés. Ce classement concerne les véhicules à moteur de moins de 3,5 tonnes :

- Les 2 roues (moto, scooter).
- Les 3 roues (certains types de scooter).
- Les 4 roues (voitures, camionnettes…).

Étant un outil indispensable au secteur automobile, le classement SRA est majoritairement utilisé par les assureurs, mais aussi par les constructeurs automobiles et les fabricants d'accessoires automobiles. À chaque véhicule importé ou fabriqué sur le territoire français est attribué un code d'identification SRA que l'on retrouve sur la carte grise. Composé de chiffres et de lettres, le code SRA est propre à chaque marque, modèle et version de véhicule.

2. En tarification automobile, il est également possible d'intégrer des zoniers liés à l'adresse du domicile de l'assuré ou à son lieu de travail.

3. Ces différents mémoires ont été référencés dans la bibliographie et sont accessibles sur le site de l'Institut des actuaires.

Le code d'identification SRA permet aux assureurs de proposer un contrat d'assurance ainsi qu'un prix adapté aux véhicules des clients, car il permet de s'informer en détail sur les caractéristiques du véhicule (marque, version, modèle, valeur, puissance et autres informations tels que la disposition et la qualité des airbags). Les assureurs ont ainsi une idée de la fiabilité du véhicule et sont en mesure de décider du prix du contrat en fonction des caractéristiques techniques du véhicule. En effet, un véhicule doté d'un bon classement SRA est jugé plus sûr et s'assure plus facilement. Ce classement aide les assureurs à maîtriser et anticiper le coût des sinistres grâce à trois indicateurs : le groupe, la classe de prix et la classe de réparation.

- Le groupe est un indicateur représentant la puissance du véhicule, mais aussi sa dangerosité. Il reflète la dangerosité intrinsèque des véhicules, indépendamment du conducteur, de l'usage ou de la zone géographique. Le groupe SRA est noté de 20 à 50, de la plus faible puissance à la plus forte puissance.
- La classe de prix du véhicule est établie par la valeur à neuf TTC du véhicule (hors option, hors remise). Cette variable est notée de A à V (A = faible valeur, V = forte valeur, HC = Hors Classe). Elle est utile à l'assureur en cas de vol ou de remplacement d'un véhicule suite à un sinistre.
- La Classe de réparation est définie par le coût HT du panier de pièces SRA et des chocs 15Km/h. Elle donne des informations sur la réparabilité. Sa valeur est comprise entre A et ZE.

De plus, le classement SRA est utile aux constructeurs automobiles, car ces derniers peuvent analyser et déterminer les équipements et accessoires automobiles les plus utiles. Le classement SRA offre une expertise complète sur les véhicules, permettant aux gérants de l'industrie de se renouveler et d'améliorer les produits existants, en fournissant les informations sur : les options à privilégier, les options inutiles ou apportant peu d'intérêt, les véhicules les plus sécurisés. En plus d'être assuré plus facilement, un véhicule doté d'un bon classement se vend mieux.

Enfin, l'expertise du SRA a été très bénéfique pour le développement des stratégies antivol, car elle a permis la généralisation des équipements anti-démarrage sur les automobiles. Le classement SRA constitue donc un véritable référentiel autant pour assureurs que les constructeurs automobiles, car il est considéré comme étant le plus fiable pour définir les normes de sécurité automobile et pour minimiser le coût des accidents routiers.

Chapitre 2

Étapes de construction d'un véhiculier

Dans ce chapitre, nous allons expliquer le principe de construction des véhiculiers avant de développer l'approche qui sera mise en place dans ce mémoire.

2.1 Méthodologie

Le véhiculier a pour but de répondre à des besoins de segmentation, en captant le maximum d'information liée aux véhicules et non prise en compte par le modèle de tarification. Il s'agit de regrouper toute l'information résiduelle liée aux véhicules, en une seule variable, que l'on rajoutera a posteriori au modèle de tarification initial. Plusieurs approches de construction de véhiculier ont été développées à ce jour et s'inspirent généralement des méthodes de construction de zoniers géographiques [13]. Les zoniers sont développés sur un portefeuille spécifique, afin de lier un risque à sa localisation sous forme de classes. Ils sont habituellement numérotés ou nommés alphabétiquement, de la zone la moins risquée à la zone la plus risquée.

Ces différentes approches reposent généralement sur l'analyse résiduelle des modèles de tarifications, sans les variables véhiculières. En effet, les variables tarifaires utilisées pour l'élaboration de la prime pure, peuvent être décomposées de la manière suivante :

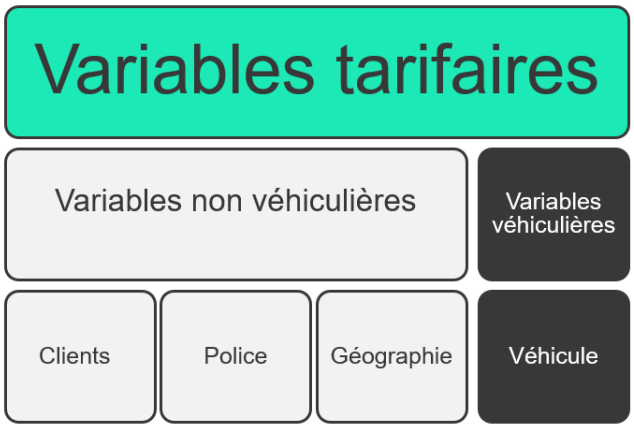


FIGURE 2.1 – Décomposition des variables tarifaires

Les variables non véhiculières : Dans ce mémoire, nous désignons par variables non véhiculières, toutes les variables liées à l'assuré, sa zone géographique et les caractéristiques de sa police. L'association de ces différentes variables permet d'obtenir "l'effet non véhicule", qui représente la part de sinistralité non liée au véhicule.

Les variables véhiculières : ce sont toutes les variables liées au véhicule (marque, carrosserie, version, vitesse maximale), permettant d'obtenir "l'effet véhicule". L'effet véhicule représente la part de la sinistralité due aux différences entre les caractéristiques de véhicule.

Les modèles de fréquence et de coût moyen composant la prime pure, prennent en compte une grande partie de ces effets (effet non véhicule et effet véhicule). Cependant, ces modèles ne contiennent pas toute l'information liée au véhicule, ce qui empêche d'avoir une segmentation optimale des véhicules. On suppose que les résidus de ces modèles contiennent outre du bruit, une part d'information liée aux véhicules qui n'a pas été expliquée.

L'information non prise en compte par les modèles de tarification classiques, constitue la part résiduelle qui n'est pas captée par ces modèles [17]. Le véhiculier est construit à partir de cette part résiduelle puis intégré dans le modèle afin d'apporter une part maximale de l'explication manquante.

Les différentes étapes de construction d'un véhiculier, peuvent être résumées dans le schéma suivant :

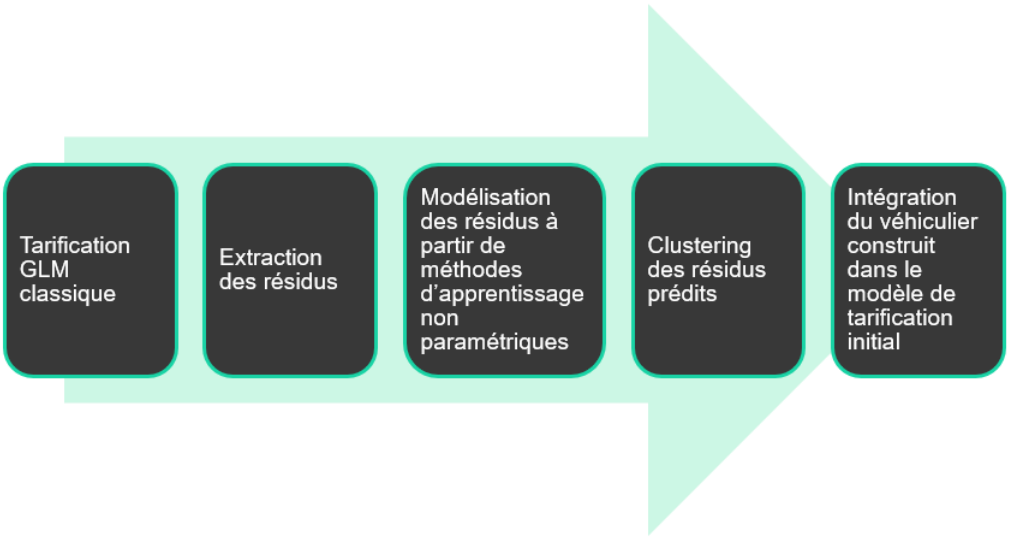


FIGURE 2.2 – Étapes de construction d'un véhiculier

La première étape consiste à construire un modèle tarifaire *fréquence x coût* à partir du modèle traditionnel de tarification, le modèle GLM.

Dans l'étape suivante, on extrait les résidus de ces modèles qui deviennent la variable cible dans la suite de l'étude. En effet, on considère que toute l'information non expliquée par le modèle initial se retrouve dans les résidus qui contiennent également un bruit lié aux variations aléatoires entre les observations.

Les résidus extraits sont ensuite prédits à l'aide de méthodes d'apprentissage non paramétriques (Arbre de régression CART, Random Forest, Gradient Boosting). Le score issu de cette modélisation peut subir un lissage spatial [14] ou être directement classifié à l'aide de méthodes de clusterisation. Dans ce mémoire, nous adoptons une approche directe.

Enfin, les classes issues du lissage seront intégrées dans le modèle de tarification en tant que variable tarifaire.

2.2 Démarche et apport du mémoire

Comme précisé précédemment, la construction du véhiculier repose sur la création d'une variable chargée de représenter tout le risque lié au véhicule. Dans ce mémoire, deux véhiculiers seront construits : l'un pour le modèle de fréquence et l'autre pour le modèle de coût.

La particularité de l'approche développée repose sur l'utilisation du Machine Learning. Nous appliquons en particulier deux méthodes : le Random Forest et le Gradient Boosting. À ce jour, peu de véhiculiers ont été construits sur la garantie RCM et les approches développées utilisent l'algorithme de régression CART et le Random Forest. Le Gradient Boosting repose sur la méthode du Boosting qui fournit des résultats plus robustes. De plus, l'objectif est de construire un véhiculier adapté au portefeuille étudié, afin d'évaluer son apport et sa pertinence à la garantie étudiée.

À cet effet, nous construisons un modèle de tarification le plus optimal possible en prenant le temps de comparer diverses méthodes pour les étapes préliminaires au GLM (discrétisation, sélection de variables), afin d'évaluer la vraie valeur ajoutée du véhiculier au modèle. Si le modèle de base est optimal, la valeur ajoutée du véhiculier sera d'autant plus pertinente.

Les étapes de construction du véhiculier, schématisés sur la figure 2.3, peuvent être résumés de la manière suivante :

- **Création d'un modèle hors variables véhiculières :** le coût et la fréquence de sinistre seront prédits uniquement avec les variables non véhiculières, à partir de deux GLM. Ceci nous permettra d'estimer les facteurs de risque non liés aux véhicules et d'isoler l'effet véhicule.
- **Extraction des résidus :** les résidus issus des modèles précédents seront modélisés par les variables véhiculières issues de la base SRA.
- **Modélisation des résidus prédits :** afin d'éviter de faire des hypothèses sur la distribution des données, nous allons utiliser le Random Forest et le Gradient Boosting qui sont des méthodes d'apprentissage non paramétriques. Ces deux méthodes seront comparées entre elles et celle qui réalise les meilleures performances sera retenue.
- **Agrégation et clusterisation des résidus prédits :** lors de cette étape, les résidus prédits seront agrégés suivant le code SRA afin d'attribuer à chaque véhicule un score unique. Nous allons ensuite regrouper les véhicules par groupes de risques homogènes en clusterisant le score prédit. La nouvelle variable représentant les classes de véhicules constituera le véhiculier.
- **Évaluation du véhiculier :** Enfin, le véhiculier sera intégré dans les modèles GLM puis nous évalueront son apport par rapport aux modèles sans véhiculier.

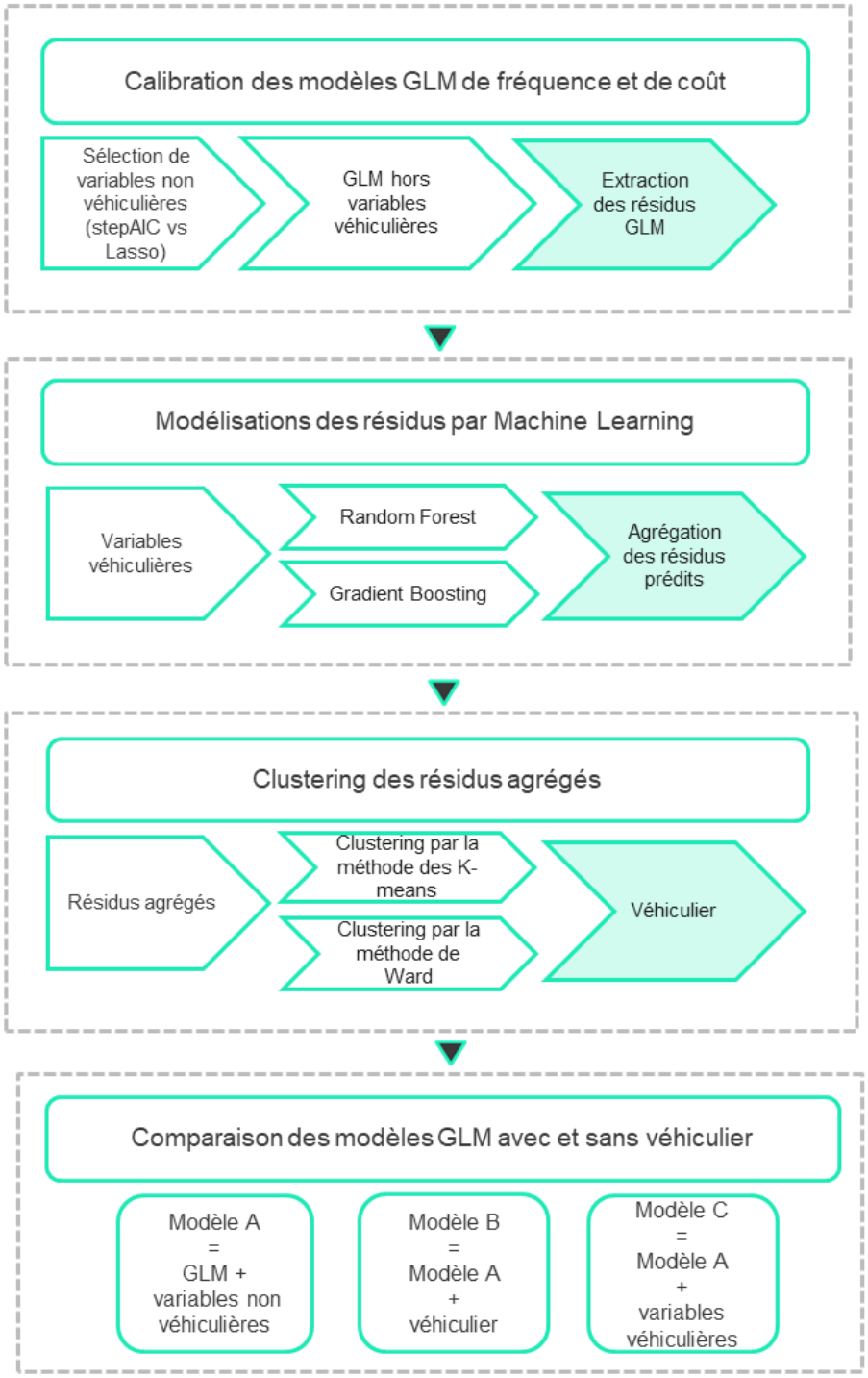


FIGURE 2.3 – Étapes de l'approche mise en œuvre

Découpage de la base de donnée

Lors de la mise en œuvre de modèles d'apprentissages statistiques, il est indispensable de découper la base de donnée en sous échantillons. La base est découpée en échantillon d'apprentissage et de validation. L'échantillon d'apprentissage représente l'échantillon sur lequel le modèle est entraîné, puis l'échantillon de validation est celui sur lequel le modèle est testé.

Dans le cadre de ce mémoire, 80% de la base initiale a été utilisée pour calibrer les modèles GLM tandis que les 20% restants ont été utilisés pour la validation de ces modèles. Ensuite, la modélisation des résidus par Random Forest et Gradient Boosting a été réalisée sur 80% de l'échantillon d'apprentissage précédent. Cet échantillon a également servi à la construction des véhiculiers de fréquence et de coût. Les modèles contenant les véhiculiers ont ensuite été validés sur l'échantillon de validation puis testés sur l'échantillon ayant servi pour la validation du GLM.

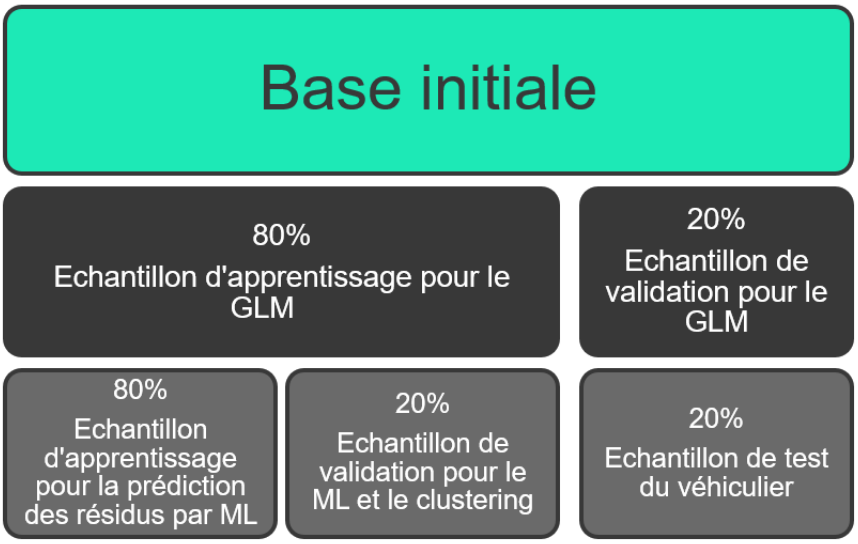


FIGURE 2.4 – Échantillonnage de la base d'étude

Chapitre 3

Traitement des données

Ce chapitre est consacré à la description des différentes bases de données disponibles à l'étude. Les différents traitements, analyses et contrôles de qualité qui ont été effectués sur ces bases, afin d'aboutir à une base unique et fiable, seront abordés.

3.1 Description des différentes bases de données

Pour l'étude, les bases de données disponibles sont une base Client et une base SRA en complément. Rappelons que le type de garantie que couvre le portefeuille étudié est la Responsabilité Civile Matérielle.

3.1.1 Base client

La base initiale comporte les informations relatives aux polices, aux clients ainsi que les informations de sinistralité associées. Elle contient également quelques caractéristiques des véhicules assurés. Elle est constituée de 561 592 observations, de 76 variables, avec une profondeur d'historique allant de 2014 à 2016.

Chaque ligne de la base représente une vision de contrat sinistré ou non, par année d'exercice. Chaque vision de contrat sera appelée *image*. À chaque modification de contrat (survenance de sinistre, suspension, résiliation, avenant, etc.), une nouvelle image est créée. Elle compte donc autant de lignes que de risque durant la période considérée, et non autant de lignes que de numéros de contrats.

Sur la figure 3.1 est présentée une liste des variables qui se trouvent dans la base.

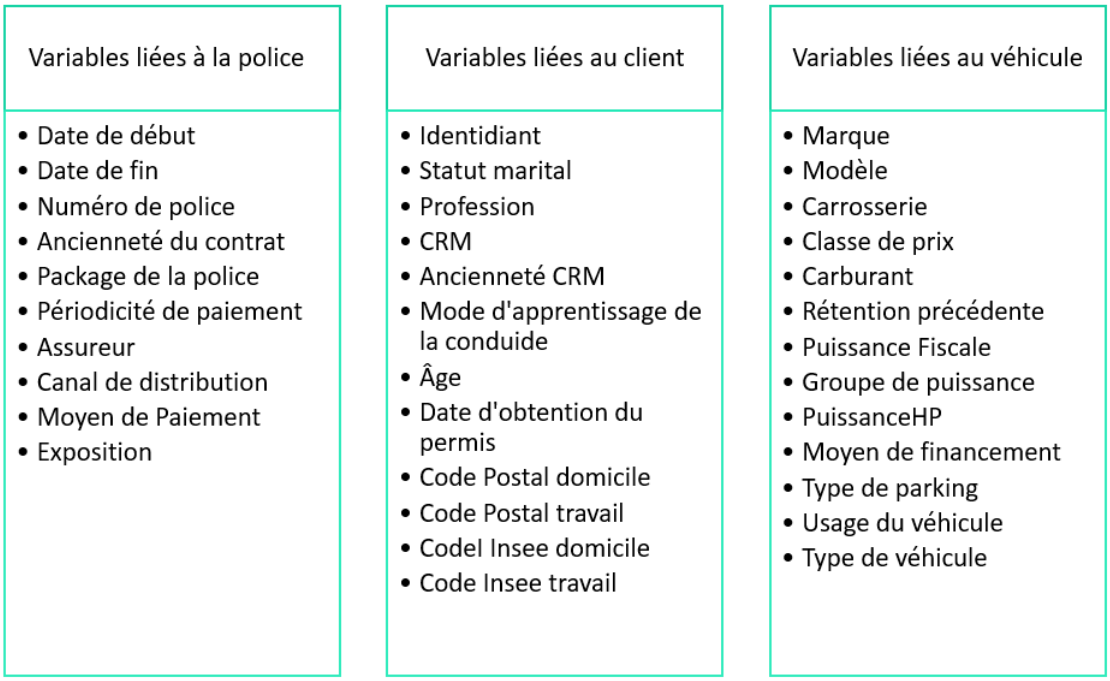


FIGURE 3.1 – Liste des variables de la base initiale

3.1.2 Base SRA

La base SRA est la base générée par l'organisme de Sécurité et Réparation Automobile (présentée en section 1.2.2) et mise à disposition de tous les assureurs automobile de la France. Elle recense toutes les caractéristiques techniques et commerciales des véhicules dans le but de maîtriser, le mieux possible, le coût des indemnisations des sinistres.

Elle permet de définir avec exactitude :

- Quels sont les véhicules ou accessoires les plus sûrs,
- Quelles sont les options à privilégier,
- Quelles sont les options inutiles ou celles qui apportent peu d'intérêt,
- Quels sont les modèles qui s'avèrent être les plus fiables en matière de sécurité.

La base SRA à notre disposition contient des informations de véhicules commercialisés entre 1936 et 2018. Elle est composée d'environ 100 000 lignes et de 90 variables dont plusieurs variables très corrélées ou mal renseignées (par exemple, la variable *charge utile* ayant environ 77 % de valeurs manquantes).

En annexe A se trouve un tableau récapitulatif des différentes variables qui composent la base SRA.

3.2 Jointure des bases

Le rapprochement des deux bases était une étape délicate qui a été réalisée en plusieurs étapes. Les principales difficultés rencontrées étaient les suivantes :

- L'absence d'une clé commune aux deux bases,
- Une écriture différente des modèles de voiture,

- La multiplicité des véhicules SRA.

Dans un premier temps, les erreurs d'écriture ont été analysées puis corrigées.

Ensuite, plusieurs clés ont été créées avec les variables véhicules communes aux deux bases. Les doublons des véhicules SRA pour chaque clé ont été supprimés, en gardant pour chaque doublon la dernière date de mise à jour.

Enfin, les bases ont été fusionnées successivement avec chaque clé, allant de la clé la plus précise à la clé la moins précise suivant le processus itératif suivant :

- Fusion i : on fusionne les deux bases avec la première clé.
- Fusion i+1 :
  - On récupère le sous-ensemble de la base n'ayant pas fusionné pour la 1ère clé et on crée de nouvelles clés, en enlevant à tour de rôle une variable de la clé précédente.
  - On réalise une deuxième fusion avec chacune des nouvelles clés et on retient la clé pour laquelle on a le plus grand pourcentage de lignes ayant fusionné.
- Fusion i+2 : On récupère à nouveau le sous-ensemble de la base n'ayant pas fusionné lors du rapprochement précédent et on reprend le même processus.

À la fin, chacune des sous-bases fusionnées ont été regroupées pour constituer la base finale. Seules 83 lignes de la base initiale n'ont pas pu être fusionnées.

Ci-dessous le récapitulatif des clés ayant été construites pour la jointure.

FIGURE 3.2 – Liste des clés de jointure

Des contrôles de qualité ont été réalisés en calculant, après chaque fusion, la proportion d'exposition totale récupérée ainsi que celle perdue par rapport à la base initiale.

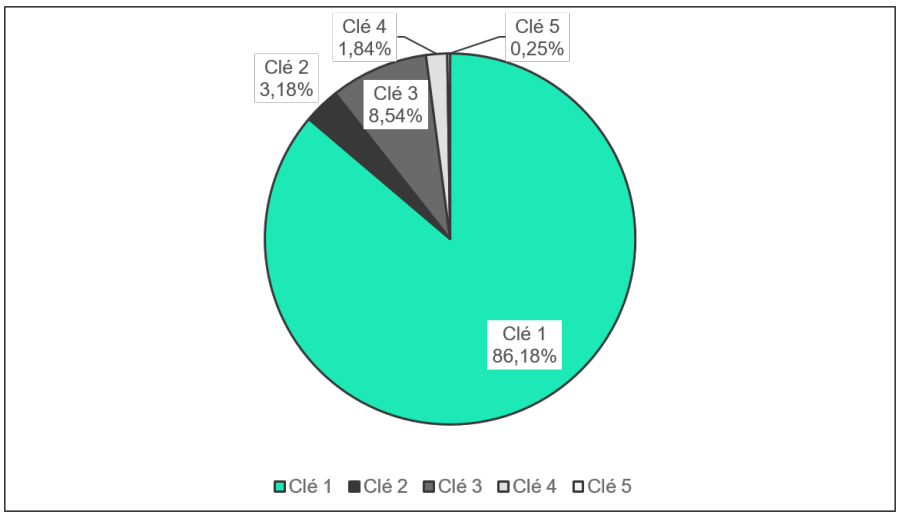


FIGURE 3.3 – Proportion d'exposition récupérée lors de chaque fusion

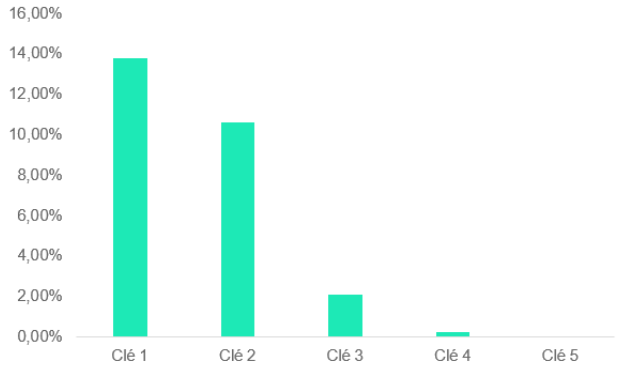


FIGURE 3.4 – Proportion d'exposition perdue lors de chaque fusion

La totalité de l'exposition initiale a été récupérée lors des fusions dont plus de 86 % récupérée lors de la première fusion. De plus, la proportion d'exposition n'ayant pas fusionné lors de chaque fusion est globalement faible.

En conclusion, la base finale est globalement représentative de la base initiale et la qualité des données reste fiable.

3.3 Traitements et contrôles de qualité

Plusieurs traitements (traitement des doublons, anonymisation,...) ont été effectués sur la base initiale ainsi que sur la base finale afin de fiabiliser les données.

Traitement de la base initiale

- Suppression des variables constantes.
- Suppression des lignes ayant une exposition nulle et un coût total de sinistre nul ou négatif.
- Suppression des doublons suivant une clé constituée à partir des variables suivantes : Numéro de police, date de début contrat, date de fin contrat et Coût total de sinistre.
- Anonymisation des modalités de variables liées à la police.

Traitement de base finale et tests de cohérence

Les principaux traitements effectués sur la base finale sont les suivants :

- Les variables de dimensions (hauteur, longueur, largeur) et la variable représentant le poids, sont mal renseignées pour les véhicules commercialisés avant 2010. Ces valeurs manquantes ont été remplacées par la moyenne selon la marque et la carrosserie.
- L'Assistance au Freinage d'Urgence (AFU) et le Système Anti-Blocage des roues (ABS) sont des variables non renseignées pour les véhicules commercialisés avant 2010. En effet, l'AFU et l'ABS constituent la nouvelle génération d'équipements de sécurité d'aide au freinage dont sont équipés les freins des véhicules neufs en série. Même si depuis 1998 tous les véhicules de la marque Mercedes-Benz sont équipés de l'AFU [1], ces variables ont plus de 20 % de valeurs manquantes dans notre base. Afin de ne pas perdre l'information relative à ces variables, la modalité "N-R" (Non-Renseigné) est attribuée, lorsque l'information est manquante.

Certains tests ont été réalisés afin de s'assurer de la cohérence des données :

- Vérification des sinistres ayant un coût total positif et pas de date de survenance.
- Vérification de la cohérence entre les dates de survenances de sinistres et la période de couverture.
- Vérification des individus de moins de 18 ans.

La base initiale était constituée de 561 592 lignes. La base finale après fusion et retraitement comporte 534 755 lignes, soit 95 % de la base initiale avec une exposition totale de plus de 200 000. Après traitement et analyse, la qualité de la base de données reste fiable. Néanmoins, nous déplorons la faible profondeur d'historique de données qui nous parait assez limitée ainsi que l'absence de données plus récentes.

1. https ://www.ornikar.com/code/cours/mecanique-vehicule/freins/aide-freinage

Chapitre 4

Études préliminaires

Ce chapitre décrit les études préliminaires qui ont été réalisées avant la mise en œuvre de l'approche. Dans un premier temps, une étude descriptive du portefeuille a été effectuée ainsi qu'une analyse de la stabilité temporelle. Ensuite, nous avons réalisé une étude des sinistres forfaitaires ainsi qu'un écrêtement des sinistres graves. Pour finir, une analyse de la sinistralité en fonction des variables tarifaires a été réalisée.

4.1 Étude du portefeuille

Le portefeuille étudié a eu au total 61 078 sinistres et au moins 68% des polices de la base totale n'ont pas eu de sinistres. La fréquence moyenne est de 10 % et le coût moyen s'élève à 1 172 €. Nous sommes en présence d'un portefeuille sur-sinistré, car on y retrouve des polices ayant eu jusqu'à huit sinistres par an, pour 274 000 polices. Cette fréquence de sinistralité semble élevée par rapport à celle observée sur le marché en 2016. En effet, en 2016, la fréquence moyenne de la garantie RCM en France était de 3.39 %[1]. Il faut noter qu'il s'agit de deux bases différentes, car les chiffres clés sont donnés à l'échelle nationale. De plus, la base étudiée est marquée par une sur-représentation de sinistres à faibles montants, et de sinistres forfaitaires (figure 4.2). Dans la suite de ce chapitre, la stabilité temporelle du portefeuille sera vérifiée et les sinistres forfaitaires de la convention IRSA seront étudiés.

4.2 Stabilité temporelle

De manière générale, la sinistralité est étudiée par année de survenance du sinistre. En effet, les indicateurs de sinistralité se stabilisent en tenant compte du recul par rapport au 1^{er} janvier de l'année de survenance. Par exemple, les sinistres de la garantie responsabilité civile corporelle peuvent prendre des années avant de se stabiliser.

La base de données étudiée est vue courant 2017, au 30/06/2017. Elle dispose d'une profondeur d'historique allant de 2014 à 2017. Cependant, les contrats survenus en 2017 représentent une proportion négligeable des contrats sinistrés (0.27 %). De plus, nous estimons que nous n'avons pas assez de recul afin de prendre en compte les sinistres tardifs de cette année. Ces contrats ne seront donc pas pris en compte dans la modélisation de la sinistralité.

Le graphique 4.1 mets en évidence la stabilité temporelle de la fréquence et de la sévérité au cours du temps. La répartition des expositions est uniforme sur les années tandis que la fréquence de sinistre

1. Données clés 2016 disponibles sur le site de la Fédération Française de l'Assurance (FFA).

et le coût ont une tendance décroissante. On remarque également l'exposition quasi négligeable de l'année 2017. Par ailleurs, on ne dénote pas de fortes variations des indicateurs de sinistralité dans le temps. Il n'y a donc pas eu de changements majeurs dans la gestion des sinistres.

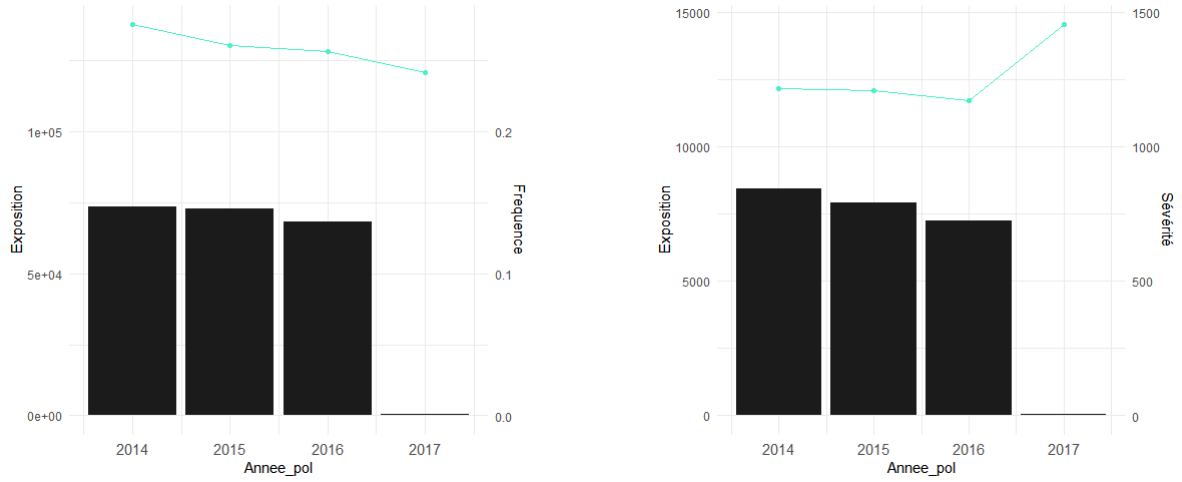


FIGURE 4.1 – Stabilité temporelle

4.3 Étude des sinistres forfaitaires de la convention IRSA

Le graphique ci-dessous présente la distribution des coûts moyens des sinistres du portefeuille. La distribution s'apparente à la distribution d'une loi gamma avec une forte représentation des sinistres à faibles coûts. Par ailleurs, on dénote une sur-représentation des montants de sinistres entre 1 200 € et 1 400 €. Mais la base étudiée ne renseigne aucune information sur les sinistres IRSA, c'est-à-dire les sinistres conventionnés ou non. Quant à la variable renseignant sur le taux de responsabilité, elle est mal renseignée avec plus de 97 % de valeurs manquantes. Ce qui ne permet pas d'identifier les demi-forfaits pour les sinistres à responsabilité partagée. Nous allons donc supposer que les montants de sinistres sur-représentés correspondent aux montants forfaitaires IRSA. Ces sinistres représentent environ 9 % de la charge de sinistre totale.

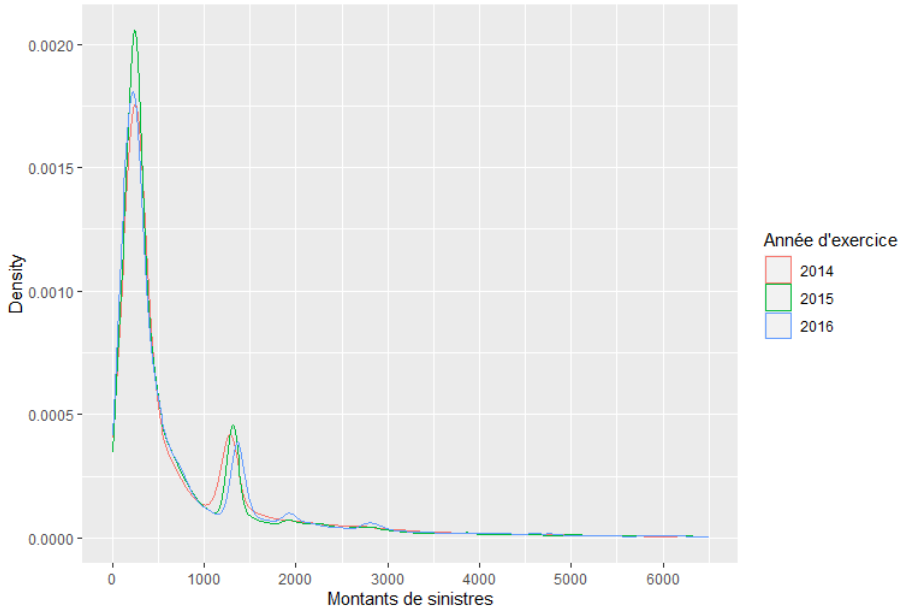


FIGURE 4.2 – Distribution annuelle des coûts moyens de sinistres

Les figures 4.3 et 4.4 représentent la fréquence moyenne des sinistres forfaitaires et celle des sinistres non forfaitaires suivant respectivement, l'ancienneté du véhicule et le groupe SRA.

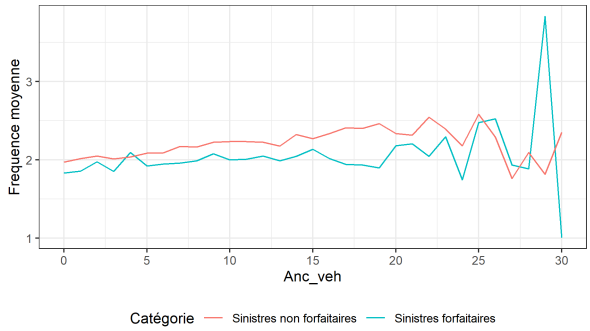


FIGURE 4.3 – Fréquence moyenne selon l'ancienneté du véhicule

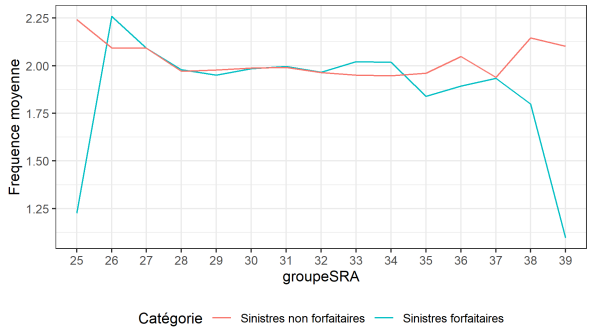


FIGURE 4.4 – Fréquence moyenne suivant le groupe SRA

Les deux catégories de sinistres présentent la même tendance et la fréquence des sinistres non forfaitaires est globalement plus élevée que celle des sinistres forfaitaires. Ce constat s'explique par le fait que les sinistres forfaitaires aient été identifiés grâce à la distribution des coûts moyens de sinistres et ne représentent qu'une faible proportion de la sinistralité totale.

Le graphique représentant la fréquence moyenne suivant l'ancienneté du véhicule montre que les véhicules récents génèrent le plus de sinistres non forfaitaires. La tendance est inversée à partir de 25 ans où les véhicules les plus âgés génèrent plus de sinistres forfaitaires. En effet, les véhicules anciens dits "véhicules de collection" sont très peu représentés et circulent moins. Ils auront donc tendance à générer des sinistres forfaitaires. Par ailleurs, les véhicules dont le groupe SRA est proche de 50 génèrent plutôt des sinistres non forfaitaires. Rappelons que le groupe SRA, noté de 20 à 50, représente la puissance du véhicule, mais également sa dangerosité. Il est donc cohérent que les véhicules de forte puissance estimés plus dangereux génèrent des montants de sinistres élevés (non forfaitaires).

L'analyse graphique comparant les sinistres forfaitaires aux sinistres non forfaitaires est insuffisante, pour mettre en évidence les types de profils qui génèrent chaque catégorie de sinistre. Cependant, il convient de faire une distinction entre les montants forfaitaires et les montants non forfaitaires. À cet effet, les montants forfaitaires seront retirés de la modélisation du coût moyen puis rajoutés plus tard à la prime pure globale prédite selon la formule suivante :

$$\text{Prime Pure} = \text{Fréquence moyenne} * (\text{Coût moyen hors forfaits} + \text{Coût moyen forfaitaire})$$

Par ailleurs, les sinistres de montants nuls ne seront pas étudiés, car il s'agit potentiellement des sinistres pour lesquels l'assuré n'est pas responsable (taux de responsabilité égal à zéro). De plus, il est important de faire remarquer que les montants de sinistres sont nets de franchise, car il n'existe pas de franchises sur la garantie Responsabilité Civile Matérielle [2].

2. Une franchise peut être prévue en cas de « prêt de volant », lorsque l'assuré prête son véhicule à un jeune conducteur ou à un conducteur novice, ou en cas de « conduite exclusive », lorsque le véhicule est conduit par une autre personne que le souscripteur et les conducteurs désignés au contrat, en contrepartie d'une tarification préférentielle[11].

4.4 Écrêtement des sinistres graves

Théorie des valeurs extrêmes

En assurance automobile, les classes de risques constituées à partir de caractéristiques de l'assuré et du véhicule en vue de la segmentation, sont supposées homogènes en termes de sinistralité. Cependant, cette hypothèse d'homogénéité des classes n'est pas toujours vérifiée en raison de la présence de sinistres graves dans le portefeuille. Pour pallier ce problème, les assureurs effectuent des écrêtements et répartissent la charge sur l'ensemble du portefeuille. En d'autres termes, les sinistres sont plafonnés à un niveau maximum.

Le choix de ce plafond peut être délicat et peut conduire à une sous-estimation ou à une surestimation des sinistres ordinaires. Ce qui aurait pour conséquence une mauvaise représentation de la sinistralité réelle par prime pure estimée. Une mauvaise estimation de la prime pure entraînerait soit à de l'anti-sélection, soit à des tarifs trop élevés et donc peu compétitifs.

Il existe plusieurs méthodes permettant d'obtenir des estimations fiables de ce plafond. Ces méthodes dites de seuillage sont issues de la théorie des valeurs extrêmes[16]. Elles permettent la détermination d'un seuil d'écrêtement au-delà duquel un événement est considéré comme atypique.

La théorie des valeurs extrêmes concerne l'étude du maximum d'une distribution et de sa loi. Elle a été développée pour l'estimation de la probabilité d'occurrence d'événements rares et permet d'obtenir des estimations fiables des valeurs extrêmes, pour lesquelles les observations sont peu nombreuses.

En pratique, le seuil est déterminé graphiquement en utilisant le graphique Quantile-Quantile (QQ-plot), outil permettant de comparer deux distributions que l'on estime semblables. Il nous permettra de vérifier l'adéquation à une loi GPD (Generalized Pareto Distribution) ou à une loi à queue épaisse.

Le graphique QQ-plot permet d'obtenir la forme de la queue de la distribution. Trois cas de figure sont possibles :

- Les données suivent la loi exponentielle : la distribution présente une queue très légère, les points du graphique présentent une forme linéaire.
- Les données suivent une distribution à queue épaisse « fat-tailed distribution » : le graphique QQ-plot est concave, cela revient à la présence d'un grand nombre de valeurs extrêmes au niveau de la queue de la distribution.
- Les données suivent une distribution à queue légère « short-tailed distribution » : le graphique QQ-plot a une forme convexe. Le nombre de valeurs extrêmes est faible.

Le QQ-plot des données représenté sur la figure 4.5 montre que nous sommes en présence d'une distribution à queue épaisse. De plus, on remarque que certains sinistres ont des coûts très élevés. Il pourrait s'agir de sinistres de carambolages, sinistres impliquant plusieurs voitures ou de sinistres de dommages au bien public.

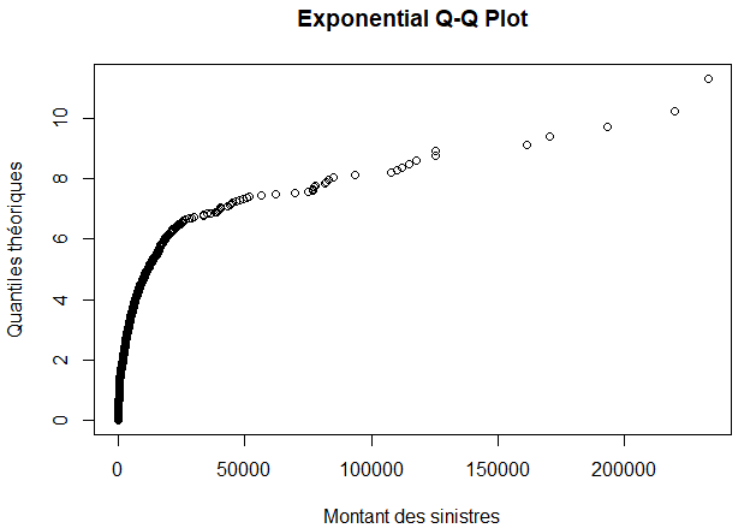


FIGURE 4.5 – Graphique Quantile-Quantile

Un autre outil graphique permettant de déterminer l'allure de la queue d'une distribution ainsi que le seuil à utiliser est la "*mean excess function*" définie par :

$$e(u) = E(X - u/X > u), u > 0$$

Cette méthode repose sur le comportement des valeurs observées X au-delà d'un seuil u donné. La fonction des excès moyens $e(u)$ est l'espérance de tous les sinistres supérieurs à u après leur avoir retranché la valeur u. À partir d'un certain seuil, la fonction des excès moyens doit devenir linéaire. Cela permet d'avoir une première idée concrète du niveau du seuil à partir duquel nos observations s'identifient à une loi des extrêmes. Cependant, cette méthode ne permet pas d'obtenir une valeur unique pour le seuil, car sur le graphique 4.6, nous constatons plusieurs plages de linéarité. Nous retenons la première valeur à partir de laquelle la fonction devient linéaire. Soit un seuil de 30 000€. Environ 99% des sinistres ont une charge inférieure à 30 000€ et les 1% restant représentent 13% de la charge de totale.

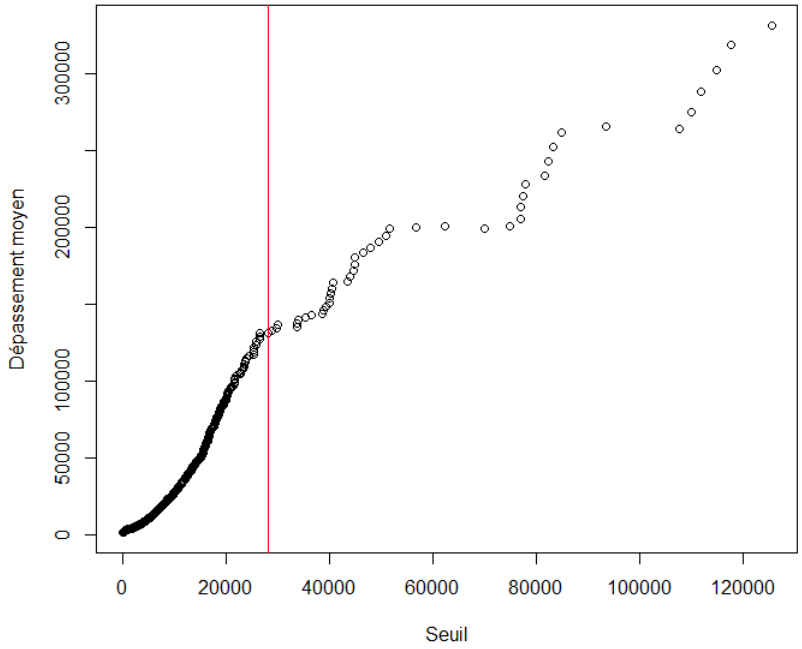


FIGURE 4.6 – Dépassement moyen en fonction du seuil

4.5 Analyses univariées

L'analyse univariée permet de visualiser l'évolution de la fréquence et du coût de sinistre suivant les variables possiblement tarifaires. Elle mettra également en évidence les segments les plus risqués du portefeuille ainsi que leur exposition.

Les graphiques d'analyse se décomposent en deux parties. La partie basse représente l'exposition de chaque modalité de la variable étudiée tandis que la partie supérieure décrit la fréquence de survenance du sinistre. Les expositions ont été multipliées par un coefficient afin de d'être mises à la même échelle que la fréquence. Ci-dessous sont présentées quelques analyses univariées. D'autres sont disponibles en annexe B.

Âge de l'assuré

Le graphique ci-dessous présente l'évolution de la fréquence et du coût moyen en fonction de l'Âge de l'assuré. L'âge moyen de la base est de 43 ans. La fréquence de sinistre est globalement décroissante de l'âge. Les classes d'âge les plus jeunes semblent être les plus à risque et les conducteurs les plus expérimentés semblent moins risqués. Cependant, on peut remarquer que la fréquence du sinistre augmente à nouveau au niveau des âges élevés.

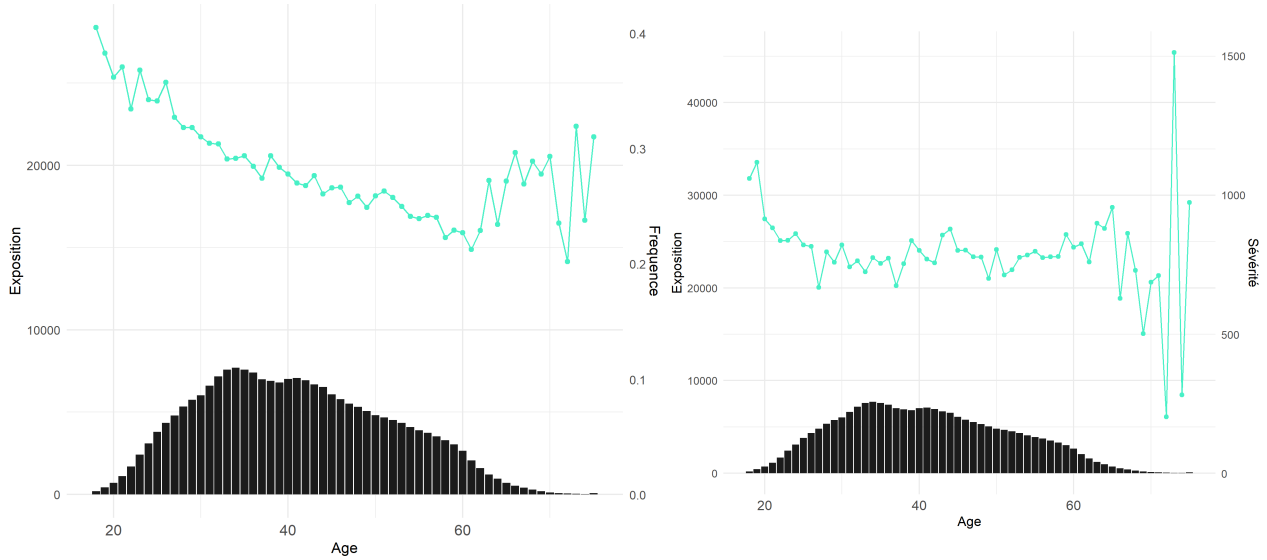


FIGURE 4.7 – Analyse univariée de l'âge de l'assuré

Ancienneté du permis de l'assuré

Comme nous l'avions observé lors de l'analyse de l'âge, la fréquence de sinistre est décroissante de l'ancienneté du permis entre 0 et 40 ans et croissante à partir de 40 ans. De plus, les nouveaux conducteurs semblent être les individus les plus risqués du portefeuille.

L'analyse de l'ancienneté du permis est similaire à celle de l'âge de l'assuré, car cette variable a été calculée à partir de la date de naissance de l'assuré. Nous remarquerons sans doute que ces deux variables sont corrélées dans la section 4.7 consacrée à l'étude de la corrélation entre les variables.

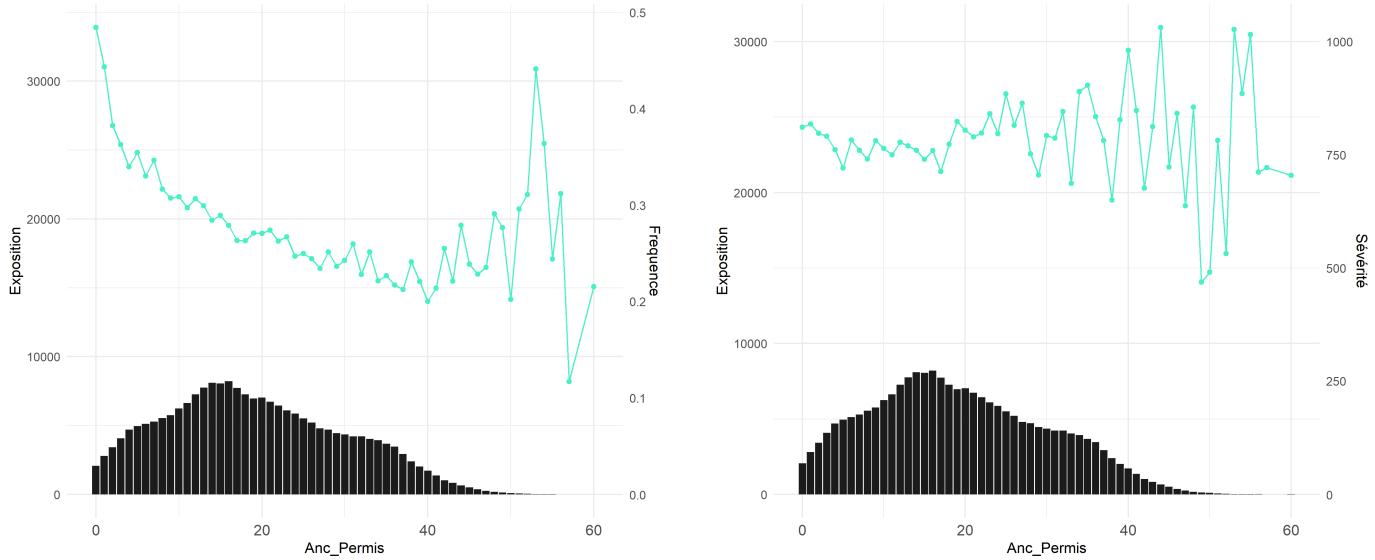


FIGURE 4.8 – Analyse univariée de l'ancienneté du permis de l'assuré

CRM (Coefficient de Réduction-Majoration)

Le graphique ci-dessous présente l'évolution de la sinistralité en fonction de la variable CRM (présentée en section 1.1.4). La fréquence de sinistre est croissante du CRM. En effet, les bons conducteurs ont un CRM à 50% et les mauvais conducteurs ont un CRM majoré à 350%. Par ailleurs, on n'observe aucune tendance particulière de la sévérité en fonction du CRM.

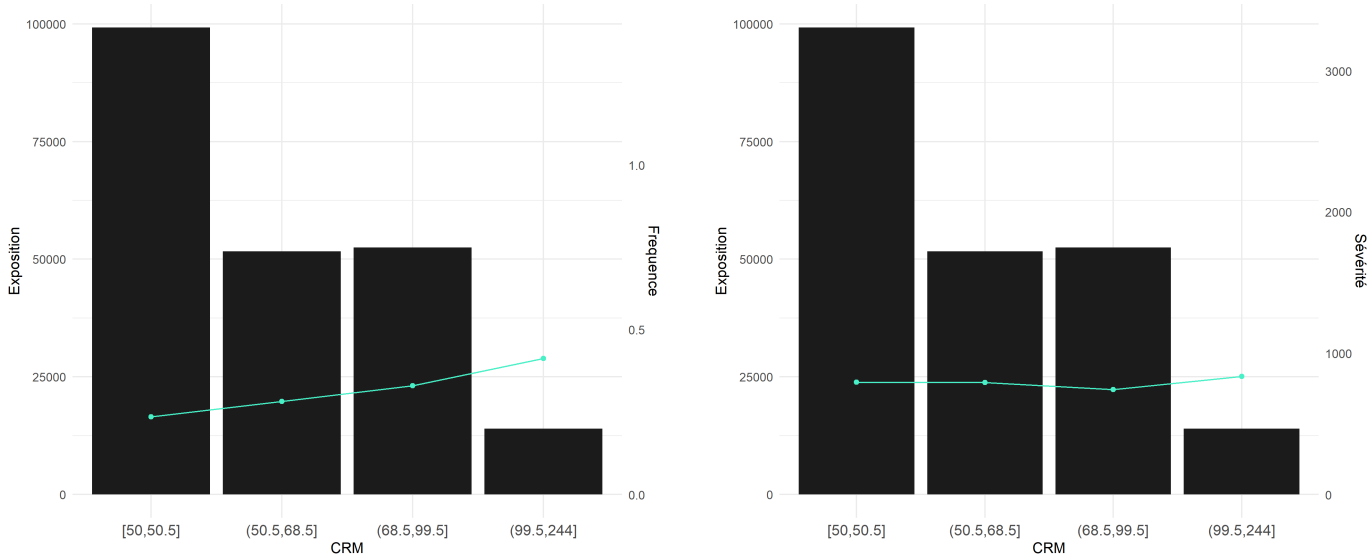


FIGURE 4.9 – Analyse univariée du CRM

Classe de prix du véhicule

Les figures 4.10 et 4.11 présentent l'évolution de la sévérité en fonction du groupe SRA et de la classe de Prix du véhicule. On observe que le coût moyen a une tendance croissante pour chacune de ces variables. Le groupe SRA (cf. section 1.2.2) représente la dangerosité intrinsèque d'un véhicule et le portefeuille semble très exposé aux véhicules de groupe élevés (30 à 34). La sinistralité chute pour les véhicules de très forte puissance (39 et plus), qui sont très peu représentés dans le portefeuille. Le même constat est observé pour la classe de Prix des véhicules.

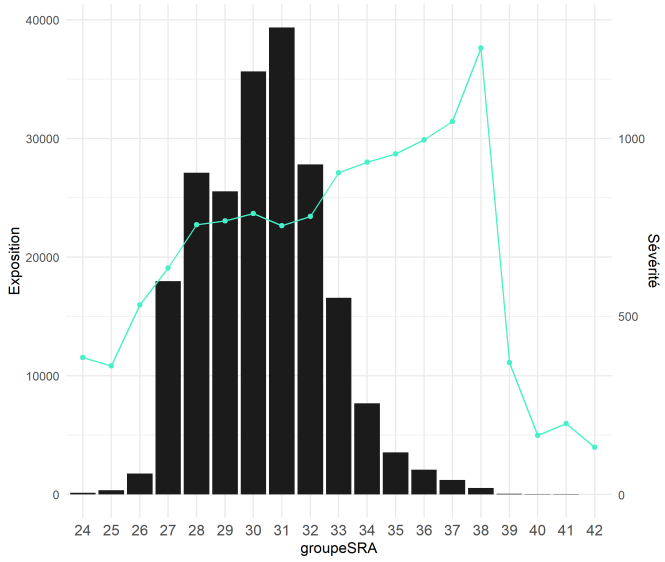


FIGURE 4.10 – Sévérité prédite en fonction du groupe SRA

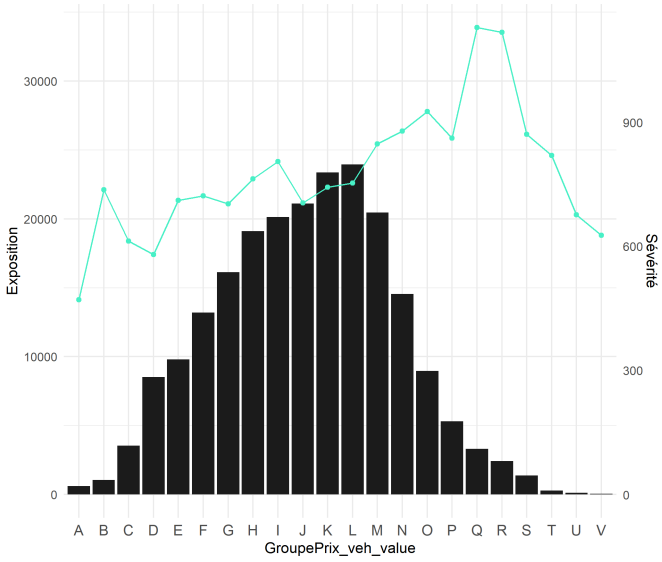


FIGURE 4.11 – Sévérité prédite en fonction de la classe de Prix

4.6 Discrétisations et Regroupements

Les étapes de discrétisations et de regroupements sont essentielles pour réduire la complexité des modèles. On cherche à calibrer un nombre limité de paramètres dans le modèle. Chaque niveau de

variable catégorielle différent du niveau de référence, est un paramètre si non groupé. De plus, une variable non significative, mais ayant un nombre de modalités importantes, peut être retenue dans le modèle alors qu'elle ne devrait pas l'être. Il est donc nécessaire de simplifier les variables granulaires grâce à des méthodes de classification/regroupement. Cette étape permet également de regrouper les modalités ayant de faibles expositions.

Il existe plusieurs méthodes de discrétisations telles que la méthode des quantiles, la méthode des amplitudes et les arbres de classification. Les arbres de régression (cf. section 6.1.1) présentent l'avantage de créer des classes homogènes en termes de sinistralité, quelle que soit la nature de la variable (quantitative ou qualitative). C'est donc la méthode de discrétisation utilisée dans ce mémoire.

Toutes les variables d'intérêt ont été classifiés sur l'échantillon d'apprentissage avant la calibration de chaque modèle afin d'éviter le sur-apprentissage.

Les variables clients [3] qui ont été classifiées sont les suivantes :

- Âge de l'assuré
- CRM
- Ancienneté du CRM
- Catégorie socioprofessionnelle (CSP)
- Statut Marital du client

Dans la section suivante, les résultats des discrétisations seront analysées afin de s'assurer de la cohérence des regroupements.

Discrétisation de l'âge de l'assuré

L'âge est une variable continue qui n'observe pas d'évolution linéaire. En fait, l'effet de l'âge sur la fréquence, illustré en Figure 4.7, présente un comportement différent, selon qu'il s'agisse de jeunes conducteurs ou de personnes plus âgées. Le résultat de la discrétisation est présenté sur la figure ci-dessous.

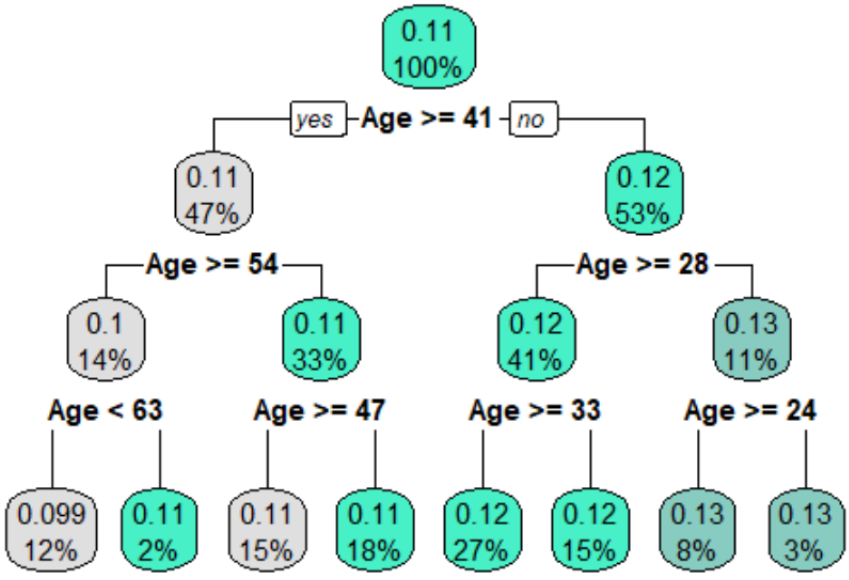


FIGURE 4.12 – Discrétisation de l'âge par arbre de régression

3. Les variables véhicules intervenant dans nos modèles ont également été regroupés.

La variable a été découpée en 8 classes. Le premier découpage est effectué à l'âge de 41 ans. Les nœuds gauches de l'arbre regroupent tous les découpages appliqués aux individus d'âge supérieur tandis que les nœuds droits regroupent les individus d'âge inférieurs.

Une fois ces classes constituées, on s'intéresse à la fréquence de sinistralité au sein de ces dernières ainsi qu'à leur effectif en terme d'exposition.

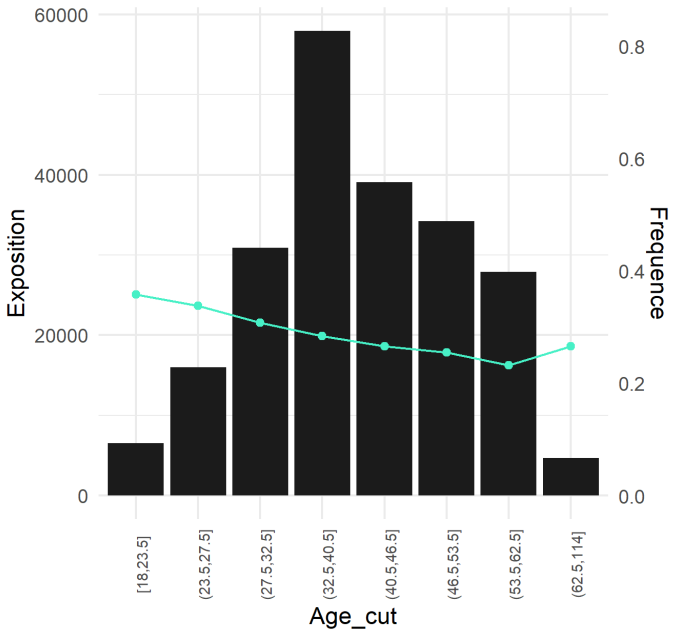


FIGURE 4.13 – Analyse univariée de l'âge discrétisée

La fréquence de sinistre est toujours décroissante de l'âge de l'assuré avec une recroissance au niveau de la dernière classe. Ce découpage est donc cohérent avec l'analyse de l'âge faite précédemment. De plus, chaque classe est bien représentée à travers son exposition.

Discrétisation du CRM et de son ancienneté

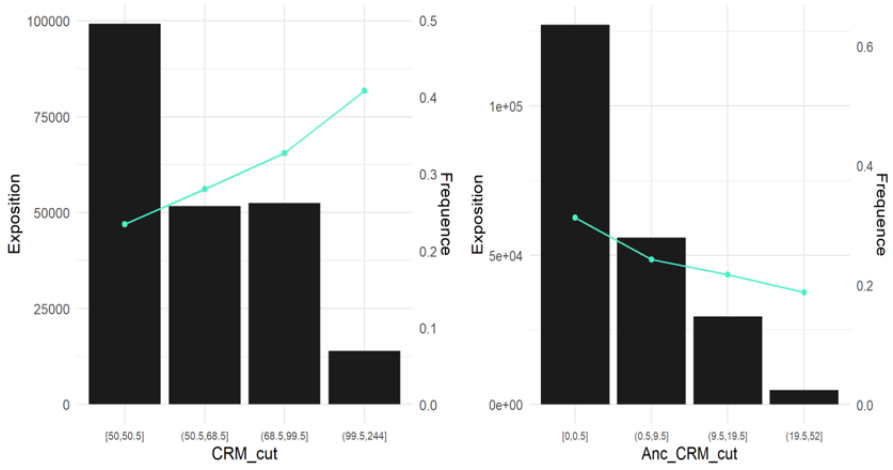


FIGURE 4.14 – Analyse de la discrétisation du CRM et de son ancienneté

Le découpage du CRM et de son ancienneté semble pertinent, car un coefficient CRM à 50 représente les meilleurs conducteurs et ne peut pas être regroupé avec d'autres coefficients. De même,

les nouveaux conducteurs ayant une ancienneté de CRM nulle ne peuvent être regroupés avec les autres conducteurs.

En outre, la fréquence de sinistralité est croissante avec le CRM et décroissante avec son ancienneté. Ce qui semble cohérent avec l'analyse univariée du CRM faite en section 4.5. En effet, un nouveau conducteur est vraisemblablement plus risqué et a une ancienneté de CRM nulle donc une fréquence de sinistre plus élevée par rapport aux conducteurs plus expérimentés.

Discrétisation de la catégorie socioprofessionnelle et de la Classe de Prix

Les variables CSP (catégorie socioprofessionnelle) et Classes de Prix sont des variables qualitatives ayant un nombre de modalités trop élevé : 15 pour la variable CSP et 26 pour la classe de Prix. De plus, l'effectif de chacune des classes n'est pas suffisant pour obtenir un modèle robuste. Un regroupement des modalités de ces classes s'impose donc.

L'arbre de régression effectue un regroupement en trois classes pour la variable CSP et en quatre classes pour la Classe de Prix.

Variable Catégorie socio-professionnelle	
Classe 1	Agriculteur, éleveur Armée, police, pompier Ouvrier Salarié cadre sup/dirigeant Retraité Sans profession
Classe 2	Enseignant Fonctionnaire cadre Salarié non cadre
Classe 3	Etudiants Profession libérale ou médicale Artisan, commerçant Taxi/VTC Fonctionnaire non cadre Salarié cadre

Variable ClassePrix_veh	
Classe 1	A,B,C,D,D,E,F,HORS CLASSE
Classe 2	G, H,I
Classe 3	J,K,L
Classe 4	M,N,O,P,Q,R,S,T,U,V

FIGURE 4.15 – Classes de CSP et Classe de Prix

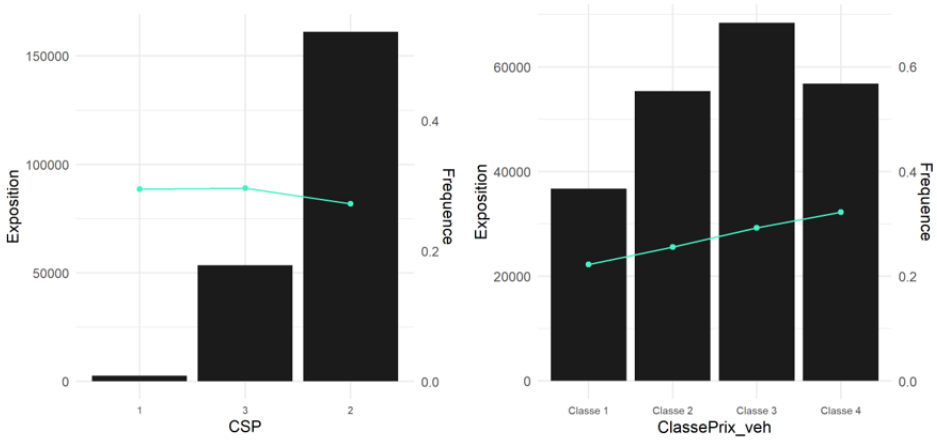


FIGURE 4.16 – Analyse de la discrétisation des variables CSP et Classe de Prix

L'analyse graphique de la variable CSP ne met pas en évidence une forte variation de la fréquence en fonction des classes créées. Les individus des classes 1 et 3 (agriculteurs, ouvriers, étudiants,

etc.) présentent le même niveau de risque pour la fréquence et sont les moins représentés dans le portefeuille, tandis que les individus de la classe 2 sont les moins risqués et les plus représentés. Par ailleurs, la fréquence présente une tendance linéaire et croissante pour les classes de prix.

L'analyse de la Figure 4.16 nous permet d'observer des classes de risques homogènes avec un volume d'individus suffisant par classes.

4.7 Étude de la corrélation

L'étude de la corrélation permet d'identifier les variables fortement corrélées entre elles. Les variables redondantes seront retirées du modèle afin d'éviter tout biais. En effet, les GLM ne prennent pas en compte les interactions entre les différentes variables (sauf s'ils sont rajoutés manuellement). Il est donc important de veiller à ce que les variables explicatives ne soient pas fortement corrélées entre elles, car cela aurait pour conséquence de fausser les résultats du modèle.

La méthode de corrélation utilisée dans ce mémoire est le V de Cramer. Cette méthode permet d'affecter une corrélation entre deux variables qualitatives. Pour cela, toutes les variables quantitatives ont été discrétisées en classes et sont considérées comme des variables qualitatives. Deux variables sont dites fortement corrélées lorsque la valeur absolue de leur corrélation est proche de 1.

En raison du nombre important de variables dont nous disposons, nous avons d'abord étudié la corrélation entre les variables clients avant d'étudier celle entre les variables véhicules. Cependant, une étude de la corrélation entre les variables non véhiculières et les variables véhiculières a révélé une absence de corrélation entre les deux catégories de variables.

Le V de Cramer

La mesure du V de Cramer se base sur le test d'indépendance du χ^2 développé par Karl Pearson en 1900 [15]. Elle permet d'évaluer l'existence d'une relation entre deux variables qualitatives. Les variables quantitatives sont assimilées à des variables qualitatives avec un grand nombre de modalités. Cependant, la valeur du χ^2 ne permet pas de quantifier la dépendance entre deux variables, car cette mesure varie entre 0 et $+\infty$. D'où l'intérêt du V de Cramer qui permet de normaliser cette valeur, maintenant comprise entre 0 et 1.

Formule du V de Cramer : $V = \sqrt{\frac{\chi^2}{n(\min(k-1,r-1))}}$

Le test d'indépendance du χ^2

Il existe plusieurs tests du χ^2 : le test du χ^2 d'adéquation, celui d'homogénéité et celui d'indépendance. Le test d'intérêt ici est le test du χ^2 d'indépendance, car il permet de vérifier à partir d'un échantillon l'hypothèse « Ho : les variables X et Y sont indépendantes ».

Soient X et Y deux variables qualitatives avec comme nombre de modalités respectives k et r $(k, r \geq 2)$ On observe le tableau de contingence suivant :

	Y_1	Y_2	...	Y_r	Total
X_1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,r}$	$n_{1,.}$
X_2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,r}$	$n_{2,.}$
...	...	...	...	...	...
X_k	$n_{k,1}$	$n_{k,2}$	...	$n_{k,r}$	$n_{k,.}$
Total	$n_{.,1}$	$n_{.,2}$	...	$n_{.,r}$	n

On a alors. $\chi^2 = \sum_{i,j} \frac{\left(n_{i,j} - \frac{n_{i,\cdot} n_{\cdot,j}}{n}\right)^2}{\frac{n_{i,\cdot} n_{\cdot\cdot\cdot}}{n}}$

Dans le but de vérifier l'hypothèse Ho on compare le résultat obtenu au fractile $\chi^2_{1-\alpha}[(k-1)(r-1)]$ [4]. En pratique, on prend $\alpha = 5\%$ pour délimiter la région critique du test. Lorsque $\chi^2 > \chi^2_{1-\alpha}[(k-1)(r-1)]$, l'hypothèse Ho est rejetée.

Corrélation entre les variables non véhiculières

On n'observe aucune corrélation entre les variables clients (variables non véhiculières) sauf entre l'ancienneté du permis et l'âge du client. Ce constat est logique, car l'ancienneté du permis a été calculé à partir de l'âge de l'assuré et de la date d'obtention du permis.

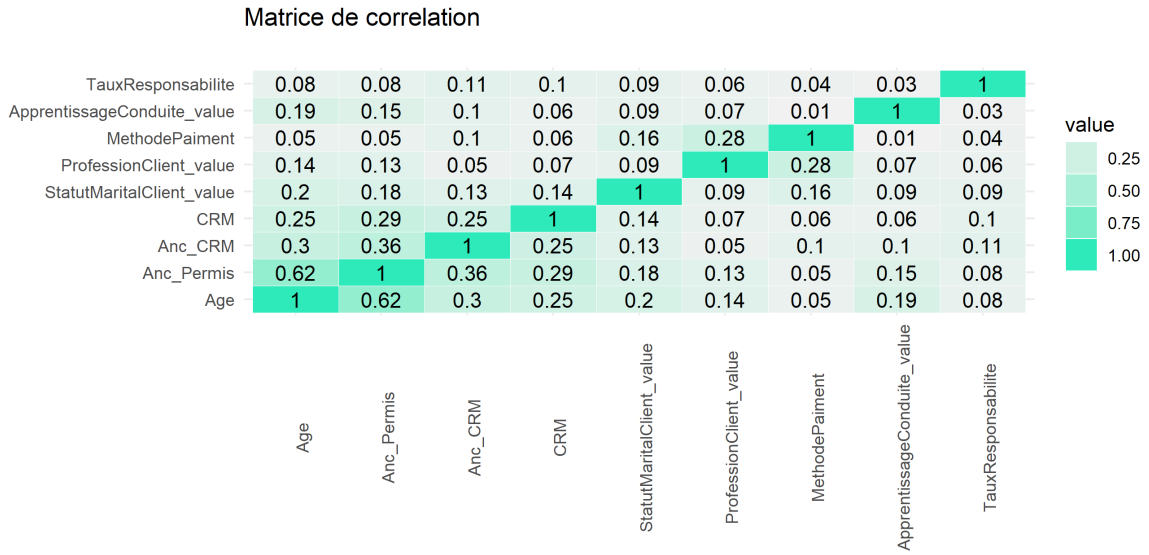


FIGURE 4.17 – Matrice de corrélation des variables clients

Corrélation entre les variables véhiculières

Certaines variables ont été exclues de l'étude, car elles sont soit redondantes, soit très corrélées avec d'autres variables. Les variables redondantes sont principalement des variables dont l'effet est déjà capté par d'autres variables, et dont la valeur ajoutée serait infime. C'est le cas par exemple de la variable représentant la version du véhicule, une variable caractère très granulaire étant une répétition du modèle. Les variables corrélées sont les variables dont le coefficient de corrélation est proche de 1. On retrouve par exemple la génération qui est très corrélée à la marque du véhicule.

Dans le tableau ci-dessous est consigné le récapitulatif des variables corrélées et leurs coefficients de corrélation avec en rouge les variables qui ont été exclues de l'étude.

4. Il s'agit du fractile d'ordre $1-\alpha$ de la loi du χ^2 à $(k-1)(r-1)$ degrés de liberté

Variables corrélées		V de Cramer
Modèle	Version	0,92
Modèle	Génération	0,51
Marque	Transmission	0,58
Note de sécurité	Charge utile	0,58
Méthode de paiement	Financement	0,53
Puissance Hp	Groupe de puissance	0,57
Puissance Hp	Puissance CEE	0,89
Carrosserie	Type de véhicule	0,97
Carburant	Alimentation	0,83
Groupe SRA	Groupe SRA Origine	0,86
Classe Réparation	Classe Réparation Origine	0,69

TABLE 4.1 – Tableau de corrélation entre les variables véhiculières

Chapitre 5

Construction des modèles de tarification GLM

Ce chapitre est consacré à la présentation de la théorie des modèles linéaires généralisés ainsi qu'à leur mise en œuvre dans le cadre de cette étude.

5.1 Théorie des modèles linéaires généralisés

5.1.1 Principe et définition

En assurance IARD, la construction d'un tarif se base sur une modélisation *fréquence*sévérité* grâce à des méthodes statistiques de régression. Le but est de chercher à établir une relation entre une variable d'intérêt Y et plusieurs variables explicatives $X = {}^t(X_1, \ldots, X_p)$ $(p \in \mathbb{N}^*)$.

Principe

Le modèle linéaire généralisé (GLM) est la méthode de régression la plus utilisée en assurance Non-Vie. Il s'agit d'une extension de la régression linéaire ordinaire. En régression linéaire classique, une variation constante d'un prédicteur entraîne une variation constante de la variable expliquée, peu importe le point de l'ensemble de définition considéré. La relation de dépendance entre la sortie et les régresseurs est donc exclusivement linéaire. Les GLM quant à eux permettent de modéliser une dépendance non linéaire dans le sens où c'est l'image de la variable réponse par une fonction arbitraire g (appelée fonction de lien) qui décrit la relation entre la combinaison linéaire des variables explicatives X et l'espérance de la variable réponse Y.

Définition

Un modèle linéaire généralisé est caractérisé par trois hypothèses :

- **Une loi de probabilité :** on suppose que les observations Y_i sont indépendantes et associées à une loi de probabilité $\mathcal{F}_{\exp}(\theta_i, \phi_i, a, b, c)$ issue d'une structure exponentielle. Avec θ_i le paramètre d'échelle, ϕ le paramètre de dispersion et a, b, c trois fonctions.
- **Une fonction déterministe :** c'est le prédicteur linéaire donné par le vecteur de variables explicatives X_i. Soit $x_i, ..., x_n$ les observations de la variable explicative X_i, nous avons :

$$\eta_i = {}^t X_i \beta$$

où

$$^t X_i = \begin{pmatrix} 1 & x_{i1} & \dots & x_{ip} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

- **Une fonction de lien :** La relation entre la composante aléatoire et le prédicteur linéaire est exprimée par une troisième composante appelée fonction de lien g qui est monotone, différentiable et inversible.
  La fonction g est telle que $E(Y_i) = g^{-1}(\eta_i)$, pour $i \in \{1, \dots, n\}$.
  Notons que les paramètres θ_i sont liés au prédicteur linéaire par la relation

$$\mu_i = E(Y_i) = b'(\theta_i) = g^{-1}(\eta_i).$$

  Si $\theta_i = \eta_i$, alors la fonction de lien est dite canonique. C'est-à-dire $\theta = g(b'(\theta)) \Leftrightarrow g(x) = (b')^{-1}(x)$.

Ci-dessous un tableau récapitulatif des fonctions de lien classiques :

Identité	$g(x) = x$
Log	$g(x) = \log(x)$
Logit	$g(x) = \log\left(\frac{x}{1-x}\right)$
Inverse	$g(x) = \frac{1}{x}$
Probit	$g(x) = \varphi(x)$

5.1.2 Les lois de probabilité

Il est considéré que la variable à expliquer Y appartient à une famille de loi exponentielle. La famille exponentielle contient toutes les lois disposant d'une fonction de densité pouvant s'écrire sous la forme suivante :

$$f(x_i; \theta) = \exp\left(\sum_{j=1}^{k} \eta_j(\theta) T_j(x_i) - B(\theta)\right) h(x_i), \quad x_i \in X$$

où $\eta(.), T(.) : \mathbb{R}^k \mapsto \mathbb{R}^k, h(.) : \mathbb{R}^d \mapsto \mathbb{R}^k$ la fonction de base, $B : \mathbb{R}^k \mapsto \mathbb{R}$ et $\theta \in \Theta \subset \mathbb{R}^k$ est le vecteur de paramètres naturel.

Le tableau 5.6 liste quelques lois usuelles de la famille exponentielle et leurs usages les plus classiques.

Loi	Fonction de lien	Moyenne	Variance $V(\mu)$	Utilisation classique
Normal $\mathcal{N}(\mu, \sigma^2)$	identité $\eta = \mu$	$\mu = x^T\beta$	1	régression standard
Bernoulli $\mathcal{B}(\mu)$	logit $\eta = \log\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{1}{1+e^{-x}T_\beta}$	$\mu(1-\mu)$	modèle de taux
Poisson $\mathcal{P}(\mu)$	$\log \eta = \log(\mu)$	$\mu = e^{x^T\beta}$	μ	fréquence de sinistre
Gamma $\mathcal{G}_2(\alpha, \beta)$	inverse $\eta = \frac{1}{\mu}$	$\mu = \left(x^T\beta\right)^{-1}$	μ^2	sévérité des sinistres

TABLE 5.1 – Quelques lois de la famille exponentielle

5.1.3 Estimation des paramètres

Une fois le modèle défini, il convient d'estimer le vecteur de coefficients associé à chaque variable explicative du modèle. L'estimation du vecteur de paramètres β et du paramètre ϕ se fait par la méthode de maximum de vraisemblance. La log-vraisemblance est :

$$\ln L(\beta, \phi) = \sum_{i=1}^{n} \frac{\theta_i(\beta) y_i - b\left(\theta_i(\beta)\right)}{a\left(\phi_i\right)} + \sum_{i=1}^{n} c\left(y_i, \phi_i\right)$$

avec $\theta_i = \left(b'\right)^{-1}\left(g^{-1}\left(\eta_i\right)\right)$. Les équations du score dans le cas général sont :

$$\forall j = 1, \ldots, d, \sum_{i=1}^{n} \frac{y_i - \mu_i}{a\left(\phi_i\right) V\left(\mu_i\right)} \frac{x_{ij}}{g'\left(\mu_i\right)} = 0$$

.

5.1.4 Mesures d'adéquation et choix du modèle

Plusieurs modélisations de la variable réponse sont envisageables. À cet effet, il existe quelques indicateurs de qualité qui permettent de vérifier l'adéquation de chaque modèle aux données et de les comparer entre eux.

La déviance

Il est difficile de se faire une idée sur l'ajustement en se basant sur la valeur de la vraisemblance puisqu'elle dépend de la taille de l'échantillon. Pour la régression logistique, un outil spécifique est introduit : la déviance.

La déviance est un indicateur statistique issu de la comparaison du modèle étudié avec un modèle saturé. Le modèle saturé est le modèle le plus complexe avec une distribution identique, une même fonction de lien et possédant autant de paramètres que d'observations fournissant ainsi une description parfaite des données.

La déviance est calculée en faisant la différence entre la log-vraisemblance du modèle saturé et la log-vraisemblance du modèle étudié multipliée par le paramètre de dispersion ϕ :

$$D = 2\phi\{\log L_{sat} - \log L\}$$

Elle peut également être définie comme la déviance standardisée multipliée par le paramètre de dispersion :

$$D = \phi D^* \text{ où } D^* = 2\{\log L_{sat} - \log L\}$$

Plus la déviance est faible, plus le modèle est meilleur, car ce critère indique un écart plus faible entre les log-vraisemblances, et ainsi une distance plus faible entre les valeurs modélisées et les valeurs observées.

Les critères AIC et BIC

La déviance ne permettant de comparer que des modèles emboîtés, son utilisation se révèle alors insuffisante. Il est important d'avoir recours à des critères de choix de modèles qui permettent de comparer des modèles qui ne sont pas forcément emboîtés.

Les critères AIC et BIC sont les plus utilisés. Ces critères sont basés sur le principe suivant : plus la vraisemblance est grande, plus grande est donc la log-vraisemblance et meilleur est le modèle (en terme d'ajustement). Cependant, la vraisemblance augmente avec la complexité du modèle et choisir le modèle qui maximise la vraisemblance revient à choisir le modèle saturé. Ce modèle est clairement sur-paramétré, il "sur-ajuste" les données (overfitting).

L'Akaïke Information Criterion (AIC) permet d'effectuer un compromis entre la réduction du biais (avec l'augmentation du nombre de paramètres) et le besoin de modéliser les données avec le plus petit nombre de paramètres.

Par définition, l'AIC pour un modèle $\mathcal{M}$ de dimension p est défini par :

$$AIC(\mathcal{M}) = -2\log(\hat{\mathcal{M}}) + 2p$$

Le Bayesian Information Criterion (BIC) est l'indicateur dérivé de l'AIC dont l'utilisation est la plus populaire. Il permet de pénaliser le modèle sur la taille de l'échantillon, et non pas seulement sur le nombre de paramètres comme l'AIC.

Le critère BIC pour un modèle $\mathcal{M}$ de dimension p est défini par :

$$BIC(\mathcal{M}) = -2\log(\mathcal{M}) + k\log(n)$$

Le meilleur modèle est celui qui possède le plus petit AIC ou BIC.

5.1.5 Sélection de variables

La sélection de variables est une étape cruciale lorsque le nombre de variables explicatives est trop important. L'objectif étant de choisir le plus petit nombre de variables expliquant au mieux le risque. Il est important de ne retenir que les variables qui ont une réelle influence sur la variable réponse. À cet effet, on distingue plusieurs algorithmes de sélection automatique des variables.

Les méthodes de sélection de variables couramment utilisées sont les méthodes pas à pas. Ces méthodes utilisent des algorithmes répétant la régression en ajoutant ou enlevant à chaque étape certaines variables. Le modèle choisi est celui comportant la sélection de variables qui minimise les critères AIC/BIC.

La méthode backward

La méthode backward encore appelée la méthode descendante est une méthode d'élimination de variables qui part du modèle complet. Le principe consiste à partir du modèle complet comportant toutes les variables explicatives pré-sélectionnées et de chercher, à chaque étape de l'algorithme, la variable la plus pertinente à retirer selon le critère choisi.

Pour savoir si une variable est significative ou non, le test suivant est réalisé :

$$H_0 : \text{Le coefficient } \beta_j \text{ associé à la variable } x_j \text{ est nul.}$$

Lorsque l'hypothèse H_0 est rejetée avec un niveau de probabilité α, la variable est significative i.e elle exerce une influence (positive ou négative) sur la variable réponse.

La méthode forward

La méthode forward est une méthode de sélection par ajouts successifs. Ici le principe consiste à partir d'un modèle vide ne comprenant que l'intercept et d'y ajouter une à une les variables les plus significatives.

L'inconvénient de cette méthode est qu'une variable qui était significative lors d'une étape de la méthode peut ne plus l'être suite à l'ajout de nouvelles variables.

La méthode mixte

Cette méthode peut être vue comme une combinaison des deux précédentes : à chaque étape, l'algorithme retire ou rajoute les variables au modèle en cherchant à minimiser le critère considéré.

Méthode Lasso

Aux méthodes de sélections automatiques, nous pouvons ajouter la régression pénalisée Lasso, une méthode de régularisation. Les méthodes de régularisation sont basées sur le principe d'ajout d'une pénalité envers la complexité du modèle dans le but de favoriser la parcimonie. Cette pénalisation est effectuée à travers l'utilisation d'une norme sur les paramètres à estimer. On distingue deux méthodes de régularisation : la régression Ridge utilisant la norme L_1 [1] et la régression Lasso utilisant la norme L_2 [2].

La régression pénalisée de Lasso consiste à ajouter une pénalisation sur les coefficients à l'aide d'un coefficient λ. L'objectif ici n'est plus de minimiser uniquement la somme des erreurs au carré, mais également la valeur des coefficients. L'algorithme utilise comme fonction de pénalisation la norme L_1 du vecteur de paramètres β qui peut combiner les capacités d'une élimination des variables classiques en supprimant automatiquement les petits coefficients et les capacités d'une norme L_2 pour stabiliser les estimations. L'estimateur Lasso est la solution du problème d'optimisation suivant :

$$\min_{\beta} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_j |\beta_j| \right\} \quad \lambda \geq 0$$

Le coefficient λ est le paramètre de régularisation qui contrôle l'importance de la pénalité. Pour $\lambda = 0$ aucune pénalité n'est appliquée, on retrouve l'estimateur des moindres carrés tandis que pour

1. Norme L1 : $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$
2. Norme L2 : $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$

$\lambda \to \infty$, toutes les variables sont associées à un coefficient estimé nul. Il est donc important de trouver la valeur optimale de λ qui permet une élimination judicieuse des variables. Pour cela, le processus de validation croisée est mis en place afin d'optimiser le choix de λ. Le principe consiste à sectionner la base en k échantillons, dont l'un sera considéré comme la base de test, les k-1 sous-ensembles restant seront utilisés comme base d'apprentissage. L'opération est itérée en considérant chacun des k échantillons comme base de validation. Dans notre étude, nous avons choisi $k = 10$.

Dans ce mémoire, nous comparerons la méthode de sélection backward à la méthode Lasso.

5.1.6 Indicateurs de performance des modèles

L'erreur quadratique moyenne RMSE

Le RMSE ou encore Root-Mean-Square-Error est un indicateur permettant de mesurer les différences entre les valeurs prédites par un modèle et les valeurs observées. Comme son nom l'indique, elle représente la racine carrée de la moyenne des erreurs au carré :

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

Avec y_i, l'observation Y de l'individu i, $\hat{y}_i$, la prédiction de y_i et n, le nombre d'observations. Entre deux modèles, le meilleur est celui qui a le RMSE le plus faible.

L'indice de Gini

Le coefficient de Gini permet de comparer des modèles et de tester l'apport de nouvelles variables sur la capacité à segmenter les risques. Il est compris entre 0 et 1 et est calculé à partir de la courbe de Lorenz [6]. Si on note $\mathcal{L} : x \longrightarrow \mathcal{L}(x)$ la courbe de Lorenz, alors l'indice de Gini qui lui est associé se définit de la manière suivante :

$$G = 1 - 2\int_0^1 \mathcal{L}(x)dx$$

5.2 Mise en place des modèles GLM hors variables véhiculières

En assurance, la méthode de tarification classiquement utilisée est l'approche *fréquence*coût*, car elle permet une estimation cohérente des risques considérés. Elle est basée sur l'hypothèse d'indépendance entre la fréquence et le coût. La prime pure est ensuite déduite en multipliant la fréquence moyenne de sinistres et le coût moyen.

La construction des modèles de fréquence et de coût s'appuie sur la théorie des modèles linéaires généralisés. Les étapes sont les suivantes :

- Choix des lois paramétriques,
- Sélection des variables tarifaires,
- Analyse des résultats et validation des modèles.

5.2.1 Adéquation de loi

Le modèle de fréquence

Dans le modèle de fréquence, la variable à expliquer est la suivante :

$$Y = \frac{\text{Nombre de sinistres}}{\text{Exposition}}$$

L'exposition est la durée d'exposition au risque exprimée en année et qui sera mise en *offset*. En effet, l'exposition est différente pour chaque observation et impacte fortement la mesure de la fréquence des sinistres. N'étant pas une variable explicative du modèle, elle est intégrée dans la spécification opérationnelle du modèle GLM (*l'offset*).

Parmi les différentes lois de probabilité, celles usuellement adoptées pour la modélisation de la fréquence sont la loi de Poisson et la loi Binomiale-négative. Il est important de choisir judicieusement la loi la plus adaptée pour la modélisation. Dans cette optique, on part d'une analyse graphique qui permet d'identifier la loi qui s'ajuste le mieux aux données. Nous traçons l'histogramme des fréquences correspondantes aux données, puis nous superposons les valeurs de la densité associée à la loi théorique en estimant éventuellement les paramètres inconnus de celle-ci.

Sur les figures 5.1 et 5.2, on peut observer que les densités des deux lois sont assez proches des données. Cependant, la loi de Poisson est celle qui s'ajuste parfaitement aux données. De plus, le test de Kolmogorov-Smirnov [3], a permis d'invalider l'hypothèse selon laquelle les données de la fréquence suivent une loi Binomiale-négative. La loi retenue pour la modélisation de la fréquence de sinistralité est donc la loi de Poisson.

3. Le test de Kolmogorov-Smirnov est présenté en annexe A.

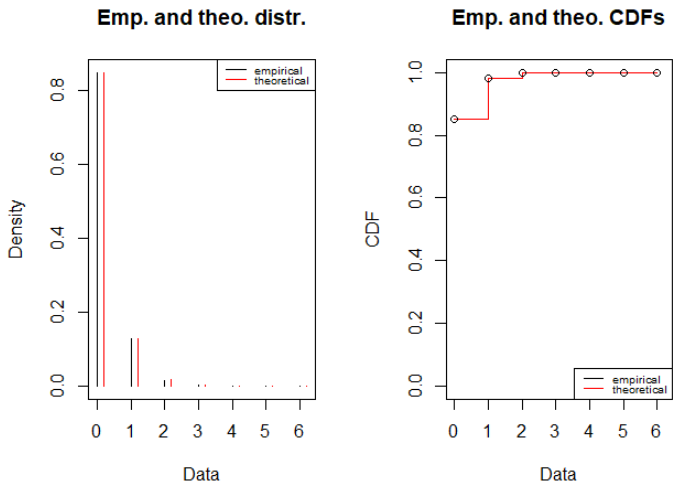


FIGURE 5.1 – Adéquation de la loi de poisson

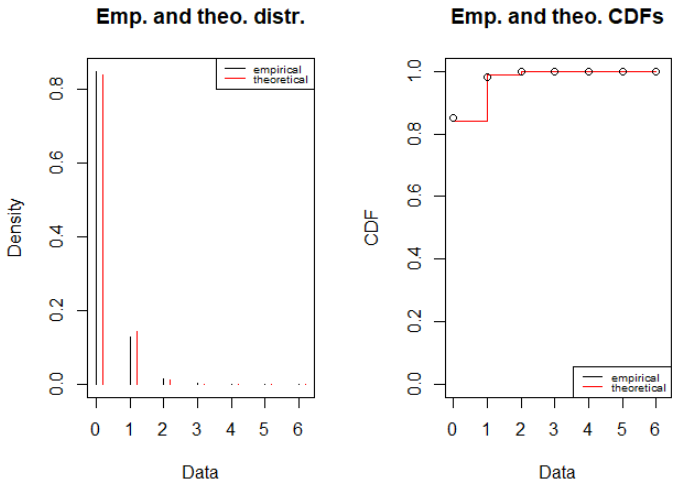


FIGURE 5.2 – Adéquation de la loi Binomiale-négative

Le modèle de coût

La modélisation du coût moyen est similaire à celle de la fréquence. La variable explicative est définie par :

$$Y = \frac{\text{Charge des sinistres}}{\text{Nombre de sinistres}}$$

La loi Gamma et la loi Log-normale sont celles souvent utilisées pour la modélisation du coût. Cependant, la densité des montants de sinistres (figure 4.2) hors forfaits présente la forme d'une distribution issue d'une loi gamma. Cette loi sera donc retenue pour la modélisation du coût moyen.

Choix de la fonction de lien

Les fonctions de lien jouent un rôle important dans la calibration des modèles GLM, car elles influencent les résultats des modèles. Dans le cadre de la tarification automobile, le logarithme est la loi souvent utilisée. En effet, elle permet :

- d'obtenir des coefficients positifs des paramètres estimés et
- d'avoir un modèle multiplicatif qui permet de connaître facilement l'effet de chaque paramètre sur la variable réponse.

5.2.2 Pré-sélection des variables tarifaires

Comme nous l'avons déjà évoqué, la sélection des variables est une étape cruciale dans la modélisation, car elle permet de simplifier le modèle en ne gardant qu'un sous-ensemble de variables explicatives (les plus significatives). En outre, elle permet aux actuaires d'identifier les facteurs de risque déterminants de la sinistralité.

Dans ce mémoire, deux des méthodes de sélection de variables présentées en section 5.1.5 ont été comparées : la méthode backward et la régression pénalisée de Lasso.

Avant d'appliquer ces méthodes de sélection automatiques, une pré-sélection des variables candidates au GLM a été effectuée grâce aux analyses descriptives (section 4.5) et à notre expertise du

métier. Rappelons que le modèle de tarification GLM sera construit uniquement à partir des variables non véhiculières, c'est-à-dire les variables liées au client et à la police.

Variables a priori tarifaires

- Âge de l'assuré
- CSP (Catégorie socioprofessionnelle)
- CRM (Coefficient de Réduction-Majoration)
- Ancienneté du CRM
- Statut Marital
- Mode d'apprentissage de la conduite
- Formule de la police
- Usage du véhicule
- Mode de stationnement

Il est rappelé que la variable liée au sexe de l'assuré a été omise de cette étude pour des raisons de législation.

Pour plus de compréhension, nous détaillons dans le tableau 4.2 les modalités des variables qualitatives.

Variables	Modalités
Statut Marital	• Célibataire • Marié (e) • Divorcé (e) • Pacsé (e) • En concubinage • Veuf/Veuve
Mode d'apprentissage conduite	• Auto-école • Conduite accompagnée
Formule de la police	• Tiers • Tiers + • Tous risques • Tous risques + • Sans libellé
Usage du véhicule	• Privés et professionnels (occasionnels) ou associatifs • Privés et trajets travail
Mode de stationnement	• Dans un garage individuel fermé • Dans un jardin clos privé • Dans un parking couvert collectif • Dans un parking découvert collectif • Sur la voie publique

TABLE 5.2 – Modalités des variables catégorielles

5.2.3 Sélection automatisée des variables

Dans cette section, nous mettons en œuvre la sélection de variable (backward et Lasso) pour chacun des modèles de fréquence et de coût. Cette démarche permettra de définir une sélection de facteurs à conserver en vue de produire des modèles optimaux.

Dans un premier temps, nous présenterons en exemple la mise en œuvre de la méthode Lasso pour le modèle de fréquence et celle de la méthode backward pour le modèle de sévérité. Ensuite, les variables sélectionnées pour chaque modèle seront listées puis les méthodes de sélection seront comparées selon le critère du RMSE et celui de la déviance.

5.2.3.1 Sélection Lasso : exemple du modèle de fréquence

La régression pénalisée Lasso consiste en l'introduction d'une pénalisation qui réduit la variabilité de l'estimation, améliorant ainsi la précision de prédiction. Cependant, comme nous l'avons précisé

en section 5.1.5, la valeur du paramètre de pénalisation λ doit être judicieusement choisi. À cet effet, nous utiliserons le procédé de validation croisée qui nous permettra d'obtenir deux valeurs de λ que nous pourrons comparer. Il s'agit du λ_{min} et du λ_{1se}.

La figure 5.3 représentant l'évolution de l'erreur en fonction des valeurs de λ permet de retrouver les valeurs de λ_{min} et λ_{1se}. La première droite verticale en pointillé indique la valeur de λ_{min}, tandis que la seconde désigne la valeur de λ_{1se}.

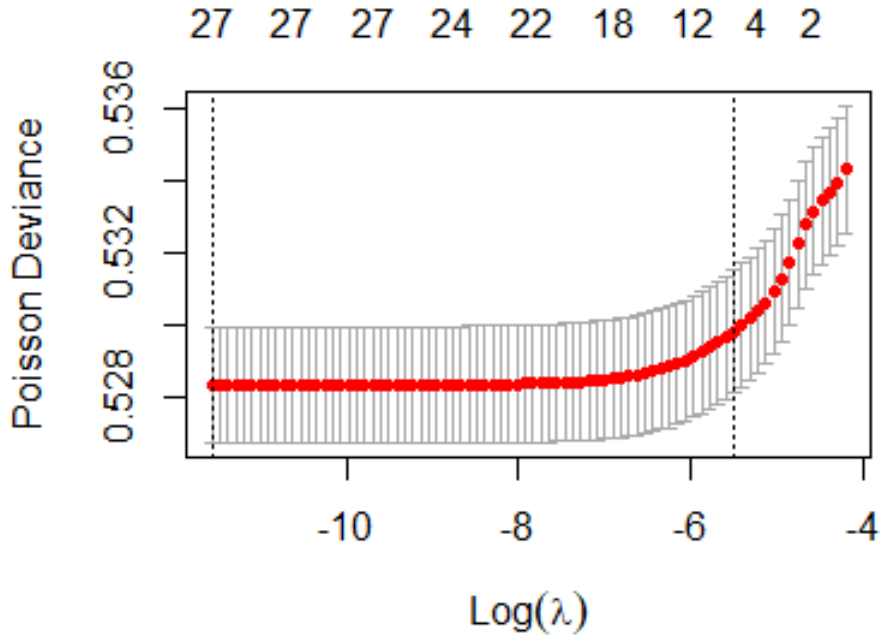


FIGURE 5.3 – Évolution de la déviance en fonction de λ

On observe que la déviance est minimale pour $log(\lambda_{min}) = -11.5547$ soit $\lambda_{min} = 9.590482e - 06$ tandis que $log(\lambda_{1se}) = -5.507546$ avec $\lambda_{1se} = 0.004056048$.

5.2.3.2 Sélection Backward : exemple du modèle de coût

La sélection backward a été réalisée à l'aide de la fonction stepAIC de la library MASS sous R. À partir du modèle fourni, l'algorithme retire à chaque étape une variable et calcule l'AIC du modèle. La variable sans laquelle la perte d'information est minimale est retenue, c'est-à-dire le modèle sans cette variable a le plus petit AIC.

Les résultats du stepAIC backward ont été consignés dans le tableau 4.3. La valeur de l'AIC est forcément décroissante au cours de la procédure, car c'est le critère de sélection de la variable la plus pertinente à chaque itération. On remarque que la perte d'information est minimale pour le modèle ne contenant pas la variable liée au mode d'apprentissage de la conduite. Ce modèle est donc celui retenu par l'algorithme.

Variable	Déviance	AIC
Mode d'apprentissage conduite	69 563	636 876
Mode de stationnement	69 642	636 876
Usage du véhicule	69 544	636 877
Statut Marital	69 552	636 877
Ancienneté du CRM	69 668	636 878
<none>	69 542	636 878
CSP	69 727	636 886
Âge	69 901	636 888
CRM	70 059	636 903
Formule police	71 194	636 969

TABLE 5.3 – Résultats du stepAIC pour la sévérité

5.2.3.3 Résultats

Nous nous intéressons maintenant aux facteurs de risques retenus par chaque procédure pour chacun des modèles de fréquence et de coût. Les résultats sont consignés dans le tableau suivant et chaque variable sélectionnée est marquée par une croix $\times$. Notons qu'une variable non significative peut tout de même être sélectionnée, car l'AIC se dégraderait sans cette variable. Les variables significatives seront donc marquées par un astérisque $*$.

Variable	Lasso λ_{min}	Lasso λ_{1se}	Backward	
Âge assuré	$\times$		$\times$	*
CSP	$\times$		$\times$	*
CRM	$\times$	$\times$	$\times$	*
Ancienneté CRM	$\times$	$\times$	$\times$	*
Statut Marital	$\times$			
Mode d'apprentissage conduite			$\times$	*
Formule police	$\times$	$\times$	$\times$	*
Usage véhicule	$\times$		$\times$	*
Mode de stationnement	$\times$	$\times$	$\times$	*

TABLE 5.4 – Sélection de variables pour la fréquence

L'observation des résultats de sélection pour la fréquence prête à croire que les méthodes sont peu sélectives, car très peu de variables ont été écartées. En effet, dès qu'au moins une modalité d'une variable est sélectionnée, la variable est retenue. Le λ_{min} et la méthode backward produisent des résultats similaires. Le λ_{1se} a été plus sélectif que les autres méthodes et écarte l'âge de l'assuré qui est tout de même une variable importante. Cette valeur de lambda réalise visiblement une pénalisation trop forte. Une pénalisation avec λ_{min} nous semble plus optimale.

Variables	Lasso λ_{min}	Lasso λ_{1se}	Backward	
Âge assuré	×		×	
CSP	×		×	*
CRM			×	*
Ancienneté CRM			×	*
Statut Marital			×	
Mode d'apprentissage conduite				
Formule police	×		×	*
Usage véhicule	×		×	
Mode de stationnement	×		×	

TABLE 5.5 – Sélection de variables pour la sévérité

La modélisation du modèle de sévérité est souvent particulière, car peu de variables sont significatives. La méthode backward retient pratiquement toutes les variables du modèle quand bien même peu d'entre elles sont significatives. λ_{min} écarte le CRM, l'ancienneté du CRM, le mode d'apprentissage de la conduite et le Statut Marital. Le λ_{1se} quant à lui ne sélectionne aucune variable.

Dans la section suivante, nous allons comparer les modèles issus d'une sélection de variable avec la méthode backward à ceux de la méthode Lasso avec une pénalisation λ_{min}. La méthode Lasso avec une pénalisation λ_{1se} est écartée, car elle effectue une sélection radicale.

5.2.3.4 Comparaison des méthodes de sélection

Nous comparons les méthodes de sélection de variables pour chaque modèle suivant le critère du RMSE et de la déviance.

Modèle	Méthode	$RMSE_{app}$	$RMSE_{test}$	$Déviance_{app}$	$Déviance_{test}$
Fréquence	StepAIC (backward)	0.3198009	0.3197025	3545451	885295.3
	Lasso λ_{min}	0.3197023	0.3197015	3545280	885245.3
Sévérité	StepAIC	0.3198005	4766.57	3545020	885106.5
	Lasso λ_{min}	0.3197919	4765.068	3542851	884551.7

TABLE 5.6 – Comparaison du stepAIC et de la régression Lasso

Remarquons que dans l'ensemble, les valeurs des indicateurs sont très proches pour chacune des méthodes. Cependant, la méthode Lasso semble être la plus performante au sens des deux critères. Pour le modèle de fréquence comme celui de sévérité, nous retiendrons donc les variables sélectionnées par la méthode Lasso.

5.3 Analyse des résultats et validation des modèles

5.3.1 Analyse de la qualité de prédiction

Une fois les facteurs de risques influençant la fréquence et le coût déterminés, les modèles GLM de fréquence et de coût ont été calibrés. il convient de mesurer le pouvoir prédictif des modèles en comparant les valeurs prédites aux valeurs observées. Il s'agit d'étudier les divergences de prédiction entre les différents modèles selon des segments spécifiques, afin d'avoir une vision plus précise des phénomènes en présence. À cet effet, les graphes de comparaison des valeurs prédites et des valeurs observées en fonction du risque étudié, ont été représentés. L'analyse de prédiction des autres variables est disponible en annexe C.

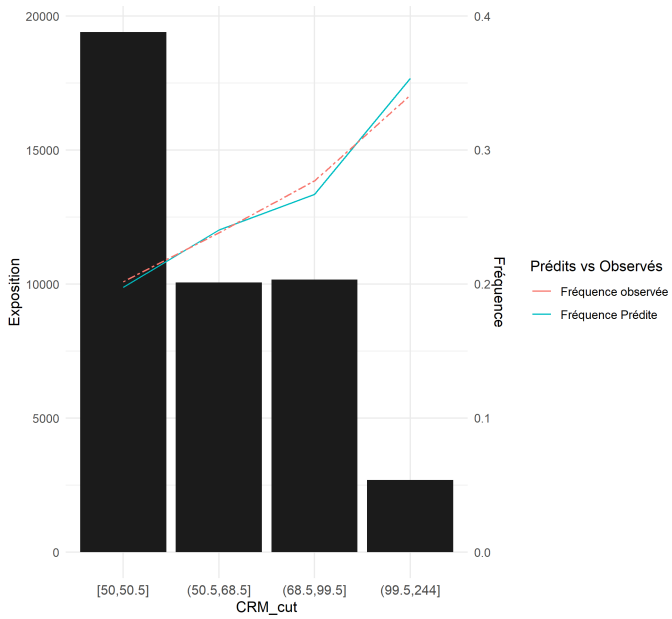


FIGURE 5.4 – Fréquence prédite en fonction du CRM

Le graphe ci-dessus compare les valeurs prédites et les valeurs observées de la fréquence moyenne

par classe de risque pour la variable CRM. Remarquons que les prédictions sont très proches des observations et suivent la même tendance. Cependant, on note une légère sous-estimation du risque au niveau de la troisième classe. Cette sous-estimation peut-être expliquée par le fait que le risque de cette classe est déjà capté par une autre variable : l'âge de l'assuré par exemple.

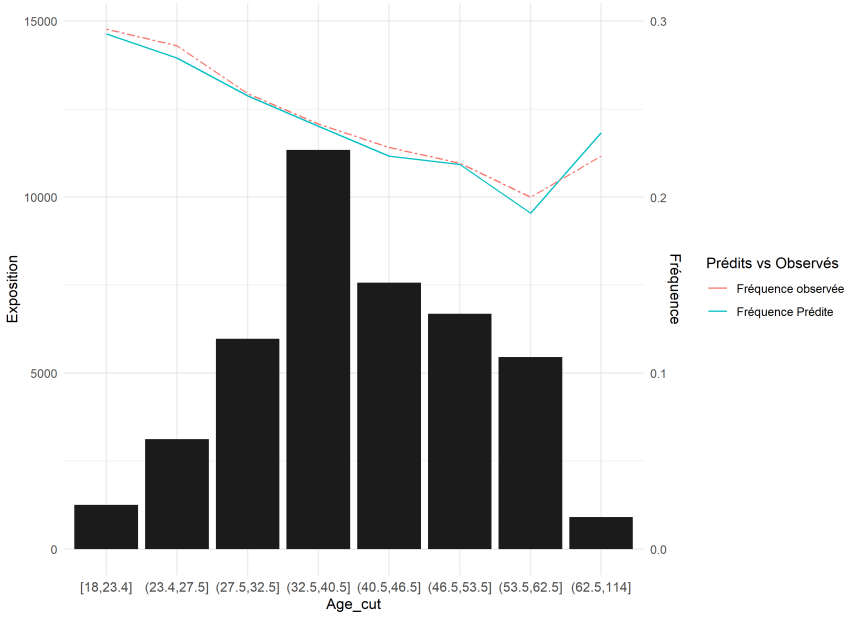


FIGURE 5.5 – Fréquence prédite en fonction de l'âge de l'assuré

Les prédictions de la fréquence en fonction de l'âge de l'assuré sont satisfaisantes. La tendance est la même pour les valeurs observées et les valeurs prédites. Le modèle capture donc bien les effets de chacune des classes.

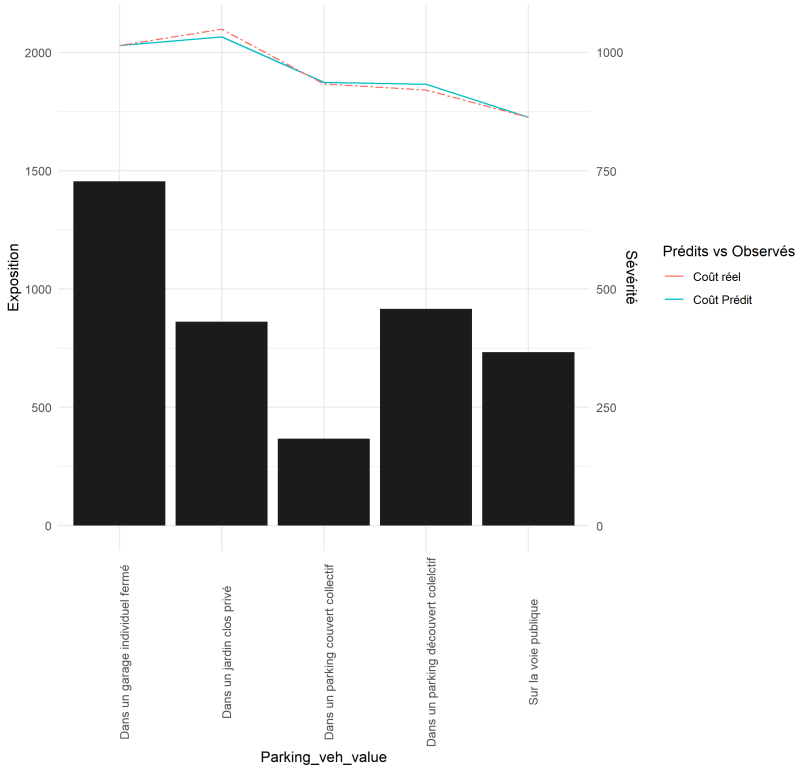


FIGURE 5.6 – Sévérité prédite en fonction du mode de stationnement

Pour la variable représentant le mode de stationnement, la prédiction de la sévérité est cohérente avec les observations de chaque classe de risque.

En somme, les valeurs observées et les valeurs prédites par les modèles présentent la même tendance. Les effets de chaque variable sont bien capturés. Des ajustements post GLM pourront être réalisé au niveau des tarifs, pour les classes où le risque a été surestimé ou sous-estimé.

5.3.2 Analyse des résidus

Une étape complémentaire dans la validation des modèles est l'analyse des résidus en fonction des prédictions de chaque observation.

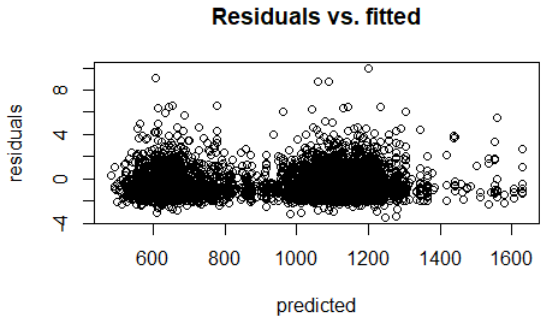


FIGURE 5.7 – Résidus du modèle de coût

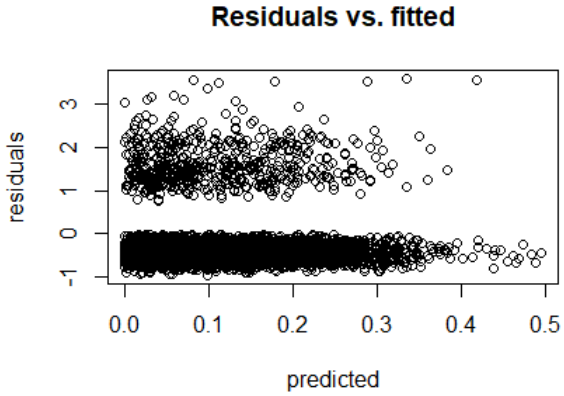


FIGURE 5.8 – Résidus du modèle de fréquence

Sur le graphe des résidus du modèle de sévérité, on peut observer que le modèle réalise de meilleures prédictions pour les montants autour de 900€, car ces résidus sont proches de zéro. Cependant, les résidus sont distribués de façon aléatoire et indépendamment des valeurs prédites. Sur le graphe des résidus du modèle de fréquence, on observe deux nuages de points distincts. Les points négatifs correspondent aux polices n'ayant pas subi de sinistres. Nous rappelons que c'est la probabilité de survenance d'un sinistre qui est modélisée sachant que 88% de la base de donnée est non sinistrée. Ceci explique dont l'allure des résidus qui est classique pour un modèle de fréquence.

Ces différentes analyses permettent donc de valider les modèles de fréquence et de coût. Dans la section suivante, les différents types de résidus seront énumérés, puis nous justifieront le choix des résidus qui seront modélisés dans la suite de l'étude.

5.4 Extraction des résidus

Les résidus d'un modèle de régression peuvent être définis comme étant les différences entre les valeurs observées et les valeurs estimées par ce modèle. Leur particularité est qu'ils représentent la partie non expliquée par l'équation de régression. On distingue plusieurs types de résidus.

Les résidus additifs : $R = y - \hat{y}$ Ces résidus sont faciles à exprimer et sont obtenus en faisant la différence entre les observations et les prédictions.

Les résidus multiplicatifs : $R = \frac{y}{\hat{y}}$ Ces résidus sont également courants et s'obtiennent en faisant le rapport des valeurs observées sur les valeurs prédites.

Les résidus de Pearson : $R = \frac{(y-\hat{y})}{\sqrt{\widehat{V(\hat{y})}}}$ Ce sont des résidus standardisés obtenus en comparant les valeurs observées y et valeurs prédites $\hat{y}$ pondérés par leur précision estimée par l'écart-type de $\hat{y}$.

Les résidus de déviance : Ces résidus mesurent la contribution de chaque observation à la déviance du modèle par rapport au modèle saturé. Ils sont complexes à déterminer et sont définis à partir d'un terme d_i représentant la contribution de l'ième observation y_i à la déviance D,

$R = \text{signe}\,(y_i - \hat{y}_i)\,\sqrt{d_i}$ avec $d_i = 2\left\{\log L\,(y_i, y_i, \phi) - \log L\,(y_i, \hat{y}_i, \phi)\right\}$ et ϕ le paramètre de dispersion.

L'inconvénient de ces résidus est qu'ils ne sont pas inversibles.

Les résidus d'Anscombe : Les résidus d'Anscombe permettent de palier aux limites résultant des résidus standards et présentent l'avantage d'être inversibles. Cette méthode proposée par Anscombe consiste à appliquer une transformation préalable afin de construire des résidus suivant une loi normale :

$$R = \frac{A\,(y_i) - A\,(\hat{y}_i)}{A'\,(y_i)\,\sqrt{V\,(\hat{y}_i)}}$$

avec $A(y) = \int_{-\infty}^{y} V(t)^{-1/3} dt$. Dans le cas de la loi de Poisson, on a : $R = \frac{3\left(y_i^{2/3} - \hat{y}_i^{2/3}\right)}{2\hat{y}_i^{1/6}}$.

La littérature sur la construction des zoniers et des véhiculiers préconise l'utilisation des résidus d'Anscombe. En effet, ces résidus sont inversibles et il est possible, à partir d'une estimation $\hat{y}$ et d'un résidu R, de produire une nouvelle estimation $\hat{\hat{y}}$ de y. Cette propriété semble utile en cas de lissage spatial (lissage par la méthode de Krigeage par exemple). Dans le cadre de cette étude, nous préconisons une structure de résidus simple et homoscédastiques. Nous utiliserons donc les résidus de Pearson.

Conclusion partielle

L'objectif de ce chapitre était de mettre en œuvre un modèle de tarification pour les modèles de fréquence et de coût en fonction des variables non véhiculières. Nous avons mis en place des modèles optimaux en nous attardant sur la sélection des variables pouvant influencer le risque étudié. Nous avons dans un premier temps procédé à une sélection judicieuse des variables que nous supposons tarifaires grâce aux analyses descriptives et univariées des variables. Nous avons ensuite procédé à la comparaison de deux méthodes de sélections automatiques de variables. La régression Lasso s'est avérée la plus pertinente au sens du RMSE et de la déviance. Enfin, nous avons pu vérifier que les prédictions des modèles sont bien en accord avec les observations.

Chapitre 6

Construction du véhiculier à partir des méthodes de Machine Learning

Ce chapitre est consacré à la mise en œuvre de l'approche décrite dans la section 2.2. Nous rappelons que la méthode de construction du véhiculier développée dans ce mémoire est basée sur une modélisation des résidus issus d'un modèle de tarification GLM. La première étape réalisée dans le chapitre précédent consistait à mettre en place les modèles GLM fréquence et coût hors variables véhiculières. Les étapes suivantes développées dans ce chapitre consistent à extraire la part résiduelle non captée par ces modèles puis à les modéliser en fonction des variables véhiculières à l'aide de méthodes de Machines Learning.

À cet effet, les méthodes de modélisations mises en place seront décrites. Ensuite, ces deux méthodes seront comparées afin de sélectionner celle qui réalise le modèle le plus optimal au sens des critères de validation de modèle. Enfin, les prédictions du meilleur modèle seront classifiées puis intégrées dans le modèle de tarification initial.

6.1 Les méthodes de Machine Learning

Les trois grandes familles de méthodes de Machine Learning les plus populaires sont :

- Les arbres de décision,
- Les réseaux de neurones artificiels,
- Les machines à vecteurs de support (SVM).

Dans le cadre de cette étude, nous privilégions des modèles simples et parcimonieux afin d'obtenir des résultats facilement interprétables. Nous nous concentrerons donc sur les méthodes d'arbres de décision, qui sont plus faciles à interpréter et plus adaptées dans le cadre de cette étude. Les réseaux neurones et les machines à vecteurs de support restent des pistes explorables pour des études ultérieures.

6.1.1 Les arbres de régression

Les arbres de régression sont des méthodes d'apprentissage statistiques basées sur l'algorithme CART (Classification And Regression Trees) et introduits par Breiman et al. (1984)[3]. L'objectif de cette méthode est de construire un arbre de décision en classifiant un ensemble d'enregistrements. Cet arbre fournit un modèle pour classer de nouveaux échantillons.

Le principe général de CART est de partitionner récursivement l'espace d'entrée X de façon binaire, puis de déterminer une sous-partition optimale pour la prédiction. En d'autres termes, il s'agit de

créer des partitions par divisions successives dans le but final de séparer au mieux les données et de former des groupes de faible variance.

On parle d'arbre de régression lorsque la variable à expliquer est numérique et d'arbre de classification lorsqu'elle est qualitative. La construction d'un arbre peut se résumer en deux phases. Une première phase qui consiste en la construction d'un arbre maximal, qui permet de définir la famille de modèles à l'intérieur de laquelle on cherchera à sélectionner le meilleur modèle et une seconde phase, dite d'élagage, qui construit une suite de sous-arbres optimaux élagués de l'arbre maximal.

Un arbre est caractérisé par plusieurs éléments :

- La racine : qui contient l'ensemble de l'échantillon à segmenter (point de départ), la procédure est ensuite itérée sur chacun des sous-ensembles créés.
- Un tronc et des branches : qui contiennent les règles de division qui permettent de segmenter la population.
- Des feuilles : qui contiennent les sous-populations homogènes créées et fournissent l'estimation de la quantité d'intérêt.

Principe

Au départ, l'ensemble des données est considéré et constitue le nœud racine. Une division de cet ensemble est ensuite effectuée. Pour effectuer cette division représentée par un nœud dans l'arbre, il faut choisir une variable et un seuil de séparation (lorsque la variable choisie est quantitative) ; dans le cadre d'une variable qualitative, on choisit une division en deux groupes de modalités. Ces paramètres de division sont choisis et optimisés de manière à maximiser l'homogénéité des deux nœuds fils issus de la division et donc minimiser leur impureté. La figure ci-dessous permet de visualiser le fonctionnement d'un arbre de décision à 2 variables explicatives quantitatives : X_1 et X_2.

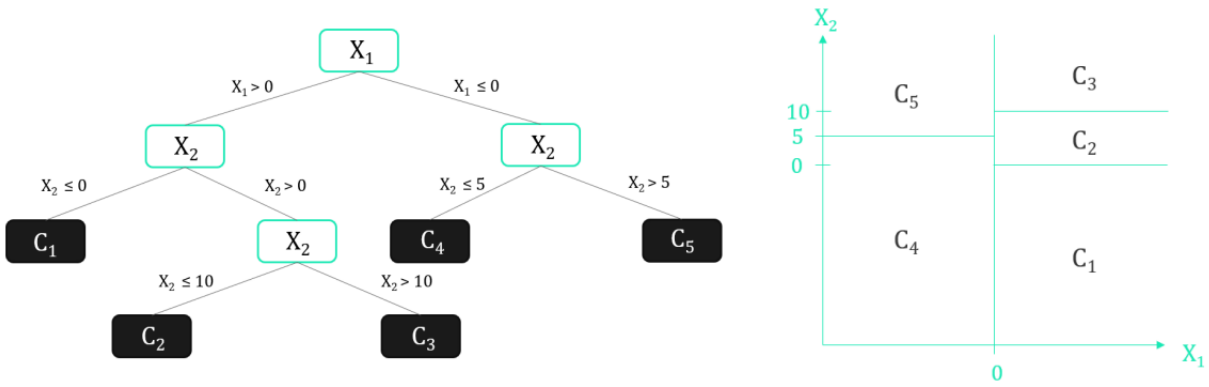


FIGURE 6.1 – Exemple d'arbre de décision

Élagage de l'arbre

L'élagage de l'arbre permet de supprimer les feuilles qui n'apporteraient rien à l'analyse. Le principe est de développer l'arbre au maximum, c'est-à-dire de construire l'arbre saturé puis de le remonter en partant des feuilles et de supprimer les nœuds dont la division n'améliore pas significativement l'arbre sur une base de validation.

L'arbre saturé est construit sur la base d'apprentissage puis élagué à l'aide d'une base de validation. Cette base de validation, indépendante, permet d'éviter le sur-apprentissage en autorisant un certain recul par rapport aux données.

6.1.2 Le bagging

Le bagging est une contraction du "bootstrap aggregation", regroupant un ensemble de méthodes introduites en 1996 par Léo Breiman[1]. C'est une technique utilisée pour améliorer la classification, notamment celle des arbres de décision. C'est un processus de tirage aléatoire avec remise sur les observations déterminé par 3 étapes clés :

- Construction de n arbres de décisions en tirant aléatoirement n échantillons d'observations,
- Entraînement de chaque arbre de décision sur ces échantillons,
- Les estimateurs ainsi obtenus sont moyennés (en cas de régression) ou utilisés pour un « vote » à la majorité (en cas de classification). La combinaison de ces multiples estimateurs indépendants permet de réduire la variance.

Mathématiquement, dans le cadre de la régression, on désigne par $(\mathbf{X}, Y)$ un vecteur aléatoire où X prend ses valeurs dans $\mathbb{R}^p$ et Y dans $\mathbb{R}$. On note $\mathcal{D}_n = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ un n-échantillon i.i.d. et de même loi que $(\mathbf{X}, Y)$ et $m(\mathbf{x}) = \mathbf{E}[Y \mid \mathbf{X} = \mathbf{x}]$ la fonction de régression.

La méthode du bagging consiste à agréger un nombre B d'estimateurs $\hat{m}_1, \ldots, \hat{m}_B$:

$$\hat{m}(\mathbf{x}) = \frac{1}{B} \sum_{k=1}^{B} \hat{m}_k(\mathbf{x})$$

.

Sous l'hypothèse que les régresseurs $\hat{m}_1, \ldots, \hat{m}_B$ sont i.i.d., avec :

$$\mathbf{E}[\hat{m}(\mathbf{x})] = \mathbf{E}\left[\hat{m}_1(\mathbf{x})\right] \quad \text{et} \quad \mathbf{V}(\hat{m}(\mathbf{x})) = \frac{1}{B}\mathbf{V}\left(\hat{m}_1(\mathbf{x})\right)$$

Ainsi, l'estimateur agrégé aura le même biais que les estimateurs $\hat{m}_k$ mais avec une variance plus faible. En pratique, la méthode du bagging donne d'excellents résultats sur les arbres de décision utilisés en forêts aléatoires.

6.1.3 Le Random Forest

Le Random Forest ou encore forêt aléatoire proposé par Leo Breiman et Adèle Cutler en 2001[2] est un algorithme d'apprentissage automatique supervisé construit à partir des algorithmes d'arbre de décision. En effet, les arbres de décision sont en général très instables et s'accompagnent du phénomène de sur-apprentissage. Ainsi, afin d'offrir une meilleure stabilité et fiabilité aux résultats, le Random Forest fonctionne sur le principe du bagging.

En plus du principe de bagging, le Random Forest ajoute de l'aléa au niveau des variables. En effet, avant la division de chaque nœud, au lieu de sélectionner la division optimale parmi les divisions possibles basées sur toutes les variables explicatives, on tire aléatoirement un certain nombre de variables explicatives, et on considère les divisions possibles basées sur ce sous-ensemble. L'ajout de cet aléa dans la construction des arbres permet de rendre ces derniers plus indépendants et de réduire la variance de l'estimation.

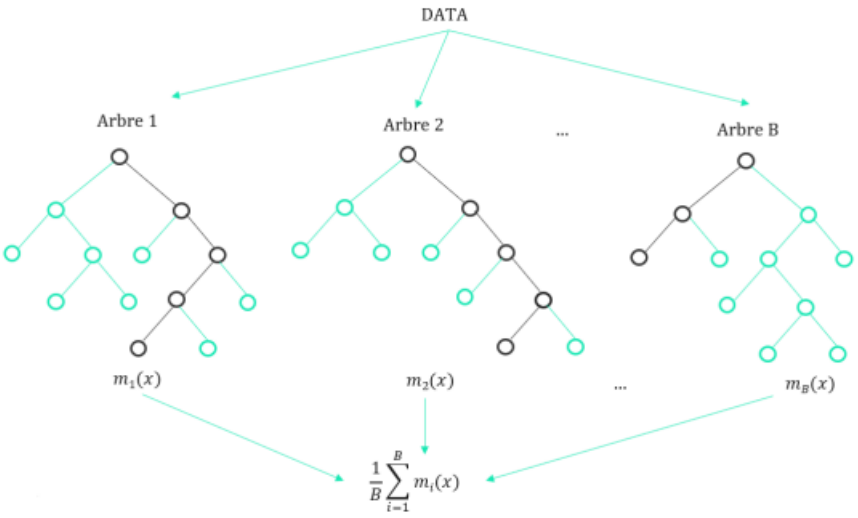


FIGURE 6.2 – Exemple de forêts aléatoires

Importance des variables

Comme en régression linéaire, il est important de déterminer les variables explicatives qui influencent le risque étudié. Dans le cadre du Random Forest, il est possible de classer les variables selon leur importance.

La mesure d'importance calculée pour une variable donnée est l'accroissement moyen de l'erreur d'un arbre dans la forêt lorsque les valeurs observées de cette variable sont permutées au hasard dans les échantillons Out Of Bag (OOB).

Soit OOB_k l'échantillon Out Of Bag associé au $k^{\text{ième}}$ arbre de la forêt. Il s'agit de l'échantillon formé par les observations qui ne figurent pas dans le $k^{\text{ième}}$ échantillon bootstrap.

L'erreur E_{OOB_k} de l'arbre k est donnée par :

$$E_{OOB_k} = \frac{1}{|OOB_k|} \sum_{i \in OOB_k} \left(h\left(X_i, \theta_k\right) - Y_i\right)^2$$

L'importance de la $j^{\text{ième}}$ variable du modèle se mesure en effectuant une permutation aléatoire des valeurs de la $j^{\text{ième}}$ variable dans l'échantillon OOB_k. On obtient alors un échantillon dit perturbé noté OOB_k^j et on calcule son erreur de prédiction $E_{OOB_k^j}$.

On réalise cette opération sur l'ensemble des échantillons bootstrap formés au préalable. Enfin, on obtient l'importance de la $j^{\text{ième}}$ variable sur la forêt en moyennant la différence d'erreurs $E_{OOB_k^j} - E_{OOB_k}$ sur tous les arbres :

$$\text{Imp}\left(X_j\right) = \frac{1}{B} \sum_{k=1}^{B} \left(E_{OOB_k}^j - E_{OOB_k}\right)$$

6.1.4 Le Gradient Boosting

Le Boosting est une technique d'agrégation développée par Freund et Schapire (1996)[9], et basé sur le même principe que le bagging : construction d'une famille de modèles qui sont ensuite agrégés par une moyenne pondérée des estimations ou un vote. Cependant, il diffère nettement sur la fa-

çon de construire la famille qui est dans ce cas récurrent. En effet, 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. En d'autres termes, il s'agit d'assembler plusieurs algorithmes ayant une performance peu élevée "weak learners" pour en créer un beaucoup plus efficace et satisfaisant "strong learner".

Le Gradient Boosting est un algorithme particulier de Boosting développé par Jerome H. Friedman en 1999 [8]. Il est basé sur la descente du gradient qui est une technique itérative qui permet d'approcher la solution d'un problème d'optimisation. En apprentissage supervisé, la construction du modèle revient souvent à déterminer les paramètres (du modèle) qui permettent d'optimiser (max ou min) une fonction.

Algorithme du Gradient Boosting

Entrée :

- un échantillon $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$
- un ensemble de règles de régression faibles
- le nombre M d'itérations, le coefficient λ

Initialisation : $\hat{g}_0(\mathbf{x}) = \arg\min_g \sum_{i=1}^{n} L(y, g)$
Itération pour $m = 1$ à M

1. calculer l'opposé du gradient aux points d'observations

$$r_{im} = -\left[\frac{\partial L(y, g)}{\partial g}\right]_{y=y_i, g=\hat{g}_{m-1}(\mathbf{x}_i)}$$

2. ajuster une règle faible de régression g_m sur l'ensemble $(\mathbf{x}_1, r_{1m}), \ldots, (\mathbf{x}_n, r_{nm})$
3. $\hat{g}_m(\mathbf{x}) = \hat{g}_{m-1}(\mathbf{x}) + \lambda g_m(\mathbf{x})$

Sortie : $\hat{g}_M(\mathbf{x})$ pour une régression, $\operatorname{sign} \hat{g}_M(\mathbf{x})$ pour une classification

Tuning des hyperparamètres

L'hyperparameter tuning ou le réglage des hyperparamètres, est une étape cruciale en Machine Learning qui permet d'optimiser l'algorithme d'apprentissage automatique.

Les algorithmes de Machine Learning dépendent de paramètres influençant l'apprentissage des données. Ces paramètres appelés hyperparamètres ne sont pas optimisés et il revient à l'utilisateur de l'algorithme de les choisir. La procédure de sélection optimale de ces hyperparamètres (tuning) nécessite un important travail statistique afin de déterminer ceux qui donneront le meilleur résultat.

Pour le Random Forest, les hyperparamètres à optimiser sont :

- le nombre d'arbres décisionnels,
- le nombre de variables à utiliser pour chaque division d'un nœud,
- la profondeur maximale des arbres,
- le nombre minimal d'observations que doit avoir un nœud pour être subdivisé.

Pour le Gradient Boosting, on peut relever 5 hyperparamètres à optimiser :

- le nombre d'arbres,
- la profondeur maximale de l'arbre (un hyperparamètre de régularisation),
- le taux d'apprentissage,
- le contrôle de complexité (gamma=γ) (un hyperparamètre de pseudo-régulation),
- le poids requis pour créer un nouveau nœud dans l'arbre.

Il existe plusieurs méthodes pour tester les paramètres d'un modèle. L'une des méthodes souvent utilisées est la recherche par quadrillage, ou Grid Search.

Le Grid Search

Le Grid search est une méthode d'optimisation qui permet de tester une série de paramètres et de comparer les performances pour en déduire le meilleur paramétrage. Pour chaque paramètre, on détermine un ensemble de valeurs que l'on souhaite tester, puis le Grid Search croise chacune de ces hypothèses et va créer un modèle pour chaque combinaison de paramètres. Ensuite, les modèles obtenus sur le dataset seront testés par validation croisée puis comparés suivant le critère du RMSE.

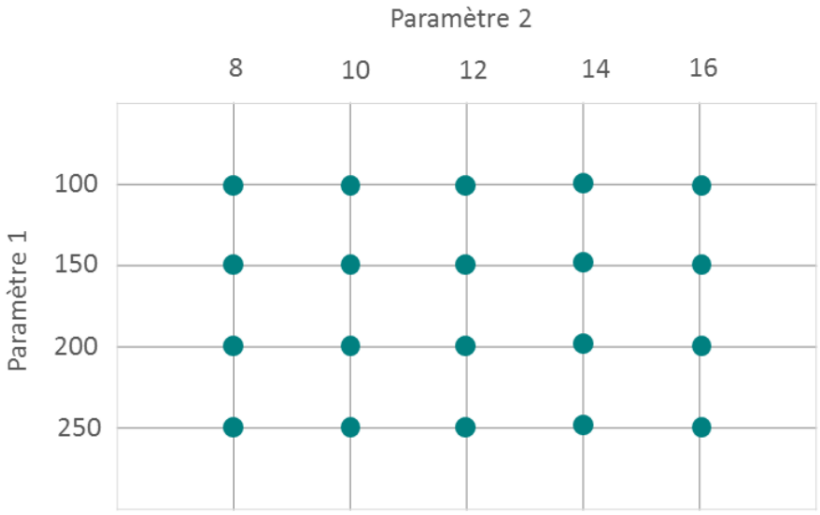


FIGURE 6.3 – Principe du Grid search

L'inconvénient du Grid search est que les paramètres à tester sont choisis à l'avance par l'utilisateur. De plus, il augmente considérablement le temps de calcul lorsque le nombre de paramètres à tester est élevé.

6.2 Modélisation des résidus par Machine learning

Cette section est dédiée à la prédiction des résidus issus des modèles de tarification. Afin de ne pas faire d'hypothèses sur la loi de la variable réponse, nous comparons les algorithmes de Random Forest et de Gradient Boosting qui sont des méthodes d'apprentissage non paramétriques contrairement au GLM. Nous n'imposerons donc aucune structure aux modèles.

Dans le chapitre précédent, le choix du type de résidus à modéliser s'est porté sur les résidus de Pearson. Cette nouvelle variable calculée devient la nouvelle cible de la modélisation. Elle sera prédite en fonction des variables véhiculières provenant de la base SRA. Rappelons que la base SRA comporte 90 variables. Suite aux analyses univariées, aux études de corrélation entre les différentes variables et l'élimination des variables représentant des dates, 24 variables véhiculières interviendront dans les modélisations.

Variables	Type
Marque	Qualitative
Carrosserie	Qualitative
Ancienneté du véhicule	Quantitative
Classe de Prix	Qualitative
Carburant	Qualitative
Poids	Quantitative
Puissance fiscale	Quantitative
Puissance Hp	Quantitative
Groupe SRA	Qualitative
Type de véhicule (Privé ou Utilitaire)	Qualitative
Vitesse maximale	Quantitative
Nombre de place	Qualitative
Cylindre M3	Quantitative
Suspension	Qualitative
Type de frein	Qualitative
Série limitée	Qualitative
Nombre de cylindres	Quantitative
Position du moteur	Qualitative
Type boîte de vitesse	Qualitative
Hauteur	Quantitative
Longueur	Quantitative
Largeur	Quantitative
Assistance au Freinage d'Urgence	Qualitative
Système Anti-Blocage des roues	Qualitative

TABLE 6.1 – Sélection de variables véhiculières

Afin d'évaluer et de comparer les algorithmes de Machine Learning, la prédiction des résidus se fera sur un nouveau partitionnement. Nous partons de la base d'apprentissage utilisée précédemment pour les modèles GLM. Nous la partitionnons de la même manière que précédemment : 80% pour l'apprentissage et 20% pour la validation.

6.2.1 Optimisation des hyperparamètres

Les modèles de Machine Learning sont calibrés pour donner les meilleurs résultats possibles. À cet effet, il est important d'ajuster les hyperparamètres afin d'optimiser les performances des modèles. La recherche par quadrillage (Grid Search) sera utilisée.

Pour une liste de possibilités, pour chacun des hyperparamètres et pour chacune des combinaisons, la méthode entraînera le modèle puis retiendra la combinaison de paramètres pour laquelle le RMSE est minimal. Pour cela, la librairie "caret" du logiciel R sera utilisée. Cette librairie permet de tester des modèles par validation croisée.

Pour le Random Forest, nous avons calibré les hyperparamètres (a, b, c) représentant respectivement le nombre d'arbres décisionnels, la profondeur maximale de chaque arbre et le nombre minimal d'observations que doit avoir un nœud pour être subdivisé.

$$a \in \{100, 200, 300, 400, 500\}, b \in \{1, 2, \ldots, 10\} \text{ et } c \in (2000, 3000, 4000, 6000)$$

La meilleure combinaison de valeur retenue au sens du critère RMSE est $(400, 1, 6000)$ pour la fréquence et $(300, 1, 3000)$ pour la sévérité.

Le même processus est réalisé pour le Gradient Boosting pour lequel les hyperparamètres (a, b), représentant respectivement le nombre d'arbres et la profondeur maximale de chaque arbre, ont été calibrés.

$$a \in \{100, 200, 300, 400, 500\} \text{ et } b \in \{1, 2, \ldots, 10\}$$

Les couples de valeurs retenus sont $(100, 6)$ pour la fréquence et $(200, 2)$ pour la sévérité.

Une fois que les algorithmes de Machine Learning sont optimisés, il est possible d'afficher l'importance des variables, donnée par le Random Forest. Ce graphique permettra d'évaluer l'impact de chaque variable dans la construction des modèles.

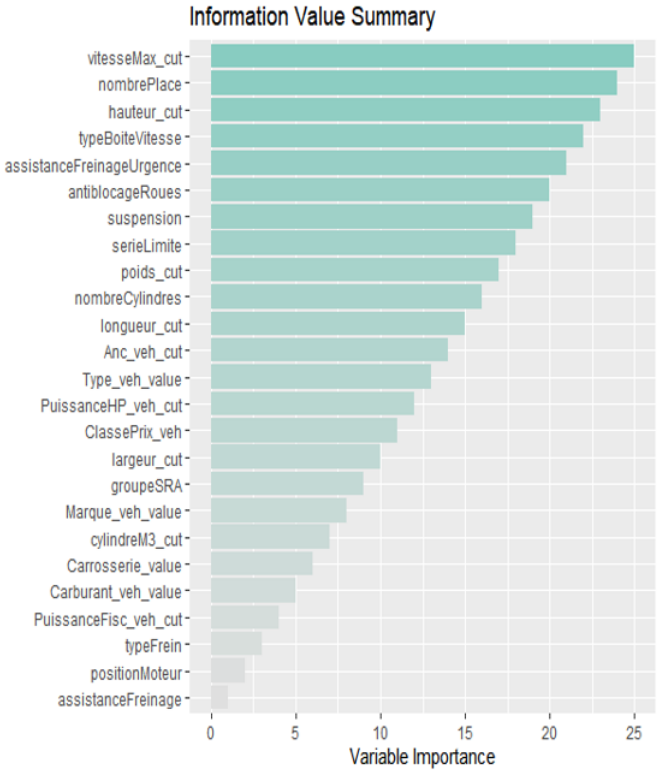


FIGURE 6.4 – Importance des variables pour le modèle de fréquence

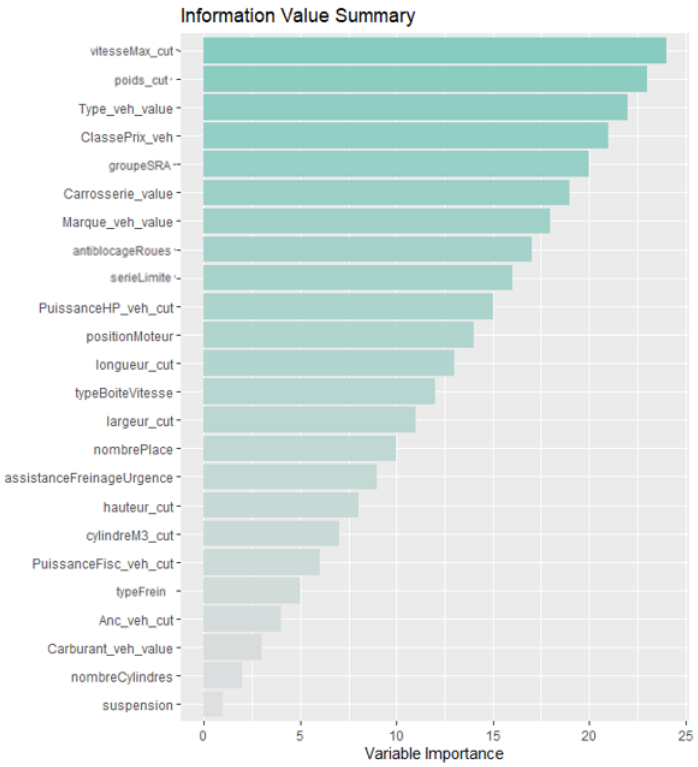


FIGURE 6.5 – Importance des variables pour le modèle de sévérité

L'importance des variables n'est pas la même pour les deux modèles. Pour le modèle de sévérité, les variables contribuant le plus à la construction du modèle, sont la vitesse maximale, le poids, le type de véhicule et le groupe SRA. La garantie couverte par le portefeuille étant la responsabilité civile matérielle, ces caractéristiques pourraient influencer les montants de sinistre indemnisés par l'assureur. En effet, la vitesse est un facteur déterminant lors d'un choc et le groupe SRA représente la dangerosité intrinsèque du véhicule.

6.2.2 Comparaison des modèles

Les modèles sont comparés suivant le critère du RMSE décrit en section 5.1.6. Cet indicateur a été calculé sur les échantillons d'entraînement et de validation.

Modèle	Algorithme	$RMSE_{app}$	$RMSE_{validation}$
Fréquence	Random Forest	1.6574	1.7354
	GBM	1.6535	1.7351
Sévérité	Random Forest	4.2097	4.5855
	GBM	4.2069	4.5835

TABLE 6.2 – Comparaison du Random Forest et du Gradient Boosting

Les valeurs du RMSE sont très proches pour chacun des modèles. Cependant, le Gradient Boosting est le modèle produisant les valeurs les plus faibles sur les échantillons d'apprentissages et de validations. Par ailleurs, on dénote un léger phénomène de sur-apprentissage, car les performances de l'échantillon de validation sont moins bonnes que celles de l'échantillon d'apprentissage. Ce phénomène est très courant pour les algorithmes basés sur les arbres de décision. Il peut être évité en réalisant une modélisation par validation croisée.

C'est donc à partir des résidus prédits par Gradient Boosting que nous allons créer les classes de véhicules.

Agrégation des résidus

Avant de regrouper les véhicules par classes de risque, il est important d'agréger les résidus modélisés par ML à la maille SRA. Ainsi nous disposerons d'une valeur unique de résidu par véhicule. Le résidu agrégé correspondra à la moyenne des résidus pour chaque code SRA.

$$R_j = \frac{\sum_{i=1}^{n} r_i^j}{n_j}$$

avec n_j le nombre d'observations du code SRA j et r_i^j le résidu associé à l'observation i ayant le code SRA j.

6.3 Classification des résidus modélisés

Une fois les résidus modélisés et agrégés par code SRA, nous allons regrouper le score prédit par groupe de risques homogènes. C'est cette nouvelle variable représentant les classes de véhicules qui représentera le véhiculer que nous intégrerons par la suite dans les modèles de fréquence et de coût. Chaque classe représentera un ensemble de véhicules ayant un niveau de risque similaire. Ce regroupement se fera à l'aide des méthodes de classification/clustering automatiques. Ce sont des techniques d'apprentissages automatiques reposant sur des algorithmes repérant des similarités dans les données en vue de les regrouper.

À cet effet, nous implémenterons deux méthodes de classification : la méthode des k-means et la méthode de Ward issue de la Classification Ascendante Hiérarchique (CAH). Pour chaque méthode, nous allons choisir le nombre de classes optimal puis nous comparerons les véhiculiers issus de chacune d'entre elles.

Les méthodes de clustering

Le clustering est une méthode d'apprentissage non supervisé qui a pour objectif de séparer des données en groupes homogènes ayant des caractéristiques communes. Contrairement à l'apprentissage supervisé qui essaie d'apprendre une relation de corrélation entre un ensemble de features X d'une observation et une valeur à prédire Y, le clustering va regrouper en plusieurs familles (clusters) les individus/objets en fonction de leurs caractéristiques. De cette manière, les individus se trouvant dans un même cluster sont similaires.

Il existe deux types de clustering :

- le clustering hiérarchique et
- le clustering non hiérarchique (partitionnement).

6.3.1 La méthode des k-means

Le k-means est un algorithme non supervisé de clustering non hiérarchique conçu en 1957 au sein des Laboratoires Bell par Stuart P.Lloyd comme technique de modulation par impulsion et codage(MIC). L'algorithme permet de regrouper les observations d'une base en k clusters distincts. Ainsi, les données similaires se retrouvent dans un même cluster. Par ailleurs, une observation ne peut se retrouver que dans un cluster à la fois (exclusivité d'appartenance). En d'autres termes, une même observation ne peut appartenir à deux clusters différents.

Principe

Les méthodes de clusterisation reposent sur des notions de distance. La méthode des k-means repose sur la minimisation de la somme des distances euclidiennes au carré entre chaque point et le centroïde (le point central) de son cluster. On parle de minimisation de la variation intra-cluster. La distance euclidienne entre deux observations x_1 et x_2 est donnée par :

$$d\left(x_1, x_2\right) = \sqrt{\sum_{j=1}^{n} \left(x_{1n} - x_{2n}\right)^2}$$

Algorithme des k-means

Entrée : le nombre de clusters k
Boucle :

1. Attribuer un cluster à chacune des observations
2. Calculer les centroïdes de chaque cluster
3. Calculer la distance euclidienne pour chaque observation avec les centroïdes de chacun des clusters
4. Attribuer à chaque observation le cluster le plus proche de lui
5. Calculer la somme de la variabilité intra-cluster
6. Réitérer jusqu'à atteindre la convergence : la convergence est atteinte lorsqu'il n'y a plus aucun changement de clusters après mise à jour.

Sortie : les k clusters finaux

Choix du nombre optimal de classes

L'algorithme des k-means n'est pas capable de déterminer le nombre de classes optimal et laisse ce choix à l'utilisateur. Cependant, ce choix n'est pas intuitif et il existe plusieurs méthodes basées sur la minimisation de la distance intra-classe.

On distingue trois méthodes pour déterminer le nombre optimal de classes :

- L'Elbow Method (méthode du coude) : basée sur la minimisation de la somme des carrés des écarts à l'intérieur des clusters.
- l'Average silhouette method : basée sur la maximisation du paramètre appelé "average silhouette".
- Gap statistic method : basée sur la comparaison de la variation totale intra-cluster pour différentes valeurs de k avec leurs valeurs attendues sous une distribution de référence nulle des données.

Dans ce mémoire, nous utilisons la méthode du coude pour déterminer le nombre optimal de classes. Elle consiste à lancer l'algorithme k-means avec différentes valeurs de k et à calculer la variance des différents clusters. Le principe est de minimiser la variance intra-classe en trouvant le nombre de clusters k pour lequel les clusters retenus minimisent la distance entre leurs centres (centroïdes) et les observations dans le même cluster.

La variance des clusters se calcule comme suit :

$$V = \sum_{j} \sum_{x_i \to c_j} D\left(c_j, x_i\right)^2$$

avec :

- c_j : le centre du cluster (le centroïde),
- x_i : la ième observation dans le cluster ayant pour centroïde c_j,
- $D\left(c_j, x_i\right)$: La distance euclidienne entre le centre du cluster et le point x_i.

Le graphique représentant l'évolution de la variance intra-classe en fonction des différents nombres de clusters k, pour le modèle de fréquence, est affiché sur la figure 6.6.

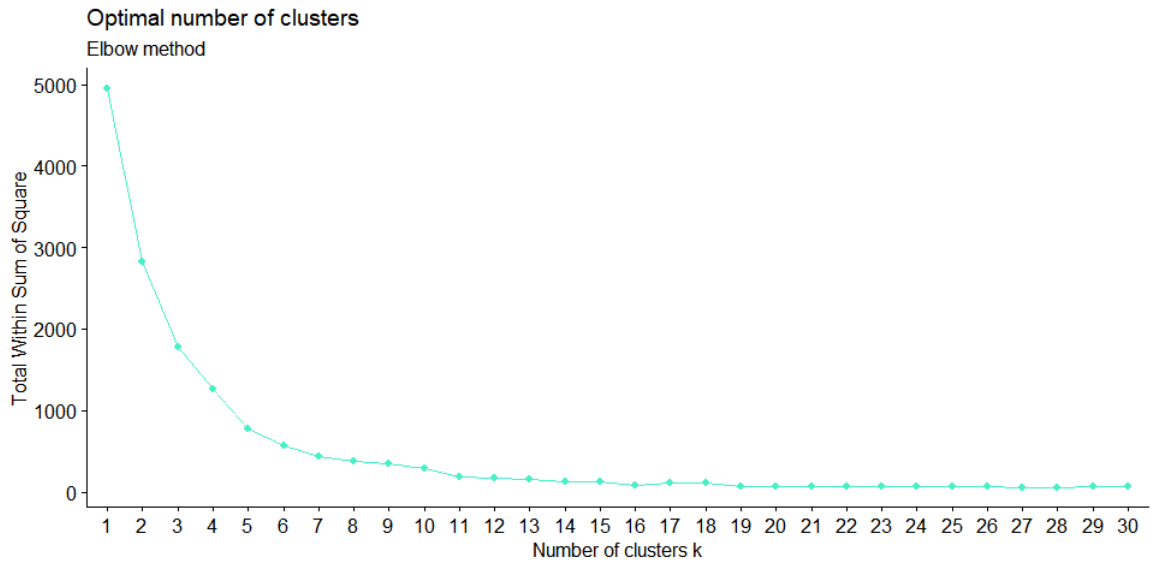


FIGURE 6.6 – Nombre de classes optimal pour le modèle de fréquence

Le nombre optimal de classes est celui à partir duquel la variance ne se réduit plus significativement. C'est ce point qui est appelé le coude.

À partir du graphique, nous pouvons dire que le nombre de classes optimal est 11 classes. Cependant, nous allons construire des véhiculier de 9, 10 et 11 classes afin de les comparer et choisir celui qui fournit le modèle optimal. Le même découpage sera effectué pour le modèle de sévérité.

6.3.2 La méthode de Ward

La méthode de Ward est un algorithme de Classification Ascendante Hiérarchique (CAH) permettant de regrouper deux classes d'une partition pour obtenir une partition plus agrégée. La CAH consiste à rassembler les individus d'une population selon un critère de ressemblance défini au préalable. Le critère de ressemblance s'exprime sous la forme d'une matrice de distances représentant la distance existant entre chaque individu pris deux à deux.

Cette matrice de distance est utilisée afin d'établir une classification en arborescence, c'est-à-dire que les individus sont rassemblés de manière itérative afin de produire un arbre de classification appelé dendogramme. La classification est ascendante, car elle part des observations individuelles ; elle est hiérarchique, car elle produit des classes ou groupes de plus en plus vastes, incluant des sous-groupes en leur sein.

Pour construire le dendrogramme, il existe plusieurs méthodes d'agrégation (méthode de Ward, distance minimale, distance maximale, etc.). Celle la plus courante est la méthode de Ward. Elle consiste à réunir les deux clusters dont le regroupement fera le moins baisser l'inertie interclasse. La distance utilisée par l'algorithme est la distance de Ward : la distance entre deux classes est celle de leurs barycentres au carré, pondérée par les effectifs des deux clusters.

Exemple : La distance de Ward entre deux classes A et B est donnée par la formule suivante :

$$d(A,B) = \frac{p_A p_B}{p_A + p_B} d(G_A, G_B)^2$$

où p_A et p_B sont les poids des deux classes et $d(G_A, G_B)$ représente la distance euclidienne de leurs centres de gravité.

Choix du nombre optimal de classes : Contrairement à la méthode des k-means, les méthodes de classification ascendantes hiérarchiques disposent d'une fonction permettant de déterminer le nombre de classes optimal. Il s'agit de la fonction *NbClust* du package du même nom sous R. Cette fonction permet d'appliquer simultanément plusieurs valeurs du nombre de classes et de déterminer la meilleure partition parmi toutes celles obtenues. Le nombre de classes optimal obtenu pour la méthode de Ward est 3 classes.

Visualisation des différentes classes construites

En vue de déterminer le nombre optimal de classes pour segmenter au mieux le risque, plusieurs valeurs ont été testées. Le nombre de classes k testé pour chaque algorithme varie entre 1 et 30 ($k \in 1, ..., 30$).

Le nombre de classes optimal qui ressort pour la méthode de Ward est 3 classes. Pour la méthode des k-means, des véhiculiers de 9, 10 et 11 classes ont été construits puis comparés afin de choisir le véhiculier le plus optimal.

Ainsi, nous pouvons représenter la répartition des résidus agrégés par clusters (exemple du modèle de fréquence).

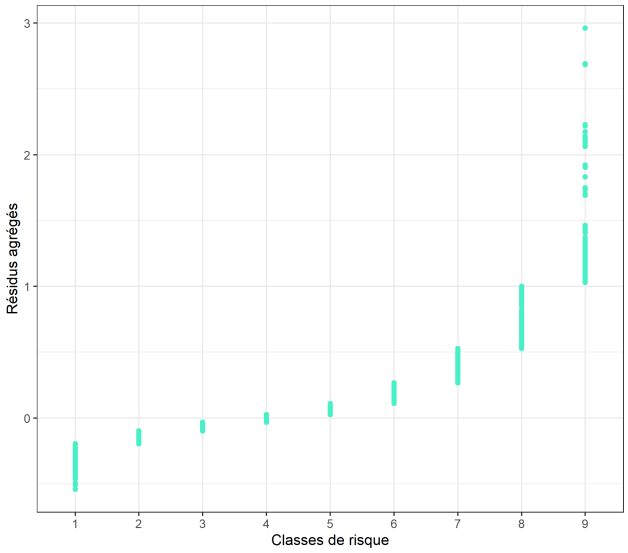


FIGURE 6.7 – Méthode des k-means avec k=9

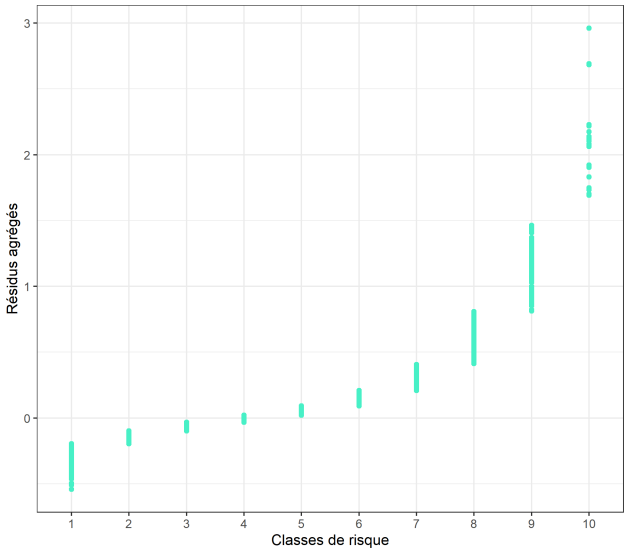


FIGURE 6.8 – Méthode des k-means avec k=10

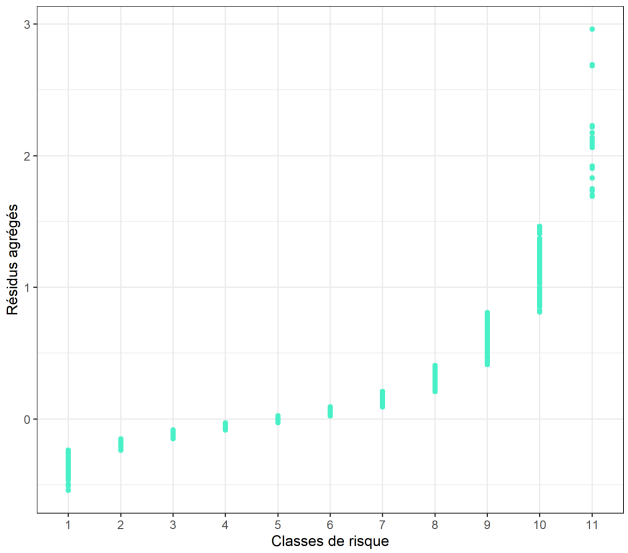


FIGURE 6.9 – Méthode des k-means avec k=11

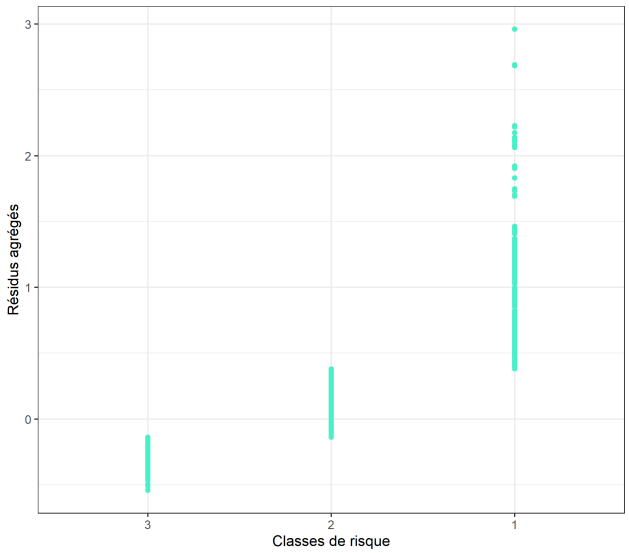


FIGURE 6.10 – Méthode de Ward avec k=3

Chaque classe est définie par un intervalle de valeurs de résidus. Les classes sont définies de sorte qu'une valeur de résidu ne puisse se retrouver dans deux classes distinctes. De plus, les classes sont numérotées par degré de risque croissant, c'est-à-dire de la moins risquée à la plus risquée.

L'étape suivante consiste à intégrer les quatre véhiculiers construits, pour chacun des modèles de fréquence et de coût, dans les modèles GLM initiaux afin de les comparer au sens des indicateurs de performance.

6.4 Intégration des véhiculiers dans les modèles de tarification GLM

Cette étape constitue la dernière étape de l'approche mise en œuvre. Elle consiste à intégrer les véhiculiers issus des deux méthodes de classification dans les modèles de tarification GLM. Il sera ainsi possible de comparer les différents véhiculiers afin de retenir celui qui réalise les meilleures performances pour chacun des modèles de fréquence et de coût.

À l'issue de la clusterisation de la section précédente, nous disposons de 4 véhiculiers distincts pour chaque modèle. Afin de tester ces véhiculiers et de choisir celui qui est le plus optimal, ils seront intégrés dans les modèles GLM construits dans le chapitre précédent. Les critères considérés pour les comparaisons sont l'AIC et la déviance.

Cependant, avant d'évaluer les performances de chaque véhiculier sur l'échantillon de validation, il est important de les créer pour cet échantillon. En d'autres termes, il s'agit de répliquer les clusters de résidus créés à partir de l'échantillon d'apprentissage sur l'échantillon de validation. Pour cela, nous avons créé une fonction qui à chaque valeur de résidu agrégé de l'échantillon de validation associe la classe correspondante dans l'échantillon d'apprentissage. En effet, les classes de la base d'apprentissage sont construites par segments de résidus (voir figures 6.7 à 6.10). Donc chaque résidu de la base de validation appartient à un segment donné et peut être associé à la classe correspondante.

Une fois les véhiculiers disponibles dans la base de validation, les performances des modèles GLM incluant ces variables peuvent être évaluées. En effet, il s'agit de faire une nouvelle régression de la fréquence et du coût (sur l'échantillon d'apprentissage) en fonction des variables non véhiculières et du véhiculier puis d'évaluer leurs prédictions sur l'échantillon de validation.

Afin de comparer les modèles incluant les différents véhiculiers, les valeurs des indicateurs statistiques permettant de les comparer ont été consignés dans le tableau suivant.

Méthode	Nombre de classes	AIC_{app}	AIC_{val}	Déviance$_{app}$	Déviance$_{val}$
k-means	k=9	109 157.5	25 538.71	78 473.09	21 276.92
	k=10	109 118.3	25 509.11	77 923.00	21 425.39
	k=11	109 001.5	25 469.99	77 803.99	19 603.92
Ward	k=3	111 123.7	26 229.45	80 142.44	21 971.73

TABLE 6.3 – Comparaison des méthodes de clusterisation pour le modèle de fréquence

Méthode	Nombre de classes	AIC_{app}	AIC_{val}	Déviance$_{app}$	Déviance$_{val}$
k-means	k=9	403 746.5	99 366.28	44 395.79	10 713.126
	k=10	403 620.7	99 307.91	44 255.93	10 672.278
	k=11	403 499.6	99 299.75	41 579.3	10 361.01
Ward	k=3	407 277.6	100 950.28	48 475.00	11 259.531

TABLE 6.4 – Comparaison des méthodes de clusterisation pour le modèle de coût

Le véhiculier minimisant l'AIC et la déviance est celui constitué de 11 classes, construit à partir de la méthode des k-means. Intuitivement, le véhiculier de 3 classes issu de la méthode de Ward n'aurait pas été retenu, car nous avons besoin de finesse dans la segmentation et ce n'est pas avec 3 classes que l'on pourra se distinguer sur le marché.

Pour les modèles de fréquence et de coût, les véhiculiers retenus sont les véhiculiers constitués de 11 classes.

Chapitre 7

Mise en place d'un modèle de tarification comparatif

L'objectif de ce chapitre est de construire un modèle GLM incluant les variables véhiculières, auquel on pourra comparer les modèles incluant le véhiculier.

Dans le but d'évaluer au mieux l'intérêt du véhiculier, il est important de disposer d'un modèle de référence auquel seront comparés les modèles contenant le véhiculier. Pour ce faire, nous construirons deux modèles GLM de fréquence et de coût intégrant toutes les variables (véhiculières et non véhiculières) ; c'est-à-dire des modèles contenant toute l'information disponible. La comparaison de ces différents modèles permettra de mesurer l'importance de résumer toute l'information liée au véhicule en une seule variable plutôt que d'utiliser directement les variables véhiculières dans les modèles de tarification.

Ce modèle sera calibré en passant par une étape de sélection de variable utilisant la régression pénalisée Lasso. Cependant, étant donné la complexité de ce modèle, une étape supplémentaire de recherche d'interactions entre les variables a été implémentée. En effet, ce modèle GLM prendra en entrée 24 variables véhiculières (voir tableau 6.1) et 9 variables non véhiculières (voir section 5.2.2).

Les modèles linéaires généralisés ne permettent pas par défaut la prise en compte des effets croisés entre variables (interactions), à moins de les spécifier manuellement dans la formule de régression. Cependant, il n'existe pas de méthode automatique permettant d'identifier efficacement l'existence d'interactions. Il serait possible d'utiliser les procédures stepwise afin de sélectionner les effets croisés pertinents, mais cette procédure serait trop coûteuse en temps, car le nombre d'itérations serait trop important.

Une méthode introduisant automatiquement les termes d'interactions entre les variables est la Régression Multivariée par Spline Adaptative (MARS). Cette méthode est une technique de régression non paramétrique pouvant être vue comme une extension des régressions linéaires en modélisant automatiquement les interactions et les non-linéarités. Dans le cadre de ce mémoire, l'algorithme MARS sera utilisé uniquement pour déterminer les éventuelles interactions qu'il faudra intégrer dans le modèle GLM.

La Régression Multivariée par Spline Adaptative (MARS)

Les MARS sont des algorithmes conçus pour les problèmes de régression non linéaire multivariée. Introduites par Jerome H. Friedman en 1991 [10], elles ne font aucune hypothèse sur la manière dont les variables sont liées aux prédicteurs et permettent à la fonction du modèle de suivre directement

les données.

L'algorithme crée automatiquement des fonctions linéaires par morceaux caractérisant les données puis les utilise de manière agrégée pour faire une prédiction. Plus précisément, il génère des fonctions bilatérales tronquées de la forme :

$$(x-t)_+ = \left\{ \begin{array}{ll} x-t & x > t \\ 0 & \text{sinon} \end{array} \right\}$$

Ces fonctions sont appelées fonctions de bases et tentent d'approcher les relations entre les variables prédictives. La sortie de chaque fonction est pondérée par un coefficient puis une prédiction est réalisée en additionnant la sortie pondérée de toutes les fonctions de base du modèle. L'approche se décompose en deux étapes distinctes.

La première étape « étape avant » génère des fonctions de bases candidates et les ajoute au modèle. Cette étape est considérée comme un passage ascendant et est similaire à l'algorithme de partitionnement récursif utilisé dans les arbres de décision. En effet, comme dans les algorithmes d'arbres de décision, on considère chaque valeur de chaque variable d'entrée dans l'ensemble de données d'apprentissage comme un candidat pour une fonction de base. Le processus d'ajout de termes est itéré jusqu'à ce que le changement d'erreur résiduelle soit trop faible pour continuer ou jusqu'à ce que le nombre maximum de termes soit atteint. Le nombre maximum de termes est spécifié au préalable par l'utilisateur avant le début de la construction du modèle.

La seconde étape « étape arrière » ou passage descendant consiste à simplifier le modèle à travers un processus d'élagage. Les termes du modèle sont supprimés un à un en partant du moins efficace, c'est-à-dire celui dont la suppression entraîne une faible (ou aucune) amélioration du modèle ; jusqu'à ce que le meilleur sous-modèle soit trouvé. Les performances du modèle à chaque suppression sont évaluées à l'aide de la validation croisée de l'ensemble de données d'entraînement selon l'erreur dite de Validation Croisée Généralisée (GCV). C'est une mesure de qualité d'ajustement qui tient non seulement compte de l'erreur des résidus, mais également de la complexité du modèle.

Pour terminer, les deux paramètres clés à prendre en compte et qui permettent de limiter le risque de sur-apprentissage sont : le degré maximum d'interactions et nombre de termes retenus dans le modèle final.

Calibration du modèle de référence

Sélection de variables

Comme dans la section 5.2.3.1, la sélection de variables est effectuée en utilisant la régression pénalisée de Lasso. Les variables retenues pour chacun des modèles sont consignés dans le tableau suivant :

Variables	Modèle de fréquence	Modèle de sévérité
Âge de l'assuré	×	×
CSP	×	
CRM	×	
Ancienneté CRM	×	
Statut Marital		
Mode d'apprentissage conduite		
Formule police	×	×
Usage véhicule	×	×
Mode de stationnement	×	×
Marque		
Carrosserie		×
Ancienneté du véhicule		×
Classe de Prix	×	×
Carburant		
Puissance fiscale		
Puissance Hp		
Type de véhicule		
Vitesse maximale	×	×
Nombre de places	×	
Poids	×	×
Hauteur		
Longueur		
Largeur		
Groupe SRA	×	×
Cylindre M3		
Suspension	×	
Assistance freinage d'urgence	×	
Antiblocage des roues	×	
Type de frein		
Série limitée		
Nombre de cylindres		
Position du moteur		
Type boîte de vitesse		

TABLE 7.1 – Sélection de variables pour le modèle comparatif

Pour résumer, 15 variables ont été retenues pour le modèle fréquence et 10 variables pour celui de sévérité. Il est important de remarquer que les variables véhiculières sélectionnées font partie des variables ayant le plus contribué à la construction des modèles de Machine Learning en section 6.2.1.

Détection des interactions

L'une des principales limites des modèles linéaires généralisés est l'impossibilité de détection des éventuelles interactions entre les variables explicatives du modèle. Pour palier à cette limite, nous nous tournons vers l'algorithme MARS, une extension des modèles linéaires s'affranchissant de l'hypothèse de linéarité tout en prenant automatiquement en compte les interactions.

La mise en place de ce processus passe par un processus de validation croisée ($k = 10$) grâce à library $caret$. Les hyperparamètres testés sont : le degré maximum d'interaction d limité à 3 $(d = 1, 2, 3)$ et le nombre de termes n limité à 100. L'algorithme retourne la valeur 1 pour le degré d'interaction. Ce qui veut dire qu'il ne détecte aucune interaction entre les variables explicatives des modèles.

Résultats

L'objectif de cette section est de mettre en place un modèle que l'assureur aurait pu utiliser s'il n'avait pas eu recours au véhiculier. Suite à une sélection judicieuse des variables candidates aux GLM et à la recherche d'interactions entre les variables grâce à la méthode de MARS, les modèles ont été calibrés. Aucune interaction n'ayant été détectée, la fréquence et le coût de sinistres ont donc été modélisés en fonction des variables sélectionnées. Les performances de ce modèle seront analysées dans la section 8.1 lors de la comparaison de tous les modèles construits.

Chapitre 8

Analyse des résultats et limites de l'approche

Dans les chapitres précédents, les différentes étapes de la construction du véhiculier ont été mises en œuvre. Dans ce chapitre, nous allons intégrer les véhiculiers sélectionnés dans les modèles GLM de tarification. Afin de mesurer l'apport prédictif de cette nouvelle variable, les modèles incluant le véhiculier et les modèles sans véhiculier seront comparés.

8.1 Comparaison des modèles avec et sans véhiculier

Les modèles qui ont été comparés sont les suivants :

- Modèle A : GLM contenant les variables non véhiculières uniquement.
- Modèle B : GLM contenant les variables non véhiculières plus le véhiculier construit.
- Modèle C : GLM comparatif contenant les variables non véhiculières et les variables véhiculières.

Les performances des modèles ont été calculées sur la base d'apprentissage puis testés sur une base de test et de validation. Les indicateurs statistiques de performance pour chaque modèle sont consignés dans le tableau suivant.

Modèles		AIC_{app}	$Déviance_{app}$	AIC_{val}	$Déviance_{val}$	AIC_{test}	$Déviance_{test}$
Fréquence	Modèle A	145 385.5	99 748.12	36 201.46	24 786.48	80 519.6	76 692.5
	Modèle B	109 001.5	77 803.99	25 469.99	19 603.92	56 241.6	52 400.5
	Modèle C	113 833.3	73 004.24	28 719.64	16 388.27	62 057.5	41 135.0
Coût	Modèle A	520 239.2	53 690.29	130 041.2	13 203.56	178 901.8	17 079.5
	Modèle B	403 499.6	41 579.3	99 299.75	10 361.01	124 755.7	12 132.6
	Modèle C	409 831.4	34 121.67	102 816.3	6 652.843	139 790.1	9 403.4

TABLE 8.1 – Comparaison des modèles GLM avec et sans véhiculier

De manière générale, au niveau du Modèle B, nous observons une diminution de l'AIC et de la déviance par rapport au modèle A. Pour le modèle de fréquence, l'AIC et la déviance connaissent une amélioration (diminution) de 30% et 21% sur l'échantillon de validation. Il en est de même pour le modèle de coût où l'AIC diminue de 24% et la déviance de 22%. Le même constat est effectué sur l'échantillon test.

Le modèle C quant à lui est préférable au modèle A, mais réalise des performances moins bonnes que le modèle B. Sa déviance est inférieure à la déviance du modèle B, car il est plus complexe, donc l'écart entre la log-vraisemblance du modèle saturé et celle du modèle étudié, est plus faible.

Afin de consolider les remarques précédentes, nous comparons les valeurs des indices de Gini pour chaque modèle. Ces résultats sont présentés dans le tableau ci-dessous. On remarque que l'indice de Gini du modèle B est nettement meilleur que celui du modèle C, et ce, sur les deux échantillons. Le modèle B est donc celui qui discrimine le mieux le risque étudié.

Modèles		Échantillon d'apprentissage	Échantillon de validation
Fréquence	Modèle C	19.83%	18.42%
	Modèle B	23.27%	21.34%
	Amélioration	17%	16%
Coût	Modèle C	43.12%	42.75%
	Modèle B	55.62%	53.09%
	Amélioration	29%	24%

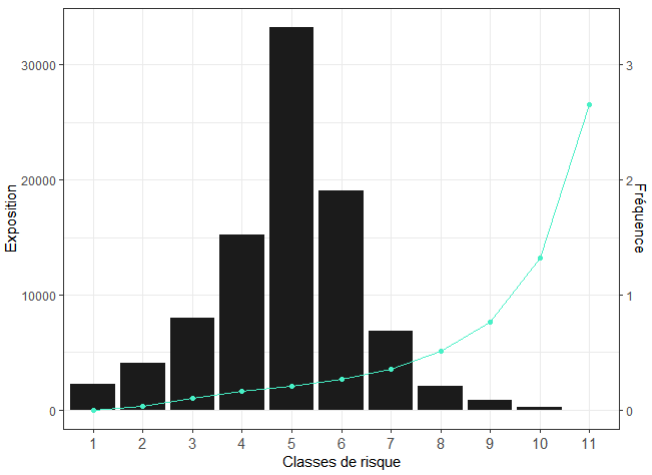
Table 8.2 – Comparaison des coefficients de GINI

Ces différentes analyses mettent en évidence l'importance du véhiculier en ce sens que le modèle B contenant le véhiculier réalise de meilleures performances que les autres modèles. Il est donc préférable de synthétiser toute l'information liée au véhicule en une seule variable plutôt que d'intégrer directement les variables véhiculières dans le modèle GLM.

8.2 Interprétation des classes du véhiculier et étude des prédictions

L'intégration du véhiculier dans les GLM permet non seulement d'optimiser les modèles de tarification, mais aussi de mieux interpréter l'évolution de la fréquence et du coût par segment de risque. En effet, il est possible d'identifier les classes de véhicules les plus risquées. Cet outil est utile à l'actuaire lors du processus de tarification, car il lui permettra d'adapter son tarif suivant les profils de risque. Par exemple, les individus dont les véhicules sont considérés comme hautement risqués selon le véhiculier, verront leur tarif augmenter. L'anti-sélection sera ainsi évitée et la clientèle fidélisée, car les individus possédant des véhicules jugés à faibles risques verront leur prime baisser.

Les graphes ci-dessous présentent l'évolution de la fréquence et du coût de sinistre par classes de risque.



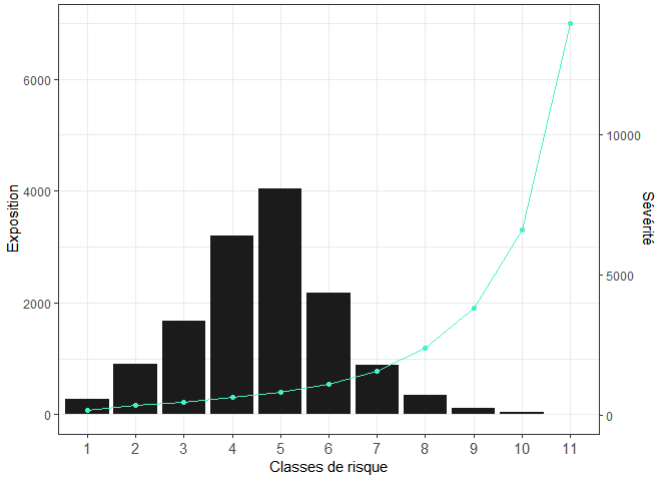


FIGURE 8.1 – Évolution de la fréquence et de la sévérité par classes de risque

La fréquence et le coût moyen sont croissants en fonction de la classe de risque. Cette observation est pertinente, car les classes sont numérotées de la moins risquée à la plus risquée. Cependant, l'exposition des trois dernières classes est très faible par rapport à l'exposition des autres classes. Il est donc possible de regrouper ces classes en une seule classe afin d'avoir une meilleure représentativité de la classe. On aboutit ainsi à un véhiculier de 9 classes.

Analyse de la qualité de prédiction

Pour finir, il est important de vérifier si les nouveaux modèles capturent bien les effets de chaque segment de risque pour les différentes variables. À cet effet, les valeurs de fréquence et de coût prédits sont comparées aux valeurs observées sur l'échantillon de validation. Sur les graphes ci-dessous, les valeurs prédites sont représentées en vert, les valeurs observées en rouge et les intervalles de confiances de niveau 95% en bleu. Les expositions ont été multipliées par un coefficient afin d'être mise à la même échelle que la fréquence.

Prédictions pour le modèle de fréquence

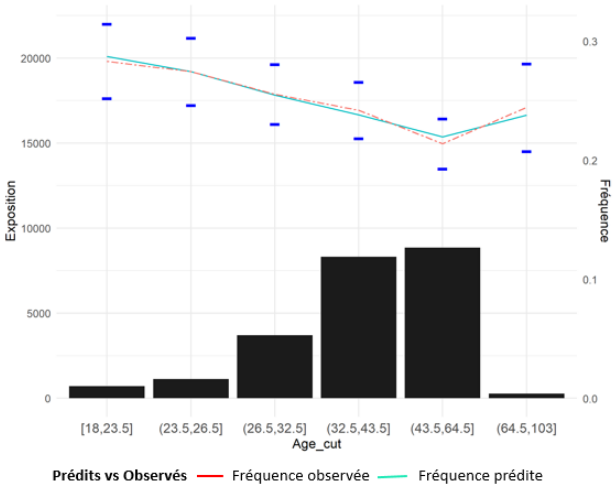


FIGURE 8.2 – Fréquence prédite pour l'âge de l'assuré

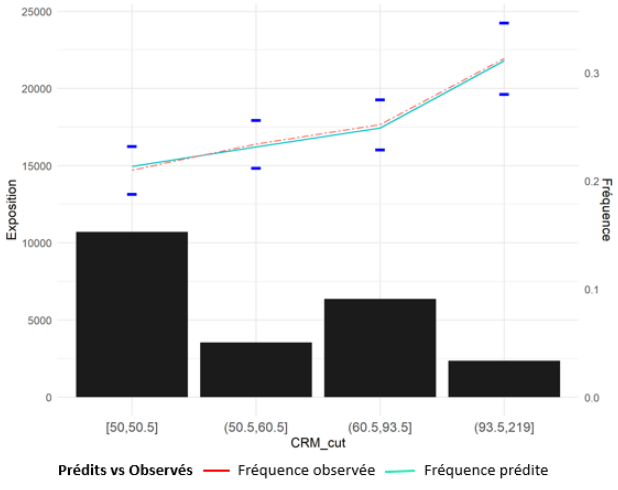


FIGURE 8.3 – Fréquence prédite pour la variable CRM

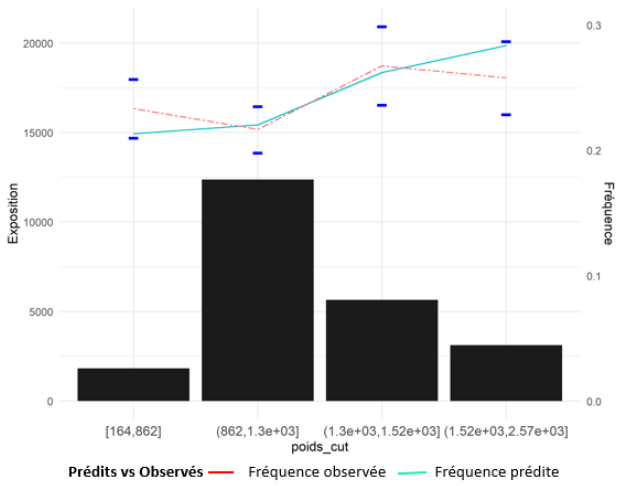


FIGURE 8.4 – Fréquence prédite pour le poids du véhicule

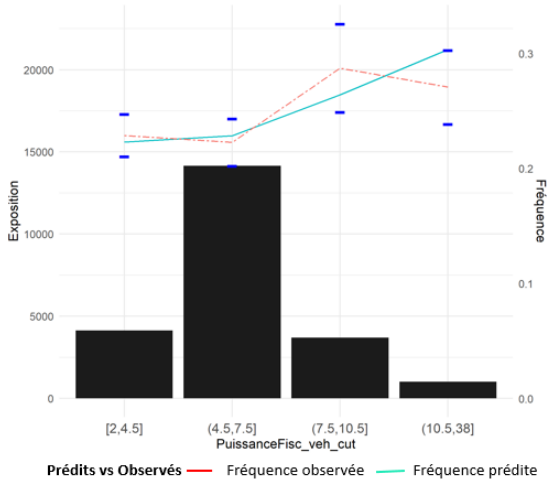


FIGURE 8.5 – Fréquence prédite pour la puissance fiscale

Prédictions pour le modèle de sévérité

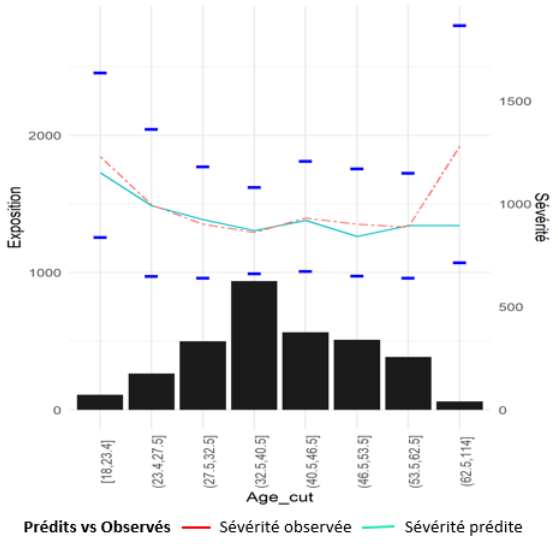


FIGURE 8.6 – Sévérité prédite pour l'âge de l'assuré

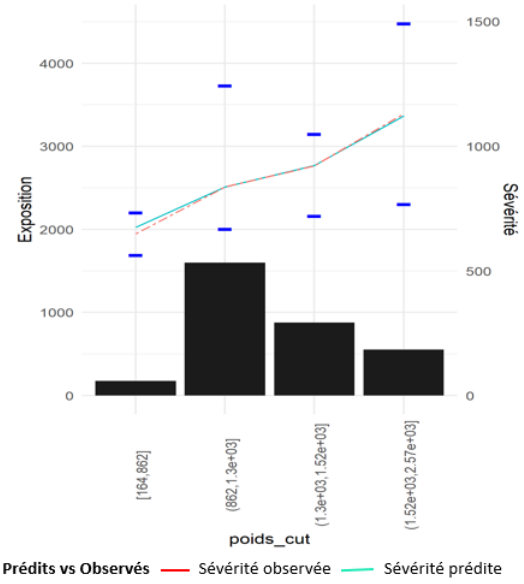


FIGURE 8.7 – Sévérité prédite pour le poids du véhicule

Pour les variables représentées [1], les prédictions sont proches des observations et présentent la même tendance. On remarque cependant une sous-estimation et une surestimation du risque pour certaines classes. Ces écarts entre valeurs prédites et valeurs observées peuvent s'expliquer par le fait que les modèles ne capturent pas tous les effets de risque de ces classes. Par ailleurs, les prédictions se retrouvent dans les intervalles de confiance, qui sont beaucoup plus larges pour le modèle de sévérité, dont la base de validation est 4 fois plus petite que celle de la base d'apprentissage.

Cependant, il est important de faire remarquer que les prédictions des variables véhiculières sont moins bonnes que celles des variables non véhiculières pour le modèle de fréquence. En effet, pour la garantie Responsabilité Civile Matérielle, la fréquence du sinistre serait plutôt influencée par les

1. Des analyses complémentaires sont disponibles en annexe D.

caractéristiques liées au conducteur (son âge, l'usage du véhicule, le mode d'apprentissage de la conduite, etc). Cette remarque se confirme au niveau du modèle de sévérité pour lequel les prédictions des variables non-véhiculières sont moins précises tandis que le modèle capture mieux bien les effets de risques pour les variables véhiculières. Le risque étudié ici est le montant des dommages causés à autrui. Ce montant dépendrait donc de la puissance du choc et serait influencé par les caractéristiques du véhicule (le poids, la puissance, la vitesse maximale, etc). De plus, une analyse des classes les plus risquées du modèle de coût, montre que les véhicules dominant dans ces classes sont des véhicules utilitaires et des véhicules de luxe de type sportifs. Les véhicules utilitaires sont des véhicules lourds et puissants qui peuvent causer beaucoup de dégâts lors d'un choc.

En outre, les graphiques comparant les prédictions des modèles contenant le véhiculier à celles du modèle comparatif [2] sont représentés en annexe D. Dans l'ensemble, les prédictions des deux modèles sont très proches. Cependant, les prédictions du modèle contenant le véhiculier semblent meilleures que celles du modèle comparatif.

2. Les prédictions du modèle comparatif sont représentées en noir et celle du modèle contenant le véhiculier en vert.

8.3 Limites de l'approche et axes d'amélioration

L'étude a démontré que l'intégration d'un véhiculier dans les modèles de tarification GLM améliore nettement les résultats et permet une bonne compréhension de la sinistralité par segments de risque. Au niveau de l'analyse des indicateurs de performance, les modèles contenant le véhiculier se sont révélés plus performants que les modèles contenant les variables véhiculières. En effet, le fait de synthétiser toute l'information liée au véhicule en une seule variable permet de simplifier les modèles et d'avoir de meilleures performances. Cependant, cette synthétisation de l'information s'accompagne d'une perte d'information qu'il ne faudrait pas négliger. Une solution serait peut-être de créer un modèle alternatif contenant les variables véhiculières les plus pertinentes en plus d'un véhiculier qui regrouperait les variables les moins importantes. Ainsi, la perte de l'information serait minimisée.

L'analyse de la qualité de prédiction mets en cause la pertinence de la construction d'un véhiculier pour la garantie Responsabilité Civile Matérielle. Pour le modèle de fréquence, les prédictions des variables véhiculières sont moins précises que celles des variables non véhiculières. Ceci peut s'expliquer par le fait que la fréquence de la garantie Responsabilité Civile Matérielle serait beaucoup plus influencée par les caractéristiques de l'assuré et de la police. Le véhiculier serait donc plus pertinent pour le modèle de sévérité en ce sens que le montant des dommages dépendrait des caractéristiques du véhicule (performances et masse).

Par ailleurs, quelques difficultés ont été rencontrées lors de la mise en œuvre de l'approche. L'une des principales difficultés était la clusterisation des résidus prédits. En effet, l'inconvénient de la méthode des k-means est qu'il revient à l'utilisateur de choisir le nombre de classes optimal. L'algorithme a donc été implémenté avec différents nombres de classes et ce processus a été coûteux en temps. D'autres valeurs de k auraient pu être testées, mais nous n'avons testé que trois valeurs afin de ne pas alourdir l'étude.

En outre, le nombre de classes obtenues pour le véhiculier paraît insuffisant par rapport à la quantité de véhicules présents sur le marché. Cependant, compte tenu du volume de données à disposition de cette étude, réaliser une segmentation avec un nombre de classes trop élevé n'aurait pas été approprié, car certaines classes n'auraient pas une exposition suffisante.

Une alternative à l'approche de ce mémoire pourrait être basée sur le lissage spatial des résidus prédits par Machine Learning. Il s'agirait de s'inspirer des méthodes de classification géographique en créant une carte de véhicule en vue d'effectuer un lissage par voisinage. Cette approche aurait pu être comparée à l'approche directe implémentée dans ce mémoire, mais elle n'a pas été mise en œuvre par manque de temps. Elle pourrait faire l'objet de travaux ultérieurs. De plus, cette approche pourrait faciliter la mise à jour du véhiculier car il suffirait de projeter les nouveaux véhicules sur la carte des véhicules afin de déterminer leur classe de risque.

Il serait également utile d'enrichir la sélection de variables véhiculières par de nouvelles variables indicatrices de la sécurité du véhicule. Il s'agit notamment des notations de sécurité des nouveaux véhicules qui sont déterminés grâce à des crash tests. En effet, il existe des organismes tels que l'Euro NCAP (European New Car Assessment Program) qui effectuent des essais de choc (crash tests) afin de tester les capacités dans le domaine de la sécurité passive des véhicules automobiles. Cette nouvelle variable pourrait jouer un rôle déterminant dans la segmentation des véhicules en assurance automobile, principalement pour la garantie dommage.

Enfin, l'automatisation du véhiculier pourrait faire l'objet d'études ultérieures. Cette automatisation permettra de mettre à jour le véhiculier afin de prendre en compte les véhicules qui ne sont pas encore commercialisés ou ceux non présents dans la base SRA au moment de l'étude.

Conclusion

Le véhiculier est un outil de segmentation utilisé par les actuaires en vue d'adapter leurs tarifs aux différents segments de risque. L'objectif de ce mémoire était de construire un véhiculier à partir de méthodes de Machine Learning sur un portefeuille couvrant la Responsabilité Civile Matérielle. Pour cela, une étude approfondie de la base a été effectuée en passant par la distinction des sinistres conventionnés et des sinistres non conventionnés. L'approche mise en œuvre dans ce mémoire est basée sur une modélisation de résidus et a été mise en œuvre en suivant les étapes suivantes :

- Construction de modèles GLM de fréquence et de coût hors variables véhiculières,
- Modélisation des résidus de ces GLM par Machine Learning,
- Création de classes de véhicules par clusterisation des résidus prédits.

À chaque étape, plusieurs méthodes ont été implémentées puis comparées afin de retenir celles qui réalisaient les meilleures performances au sens des indicateurs statistiques. Par exemple, le Gradient Boosting a été retenu contre le Random Forest pour la prédiction des résidus et la méthode des k-means a été retenue contre la méthode de Ward pour la clusterisation.

Toute l'information liée au véhicule a donc été regroupée en une seule variable tarifaire qui a été intégrée dans les modèles GLM initiaux. La comparaison des GLM avec véhiculiers et des GLM sans véhiculiers a montré que le véhiculier améliore les performances du modèle initial. Cependant, l'information perdue en synthétisant l'information des variables véhiculières ne devrait pas être négligée. De plus, l'étude a révélé que la construction d'un véhiculier pour la garantie Responsabilité Civile Matérielle, serait plus pertinente pour le modèle de sévérité. En outre, la construction d'un véhiculier est un processus long et coûteux en temps.

Toutefois, l'option du véhiculier n'est pas à écarter, car en plus d'optimiser les modèles de tarification, le véhiculier favorise l'interprétation des résultats lors de la tarification par un assureur. Il devient donc plus aisé de distinguer les classes les moins risquées et celles les plus risquées en termes de fréquence et de coûts. De plus, le véhiculier réduit considérablement les temps de calcul d'un modèle GLM en ce sens qu'il permet de simplifier les modèles. Sur le long terme, le véhiculier peut s'avérer utile, surtout lorsque l'on dispose d'un nombre important de variables véhiculières.

En somme, les résultats issus de notre approche sont concluants. Cependant, il existe des axes d'amélioration. Il serait utile de comparer l'approche développée dans ce mémoire à d'autre approches telles que celles basées sur un lissage spatial afin de décréter la méthode la plus optimale. De plus, il faudrait mettre en œuvre un processus optimal pour mettre à jour le véhiculier afin de prendre en compte les véhicules qui n'étaient pas présents dans la base au moment de l'étude.

Bibliographie

[1] Leo Breiman. "Bagging predictors". *Machine learning* 24.2 (1996), p. 123-140.

[2] Leo Breiman. "Random forests". *Machine learning* 45.1 (2001), p. 5-32.

[3] Leo Breiman et al. *Classification and regression trees*. Routledge, 2017.

[4] Arthur Charpentier et Michel Denuit. "Mathématiques de l'assurance non-vie". *Economica, Paris* (2005).

[5] "Convention d'Indemnisation Directe de l'Assuré et de Recours entre Sociétés d'Assurance Automobile."

[6] Michel Denuit, Dominik Sznajder et Julien Trufin. "Model selection based on Lorenz and concentration curves, Gini indices and convex order". *Insurance : Mathematics and Economics* 89 (2019), p. 128-139.

[7] Christophe Dutang. "Actuariat de l'assurance non-vie". *Cours ENSAE* (2020).

[8] Yoav Freund et Robert E Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of computer and system sciences* 55.1 (1997), p. 119-139.

[9] Jerome H Friedman. "Greedy function approximation : a gradient boosting machine". *Annals of statistics* (2001), p. 1189-1232.

[10] Jerome H Friedman. "Multivariate adaptive regression splines". *The annals of statistics* (1991), p. 1-67.

[11] "https ://www.argusdelassurance.com/cas-pratique/la-franchise-en-assurance-automobile.51279".

[12] Julie Lavenu. "Les méthodes de machine learning peuvent-elles être plus performantes que l'avis d'experts pour classer les véhicules par risque homogène ?" Mém. de mast. ISFA, (2016).

[13] Antoine Pesnaud. "Création de zoniers en assurance habitation à l'aide de variables externes et de méthodes de Data Science". Mém. de mast. ISUP, (2019).

[14] Matthieu Quilfen. "Classification des Véhicules en Assurance Automobile". Mém. de mast. ISFA, (2018).

[15] CR Rao. "Karl Pearson chi-square test the dawn of statistical inference". *Goodness-of-fit tests and model validity*. Springer, 2002, p. 9-24.

[16] Christian-Yann Robert. “Extreme-Value Theorie”. *Cours ENSAE* (2020).

[17] Magali Ruimy. “Elaboration d’un véhiculier en assurance automobile”. Mém. de mast. ISUP, (2016).

Table des figures

Liste des tableaux

Annexe A

Historique des forfaits IRSA-IDA

Année	2014	2015	2016	2017	2018	2019	2020	2021
Forfaits IRSA	1276	1308	1354	1420	1446	1482	1568	1678

Table A.1 – Montant des forfaits IRSA depuis 2014

Le test de Kolmogorov-smirnov

Le test de Kolmogorov-Smirnov (K-S) est un test d'adéquation non paramétrique permettant de tester l'hypothèse que deux échantillons sont issus de la même loi (hypothèse H_0). Il consiste à mesurer l'écart maximum qui existe soit entre une fonction de répartition empirique (donc des fréquences cumulées) et une fonction de répartition théorique, soit entre deux fonctions de répartition empiriques. La distance entre deux distributions A et B peut être définie par :

$$D = \max |F_A(x) - F_B(x)|$$

	D	P-valeur
Loi de Poisson	0.0064573	0.8222
Loi Binomiale-négative	0.77011	2.2e-16

Table A.2 – Tests d'adéquation de Kolmogorov-Smirnov pour la fréquence

D'après le tableau ci-dessus, l'hypothèse H_0 est acceptée pour la loi de Poisson et rejeté pour la loi Binomiale-négative, au seuil de 1%.

Liste des variables de la base SRA

Liste des variables contenues dans la base SRA
Marque
Modèle
Carrosserie
Version
Type
Alimentation
Énergie
Nombre de cylindres
Disposition des cylindres
Cylindrée (en m^3)
Puissance (en watt)
Régime (en DIN)
Transmission
Boîte de vitesse
Nombre de rapports
Suspension
Type de freinage
Dimensions
Nombre de places
Longueur
Largeur
Hauteur
Poids
Équipement
Airbags
Antiblocage roues
Classe de prix actuelle
Classe de prix d'origine
Groupe SRA
Classe de réparation d'origine
Classe de réparation actuelle
Date de début de commercialisation
Date de fin de commercialisation

Table A.3 – Base SRA

Annexe B

Compléments sur les analyses univariées

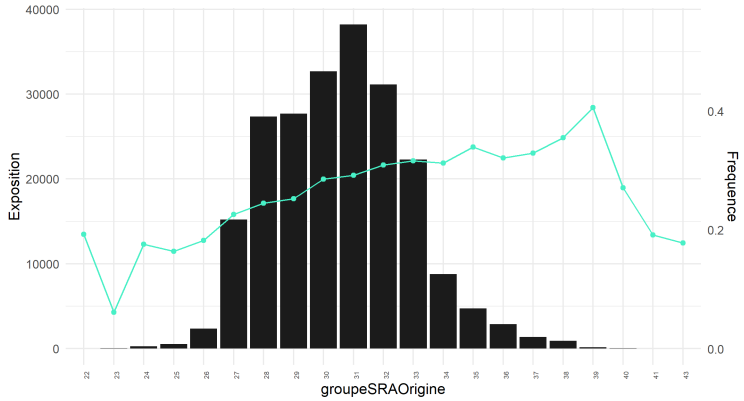


FIGURE B.1 – Fréquence en fonction du groupe SRA

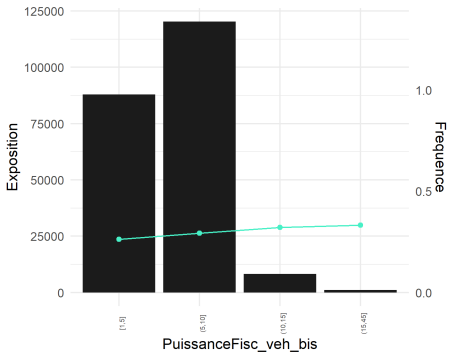


FIGURE B.2 – Fréquence en fonction de la puissance fiscale

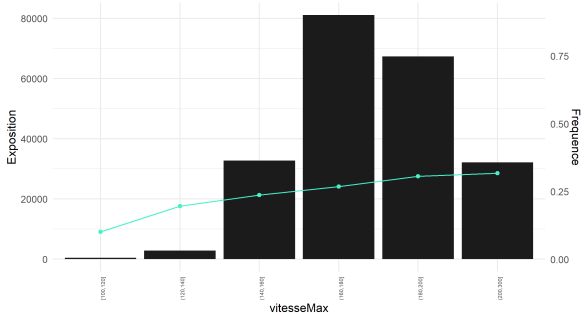


FIGURE B.3 – Fréquence en fonction de la vitesse maximale

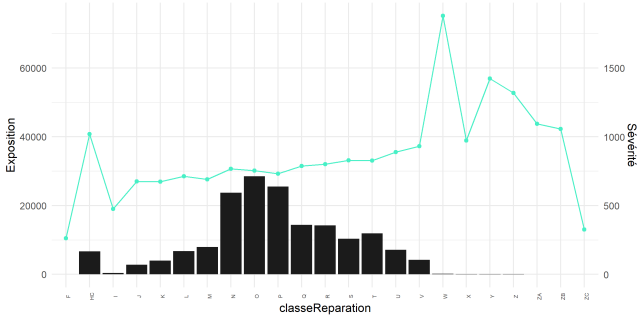


FIGURE B.4 – Sévérité en fonction de la classe de réparation

Annexe C

Compléments sur les analyses de prédiction des modèles contenant le véhiculier

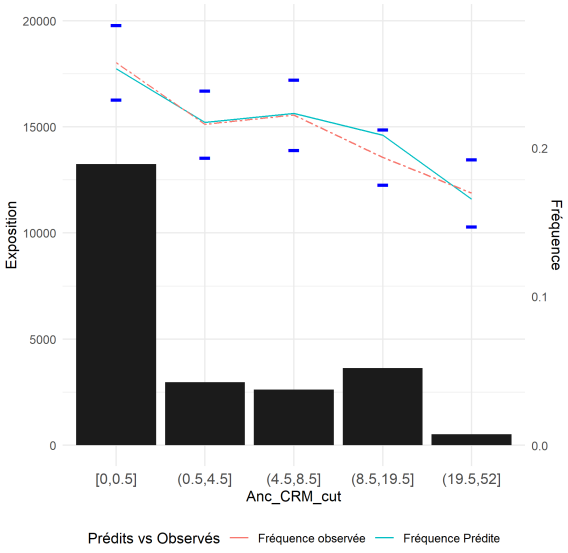


FIGURE C.1 – Fréquence prédite en fonction de l'ancienneté du CRM

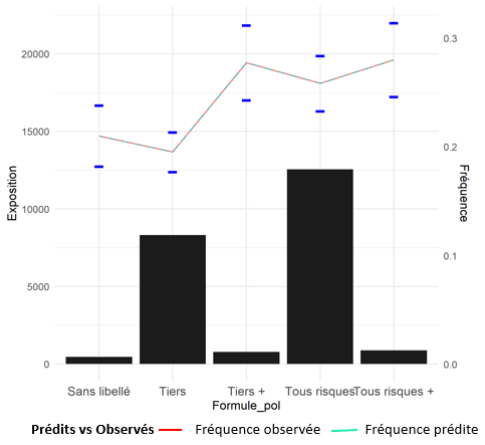


FIGURE C.2 – Fréquence prédite en fonction de la formule de la police

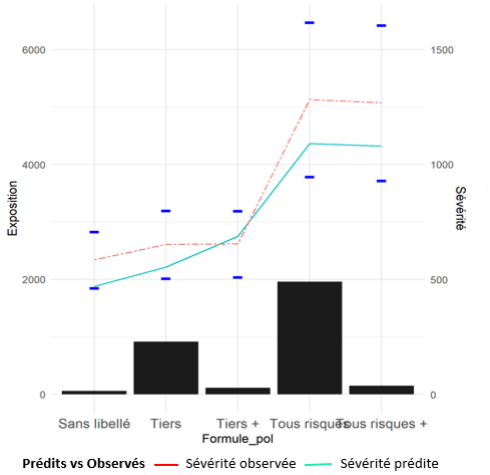


FIGURE C.3 – Sévérité prédite en fonction de la formule de la police

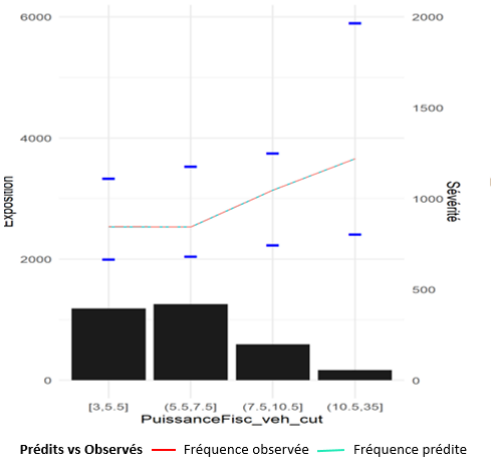


FIGURE C.4 – Sévérité prédite en fonction de la Puissance fiscale

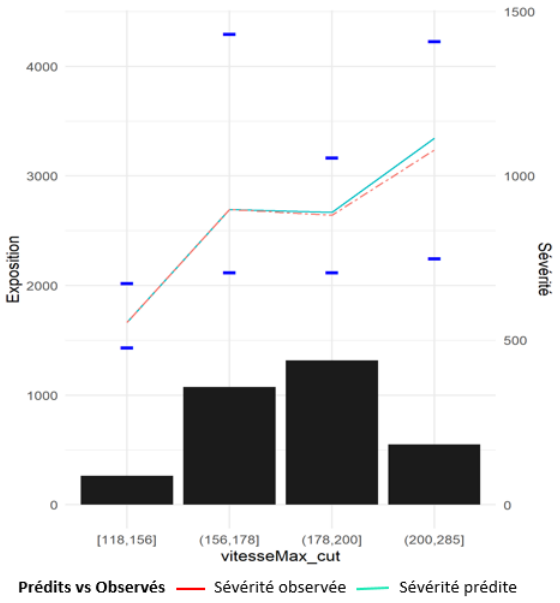


FIGURE C.5 – Sévérité prédite en fonction de la vitesse maximale

Comparaison des prédictions des modèles B et C

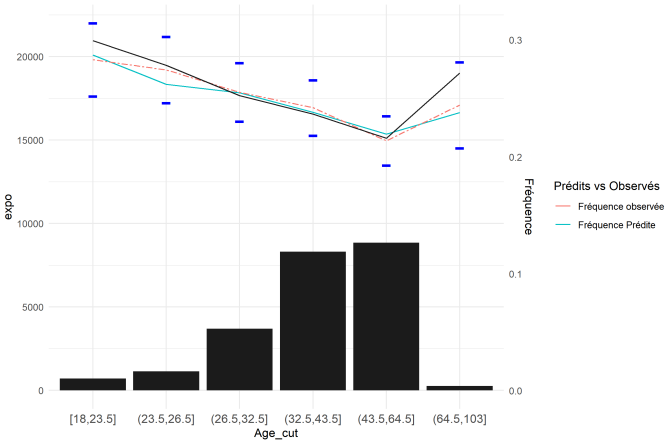


FIGURE C.6 – Fréquence prédite en fonction de l'âge de l'assuré

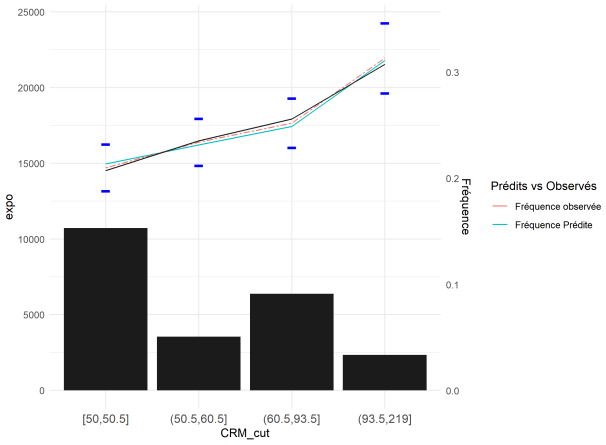


FIGURE C.7 – Fréquence prédite en fonction du CRM

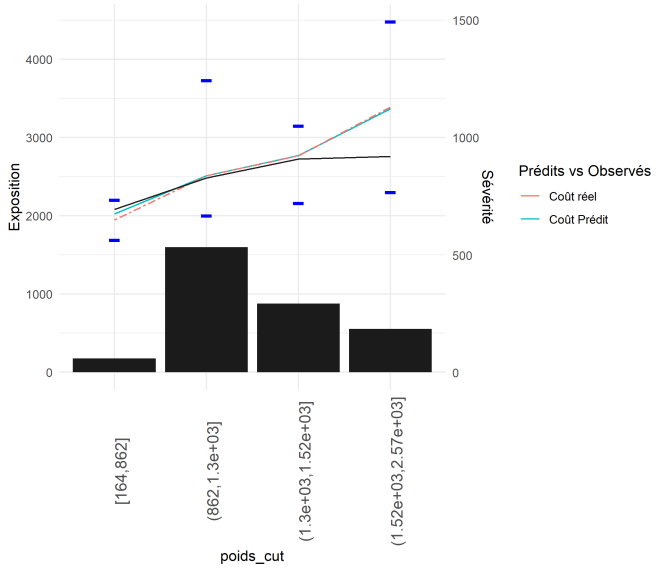


FIGURE C.8 – Sévérité prédite en fonction du poids

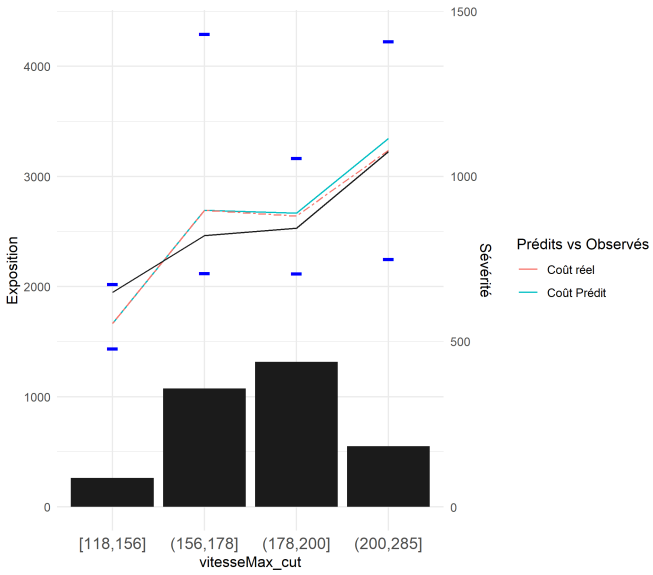


FIGURE C.9 – Sévérité prédite en fonction de la vitesse maximale

Annexe D

Zoom sur la matrice de corrélation des variables véhiculières