Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 18 Septembre 2024

Par : Thibaut DOMAVO
Titre : Construction et exploitation d'un véhiculier dans un modèle de Machine Learning interprétable pour l'optimisation tarifaire automobile

Confidentialité : Non

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

Membres présents du jury de l'Institut des Actuaires :
Samuel STOCKSIEKER
Romain LAILY
Yufei LUO
Signature :

Membres présents du jury de l'EURIA :
Jean-Marc DERRIEN

Entreprise :
GENERALI FRANCE
Signature :

Directeur de mémoire en entreprise :
Christophe MOUREN
Signature :

Invité :

Signature :

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

Signature du responsable entreprise:

Signature du candidat:

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

Les rapports annuels 2020-2022 de France Assureurs signalent une tendance croissante des prestations en assurance automobile, incitant les assureurs à ajuster leurs primes pour mieux aligner celles-ci avec le profil de risque des assurés. Cette situation nécessite une réévaluation et une amélioration des méthodes de tarification actuelles afin d'identifier plus précisément les risques encourus par les assurés. La littérature sur les mémoires d'actuariat souligne que l'intégration d'une variable telle que le véhiculier, représentant le risque associé aux véhicules, dans les modèles de tarification améliore significativement leurs capacités prédictives. Par ailleurs, l'essor des modèles de Machine Learning présente une opportunité stratégique majeure pour les assureurs, particulièrement en raison des enjeux critiques liés à la gestion des risques. Ces modèles avancés peuvent optimiser la prédiction des indicateurs de risques clés en assurance automobile, tels que la fréquence et le coût moyen des sinistres. Cependant, l'utilisation de ces modèles sophistiqués pose un défi important en termes de transparence, souvent qualifiés de "boîte noire", ce qui complique l'interprétation des résultats et la prise de décision éclairée. L'objectif principal de ce mémoire est d'optimiser la tarification en assurance automobile en développant et intégrant un véhiculier afin d'améliorer les modèles prédictifs existants. Ensuite, cette étude se concentre sur la gestion des risques associés aux profils des assurés en utilisant un modèle de Machine Learning interprétable, le Modèle Additif Neuronal. Ce modèle est spécialement conçu pour concilier performance et transparence, en acceptant une légère réduction de précision pour favoriser une meilleure interprétabilité. Cette approche vise à faciliter les décisions relatives aux politiques tarifaires, garantissant ainsi une meilleure adéquation entre les primes proposées et les risques réels. En conclusion, ce mémoire propose des améliorations méthodologiques pour la tarification en assurance automobile, en mettant l'accent sur l'intégration de données détaillées et l'application d'un modèle de Machine Learning interprétable, dans le but de renforcer à la fois la précision des prédictions et la transparence des décisions tarifaires.

Mots clefs : *Assurance automobile, Tarification, Machine Learning, Boîte Noire, Modèle Additif Neuronal, Interprétabilité*

Abstract

France Assureurs' 2020-2022 annual reports point to a growing trend in automobile insurance benefits, prompting insurers to adjust their premiums to better align them with the risk profile of policyholders. This situation requires a re-evaluation and improvement of current pricing methods in order to more accurately identify the risks incurred by policyholders. The literature on Actuariat's dissertations emphasizes that the integration of a variable such as vehicle classification, representing the risk associated with vehicles, into pricing models significantly improves their predictive capabilities. Moreover, the rise of machine learning models presents a major strategic opportunity for insurers, particularly because of the critical issues related to risk management. These advanced models can optimize the prediction of key risk indicators in automobile insurance, such as the frequency and average cost of claims. However, the use of these sophisticated models poses a significant challenge in terms of transparency, often referred to as the "black box", which complicates the interpretation of results and informed decision-making. The main objective of this thesis is to optimize car insurance pricing by developing and integrating a vehicle classification to improve existing predictive models. Then, this study focuses on the management of risks associated with the profiles of insured persons using an interpretable Machine Learning model, the Neural Additive Model. This model is specially designed to reconcile performance and transparency, accepting a slight reduction in precision to promote better interpretability. This approach aims to facilitate pricing policy decisions, thus ensuring a better match between the proposed premiums and actual risks. In conclusion, this thesis proposes methodological improvements for car insurance pricing, focusing on the integration of detailed data and the application of an interpretable Machine Learning model, with the aim of strengthening both the accuracy of predictions and the transparency of tariff decisions.

Keywords : *Automobile Insurance, Pricing, Machine Learning, Black Box, Neural Additive Model, Interpretability*

Remerciements

Au terme de ce travail, je tiens à exprimer mes sincères remerciements à Christophe MOUREN, manager de l'équipe Assurance Dommages de la Direction des Partenariats de Generali.

Je souhaite également remercier mes tuteurs d'entreprise Guillaume DURAND et Thi To Vong NGUYEN pour leur accompagnement, leurs conseils, l'aide apportée à la réalisation de cette étude, ainsi que pour leurs relectures qui ont permis d'obtenir ce document. Je remercie également toutes les personnes de la direction qui m'ont aidé dans la réalisation de ces travaux.

J'aimerais aussi remercier mes tuteurs académiques Anaëlle LE BERRE et Pierre AILLIOT qui, malgré leurs multiples occupations, m'ont accompagné et conseillé dans la rédaction de ce mémoire.

Je remercie également le corps professoral et administratif de l'EURIA pour leur accompagnement pendant mes trois dernières années d'études.

Note de Synthèse

Contexte

Face à une augmentation continue des prestations automobiles, les assureurs doivent sans cesse réévaluer leurs modèles afin de mieux aligner les primes avec les risques encourus par les assurés. Cette exigence de précision dans la tarification des produits d'assurance est un défi constant pour les actuaires, qui doivent concilier la performance prédictive des modèles avec leur capacité à être interprétés de manière transparente. Une solution à cette problématique étudiée dans ce mémoire est l'intégration d'une variable spécifique, le véhiculier, qui agrège les caractéristiques du véhicule pour mieux capturer le risque inhérent. L'avènement des méthodes de Machine Learning, telles que l'utilisation des réseaux neuronaux et des modèles XGBoost et Random Forest, a permis de repousser les limites de la prédiction. Toutefois, ces modèles, qualifiés de "boîtes noires", souffrent d'un manque d'interprétabilité, ce qui limite leur adoption dans des secteurs comme l'assurance, où la transparence des décisions est essentielle. C'est dans ce contexte qu'émergent les modèles hybrides combinant la puissance des modèles de Machine Learning avec l'interprétabilité des modèles additifs généralisés. La construction du Modèle Additif Neuronal (IGANN) consistant à remplacer les fonctions splines par des réseaux de neurones constitue la deuxième solution complémentaire à la problématique de cette étude.

Les données utilisées dans ce mémoire sont celles de la direction des partenariats de Generali. Les modélisations effectuées sont la modélisation de la fréquence de sinistres et du coût moyen et concernent essentiellement la garantie Dommages Tous Accidents (DTA).

Modèle Linéaire Généralisé

Le modèle GLM a été développé pour contourner trois hypothèses nécessaires à l'utilisation du modèle de régression linéaire.

Le modèle GLM peut être exprimé sous la forme suivante :

$$g(\mathbb{E}[Y_i]) = \theta_0 + \theta_1 x_i^1 + \theta_2 x_i^2 + \ldots + \theta_p x_i^p \tag{1}$$

La fonction g, appelée fonction de lien, permet d'introduire une non-linéarité entre les variables explicatives et la variable à expliquer. La variable aléatoire Y ne suit plus nécessairement une loi normale, mais elle doit appartenir à un groupe spécifique de lois de probabilité appelé famille exponentielle.

Pour le modèle GLM, la **loi de Poisson** a été sélectionnée pour modéliser la fréquence de sinistres et la **loi Gamma** a été retenue pour modéliser le coût moyen.
Dans le contexte de l'assurance non-vie, les primes tarifaires sont calculées en multipliant plusieurs facteurs, ce qui correspond à un modèle Log-GLM. Ainsi, la fonction de **lien logarithmique** sera utilisée quel que soit le modèle GLM afin de disposer d'une structure de tarification multiplicative.

Construction d'un véhiculier et impact sur la tarification

La figure suivante illustre le résumé des différentes étapes de la construction du véhiculier.

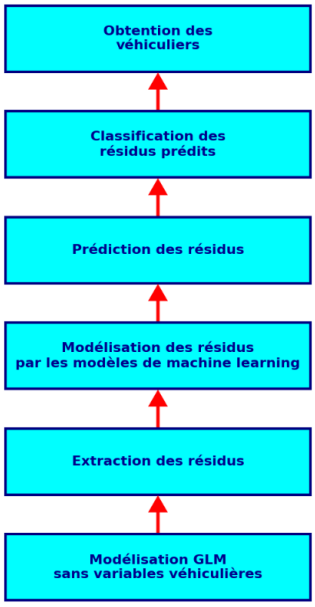

Figure 1: Étapes de construction du véhiculier

Comme indiqué sur la figure 1, pour la construction du véhiculier, nous réalisons premièrement les modèles GLM des deux indicateurs de risques. Ces premiers modèles étant obtenus avec toutes les variables non véhiculières disponibles, il est ensuite nécessaire de procéder à une sélection de variables pour ces modèles. À cet effet, différentes méthodes peuvent être utilisées. **La méthode Forward** a été employée en utilisant le critère AIC pour le modèle de fréquence et la métrique RMSE pour le coût moyen des sinistres. Cette sélection a révélé que la variable `CRM` a été choisie en première

position pour les deux modèles.

Nous procédons ensuite à l'extraction des résidus qui constitue l'étape suivante de la construction du véhiculier. Les résidus d'un modèle de prédiction représentent la part non expliquée par les variables explicatives du modèle. Dans notre cas, les résidus peuvent être expliqués par l'absence de l'utilisation des variables véhiculières dans les modèles. Par défaut, les résidus sont définis comme la différence entre les valeurs observées y_i et les valeurs prédites $\hat{y}_i$, et on les appelle **résidus additifs** mais pour la construction du véhiculier, nous avons plutôt choisi d'utiliser les **résidus de Pearson**, qui sont relativement simples et possèdent la propriété d'homoscédasticité.

Les résidus extraits, l'étape suivante consiste à les modéliser à l'aide de modèles de machine learning. Les résidus de Pearson issus des modèles GLM de fréquence et de coût moyen sont modélisés dans notre étude à l'aide de modèles de machine learning, car ces derniers ne font aucune hypothèse sur la loi de la variable à expliquer. Les algorithmes de machine learning repèrent des motifs récurrents dans l'ensemble de données sur lequel le modèle est appliqué. Lorsque l'objectif principal est de faire des prédictions précises, les modèles de machine learning deviennent alors essentiels. Contrairement aux modèles classiques comme le GLM, les modèles d'apprentissage statistique capturent non seulement les relations non linéaires mais aussi les interactions entre variables, intégrant une complexité dans le modèle.

Les modèles sélectionnés pour la modélisation des résidus sont XGBoost (modèles ensemblistes adaptatifs) et Random Forest (modèles ensemblistes parallèles).
La variable à expliquer dans les modèles de machine learning utilisés à cette étape est la distribution des résidus de Pearson des modèles GLM ajustés sans les variables véhiculières. Ces résidus sont donc attribués à l'absence de variables liées aux véhicules. Ainsi, les variables explicatives utilisées sont les variables véhiculières.
Les modèles XGBoost et Random Forest étant des modèles hyperparamétriques, il est nécessaire de rechercher des hyperparamètres optimaux pour maximiser leurs performances : l'optimisation des hyperparamètres est cruciale. Dans cette étude, l'optimisation des hyperparamètres a été réalisée avec la méthode Grid Search.
Les modèles sont ajustés avec les hyperparamètres obtenus par la méthode Grid Search dont le récapitulatif est présenté dans le tableau suivant :

Métriques	**MAE**		**RMSE**	
Modèles	**Random Forest**	**XGBoost**	**Random Forest**	**XGBoost**
FREQ	0,2637	0,0199	1,0526	0,3054
CM	0,6643	0,6617	0,9304	0,9293

Table 1: Comparaison des modèles

Sur la base du tableau 1, le modèle XGBoost s'affiche comme le meilleur choix pour la

modélisation des résidus du modèle GLM de fréquence de sinistres et de coût moyen. Ainsi, XGBoost est le modèle retenu en raison de ses meilleures performances globales.

L'étape de modélisation des résidus permet d'obtenir les résidus prédits. La phase finale de la construction du véhiculier consiste à classifier ces résidus prédits.

Le but étant de construire des classes de risques homogènes associées aux véhicules, la méthode de classification utilisée est la classification non supervisée. Celle-ci se divise en deux sous-catégories : la classification non supervisée hiérarchique et la classification non supervisée non hiérarchique. Parmi les méthodes de classification non supervisée hiérarchique, on peut citer la **Classification Ascendante Hiérarchique (CAH)**, tandis que la méthode **K-Means** est un exemple très connu de classification non supervisée non hiérarchique.
La construction des véhiculiers dans cette étude se base sur la combinaison des méthodes K-Means et CAH appliquées sur les différents résidus prédits permettant d'obtenir les nombres de classes définis. L'application de la méthode CAH et l'analyse des dendrogrammes fournis permettent de retenir approximativement 10 classes pour le modèle de fréquence de sinistres et 15 classes pour le modèle de coût moyen. La méthode du coude appliquée sur la méthode K-Means permet de choisir définitivement **14 classes pour le véhiculier du modèle de coût moyen et 11 classes pour le véhiculier du modèle de fréquence de sinistres.**

Les classifications obtenues représentent les véhiculiers. Ils sont ensuite intégrés dans la base de données pour évaluer leur impact sur la modélisation GLM de la fréquence de sinistres et du coût moyen. L'objectif initial était de comparer le véhiculier construit avec celui proposé par la SRA, à travers les variables `Classe SRA` et `Groupe SRA`. La comparaison des modèles est résumée dans le tableau suivant :

Métriques	**RMSE**				**MAE**			
	CM	**VAR**	**FREQ**	**VAR**	**CM**	**VAR**	**FREQ**	**VAR**
GLM SVV	2292,19	100%	0,1403	100%	1588,06	100%	0,0380	100%
GLM SRA	2252,97	**-1,74%**	0,1403	100%	1571,79	**-1,03%**	0,0380	100%
GLM VEH	1756,80	**-30,47%**	0,1395	**-0,57%**	1249,45	**-27,10%**	0,0377	**-0,79%**

Table 2: Comparaison des modèles

Légende : **VAR** : Variation, **GLM SVV** : GLM Sans Variables Véhiculières, **GLM SRA** : GLM avec le véhiculier de la SRA, **GLM VEH** : GLM avec le véhiculier construit.

Le modèle **GLM VEH** affiche des performances nettement supérieures par rapport au modèle GLM SVV. Le RMSE pour le coût moyen diminue de 30,47% et le MAE diminue de 27,10%. Pour la fréquence de sinistres, le RMSE diminue de 0,57% et le

MAE diminue de 0,79%. Ces réductions indiquent que le véhiculier construit pour le coût moyen capture efficacement les risques associés aux véhicules, améliorant ainsi le modèle sans variables véhiculières. Cela s'explique par le fait que le ***véhiculier construit*** prend en compte des variables supplémentaires contenant des informations sur les caractéristiques des véhicules explicatives du risque qui ne sont pas présentes dans le modèle GLM SVV.

Les améliorations significatives dans les métriques de performance pour le coût moyen indiquent que le véhiculier capture mieux les risques spécifiques aux véhicules. Ces résultats mettent en évidence l'importance de la construction du véhiculier, une variable spécifique adaptée aux caractéristiques des véhicules pour améliorer la précision de la tarification dans les contrats d'assurance automobile. La création de ce véhiculier sur mesure est donc une stratégie efficace pour mieux modéliser les risques associés aux véhicules, offrant ainsi une tarification plus précise et potentiellement plus juste pour les assureurs.

La caractérisation du véhiculier montre que les classes des véhiculiers dont les coefficients sont les plus élevés dans les résultats des modèles GLM sont les classes les plus risquées. Pour illustrer cela, les graphiques ci-dessous comparent les valeurs prédites avec celles observées, en fonction des différentes classes de véhiculier, renommées selon le risque associé.

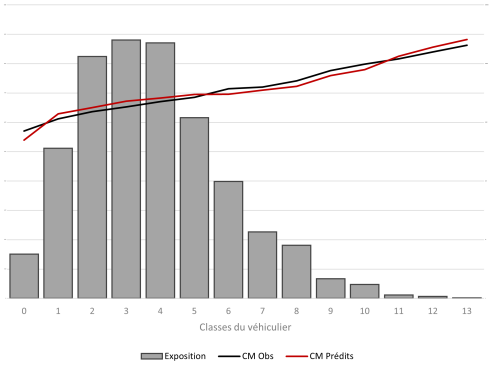

Figure 2: Coût moyen observé vs Coût moyen prédit

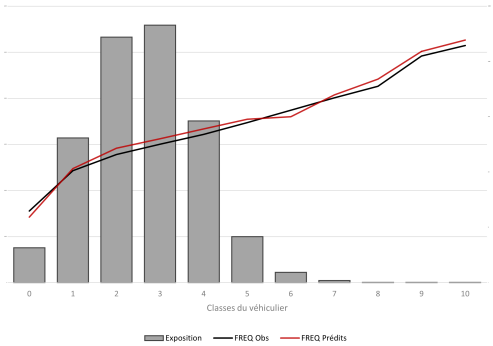

Figure 3: Fréquence de sinistres observée vs Fréquence de sinistres prédite

L'analyse de ces figures montre que, plus la classe du véhiculier est élevée, plus la fréquence de sinistres et le coût moyen augmentent, avec une correspondance presque parfaite entre les deux courbes, illustrant ainsi la performance du véhiculier dans la modélisation.

Concept de Modèles Additifs Généralisés

Le modèle additif généralisé (GAM) est une extension du modèle linéaire généralisé (GLM). L'idée derrière le modèle GAM est de laisser les données s'exprimer en modélisant la contribution des variables explicatives à travers des fonctions de ces variables. Cela permet de capturer les relations non linéaires qui peuvent exister entre certaines covariables et la variable réponse. Une formulation mathématique simple du modèle GAM est donnée par :

$$g(\mu_i) = f_1(x_i^1) + f_2(x_i^2) + \ldots + f_p(x_i^p) + \epsilon_i \tag{2}$$

Le modèle ainsi présenté, avec f_j représentant les fonctions des variables explicatives, correspond à un modèle non paramétrique nécessitant l'estimation des fonctions utilisées. Les fonctions splines ayant de bonnes propriétés dans ce contexte sont utilisées dans les modèles GAM dont la forme générale est la suivante :

$$f(x) = \sum_{k=1}^{q} b_k(x)\theta_k$$

La théorie mathématique sur le modèle GAM montre alors que le modèle GAM est un modèle GLM surparamétré dont la forme matricielle est la suivante :

$$g(\mu_i) = \mathbf{X}_i\theta + \epsilon_i \tag{3}$$

Le paramètre θ est ensuite déterminé par maximum de vraisemblance. Cependant, la log-vraisemblance utilisée dans ce cas est une log-vraisemblance pénalisée dans l'objectif de contrôler le niveau de lissage des courbes f_j et le nombre de fonctions de base. La maximisation de cette log-vraisemblance pénalisée peut être résolue par une méthode itérative proposée par Wood (2006) appelée méthode P-IRLS (Penalized Iteratively Re-weighted Least Squares), qui est une variante de la méthode IRLS (Iteratively Re-weighted Least Squares) de Nelder et Wedderburn (1972).

Modèle Additif Neuronal

Les modèles GAM offrent deux avantages majeurs : ils permettent de prendre en compte les relations non linéaires entre la variable à expliquer et les covariables à travers les fonctions splines des covariables, et ils offrent la possibilité de visualiser et donc d'interpréter ces fonctions splines. Il est crucial de comprendre ces relations, raison pour laquelle de nombreuses études se concentrent sur le développement de méthodes d'interprétation. Ces recherches ont également ouvert de nouvelles perspectives pour l'interprétation des modèles de machine learning, notamment en les combinant avec les modèles GAM. L'idée derrière cette combinaison est de remplacer les fonctions splines des modèles GAM par des modèles de machine learning simples, tels que les arbres de décision et les réseaux de neurones. Nous nous intéresserons particulièrement à un modèle combinant le modèle GAM et les réseaux de neurones, que nous appellerons

dans la suite **Modèle Additif Neuronal** (NAM).

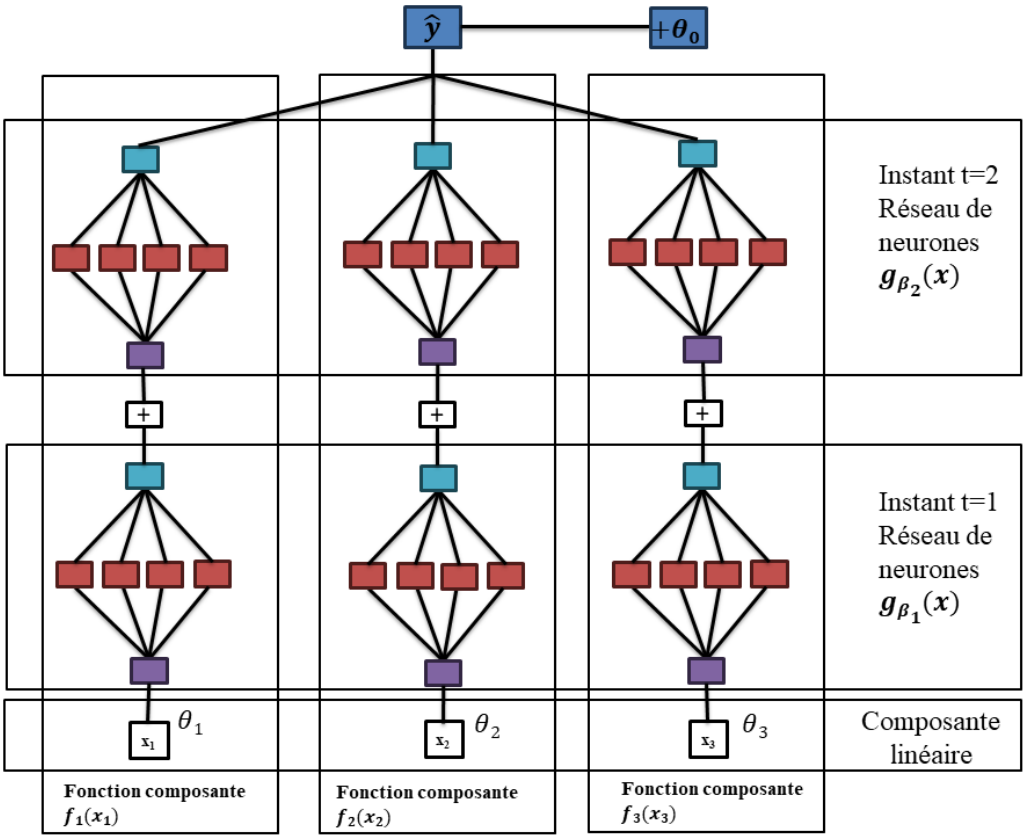

Figure 4: Illustration du modèle IGANN pour deux instants d'agrégation

Les modèles NAM sont encore en développement, avec diverses versions proposées par différents chercheurs. Parmi celles-ci, on distingue le modèle NAM développé par l'équipe de "Google Research Team" (Agarwal et al., 2020), le modèle CANN (Combined Actuarial Neural Network) de Schelldorfer et Wüthrich (2019), et le modèle **IGANN** (Interpretable Generalized Additive Neural Networks) proposé en 2023 par Kraus et al. Nous utiliserons ici le modèle IGANN, qui est directement implémenté en `Python` via le package `igann`.

La forme mathématique du modèle IGANN est exprimée par l'équation suivante :

$$\hat{y} = \theta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d)$$

Dans le modèle IGANN, les réseaux de neurones sont agrégés, chaque réseau étant constitué d'une couche cachée. La particularité de ce réseau de neurones réside dans le fait que chaque élément du vecteur d'entrée est modélisé par un sous-réseau de

neurones à une couche cachée. L'objectif est de capturer l'effet de chaque covariable sur la variable à prédire. Pour obtenir cet effet additif, la sortie du modèle est calculée comme la somme des sorties de chacun de ces sous-réseaux de neurones. Ces réseaux de neurones s'agrègent après estimation de la composante linéaire du modèle. En effet, le modèle part d'une composante linéaire et agrège ensuite les réseaux de neurones à chaque itération jusqu'à une condition d'arrêt (comme la stabilité d'une métrique) ou jusqu'à l'atteinte du nombre d'itérations T définies dans l'algorithme du modèle IGANN.

Le modèle IGANN propose trois types d'interprétation. La première est une interprétation globale basée sur l'analyse des courbes des fonctions composantes. Ces courbes permettent de comprendre l'influence des covariables sur la variable cible. La deuxième interprétation repose sur une approche locale de l'interprétabilité. **Ces courbes ne se limitent pas à représenter des relations, elles offrent une description exacte de la façon dont le modèle IGANN calcule une prédiction**. Pour une observation donnée, il suffit de récupérer les valeurs correspondantes sur l'axe des ordonnées de chaque courbe associée aux variables de l'observation, puis de les additionner, en incluant l'intercept θ_0 du modèle, pour obtenir la prédiction. La troisième interprétation offerte par le modèle est une méthode d'évaluation globale de l'importance des variables ; le modèle IGANN a l'avantage de pouvoir afficher directement l'importance relative de chaque variable sur les courbes des fonctions composantes.

Les graphiques suivants illustrent les courbes des fonctions composantes associées aux deux premières variables les plus importantes du modèle IGANN ajusté sur le coût moyen.

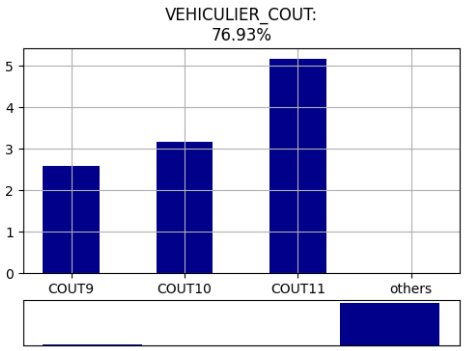

Figure 5: Véhiculier

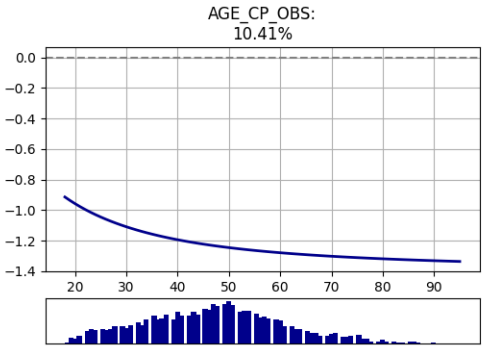

Figure 6: Âge du conducteur

Le véhiculier que nous avons construit joue un rôle crucial dans tous les modèles, en particulier dans le modèle IGANN. Le véhiculier s'affiche comme une variable déterminante pour la modélisation du coût moyen, avec une importance évaluée à 76,93% dans le modèle, suivie par l'âge du conducteur (10,41%). Les relations observées entre ces covariables et la variable dépendante confirment des tendances attendues. Le graphique sur le véhiculier montre une relation croissante avec l'augmentation du risque

de la classe en exerçant un impact positif sur le coût moyen.

L'âge du conducteur, quant à lui, a un impact négatif sur la prédiction du coût moyen. Comme la prédiction résulte de l'addition des contributions individuelles, le coût moyen diminue à mesure que l'âge du conducteur augmente.

Comparaison des résultats

Le tableau suivant présente les performances prédictives des modèles.

Modèles	**Coût moyen**		**Fréquence de sinistres**	
	MAE	**RMSE**	**MAE**	**RMSE**
GLM	1244,5	1752,47	0,038	0,1403
GAM	1237,35	1745,68	0,037	0,1395
IGANN	**1188,84**	**1711,87**	**0,036**	**0,1395**
XGBoost	**1177,79**	**1681,60**	**0,036**	**0,1285**

Table 3: Comparaison des performances des modèles de coût moyen et de fréquence de sinistres

Le tableau ci-dessus compare les performances de quatre modèles de prédiction : GLM, GAM, IGANN, et XGBoost, en se basant sur deux indicateurs : le MAE (erreur absolue moyenne) et le RMSE (racine carrée de l'erreur quadratique moyenne).

Sur l'ensemble de test, le modèle XGBoost se distingue comme le modèle le plus performant, avec les erreurs les plus faibles, ce qui montre sa capacité à prédire avec précision les valeurs cibles. Le modèle IGANN suit de près, offrant également des résultats satisfaisants, bien que légèrement en retrait par rapport à XGBoost. En revanche, les modèles GAM et GLM affichent des erreurs plus élevées, ce qui indique une moindre capacité prédictive que les modèles XGBoost et IGANN.

Dans le même objectif de comparaison des modèles, nous nous intéressons à la comparaison des courbes des valeurs prédites des modèles GLM, GAM et IGANN et la courbe des valeurs observées en fonction des classes d'âge.

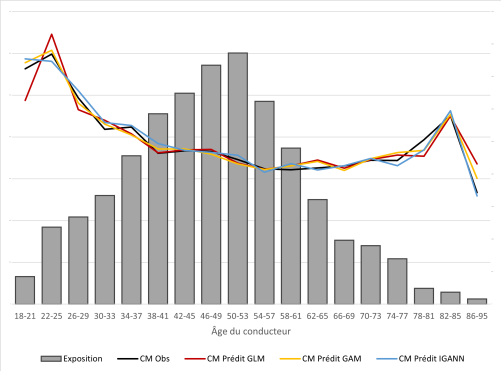

Figure 7: Coût moyen

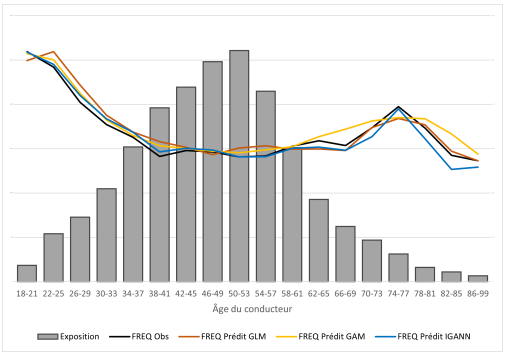

Figure 8: Fréquence de sinistres

La figure 7 montre que les trois modèles comparés affichent dans l'ensemble un bon ajustement entre les valeurs observées et prédites, avec toutefois quelques remarques à retenir. Le modèle GLM peine à généraliser les prédictions pour les conducteurs âgés de 18 à 33 ans. Il sous-estime les prédictions pour la classe d'âge 18-21 ans et surestime celles de la classe 22-25 ans. L'analyse montre que les prédictions sont globalement meilleures pour les classes d'âge intermédiaires, avec une légère amélioration supplémentaire du modèle IGANN par rapport aux deux autres modèles pour ces classes. Une tendance similaire est observée pour les prédictions concernant les classes d'âge plus élevées.

Le graphique 8, qui montre la fréquence de sinistres, confirme la légère amélioration de l'ajustement du modèle IGANN par rapport aux deux autres modèles. En effet, la fréquence de sinistres pour les jeunes conducteurs est surestimée par l'ensemble des modèles, mais de manière moins prononcée par le modèle IGANN. De même, pour les classes d'âge intermédiaires, la courbe du modèle IGANN est plus proche de celle des fréquences observées que les autres modèles. La principale différence se trouve au niveau des deux dernières classes d'âge, où le modèle GLM généralise mieux. Le modèle GAM, quant à lui, peine à généraliser la fréquence de sinistres pour les classes d'âge les plus élevées en surestimant la fréquence de sinistres pour ces classes.

Conclusion

Ce mémoire a exploré de manière approfondie l'intégration du **véhiculier** et des modèles de Machine Learning interprétables, en particulier le Modèle Additif Neuronal (**IGANN**), dans le cadre de la tarification de la garantie **DTA** en assurance automobile. L'étude a montré que l'ajout du véhiculier construit permet de mieux capturer les risques inhérents aux caractéristiques des véhicules. Ce véhiculier a démontré son efficacité en réduisant les erreurs de prédiction par rapport au modèle GLM sans cette variable, et même par rapport à l'utilisation du véhiculier standard fourni par la SRA, constitué du groupe SRA et de la classe SRA.

En appliquant le modèle IGANN à la garantie DTA, il a été démontré que le modèle IGANN permet non seulement de réduire les erreurs de prédiction, mais aussi de conserver trois types d'interprétabilité (importance des variables, prédiction locale et relation entre les covariables et la variable dépendante) pour les actuaires, rendant les décisions tarifaires plus transparentes. La comparaison du modèle IGANN avec les autres modèles en termes de capacité prédictive a montré que le modèle IGANN se situe entre le modèle XGBoost avec les meilleures performances et les modèles classiques que sont les modèles GAM et GLM. Le modèle IGANN est ainsi un bon modèle intermédiaire alliant prédiction et interprétabilité.
Il est essentiel de rappeler que l'objectif n'est pas de remplacer les modèles GLM, qui demeurent des outils fondamentaux en actuariat et ne sont pas près de disparaître. Ces modèles, bien qu'ils présentent certaines limites, restent indispensables pour leur simplicité et leur interprétabilité. Ce mémoire propose plutôt une ouverture vers d'autres approches de modélisation, telles que le modèle IGANN, qui peuvent être explorées en complément des GLM pour répondre aux défis actuels et futurs du secteur.

Limites de l'étude et pistes d'amélioration

Les principales limitations de cette étude peuvent être résumées comme suit :

- L'utilisation du Zonier actuel dans la tarification, alors qu'il aurait été possible de développer un Zonier personnalisé basé sur le portefeuille, ce qui aurait enrichi la tarification.
- L'absence de prise en compte du lissage des résidus lors de la construction du véhiculier.
- Une méthode spécifique pour introduire les interactions dans le modèle GLM aurait pu être privilégiée par rapport à celle utilisée dans cette étude.
- Les modèles incluant le véhiculier construit n'ont pas été comparés à ceux utilisant l'ensemble des variables véhiculières en tant que risques associés au véhicule.
- Le modèle IGANN, en raison de sa conception, n'inclut pas les interactions entre les variables et se limite à une seule couche cachée dans la structure du réseau de neurones ; des études supplémentaires sont nécessaires pour intégrer ces éléments.
- Le modèle IGANN aurait dû être comparé à un modèle combinant les arbres de décision avec le modèle GAM.
- Une étude complète aurait pu être menée en calculant les primes pures finales et en réalisant des analyses détaillées sur celles-ci.

Executive Summary

Context

Faced with a continuous increase in automotive claims, insurers must constantly reassess their models to better align premiums with the risks faced by policyholders. This demand for precision in insurance product pricing presents a constant challenge for actuaries, who must balance the predictive performance of models with their ability to be interpreted transparently. One solution to this problem, studied in this thesis, is the integration of a specific variable, the "vehicular," which aggregates vehicle characteristics to better capture inherent risk. The advent of Machine Learning methods, such as the use of neural networks and models like XGBoost and Random Forest, has pushed the boundaries of prediction. However, these models, often referred to as "black boxes," suffer from a lack of interpretability, limiting their adoption in sectors like insurance, where decision transparency is crucial. It is in this context that hybrid models emerge, combining the power of Machine Learning models with the interpretability of generalized additive models. The construction of the Interpretable Generalized Additive Neural Network (IGANN), which replaces spline functions with neural networks, constitutes the second complementary solution to the problem of this study.

The data used in this thesis come from the partnership division of Generali. The models developed include the modeling of claim frequency and average cost, focusing mainly on the "Dommages Tous Accidents" (DTA) coverage.

Generalized Linear Model

The GLM model was developed to bypass three assumptions required for the use of the linear regression model.

The GLM model can be expressed as follows:

$$g(\mathbb{E}[Y_i]) = \theta_0 + \theta_1 x_i^1 + \theta_2 x_i^2 + \ldots + \theta_p x_i^p \tag{4}$$

The function g, called the link function, introduces non-linearity between the explanatory variables and the response variable. The random variable Y no longer necessarily follows a normal distribution, but it must belong to a specific group of probability

distributions called the exponential family.

For the GLM model, the **Poisson distribution** was selected to model claim frequency, and the **Gamma distribution** was chosen to model the average cost.
In the context of non-life insurance, premiums are calculated by multiplying several factors, which corresponds to a Log-GLM model. Thus, the **logarithmic link function** will be used regardless of the GLM model to maintain a multiplicative pricing structure.

Construction of a Vehicular Model and Impact on Pricing

The following figure illustrates a summary of the different steps in constructing the vehicular model.

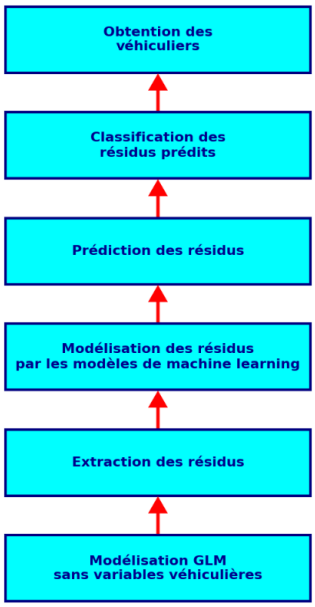

Figure 9: Steps in constructing the vehicular model

As shown in Figure 1, to construct the vehicular model, we first develop GLM models for the two risk indicators. These initial models, obtained using all available non-vehicular variables, require a variable selection process. Various methods can be used for this purpose. **The Forward Selection method** was employed, using the AIC criterion for the frequency model and the RMSE metric for the average cost of claims. This selection revealed that the variable `CRM` was chosen first for both models.

Next, we extract the residuals, which constitute the following step in constructing the vehicular model. Residuals from a prediction model represent the portion unexplained by the explanatory variables of the model. In our case, residuals may be explained by

the absence of vehicular variables in the models. By default, residuals are defined as the difference between observed values y_i and predicted values $\hat{y}_i$, and they are referred to as **additive residuals**, but for the construction of the vehicular model, we chose to use **Pearson residuals**, which are relatively simple and possess the property of homoscedasticity.

Once residuals are extracted, the next step is to model them using machine learning models. Residuals from the GLM models of frequency and average cost are modeled in our study using machine learning models, as these do not assume a particular distribution for the response variable. Machine learning algorithms identify recurring patterns in the dataset to which the model is applied. When the primary objective is to make accurate predictions, machine learning models become essential. Unlike traditional models like GLM, statistical learning models capture not only nonlinear relationships but also interactions between variables, adding complexity to the model.

The models selected for residual modeling are XGBoost (adaptive ensemble models) and Random Forest (parallel ensemble models).
The response variable in the machine learning models used at this stage is the distribution of residuals from the GLM models fitted without vehicular variables. These residuals are therefore attributed to the absence of vehicle-related variables. Consequently, the explanatory variables used are the vehicular variables.
Since XGBoost and Random Forest are hyperparametric models, it is necessary to search for optimal hyperparameters to maximize their performance: hyperparameter optimization is crucial. In this study, hyperparameter optimization was performed using the Grid Search method.
The models are fitted with the hyperparameters obtained through the Grid Search method, summarized in the following table:

Metrics	**MAE**		**RMSE**	
Models	**Random Forest**	**XGBoost**	**Random Forest**	**XGBoost**
FREQ	0.2637	0.0199	1.0526	0.3054
CM	0.6643	0.6617	0.9304	0.9293

Table 4: Comparison of models

Based on Table 1, the XGBoost model emerges as the best choice for modeling residuals from the GLM frequency and average cost models. Thus, XGBoost is the selected model due to its overall superior performance.

The residual modeling step allows for the prediction of residuals. The final phase of vehicular model construction involves classifying these predicted residuals.

The goal is to construct homogeneous risk classes associated with vehicles, and the

classification method used is unsupervised classification. This method is divided into two subcategories: hierarchical unsupervised classification and non-hierarchical unsupervised classification. Among hierarchical unsupervised classification methods, we can mention **Hierarchical Ascendant Classification (HAC)**, while the **K-Means** method is a well-known example of non-hierarchical unsupervised classification.
The construction of vehicular models in this study is based on the combination of K-Means and HAC methods applied to the various predicted residuals to obtain the defined number of classes. The application of the HAC method and analysis of the provided dendrograms allow for the retention of approximately 10 classes for the claim frequency model and 15 classes for the average cost model. The elbow method applied to the K-Means method ultimately selects **14 classes for the average cost model vehicular model and 11 classes for the claim frequency vehicular model.**

The obtained classifications represent the vehicular models. They are then integrated into the database to evaluate their impact on the GLM modeling of claim frequency and average cost. The initial objective was to compare the constructed vehicular model with the one proposed by SRA, through the variables `SRA Class` and `SRA Group`. The model comparison is summarized in the following table:

Metrics	**RMSE**				**MAE**			
	CM	**VAR**	**FREQ**	**VAR**	**CM**	**VAR**	**FREQ**	**VAR**
GLM SVV	2292.19	100%	0.1403	100%	1588.06	100%	0.0380	100%
GLM SRA	2252.97	**-1.74%**	0.1403	100%	1571.79	**-1.03%**	0.0380	100%
GLM VEH	1756.80	**-30.47%**	0.1395	**-0.57%**	1249.45	**-27.10%**	0.0377	**-0.79%**

Table 5: Comparison of models

Legend: **VAR**: Variation, **GLM SVV**: GLM Without Vehicular Variables, **GLM SRA**: GLM with SRA vehicular model, **GLM VEH**: GLM with constructed vehicular model.

The **GLM VEH** model shows significantly better performance compared to the GLM SVV model. The RMSE for average cost decreases by 30.47%, and the MAE decreases by 27.10%. For claim frequency, the RMSE decreases by 0.57% and the MAE decreases by 0.79%. These reductions indicate that the constructed vehicular model for average cost effectively captures the risks associated with vehicles, improving the model without vehicular variables. This could be explained by the fact that the ***constructed vehicular model*** takes into account additional variables containing information about vehicle characteristics that explain the risk, which are not present in the GLM SVV model.

The significant improvements in performance metrics for average cost indicate that the vehicular model better captures vehicle-specific risks. These results highlight the

importance of constructing the vehicular model, a specific variable tailored to the characteristics of vehicles, to improve pricing accuracy in auto insurance contracts. The ***constructed vehicular model*** is also more effective than the SRA vehicular model because the SRA classification is global, while the ***constructed vehicular model*** is based on the insurer's portfolio, providing a classifier better suited to the data, leading to better performance in pricing models. Creating this custom vehicular model is therefore an effective strategy to better model vehicle-related risks, offering more accurate and potentially fairer pricing for insurers.

The characterization of the vehicular model shows that the vehicular model classes with the highest coefficients in the GLM model results are the most risky classes. To illustrate this, the following graphs compare the predicted values with the observed values, based on different vehicular model classes, renamed according to the associated risk.

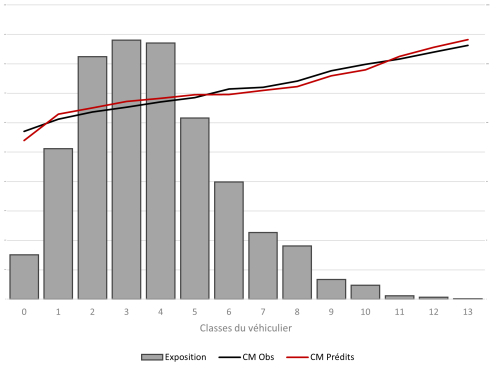

Figure 10: Observed vs Predicted Average Cost

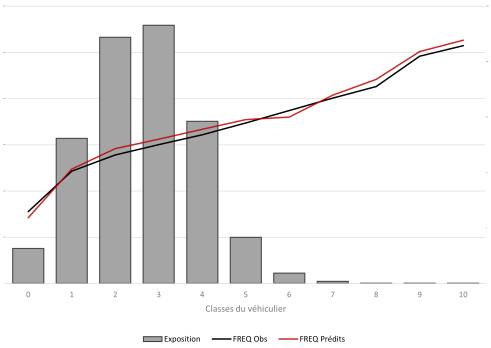

Figure 11: Observed vs Predicted Claim Frequency

The analysis of these figures shows that the higher the vehicular model class, the higher the claim frequency and average cost, with an almost perfect match between the two curves, thus illustrating the performance of the vehicular model in modeling.

Concept of Generalized Additive Models

The generalized additive model (GAM) is an extension of the generalized linear model (GLM). The idea behind the GAM is to allow the data to speak for themselves by modeling the contribution of explanatory variables through functions of these variables. This allows capturing nonlinear relationships that may exist between certain covariates and the response variable. A simple mathematical formulation of the GAM is given by:

$$g(\mu_i) = f_1(x_i^1) + f_2(x_i^2) + \ldots + f_p(x_i^p) + \epsilon_i \tag{5}$$

The model presented here, with f_j representing the functions of the explanatory variables, corresponds to a non-parametric model requiring the estimation of the functions used.

Spline functions, which have good properties in this context, are used in GAM models with the following general form:

$$f(x) = \sum_{k=1}^{q} b_k(x)\theta_k$$

The mathematical theory behind the GAM shows that the GAM is an over-parameterized GLM, whose matrix form is as follows:

$$g(\mu_i) = \mathbf{X}_i\theta + \epsilon_i \tag{6}$$

The parameter θ is then determined by maximum likelihood. However, the log-likelihood used in this case is a penalized log-likelihood to control the level of smoothing of the f_j curves and the number of basis functions. The maximization of this penalized log-likelihood can be solved using an iterative method proposed by Wood (2006) called the Penalized Iteratively Re-weighted Least Squares (P-IRLS) method, which is a variant of the Iteratively Re-weighted Least Squares (IRLS) method by Nelder and Wedderburn (1972).

Neural Additive Model

GAM models offer two major advantages: they allow for consideration of nonlinear relationships between the response variable and covariates through spline functions of the covariates, and they offer the possibility to visualize and thus interpret these spline functions. Understanding these relationships is crucial, which is why many studies focus on developing interpretation methods, as discussed in the previous section. These studies have also opened new perspectives for interpreting machine learning models, particularly by combining them with GAM models. The idea behind this combination is to replace the spline functions of GAM models with simple base models, such as decision trees and neural networks. We are particularly interested in a model combining the GAM and neural networks, which we will refer to as the **Neural Additive Model** (NAM).

NAM models are still under development, with various versions proposed by different researchers. Among these, we distinguish the NAM model developed by the "Google Research Team" (Agarwal et al., 2020), the CANN (Combined Actuarial Neural Network) model by Schelldorfer and Wüthrich (2019), and the **IGANN** (Interpretable Generalized Additive Neural Networks) model proposed in 2023 by Kraus et al. We will use the IGANN model here, which is directly implemented in Python via the `igann` package.

The mathematical form of the IGANN model is expressed by the following equation:

$$\hat{y} = \theta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d)$$

In the IGANN model, neural networks are aggregated, with each network consisting of a single hidden layer. The peculiarity of this neural network lies in the fact that each element of the input vector is modeled by a sub-neural network with a single hidden layer. The goal is to capture the effect of each covariate on the target variable. To obtain this additive effect, the model output is calculated as the sum of the outputs of each of these sub-neural networks. These neural networks are aggregated after estimating the linear component of the model. Indeed, the model starts with a linear component and then aggregates the neural networks at each iteration until a stopping condition is met (such as the stability of a metric) or until the number of iterations T defined in the IGANN model algorithm is reached.

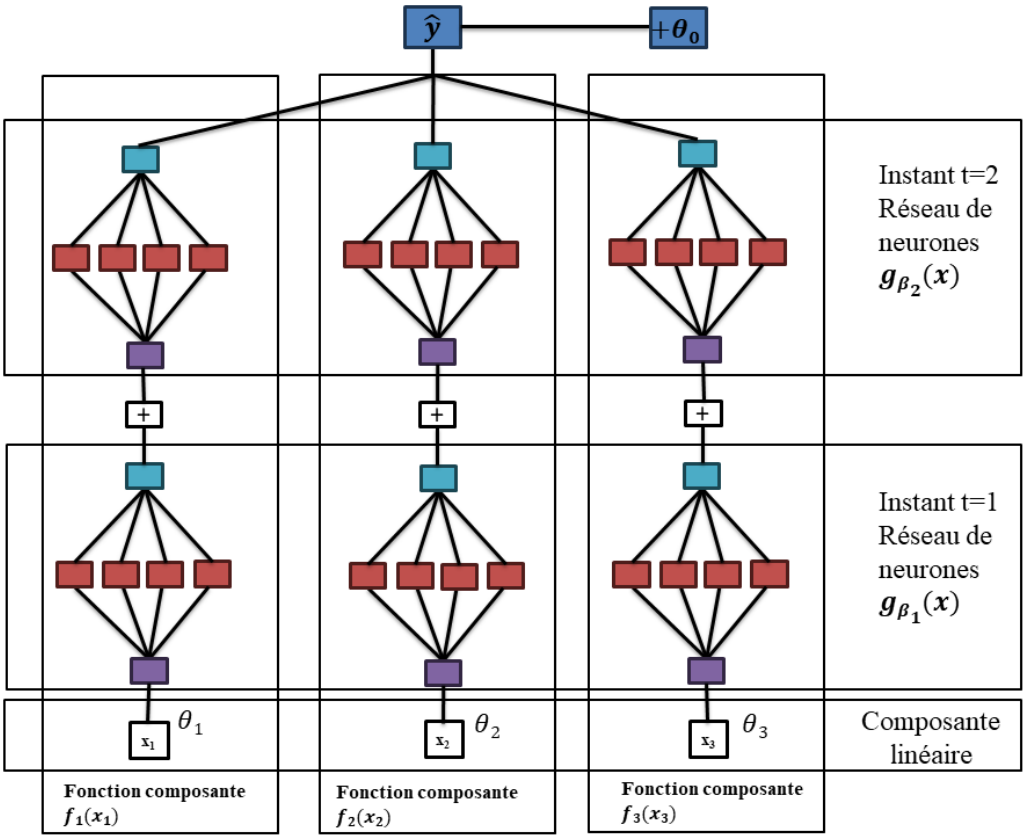

Figure 12: Illustration of the IGANN model for two aggregation instances

The IGANN model offers three types of interpretation. The first is a global interpretation based on the analysis of the curves of the component functions. These curves allow understanding the influence of covariates on the target variable. The second

interpretation is based on a local approach to interpretability. **These curves do not merely represent relationships; they provide an exact description of how the IGANN model calculates a prediction**. For a given observation, one simply needs to retrieve the corresponding values on the y-axis of each curve associated with the variables of the observation, then sum them, including the model's intercept θ_0, to obtain the prediction. The third interpretation offered by the model is a method of globally assessing variable importance; the IGANN model has the advantage of directly displaying the relative importance of each variable on the component function curves.

The following graphs illustrate the component function curves associated with the two most important variables of the IGANN model fitted on the average cost.

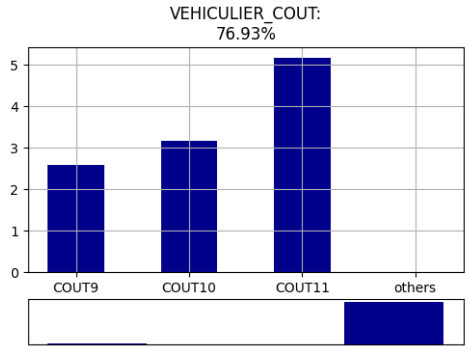

Figure 13: Vehicular Model

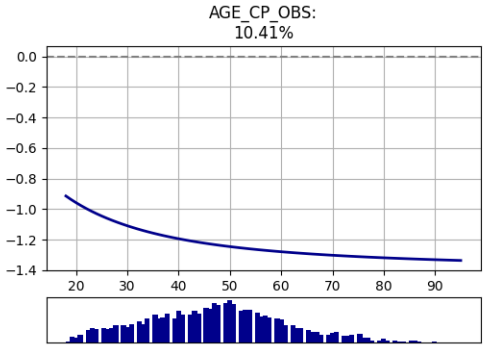

Figure 14: Driver's Age

The vehicular model we constructed plays a crucial role in all models, particularly in the IGANN model. The vehicular model stands out as a key variable for modeling average cost, with an importance assessed at 76.93% in the model, followed by the driver's age (10.41%). The observed relationships between these covariates and the dependent variable confirm expected trends. The graph on the vehicular model shows an increasing relationship with the rise in class risk, exerting a positive impact on the average cost.

The driver's age, meanwhile, has a negative impact on the average cost prediction. As the prediction results from the sum of individual contributions, the average cost decreases as the driver's age increases.

Comparison of Results

The following table presents the predictive performance of the models.

Models	Average Cost		Claim Frequency	
	MAE	**RMSE**	**MAE**	**RMSE**
GLM	1244.5	1752.47	0.038	0.1403
GAM	1237.35	1745.68	0.037	0.1395
IGANN	**1188.84**	**1711.87**	**0.036**	**0.1395**
XGBoost	**1177.79**	**1681.60**	**0.036**	**0.1285**

Table 6: Comparison of average cost and claim frequency model performances

The table above compares the performance of four predictive models: GLM, GAM, IGANN, and XGBoost, based on two indicators: MAE (mean absolute error) and RMSE (root mean square error).

On the test set, the XGBoost model stands out as the best-performing model, with the lowest errors, demonstrating its ability to accurately predict target values. The IGANN model follows closely, also offering satisfactory results, though slightly behind XGBoost. On the other hand, the GAM and GLM models exhibit higher errors, indicating lower predictive capability compared to the XGBoost and IGANN models.

In the same goal of model comparison, we focus on comparing the predicted value curves of the GLM, GAM, and IGANN models with the observed value curve based on age classes.

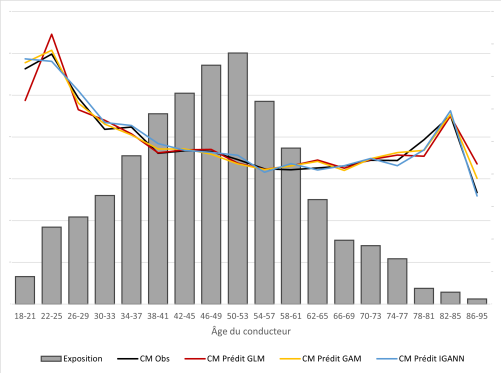

Figure 15: Average Cost

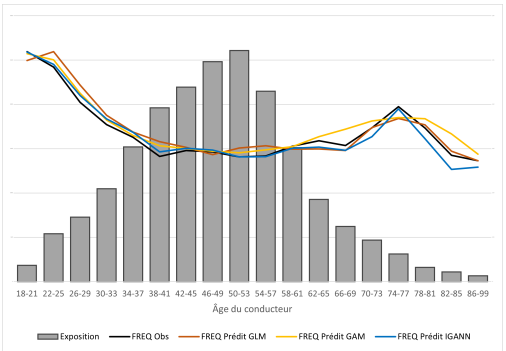

Figure 16: Claim Frequency

Figure 7 shows that the three compared models generally display a good fit between observed and predicted values, with a few remarks to note. The GLM model struggles to generalize predictions for drivers aged 18 to 33 years. It underestimates predictions for the 18-21 age class and overestimates those for the 22-25 age class. Analysis shows that predictions are generally better for intermediate age classes, with a slight additional improvement in the IGANN model compared to the other two models for these classes. A similar trend is observed for predictions regarding older age classes.

The graph in Figure 8, which shows claim frequency, confirms the slight improvement in fit of the IGANN model compared to the other two models. Indeed, claim frequency for young drivers is overestimated by all models, but less so by the IGANN model. Similarly, for intermediate age classes, the IGANN model curve is closer to that of observed frequencies than the other models. The main difference lies in the last two age classes, where the GLM model generalizes better. The GAM model, on the other hand, struggles to generalize claim frequency for the highest age classes by overestimating claim frequency for these classes.

Conclusion

This thesis has thoroughly explored the integration of the **vehicular model** and interpretable Machine Learning models, particularly the Neural Additive Model (**IGANN**), in the context of pricing the **DTA** coverage in auto insurance. The study showed that adding the constructed vehicular model better captures the risks inherent in vehicle characteristics. This vehicular model demonstrated its effectiveness by reducing prediction errors compared to the GLM model without this variable and even compared to the use of the standard vehicular model provided by SRA, which consists of the SRA group and SRA class.
Applying the IGANN model to the DTA coverage demonstrated that the IGANN model not only reduces prediction errors but also maintains three types of interpretability (variable importance, local prediction, and the relationship between covariates and the dependent variable) for actuaries, making pricing decisions more transparent. The comparison of the IGANN model with other models in terms of predictive capability showed that the IGANN model falls between the XGBoost model with the best performance and the classical models, which are the GAM and GLM models. Thus, the IGANN model is a good intermediate model combining prediction and interpretability.
It is essential to note that the goal is not to replace GLM models, which remain fundamental tools in actuarial science and are not likely to disappear soon. These models, while presenting certain limitations, remain indispensable for their simplicity and interpretability. This thesis rather proposes an opening towards other modeling approaches, such as the IGANN model, which can be explored in addition to GLMs to address current and future challenges in the sector.

Limitations of the Study and Areas for Improvement

The main limitations of this study can be summarized as follows:

- The use of the current zoning system in pricing, whereas it would have been possible to develop a customized zoning system based on the portfolio, which would have enriched the pricing.
- The lack of consideration for smoothing residuals in constructing the vehicular model.

- A specific method for introducing interactions in the GLM model could have been preferred over the one used in this study.
- Models including the constructed vehicular model were not compared to those using all vehicular variables as risks associated with the vehicle.
- The IGANN model, due to its design, does not include interactions between variables and is limited to a single hidden layer in the neural network structure; further studies are needed to integrate these elements.
- The IGANN model should have been compared to a model combining decision trees with the GAM model.
- A comprehensive study could have been conducted by calculating final pure premiums and conducting detailed analyses on them.

Table des matières

Introduction

Face à une augmentation continue des prestations automobiles, les assureurs doivent sans cesse réévaluer leurs modèles afin de mieux aligner les primes avec les risques encourus par les assurés. Cette exigence de précision dans la tarification des produits d'assurance est un défi constant pour les actuaires, qui doivent concilier la performance prédictive des modèles avec leur capacité à être interprétés de manière transparente. L'enjeu est particulièrement crucial dans le secteur de l'assurance automobile, où l'optimisation tarifaire est au cœur des préoccupations pour maintenir l'équilibre financier des compagnies d'assurance.

La première solution à cette problématique étudiée dans ce mémoire est l'intégration d'une variable spécifique, le véhiculier, qui agrège les caractéristiques du véhicule pour mieux capturer le risque inhérent. Le véhiculier permet une meilleure segmentation des risques liés aux types de véhicules assurés. En intégrant ce véhiculier dans les modèles de tarification, il est possible de mieux capturer les spécificités des véhicules assurés et ainsi d'améliorer la prédiction des modèles. Dans la littérature des mémoires d'actuariat, l'utilisation d'un véhiculier personnalisé, construit à partir des données spécifiques au portefeuille de l'assureur, a montré une amélioration significative des performances des modèles de tarification. Par exemple, un modèle linéaire généralisé (GLM) intégrant ce véhiculier a permis de réduire les erreurs de prédiction comparativement aux modèles sans cette variable.

Historiquement, ces modèles GLM sont largement utilisés pour leur simplicité et leur capacité à fournir des résultats interprétables. Cependant, ils montrent parfois des limites lorsqu'il s'agit de capturer des relations complexes dans la modélisation. Pour répondre à ces défis, les modèles additifs généralisés (GAM) ont été introduits, offrant une plus grande flexibilité grâce à l'utilisation de fonctions splines pour modéliser des relations non linéaires. Parallèlement, l'avènement des méthodes de Machine Learning, telles que l'utilisation des réseaux neuronaux et des modèles XGBoost et Random Forest, a permis de repousser les limites de la prédiction. Ces modèles de Machine Learning sont désormais capables de traiter des volumes de données massifs et de détecter des interactions complexes entre les variables, surpassant souvent les modèles traditionnels en termes de performance prédictive. Toutefois, ces modèles, qualifiés de "boîtes noires", souffrent d'un manque d'interprétabilité, ce qui limite leur adoption

dans des secteurs comme l'assurance, où la transparence des décisions est essentielle. C'est dans ce contexte qu'émergent des modèles hybrides combinant la puissance des modèles de Machine Learning avec l'interprétabilité des modèles additifs généralisés. Le Modèle Additif Neuronal (IGANN) en est un exemple notable. Ce modèle combine la flexibilité des réseaux neuronaux avec la structure interprétable des modèles GAM, offrant ainsi un compromis précieux entre précision prédictive et transparence. La construction du modèle IGANN, consistant à remplacer les fonctions splines par des réseaux de neurones, constitue la deuxième solution complémentaire à la problématique de cette étude.

Pour répondre à cette problématique à travers les deux solutions proposées, ce mémoire est structuré en quatre chapitres. Le premier chapitre pose le contexte et les enjeux de l'étude, en se concentrant sur les défis actuels de la tarification en assurance automobile, notamment la gestion des risques liés aux véhicules. Le deuxième chapitre est consacré à la construction de la base de données utilisée pour cette étude et aux analyses univariées, bivariées et multidimensionnelles. Le troisième chapitre se penche sur la construction du véhiculier et son intégration dans le modèle GLM avec une évaluation des impacts de l'introduction de cette variable dans le modèle. Enfin, le quatrième chapitre explore le Modèle Additif Neuronal (IGANN) et son application dans le contexte de la tarification en assurance automobile, en proposant une analyse comparative des modèles et en mettant en lumière les avantages en termes d'interprétabilité et de performance du modèle IGANN.

Chapitre 1

Contexte et enjeux de l'étude

Ce premier chapitre introduit les concepts clés nécessaires à la compréhension de notre étude, tout en définissant les objectifs du mémoire. Il se divise en trois parties distinctes. Dans la première partie, nous clarifions la notion d'assurance, puis nous nous concentrons sur le domaine spécifique de l'assurance automobile, principal produit d'assurance de cette étude. La deuxième partie pose les bases de la tarification en explorant les principes de mutualisation et de segmentation. Enfin, la dernière partie met l'accent sur le risque associé aux véhicules et sa caractérisation à travers la construction d'un véhiculier.

1.1 Généralités sur l'assurance

1.1.1 Notion d'assurance

La notion d'assurance ne peut être dissociée de celle du risque. En effet, le risque imprègne toutes les activités humaines, et revêt une importance particulière, car elle implique la possibilité d'un événement néfaste, dont la gravité et le moment précis de survenance demeurent inconnus. Les trois principales caractéristiques du risque englobent le préjudice potentiel, la probabilité de son occurrence, et l'étendue de ce préjudice. Le risque émerge de la réalité où le préjudice a une chance de se matérialiser, entraînant des dommages d'une ampleur déterminée.

L'aléa se manifeste dans toutes les activités humaines, ce qui justifie la nécessité de détenir une couverture au préalable pour se protéger en cas de survenance du risque, dont le moment précis demeure incertain. Le concept de risque est étroitement lié à celui de l'**assurance**, qui constitue un moyen permettant à une personne physique ou morale de réparer les dommages qu'elle subit ou qu'elle inflige à autrui.
L'**assurance** est définie, juridiquement, comme un accord contractuel où l'assureur s'engage, moyennant le paiement d'une prime, à fournir une protection au souscripteur contre les conséquences d'un événement aléatoire prévu dans le contrat.

La personne physique ou morale, désignée sous le terme **assuré**, qui cherche à se prémunir contre les risques, effectue un paiement à une entité morale. Lorsque le risque se matérialise, cette entité, nommée **assureur**, prend en charge la réparation du préjudice. La somme d'argent perçue par l'assureur est appelée **cotisation** ou **prime**, en fonction de la typologie de l'entité.
Certaines notions doivent être clarifiées à ce stade. En effet, le bénéficiaire de la réparation peut être une personne distincte de l'assuré, de même que le souscripteur de l'assurance peut différer de l'assuré. À titre d'exemple, nous pouvons considérer le cas d'une entreprise (le souscripteur) qui verse une prime ou une cotisation à un tiers pour obtenir une indemnisation des sinistres en faveur des enfants (les bénéficiaires) de son salarié (l'assuré). Selon son fonctionnement, le but essentiel de souscrire à une assurance est de se protéger contre un événement futur, incertain et particulièrement aléatoire.

Les transactions, à l'exception des échanges de troc, impliquent un prix de vente en échange d'un service ou d'un produit, représentant la valeur de l'élément échangé. Dans les activités courantes, le prix de vente ou le coût de production est généralement connu au moment des transactions et est payé par le consommateur.
Cependant, le fonctionnement de l'assurance présente une particularité, car l'assureur reçoit initialement une somme d'argent, et c'est seulement dans le futur qu'il connaîtra le prix de vente réel de son service. Ce processus inverse le cycle de production dans le domaine de l'assurance. Cette **inversion du cycle du production** oblige l'assureur à développer des méthodes pour déterminer les montants et les facteurs explicatifs des sinistres, afin de proposer une prime ou une cotisation équitable. Dans cette démarche, l'assureur s'appuie sur des données historiques pour tenter d'anticiper l'avenir.

Tous les risques ne peuvent être assurés. Certains ne le sont pas en raison d'une probabilité élevée de survenance du risque. Dans d'autres cas, les montants des sinistres sont trop faibles, ce qui conduit l'assureur à établir une franchise, excluant ainsi ces situations de couverture. Enfin, il existe une dernière catégorie de risques non couverts, liés à des préjudices illégaux, criminels ou en violation des politiques publiques. Ces risques échappent à la sphère d'assurabilité.
La notion d'assurance s'étend au-delà de la couverture des risques de dommages aux biens, incluant également les risques de responsabilité, les risques liés à la vie humaine ainsi que ceux associés aux préjudices corporels résultant de maladies et d'accidents.

1.1.2 Assurance automobile

Le secteur de l'assurance est divisé en deux groupes : l'**assurance non-vie** et l'assurance vie. L'assurance non-vie est constituée de l'assurance de biens, de responsabilité et les assurances de dommages corporels. Nous pouvons également distinguer deux types d'assurances que sont l'assurance de dommages et l'assurance de personnes.
La figure 1.1 propose une présentation détaillée de toutes ces classifications.

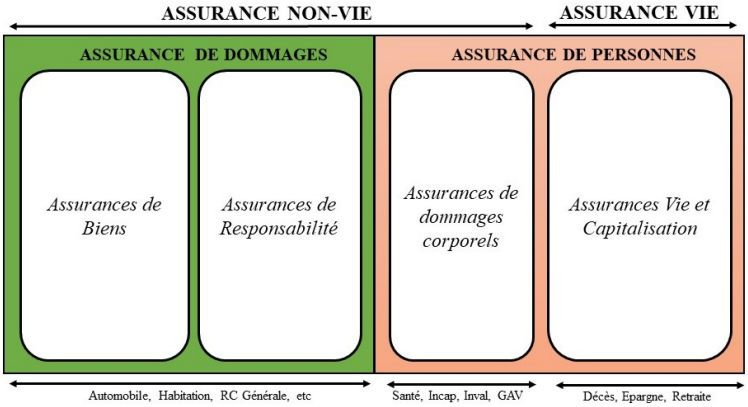

Figure 1.1: Typologie des assurances[1]

Certains contrats d'assurance non-vie sont obligatoires, tels que l'**assurance automobile** et l'assurance multirisque habitation, qui sont d'ailleurs les principaux produits d'assurance de dommages.
Depuis 1958, l'assurance automobile est devenue une assurance obligatoire en France, garantissant la Responsabilité Civile qui consiste en la réparation des dommages matériels ou corporels causés à autrui. Outre cette garantie obligatoire, l'assurance automobile propose d'autres catégories de garanties, telles que les dommages au véhicule de l'assuré, dont la garantie **Dommage Tous Accidents (DTA)**, les garanties corporelles du conducteur, et d'autres garanties facultatives, comme la Protection Juridique. La Défense Pénale et Recours Suite à un Accident (DPRSA) constitue également une garantie obligatoire dans le cadre d'un contrat d'assurance automobile.
La garantie DTA offre à l'assuré la possibilité de couvrir les dommages subis par son véhicule, quel que soit le type d'accident. Cette garantie est destinée à la réparation des dégâts causés au véhicule du conducteur, sans dépendre de la responsabilité de l'assuré ni de l'identification du tiers impliqué dans l'accident. Chez Generali, la DTA est appliquée dans les situations suivantes :

- Impact avec un objet fixe ou en mouvement, qu'il s'agisse d'un renversement du véhicule ou d'un simple basculement, même en l'absence de collision préalable, lorsque ces incidents affectent le véhicule assuré, ainsi que la remorque ou la caravane spécifiée dans les Dispositions Particulières.
- La détérioration du châssis, de l'essieu, de la roue, ou la rupture de l'attelage en cours de circulation, engendrant des dommages à la remorque ou à la caravane spécifiée dans les Dispositions Particulières, s'ils sont consécutifs.
  (Dispositions Générales, Assurance Auto, Generali)

Le secteur de l'assurance automobile se distingue par des contrats relativement courts, généralement d'une durée d'un an renouvelable par tacite reconduction, à

[1]Incap : Incapacité, Inval : Invalidité, GAV : Garantie Accidents de la Vie.

moins qu'une des parties ne décide de résilier le contrat. L'assureur peut résilier le contrat pour diverses raisons, telles qu'une sinistralité élevée de l'assuré, tout comme l'assuré peut résilier s'il trouve une offre plus avantageuse auprès d'un autre assureur. La loi Hamon offrant à l'assuré la possibilité de résilier son contrat sans frais à la fin de la première année et l'émergence de comparateurs d'assurance automobile en ligne intensifient la concurrence, rendant ainsi ce secteur très dynamique et concurrentiel.

Selon *Assurland* qui est d'ailleurs un comparateur de produits d'assurance en ligne, le taux de variation des primes annuelles d'assurance automobile est estimé à 3,3% en 2023. Cette augmentation est attribuée à une hausse de la sinistralité et à une augmentation des coûts de réparation des dommages. Les études d'*Assurland* indiquent spécifiquement une augmentation des coûts des sinistres de 8,42% en 2023. La Sécurité Routière rapporte une hausse des accidents de la circulation au cours du dernier trimestre de 2023, ce qui pourrait entraîner une augmentation attendue de 3,5% des primes d'assurance automobile en 2024.

Le secteur de l'assurance de dommages et de responsabilité montre une tendance à la hausse des prestations réalisées ces dernières années. Les chiffres clés de France Assureurs [2] indiquent qu'en 2022, l'équilibre technique de ce secteur d'assurance s'est détérioré, avec une augmentation des prestations deux fois supérieure à celle des primes. Tout ceci ajouté à la compétitivité et la dynamicité du secteur de l'assurance automobile oblige les assureurs à exploiter de manière judicieuse les données devenues essentielles pour non seulement optimiser la gestion de leur portefeuille, mais pour surtout perfectionner les offres tarifaires afin de rester compétitifs, notamment grâce aux nouvelles techniques de **Machine Learning**.

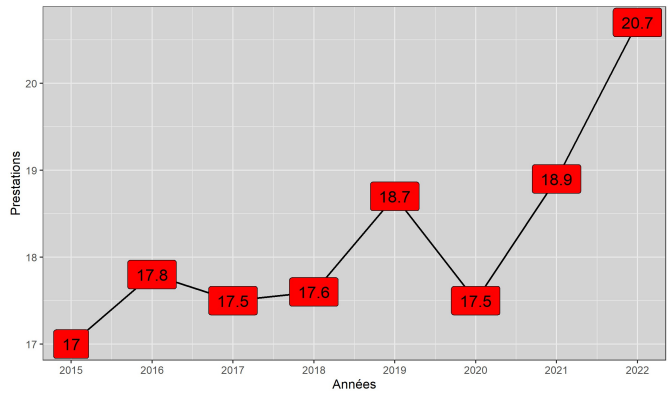

Figure 1.2: Prestations en assurance automobile

Source : Rapports Annuels France Assureurs 2020-2022

[2] France Assureurs, anciennement désigné sous l'appellation FFA (Fédération Française de l'Assurance), est l'instance principale représentant les entreprises d'assurance en France.

1.2 Fondements de la tarification en assurance automobile

La conception d'un produit d'assurance implique l'intervention de plusieurs services de l'assureur. Les services marketing, commercial et juridique définissent l'offre et les conditions contractuelles en mettant en place les conditions de souscription, les garanties, les exceptions, les types d'indemnisation (forfaitaire, proportionnelle à la valeur du bien, VRADE [3]), les franchises, les plafonds, etc. Le service Actuariat détermine les primes ou cotisations à payer par les assurés ainsi que le Business Plan. La distribution est ensuite effectuée par divers canaux : les réseaux bancaires, les agents généraux, les courtiers, les sites Internet, les salariés, etc.
La **tarification** fait partie des deux missions principales de l'actuaire en assurance non-vie, la deuxième étant le calcul des provisions techniques qui représentent les engagements de l'assureur. L'actuaire garantit avec ces deux missions, l'équilibre technique de l'assureur.
La prime versée par l'assuré est la prime commerciale, composée d'une prime de base appelée **prime pure**, représentant la part principale, ainsi que d'autres montants supplémentaires tels que les chargements de sécurité et les frais. Les chargements de sécurité représentent la marge bénéficiaire de l'assureur et la rémunération de ses fonds propres, tandis que les frais comprennent les commissions, les coûts de réassurance et les frais généraux. La prime pure est l'objet de la tarification de l'actuaire.

Afin de calculer les primes ou cotisations, l'actuaire se sert de deux outils essentiels :

- Les informations accessibles dans le domaine de l'assurance : elles revêtent une importance particulière. Elles reflètent le passé et offrent à l'assureur la possibilité de projeter les tendances futures. Il est impératif que ces données soient cohérentes et précises afin de garantir une utilisation adéquate dans l'évaluation des risques et la détermination de primes justes.

- Les modèles mathématiques : ils constituent un autre instrument crucial, exploitant les données dans le but d'anticiper les évolutions futures.

1.2.1 Principe de mutualisation

Les assurés ont la possibilité de cotiser collectivement et, après la survenance d'un sinistre, de recevoir une prestation pour la réparation des dommages, mettant ainsi en œuvre le principe de mutualisation des risques.
La prime pure est la représentation de l'espérance des montants des sinistres. Lorsque les risques portés par les assurés sont homogènes dans le portefeuille, la prime est l'espérance mathématique du coût total des sinistres. Son expression est la suivante :

$$\text{Prime Pure} = \mathbb{E}[S] \quad \text{avec } S = \sum_{i=1}^{N} X_i \tag{1.1}$$

[3] Valeur de Remplacement à Dire d'Expert

où S désigne le montant total des sinistres, N représente le nombre total de sinistres, et X_i est le montant du $i^{\text{ème}}$ sinistre.

Nous faisons l'hypothèse que les montants des sinistres X_i sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d) de même loi notée X. Nous supposons aussi que le nombre total de sinistres N est une variable aléatoire indépendante des montants des sinistres. L'espérance du montant total des sinistres est alors déterminée par :

$$\begin{aligned} \mathbb{E}[S] &= \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] \\ \mathbb{E}[S] &= \mathbb{E}[N]\mathbb{E}[X] \end{aligned} \tag{1.2}$$

$$\textbf{Prime Pure} = \mathbb{E}[\mathbf{N}]\mathbb{E}[\mathbf{X}] \tag{1.3}$$

La prime pure est déterminée en multipliant l'espérance du nombre total de sinistres, également appelée **fréquence**, par le coût moyen d'un sinistre, désigné comme **sévérité**. En pratique, la fréquence de sinistres et la sévérité sont modélisées séparément avant d'être combinées pour calculer la prime pure. Cette approche de mutualisation implique l'application d'une même prime à tous les assurés présentant des profils de risque similaires. Par conséquent, la prime varie en fonction de l'évolution de la fréquence de sinistres ou de la sévérité, augmentant en cas de hausse de l'un ou de l'autre et diminuant dans le cas contraire.

1.2.2 Principe de segmentation

Le principe de mutualisation est bénéfique pour tous les assurés lorsque leurs profils de risque sont similaires, mais cette situation n'est pas toujours la norme au sein d'un portefeuille d'assurés. Les assurés pouvant donc présenter des profils de risque différents, le principe de segmentation permet de résoudre cette disparité. Il consiste à former des groupes homogènes de risques en établissant des primes qui reflètent les caractéristiques spécifiques de chaque assuré. Cette approche permet d'ajuster les coûts d'assurance de manière plus précise, favorisant ainsi une répartition équitable des charges parmi les assurés en fonction de leur profil de risque individuel.

Figure 1.3: Principe de segmentation des risques

Le profil est représenté par les variables tarifaires. En assurance automobile, les risques de l'assuré peuvent être regroupés en trois catégories :

- Le risque lié au profil du conducteur, déterminé par des facteurs tels que l'âge, l'ancienneté du permis, la catégorie socio-professionnelle (CSP), etc.,
- Le risque associé aux caractéristiques du véhicule de l'assuré : puissance, nombre de chevaux, système d'alimentation, etc.,
- Le risque lié à la situation géographique du conducteur : lieu de résidence et lieu de circulation du véhicule.

La tarification dans le cadre de la segmentation des risques repose sur la même théorie du risque, en multipliant la fréquence de sinistres par la sévérité pour obtenir la prime pure. Cependant, dans ce contexte, la fréquence et la sévérité sont des espérances conditionnelles mathématiques, et la prime pure est également conditionnelle au profil de l'assuré. Elle est ainsi définie par :

$$\text{Prime Pure}|\text{C} = \mathbb{E}[S|C] \quad \text{avec } S = \sum_{i=1}^{N} X_i \tag{1.4}$$

où $C = (c_1, c_2, \ldots, c_p)$ représente les caractéristiques du risque.

Sous les mêmes hypothèses concernant le nombre total de sinistres N et les montants des sinistres X_i, nous avons :

$$\mathbf{Prime\ Pure}|\mathbf{C} = \mathbb{E}[\mathbf{N}|\mathbf{C}]\mathbb{E}[\mathbf{X}|\mathbf{C}] \tag{1.5}$$

La modélisation de la fréquence de sinistres et de la sévérité repose sur l'utilisation de modèles de tarification puissants visant à améliorer les performances, permettant ainsi de déterminer une prime étroitement liée au risque de l'assuré pour éviter l'antisélection[4]. Les assureurs ont donc tout intérêt à développer des modèles qui prédisent de manière aussi précise que possible les risques. Les modèles de Machine Learning, qui atteignent des performances remarquables, sont largement utilisés pour prédire ces deux facteurs de risque. Cependant, ces modèles sont considérés comme des boîtes noires car ils ne sont pas facilement interprétables comme les modèles classiques de tarification, tels que les modèles linéaires généralisés. Ce mémoire a pour objectif final de se pencher sur cet aspect boîte noire des modèles de Machine Learning en construisant un modèle de machine learning qui est tout aussi performant qu'interprétable. Ainsi, la fréquence et la sévérité seront mieux prédites tout en conservant la propriété d'interprétabilité, permettant aux dirigeants de prendre des décisions sur la conception du produit et les profils à assurer, par exemple.

1.3 Risque inhérent au véhicule

Pour une tarification plus précise, il est essentiel d'effectuer une sélection rigoureuse des variables tarifaires et d'agréger certaines d'entre elles afin d'obtenir une caractérisation

[4] L'antisélection se produit lorsque l'assureur se retrouve avec un niveau de risque des assurés plus élevé que prévu initialement.

plus fine. Les variables liées à la situation géographique peuvent être regroupées en une variable appelée zonier, tandis que celles associées aux caractéristiques du véhicule peuvent être combinées en une variable appelée **véhiculier**, premier objectif de notre étude.
Les variables liées au véhicule peuvent ainsi être agrégées en une variable que nous appelons véhiculier, qui saisit le risque inhérent au type de véhicule de l'assuré. L'utilisation du véhiculier permet une segmentation plus précise du risque, favorisant une proposition de prime en adéquation avec le risque porté par l'assuré à travers son véhicule. Selon les mémoires d'Actuariat de GNANSOUNOU (2022) et FADIL (2020), l'utilisation du véhiculier construit sur la base de données des assurés peut également améliorer les performances des modèles de tarification.

Les assureurs utilisent généralement un véhiculier standard fourni par l'association de **Sécurité et Réparation Automobile** (SRA). Cette association professionnelle, créée en 1977, regroupe toutes les sociétés d'assurance automobile. La mission principale de la SRA est de contribuer à la réduction du nombre et du coût des sinistres, en diffusant des informations sur les véhicules aux sociétés d'assurances, en promouvant la sécurité routière, en luttant contre le vol de véhicules et en maîtrisant les coûts de réparation. La diffusion d'informations sur les véhicules est particulièrement pertinente pour notre étude. En effet, la SRA fournit une base de données, que nous appelons **Base SRA**, contenant les caractéristiques des véhicules de types 4, 3 et 2 roues. De plus, elle établit trois classifications des véhicules : le groupe SRA, qui indique la dangerosité du véhicule, la classe SRA, qui reflète la valeur d'achat à neuf toutes taxes comprises du véhicule, et la classe de réparation, basée sur la valeur hors taxes du panier de pièces SRA et des chocs à 15 km/h. Le groupe SRA est une variable qui prend des valeurs entières naturelles comprises entre 20 et 50. La classe de prix prend des modalités qualitatives allant de "A" à "Z" puis les modalités "ZA" et "HC". Le véhiculier utilisé par la plupart des assureurs est celui constitué du groupe et de la classe SRA.

Résumé du chapitre

Ce chapitre définit les termes essentiels pour comprendre les thématiques des chapitres suivants. Il aborde la segmentation et la mutualisation des risques à travers les fondements de la tarification, l'inversion du cycle de production en assurance, et l'augmentation des prestations en assurance automobile. Ces éléments soulignent la nécessité d'une tarification robuste adaptée aux profils de risque des assurés. Pour cela, notre étude utilise la Base SRA et des méthodes de modélisation et de classification pour créer un véhiculier, permettant une segmentation efficace du risque et l'amélioration des modèles de tarification. La réalisation de nos objectifs s'appuie sur la garantie DTA. Le chapitre suivant est consacré à l'analyse exploratoire des données, essentielle pour cet objectif.

Chapitre 2

Analyse exploratoire des données

Ce chapitre est consacré à l'analyse exploratoire des données, étape indispensable pour une compréhension approfondie du portefeuille d'assurance automobile étudié. Nous débuterons par la construction de notre base de données, en détaillant les différentes sources de données et les méthodes de jointure utilisées. Ensuite, nous traiterons les données pour garantir leur qualité et cohérence. Nous réaliserons ensuite une analyse univariée pour examiner la distribution individuelle de quelques variables, puis une analyse bivariée pour explorer les relations entre paires de variables. L'analyse multidimensionnelle sera également abordée avec la technique de l'analyse en composantes principales (ACP) pour réduire la dimensionnalité des données tout en conservant un maximum d'information. Cette analyse globale permettra de poser des bases solides pour les étapes de modélisation et de tarification ultérieures.

2.1 Construction de la base de données

2.1.1 Présentation des bases de données

Les données utilisées dans ce mémoire proviennent du portefeuille de la compagnie d'assurance l'**Equité**, filiale de Generali et représentante de sa Direction des partenariats en France. L'Equité est le porteur de risque des contrats d'assurance de dommages ou de personnes détenus par les courtiers partenaires avec lesquels elle collabore. Les partenaires fournissent à l'Équité, conformément aux accords contractuels des bases de données. Il s'agit de la `Base Portefeuille` contenant les images des contrats. Elle inclut toutes les informations relatives aux contrats, aux assurés, aux véhicules et aux lieux de résidence des assurés. Ils fournissent également une `Base Liaison`, qui comprend les informations sur les sinistres, et celles des contrats associés aux sinistres ainsi que les numéros d'immatriculation. Enfin les partenaires fournissent une base de données de sinistralité nommée `Base Sinistres`. Elle renferme les informations sur les sinistres survenus et sur les contrats associés à ces sinistres, à l'exception des numéros de contrat. Le tableau suivant présente les principales informations des bases de données, les autres variables sont présentées en annexe.

Informations	**Description**
Contrat	Numéro de contrat : **Clé Primaire**
	Date de début d'image : **Clé Secondaire**
	Date de fin d'image : **Clé Secondaire**
	Numéro de Police Mère : Regroupement de numéros de contrats
	Année d'Exercice
	Nom du Partenaire
	Type de Formule
	Type de Produits : Contrats quatre Roues Standards et Contrats quatre Roues Malussées
	Date d'effet du contrat
Assuré	Usage du Véhicule
	Situation Familiale
	Âge du conducteur principal
	Catégorie Socio-Professionnelle (CSP)
	Coefficient Réduction Majoration (CRM)
	Nombre d'années de détention du contrat
	Antécédents assurance
	Nombre de conducteurs
	Ancienneté du Permis
	Lieu de garage du véhicule
	...
Véhicule	Code Auto du véhicule : **Clé Primaire**
	Numéro d'immatriculation du véhicule : **Clé Secondaire**
	Marque du véhicule
	Modèle du véhicule
	Alimentation du véhicule
	Puissance du véhicule
	Groupe SRA du véhicule
	Classe SRA du véhicule
	Classe de répartition du véhicule
	Ancienneté du véhicule
	...
Lieu de résidence	Code Postal de stationnement
	Ville de stationnement
	Zonier
Sinistres	Numéro du sinistre : **Clé Primaire**
	Date de survenance du sinistre : **Clé Secondaire**
	Type de sinistre
	Coût du sinistre

Table 2.1: Récapitulatif des informations liées aux contrats, aux assurés, aux véhicules, aux lieux de résidence et aux sinistres

Nous avons utilisé chacune des bases de données fournies par les partenaires, que nous avons ensuite agrégées pour construire la base de données finale. Les bases de données contenant les images des contrats des différents partenaires ont été fusionnées en une seule `Base Portefeuille`. Chaque ligne de cette `Base Portefeuille` correspond à une périodicité mensuelle des contrats. Le même processus d'agrégation a été appliqué aux `Bases Liaison` et `Bases Sinistres`. Nous disposons ainsi initialement de trois bases de données.

2.1.2 Jointure des bases de données

La construction de la base de données finale débute par l'identification des clés primaires et secondaires des bases de données disponibles.
La `Base Portefeuille` contient la clé primaire `Numéro de Contrat` ainsi que les clés secondaires `Date de début d'image`, `Date de fin d'image` et `Numéro d'immatriculation`. En parallèle, dans la `Base Sinistres`, on retrouve la clé primaire `Numéro de Sinistre` avec les clés secondaires `Numéro d'immatriculation` du véhicule et `Date de survenance du sinistre`. Pour construire la base finale, il est nécessaire de les relier entre elles. Cependant, la `Base Sinistres` ne possédant pas la clé primaire `Numéro de contrat`, nous utilisons la `Base Liaison` comme base intermédiaire afin de résoudre cette contrainte.
La première étape de la construction de la base de données consiste à attribuer à chaque ligne de la `Base Sinistres`, le numéro de contrat correspondant. Cela implique une jointure entre la `Base Liaison` et la `Base Sinistres` pour ajouter la variable `Numéro de contrat` dans la `Base Sinistres`. Ce processus se déroule en deux phases :

- La première jointure est réalisée avec la clé primaire `Numéro de sinistre` présente dans les deux bases. Toutefois, cette première jointure ne couvre pas toutes les correspondances de numéros de sinistre, ce qui nécessite une seconde jointure pour améliorer le taux de correspondance.

- Ensuite, une deuxième jointure est effectuée avec les clés secondaires `Date de survenance du sinistre` et `Numéro d'immatriculation` pour compléter les correspondances manquantes de la première jointure.

La `Base Sinistres` obtenue à la fin de la première étape contient la variable `Numéro de contrat`, qui sera utilisée pour effectuer la jointure avec la `Base Portefeuille`. Nous procédons ensuite à la deuxième étape de la construction de la base de données en réalisant cette jointure. Cette étape se déroule également en deux phases :

- Tout d'abord, nous effectuons une première jointure en utilisant les clés `Numéro de contrat`, `Date de début d'image` et `Date de fin d'image` de la `Base Portefeuille` ainsi que les clés `Numéro de contrat` et `Date de survenance du sinistre` de la `Base Sinistres`. La première condition de jointure est la correspondance des numéros de contrat entre les deux bases, et la seconde condition est que la date de survenance du sinistre soit comprise entre les dates de début et de fin d'image.

- Ensuite, nous réalisons une seconde jointure utilisant le numéro d'immatriculation à la place du numéro de contrat. Les autres clés et conditions de jointure restent inchangées. Cette deuxième jointure vise à améliorer la première, de manière similaire à l'étape précédente.

À ce stade, nous disposons d'une base de données presque complète. Dans l'étape suivante, nous avons réalisé les phases suivantes :

- Calcul du taux de correspondance par partenaire et sélection des partenaires avec un taux de correspondance supérieur à 97%. Au total, neuf partenaires ont été retenus. Le **taux de correspondance** mesure la proportion de lignes de la `Base Sinistres` ayant une correspondance dans la `Base Portefeuille`

- Conversion de la périodicité mensuelle à la périodicité annuelle, chaque ligne représentant désormais les informations annuelles du contrat. Si des avenants surviennent pendant l'année pour un assuré, il y aura autant de lignes que de contrats pour cet assuré au cours de l'année concernée

- Calcul de l'exposition, définie comme la période pendant laquelle l'assuré était effectivement couvert par l'assurance sur une durée spécifique. Sa formule dans notre cas est la suivante :

$$\textbf{Exposition} = \frac{\text{Durée du contrat sur l'année d'exercice}}{\text{Nombre de jours dans l'année d'exercice}}$$

- Élimination des variables liées aux contrats (à l'exception des variables concernant l'année d'exercice et le type de produits) et des variables d'identification liées à l'assuré comme le numéro d'immatriculation ou la date de naissance.

Pour construire le véhiculier, il est essentiel d'utiliser les variables spécifiques aux véhicules. À ce stade, notre base de données ne contient pas toutes les variables véhiculières. Ainsi, la dernière étape consiste à les intégrer à notre base de données. La Base SRA regroupe toutes les informations relatives aux véhicules que nous utilisons à cet effet. Elle a permis d'enrichir notre base de données avec de nouvelles variables véhiculières. Cette intégration a été réalisée en effectuant une jointure entre la Base SRA et notre base existante, utilisant la clé `Code Auto du véhicule`.

2.2 Analyse exploratoire des données

La base de données étant construite sur les neuf partenaires retenus, les résultats issus de ce choix sont résumés dans le tableau suivant :

Description	Valeur
Nombre de lignes	413 562
Nombre de variables	82
Proportion d'exposition	90%
Proportion des sinistres	98%
Proportion du montant des sinistres	97%
Années d'Exercice	2018 à 2023

Table 2.2: Résultats issus de la construction de la base de données

La proportion d'exposition représente la part d'exposition récupérée sur l'exposition totale de la `Base Portefeuille` en considérant uniquement les partenaires retenus. La proportion des sinistres ou du montant des sinistres est également définie de la même manière.
Ce mémoire se focalise sur la garantie **Dommages Tous Accidents (DTA**) définie au préalable, qui se compose de deux types de sinistres : dommages en stationnement et dommages hors stationnement impliquant ou non un autre véhicule.

2.2.1 Traitement des données

Le traitement des données est une étape cruciale dans toute analyse statistique et dans la préparation de données pour des modèles prédictifs ou des études empiriques. Il comprend plusieurs phases essentielles visant à garantir la qualité, la cohérence et la pertinence des données utilisées.
Le traitement de nos données a couvert plusieurs aspects. Nous avons débuté par la gestion des valeurs manquantes en calculant le taux de données manquantes pour chaque variable de la base de données. Ce calcul nous a permis d'éliminer les variables avec un taux de valeurs manquantes supérieur à 30%. Principalement, les variables supprimées étaient liées à l'assuré. Les variables restantes affichent toutes un taux de complétion d'au moins 70%.
Ensuite, nous avons traité les variables en fonction de leurs modalités. Pour les variables qualitatives restantes avec des valeurs manquantes, nous avons attribué la modalité représentant le mode aux valeurs manquantes dans certains cas. Pour deux variables qualitatives, nous avons utilisé la méthode du "Hot Deck", où les valeurs manquantes ont été aléatoirement attribuées à l'une des modalités de la variable concernée. Ces approches ont été adaptées à la typologie et à la nature de chaque variable. Pour les variables quantitatives, nous avons utilisé des méthodes d'imputation basées sur la moyenne ou la médiane, en fonction du type de variable.
Par la suite, nous avons effectué des contrôles de cohérence et des tests. À titre d'exemple, pour la variable `Âge`, nous avons supprimé les lignes où l'âge de l'assuré était inférieur à

18 ans. Nous avons également vérifié que l'âge évoluait correctement dans le temps pour cette variable. Avant de calculer l'exposition, nous avons vérifié que la date d'effet des contrats précédait la date de début d'image pour toutes les lignes. Nous avons également vérifié la nature des modalités de toutes les variables et apporté les corrections nécessaires en conséquence.
Le traitement des données s'est poursuivi par le regroupement de modalités de type métier sur certaines variables, notamment celles concernant l'usage du véhicule, la catégorie socio-professionnelle et le lieu de garage.
Enfin, nous avons supprimé les variables inutiles telles que la date de naissance du conducteur, la date de mise en circulation du véhicule, et la marque du véhicule qui présentaient trop de modalités ne pouvant être regroupées. De même, les variables n'ayant qu'une seule modalité, comme celle indiquant la présence de l'assistance au freinage avec uniquement la modalité `Non`, ont été retirées.

2.2.2 Analyses préliminaires

L'analyse univariée étudie une seule variable à la fois, en examinant sa distribution, ses mesures centrales et sa dispersion. Elle permet de comprendre les caractéristiques essentielles d'une variable sans considération pour d'autres variables du jeu de données. Nous allons décrire quelques variables essentielles avant de passer à l'analyse bivariée et multidimensionnelle qui sont plus importantes.
Tous les graphes de cette partie et dans l'analyse bivariée sont obtenus de la fonction `ggplot` du package `plotnine`.

Description du coût des sinistres

L'analyse du coût des sinistres est essentielle pour comprendre la variabilité des sinistres. Le boxplot suivant illustre cette répartition.

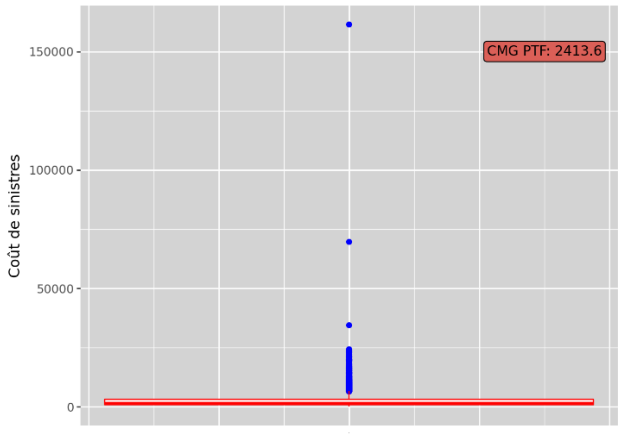

Figure 2.1: Boxplot du coût de sinistres

Légende : **CMG PTF** : Coût Moyen Global du Portefeuille.

Le boxplot illustre la répartition du coût des sinistres, un élément clé dans la tarification automobile. Le graphique révèle que le coût moyen global du portefeuille s'établit à 2413,6 unités monétaires. Cet outil visuel est particulièrement utile pour examiner la dispersion des données et repérer les valeurs extrêmes, lesquelles peuvent fortement influencer les modèles de tarification. Le graphique indique que la plupart des sinistres ont un coût faible, ce qui est typique pour des incidents mineurs ou de faible envergure. L'axe des ordonnées met en évidence que les coûts peuvent atteindre des montants très élevés, soulignant une grande variabilité. Plusieurs coûts de sinistre se distinguent nettement du corps principal de la distribution, représentant des sinistres avec des coûts exceptionnellement élevés, dépassant 150 000 unités monétaires. Bien que rares, ces valeurs extrêmes sont cruciales dans le contexte de la tarification, car elles correspondent à des sinistres graves qui doivent être pris en compte dans le calcul des primes. Ces sinistres particulièrement coûteux peuvent être dus à des événements spécifiques comme des catastrophes climatiques majeures, des accidents graves ou des fraudes. Pour assurer une tarification précise et pertinente, nous décidons d'écrêter ces sinistres.

Description du Zonier

Ce mémoire ne prévoit pas la construction d'un Zonier. Toutefois, nous avons inclus le Zonier utilisé pour la tarification actuelle.

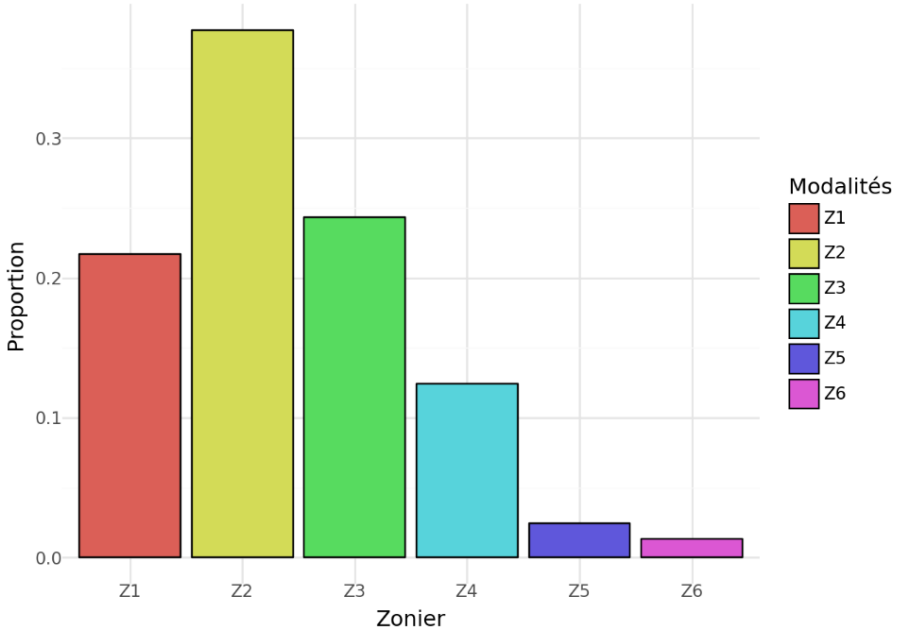

Figure 2.2: Histogramme de la variable `Zonier`

L'histogramme ci-dessus représente la répartition de la variable `Zonier` en termes de proportion. Les zones Z1, Z2, et Z3 représentent la majorité des observations, avec des proportions respectives d'environ 22%, 40%, et 25%. La zone Z2 est la plus représentée, ce qui pourrait indiquer une concentration de risques dans cette zone. En revanche, les zones Z5 et Z6 sont les moins représentées, avec des proportions d'environ 2% et 1% respectivement. Cette rareté pourrait signaler des caractéristiques spécifiques ou des risques particuliers qui, bien que peu fréquents, pourraient être significatifs.
Dans le cadre de la tarification, il est crucial de prendre en compte la variable `Zonier`. La segmentation des risques basée sur cette répartition permet d'identifier des risques communs dans les zones largement représentées et des risques uniques dans les zones moins représentées.

Description de l'âge des assurés

L'analyse de la variable `Âge` permet de se faire une idée précise de la population d'assurés du portefeuille étudié. En examinant la répartition des âges, on peut mieux comprendre les caractéristiques de cette population et identifier les segments d'âge prédominants. Le graphique suivant présente l'histogramme illustrant la distribution des âges des assurés.

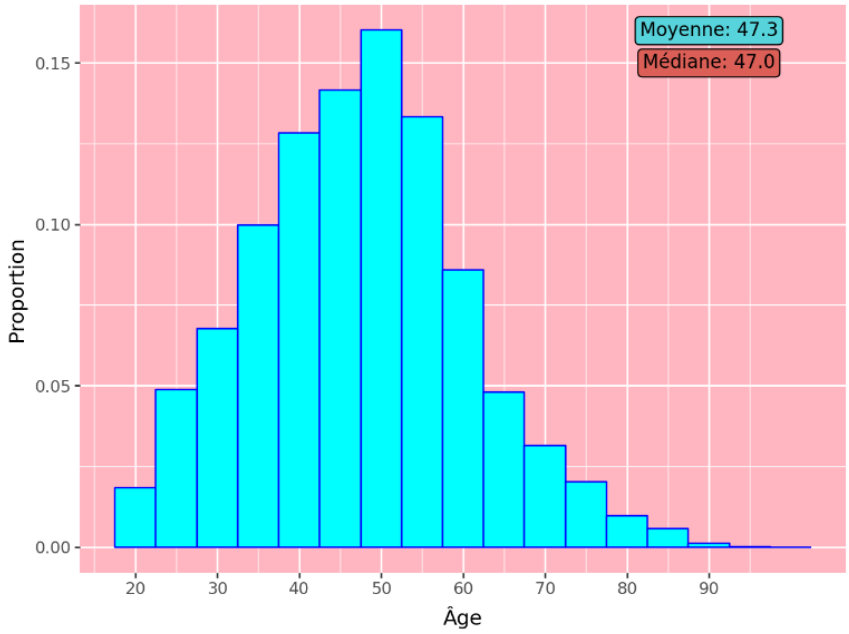

Figure 2.3: Histogramme de la variable `Âge`

De la figure 2.3, on observe que la distribution des âges suit une forme approximativement en cloche, avec une concentration élevée d'assurés dans la tranche d'âge de 30 à 60 ans. Plus précisément, la proportion maximale d'assurés se situe autour de 50 ans, ce qui indique que la majorité des assurés sont des adultes d'âge moyen. Les proportions

diminuent progressivement de chaque côté de ce pic central, suggérant une moindre représentation des jeunes adultes (moins de 30 ans) et des personnes âgées (plus de 60 ans). Les groupes d'âge ayant des proportions élevées sont notamment ceux autour de la cinquantaine. La moyenne de la distribution d'âge est 47.1 très proche de la médiane à 47.
Nous présentons la distribution de la fréquence de sinistres et du nombre de sinistres qui sont également importants en annexe de ce document.

L'analyse bivariée explore la relation entre deux variables afin d'identifier les tendances, permettant ainsi de tirer des conclusions éclairées dans notre étude.

Évolution temporelle de la fréquence de sinistres

Nous analysons la fréquence de sinistres dans le temps en lien avec le nombre de sinistres et l'exposition à travers le graphe suivant :

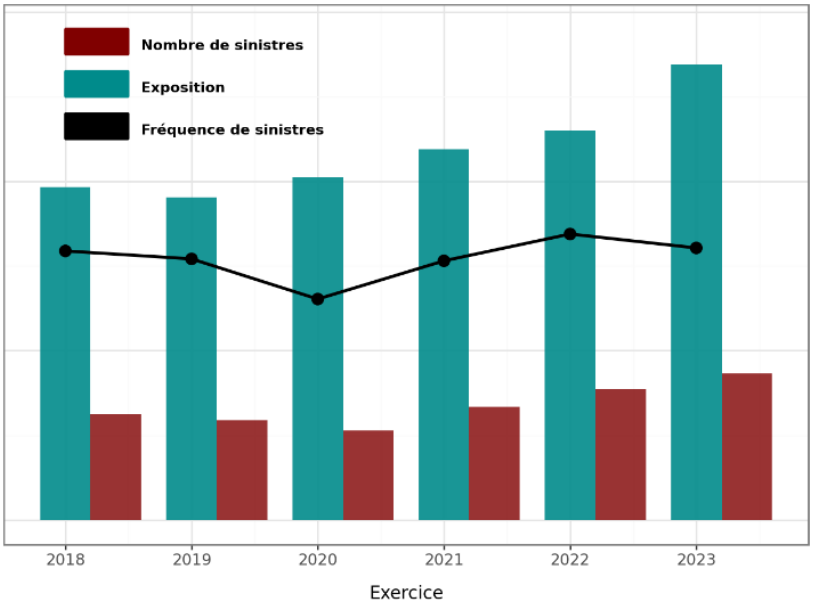

Figure 2.4: Évolution temporelle de la fréquence de sinistres

Les valeurs de la fréquence de sinistres et du nombre de sinistres ont été modifiées sur ce graphique.
La fréquence de sinistres représente le nombre de sinistres survenus par unité d'exposition. Elle est donnée par la formule suivante :

$$\textbf{Fréquence de sinistres} = \frac{\text{Nombre de sinistres}}{\text{Exposition}}$$

La fréquence globale des sinistres dans le portefeuille 3,09%.

De l'analyse de la figure 2.4, on remarque une augmentation globale du nombre de sinistres au fil des années avec un petit frein en 2020. Cette tendance suggère une hausse généralement continue des incidents déclarés au cours des périodes en étude.
Contrairement au nombre de sinistres, l'exposition a connu une augmentation plus marquée au fil des ans. En 2018, le niveau d'exposition était le plus bas de toutes les années observées, et il augmente de manière significative jusqu'en 2023. Cette hausse de l'exposition s'explique par une augmentation dans le temps, du nombre de contrats d'assurance dans le portefeuille.
La courbe de la fréquence de sinistres montre une variation faible hormis le pic vers le bas de l'année 2020. Malgré les variations du nombre de sinistres et de l'exposition, la fréquence de sinistres reste relativement sans variation extrême, oscillant légèrement autour d'une valeur fixe sans fluctuations importantes. La remarque notable concerne la diminution de la fréquence de sinistres en 2020 expliquée par la crise du COVID.

Évolution temporelle du coût moyen de sinistres

Nous allons également étudié l'évolution temporelle du coût moyen des sinistres représentée sur la figure ci-dessous :

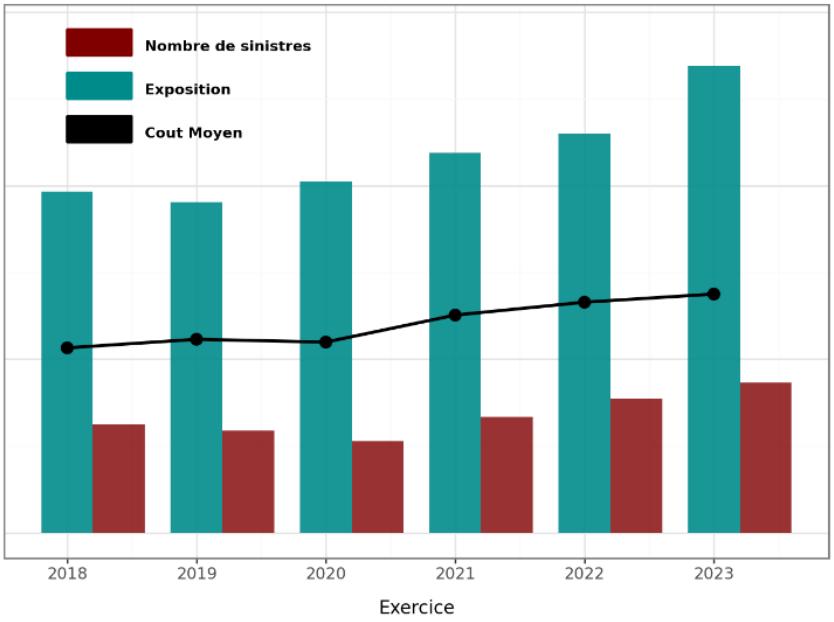

Figure 2.5: Évolution temporelle du coût moyen de sinistres

Le coût moyen de sinistres est une mesure utilisée pour évaluer le montant moyen des indemnités versées pour un sinistre dans un portefeuille d'assurés. Il est calculé en divisant le montant total des indemnités versées par le nombre de sinistres survenus pendant une période donnée. La formule pour calculer le coût moyen des sinistres est la suivante :

$$\textbf{Coût moyen de sinistres} = \frac{\text{Coût de sinistres}}{\text{Nombre de sinistres}}$$

La courbe d'évolution du coût moyen des sinistres de la figure 2.5 montre une oscillation des valeurs entre 2000 et 3000 unités monétaires. On observe une augmentation générale de 29.09% environ (passage de 2131,31 à 2751,31) mais modérée de 4,84% en moyenne d'augmentation annuelle, avec un léger ralentissement en 2020.

L'augmentation globale du coût moyen est notamment due à l'inflation observée ces dernières années, augmentant ainsi le coût de réparation des véhicules. Les assureurs doivent donc également être en mesure de proposer en face des primes qui reflètent cette augmentation. Il faut noter cependant que ces coûts moyens de sinistres n'ont pas subi de variations extrêmes. La faible variabilité des coûts moyens permet aux compagnies d'assurance de mieux prévoir leurs dépenses et de maintenir une gestion efficace des risques.

Coût de sinistres et Classe SRA

Nous montrons la relation entre les coûts de sinistres et la classe SRA à travers le graphe des boxplots suivant :

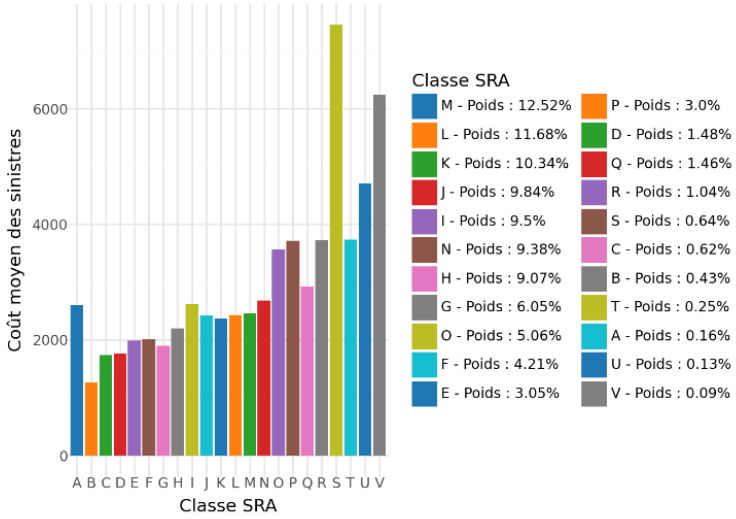

Figure 2.6: Coût moyen et Classe SRA

Le graphique ci-dessus présente un ensemble d'histogrammes, chaque classe SRA représentant un niveau de risque associé aux véhicules. Le poids de la classe J représente sa proportion dans l'ensemble des données. Les autres poids sont définis de la même manière. L'analyse de la figure montre que la classe S est une classe spéciale avec un coût moyen dépassant 6000 unités monétaires. L'investigation sur cette particularité révèle

que la majorité des coûts de sinistres les plus extrêmes (en particulier celui dépassant 150000 unités monétaires) sont associés à la classe S. De plus, cette classe présente une grande variabilité des coûts, ce qui entraîne un coût moyen beaucoup plus élevé par rapport aux autres classes. On observe néanmoins une tendance générale à la hausse des coûts moyens, conforme à l'observé. En effet, les classes SRA sont associées à la valeur des véhicules à neuf, et les coûts moyens augmentent au fur et à mesure que l'on passe d'une classe à l'autre, les réparations des pièces de véhicule devenant plus chères à mesure que la valeur à neuf du véhicule augmente. Ce graphique permet également de remarquer une tendance inverse entre le coût moyen des sinistres et le poids des classes S,T,U,V qui sont des classes associées à des coûts moyens élevés, indiquant une bonne gestion des risques.

L'analyse des coûts de sinistres à travers la classe SRA met en évidence l'importance cruciale d'une bonne segmentation des risques des véhicules. Une segmentation efficace permet d'évaluer précisément les risques, d'optimiser les primes, d'améliorer la rentabilité et de gérer les sinistres de manière plus ciblée et efficace. En comprenant les différences de coûts de sinistres entre les classes SRA, les assureurs peuvent personnaliser les primes pour refléter plus précisément le niveau de risque, évitant ainsi de surcharger ou de sous-charger certains segments. Cette analyse fondamentale motive le premier objectif de ce mémoire, qui est de développer une segmentation optimale des risques des véhicules en construisant un véhiculier. Ce véhiculier permettra une évaluation plus précise des risques associés aux véhicules.

Nous présentons en annexe deux graphiques supplémentaires qui sont également importants sur l'analyse bivariée : l'un illustre la relation entre le coût moyen et l'âge, et l'autre montre la relation entre le coût moyen et le Zonier.

2.2.3 Analyse multidimensionnelle

L'analyse multidimensionnelle aide à découvrir des relations significatives entre les variables et à réduire la complexité des données. Elle inclut diverses méthodes telles que l'analyse des corrélations entre les variables et l'analyse en composantes principales (ACP) que nous allons développer dans cette section.

Analyse de la corrélation entre les variables quantitatives

Les graphiques présentant les matrices de corrélation de cette partie ont été générés grâce à la fonction `heatmap` du package `seaborn`.

La corrélation est une statistique qui indique la force et la direction de la relation entre deux variables quantitatives. Elle est couramment utilisée pour déterminer si et comment deux variables sont liées. Il est essentiel d'identifier les variables significativement corrélées entre elles afin d'effectuer des suppressions de variables et d'éviter la redondance d'informations.

La corrélation est généralement mesurée par le **coefficient de corrélation de Pearson**, qui prend des valeurs comprises entre -1 et +1. Lorsque la valeur du coefficient est de +1, la corrélation est positive, parfaite et les deux variables augmentent ensemble linéairement tandis qu'un coefficient de -1 indique une corrélation négative parfaite, où une variable augmente pendant que l'autre diminue de manière linéaire. La valeur 0 du coefficient indique qu'il n'y a pas de corrélation linéaire entre les deux variables.

La formule théorique du coefficient de corrélation de Pearson est donnée par :

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

où :

- $\rho_{X,Y}$ est le coefficient de corrélation entre les variables X et Y,
- $\text{Cov}(X,Y)$ est la covariance entre X et Y,
- σ_X et σ_Y sont les écarts-types des variables X et Y.

Lorsqu'il s'agit des échantillons de variables ou des distributions, nous parlons de **coefficient de corrélation empirique** dont la formule est donnée par :

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

où :

- r_{xy} est le coefficient de corrélation empirique,
- x_i et y_i sont les valeurs des distributions X et Y pour l'observation i,
- $\bar{x}$ et $\bar{y}$ sont les moyennes empiriques des distributions X et Y,
- n est le nombre d'observations.

La corrélation n'implique pas la causalité. Autrement dit, même si deux variables fortement corrélées, cela ne signifie pas qu'une des variables cause l'autre. La corrélation peut être influencée par des variables tierces ou des facteurs cachés.

Les coefficients de corrélation obtenus pour les variables quantitatives sont résumés dans la matrice de corrélation suivante :

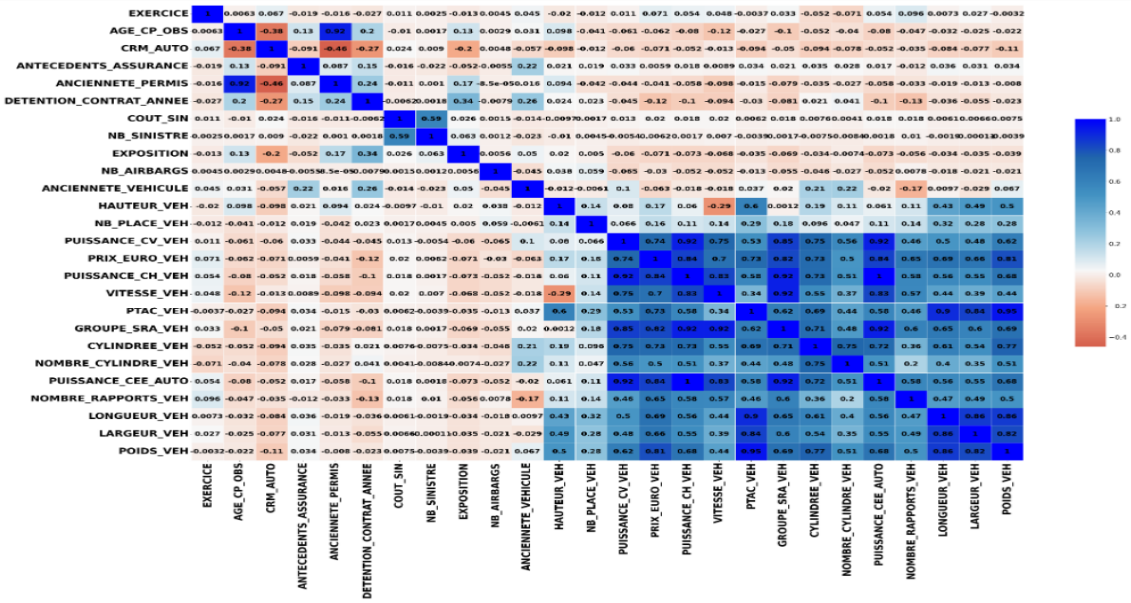

Figure 2.7: Matrice de corrélation des variables quantitatives

La matrice de corrélation présentée ci-dessus fournit une vue d'ensemble des relations linéaires entre différentes variables utilisées dans l'étude.

Les variables liées aux caractéristiques des véhicules montrent des corrélations particulièrement élevées entre elles. On observe par exemple une forte corrélation positive entre `PUISSANCE_CV_VEH` (Puissance en chevaux du véhicule) et `PRIX_EURO_VEH` (Prix du véhicule en euros), avec un coefficient de 0,84 indiquant que des véhicules plus puissants sont également plus chers. La variable `PTAC_VEH` (Poids Total Autorisé en Charge du véhicule) est fortement corrélée avec les variables `LONGUEUR_VEH` (0,9) et `LARGEUR_VEH` (0,84), ce qui est logique car un véhicule plus grand a tendance à avoir un PTAC plus élevé.

La corrélation observée entre les variables véhiculières souligne l'importance de considérer ces variables de manière combinée plutôt qu'individuellement. Par exemple, lors de la segmentation des risques ou de la tarification, il serait pertinent d'inclure des combinaisons de ces variables en une variable comme le véhiculier pour obtenir une évaluation plus précise. L'identification de ces corrélations permet également d'éviter la redondance des informations. Lors de la construction de modèles prédictifs, il est essentiel de choisir des variables non redondantes pour améliorer l'efficacité et la précision des modèles.

La figure suivante fait un zoom sur la corrélation des variables véhiculières afin de montrer les coefficients de corrélation particulièrement élevés sur ces variables.

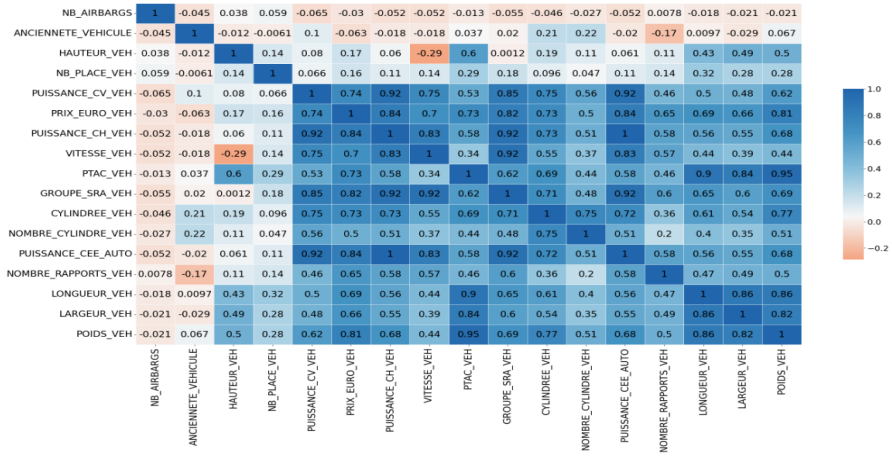

Figure 2.8: Matrice de corrélation des variables quantitatives

Analyse de la corrélation entre les variables qualitatives

La statistique Chi-deux (χ^2) est une mesure utilisée pour évaluer l'indépendance entre deux variables qualitatives. Cette statistique compare les effectifs observés dans une table de contingence avec les effectifs attendus si les variables étaient indépendantes. Chaque cellule du tableau de contingence contient le nombre d'observations correspondant à une combinaison spécifique des modalités des deux variables. La valeur du Chi-deux est donnée par ma formule suivante :

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - A_{ij})^2}{A_{ij}}$$

où :

- O_{ij} est l'effectif observée dans la cellule de la i-ème ligne et j-ème colonne du tableau de contingence,
- A_{ij} est l'effectif attendu dans la cellule de la i-ème ligne et j-ème colonne du tableau de contingence, calculée comme suit :

$$A_{ij} = \frac{R_i \times C_j}{n}$$

- R_i est le total des effectifs observés pour la i-ème ligne dans le tableau de contingence
- C_j est le total des effectifs observés pour la j-ème colonne dans le tableau de contingence,
- n est le nombre d'observations,

- r et c représentent respectivement le nombre de lignes et de colonnes du tableau de contingence.

Le test du χ^2 avec l'hypothèse nulle d'indépendance des deux distributions utilise cette statistique pour calculer une p-value permettant de conclure sur l'acceptation de l'hypothèse H_0. Nous rejetons alors l'hypothèse nulle d'indépendance des deux distributions au niveau de confiance de $(1-\alpha)$ si la p-value est inférieure à α.

Il est également possible d'employer une méthode plus précise pour mesurer la relation entre deux variables qualitatives en utilisant la notion du **V de Cramer**. Le V de Cramer mesure la force de l'association entre deux variables catégorielles. Il est calculé à partir de la statistique du Chi-deux et varie de 0 (aucune association) à 1 (association parfaite). La formule du V de Cramer est la suivante :

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

où $\min(c-1, r-1)$ est le nombre de degrés de liberté du test du Chi-deux.

La matrice suivante présente les valeurs du V de Cramer des variables qualitatives de la base de données.

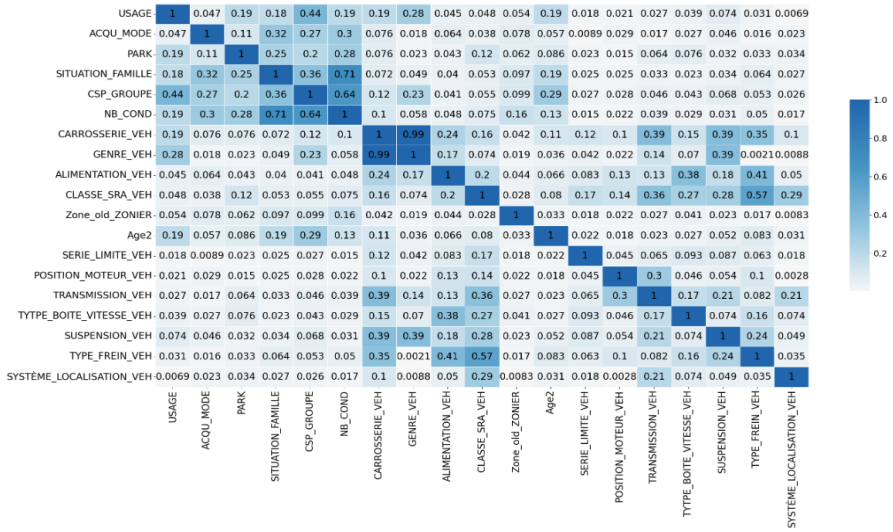

Figure 2.9: Matrice de V de Cramer des variables qualitatives

L'analyse de la matrice des valeurs du V de Cramer montre que les associations entre les variables qualitatives de la base de données sont majoritairement faibles, sauf l'association entre le genre du véhicule et sa carrosserie, qui est presque parfaite $(0,99)$. Nous testerons successivement l'entrée de ces deux variables dans les modèles afin de déterminer s'il faut garder les deux ou en conserver une seule en fonction des résultats obtenus.

Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) a pour but principal la représentation de n individus dans un sous-espace de k variables (généralement 2 ou 3) à partir d'un espace initial de p variables, avec $(k < p)$. L'ACP vise à trouver k nouvelles variables issues de la combinaison linéaire des variables d'origine, tout en conservant le maximum d'information initiale. Ces nouvelles variables sont appelées **composantes principales**. Algébriquement, minimiser la perte d'information revient à définir un sous-espace de dimension k de sorte que le nuage de points projeté conserve une inertie maximale. La formule mathématique de l'inertie totale est donnée par :

$$\mathrm{I}_g = \frac{1}{n}\sum_{i=1}^{n} \mathrm{d}^2\,(e_i, g) \quad \text{ou} \quad \mathrm{I}_g = \sum_{i=1}^{n} p_i\,\mathrm{d}^2\,(e_i, g)$$

où p_i représente le poids du i-ème individu lorsque les poids des individus sont différents, e_i est le vecteur des coordonnées du i-ème individu, g est le vecteur des coordonnées du centre de gravité des individus, et d désigne la distance considérée.

La valeur de l'inertie indique ainsi la dispersion totale du nuage de points. Par conséquent, sa valeur est également obtenue en faisant la somme des variances des variables initiales. Cependant, comme l'ACP utilise les variables centrées et réduites, l'inertie est alors équivalente au nombre de variables.
La détermination des nouvelles variables passe par la recherche d'**axes principaux**, qui doivent également conserver la même inertie maximale. Le nouveau repère, dans lequel le nuage de points est projeté, est défini par ces axes principaux, classés par ordre d'importance en fonction de l'inertie expliquée. Ce classement est effectué selon les valeurs propres de la matrice de variance-covariance Γ. Le premier axe principal correspond ainsi au vecteur propre normé associé à la plus grande valeur propre de la matrice Γ, et de même pour les axes suivants. La première composante principale est alors déterminée par les coordonnées de la projection des individus sur le premier axe principal, et cette définition s'applique également aux autres composantes principales suivantes.
Les autres notions importantes en ACP seront expliquées à travers les résultats de l'ACP que nous avons appliquée aux variables quantitatives véhiculières de notre base de données.
Le premier objectif étant de réduire l'espace de représentation des données, nous présentons initialement le pourcentage de variance expliquée par les dix premiers axes principaux. La variance expliquée par la i-ème composante principale est donnée par la valeur propre qui lui est associée. Ainsi, la part d'inertie ou de variance expliquée par l'axe i est calculée selon la formule suivante :

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots + \lambda_i + \lambda_{i+1} + \ldots + \lambda_r}$$

avec λ_i la valeur propre associée à l'axe principal i et r le nombre d'axes principaux.

L'histogramme suivant montre les pourcentages de variance expliquée de l'ACP réalisée.

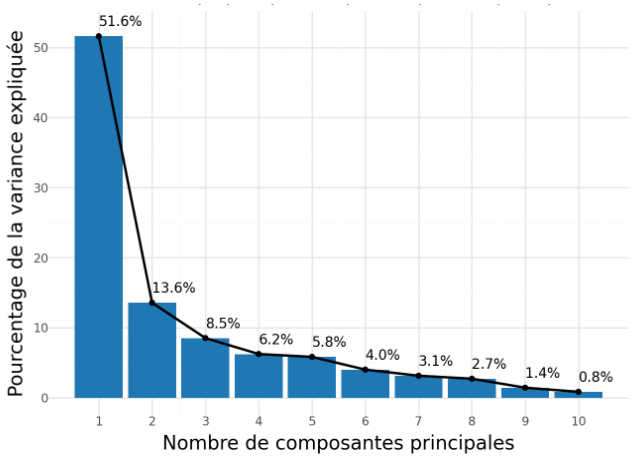

Figure 2.10: Pourcentage de variance expliquée par les composantes principales

Les résultats de cette section sont réalisés avec la fonction `PCA` du package `sklearn`.

La première composante principale explique 51,6% de la variance totale, indiquant qu'elle capture plus de la moitié de l'information initiale contenue dans les données. Cette proportion élevée souligne l'importance de cette composante. La deuxième composante principale explique 13,6% de la variance. Bien que ce chiffre soit inférieur à celui de la première composante, il reste important et contribue significativement à l'explication de la variance. On observe ensuite une diminution progressive de la variance expliquée par les composantes principales successives. Cette tendance est typique en ACP, où chaque composante principale successive explique une part de variance de plus en plus réduite. Les deux premières composantes principales expliquent 65,2% de la variance totale, indiquant qu'elles capturent la majeure partie de l'information présente dans les données. Le choix du nombre de composantes peut être déterminé en appliquant le **critère du coude**, qui consiste à identifier le point où la courbe de la variance expliquée commence à se stabiliser. Appliqué à notre graphique, ce critère renforce le choix des deux premiers axes pour la projection du nuage de points.

Le deuxième objectif de l'ACP dans cette étude est la détermination des variables qui sont représentatives des données étudiées. Les deux premiers axes ayant été choisis comme espace de projection du nuage de points, les variables les plus importantes sont celles qui sont en proximité des axes principaux. Cette proximité est mesurée par le coefficient de corrélation linéaire entre les variables et les composantes principales. Le cercle de corrélation présenté ci-dessous montre la représentation des variables sur les axes principaux.

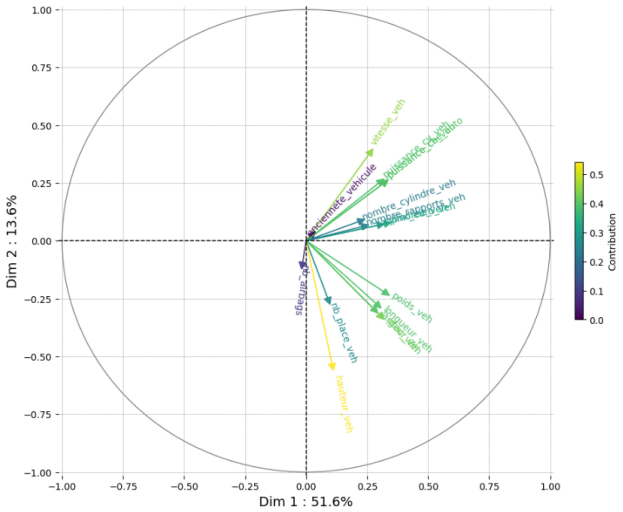

Figure 2.11: Cercle de corrélations

Le cercle de corrélation permet d'identifier deux catégories principales de variables relatives aux véhicules. La première catégorie concerne les performances du véhicule, mesurées par des variables telles que la vitesse, la puissance ou le nombre de cylindres. En revanche, on distingue également une catégorie relative aux caractéristiques dimensionnelles et de masse du véhicule, déterminées par des variables telles que la largeur, la longueur ou le poids.
Les variables liées au nombre d'airbags et à l'ancienneté du véhicule ne s'intègrent dans aucune des catégories mentionnées. En effet, ces variables, situées près de l'origine du cercle, sont les moins représentatives des deux axes principaux. Ainsi, elles ne contribuent pas significativement à la variation des données selon ces deux axes. Ces variables seront testées successivement dans les modèles afin de déterminer s'il faut les garder, en conserver une d'entre elles, ou n'en conserver aucune en fonction des performances observées.

2.3 Étude des coûts de sinistres extrêmes

Les valeurs extrêmes, aussi appelées queues de distribution, désignent des événements rares mais à fort impact situées aux extrémités d'une distribution statistique. En assurance automobile, ces valeurs extrêmes correspondent aux sinistres les plus coûteux. Bien qu'ils soient peu fréquents, ces sinistres peuvent engendrer des coûts considérables. Les assureurs doivent prendre en compte ces coûts potentiels pour fixer des primes justes et compétitives, évitant ainsi une sous-estimation des sinistres extrêmes, ce qui entraînerait des pertes financières importantes, ou une surestimation rendant les primes prohibitives pour les clients.

Pour gérer ces sinistres graves, les assureurs déterminent un seuil au-delà duquel les

montants des sinistres sont considérés comme des valeurs extrêmes. Cette détermination se base souvent sur l'expérience et le jugement professionnel. Dans cette étude, nous allons utiliser la théorie des valeurs extrêmes pour déterminer ce seuil. La détermination du seuil suppose la présence de valeurs extrêmes dans la distribution. Les graphiques suivants illustrent les valeurs extrêmes dans la distribution des coûts des sinistres.

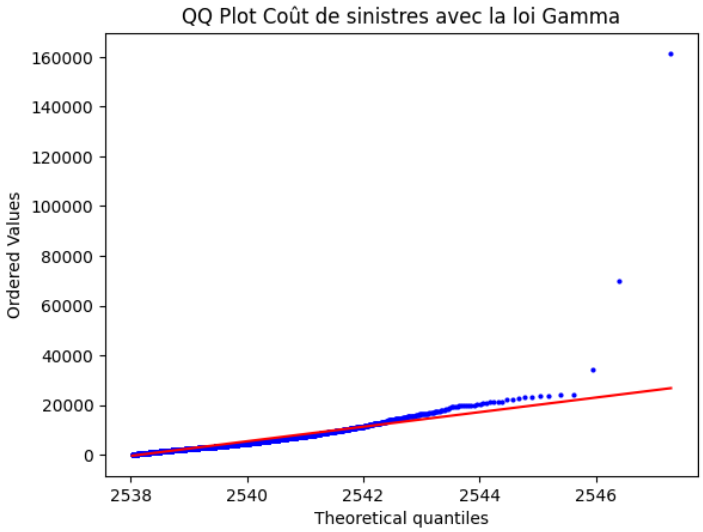

Figure 2.12: Mise en évidence des valeurs extrêmes

Le graphique `QQ Plot`, a été conçu avec la fonction `probplot` du package `scipy.stats`. Les données de l'axe des abscisses du `QQ Plot` ont donc été modifiées.

L'analyse de la figure 2.1 de la section du chapitre 2 sur l'analyse univariée montre que les points situés au-dessus du boxplot représenteraient des valeurs extrêmes. Ces valeurs sont nettement supérieures à la majorité des données, indiquant l'existence de sinistres très coûteux. Cette hypothèse est confirmée par l'analyse du `QQ Plot`, où l'on observe que les valeurs élevées s'écartent de la distribution théorique de la loi Gamma.

La théorie des valeurs extrêmes vise principalement à ajuster la queue de distribution à une loi particulière permettant certaines estimations. Elle propose deux méthodes :

- L'ajustement des maxima par blocs à la loi généralisée des extrêmes (GEV).
- La méthode du dépassement des seuils (POT) utilisant la loi généralisée de Pareto (GPD).

Nous utiliserons la méthode POT pour déterminer le seuil des sinistres graves.

Étant donnée une variable aléatoire X, les événements extrêmes sont sélectionnés en fixant un seuil s suffisamment élevé et en retenant les valeurs de X qui dépassent ce

seuil. La loi des excès résiduels $X_s = [X - s|X > s]$ est alors définie et sa fonction de répartition est :

$$F_{X_s}(y) = \mathbb{P}[X - s < y \mid X > s] = 1 - \frac{\overline{F}_X(s+y)}{\overline{F}_X(s)}$$

avec $y > 0$ et $\overline{F}_X$ la fonction de survie de la loi de X

Avec un nombre de données et un seuil s tendant vers l'infini, le théorème de Pickands assure que F_{X_s} suit une loi GPD dont la fonction de répartition est :

$$G(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}$$

où ξ et σ sont les paramètres de la loi.

Le comportement de G dépend du signe de ξ :

- Si $\xi < 0$, les quantiles associés de la loi GPD sont bornés par $s - \sigma/\xi$, indiquant une absence de queue de distribution.
- Si $\xi = 0$, la distribution converge vers la loi exponentielle de paramètre $1/\sigma$, indiquant une queue de distribution légère.
- Si $\xi > 0$, les quantiles augmentent de plus en plus, similaire à la loi de Fréchet, indiquant une queue de distribution épaisse.

La fixation du seuil s rencontre deux contraintes opposées. Si s est trop petit, les valeurs ne sont pas extrêmes et la densité de probabilité de l'échantillon ne s'approche pas d'une loi de Pareto. Si s est trop grand, il y a peu de données pour ajuster les données à la loi GPD. Une méthode pour optimiser le choix de s consiste à examiner l'espérance de la variable X_s, appelée moyenne des excès. Comme X_s suit une loi GPD, son espérance est :

$$\mathbb{E}[Y] = \frac{\sigma_s}{1 - \xi}$$

Le coefficient σ_s dépend du choix de s.
La stabilité de la loi GPD permet en particulier d'établir que pour tout $s > s_0$:

$$\mathbb{E}[X - s \mid X > s] = \frac{\sigma_s}{1 - \xi} = \frac{\sigma_{s_0} + \xi(s - s_0)}{1 - \xi} \quad \text{avec} \quad \xi < 1$$

Cette relation montre que pour tout seuil $s > s_0$, $\mathbb{E}[X - s \mid X > s]$ doit être une fonction linéaire de s. On trace donc la courbe de la fonction $g(s) = \mathbb{E}[X - s \mid X > s]$ et on recherche le seuil à partir duquel la fonction g devient linéaire. La linéarité de cette courbe permet de déterminer la valeur de ξ. Lorsque g ne varie pas quand s croît (allure

horizontale linéaire), la loi de Gumbel s'ajuste correctement aux données extrêmes avec $\xi \approx 0$.

La figure suivante montre la fonction des excès moyens.

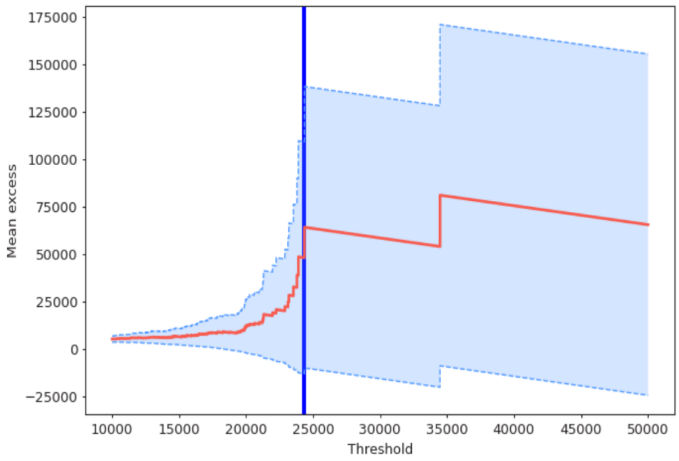

Figure 2.13: Courbe de la fonction des excès moyens

La fonction `plot_mean_residual_life` du package `pyextremes` a été utilisée pour construire ce graphique.

La visualisation de la courbe de la fonction des excès moyens montre que l'allure horizontale commence à s'observer au niveau du seuil $s = 24300$. Nous choisissons ce point car la loi des coûts de sinistres s'ajuste bien à la loi Gamma en dehors des valeurs extrêmes (voir le graphique du `QQ Plot` de la figure 2.12), et la loi Gamma a une queue de distribution légère comme la loi de Gumbel. Les montants de sinistres au-delà du seuil choisi $s = 24300$ représentent 1.48% du montant total des sinistres. Ces montants de sinistres ne seront pas pris en compte dans la modélisation du coût moyen des sinistres. Mais elles seront ensuite écrêtées sur l'ensemble de la base de données pour le calcul de la prime annuelle.

Résumé du chapitre

L'analyse exploratoire menée dans ce chapitre a permis de comprendre en détail le processus de construction de la base de données utilisée dans cette étude. Une fois cette base obtenue, il a été essentiel de réaliser des analyses des variables afin de se familiariser avec elles et de mieux les comprendre. Ces analyses ont également permis d'identifier les diverses relations entre les variables. Nous avons aussi concentré nos efforts sur la recherche de potentielles variables à ne pas inclure dans les modélisations futures. Cet ensemble de travaux constitue une base solide pour aborder le premier objectif de construction du véhiculier, présenté dans le chapitre suivant.

Chapitre 3

Construction du véhiculier et impacts sur la tarification

Ce chapitre est dédié à la construction du véhiculier en utilisant la base de données construite, les modèles linéaires généralisés (GLM), les techniques d'apprentissage statistique et de classification non supervisée. Nous commencerons par l'ajustement d'un modèle GLM de fréquence de sinistres et de coût moyen sans variables véhiculières, en choisissant les lois de probabilité et les fonctions de lien appropriées. Ensuite, nous extrairons les résidus que nous modéliserons avec les modèles de machine learning grâce aux variables véhiculières et nous effectuerons ensuite une classification des résidus prédits qui constitue le véhiculier. Nous évaluerons enfin l'impact de ce véhiculier sur la tarification en comparant les performances des modèles utilisant notre véhiculier personnalisé avec ceux utilisant le véhiculier standard de la SRA.

3.1 Présentation du Modèle Linéaire Généralisé

L'évaluation de la prime pure implique la nécessité de modéliser à la fois la fréquence de sinistres et leur sévérité. Le Modèle Linéaire Généralisé (GLM) se présente comme l'outil le plus couramment utilisé pour ajuster ces deux composantes. Leur développement a été initié par Nelder et Wedderburn en 1972. Le modèle GLM représente une extension du modèle de régression linéaire, lequel sera présenté au préalable.

3.1.1 Modèle de régression linéaire

Le modèle de régression linéaire s'exprime généralement par l'écriture suivante :

$$Y_i = \theta_0 + \theta_1 x_i^1 + \theta_2 x_i^2 + \ldots + \theta_p x_i^p + \epsilon_i \tag{3.1}$$

Les variables x^j sont observées, non aléatoires et les variables $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ sont des variables aléatoires i.i.d de loi normale $\mathcal{N}(0, \sigma^2)$ avec σ^2 inconnue. Les termes $\theta_0, \theta_1, \ldots, \theta_p$ sont les paramètres inconnus du modèle et doivent être estimés, ce qui constitue l'objectif

principal du modèle. Les variables Y_i sont alors des variables aléatoires i.i.d et suivant une loi normale.
Nous disposons ensuite de la variable à expliquer y, réalisation de la variable aléatoire $Y = (Y_1, Y_2, \ldots, Y_n)$. Nous disposons également de la matrice X de taille $\mathbb{R}^{n\times(p+1)}$. La première colonne de la matrice X représente l'observation de n valeurs de la variable constante prenant la valeur $\mathbf{1}$, tandis que les p colonnes suivantes représentent respectivement n observations pour chacune des variables explicatives x^j.

$$y = X\theta + e \tag{3.2}$$

avec $\theta = (\theta_0, \theta_1, \ldots, \theta_p)^T$[1] formant le vecteur des paramètres du modèle et e, le vecteur de n réalisations de la loi normale $\mathcal{N}(0, \sigma^2)$.

L'estimation des paramètres du modèle peut être réalisée par la Méthode des Moindres Carrés (MCO) ou la Méthode du Maximum de Vraisemblance. La MCO consiste à minimiser la somme des carrés des erreurs sans nécessiter d'hypothèse de normalité sur ces résidus. Le problème de minimisation est formulé comme suit :

$$\inf_{\theta\in\mathbb{R}^{(p+1)}} \|e\|_2^2 = \inf_{\theta\in\mathbb{R}^{(p+1)}} \|y - X\theta\|_2^2 \tag{3.3}$$

La méthode du Maximum de Vraisemblance repose sur l'hypothèse de normalité des résidus, et l'estimateur est obtenu en résolvant le système d'équations constitué des équations suivantes :

$$\frac{\partial \ln \mathcal{L}(\theta|y)}{\partial \theta_i} = 0, \quad \text{pour } i = 0, 1, \ldots, p, \quad \text{avec} \quad \mathcal{L}(\theta|y) = \prod_{i=1}^{n} f(y_i, \theta) \tag{3.4}$$

Les deux méthodes fournissent le même estimateur pour θ. Notant $\hat{\theta}$ l'estimateur de θ et $\hat{\theta}(y)$ une réalisation de cet estimateur, nous avons :

$$\hat{\theta} = (X'X)^{-1}X'Y \quad \text{et} \quad \hat{\theta}(y) = (X'X)^{-1}X'y \tag{3.5}$$

L'estimateur $\hat{\theta}$ est donc sans biais et suit une loi normale $\mathcal{N}(\theta, \sigma^2(X'X)^{-1})$.
La qualité d'ajustement du modèle aux données est mesurée par le critère R^2, défini comme :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad \text{avec } \overline{y} = \sum_{i=1}^{n} y_i \tag{3.6}$$

Ce critère, également appelé coefficient de détermination, représente la part de variation de y expliquée par le modèle ajusté. Un modèle est considéré comme meilleur lorsque ce critère est proche de 1. Cependant, le coefficient de détermination augmente avec le

[1] Dans la suite, la transposé du vecteur u (resp. de la matrice M) sera notée u^T (resp. M^T)

nombre de variables. Le coefficient de détermination ajusté R_a^2 a été introduit dans ce but et s'exprime comme suit :

$$R_a^2 = \frac{(n-1)R^2 - p}{n-p-1} \tag{3.7}$$

3.1.2 Modèle Linéaire Généralisé

Le modèle de régression linéaire est contraint par des hypothèses qui doivent être satisfaites pour une utilisation valide. Ces hypothèses incluent la linéarité de la relation entre la variable à expliquer et les variables explicatives, la normalité de la variable à expliquer, la constance de sa variance et l'indépendance entre les variables Y_i. Cependant, ces conditions ne sont pas toujours vérifiables, en particulier dans le contexte de la modélisation de la fréquence et de la sévérité. Le modèle GLM a été développé pour contourner ces trois premières hypothèses nécessaires à l'utilisation du modèle de régression linéaire.

En reprenant les mêmes notations du modèle linéaire, le modèle GLM peut être exprimé sous la forme suivante :

$$g(\mathbb{E}[Y_i]) = \theta_0 + \theta_1 x_i^1 + \theta_2 x_i^2 + \ldots + \theta_p x_i^p \tag{3.8}$$

La fonction g, appelée fonction de lien, permet d'introduire une non-linéarité entre les variables explicatives et la variable à expliquer. La variable aléatoire Y ne suit plus nécessairement une loi normale, mais elle doit appartenir à un groupe spécifique de lois de probabilité appelé famille exponentielle.

Famille exponentielle et choix de la loi de probabilité

La **famille exponentielle** concerne les lois usuelles à l'exemple de la loi Normale, la loi Gamma, la loi de Poisson, la loi de Bernoulli. Nous allons représenter les variables aléatoires par leur fonction de densité de probabilité notée f_Y et par leur loi de probabilité donnée par $\mathbb{P}_Y$ si Y n'admet pas de densité. Ainsi, la variable aléatoire Y appartient à la famille exponentielle si sa fonction de densité ou sa loi de probabilité s'écrit sous la forme exponentielle que nous présentons ci-dessous. Nous allons donner la forme exponentielle admise par la fonction de densité d'une variable aléatoire appartenant à cette famille exponentielle. Cette forme doit être également admise par la loi de probabilité donnée par la quantité $\mathbb{P}_Y(Y = y)$.

$$f_Y(y, \omega, \phi) = \exp\left(\frac{1}{\gamma(\phi)}(y\omega - b(\omega)) + c(y, \phi)\right) \tag{3.9}$$

Sous cette forme, la fonction c est une fonction dérivable, b est une fonction dérivable d'ordre 3 et sa dérivée première notée b' admet une fonction réciproque. Le paramètre ω (resp. ϕ) est appelé paramètre naturel (resp. paramètre de nuisance ou de dispersion) de la loi.

Si on définit $\mathbb{E}[Y] = \mu$, et si Y appartient à la famille exponentielle, nous avons les propriétés suivantes :

- $\mathbb{E}[Y] = \mu = b'(\omega)$
- $\mathbb{V}[Y] = b''(\omega)\gamma(\phi)$, b'' est la dérivée seconde de la fonction b

Dans le cadre du modèle GLM, le choix de la loi de probabilité issue de la famille exponentielle revêt une importance cruciale. Ce choix est guidé par la nature de la distribution de la variable à expliquer. Par exemple, dans le contexte de la modélisation de la durée de vie, la loi exponentielle est fréquemment adoptée. De même, pour une variable à expliquer de nature binaire, la loi de Bernoulli est préférée, tandis que la loi de Poisson est privilégiée pour la modélisation de données de comptage.

Choix de la fonction de lien

A l'instar de la loi de probabilité de la variable à expliquer, le modèle est défini par la composante déterministe et la fonction de lien.

La **composante déterministe** représente le prédicteur linéaire du modèle GLM que nous noterons dans la suite par η. Il est défini par $\eta = X\theta$.

La **fonction de lien** g est une fonction réelle, monotone et différentiable, définie telle que la forme matricielle du modèle GLM s'écrive de la manière suivante :

$$g(\mathbb{E}[Y]) = \eta \quad \text{ou} \quad g(\mu) = \eta \tag{3.10}$$

Chaque loi de probabilité appartenant à la famille exponentielle admet une fonction de lien canonique. Cette fonction de lien canonique est définie de sorte que g soit la fonction qui vérifie l'égalité : $g(\mu) = \omega$. D'après les propriétés définies ci-dessus, $\mu = b'(\omega)$ et on a donc $b'^{-1}(\mu) = \omega$ puisque b' admet une fonction réciproque. En conclusion, la fonction de lien canonique est donc celle qui vérifie l'égalité : $g(\mu) = b'^{-1}(\mu)$.

En utilisant la loi normale avec sa fonction de lien canonique identité, on retrouve le modèle linéaire classique. La fonction de lien canonique de la loi de Bernoulli est la fonction logit donnée par $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$ tandis que la fonction logarithmique définie par $g(\mu) = \ln(\mu)$ est la fonction de lien canonique de la loi de Poisson. En ce qui concerne la loi Gamma, sa fonction de lien canonique est la fonction inverse, exprimée par $g(\mu) = \frac{1}{\mu}$.

Le choix de la fonction de lien dépend des objectifs visés. La fonction de lien canonique est le choix par défaut en pratique si aucun objectif de modélisation n'implique d'utiliser d'autres fonctions de lien.

Estimation des paramètres du modèle

L'objectif du modèle, une fois la loi de la famille exponentielle et la fonction de lien définies, consiste à déterminer les estimateurs des paramètres du modèle.
Nous adoptons la méthode du maximum de vraisemblance pour ce faire, en considérant la fonction de lien g comme la fonction de lien canonique. En prenant une réalisation $y = (y_1, y_2, \ldots, y_n)$ de la variable aléatoire $(Y_1, Y_2, \ldots, Y_n)$ dont les composantes sont des variables aléatoires i.i.d suivant la loi de la famille exponentielle de densité f_Y, nous définissons la vraisemblance comme suit :

$$\mathcal{L}(y, \theta, \phi) = \prod_{i=1}^{n} f(y_i, \omega_i, \phi) = \prod_{i=1}^{n} f(y_i, x_i\theta, \phi), \quad \text{car} \quad \omega_i = g(\mu) = x_i\theta \tag{3.11}$$

avec x_i, le vecteur défini par $x_i = (1, x_i^1, x_i^2, \ldots, x_i^p)$
La log-vraisemblance est donc donnée par :

$$l(y, \theta, \phi) = \sum_{i=1}^{n} \left[\frac{1}{\gamma(\phi)} (y_i x_i \theta - b(x_i\theta)) + c(y_i, \phi) \right] \tag{3.12}$$

Les paramètres θ et ϕ du modèle sont obtenus en résolvant le système d'équations défini par :

$$\begin{cases} \dfrac{\partial l(y, \theta, \phi)}{\partial \theta_j} = 0, \quad \text{pour } j = 0, 1, \ldots, p \\ \dfrac{\partial l(y, \theta, \phi)}{\partial \phi} = 0 \end{cases} \tag{3.13}$$

La résolution se fait en déterminant d'abord l'estimateur de θ grâce aux $p+1$ premières équations du système, puis en utilisant cet estimateur pour retrouver celui de ϕ à l'aide de la dernière équation du système.

Ainsi, nous avons :

$$\frac{\partial l(y, \theta, \phi)}{\partial \theta_j} = 0 \Rightarrow \sum_{i=1}^{n} x_i^j \left[\frac{1}{\gamma(\phi)} (y_i - b'(x_i\theta)) \right] = 0 \tag{3.14}$$

Nous obtenons alors l'équation suivante :

$$\sum_{i=1}^{n} x_i^j \left[y_i - b'(x_i\theta) \right] = 0 \quad \text{avec pour} \quad j = 0, \quad x_i^j = 1 \quad \forall \quad i \tag{3.15}$$

L'équation (3.15) montre qu'on peut retrouver le modèle de régression linéaire si la loi de probabilité considérée est la loi normale. En effet, la fonction b' de la loi normale est une fonction identité et comme $b'(x_i\theta) = x_i\theta$, on retrouve exactement l'équation du maximum de vraisemblance permettant de déterminer l'estimateur $\hat{\theta}$ du modèle de

régression linéaire.

Lorsque la fonction b' n'est pas une fonction identité, l'équation (3.15) n'est pas linéaire en θ, et pour retrouver l'estimateur de θ, on utilise des algorithmes d'optimisation qui sont itératifs, comme l'algorithme IRLS (Iteratively Re-weighted Least Squares) de Nelder et Wedderburn.

Dès lors que nous obtenons l'estimateur $\hat{\theta}$, nous pouvons calculer un estimateur $\hat{\omega}_i$ pour ω_i en le définissant comme $\hat{\omega}_i = x_i\hat{\theta}$. Cela nous permet alors de prédire la valeur de y_i au point x_i en utilisant l'estimateur de la moyenne : $\hat{\mu}_i = g^{-1}(x_i\hat{\theta})$

Propriétés et intervalle de confiance de l'estimateur du maximum de vraisemblance

L'estimateur $\hat{\theta}$ du maximum de vraisemblance (EMV) vérifie sous certaines hypothèses de régularité, les propriétés suivantes :

- Convergence en probabilité de $\hat{\theta} \xrightarrow{\mathbb{P}} \theta$
- $\hat{\theta}$ est asymptotiquement sans biais
- Normalité asymptotique de $\hat{\theta}$: $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{Loi} \mathcal{N}(0, I^{-1}(\theta))$
  avec $I(\theta) = -\frac{1}{n}\mathbb{E}\left[\frac{\partial^2 l(y,\theta,\phi)}{\partial\theta\partial\theta'}\right]$, la matrice de l'information de Fisher

En utilisant la fonction de lien canonique, nous avons :

$$I_n(\theta) = \frac{1}{\gamma(\phi)} X'\mathbb{V}[Y]X \tag{3.16}$$

avec $I_n(\theta) = nI(\theta)$

Nous allons utiliser la dernière propriété de normalité asymptotique pour construire un intervalle de confiance pour θ_j.

La normalité asymptotique est toujours vérifiée si on évalue la matrice de l'information de Fisher au point $\hat{\theta}$. En effet on a : $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{Loi} \mathcal{N}(0, I^{-1}(\hat{\theta}))$.

Nous pouvons alors déterminer un intervalle de confiance asymptotique de niveau confiance $(1-\alpha)$ de chaque paramètre θ_j. L'intervalle de confiance est donné par l'expression suivante :

$$IC_{1-\alpha}(\theta_j) = \left[\hat{\theta} - q_{1-\frac{\alpha}{2}}\sqrt{v_{jj}} \quad ; \quad \hat{\theta} + q_{1-\frac{\alpha}{2}}\sqrt{v_{jj}}\right] \tag{3.17}$$

où v_{jj} est la variance de $\hat{\theta}_j$, donnée par le $j^{\text{ème}}$ élément de la diagonale de la matrice $I_n^{-1}(\hat{\theta})$, et $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi normale $\mathcal{N}(0,1)$.

Notion de déviance

Nous désignons par M_S le modèle saturé, qui correspond au modèle le plus complet comportant autant de variables explicatives que d'observations. La déviance d'un modèle M est fonction de la différence entre la log-vraisemblance du modèle M et celle du modèle M_S, définie par :

$$D(M) = -2\left[l(y,\hat{\theta}) - l(y,\hat{\theta}_S)\right]$$

Le modèle M serait alors considéré comme mauvais si sa déviance est grande.

Le modèle est de mauvaise qualité au niveau de confiance $(1-\alpha)$ en pratique si la valeur observée de la déviance est supérieure au quantile d'ordre $(1-\alpha)$ de la loi du Chi-deux de degré de liberté $(n-(p+1))$, avec n le nombre d'observations et $(p+1)$ le nombre de paramètres du modèle.

Sélection de modèle

La sélection parmi plusieurs modèles non emboîtés se fait en fonction de la déviance, le modèle avec la plus petite déviance étant considéré comme le meilleur.
Le modèle ayant une déviance nulle est le modèle saturé M_S, tandis que celui avec la déviance la plus élevée est le modèle M_0, qui ne contient qu'un seul paramètre : l'intercept.
En pratique, nous cherchons un modèle avec le moins de paramètres possible et une déviance faible, proche de zéro. Ce modèle sera désigné comme le modèle optimal.
Le modèle optimal est choisi en utilisant des critères d'information, les plus connus étant le BIC (Bayesian Information Criterion) et l'AIC (Akaike Information Criterion). Ces critères pénalisent les modèles en tenant compte du nombre de paramètres et de la déviance.
Ils sont définis par les formules suivantes :

$$\mathbf{AIC} = 2(p+1) - 2l(y,\theta)$$
$$\mathbf{BIC} = (p+1)\ln(n) - 2l(y,\theta)$$

Le BIC pénalise plus sévèrement les modèles complexes en ajoutant un terme de pénalisation proportionnel au logarithme de la taille de l'échantillon. En effet, lorsque $\ln(n) > 2$, la pénalisation du BIC est plus forte que celle de l'AIC. L'utilisation du BIC favorise ainsi les modèles avec moins de variables (ou de paramètres), dits "parcimonieux". L'AIC pénalise les modèles complexes de manière moins sévère que le BIC, ce qui peut parfois entraîner la sélection de modèles plus complexes. En favorisant les modèles plus simples, le BIC adopte une approche plus prudente dans la sélection de modèles, contribuant ainsi à éviter le surajustement. Le choix entre le BIC et l'AIC n'est donc pas simple et dépend du contexte spécifique de l'analyse ainsi que des objectifs de modélisation. Cependant, quel que soit le critère choisi, le modèle optimal est celui qui présente la plus petite valeur de l'AIC ou du BIC.

Sélection de variables

La sélection de variables peut être réalisée à l'aide d'algorithmes appelés "méthodes pas à pas". Ces méthodes se basent sur un critère spécifique et sélectionnent les variables susceptibles de constituer le modèle optimal. Elles se déclinent en trois types :

1. **Méthode Backward ou méthode d'élimination en arrière :**

   Cette méthode débute avec un modèle incluant toutes les variables disponibles. L'algorithme procède ensuite à l'élimination de la variable la moins importante. Cette élimination se base généralement sur la valeur de la p-value, en retirant la variable dont la p-value associée est la plus élevée. Cette opération se répète jusqu'à ce que toutes les p-values soient inférieures à 5%.

   Si l'on utilise les critères d'information AIC ou BIC, la variable supprimée est celle dont l'élimination entraîne la plus grande diminution de la valeur de l'AIC ou du BIC. Le processus s'arrête lorsque la valeur de l'AIC ou du BIC cesse de diminuer.

2. **Méthode Forward ou méthode de sélection avant :**

   La méthode Forward consiste à partir d'un modèle vide ne contenant que l'intercept. À chaque étape, l'algorithme ajoute la variable la plus importante qui améliore le modèle selon le critère d'information choisi. Le processus s'arrête lorsque le modèle ne s'améliore plus selon ce critère.

   Cette méthode présente l'avantage d'être plus économique car elle permet de limiter le nombre de variables utilisées au strict nécessaire. Cependant, une fois qu'une variable est ajoutée, elle est conservée, ce qui peut entraîner la présence de variables non significatives dans le modèle final. La troisième méthode permet de remédier à cet inconvénient.

3. **Méthode de sélection mixte :**

   Cette méthode combine les deux approches précédentes. L'algorithme débute avec le modèle nul. Lors de la première étape, une sélection avant est effectuée. Aux étapes suivantes, une sélection avant est suivie d'une élimination en arrière. La méthode Stepwise peut également être appliquée dans le sens inverse, en débutant avec un modèle contenant toutes les variables. On effectue alors une élimination en arrière à la première étape, puis une élimination en arrière suivie d'une sélection en avant à chaque étape suivante.

Résumé intermédiaire

La construction du véhiculier sera réalisée pour les deux modèles : le modèle de fréquence et le modèle de coût-moyen. La première étape consiste à ajuster les modèles sans inclure les variables explicatives liées aux véhicules. Les sections suivantes illustrent la construction du véhiculier et représentent le cœur de ce chapitre.

La figure suivante illustre le résumé des différentes étapes de la construction du véhiculier.

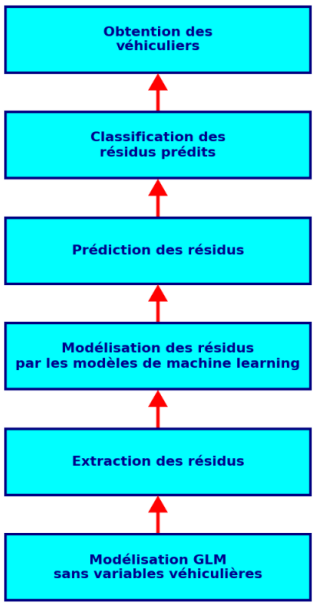

Figure 3.1: Etapes de construction du véhiculier

3.2 Ajustement du modèle de fréquence de sinistres et de coût moyen

Les modèles GLM seront utilisés dans cette première étape.

3.2.1 Choix de la loi de probabilité

Le premier objectif dans l'utilisation d'un modèle GLM est le choix de la loi de probabilité appropriée. Deux lois de la famille exponentielle ont été considérées pour chaque modèle afin de déterminer la meilleure option.

Pour la fréquence de sinistres, la loi de Poisson et la loi Binomiale Négative ont été sélectionnées pour ajuster le modèle GLM.
Les lignes qui suivent démontrent l'appartenance de la loi de Poisson à la famille exponentielle.
La loi de probabilité de Poisson est donnée par :

$$P[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}$$

où λ est le paramètre de la loi.
Cette formule peut également être réécrite sous la forme suivante :

$$P[X = k] = \exp\left(k\ln(\lambda) - \lambda + \ln\left(\frac{1}{k!}\right)\right)$$

La loi de Poisson est donc une loi de la famille exponentielle associée aux paramètres suivants :

$$\begin{cases} \omega = \ln(\lambda) \\ b(\omega) = \exp(\omega) = \lambda \\ \gamma(\phi) = 1 \\ c(k,\phi) = \ln\left(\frac{1}{k!}\right) \end{cases}$$

En ce qui concerne la modélisation du coût moyen, les lois Gamma et Inverse Gaussienne ont été choisies. Nous démontrons également que la loi Gamma appartient à la famille exponentielle.
La fonction de densité de la loi Gamma est donnée par :

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}_{x>0}$$

où $\beta > 0$ et $\alpha > 0$ sont respectivement les paramètres d'échelle et de forme de la loi. Cette fonction de densité peut également être exprimée sous la forme suivante :

$$\begin{aligned} f(x) &= \exp\left(\alpha\ln(\beta) + (\alpha-1)\ln(x) - \beta x - \ln(\Gamma(\alpha))\right) \\ &= \exp\left(\alpha\ln(\alpha) - \alpha\ln(\alpha) + \alpha\ln(\beta) - \beta x + (\alpha-1)\ln(x) - \ln(\Gamma(\alpha))\right) \\ &= \exp\left(\alpha\ln\left(\frac{\beta}{\alpha}\right) - \beta x + \alpha\ln(\alpha) + (\alpha-1)\ln(x) - \ln(\Gamma(\alpha))\right) \\ f(x) &= \exp\left(-\alpha\left(\frac{\beta}{\alpha}x - \ln\left(\frac{\beta}{\alpha}\right)\right) + \alpha\ln(\alpha) + (\alpha-1)\ln(x) - \ln(\Gamma(\alpha))\right) \end{aligned}$$

La loi Gamma est donc également une loi de la famille exponentielle avec les paramètres suivants :

$$\begin{cases} \omega = \frac{\beta}{\alpha} \\ b(\omega) = \ln\left(\frac{\beta}{\alpha}\right) \\ \gamma(\phi) = -\frac{1}{\alpha} \\ c(x,\phi) = \alpha\ln(\alpha) + (\alpha-1)\ln(x) - \ln(\Gamma(\alpha)) \end{cases}$$

Avant d'être utilisées dans les modèles, les lois considérées doivent être ajustées à la distribution correspondante.
Dans le cas du modèle GLM de fréquence de sinistres, les lois de Poisson et Binomiale Négative ont été ajustées à la distribution de fréquence de sinistres afin de vérifier leur adaptation. L'ajustement de la loi de Poisson à la distribution de la fréquence nécessite

l'estimateur du paramètre λ. Afin de comparer la loi théorique de Poisson et la distribution, nous avons utilisé l'estimateur $\hat{\lambda}$ du maximum de vraisemblance (EMV) qui sert à simuler la loi théorique qui sera comparée à la distribution. L'estimateur EMV du paramètre λ est donné par la moyenne empirique. En effet nous avons :
La fonction de vraisemblance est :

$$L(k_1, \ldots, k_n; \lambda) = \prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} k_i}}{\prod_{i=1}^{n} k_i!}$$

La log-vraisemblance est définie par :
$\ln L(k_1, \ldots, k_n; \lambda) = -n\lambda + \sum_{i=1}^{n} k_i \ln(\lambda) - \sum_{i=1}^{n} \ln(k_i!)$.

La dérivée de la log-vraisemblance est : $\frac{\partial \ln L}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} k_i$

En annulant cette dérivée, nous trouvons l'estimateur du maximum de vraisemblance : $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} k_i$

La dérivée seconde permet de confirmer l'estimateur obtenu : $\frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{1}{\lambda^2} \sum_{i=1}^{n} k_i \leq 0$

Nous utilisons alors la moyenne empirique pour l'ajustement de la loi théorique à la distribution de la fréquence de sinistres. Cet estimateur est également renvoyé par le logiciel `R` lorsqu'on détermine le paramètre de la loi théorique de Poisson ajustée à la fréquence de sinistres. Les paramètres de la loi Binomiale Négative ajustée à la distribution ont également été obtenus par le logiciel `R`.
Ces paramètres sont ensuite utilisés pour générer le graphique d'adaptation des lois de Poisson et Binomiale Négative à la distribution empirique de la fréquence de sinistres avec le package `plot_ly` du logiciel `R`.

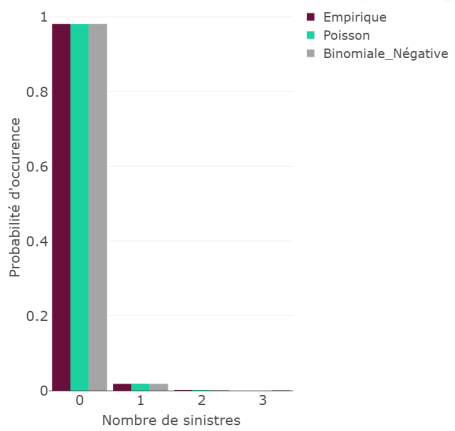

Figure 3.2: Ajustement des lois de Poisson et Binomiale Négative à la distribution de la fréquence de sinistres

La loi de Poisson est adéquate pour les sinistres rares où la variance est proche de la moyenne. Cependant, lorsqu'il y a une sur-dispersion, la loi Binomiale Négative est plus appropriée. La comparaison visuelle sur le graphique montre que la loi Binomiale Négative offre un ajustement plus précis des fréquences observées pour les valeurs de sinistres supérieures à 1. Pour les données de sinistres avec une plus grande variabilité, la loi Binomiale Négative est plus adaptée. Ce modèle tient compte de la sur-dispersion avec une variance plus grande que l'espérance. La loi de Poisson reste utile pour des ajustements de distribution simples de données n'affichant pas de sur-dispersion significative.

Le même exercice a été effectué pour le modèle de coût moyen. Les lois Gamma et Inverse Gaussienne ont été ajustées à la distribution du coût moyen et un graphe de comparaison des deux modèles a été généré avec le package `plotly` de `Python`.

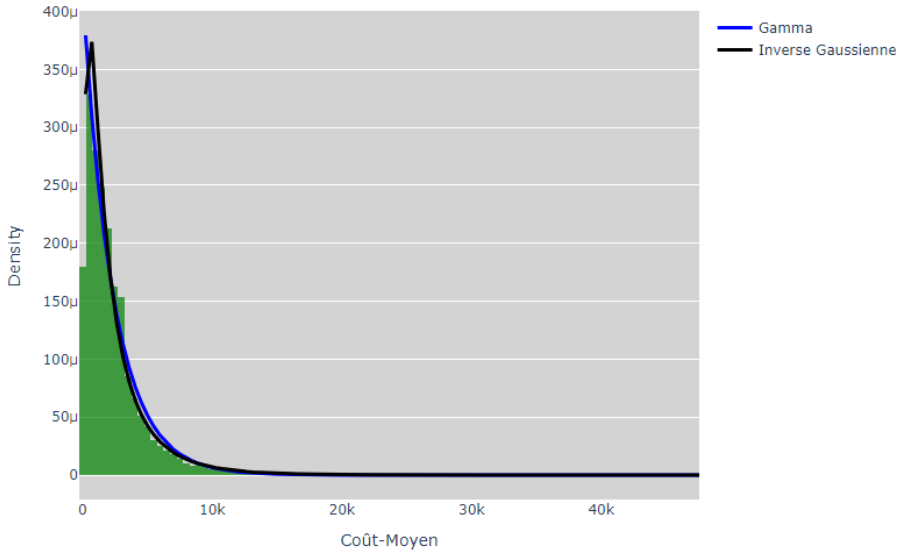

Figure 3.3: Ajustement des lois Gamma et Inverse Gaussienne à la distribution de coût moyen de sinistres

Les deux lois capturent bien la forme de la distribution empirique. Elles montrent une décroissance rapide de la densité à mesure que le coût moyen augmente, ce qui correspond bien aux données observées. Elles offrent donc toutes deux un bon ajustement aux données empiriques et leurs différences dans la modélisation des queues de distribution doivent alors être prises en compte. La queue de distribution de la loi Gamma est plus légère tandis que la loi Inverse Gaussienne admet une queue de distribution plus lourde. Cependant, les valeurs extrêmes ayant été supprimées de la distribution des coûts de sinistres, nous sommes désormais en présence de sinistres courants et les deux lois restent alors de bons choix pour l'ajustement de la distribution du coût moyen de sinistres.

Le choix de la loi du modèle GLM ne dépend pas uniquement de la variable à expliquer, il dépend aussi des variables explicatives qui ne sont pas utilisées dans les comparaisons effectuées dans cette sous-section. Le choix définitif se fera sur la base des résultats issus des modèles ajustés à partir de ces différentes lois.

3.2.2 Choix de la fonction de lien

Les modèles considérés sont les GLM avec les lois de Poisson, Gamma, Binomiale Négative et Inverse Gaussienne.
Le tableau suivant présente leurs fonctions de lien canonique :

Loi de probabilité	Fonction de lien canonique	Nom du lien canonique
Poisson	$g(\mu) = \ln(\mu)$	Log
Gamma	$g(\mu) = \frac{1}{\mu}$	Inverse
Binomiale négative	$g(\mu) = \ln(\mu)$	Log
Inverse Gaussienne	$g(\mu) = \frac{1}{\mu^2}$	Inverse du carré

Table 3.1: Fonctions de lien canonique

La fonction de lien canonique est choisie naturellement en raison de ses propriétés mathématiques qui simplifient les calculs et les inférences dans le cadre des modèles de régression généralisée.

L'utilisation de la fonction de lien logarithmique permet de modéliser les effets multiplicatifs de manière directe et intuitive. Chaque coefficient dans le modèle peut être interprété comme un multiplicateur de la fréquence ou du coût moyen, ce qui facilite la justification des tarifs. Dans le contexte de l'assurance non-vie, les primes tarifaires sont calculées en multipliant plusieurs facteurs, ce qui correspond à un modèle Log-GLM. **Ainsi, la fonction de lien logarithmique sera utilisée quel que soit le modèle GLM afin de disposer d'une structure de tarification multiplicative.**

3.2.3 Modélisation

Les techniques employées pour la modélisation des deux indicateurs de risque dans cette section ne sont pas présentées en exhaustivité, les autres techniques comme la prise en compte des interactions seront présentées dans le modèle GLM final après la construction du véhiculier, qui sera comparé au modèle additif neuronal du chapitre suivant.

Les variables retenues pour les modélisations dans cette première étape de la construction du véhiculier sont présentées dans le tableau 3.2.

Les assurés pour une année d'exercice donnée ne sont pas nécessairement observés sur toute la période d'exercice. L'objectif étant d'obtenir une prime annuelle, la prise en

	Fréquence de sinistres	***Coût moyen***
Variables	**Nombre de sinistres**	**Coût de sinistres**
	Exposition	**Nombre de sinistres**
	Âge du conducteur principal	Âge du conducteur principal
	Exercice	Exercice
	Usage du véhicule	Usage du véhicule
	Mode d'acquisition du véhicule	Mode d'acquisition du véhicule
	Mode de stationnement	Mode de stationnement
	Type de produits	Type de produits
	Situation familiale	Situation familiale
	CRM	CRM
	Antécédents d'assurance	Antécédents d'assurance
	Zonier	Zonier
	CSP	CSP
	Nombre de conducteurs	Nombre de conducteurs
	Ancienneté du permis	Ancienneté du permis
	Nombre d'années de détention du contrat	Nombre d'années de détention du contrat

Table 3.2: Table des variables utilisées pour la modélisation

compte de l'exposition permet d'estimer le nombre moyen de sinistres sur une année en effectuant un calcul au prorata temporis. En effet, pour un assuré observé uniquement sur la moitié de la période d'exercice et ayant eu un sinistre, la prise en compte du chiffre 1 comme nombre moyen de sinistres ne reflète pas la période annuelle car l'assuré n'a été observé que sur la moitié de l'exercice. On considère alors que si l'assuré avait été présent sur toute la période d'exercice, il aurait eu 2 sinistres au lieu d'un seul : on divise ainsi le nombre de sinistres par l'exposition qui équivaut à $0,5$ dans cet exemple.

Le modèle GLM doit donc prendre en compte cette notion d'exposition. Le modèle s'écrit alors sous la forme mathématique suivante :

$$\mathbb{E}[Y_i] = \exp(\theta_0 + \theta_1 x_i^1 + \theta_2 x_i^2 + \ldots + \theta_p x_i^p + \ln(Exposition_i)) \tag{3.18}$$

En pratique, ceci est réalisé en forçant le modèle à attribuer un coefficient de 1 au logarithme de la variable `Exposition`. Cette manipulation est effectuée en fournissant le logarithme de la variable `Exposition` à l'argument `offset` lors de l'ajustement du modèle.

La formule mathématique de la modélisation du coût moyen ne subit aucune modification. En pratique, nous spécifions dans l'ajustement du modèle le poids des observations, qui correspond à la variable `Nombre de sinistres`.

Les modèles étant définis, nous choisissons les bases d'apprentissage et de test. Nous

optons pour une base d'apprentissage et une base de test représentant respectivement 80% et 20% du nombre de lignes de la base initiale. Cependant, le choix de ces bases doit être aléatoire. Les données sont d'abord mélangées de manière aléatoire. Ensuite, un pourcentage spécifique (80% dans notre cas) des lignes de données est choisi aléatoirement pour former la base d'apprentissage. Les données restantes, non sélectionnées, constituent la base de test. Cette méthode permet de créer deux ensembles distincts : l'un pour l'entraînement du modèle et l'autre pour l'évaluation de ses performances. Dans cette étude, ce processus a été itéré plusieurs fois avec des graines de reproductibilité différentes. Les performances sur ces différentes bases ont été calculées grâce aux métriques MAE (Mean Absolute Error) et RMSE (Root Mean Square Error) définies ci-dessous :

- Le MAE mesure la moyenne des différences absolues entre les valeurs prédites par le modèle ($\hat{y}_i$) et les valeurs réelles (y_i). Il donne une idée de la valeur moyenne des erreurs de prédiction, sans tenir compte de la positivité ou de la négativité de ces erreurs, ce qui le rend facile à interpréter. Sa formule est :

$$\textbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  où n est le nombre total d'observations.

- La RMSE évalue la racine carrée de la moyenne des carrés des différences entre les valeurs prédites par le modèle ($\hat{y}_i$) et les valeurs réelles (y_i). Cette métrique est plus sensible aux grandes erreurs, car elle pénalise les erreurs les plus importantes en les amplifiant, offrant ainsi une mesure qui reflète mieux la variabilité des erreurs. Sa formule est :

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

Cette technique de sélection des bases d'apprentissage et de test a permis de comparer les modèles GLM Gamma et GLM Inverse Gaussienne pour la modélisation du coût moyen, ainsi que les modèles GLM Poisson et GLM Binomiale Négative avec les métriques RMSE et MAE, mais aussi avec les critères d'information AIC et BIC. Les résultats obtenus permettent d'effectuer le choix des modèles GLM Gamma et GLM Poisson. Ces deux modèles seront donc ceux utilisés dans la suite de ce mémoire.

De plus, cette méthode garantit que les modèles ajustés surpassent le modèle de référence, désigné ci-après comme le modèle `naïf`. Ce dernier prédit, pour chaque assuré, la fréquence et le coût moyen globaux du portefeuille en exploitant le principe de mutualisation des risques. Les performances du modèle `naïf` sont comparées à celles des modèles ajustés en utilisant les métriques RMSE et MAE. Les modèles de fréquence et de coût moyen retenus à cette étape sont effectivement meilleurs que les modèles naïfs, avec des RMSE et des MAE plus petits. La comparaison des métriques est présentée dans le tableau suivant.

Modèles	MAE	RMSE
Modèle de fréquence Poisson retenu	**0,03801**	**0,1403**
Modèle de fréquence Binomial Négative	0,03804	0,1403
Modèle naïf (fréquence)	0.04893	0,1412
Modèle de coût moyen Gamma retenu	**1593,7**	**2298,0**
Modèle de coût moyen Inverse Gaussienne	1594.4	2299.6
Modèle naïf (coût moyen)	1665,0	2372,0

Table 3.3: Comparaison des performances des modèles

Les premiers modèles ont été obtenus, et il est maintenant nécessaire de procéder à une sélection de variables pour ces modèles. À cet effet, différentes méthodes peuvent être utilisées. La méthode Forward a été employée en utilisant le critère AIC pour le modèle de fréquence et la métrique RMSE pour le coût moyen des sinistres. Cette sélection a révélé que la variable `CRM` a été choisie en première position pour les deux modèles.
Le Coefficient de Réduction et Majoration (CRM) est un élément crucial pour déterminer la prime d'assurance automobile. Mis en place le 6 juillet 2006, le CRM dépend des antécédents de sinistralité de l'assuré. Supposons que nous partions d'un portefeuille d'assurés à l'instant $t = 0$. À l'instant $t = 2$ ans, une amélioration du CRM (diminution du coefficient) est appliquée aux bons conducteurs n'ayant été impliqués dans aucun sinistre responsable ou partiellement responsable, et une réduction de 10% de la prime est appliquée. À l'inverse, une dégradation du CRM (augmentation du coefficient) est appliquée, et la prime est majorée de 20% pour chaque sinistre matériel et de 30% pour chaque sinistre corporel. Ces augmentations sont plafonnées à 250% du montant de la prime. L'application de ces augmentations ou diminutions s'effectue continuellement dans le temps.

Le tableau suivant présente le résumé de la sélection de variables. Pour rappel, toutes les méthodes appliquées à l'ajustement des modèles ne seront pas présentées dans cette partie.

Choix	**AIC**		**BIC**		**MAE**		**RMSE**	
Modèles	**MSSV**	**MASV**	**MSSV**	**MASV**	**MSSV**	**MASV**	**MSSV**	**MASV**
FREQ	58446,60	58431,48	58861,65	58729,46	0,0380	0,0380	0,1403	0,1403
CM	106935,33	106951,00	107197,25	107152,48	1593,77	1588,06	2298,02	2292,19

Table 3.4: Comparaison des modèles

Légende : FREQ : Fréquence, CM : Coût Moyen, MSSV : Modèle Sans Sélection de Variables, MASV : Modèle Avec Sélection de Variables.

Les autres variables sélectionnées pour le modèle de fréquence sont : ***Nombre d'années de détention du contrat***, ***Mode d'acquisition du véhicule***, ***Zonier***, ***Antécédents d'assurance***, ***Situation familiale***, ***Âge du conducteur principal***, ***Ancienneté du permis***, ***Exercice*** et ***Type de produits***. En ce qui concerne le modèle de coût moyen, les variables suivantes ont été sélectionnées : ***Exercice***, ***Zonier***, ***Âge du conducteur principal***, ***Mode de stationnement***, ***Mode d'acquisition du véhicule***, ***Nombre d'années de détention du contrat***, ***Nombre de conducteurs***, ***Usage du véhicule***.

Il convient de noter qu'une méthode de sélection de variables pourrait également être réalisée avec la régression LASSO pour comparer les résultats avec ceux obtenus par la méthode de sélection Forward. La régression LASSO ajoute une pénalisation dans le modèle GLM, empêchant ainsi les coefficients de prendre des valeurs élevées, ce qui réduit la variabilité des estimateurs.

Résumé intermédiaire

Cette première étape de construction du véhiculier a permis la mise en place des premiers modèles GLM à travers les différents choix effectués comme le choix des variables ou encore les lois de probabilité à utiliser. La deuxième étape de la construction du véhiculier est l'extraction des résidus présentée dans la suite.

3.3 Extraction des résidus

Les résidus d'un modèle de prédiction représentent la part non expliquée par les variables explicatives du modèle. Dans notre cas, les résidus peuvent être expliqués par l'absence de l'utilisation des variables véhiculières dans les modèles. Par défaut, les résidus sont définis comme la différence entre les valeurs observées y_i et les valeurs prédites $\hat{y}_i$, et on les appelle **résidus additifs**.

Les résidus additifs :

$$R_i = y_i - \hat{y}_i$$

Les résidus additifs sont faciles à comprendre, à calculer et à interpréter. Mais on peut définir d'autres types de résidus, notamment :

Les résidus de Pearson :

$$R_i = \frac{y_i - \hat{y}_i}{\sqrt{\mathbb{V}(\hat{y}_i)}}$$

Les résidus de Pearson correspondent aux résidus additifs normalisés avec l'estimateur de l'écart-type des valeurs prédites. Ils prennent en compte la variance et permettent donc de corriger l'hétéroscédasticité que présentent les résidus additifs.

Les résidus de Déviance :

$$R_i = \text{sign}(y_i - \hat{y}_i)\sqrt{d_i}$$

d_i est défini tel que

$$\sum_{i=1}^{n} d_i = \text{Déviance}$$

En fonction de la loi de probabilité utilisée dans le modèle GLM, d_i diffère en termes de formule. Dans le cas de la loi de Poisson, il est défini par :

$$d_i = \sqrt{\left| y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i) \right|}$$

Ces résidus permettent de mesurer la contribution de chaque observation à la déviance du modèle.

Les résidus d'Anscombe : Ils se basent sur la fonction de variance de la variable y de modélisation et transforment les résidus de sorte qu'ils aient une distribution normale. Sa formule est donnée par :

$$R_i = \frac{A(y_i) - A(\hat{y}_i)}{A'(y_i)\sqrt{\mathbb{V}[\hat{y}_i]}}$$

avec :

- $\mathbb{V}(.)$ la fonction de la variance, $\mathbb{V}(y) = y$ pour la loi de Poisson
- $A'(y) = \mathbb{V}(y)^{-\frac{1}{3}}$ et $A'(y_i)\sqrt{\mathbb{V}[\hat{y}_i]}$ est l'estimateur de l'écart-type de $A(y_i)$.

L'expression de $A(y_i)$ dépend également de la loi de probabilité et dans le cas de la loi de Poisson, son expression est donnée par :

$$R_i = \frac{3\left(y_i^{2/3} - \hat{y}_i^{2/3}\right)}{2\hat{y}_i^{1/6}}$$

Les résidus d'Anscombe sont très utiles dans le cas de l'utilisation de la méthode de lissage spatial, notamment le lissage par krigeage, car elle possède la propriété de régénérer un nouvel estimateur $\hat{\hat{y}}$ de y en partant d'une valeur prédite $\hat{y}$ et d'un résidu d'Anscombe. La méthode de lissage ne sera pas utilisée pour la construction du véhiculier dans ce mémoire. Nous avons plutôt choisi d'utiliser les résidus de Pearson, qui sont relativement simples et possèdent la propriété d'homoscédasticité. Les mémoires d'actuariat de François-Xavier CHAMOULAUD (2019) et Khalil FADIL (2020) proposent en effet une construction du véhiculier par méthode lissage qui peuvent être consultés pour avoir plus d'idées sur les techniques de lissage des résidus.

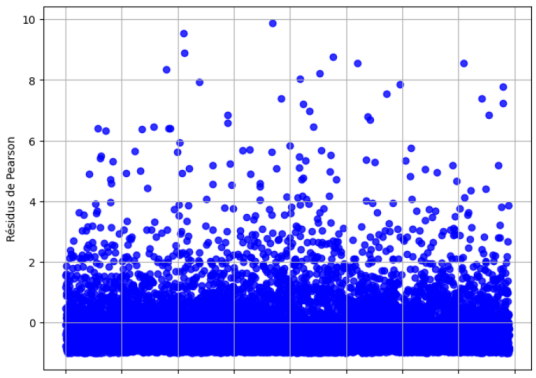

Figure 3.4: Résidus de Pearson : Coût moyen

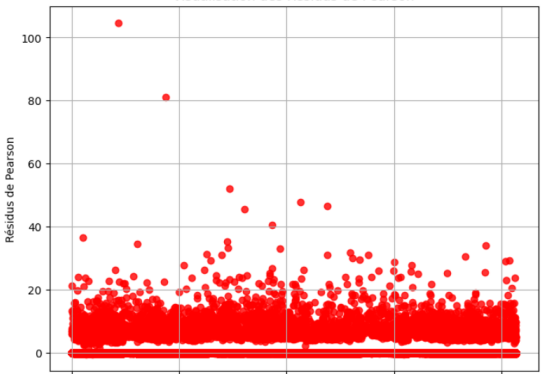

Figure 3.5: Résidus de Pearson : Fréquence de sinistres

Résumé intermédiaire

Les résidus extraits, l'étape suivante consiste à les modéliser à l'aide de modèles de machine learning. La théorie des modèles de machine learning sera présentée dans la section suivante.

3.4 Modèles d'apprentissage statistique

Le Machine Learning (apprentissage statistique) est une discipline scientifique qui fait partie intégrante de l'intelligence artificielle. Les algorithmes de machine learning repèrent des motifs récurrents dans l'ensemble de données sur lequel le modèle est appliqué. Ces données sont généralement sous la forme d'images, de textes ou de nombres. En découvrant des motifs dans ces données, les algorithmes apprennent et peuvent alors améliorer leur capacité à exécuter une tâche particulière. Les algorithmes de machine learning acquièrent donc de façon autonome la capacité d'accomplir des tâches. En particulier, ils permettent de réaliser des prédictions à partir de données, en optimisant leurs performances progressivement.

Lorsque l'objectif principal est de faire des prédictions précises, les modèles de machine learning deviennent alors essentiels. Dans ce contexte, les données d'entraînement sont utilisées pour identifier des patterns dans les données. Les modèles de machine learning utilisent une fonction de coût qui mesure l'erreur entre les prédictions du modèle et les valeurs réelles, l'objectif étant de minimiser cette erreur afin d'améliorer la précision du modèle. Ils prennent également en compte les algorithmes d'optimisation, tels que la descente de gradient pour minimiser la fonction de coût.

Il existe des modèles supervisés, entraînés sur des données étiquetées avec des sorties connues, et des modèles non supervisés, qui découvrent des structures cachées dans des données non étiquetées dont la finalité est justement l'étiquetage de ces données.

Contrairement aux modèles classiques, les modèles d'apprentissage statistique capturent non seulement les relations non linéaires mais aussi les interactions entre variables, intégrant une complexité dans le modèle.

L'application des modèles d'apprentissage statistique dans le domaine de la tarification non-vie revêt un avantage particulier. Les assureurs peuvent alors établir des primes plus justes et compétitives, optimisant par conséquent la gestion des risques.

Les modèles de machine learning peuvent être classés en deux catégories :

- Modèles simples : les Arbres de Décision (CART), les Réseaux de Neurones (RN) et les Machines à Vecteurs de Support (SVM)
- Modèles d'agrégation : Bagging, Forêts aléatoires (Random Forest), AdaBoost, LPBoost, XGBoost, GradientBoost, BrownBoost ...

Le modèle d'arbres de décision sera présenté dans cette section, les réseaux de neurones seront présentés dans le chapitre suivant dans le cadre de l'étude du Modèle Additif Neuronal. Les autres modèles présentés sont les modèles d'agrégation et principalement les modèles Random Forest, Gradient Boosting et XGBoost. Les modèles présentés relèvent exclusivement de l'apprentissage supervisé.

3.4.1 Modèle d'arbres de décision

Le modèle d'arbres de décision introduit par Breiman et al. en 1984 sous l'appellation "Classification and Regression Tree" (CART) s'applique, comme son nom l'indique, en régression (variable réponse Y quantitative) et en classification (variable réponse Y qualitative). Le modèle CART a pour principe le partitionnement de l'espace constitué par les variables explicatives (représenté par X dans la suite) de sorte que les valeurs de la variable réponse Y dans chaque partition soient presque égales[2]. Le modèle CART se présente ainsi comme un modèle facile à interpréter offrant une visualisation de la formation des partitions et intéressant pour la prise de décision. Le graphique 3.6 montre un exemple d'arbres de décision.

[2]Il existe également d'autres méthodes de constitution des partitions comme dans les modèles C4.5, CHAID...

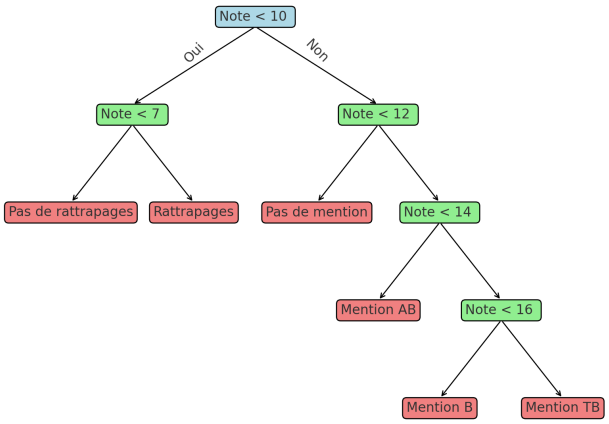

Figure 3.6: Exemple d'arbres de décision

Dans la suite, les modèles sont définis par le vecteur aléatoire $(X, Y) \in \mathbb{R}^p \times \mathbb{G}$, avec p le nombre de variables explicatives, $\mathbb{G} = \mathbb{R}$ dans le cas d'une régression et $\mathbb{G} = \{1, \ldots, K\}$ dans le cas d'une classification.
La base d'entraînement est définie par : $L_n = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{G}, i = 1, \ldots, n\}$ avec (x_i, y_i) des réalisations indépendantes de la loi du vecteur aléatoire (X, Y).
Le partitionnement débute avec l'ensemble des données d'entrée contenant toutes les observations de la base d'entraînement, qui représente la racine de l'arbre notée n_1 à laquelle se grefferont ensuite les branches issues des découpages.
Partant de cette racine, l'ensemble des données de la base d'entraînement est partitionné en deux groupes suivant une variable explicative X^j choisie. Le premier groupe est défini par : $\{X^j \leq s\}$ et le deuxième par : $\{X^j > s\}$. Le choix de la variable et du seuil est effectué à travers la minimisation de la fonction de coût $C(j, s)$ qui revient à minimiser la variance des deux groupes constitués dans le cas d'une régression. La fonction de coût est alors définie par :

$$C(s, j) = \sum_{i \in n_{1,-}(j,s)} (y_i - \overline{y}_-)^2 + \sum_{i \in n_{1,+}(j,s)} (y_i - \overline{y}_+)^2$$

avec :

- $n_{1,-}(j, s) = \{i \in \{1, \ldots, n\} \ / \ x_i^j \leq s\}$, $n_{1,+}(j, s) = \{i \in \{1, \ldots, n\} \ / \ x_i^j > s\}$ et $x_i = \{x_i^1, \ldots, x_i^p\}$
- $\overline{y}_- = \frac{1}{\text{card}(n_{1,-}(j,s))} \sum_{i \in n_{1,-}(j,s)} y_i$ et $\overline{y}_+ = \frac{1}{\text{card}(n_{1,+}(j,s))} \sum_{i \in n_{1,+}(j,s)} y_i$

Dans le cas d'une classification, la minimisation de la fonction de coût consiste à minimiser l'indice de Gini afin de rendre les classes formées aussi homogènes que possible. La fonction de coût est alors définie par :

$$C(s,j) = \sum_{k=1}^{K} \hat{p}^k_{n_{1,-}(j,s)} \left(1 - \hat{p}^k_{n_{1,-}(j,s)}\right) + \sum_{k=1}^{K} \hat{p}^k_{n_{1,+}(j,s)} \left(1 - \hat{p}^k_{n_{1,+}(j,s)}\right)$$

avec $\hat{p}^k_{n_{1,-}(j,s)}$ la proportion de la modalité k de la variable réponse dans le groupe $n_{1,-}(j,s)$.

L'algorithme itère cette procédure sur chacun des sous-arbres retenus et ceci jusqu'à une condition d'arrêt comme le nombre d'observations dans les nœuds qui ne doit pas être inférieur à une valeur fixée. Aussi, le découpage ne peut être appliqué sur un nœud pur défini comme un nœud contenant des observations identiques. Les nœuds obtenus à la fin sont les nœuds terminaux et correspondent aux feuilles de l'arbre. La prédiction pour une observation de la feuille n de l'arbre maximal T_{max} obtenu est donnée par la moyenne des observations de la feuille pour une régression et par le mode des observations pour la classification.

Le modèle nul constitué que de la racine a une variance nulle mais un biais important tandis que le modèle T_{max} a un biais moins important mais une variance grande car elle dépend fortement des observations utilisées et présente alors un problème de surapprentissage. Il convient alors de définir un sous-arbre de T_{max} qui cherche le compromis entre ces deux modèles en définissant une suite de sous-arbres imbriqués issus de l'arbre maximal : c'est l'étape d'élagage. La suite de sous-arbres élagués est obtenue dans le cas de la régression en minimisant un critère avec pénalisation de tous les sous-arbres imbriqués T de l'arbre maximal dont la formule est la suivante :

$$\mathrm{C}_\alpha(T) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{T,i})^2 + \alpha |T|$$

où : $|T|$ est le nombre de feuilles de T, $\hat{y}_{T,i}$ est la prédiction renvoyée par l'arbre T pour l'observation y_i et $\alpha \geq 0$.

Quand $\alpha = 0$, on retrouve l'arbre maximal. Si α augmente, alors le nombre de feuilles de la suite de sous-arbres diminue. Le choix de la valeur α se fait alors grâce à une méthode très utilisée en statistique, la validation croisée. On définit une base de test $\tilde{L}_{n_t} = \{(\tilde{x}_i, \tilde{y}_i) \in \mathbb{R}^p \times \mathbb{G}, i = 1, \ldots, n_t\}$ exactement comme la base d'apprentissage, mais avec des données n'ayant pas servi à la constitution des arbres. Le meilleur arbre de la suite des sous-arbres $(T_j)_{1 \leq j \leq J}$ est l'arbre dont l'indice est défini par :

$$j^* = \mathrm{argmin}_{1 \leq j \leq J} \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} (\tilde{y}_i - \hat{y}_{T,j,i})^2 \right\}$$

avec $\hat{y}_{T,j,i}$ la prédiction de l'observation y_i avec l'arbre T_j.

Le modèle CART présente l'avantage d'être facilement interprété à l'aide du graphique visuel de la formation des classes mais il manque de robustesse et dépend fortement des données utilisées pour la modélisation. Il est alors utilisé dans les modèles ensemblistes qui sont présentés dans la suite.

3.4.2 Random Forest

Les modèles ensemblistes ont pour principe général l'agrégation des modèles de base dans un objectif de réduction de la variance et de renforcement de la robustesse de ces modèles de base qui sont généralement des arbres de décision.
Le modèle Random Forest appartient à une catégorie de modèles d'agrégation appelés modèles parallèles qui combinent plusieurs modèles de base construits de manière indépendante.

Le modèle Random Forest construit les modèles de base sur différents échantillons de même taille issus de la base d'entraînement. La prédiction finale est alors donnée par la moyenne des valeurs prédites par chaque modèle dans le cas d'une régression et par le mode des classes prédites dans le cas d'une classification.
En supposant que la variance de tous les prédicteurs de base est σ^2 et que le coefficient de corrélation associé à deux prédicteurs de base est ρ, le prédicteur obtenu a une variance équivalente à $\left(\frac{\sigma^2(1-\rho)}{E} + \rho\sigma^2\right)$ avec E le nombre de modèles de base. La variance diminue avec l'augmentation du nombre de modèles de base et avec la diminution du paramètre ρ. Il faudrait alors choisir des modèles de base moins corrélés comme les arbres de décision pour obtenir des prédicteurs avec moins de variabilité.

Dans ce contexte de réduction de la variance des prédicteurs, Breiman propose les modèles de Forêts aléatoires avec pour objectif l'ajout d'aléa concernant le choix des variables de découpage dans les modèles CART utilisés comme modèles de base. Les m variables qui sont proposées comme variables de découpage sont alors choisies aléatoirement parmi les variables explicatives. Le choix est fait à chaque construction d'un arbre et de manière indépendante. Le paramètre m est alors un paramètre sensible du modèle Random Forest. Cette méthode est intéressante quand on est en présence d'un grand nombre de variables explicatives.

3.4.3 Modèles Gradient Boosting Machine et XGBoost

La deuxième catégorie des modèles d'agrégation est celle des modèles adaptatifs dont les modèles de base sont dépendants les uns des autres, contrairement aux modèles parallèles. Les modèles de base sont construits successivement en fonction de la performance du modèle précédent. Cette catégorie de modèles transforme les apprenants de base faibles, appelés weak learners, en un modèle plus robuste appelé strong learner. Ce mécanisme se réalise avec une correction progressive des erreurs commises par les modèles précédents : c'est la méthode du boosting.

Il existe plusieurs modèles découlant de cette méthode, dont le modèle de Gradient Boosting Machine (GBM) initié par Friedman. Le modèle GBM est une variante du modèle Adaboost[3] qui apporte des améliorations successives des modèles en appliquant une descente de gradient pour la minimisation d'une fonction de perte. En effet, l'erreur que corrige le modèle de manière progressive est en réalité le gradient associé à la fonction de perte par rapport aux prédictions renvoyées par le modèle. Cette fonction de perte doit être convexe et différentiable. Il peut s'agir de la fonction d'erreur quadratique en régression et de la fonction logistique dans le cas de la classification. En théorie mathématiques, la dérivée de la fonction de coût indique la direction dans laquelle doivent être ajustées les prédictions afin de maximiser l'erreur de prédiction. Cette théorie donne alors l'idée à Friedman d'ajuster plutôt les prédictions vers l'opposé du gradient correspondant au gradient négatif afin de minimiser la fonction de perte et de gagner en performance.
L'algorithme suivant décrit le processus pour le cas de la régression.

Algorithme de Gradient Boosting Machine

1: **Initialisation :** $g_0(\cdot) = \arg\min_c \frac{1}{n} \sum_{i=1}^{n} f(y_i, c)$ où f est la fonction de perte.
2: **for** $m = 1$ to M **do**
3: Calcul de l'opposé du gradient de la fonction de perte et évaluation aux points $g_{m-1}(x_i)$:

$$u_i = - \frac{\partial}{\partial g(x_i)} f(y_i, g(x_i)) \Big|_{g(x_i)=g_{m-1}(x_i)}, \quad i = 1, \ldots, n.$$

4: Ajustement d'un modèle d'arbres de décision sur la base d'entraînement $(x_1, u_1), \ldots, (x_n, u_n)$, désigné par Ψ_m.
5: Mise à jour du modèle : $g_m(x) = g_{m-1}(x) + \lambda \Psi_m(x)$.
6: **end for**
7: **Sortie :** Le modèle $(g_M(x))$.

La valeur du paramètre λ dépend du nombre de modèles de base (nombre d'itérations) choisi. En effet, plus la valeur de λ est petite, plus il est nécessaire d'utiliser un grand nombre d'itérations.

D'autres modèles de boosting découlent du modèle GBM, tels que les modèles Gradient Tree Boosting ou Stochastic Gradient Boosting. Cependant, le modèle eXtreme Gradient Boosting (XGBoost), développé en 2016 par C. Guestrin et Tianqi Chen, est l'une des améliorations les plus remarquables du modèle GBM et a été largement utilisé avec succès dans les compétitions de Kaggle de ces dernières années. Les principales améliorations du modèle XGBoost concernent :

[3]Le modèle Adaboost est un modèle de Boosting qui améliore itérativement les modèles de base en fixant des poids identiques aux observations au début puis augmente ces poids à l'itération suivante si l'observation a été mal prédite et garde inchangé le poids si la prédiction est bonne, obtenant ainsi une agrégation des modèles qui fournit une bonne performance.

La parallélisation

Dans les modèles d'arbres de décision standard, la recherche de la meilleure séparation à chaque nœud se fait de manière séquentielle. En revanche, dans le modèle XGBoost, cette recherche est effectuée simultanément sur plusieurs threads[4], ce qui réduit le temps de construction des arbres.

Le modèle XGBoost utilise la technique de "bloc par colonne", où les données sont divisées en blocs et traitées simultanément et indépendamment par les threads. Cette méthode permet une meilleure utilisation de la mémoire cache du CPU (Central Processing Unit). Le CPU, souvent appelé le "cerveau" de l'ordinateur, exécute les programmes et constitue le principal composant de calcul.

Les GPU (Graphics Processing Units)[5] sont également très performants pour les calculs parallèles. XGBoost accélère ainsi l'entraînement des modèles en tirant parti de cette capacité des GPU à effectuer des calculs parallèles, particulièrement avantageux pour des données de grande volumétrie.

La gestion des valeurs manquantes

Le modèle XGBoost simplifie le prétraitement des données en déterminant automatiquement les directions optimales pour les valeurs manquantes lors de l'ajustement du modèle. Pendant la construction des sous-arbres, les données manquantes sont identifiées par le modèle, qui décide alors de leur direction. À chaque division, la direction optimale est choisie en évaluant les gains potentiels de l'attribution d'une valeur manquante à la branche de droite ou à celle de gauche. La direction choisie pour une valeur manquante correspond à celle offrant le meilleur gain, limitant ainsi l'influence des valeurs manquantes sur l'efficacité globale du modèle.

La régularisation

La fonction de perte utilisée dans le modèle XGBoost est une fonction de perte modifiée correspondant à la somme de la fonction de perte standard (fonction de perte quadratique en régression) et des termes de régularisation Lasso (L_1) et Ridge (L_2). L'ajout de la régularisation dans la fonction de perte favorise la robustesse du modèle et sa bonne généralisation sur des données non utilisées lors de la modélisation, limitant ainsi le surapprentissage. La régularisation du modèle XGBoost est une combinaison linéaire des termes de régularisation Lasso (γ) et Ridge (λ), dont la formule est :

[4]Les threads permettent l'exécution de plusieurs opérations au sein d'un même programme, exploitant ainsi le plein potentiel des processeurs multi-cœurs.

[5]Les GPU sont des processeurs ou des unités de traitement graphique et d'images.

$$\Omega(T) = \gamma|T| + \frac{1}{2}\lambda \sum_j w_j^2$$

où $|T|$ est le nombre de nœuds terminaux ou feuilles de l'arbre de décision et w_j est le poids associé à la j-ième feuille.

Les poids des observations

Le modèle XGBoost donne la possibilité d'attribuer des poids aux observations de la base d'entraînement pour deux principales raisons :

- *Déséquilibre dans les données* : Dans le cas de la classification, par exemple, où une classe de la variable réponse Y est beaucoup plus représentée que toutes les autres.
- *Importance relative des observations* : Lorsque les observations n'ont pas la même importance, comme dans le cas de la modélisation du coût moyen de sinistres où les observations sont pondérées par le nombre de sinistres.

Grâce à ces améliorations, le modèle XGBoost se distingue par ses performances élevées en termes de prédiction. Ces améliorations ajoutent davantage de paramètres, nécessitant une optimisation pour atteindre les meilleures performances du modèle. L'optimisation des paramètres sera directement appliquée aux modélisations présentées dans la section suivante.

Résumé intermédiaire

Les modèles de machine learning présentés dans cette section seront utilisés pour modéliser les résidus de Pearson qui ont été extraits des modèles GLM. La section suivante est consacrée à la modélisation et la prédiction de ces résidus.

3.5 Ajustement d'un modèle d'apprentissage statistique sur les résidus

Les résidus issus des modèles GLM de fréquence et de coût moyen sont modélisés dans notre étude à l'aide de modèles de machine learning, car ces derniers ne font aucune hypothèse sur la loi de la variable à expliquer.

La variable à expliquer dans les modèles de machine learning utilisés à cette étape est la distribution des résidus de Pearson des modèles GLM ajustés sans les variables véhiculières. Ces résidus sont donc attribués à l'absence de variables liées aux véhicules. Ainsi, les variables explicatives utilisées sont les variables véhiculières. Il est important

de rappeler que les modèles de machine learning sont plus robustes face aux variables explicatives corrélées. Par conséquent, seules les exclusions de variables véhiculières déterminées au niveau de l'ACP du chapitre précédent seront appliquées.

Le tableau ci-dessous présente les variables explicatives utilisées dans la modélisation des résidus.

Variables explicatives	Nature de la variable
Nombre de places du véhicule	Quantitative
Puissance en CV du véhicule	Quantitative
Prix du véhicule	Quantitative
Puissance en CH du véhicule	Quantitative
Vitesse du véhicule	Quantitative
Poids Total Autorisé à Charge du véhicule	Quantitative
Cylindrée du véhicule	Quantitative
Nombre de cylindres du véhicule	Quantitative
Puissance en CEE du véhicule	Quantitative
Nombre de rapports du véhicule	Quantitative
Longueur du véhicule	Quantitative
Largeur du véhicule	Quantitative
Poids du véhicule	Quantitative
Carrosserie du véhicule	Qualitative
Transmission du véhicule	Qualitative
Type d'alimentation du véhicule	Qualitative
Type de boîte vitesse du véhicule	Qualitative
Suspension du véhicule	Qualitative
Type de frein du véhicule	Qualitative

Table 3.5: Variables explicatives utilisées dans la modélisation des résidus

Les modèles sélectionnés pour la modélisation des résidus sont XGBoost (modèles ensemblistes adaptatifs) et Random Forest (modèles ensemblistes parallèles). Au total, cela représente quatre modèles : deux modèles pour les résidus de la fréquence de sinistres et deux autres pour les résidus du coût moyen. Ces modèles seront comparés en termes de performance, et le meilleur sera retenu pour obtenir les résidus prédits, utiles pour la suite de la construction du véhiculier.

Les modèles XGBoost et Random Forest étant des modèles hyperparamétriques, il est nécessaire de rechercher des hyperparamètres optimaux pour maximiser leurs performances : l'optimisation des hyperparamètres est cruciale.

Il existe différentes méthodes pour l'optimisation des hyperparamètres, telles que la validation croisée ou K-Fold, le Random Search ou le Grid Search. Dans cette étude, l'optimisation des hyperparamètres sera réalisée avec la méthode Grid Search. Le principe de la méthode Grid Search est simple : l'utilisateur définit une plage de valeurs pour chaque hyperparamètre à optimiser. L'algorithme crée alors une matrice de toutes les combinaisons possibles des valeurs des hyperparamètres considérés. Chaque combinaison possible est ensuite utilisée pour entraîner et évaluer le modèle, ce qui permet de calculer une métrique pour chaque combinaison. La métrique dépend du type de modélisation et peut être l'accuracy pour un modèle de classification ou le RMSE pour un modèle de régression. La meilleure combinaison est celle qui offre la meilleure performance selon la métrique définie.

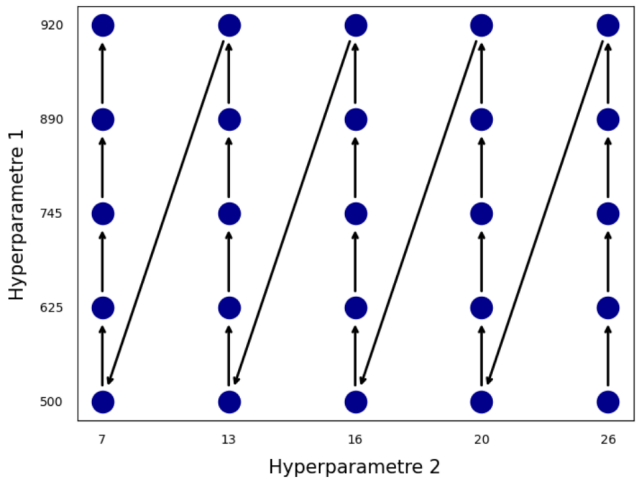

Figure 3.7: Illustration de la méthode Grid Search

Les hyperparamètres qui ont fait l'objet d'optimisation pour les modèles Random Forest et XGBoost sont les suivants :

- **Random Forest** :
  - Nombre de variables à sélectionner aléatoirement pour la construction des arbres, capté par l'argument `max_features`
  - Nombre d'arbres de décision, spécifié par l'argument `n_estimators`
  - Profondeur maximale de chaque arbre de décision, avec l'argument `max_depth`
- **XGBoost** :
  - Taux d'apprentissage dont l'argument est `learning_rate`
  - Termes de régularisation L_1 et L_2, associés aux arguments `alpha` et `lambda`
  - Profondeur maximale des arbres de décision, spécifiée par l'argument `max_depth`

– Nombre de variables à sélectionner aléatoirement pour chaque arbre, spécifié par l'argument `colsample_bytree`

Les modèles sont ajustés avec les hyperparamètres obtenus par la méthode Grid Search dont le récapitulatif est présenté dans le tableau suivant :

Métriques	**MAE**		**RMSE**	
Modèles	**Random Forest**	**XGBoost**	**Random Forest**	**XGBoost**
FREQ	0,2637	0,0199	1,0526	0,3054
CM	0.6643	0.6617	0.9304	0.9293

Table 3.6: Comparaison des modèles

Sur la base du tableau 3.6, le modèle XGBoost s'affiche comme le meilleur choix pour la modélisation des résidus du modèle GLM de fréquence de sinistres et de coût moyen. En effet, il montre des performances supérieures ou similaires à celles de Random Forest pour les métriques MAE et RMSE. En particulier, pour la modélisation des résidus du modèle GLM de fréquence de sinistres, XGBoost présente des métriques beaucoup plus faibles par rapport au modèle Random Forest. En ce qui concerne la modélisation des résidus du modèle GLM de coût moyen, bien que les performances soient très proches, XGBoost reste légèrement supérieur. Ainsi, XGBoost est le modèle retenu en raison de ses meilleures performances globales. À l'aide du modèle XGBoost, les résidus prédits sont obtenus et serviront dans la suite de la construction du véhiculier.

Résumé intermédiaire

L'étape de modélisation des résidus permet d'obtenir les résidus prédits. La phase finale de la construction du véhiculier consiste à classifier ces résidus prédits, ce qui fera l'objet de la section suivante.

3.6 Classification des résidus prédits

L'objet de cette section est de regrouper les résidus prédits en fonction de leur similarité. Il est important de noter que, dans certaines études de construction de véhiculier, les résidus sont d'abord agrégés par Code SRA avant d'être classés, comme l'a décrit Leslie GNANSOUNOU dans son mémoire d'actuariat (2022). Cependant, dans notre cas, nous avons choisi de laisser les méthodes de classification déterminer les classes optimales pour ces résidus.

Les méthodes de clustering (classification) se divisent en deux catégories : la classification supervisée et la classification non supervisée. La classification supervisée utilise les caractéristiques des observations ainsi que les classes préalablement définies

pour apprendre et fournir une prédiction. En revanche, la classification non supervisée détermine les classes possibles de regroupement des observations uniquement à partir de leurs caractéristiques, sans connaître les classes à l'avance. L'objectif final de la classification non supervisée est de fournir des classes de regroupement possibles pour l'ensemble des individus.

Le but étant de construire des classes de risques homogènes associées aux véhicules, la méthode de classification utilisée est la classification non supervisée. Celle-ci se divise en deux sous-catégories : la classification non supervisée hiérarchique et la classification non supervisée non hiérarchique. Parmi les méthodes de classification non supervisée hiérarchique, on peut citer la **Classification Ascendante Hiérarchique (CAH)**, tandis que la méthode **K-Means** est un exemple très connu de classification non supervisée non hiérarchique.

L'algorithme de clustering K-Means est l'une des méthodes de classification non supervisée les plus populaires. La méthode K-Means permet de regrouper les observations en K clusters (classes) définis par l'utilisateur. Son principe repose sur la minimisation de la **variance intra-classe**. La variance intra-classe associée à une classe est la somme des distances euclidiennes au carré de chaque observation par rapport au centroïde de la classe. Le centroïde est le centre de gravité des observations de la classe. La formule de la variance intra-classe est donnée par :

$$V = \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - c_k\|^2$$

où C_k est le k-ième cluster, $\mathbf{x}_i$ est le vecteur de caractéristiques de la i-ème observation, c_k est le centroïde du k-ième cluster et $\|\cdot\|$ représente la norme euclidienne.

L'algorithme calcule les distances euclidiennes entre les centroïdes et les observations afin d'attribuer les classes. Initialement, il n'y a pas de classes définies, donc les premiers centroïdes sont choisis aléatoirement parmi les observations. Ensuite, pour chaque observation, on calcule les distances par rapport à tous les centroïdes, et l'observation est assignée à la classe dont le centroïde est le plus proche. Après cette première étape, des classes sont établies et les nouveaux centroïdes de chaque classe, représentant le centre de gravité des observations, peuvent être calculés. Le processus d'affectation est alors répété jusqu'à ce qu'aucune observation ne change de classe.

La méthode K-Means présente deux contraintes majeures. La première est que l'algorithme dépend fortement de l'initialisation des centroïdes. Pour un nombre de classes donné, on effectue une validation croisée en changeant plusieurs fois la graine de reproductibilité qui détermine le choix aléatoire des premiers centroïdes et on conserve à la fin la graine qui minimise la variance intra-classe. La seconde contrainte est que le nombre de classes doit être fixé à l'avance, l'algorithme n'étant pas capable de déterminer le nombre optimal de classes. Pour remédier à cela, nous utilisons la

méthode CAH (Classification Ascendante Hiérarchique), qui produit un dendrogramme permettant d'identifier le nombre optimal de classes.

La méthode CAH regroupe successivement les observations en fonction de leur similarité pour aboutir à une seule classe finale. Initialement, chaque observation est considérée comme une classe distincte. La première étape consiste à associer les deux individus les plus proches selon la matrice de distances. À chaque étape, deux classes sont fusionnées, celles ayant la distance la plus faible. Pour cela, il est nécessaire de définir la notion de distance entre deux classes. Parmi ces différentes distances possibles, la **distance de Ward** est la plus couramment utilisée, car elle minimise la variance intra-classe. La distance de Ward entre deux clusters A et B est définie par :

$$D_{Ward}(A, B) = \frac{n_A \cdot n_B}{n_A + n_B} \|c_A - c_B\|^2$$

où :

- n_A et n_B sont les effectifs des clusters A et B,
- c_A et c_B sont les centroïdes des clusters A et B

Les dendrogrammes fournis par la méthode CAH sur les deux distributions de résidus prédits (fréquence de sinistres et coût moyen) sont présentés ci-dessous :

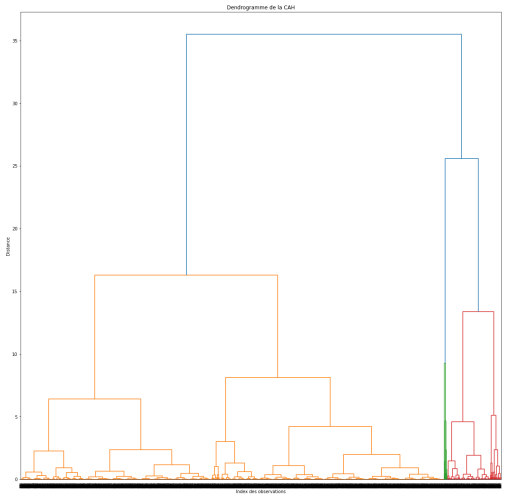

Figure 3.8: Dendrogramme sur les résidus prédits pour le véhiculier du modèle de coût moyen

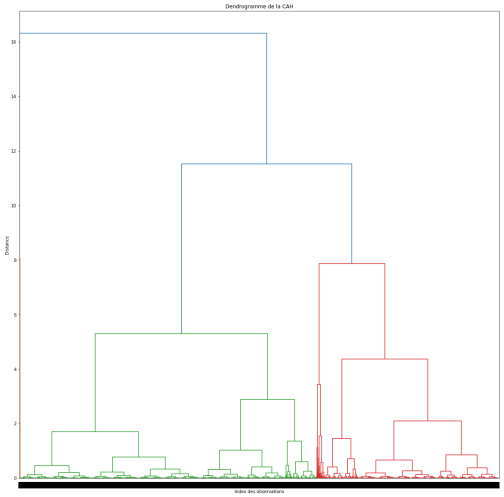

Figure 3.9: Dendrogramme sur les résidus prédits pour le véhiculier du modèle de fréquence de sinistres

L'analyse des dendrogrammes permet de retenir approximativement 10 classes pour le modèle de fréquence de sinistres et 15 classes pour le modèle de coût moyen. Ces nombres de classes seront ensuite confirmés ou infirmés par la méthode du coude appliquée sur la méthode K-Means avec le calcul de la moyenne des variances intra-classe. La méthode

du coude dans ce cas consiste à calculer, pour plusieurs nombres de classes définis, la moyenne des variances intra-classe et à choisir le nombre de classes à partir duquel cette moyenne devient stable. Les deux graphiques suivants présentent l'évolution de la moyenne des variances intra-classe en fonction du nombre de classes défini.

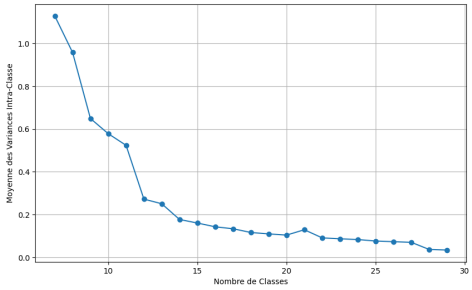

Figure 3.10: Évolution de la moyenne des variances intra-classe en fonction du nombre de classes pour le véhiculier du modèle de coût moyen

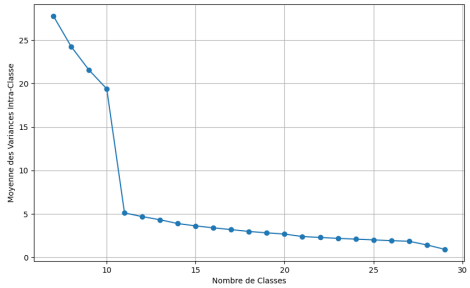

Figure 3.11: Évolution de la moyenne des variances intra-classe en fonction du nombre de classes pour le véhiculier du modèle de fréquence de sinistres

La méthode du coude appliquée aux figures 3.10 et 3.11 permet de choisir **14 classes pour le véhiculier du modèle de coût moyen et 11 classes pour le véhiculier du modèle de fréquence de sinistres**.

Résumé intermédiaire

L'étape finale de la construction des véhiculiers s'achève en réalisant une combinaison des méthodes K-Means et CAH sur les différents résidus prédits permettant d'obtenir les nombres de classes définis. Les classifications obtenues représentent les véhiculiers. Ils sont ensuite intégrés dans la base de données pour évaluer leur impact sur la modélisation de la fréquence de sinistres et du coût moyen, présenté ci-après.

3.7 Impacts du véhiculier sur la tarification et modèle GLM final

Après l'obtention des véhiculiers, il est crucial de mesurer l'impact de l'utilisation de ces variables construites sur les modélisations GLM des deux indicateurs de risque.

Les véhiculiers sont des variables qualitatives intégrées dans les modèles GLM de fréquence de sinistres et de coût moyen. L'objectif initial était de comparer le véhiculier construit avec celui proposé par la SRA, à travers les variables `Classe SRA` et `Groupe`

`SRA`. Il s'agit donc de comparer trois modèles : le modèle sans variables véhiculières, le modèle avec les variables explicatives non véhiculières et le véhiculier de la SRA, et enfin, le modèle ajusté avec les variables explicatives non véhiculières et le véhiculier construit. La comparaison des modèles est résumée dans le tableau suivant :

Métriques	**RMSE**				**MAE**			
	CM	**VAR**	**FREQ**	**VAR**	**CM**	**VAR**	**FREQ**	**VAR**
GLM SVV	2292,19	100%	0,1403	100%	1588,06	100%	0,0380	100%
GLM SRA	2252,97	**-1,74%**	0,1403	100%	1571,79	**-1,03%**	0,0380	100%
GLM VEH	1756,80	**-30,47%**	0,1395	**-0,57%**	1249,45	**-27,10%**	0,0377	**-0,79%**

Table 3.7: Comparaison des modèles

Légende : **VAR** : Variation, **GLM SVV** : GLM Sans Variables Véhiculières, **GLM SRA** : GLM avec le véhiculier de la SRA, **GLM VEH** : GLM avec le véhiculier construit.

D'après l'analyse du tableau 3.7, le modèle GLM SVV affiche un RMSE de $2292,19$ pour le coût moyen et $0,1403$ pour la fréquence de sinistres, avec un MAE de $1588,06$ pour le coût moyen et $0,0380$ pour la fréquence de sinistres. Ce modèle est celui de référence par rapport auquel les deux autres modèles seront comparés, toutes les variations (VAR) sont alors fixées à 100% pour le modèle GLM SVV.

En comparant le modèle GLM SVV avec le modèle **GLM SRA**, une légère amélioration est observée. Pour le coût moyen (CM), le RMSE diminue de 1,74% et le MAE diminue de 1,03%. Cependant, pour la fréquence de sinistres (FREQ), le RMSE et le MAE restent identiques à ceux du modèle de référence, indiquant que l'ajout du véhiculier de la SRA n'a pas eu d'impact sur la modélisation de la fréquence de sinistres selon ces deux métriques. Ces résultats suggèrent que bien que le véhiculier de la SRA apporte une légère amélioration dans la prédiction du coût moyen, son impact en termes de performance sur la fréquence de sinistres peut être négligeable.

Le modèle **GLM VEH** affiche des performances nettement supérieures par rapport au modèle GLM SVV. Le RMSE pour le coût moyen diminue de 30,47% et le MAE diminue de 27,10%. Pour la fréquence de sinistres, le RMSE diminue de 0,57% et le MAE diminue de 0,79%. Ces réductions indiquent que le véhiculier construit pour le coût moyen capture efficacement les risques associés aux véhicules, améliorant ainsi le modèle sans variables véhiculières. Cela pourrait s'expliquer par le fait que le ***véhiculier construit*** prend en compte des variables supplémentaires contenant des informations sur les caractéristiques des véhicules explicatives du risque qui ne sont pas présentes dans le modèle GLM SVV.

Les améliorations significatives dans les métriques de performance pour le coût moyen

indique que le véhiculier capture mieux les risques spécifiques aux véhicules. Ces résultats mettent en évidence l'importance de la construction du véhiculier, une variable spécifique adaptée aux caractéristiques des véhicules pour améliorer la précision de la tarification dans les contrats d'assurance automobile. La création de ce véhiculier sur mesure est donc une stratégie efficace pour mieux modéliser les risques associés aux véhicules, offrant ainsi une tarification plus précise et potentiellement plus juste pour les assureurs.

L'importance du ***véhiculier construit*** se reflète également dans la sélection des variables. En effet, lorsqu'une sélection de variables est effectuée dans les deux modèles utilisant les variables non véhiculières, ***le véhiculier construit*** et le véhiculier de la SRA, ***le véhiculier construit*** est la première variable choisie par la méthode, témoignant à nouveau de l'efficacité de cette variable.

En abordant la sélection des variables, nous présenterons dans cette partie, les méthodes utilisées dans la modélisation. Tout d'abord, il convient de souligner qu'avant la première modélisation et donc avant la sélection des variables, une gestion des modalités de référence a été effectuée. Le choix a été fait de mettre en référence, pour toutes les variables qualitatives, la modalité la plus représentative en termes d'effectifs.

Ensuite, nous avons effectué une sélection de variables exactement comme dans le cas des modélisations sans variables véhiculières. Les variables retenues pour les modèles sont les mêmes qu'auparavant, auxquelles ont été ajoutées les véhiculiers.

Nous passons ensuite à l'étape d'introduction des interactions dans le modèle. Le modèle GLM ne prenant pas en compte directement les interactions entre les variables, il est nécessaire de les introduire manuellement pour détecter les interactions optimales. Cependant, le temps nécessaire pour cette opération augmente rapidement non seulement en raison du nombre d'interactions possibles, mais également à cause du temps de calcul requis par le logiciel. L'algorithme MARS (Régression Multivariée par Spline Adaptative) a été proposé par GNANSOUNOU (2022) dans son mémoire d'actuariat pour détecter directement les interactions optimales, que le lecteur peut consulter. Néanmoins, il a été décidé dans cette étude de prendre en considération uniquement les interactions entre les variables quantitatives et qualitatives. La démarche consiste à introduire une première interaction entre la variable quantitative la plus importante, le CRM dans notre cas, avec une première variable qualitative. L'interaction est retenue dans le modèle si son intégration améliore simultanément l'AIC, le RMSE et le MAE. Ensuite, nous introduisons successivement l'interaction entre le CRM et les autres variables qualitatives, la décision de retenir l'interaction restant la même. Cette opération est ensuite répétée pour les variables quantitatives importantes suivantes. Cette méthode souffre de la dépendance à l'ordre dans lequel les interactions sont testées et les interactions testées dans notre cas n'ont pas été significatives en termes d'amélioration des performances des modèles.
Le modèle GLM est très interprétable, il est donc essentiel de comprendre comment

s'effectuent les prédictions à travers l'interprétation des coefficients significatifs du modèle. De plus, l'utilisation de la fonction logarithmique permet d'obtenir un modèle multiplicatif qui s'interprète plus facilement.

3.7.1 Interprétation du modèle de coût moyen

Le tableau suivant illustre les coefficients des variables significatives du modèle de coût moyen :

Variables	θ	$\exp(\theta)$	**p-value**
Âge[T.<=25]	0,3066	**1,358**	0,0000
Âge[T.26-35]	0,0532	**1,054**	0,0494
Âge[T.51-65]	-0,0827	**0,920**	0,0005
Âge[T.66-75]	-0,0942	**0,910**	0,0266
Âge[T.76-80]	-0,0351	**0,965**	0,0354
Âge[T.>80]	-0,0529	**0,948**	0,0182
PARK[T.Garage/Parking_Clos]	0,227	**1,254**	0,0000
PARK[T.Parking_Collectif]	0,1279	**1,136**	0,0005
Zone_ZONIER[T.Z1]	0,0124	**1,012**	0,0346
Zone_ZONIER[T.Z3]	0,0066	**1,068**	0,0086
Zone_ZONIER[T.Z4]	0,1405	**1,150**	0,0000
Zone_ZONIER[T.Z5]	0,5273	**1,694**	0,0000
Zone_ZONIER[T.Z6]	0,2897	**1,336**	0,0007
CRM_AUTO	0,3559	**1,427**	0,0000
EXERCICE	0,0579	**1,059**	0,0000

Table 3.8: Résultats de sortie du modèle GLM de coût moyen

L'interprétation des coefficients du modèle se fait à partir de l'exponentielle des coefficients constituant la grille tarifaire.

La variable Âge a pour modalité de référence la classe d'âge 36-50. Les autres classes d'âge sont donc interprétées en fonction de cette classe de référence. Le tableau montre une tendance initialement baissière puis une petite correction haussière du coût moyen des sinistres suivant l'augmentation de l'âge des conducteurs. En effet, le coût moyen des sinistres pour les jeunes conducteurs de moins de 25 ans est supérieur de 35,8% par rapport aux conducteurs de la tranche 36-50. Cette augmentation, bien que plus modérée, est également observée pour la tranche d'âge comprise entre 26 et 35 ans, avec une hausse de 5,4% par rapport à la classe d'âge de référence. Ensuite, nous observons une décroissance des coûts moyens pour les tranches d'âge plus âgées :

- Les conducteurs âgés de 51 à 65 ans ont un coût moyen de sinistres inférieur de 8%

par rapport à la référence.

- Les conducteurs âgés de 66 à 75 ans montrent une réduction de 9,0%.
- Pour les conducteurs âgés de 76 à 80 ans, la diminution est de 3,5%.
- Enfin, les conducteurs de plus de 80 ans ont un coût moyen de sinistres réduit de 5,2%.

Il est important de noter que bien que la tendance soit généralement à la baisse à partir de la tranche d'âge `51-65`, cette diminution ralentit chez les conducteurs de plus de 76 ans. Cette observation suggère que le coût moyen des sinistres est plus élevé pour les jeunes conducteurs débutants, probablement en raison de leur manque d'expérience et de prudence sur la route, ce qui entraîne plus de dégâts par sinistre. Pour les conducteurs plus âgés, bien que le coût moyen des sinistres soit plus faible par rapport à la classe de référence, il y a une légère augmentation chez les conducteurs les plus âgés (76 ans et plus). Cela peut être attribué à un déclin des réflexes et de la capacité de réaction, ce qui augmente les dégâts causés. Cette tendance est également confirmée dans le modèle additif généralisé présenté dans le chapitre suivant.

La variable `PARK` indique les types de parking. Par rapport à la catégorie de référence qui est la `Voie publique`, les résultats montrent que les véhicules garés dans un garage ou un parking clos ont un coût moyen des sinistres supérieur de 25,4% et que les véhicules garés dans un parking collectif ont un coût moyen des sinistres supérieur de 13,6%.

La variable `Zone_ZONIER` représente différentes zones géographiques du Zonier. Par rapport à la catégorie de référence représentée ici par la `Zone 2`, la `Zone 1` montre une légère augmentation de 1,2% du coût moyen des sinistres et la `Zone 3` montre une augmentation de 6,8%. De même, la `Zone 4` montre une augmentation de 15,1%, la `Zone 5` une augmentation de 69,4% et la `Zone 6` montre une augmentation de 33,6%. On observe, sauf dans les Zones 1 et 6, une augmentation générale des coûts moyens de sinistres en passant d'une zone risquée à une autre.

La variable `CRM_AUTO` montre que pour chaque augmentation d'un point du CRM, le coût moyen de sinistres augmente de 42,7%.

Enfin, la variable `EXERCICE` indique que le coût moyen des sinistres augmente de 5,9% d'un exercice à un autre, résultat confirmant les analyses sur l'évolution temporelle du coût moyen de la figure 2.5 du chapitre précédent.

Les modalités de la variable `VEHICULIER_COUT` sont toutes significatives, le tableau suivant présente les coefficients des modalités du véhiculier.

Variable	θ	$\exp(\theta)$	**p-value**
VEHICULIER_COUT[T.0]	0,2109	**1,2347**	0,0000
VEHICULIER_COUT[T.9]	-0,1818	**0,8339**	0,0000
VEHICULIER_COUT[T.11]	0,3766	**1,4574**	0,0000
VEHICULIER_COUT[T.4]	-0,4654	**0,6279**	0,0000
VEHICULIER_COUT[T.5]	0,5921	**1,8079**	0,0000
VEHICULIER_COUT[T.8]	0,7910	**2,2055**	0,0000
VEHICULIER_COUT[T.2]	1,0117	**2,7503**	0,0000
VEHICULIER_COUT[T.7]	-0,9512	**0,3862**	0,0000
VEHICULIER_COUT[T.12]	1,0610	**2,8892**	0,0000
VEHICULIER_COUT[T.1]	1,0905	**2,9757**	0,0000
VEHICULIER_COUT[T.10]	1,1587	**3,1857**	0,0000
VEHICULIER_COUT[T.3]	1,7976	**6,0351**	0,0000
VEHICULIER_COUT[T.6]	1,9395	**6,9552**	0,0000

Table 3.9: Résultats du modèle CM pour le véhiculier

L'analyse du tableau 3.9 montre que les coefficients des modalités les plus élevés sont observés pour les classes "3" et "6", nous réalisons pour cela une caractérisation du véhiculier construit. En faisant un classement des classes de véhicules en fonction des valeurs de coût moyen prédites ou en fonction des valeurs de coût moyen observées, les classes "3" et "6" se sont démarquées avec les plus grands coûts moyens et le classement respectent hiérarchiquement les classes en comparaison avec les coefficients du modèle. Une investigation a révélé que les classes "3" et "6" regroupent des véhicules luxueux, puissants et rapides, possédant des caractéristiques impressionnantes en termes de cylindrée. Ces véhicules sont des modèles haut de gamme avec des équipements luxueux, ce qui entraîne des coûts de réparation et de remplacement plus élevés. De plus, leur technologie avancée, l'image de marque prestigieuse, la valeur de revente élevée et la disponibilité limitée des pièces détachées contribuent à ces coûts moyens élevés. Les classes "3" et "6" englobent des véhicules qui, par leur nature luxueuse et leurs caractéristiques techniques avancées, engendrent des coûts plus élevés, confirmant que le véhiculier est un bon indicateur du risque associé pour le coût moyen.

Le graphique ci-dessous compare le coût moyen des sinistres prédit avec celui observé, en fonction des différentes classes de véhiculier, renommées selon le risque associé. Ce graphique montre une augmentation du coût moyen prédit et observé avec l'augmentation des classes du véhiculier.

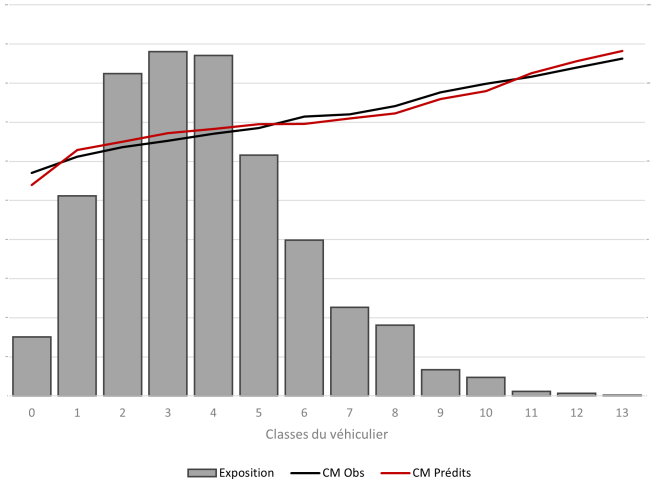

Figure 3.12: Coût moyen observé vs Coût moyen prédit

Dans la suite de l'étude, pour la variable `Zonier`, nous avons regroupé les Zones 1 et 2 ainsi que les Zones 5 et 6, nous avons regroupé également les classes "11", "12", "13" du véhiculier pour obtenir plus de représentativité en termes d'exposition.

3.7.2 Interprétation du modèle de fréquence de sinistres

Nous présentons à la suite l'interprétation du modèle de fréquence de sinistres. Le tableau ci-dessous résume les résultats du modèle à travers les coefficients significatifs du modèle de fréquence de sinistres :

Variables	θ	$\exp(\theta)$	**p-value**
Zone_ZONIER[T.Z1]	0,1222	**1,1299**	0,0004
Zone_ZONIER[T.Z3]	0,1243	**1,1324**	0,0002
Zone_ZONIER[T.Z4]	0,1939	**1,2140**	0,0000
Zone_ZONIER[T.Z5]	0,4084	**1,5045**	0,0000
Zone_ZONIER[T.Z6]	-0,0666	**0,9355**	0,5964
ACQU_MODE[T.Leasing]	0,1068	**1,1127**	0,0007
ACQU_MODE[T.Crédit]	0,0675	**1,0699**	0,0455
DETENTION_CONTRAT_ANNEE	-0,0321	**0,9684**	0,0001
CRM_AUTO	0,5227	**1,6867**	0,0000
ANTECEDENTS_ASSURANCE	-0,0027	**0,9973**	0,0007
ANCIENNETE_PERMIS	-0,0073	**0,9927**	0,0382

Table 3.10: Résultats de sortie du modèle GLM de fréquence de sinistres

Le tableau 3.10 présente les coefficients significatifs des variables du modèle de fréquence de sinistres. La variable `Zone_ZONIER` représente également ici les différentes zones géographiques. Par rapport à la catégorie de référence représentée ici par la `Zone 2` :

- La `Zone 1` montre une augmentation de 12,99% de la fréquence de sinistres.
- La `Zone 3` montre une augmentation de 13,24%.
- La `Zone 4` montre une augmentation de 21,40%.
- La `Zone 5` montre une augmentation significative de 50,54%.
- La `Zone 6` montre une réduction de 6,45% et n'est pas significatif.

La variable `ACQU_MODE` indique le mode d'acquisition du véhicule. Par rapport à la catégorie de référence, qui est l'acquisition du véhicule au comptant, les véhicules en `Leasing` ont une fréquence de sinistres supérieure de 11,27% et les véhicules acquis par `Crédit` ont une fréquence de sinistres supérieure de 6,99%.

La variable `DETENTION_CONTRAT_ANNEE` montre que pour chaque année supplémentaire de détention du contrat, la fréquence de sinistres diminue de 3,16%.

La variable `CRM_AUTO` montre que pour chaque augmentation d'un point du CRM, la fréquence de sinistres augmente de 68,97%.

La variable `ANTECEDENTS_ASSURANCE` indique que l'augmentation d'un mois d'antécédent d'assurance réduit la fréquence de sinistres de 0,27% .

Enfin, la variable `ANCIENNETE_PERMIS` indique que chaque année supplémentaire d'ancienneté du permis réduit la fréquence de sinistres de 0,73%.

Nous présentons également les coefficients associés à la variable véhiculier du modèle de fréquence de sinistres.

Variables	θ	$\exp(\theta)$	**p-value**
VEHICULIER_FREQ[T.3]	-0,4127	**0,6619**	0,0000
VEHICULIER_FREQ[T.1]	0,0381	**1,0388**	0,0000
VEHICULIER_FREQ[T.10]	-1,0599	**0,3468**	0,0000
VEHICULIER_FREQ[T.7]	0,0843	**1,0879**	0,0000
VEHICULIER_FREQ[T.6]	-2,7328	**0,0652**	0,0000
VEHICULIER_FREQ[T.5]	0,8012	**2,2282**	0,0000
VEHICULIER_FREQ[T.2]	0,9657	**2,6266**	0,0000
VEHICULIER_FREQ[T.8]	1,1321	**3,1021**	0,0000
VEHICULIER_FREQ[T.4]	1,6523	**5,2174**	0,0000
VEHICULIER_FREQ[T.9]	1,8082	**6,0994**	0,0000

Table 3.11: Résultats du modèle FREQ pour le véhiculier

Tout comme pour le cas du coût moyen, le classement des classes du véhiculier en fonction des fréquences prédites et observées met en évidence une distinction claire des classes "8", "9", et "4", ainsi qu'un ordre cohérent basé sur les coefficients obtenus du modèle. Une investigation plus approfondie de ces classes révèle qu'elles sont associées aux véhicules ayant les puissances et vitesses les plus élevées. En conséquence, les classes de véhiculier les plus risquées sont liées aux véhicules les plus dangereux, ce qui est logique.
Le graphique ci-dessous compare les fréquences de sinistres observées à celles prédites. Les classes ont été renommées en fonction de leur niveau de risque. L'analyse montre que, plus la classe du véhiculier est élevée, plus la fréquence de sinistres augmente.

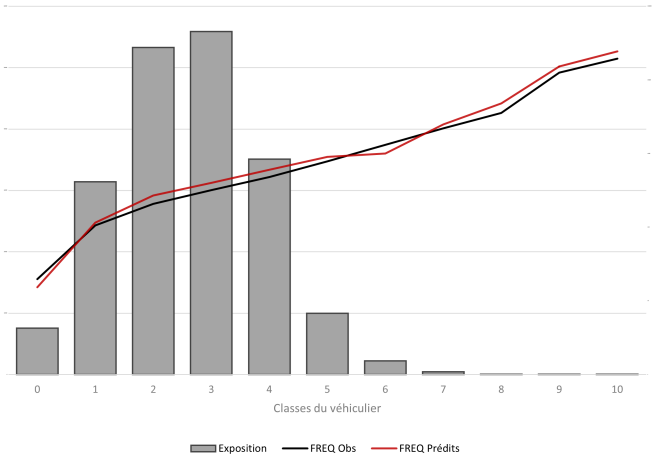

Figure 3.13: Fréquence de sinistres observée vs Fréquence de sinistres prédite

Dans la suite de l'étude, pour la variable `Zonier`, nous avons regroupé les Zones 1 et 2 ainsi que les Zones 5 et 6, nous avons également regroupé les classes de "7" à "10 pour

obtenir plus de représentativité en termes d'exposition.

En général, les résultats des deux modèles sont conformes du point de vue de l'observé.

3.7.3 Analyse : tarification et véhiculier

L'utilisation du véhiculier par l'assureur offre, avant tout, un meilleur alignement avec le marché et une meilleure réponse aux attentes des assurés. La construction de cette variable agrégée améliore la précision de la tarification, ce qui favorise l'acceptation des primes par les clients, qui perçoivent une certaine équité dans le calcul de leurs primes. Cette perception d'équité augmente la compétitivité de l'assureur sur le marché tout en lui permettant de maintenir une meilleure rentabilité.

Dans cette optique de rentabilité, en segmentant efficacement son portefeuille à l'aide de cette variable, l'assureur est en mesure de décider, pour chaque nouvelle souscription, d'accepter ou de refuser une demande en fonction de la classe du véhicule. Cette approche est couramment appliquée dans les Notes Techniques, qui décrivent les types de véhicules acceptés ou refusés, les conditions de souscription, ainsi que les garanties obligatoires... Ces notes, qui accompagnent les primes établies par l'actuaire, permettent à l'assureur de contrôler son chiffre d'affaires et d'optimiser au mieux la composition de son portefeuille.

Le véhiculier simplifie également le processus de modélisation en réduisant le nombre de variables liées aux caractéristiques des véhicules. En regroupant plusieurs variables en une seule (le véhiculier), l'actuaire diminue le risque de colinéarité entre les variables explicatives et limite les risques de surajustement du modèle.

Enfin, en renforçant la perception d'équité des clients à travers une tarification cohérente basée sur le véhiculier, l'assureur améliore l'expérience client et encourage la fidélité. Les clients, voyant que leur prime est calculée de manière justifiée en fonction des caractéristiques de leur véhicule, sont plus enclins à rester assurés auprès de l'entreprise. Cela permet à l'assureur de mettre en place des campagnes publicitaires efficaces, ciblant des segments de clientèle à forte valeur ajoutée et contribuant ainsi à accroître la rentabilité du portefeuille global.

Résumé du chapitre

Ce chapitre avait pour objectif de construire et de mesurer l'impact du véhiculier sur la tarification. Tout au long de ce chapitre, les différentes méthodes de construction et d'évaluation de l'apport du véhiculier ont été appliquées pour aboutir à une variable qui capture le risque associé aux véhicules. L'importance de cette variable dans la tarification a ensuite été démontrée à l'aide de métriques, en la comparant avec d'autres modèles ne prenant pas en compte le véhiculier construit. La caractérisation des véhiculiers construits montre que pour le modèle de coût moyen, les classes les plus risquées sont associées aux véhicules les plus luxueux et chers dont les coûts de réparation en cas d'accident sont élevés et les classes plus risquées pour la fréquence sont associées aux véhicules les plus dangereux . Le deuxième objectif de ce mémoire, consistant en la construction d'un modèle additif neuronal, est développé dans le dernier chapitre qui suit.

Chapitre 4

Modèle Additif Neuronal et impacts sur la tarification

Le dernier chapitre qui suit présente le modèle additif neuronal (NAM) et l'impact de son application à la tarification. Le modèle NAM, étant une combinaison du modèle additif généralisé (GAM) et de réseaux de neurones, nous commençons ce chapitre par l'introduction du modèle GAM que nous appliquons également à la tarification. Ensuite, il sera présenté rapidement les méthodes d'interprétation des modèles de machine learning pour aboutir à la présentation du modèle NAM. À ce niveau, nous présenterons la théorie mathématique puis son application à la modélisation des deux indicateurs de risque. Nous terminerons enfin par l'interprétation du modèle NAM et la comparaison des modèles.

4.1 Modèles Additifs généralisés

4.1.1 Mise en situation

Le modèle additif généralisé (GAM) est une extension du modèle linéaire généralisé (GLM). En 1986, Hastie et Tibshirani ont introduit cette version avancée des modèles GLM, qui se distingue par la forme du prédicteur linéaire. Dans un modèle GLM, le prédicteur linéaire est défini de sorte que la contribution de chaque variable explicative à l'espérance de la variable réponse soit mesurée par un paramètre estimé par maximum de vraisemblance dont la formule a été donnée dans le chapitre précédent. Le modèle GAM, en revanche, permet de s'affranchir de cette contrainte en offrant une plus grande souplesse et flexibilité. L'idée derrière le modèle GAM est de laisser les données s'exprimer en modélisant la contribution des variables explicatives à travers des fonctions de ces variables. Cela permet de capturer les relations non linéaires qui peuvent exister entre certaines covariables et la variable réponse. Une formulation mathématique simple du modèle GAM est donnée par :

$$g(\mu_i) = f_1(x_i^1) + f_2(x_i^2) + \ldots + f_p(x_i^p) + \epsilon_i \tag{4.1}$$

Le modèle ainsi présenté, avec f_j représentant les fonctions des variables explicatives, correspond à un modèle non paramétrique nécessitant l'estimation des fonctions utilisées. Le passage à une approche non paramétrique soulève alors la question cruciale de la définition des types de fonctions à intégrer dans le modèle. Le prédicteur linéaire des modèles GAM est en effet une combinaison linéaire de **fonctions lisses**. L'utilisation de fonctions lisses pose le problème du choix des types de fonctions lisses à utiliser et du **degré de lissage ou de régularité** à accorder à ces fonctions lisses.

4.1.2 Les fonctions splines

On s'intéresse à l'approximation d'un ensemble de points d'une covariable par une fonction lisse f.

L'idée générale est de trouver une base de fonctions dont la combinaison linéaire constitue la fonction lisse. La forme générale de la fonction lisse est alors écrite de la manière suivante :

$$f(x) = \sum_{j=1}^{k} b_j(x)\theta_j$$

avec b_j la j-ième fonction de base, θ_j le paramètre associé à b_j, et k le nombre de fonctions de base considérées.

En choisissant une base polynomiale de degré 4, $b_j = x^{j-1}$ avec $j \in \{1, 2, 3, 4, 5\}$, la fonction f devient :

$$f(x) = \sum_{j=1}^{5} \theta_j x^{j-1} = \theta_1 + \theta_2 x + \theta_3 x^2 + \theta_4 x^3 + \theta_5 x^4$$

Ainsi, en considérant une seule covariable, l'équation 4.1 devient :

$$g(\mu_i) = \theta_1 + \theta_2 x_i + \theta_3 x_i^2 + \theta_4 x_i^3 + \theta_5 x_i^4 + \epsilon_i$$

La base polynomiale est rarement utilisée en pratique pour approximer des points. En effet, la base polynomiale a de très bonnes propriétés sur un domaine restreint de points, $D = [0, 1]$, par exemple. Mais lorsque le domaine devient plus grand ou que les points sont issus d'une courbe, la base polynomiale est inefficace car elle est sujette au surapprentissage et à de grandes variations, rendant difficile l'approximation de la courbe dont sont issus les points d'approximation.

Nous introduisons alors les **fonctions splines**, qui sont les fonctions lisses les plus utilisées dans le modèle GAM et qui possèdent de meilleures propriétés que les fonctions lisses de base polynomiale.

Les fonctions splines d'interpolation linéaire

De manière générale, une fonction spline est une fonction définie par morceaux sur chacun desquels une fonction polynomiale est définie.

Comme dans le cas précédent, on s'intéresse à l'interpolation des points de coordonnées $\{(x_i, y_i), i = 1, 2, \ldots, n \ / \ x_i < x_{i+1}\}$ par une fonction f de sorte que $f(x_i) = y_i$. La résolution de ce problème admet une solution évidente, celle de relier tous les points par des droites. La courbe ainsi définie est issue de la fonction **spline d'interpolation linéaire**. La fonction spline d'interpolation linéaire est définie par morceaux avec des fonctions polynomiales de degré 1 dont la formule est donnée ci-dessous.

$$f(x) = \left(\frac{x_{i+1} - x}{h_i}\right) f(x_i) + \left(\frac{x - x_i}{h_i}\right) f(x_{i+1}) \quad \text{si } x_i \leq x \leq x_{i+1},$$

où $h_i = x_{i+1} - x_i$.

Avec cette nouvelle fonction d'interpolation, les fluctuations ne sont plus abruptes et la représentation des parties plates est possible, contrairement au cas de l'utilisation des fonctions lisses de base polynomiale. Mais ces nouvelles fonctions souffrent également de la mauvaise représentation dans le cas où les points sont issus d'une fonction h dont l'allure est courbée puisque nous avons dans ce cas $\forall x, |f(x) - h(x)| \geq 0$.

L'introduction des fonctions splines cubiques permet de corriger un peu ce biais.

Les fonctions splines cubique

L'idée mise en œuvre dans les fonctions splines cubiques qui améliore les résultats est la propriété de continuité de la fonction spline. En effet, il est exigé de la fonction spline que ses dérivées première et seconde soient continues sur les intervalles définis, mais également aux nœuds d'interpolation qui coïncident ici avec les points à interpoler. Il suffit alors de s'assurer de la continuité de la dérivée seconde pour obtenir les autres continuités. Pour y arriver, on force la dérivée seconde à être linéaire dans un intervalle donné et on construit la dérivée seconde de la fonction spline cubique de la manière suivante :

$$f''(x) = \left(\frac{x_{i+1} - x}{h_i}\right) f''(x_i) + \left(\frac{x - x_i}{h_i}\right) f''(x_{i+1}) \quad \text{si } x_i \leq x \leq x_{i+1},$$

De cette relation, on déduit alors la fonction f en intégrant deux fois la fonction f''.

$$\begin{aligned} f(x) &= \int \left(\int f''(x)\,dx \right) dx \\ &= \int \left(\int \left[\left(\frac{x_{i+1}-x}{h_i} \right) f''(x_i) + \left(\frac{x-x_i}{h_i} \right) f''(x_{i+1}) \right] dx \right) dx \\ &= \int \left[-\frac{(x_{i+1}-x)^2}{2h_i} f''(x_i) + \frac{(x-x_i)^2}{2h_i} f''(x_{i+1}) + C_1 \right] dx \\ &= \frac{(x_{i+1}-x)^3}{6h_i} f''(x_i) + \frac{(x-x_i)^3}{6h_i} f''(x_{i+1}) + C_1 x + C_2 \end{aligned} \tag{4.2}$$

Les constantes d'intégration C_1 et C_2 sont obtenues grâce aux conditions $f(x_i) = y_i$ et $f(x_{i+1}) = y_{i+1}$ dont les formules sont les suivantes :

$$\begin{aligned} C_1 &= \frac{y_{i+1}}{h_i} - \frac{y_i}{h_i} + f''(x_i)\frac{h_i}{6} - f''(x_{i+1})\frac{h_i}{6}, \\ C_2 &= \frac{y_i x_{i+1}}{h_i} - \frac{y_{i+1} x_i}{h_i} - f''(x_i)\frac{x_{i+1} h_i}{6} + f''(x_{i+1})\frac{x_i h_i}{6} - f''(x_{i+1})\frac{h_i^2}{6}. \end{aligned}$$

Ces constantes sont alors remplacées dans la dernière égalité de l'équation 4.2 et ensuite on développe les termes pour obtenir la forme simplifiée :

$$f(x) = a_i^-(x) y_i + a_i^+(x) y_{i+1} + c_i^-(x) f''(x_i) + c_i^+(x) f''(x_{i+1}) \quad \text{si } x_i \leq x \leq x_{i+1},$$

avec

$$\begin{aligned} a_i^-(x) &= \frac{x_{i+1}-x}{h_i}, & c_i^-(x) &= \frac{(x_{i+1}-x)^3/h_i - h_i(x_{i+1}-x)}{6}, \\ a_i^+(x) &= \frac{x-x_i}{h_i}, & c_i^+(x) &= \frac{(x-x_i)^3/h_i - h_i(x-x_i)}{6}. \end{aligned}$$

Il ne reste donc plus qu'à déterminer les dérivées secondes de f aux points d'interpolation qui ne sont pas connues. On utilise pour cela la condition de continuité des dérivées premières aux nœuds d'interpolation. La condition de continuité au point x_{i+1} de f' s'exprime par l'égalité : $f'(x_{i+1})_{[x_i, x_{i+1}]} = f'(x_{i+1})_{[x_{i+1}, x_{i+2}]}$. Appliquée à tous les points d'interpolation, on obtient le système d'équations suivant qui permet de déterminer les dernières inconnues de la fonction f :

$$\begin{cases} \dfrac{h_i}{6} f''(x_i) + \dfrac{x_{i+2}-x_i}{3} f''(x_{i+1}) + \dfrac{h_{i+1}}{6} f''(x_{i+2}) = \dfrac{y_{i+2}-y_{i+1}}{h_{i+1}} - \dfrac{y_{i+1}-y_i}{h_i}, \\ \text{avec } i = 1, 2, \ldots, n-2, \\ f''(x_1) = 0, \\ f''(x_n) = 0. \end{cases} \tag{4.3}$$

La fonction spline définie par tout ce procédé est appelée fonction **spline cubique naturelle**.

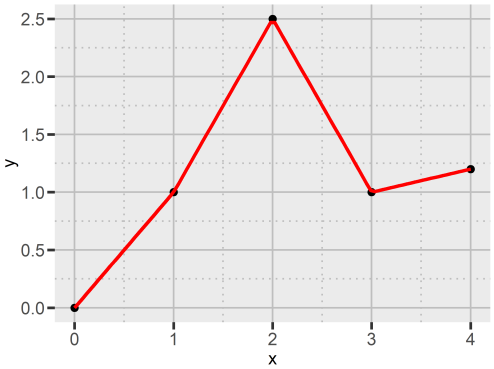

Figure 4.1: Spline linéaire

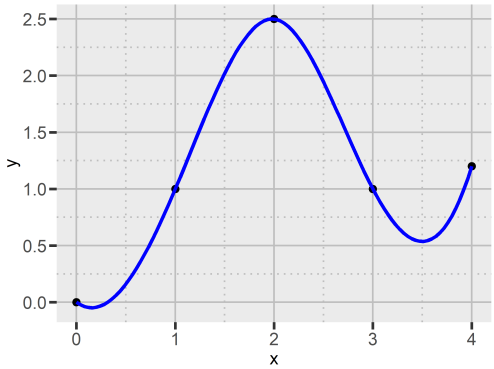

Figure 4.2: Spline cubique naturelle

4.1.3 Modèle additif univarié

Les définitions successives des modèles additifs univariés et multivariés permettent de comprendre le modèle additif généralisé. La formulation mathématique du modèle additif univarié est :

$$y_i = f(x_i) + \epsilon_i \tag{4.4}$$

C'est un modèle simple avec une seule variable explicative et les termes d'erreur ϵ_i suivent une loi normale d'espérance nulle et de variance σ^2 inconnue. Aussi, f est une fonction spline cubique naturelle. L'objectif est alors d'écrire f sous la forme suivante :

$$f(x) = \sum_{k=1}^{q} b_k(x)\theta_k$$

Les données de la variable explicative sont alors divisées en q intervalles. Il n'y a pas de règles spécifiques, mais généralement les données sont divisées suivant les quartiles. En choisissant q intervalles inférieurs au nombre de données, l'objectif n'est plus l'interpolation mais l'estimation de paramètres inconnus θ_j et δ_j correspondant aux valeurs de $f(\omega_j)$ et $f''(\omega_j)$, les ω_j étant les nœuds définis par les q intervalles. La fonction spline cubique f définie précédemment s'écrit alors comme :

$$f(x) = a_j^-(x)\theta_j + a_j^+(x)\theta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1} \quad \text{si } w_j \leq x \leq w_{j+1} \tag{4.5}$$

En utilisant la condition de continuité de f' et également que $f''(\omega_1) = f''(\omega_q) = 0$, on a l'égalité :

$$\frac{1}{h_j}\theta_j - \left(\frac{1}{h_j} + \frac{1}{h_{j+1}}\right)\theta_{j+1} + \frac{1}{h_{j+1}}\theta_{j+2} = \frac{h_j}{6}\delta_j + \left(\frac{h_j}{3} + \frac{h_{j+1}}{3}\right)\delta_{j+1} + \frac{h_{j+1}}{6}\delta_{j+2}$$

avec $j = 1, 2, \ldots, n-2$, que l'on peut écrire sous la forme matricielle suivante : $\mathbf{D}\theta = \mathbf{B}\delta^{-}$ et $\delta^{-} = (\delta_2, \delta_3, \ldots, \delta_{q-1})^T$. Enfin, on a : $\delta = \mathbf{F}\theta$ avec : $\mathbf{F} = \begin{bmatrix} 0 \\ \mathbf{F}^{-} \\ 0 \end{bmatrix}$ où $\mathbf{F}^{-} = \mathbf{B}^{-1}\mathbf{D}$.
L'équation 4.5 devient donc :

$$f(x) = a_j^{-}(x)\theta_j + a_j^{+}(x)\theta_{j+1} + c_j^{-}(x)\mathbf{F}_j\theta + c_j^{+}(x)\mathbf{F}_{j+1}\theta \quad \text{si } w_j \leq x \leq w_{j+1},$$

Nous admettons ensuite la forme de f qui est la suivante :

$$b_k(x) = \begin{cases} c_j^{-}(x)\mathbf{F}_{j,k} + c_j^{+}(x)\mathbf{F}_{j+1,k} + a_j^{+}(x) & \text{si } k = j+1 \\ c_j^{-}(x)\mathbf{F}_{j,k} + c_j^{+}(x)\mathbf{F}_{j+1,k} + a_j^{-}(x) & \text{si } k = j \\ c_j^{-}(x)\mathbf{F}_{j,k} + c_j^{+}(x)\mathbf{F}_{j+1,k} & \text{sinon} \end{cases} \tag{4.6}$$

La forme des matrices $\mathbf{B}$ et $\mathbf{D}$ et les démonstrations aboutissant au dernier résultat peuvent être consultées dans le mémoire sur les GAM de Steven Côté (2016).

Pour rappel, la définition de la fonction lisse dans le modèle concerne deux points. Le premier étant le choix de la fonction lisse à utiliser définie finalement comme la fonction spline. Le deuxième point n'a pas encore été abordé, il s'agit du degré de lissage à accorder à la fonction lisse. Nous introduisons alors une pénalisation dans le modèle qui permet de contrôler le niveau de régularité de la fonction et en même temps le nombre de nœuds final. Le problème de minimisation associé à l'équation du modèle est la détermination de la fonction f qui minimise :

$$\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda \int_{x_1}^{x_n} \{f''(x)\}^2 dx$$

Il est montré que la fonction spline cubique naturelle est la meilleure fonction lisse qui minimise ce critère, démonstration détaillée dans le livre de Wood (2006) sur les modèles GAM. En admettant que $\int_{x_1}^{x_n}\{f''(x)\}^2 dx = \lambda\theta^\top\mathbf{S}\theta$ et en posant $f(x_i) = \mathbf{X}_i\theta$, avec $\mathbf{X}_i = [b_1(x_i), b_2(x_i), \ldots, b_q(x_i)]$, alors l'ajustement du modèle revient à estimer les paramètres d'une régression pénalisée avec le critère suivant :

$$\|\mathbf{y} - \mathbf{X}\theta\|^2 + \lambda\theta^\top\mathbf{S}\theta$$

avec $\mathbf{S} = \mathbf{D}^\top\mathbf{B}^{-1}\mathbf{D}$ la matrice de pénalisation, $\mathbf{X}$ est appelé **matrice de design** dont la i-ième ligne est donnée par $\mathbf{X}_i$ et $\|.\|$ est la norme euclidienne.
La minimisation du critère basé sur les paramètres θ donne l'estimateur suivant :

$$\hat{\theta} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^\top\mathbf{y} \tag{4.7}$$

La démonstration est proposée dans le mémoire d'actuariat de Borel Wafo Kankeu (2023). Le choix du paramètre de pénalisation est également important dans le modèle. En effet, une grande valeur de λ conduit à un lissage très important de f, tandis qu'une

valeur faible de λ ne permettrait pas d'obtenir une fonction suffisamment lisse. Le choix du paramètre de pénalisation est alors effectué en minimisant le score GCV (Validation Croisée Généralisée) défini par :

$$\nu_g = \frac{n \sum_{i=1}^{n} (y_i - \hat{f}_i)^2}{[\mathrm{tr}(\mathbf{I} - \mathbf{A})]^2}$$

avec $\mathbf{A} = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{S})^{-1} \mathbf{X}^\top$ la matrice chapeau du modèle et tr la trace.

4.1.4 Modèle additif multivarié

Les bases du modèle additif étant posées, il convient donc de définir le modèle additif multivarié. La formulation mathématique du modèle additif multivarié est donnée par :

$$y_i = \alpha + f_1(x_i^1) + f_2(x_i^2) + \ldots + f_p(x_i^p) + \epsilon_i \tag{4.8}$$

Le modèle additif multivarié est défini comme le modèle additif univarié mais avec plusieurs variables explicatives, x_i^j désignant la i-ième observation de la j-ième variable explicative. Chaque covariable explique alors la variable réponse au travers de la fonction spline cubique.

Le modèle additif multivarié correspond donc aussi à une régression pénalisée, avec un paramètre de pénalisation pour chaque fonction spline. Ainsi, en considérant deux variables explicatives, le paramètre θ du modèle à estimer est donné par la formule :

$$\hat{\theta} = (\mathbf{X}^\top \mathbf{X} + \lambda_1 \mathbf{S_1} + \lambda_2 \mathbf{S_2})^{-1} \mathbf{X}^\top \mathbf{y} \tag{4.9}$$

avec :

- la formulation mathématique :

$$y_i = \alpha + f_1(x_i^1) + f_2(x_i^2) + \epsilon_i \tag{4.10}$$

- λ_1, λ_2 et $\mathbf{S}_1$, $\mathbf{S}_2$ sont les paramètres et matrices de pénalisation associés respectivement aux fonctions splines cubiques f_1 et f_2.
- La matrice $\mathbf{X}$ est définie par $\mathbf{X} = [\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2]$, où $\mathbf{X}_1$ et $\mathbf{X}_2$ sont les matrices des fonctions splines cubiques f_1 et f_2 respectivement.

La détermination des paramètres de pénalisation λ est également effectuée par la minimisation du score GCV.

Le modèle additif multivarié se comporte ainsi comme le modèle additif à la différence que le modèle additif multivarié souffre d'un problème d'identifiabilité. En effet, en rajoutant une constante et la retranchant dans la formulation mathématique du modèle additif multivarié avec deux variables explicatives, l'équation 4.10 devient :

$$y_i = \alpha + \tilde{f}_1(x_i^2) + \tilde{f}_2(x_i^1) + \epsilon_i$$

où $\tilde{f}_1(x_i^1) = f_1(x_i^1) + \text{cste}$ et $\tilde{f}_2(x_i^1) = f_2(x_i^1) - \text{cste}$

Ce problème se résout en appliquant une modification sur la matrice de f_1 avec la contrainte $\sum_{i=1}^{n} f(x_i) = 0$.

Mais on peut également redéfinir la formulation mathématique du modèle additif multivarié permettant de s'affranchir de ce problème d'identifiabilité. On définit pour cela la forme générale du modèle qui suit :

$$y_i = \alpha + f_1(x_i^1, x_i^2) + \epsilon_i$$

Les fonctions splines utilisées dans ce cas sont les **"Thin plate regression splines"** développées par Duchon (1977) dont la théorie est présentée dans le mémoire d'actuariat de Borel Wafo Kankeu (2023).

4.1.5 Modèle additif généralisé

Le modèle GAM est une extension du modèle additif multivarié avec l'introduction de la fonction de lien g et de la notion de loi de famille exponentielle que doit suivre la variable réponse Y de paramètre d'échelle ϕ, exactement comme l'extension du modèle linéaire au modèle GLM. La formulation mathématique du modèle GAM est donnée par :

$$g(\mu_i) = \mathbf{P}_i\beta + f_1(x_i^1) + f_2(x_i^2) + \ldots + f_p(x_i^p) + \epsilon_i \tag{4.11}$$

De même que dans le modèle GLM, $\mu_i = \mathbb{E}[Y]$. Le terme $\mathbf{P}_i\beta$ correspond à la partie paramétrique du modèle, avec $\mathbf{P}_i$ représentant la i-ième ligne de la matrice de design $\mathbf{P}$ et β les paramètres associés à cette partie paramétrique. Les f_j sont les fonctions splines cubiques, chacune étant écrite dans une base de fonctions comme définie jusqu'à présent. Nous donnons les définitions suivantes qui induiront les résultats de l'ajustement du modèle :

- b_{jk} est la k-ième fonction de base de f_j, $\mathbf{X}_j$ est la matrice de f_j et les contraintes d'identifiabilité sont appliquées.
- La matrice de design du modèle est obtenue par la concaténation en colonnes des matrices $\mathbf{P}$ et $\mathbf{X}_j$, et est donnée par : $\mathbf{X} = [\mathbf{P}, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]$.
- λ_j et $\mathbf{S}_j$ sont respectivement le paramètre de pénalisation et la matrice de pénalisation associés à f_j.

Avec ces définitions, le modèle GAM est un modèle GLM surparamétré dont la forme matricielle est la suivante :

$$g(\mu_i) = \mathbf{X}_i\theta + \epsilon_i \tag{4.12}$$

avec $\mathbf{X}_i$ la i-ième ligne de la matrice $\mathbf{X}$ définie et θ le paramètre du modèle constitué du paramètre β et les paramètres θ_j associés aux f_j.

Le paramètre θ est ensuite déterminé par maximum de vraisemblance. Cependant, la log-vraisemblance utilisée dans ce cas est une log-vraisemblance pénalisée dans l'objectif de contrôler le niveau de lissage des courbes f_j et le nombre de fonctions de base. La log-vraisemblance s'écrit sous la forme suivante :

$$l_p(y, \theta, \phi) = l(y, \theta, \phi) - \frac{1}{2} \sum_j \lambda_j \theta^\top \mathbf{S}_j \theta$$

La maximisation de cette log-vraisemblance pénalisée peut être résolue par une méthode itérative proposée par Wood (2006) appelée méthode P-IRLS (Penalized Iteratively Re-weighted Least Squares), qui est une variante de la méthode IRLS (Iteratively Re-weighted Least Squares) de Nelder et Wedderburn (1972).

4.1.6 Application à la tarification

Les modèles GAM offrent l'avantage d'utiliser des fonctions lisses pour représenter les relations entre les variables explicatives et la variable cible. Ces fonctions lisses permettent de visualiser les courbes qui traduisent ces relations de manière claire et interprétable. L'ajustement du modèle permet donc d'interpréter objectivement l'influence des covariables sur la variable à prédire.

Dans cette étude, nous avons appliqué les modèles GAM à la tarification en utilisant le package `pygam` de Python. Cette sous-section est dédiée à l'interprétation des graphiques obtenus suite à la modélisation des deux indicateurs de risque considérés tout au long de l'étude : le coût moyen et la fréquence de sinistres. Nous commencerons par l'interprétation des courbes des fonctions lisses pour le modèle de coût moyen, suivie de celles pour le modèle de fréquence de sinistres.

Les courbes des fonctions lisses issues du modèle de coût moyen sont présentées ci-dessous. En examinant ces courbes, nous pouvons identifier les tendances et les comportements des variables explicatives par rapport au coût moyen.

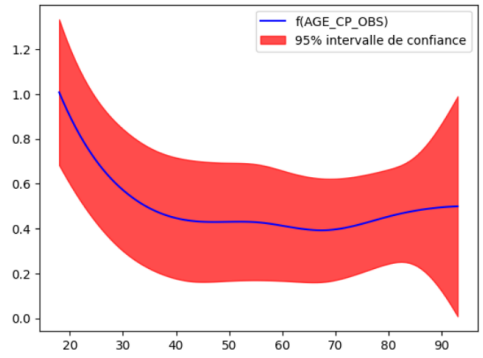

Figure 4.3: Âge du conducteur

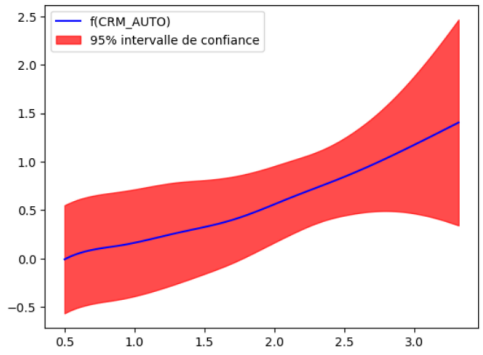

Figure 4.4: Coefficient Bonus-Malus

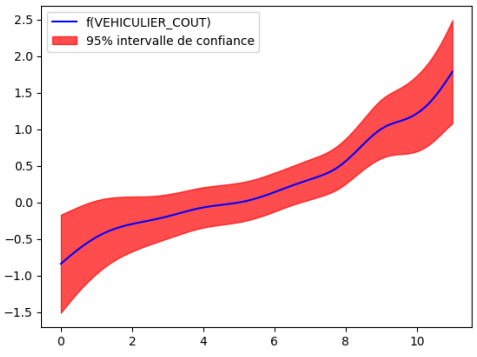

Figure 4.5: Véhiculier

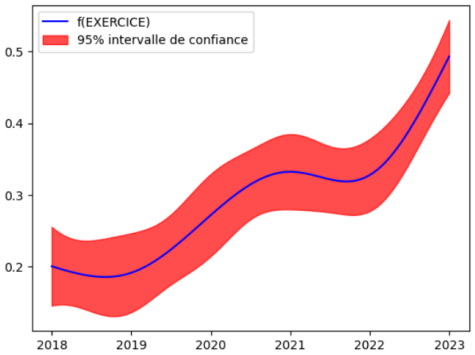

Figure 4.6: Exercice

Pour tous les graphiques d'interprétation, la ligne bleue représente la tendance estimée par le modèle, tandis que la bande rouge indique l'intervalle de confiance à 95%.

Âge du conducteur

La figure 4.3 montre l'effet de l'âge du conducteur sur la variable réponse modélisée par le GAM. Contrairement au modèle GLM, où il est nécessaire de créer des classes d'âge en raison de la non-linéarité de l'effet de l'âge sur les indicateurs de risque, le modèle GAM peut estimer directement les relations non linéaires. Par conséquent, l'âge du conducteur a été introduit sous forme quantitative dans le modèle GAM. L'analyse de cette figure révèle que le coût moyen diminue avec l'âge de 20 ans à environ 70 ans, où il atteint son minimum, avant d'augmenter légèrement jusqu'à 90 ans. Cette tendance confirme les résultats observés dans l'interprétation des coefficients significatifs du modèle GLM dans le chapitre précédent. En effet, les jeunes conducteurs (20-30 ans) et les conducteurs plus âgés (au-dessus de 70 ans) présentent un effet plus élevé sur la variable réponse, indiquant un risque accru dans ces groupes d'âge comparé aux conducteurs d'âge moyen. Les jeunes conducteurs étant novices dans la pratique de la conduite, ils commettent

donc plus de dégâts, tandis que les conducteurs très âgés perdent certains réflexes et capacités pendant la conduite.

Coefficient Bonus-Malus

La figure 4.4 montre l'effet du coefficient Bonus-Malus sur la variable réponse. Elle indique que l'effet estimé augmente de manière quasi-linéaire avec le coefficient Bonus-Malus. Un coefficient Bonus-Malus plus élevé est associé à un risque accru, et donc à un coût moyen plus élevé. La capacité du modèle GAM à identifier cette linéarité témoigne de sa robustesse dans l'identification des relations entre les variables explicatives et la variable cible.

Véhiculier

L'analyse de la figure 4.5 suggère que l'effet estimé du véhiculier augmente en fonction de l'augmentation des classes du véhiculier. Le résultat est cohérent en comparaison avec les résultats issus de l'analyse du véhiculier. Le coût moyen augmente en fonction de l'augmentation du risque de la classe.

Exercice

La figure 4.6 présente l'effet de l'année d'exercice sur la variable réponse. L'effet montre une tendance générale à la hausse au fil des ans, avec des fluctuations mineures. Cette augmentation reflète la hausse du coût moyen des sinistres due à l'inflation, évoquée comme problématique dans cette étude, et observée dans l'évolution temporelle du coût moyen de sinistres dans le chapitre sur l'analyse exploratoire ainsi que dans l'interprétation du coefficient de l'année d'exercice du modèle GLM du chapitre précédent.

Nous présentons ensuite les graphiques d'interprétation du modèle GAM de fréquence de sinistres.

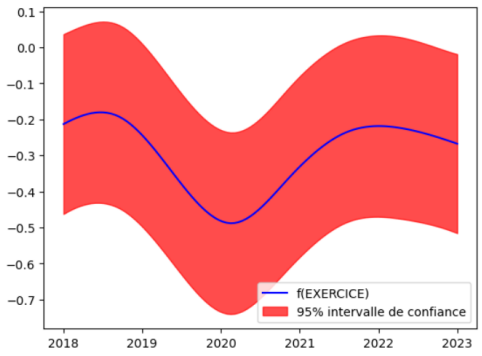

Figure 4.7: Exercice

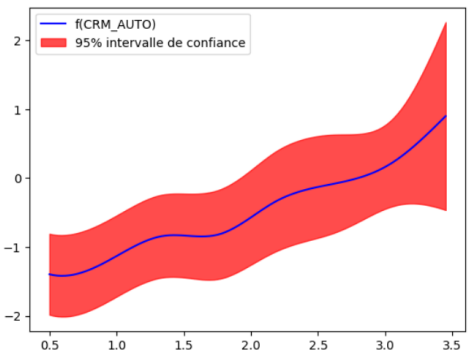

Figure 4.8: Coefficient Bonus-Malus

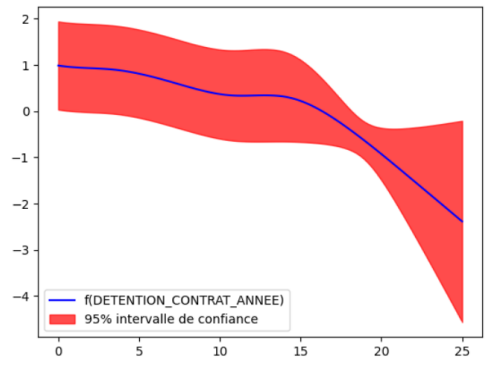

Figure 4.9: Nombre d'années de détention du contrat

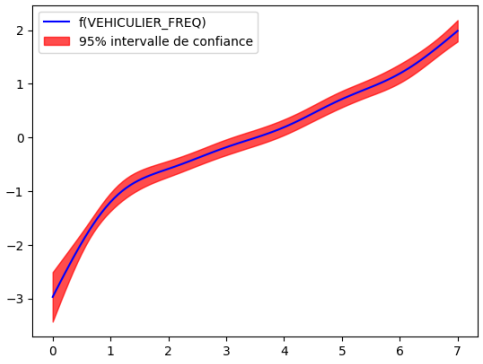

Figure 4.10: Véhiculier

Les figures ci-dessus présentent les effets de certaines variables sur l'estimation de la fréquence de sinistres. La figure 4.7 montre une baisse significative de la courbe pour l'année 2020. En se référant au graphique sur l'évolution temporelle de la fréquence de sinistres décrit dans le chapitre 2, il est logique d'observer cette même tendance, notamment pour l'année 2020 marquée par la crise sanitaire du COVID-19 et une fréquence de sinistres plus faible comparativement aux années précédentes ou suivantes. Cette particularité, captée par le modèle GAM, témoigne à nouveau de son efficacité dans la détermination des relations entre les variables explicatives et la variable cible.

Concernant l'effet estimé du coefficient Bonus-Malus, on retrouve une tendance presque linéaire par rapport à la fréquence de sinistres qui a été déjà constatée dans les autres analyses.

Les deux autres graphiques montrent également les effets attendus de l'observé sur la relation entre le nombre d'années de détention du contrat et la fréquence de sinistres, ainsi que celle entre le véhiculier et la fréquence de sinistres. En effet, la figure 4.9 indiquent une décroissance générale, montrant que la fréquence de sinistres tend à diminuer avec l'augmentation du nombre d'années de détention du contrat.
La dernière figure 4.10 montre une tendance attendue de la fréquence de sinistres en fonction des classes du véhiculier. La fréquence de sinistres augmente avec la croissance du risque portée par les classes.

Les résultats obtenus pour l'interprétation du coût moyen et de la fréquence de sinistres sont en général conformes à l'observé.

4.2 Interprétabilité des modèles de Machine Learning

Les modèles de machine learning sont devenus incontournables dans de nombreux domaines où la recherche de performance prédictive est primordiale. Cette amélioration

des performances a toutefois un coût, car ces modèles sont souvent perçus comme des "boîtes noires". Cet aspect de "boîte noire" se réfère à la perte d'interprétabilité que peuvent présenter ces modèles.

Cependant, leur efficacité est indéniable et leur utilité est prouvée dans certains secteurs, comme l'assurance. Néanmoins, l'utilisation des modèles de machine learning pour la tarification automobile nécessite une interprétabilité pour prendre des décisions tarifaires ou commerciales éclairées.

Dans des secteurs tels que la banque et la finance, où les réglementations sont strictes, les acteurs sont souvent réticents à utiliser des modèles de machine learning sans pouvoir expliquer leur fonctionnement. Cette réticence est compréhensible : il serait inconcevable d'expliquer à un client que son prêt immobilier a été refusé simplement parce que le modèle XGBoost utilisé prédit une probabilité de défaut élevée, sans fournir d'explications supplémentaires et humaines.

L'interprétabilité est alors définie comme la capacité d'une personne à saisir les raisons d'une décision du modèle. Elle correspond au degré de précision avec lequel les humains peuvent anticiper les résultats qu'il ne faut pas confondre avec l'explicabilité qui va plus loin en détaillant le processus par lequel le modèle arrive à ces résultats.

Face à cette contrainte d'interprétabilité, de nombreuses études scientifiques ont été menées, proposant diverses méthodes d'interprétation. Ces méthodes se divisent en deux catégories principales : les méthodes basées sur la structure du modèle et celles basées sur les prédictions, appelées méthodes d'interprétation post hoc. Étant donné la complexité des structures des modèles de machine learning, nous nous concentrons uniquement sur les méthodes d'interprétation post hoc.

Cette étude ne couvrira pas l'exhaustivité des méthodes d'interprétation ni le détail de leur théorie mathématique. Les lecteurs intéressés peuvent consulter le mémoire d'actuariat de Franklin FEUKAM KONHOUE (2023), qui se penche sur l'interprétabilité des modèles de tarification en actuariat, pour plus d'informations.

Les méthodes d'interprétation post hoc sont subdivisées en trois grandes catégories. La première catégorie inclut les méthodes locales et globales. Les méthodes locales sont valides pour des données dans un voisinage localement restreint, tandis que les méthodes globales sont applicables à la totalité des données. La seconde catégorie distingue les méthodes spécifiques, qui sont valides que pour certains types de modèles, et les méthodes agnostiques, applicables à tout type de modèle. La troisième catégorie se compose des méthodes dépendantes des données et des méthodes indépendantes des données, la distinction étant que les méthodes dépendantes nécessitent des données additionnelles pour l'application des méthodes d'interprétation, contrairement aux méthodes indépendantes des données.

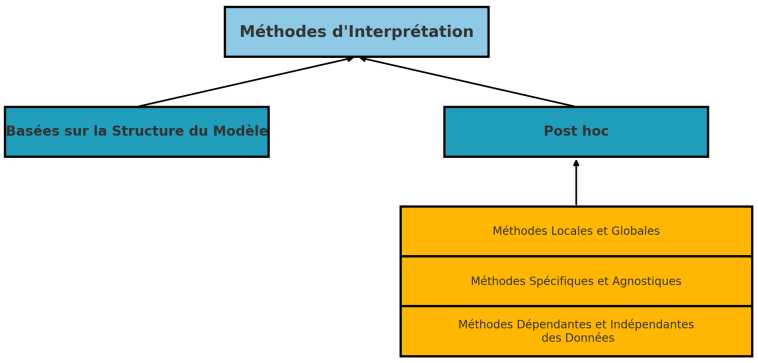

Figure 4.11: Catégorisation des méthodes d'interprétation

Nous présentons à la suite la méthode d'interprétation SHAP (SHapley Additive exPlanations), PDP (Partial Dependence Plot) et PFI (Permutation Features Importances) avec leur application.

4.2.1 Partial Dependence Plot

La méthode PDP, ainsi que la méthode ALE-Plot (qui ne sera pas présentée ici), sont les méthodes d'interprétation les plus répandues pour mesurer l'effet des variables explicatives sur la variable cible. Ces méthodes ont l'avantage de fournir un graphique illustrant la relation entre les covariables et la variable réponse. Les courbes de dépendance partielle, introduites par Friedman (2001), illustrent l'effet marginal d'une ou deux variables sur la prédiction finale du modèle. La méthode PDP est une méthode globale et, pour rappel, la théorie mathématique et les détails sur les méthodes d'interprétation ne sont pas présentés dans ce mémoire.

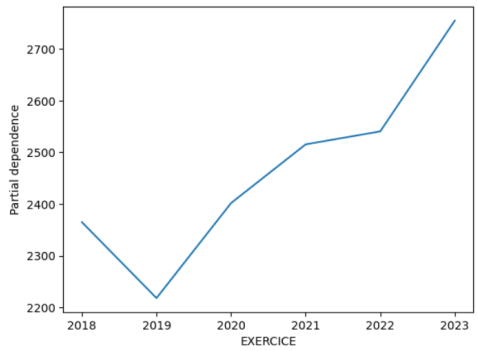

Figure 4.12: Exercice

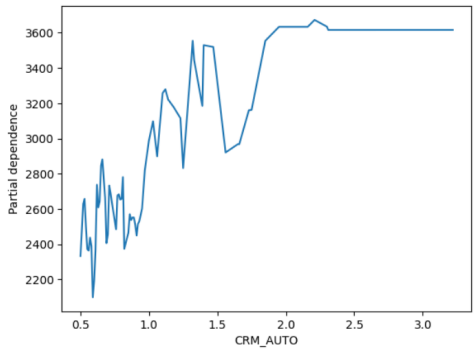

Figure 4.13: Coefficient Bonus-Malus

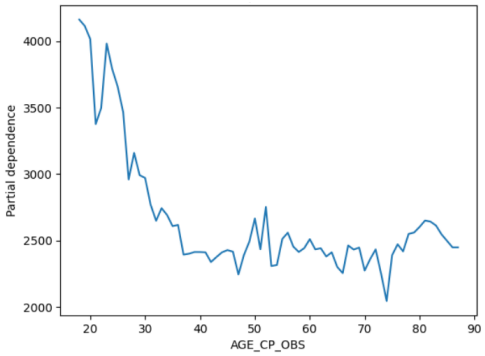

Figure 4.14: Âge du conducteur

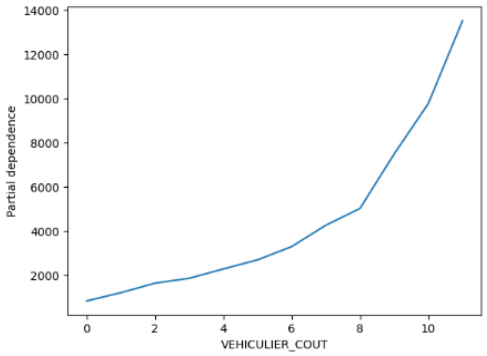

Figure 4.15: Véhiculier

Les graphes ci-dessus présentent les courbes de dépendance partielle du modèle XGBoost pour le coût moyen. Les courbes des graphiques 4.12, 4.13 et 4.14 montrent des effets globalement attendus de la variation de l'exercice, du coefficient CRM et de l'âge du conducteur, comparativement aux analyses des graphes du modèle GAM. Ces graphiques présentent quelques différences, notamment au niveau des figures 4.13 et 4.14 qui montrent beaucoup de variations, mais suivent néanmoins la tendance générale escomptée. La figure 4.15 présente la courbe de dépendance partielle du véhiculier. Ce graphique confirme la tendance haussière de la prédiction moyenne du coût moyen en fonction de l'augmentation du risque associé aux classes du véhiculier. La prédiction moyenne du cout moyen est d'autant plus grande que la classe augmente, en effet, elles ont été renommées en fonction de leur risque porté.

Nous présentons ensuite les courbes PDP relatives au modèle XGBoost de fréquence de sinistres que nous ne commenterons pas, les résultats étant ceux attendus en comparaison avec les analyses des graphes issus du modèle GAM.

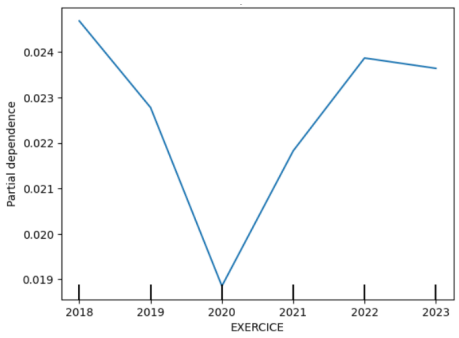

Figure 4.16: Exercice

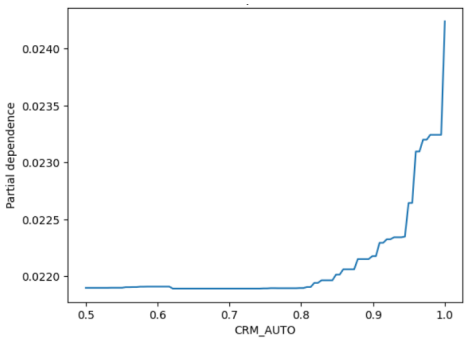

Figure 4.17: Coefficient Bonus-Malus

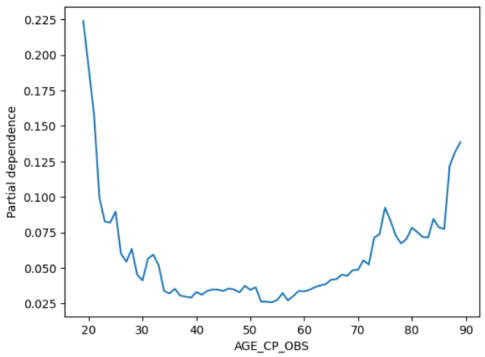

Figure 4.18: Âge du conducteur

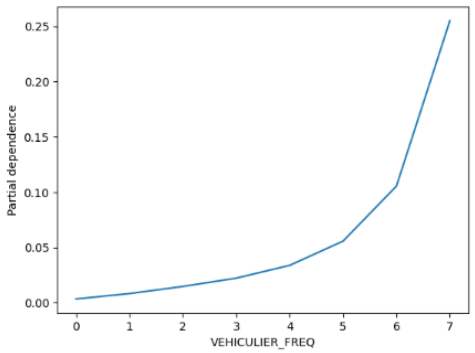

Figure 4.19: Véhiculier

Les graphiques ont été obtenus avec la fonction `PartialDependenceDisplay` de `Python`.

La méthode PDP est une méthode globale et facile à interpréter. L'interprétation du graphique correspond à la variation de la prédiction moyenne de la variable cible en fonction de la variation de la variable explicative, en considérant toutes les autres variables explicatives comme constantes fixée à leur moyenne. C'est d'ailleurs le principal inconvénient de cette méthode, car l'interprétation d'une variable repose sur la fixation des autres variables à leur moyenne. Cela n'est pas réaliste, car on interpréterait alors une moyenne de prédiction, par exemple, pour un assuré de 18 ans ayant un permis de conduire depuis 5 ans par exemple.

4.2.2 Permutation Features Importances

La méthode PFI introduite par Breiman en 2001 (Permutation Feature Importance) est l'une des premières méthodes utilisées pour mesurer l'importance de chaque variable explicative dans l'explication de la variable à prédire. Elle permet d'obtenir un classement des variables selon leur importance dans le modèle. La méthode PFI repose sur la perturbation des valeurs de la variable dont on mesure l'importance. Cette perturbation consiste à permuter l'ordre des valeurs prises uniquement par cette variable et à mesurer ensuite la variation de l'erreur de prédiction, telle que le MAE ou le RMSE. Définie de cette manière, une variable X est considérée plus importante qu'une autre variable Y si l'erreur de prédiction après la perturbation de la variable X est plus élevée que celle après la perturbation de la variable Y.

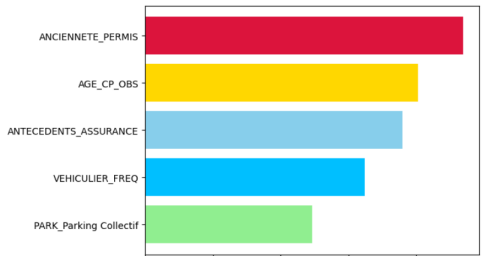

Figure 4.20: Top 5 des variables importantes pour le modèle de fréquence de sinistres

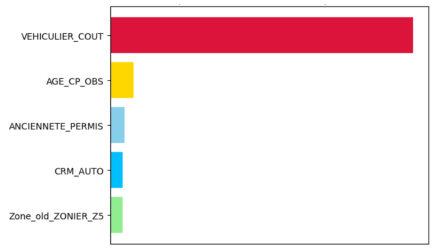

Figure 4.21: Top 5 des variables importantes pour le modèle de coût moyen

Les graphiques sont obtenus avec la fonction `permutation_importance` de `Python` sur le modèle XGBoost.

L'analyse de la figure 4.21, qui montre l'importance des variables dans le modèle de coût moyen, confirme la pertinence de la variable "véhiculier" mise en évidence dans le chapitre précédent. En effet, la variable "véhiculier" est de loin la plus importante dans la modélisation du coût moyen, suivie par les variables liées à l'âge du conducteur, l'ancienneté de son permis et le coefficient CRM. Le coefficient CRM avait déjà été identifié comme important lors des premières analyses de sélection de variables pour les modélisations des indicateurs de risque sans les variables véhiculières.
En ce qui concerne le modèle de fréquence de sinistres, il est logique de constater que les variables les plus importantes sont des variables beaucoup plus liées au profil de l'assuré.

La mesure de l'importance des variables à l'aide de la méthode PFI a l'avantage d'être simple et facile à interpréter. Cependant, cette simplicité constitue également son point faible. Une grande erreur induite par la perturbation des valeurs d'une variable ne fournit pas directement une explication claire de son importance sur la variable cible.

4.2.3 SHapley Additive ex-Planations

La méthode SHAP est fondée sur le concept des valeurs de Shapley. Pour un assuré i dans la base de données, la valeur de Shapley ϕ_{ij} associée à la variable j indique l'impact de la valeur x_{ij} (valeur prise par la variable j pour cet assuré) sur la prédiction $\hat{y}_i$, en comparaison à la prédiction moyenne de l'ensemble des assurés. La méthode SHAP utilise les valeurs de Shapley pour évaluer l'impact de chaque variable explicative sur une prédiction spécifique fournie par le modèle pour un individu. Cette approche permet une interprétation locale. La méthode SHAP offre également une interprétation globale mesurant l'importance des variables, servant ainsi d'alternative à la méthode PFI.

En pratique, l'importance des variables est fournie par la fonction `shap.summary_plot`, tandis que l'interprétation locale d'une prédiction est réalisée via la fonction `shap.plots.waterfall` de `Python`.

Nous présentons ici les graphiques relatifs à l'importance des variables. En effet, l'interprétation locale d'une prédiction en fonction des contributions des variables varie d'un individu à l'autre.

Figure 4.22: Importance des variables pour le modèle de coût moyen

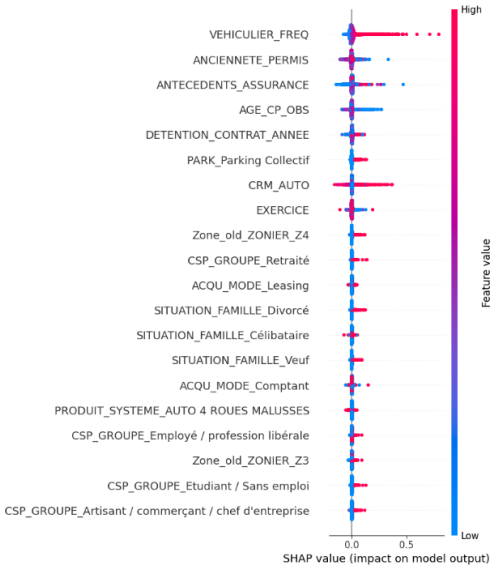

Figure 4.23: Importance des variables pour le modèle de fréquence de sinistres

Les graphiques ci-dessus présentent les variables du modèle XGBoost classées par ordre d'importance. Les résultats obtenus sont globalement conformes au classement des cinq variables les plus importantes fourni par la méthode PFI.

Nous rappelons que toutes les méthodes d'interprétation n'ont pas été abordées dans cette section. La principale difficulté de ces méthodes réside généralement dans la présence de corrélations entre les variables explicatives. Cependant, les méthodes fournissant l'importance des variables restent cohérentes pour mieux comprendre le fonctionnement de ces modèles. L'interprétabilité des modèles de machine learning est donc un sujet ouvert, avec des recherches en plein essor. Ces recherches ont permis de découvrir de nouvelles approches pour proposer l'interprétabilité de ces modèles et tenter de les rendre plus compréhensibles. Nous allons présenter une combinaison entre les modèles GAM et le modèle de réseaux de neurones, qui constitue une solution pour essayer de rendre les modèles d'apprentissage statistique plus interprétables.

4.3 Modèles Additifs Généralisés et Machine Learning

Les modèles GAM offrent deux avantages majeurs : ils permettent de prendre en compte les relations non linéaires entre la variable à expliquer et les covariables à travers les fonc-

tions splines des covariables, et ils offrent la possibilité de visualiser et donc d'interpréter ces fonctions splines. Il est crucial de comprendre ces relations, raison pour laquelle de nombreuses études se concentrent sur le développement de méthodes d'interprétation, comme discuté dans la section précédente. Ces recherches ont également ouvert de nouvelles perspectives pour l'interprétation des modèles de machine learning, notamment en les combinant avec les modèles GAM. L'idée derrière cette combinaison est de remplacer les fonctions splines des modèles GAM par des modèles de machine learning simples, tels que les arbres de décision et les réseaux de neurones. Nous nous intéresserons particulièrement à un modèle combinant le modèle GAM et les réseaux de neurones, que nous appellerons dans la suite **Modèle Additif Neuronal**.
Le modèle additif à structure arborescente est une combinaison du modèle GAM et des arbres de décision. Nous n'appliquerons pas ce modèle, mais son étude et son application ont été réalisées dans le mémoire d'actuariat de Markéta KRÚPOVÁ (2022), intitulé : "Construction d'un modèle de Machine Learning interprétable pour la tarification en assurance". Les lecteurs intéressés sont invités à consulter ce document pour obtenir davantage d'informations sur le modèle additif à structure arborescente.

4.3.1 Réseaux de Neurones

Les réseaux de neurones, également appelés "méthodes connexionnistes", ont été initialement conçus pour résoudre des problèmes en s'inspirant des modèles utilisés en neurobiologie. En effet, dans les années 1940, deux idéologies distinctes ont émergé. La première, portée par Von Neumann et al., exploitait une approche symbolique qui a conduit à l'émergence des concepts de mémoire, de processeur d'ordinateur, et d'une intelligence artificielle dite symbolique. La seconde, proposée par Mac Culloch et al., était basée sur une approche connexionniste, qui fonctionnait comme un mimétisme du comportement des neurones.

Les réseaux de neurones sont largement utilisés pour résoudre divers problèmes, tels que l'identification de formes, l'écriture, la reconnaissance vocale et même la reconnaissance d'images. Ils peuvent également être employés dans le cadre de l'apprentissage profond ou Deep Learning, dont l'objectif est d'obtenir des performances extraordinaires, révolutionnant ainsi le domaine de l'intelligence artificielle.

Nous aborderons uniquement la théorie des réseaux de neurones multicouches, également appelés réseaux de neurones artificiels. Initialement inspirés par le fonctionnement des neurones biologiques, ces réseaux de neurones ont ensuite été intégrés aux méthodes statistiques.

En 1943, Mac Culloch et Pitts ont proposé la première méthode connexionniste, basée sur le mimétisme d'un neurone du cerveau humain, désignée sous le terme de neurone formel.

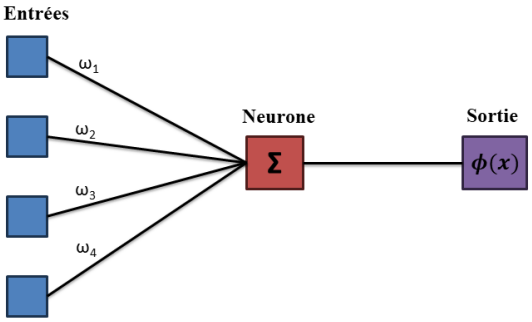

Figure 4.24: Neurone formel

Le neurone reçoit des influx nerveux, qui représentent les informations d'entrée. Il effectue alors une sommation des influx reçus et émet à son tour un influx, à condition que la somme des entrées dépasse un seuil d'activation. Cette sommation des entrées s'effectue par pondération des entrées avec des **poids** définis pour chacune d'elles. La pondération pour une entrée de la forme $x = (x_1, x_2, \ldots, x_n)$ est donnée par $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$. Étant donné le seuil θ, la **fonction d'activation** devient une notion importante dans ce contexte de construction du neurone formel. Le neurone émet un influx si la somme pondérée des entrées dépasse le seuil, sinon il n'émet rien. La fonction d'activation classique est alors la fonction indicatrice. En définissant la somme pondérée des entrées par $\omega \cdot x := \sum_{i=1}^{n} \omega_i x_i$ et en utilisant la fonction d'activation indicatrice, la sortie du neurone formel est donnée par : $\phi(\omega \cdot x - \theta) = \mathbb{1}_{\omega \cdot x - \theta > 0}$. Nous pouvons réécrire cela sous la forme : $\phi(w \cdot \tilde{x}) = \mathbb{1}_{w \cdot \tilde{x} > 0}$ avec $w = (w_1 = \omega_1, w_2 = \omega_2, \ldots, w_n = \omega_n, \omega_{n+1} = \theta)$, $\tilde{x} = (x_1, x_2, \ldots, x_n, x_{n+1} = -1)$ et $w \cdot \tilde{x} := \sum_{i=1}^{n+1} w_i \tilde{x}_i$.

Il existe également d'autres fonctions d'activation couramment utilisées, telles que :

- Fonction d'activation linéaire : $\phi(x) = x$
- Fonction d'activation sigmoïde : $\phi(x) = \frac{1}{1+\exp(-x)}$
- Fonction d'activation ReLU : $\phi(x) = \max(0, x)$
- Fonction d'activation tangente hyperbolique : $\phi(x) = \frac{1-\exp(-2x)}{1+\exp(-2x)}$
- Fonction d'activation ELU : $\phi(x) = x\mathbb{1}_{x>0} + \alpha(\exp(x) - 1)\mathbb{1}_{x<0}$

La première application concrète du neurone formel a été réalisée avec le Perceptron de Rosenblatt, qui utilise un ensemble de ces neurones formels pour l'identification d'images. Dans cette approche, les neurones formels sont disposés en plusieurs couches successives et sont interconnectés, à l'image des neurones humains. Les **réseaux de neurones artificiels** actuels sont ainsi une extension directe du Perceptron de Rosenblatt.

Nous allons nous concentrer sur les réseaux de neurones à une couche cachée, la généralisation étant réalisée de la même manière avec les réseaux de neurones à plusieurs couches cachées.

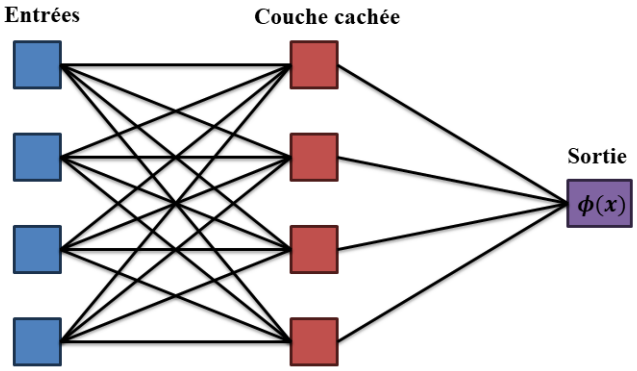

Figure 4.25: Réseaux de neurones à une couche cachée

Le réseau de neurones à une seule couche cachée est une extension du neurone formel de Mac Culloch et Pitts. En effet, nous sommes désormais en présence de plusieurs neurones dans la première couche. Chaque neurone de la couche cachée reçoit un influx provenant de chaque entrée. Ensuite, comme dans le cas du neurone simple, chaque neurone de la couche cachée émet également un influx, déterminé par la fonction d'activation ϕ_c du neurone. La sortie du neurone est obtenue en appliquant la fonction d'activation ϕ_c à la somme pondérée des poids des influx qu'il reçoit. À ce stade, chaque neurone a émis une sortie qui correspond à un influx reçu par la couche de sortie. La couche de sortie est généralement composée d'un seul neurone qui calcule la sortie finale du réseau. Lorsqu'il s'agit d'un problème de régression, une fonction d'activation linéaire est utilisée pour la couche de sortie, tandis que dans le cas d'une classification binaire, la fonction sigmoïde est souvent employée pour obtenir une sortie qui est une probabilité appartenant à l'intervalle $[0, 1]$. Au niveau de la couche de sortie, le même mécanisme est appliqué en appliquant la fonction d'activation ϕ_s à la somme pondérée des influx provenant de la première couche.

Nous allons maintenant nous intéresser à l'application de cette construction à un perceptron simple. La formalisation mathématique du perceptron permet de comprendre son application à une base de données d'apprentissage.

Soit l le nombre d'entrées, m le nombre d'unités dans la couche cachée, n le nombre d'unités dans la couche de sortie, ϕ_c et ϕ_s les fonctions d'activation réelles d'une variable réelle, $M^c = \left(M^c_{ij}\right)_{i=1,\dots,l+1,j=1,\dots,m}$ la matrice des poids de la couche cachée, et $M^s = \left(M^s_{jk}\right)_{j=1,\dots,m+1,k=1,\dots,n}$ la matrice des poids de la couche de sortie. Le perceptron à une couche cachée est alors défini par la fonction $P(M^c, M^s, \phi_c, \phi_s)$ de $\mathbb{R}^l$ dans $\mathbb{R}^n$, telle que

$$P(M^c, M^s, \phi_c, \phi_s)(x) = s$$

avec $x = (x_1, x_2, \ldots, x_l)$ et $s = (s_1, s_2, \ldots, s_n)$, où

$$s_k = \phi_s \left(\sum_{j=1}^{m} M^s_{jk} \, \phi_c \left(\sum_{i=1}^{l+1} M^c_{ij} x_i \right) - M^s_{m+1,k} \right)$$

et $x_{l+1} = -1$.

Nous considérons ensuite une base de données d'apprentissage définie par : $L_p = \{(x^r, y^r = f(x^r)) \in \mathbb{R}^l \times \mathbb{R}^n, r = 1, \ldots, p\}$. L'objectif est d'approximer la fonction f à l'aide d'un perceptron en utilisant les p données (x^r, y^r) de la fonction. Nous fixons alors les paramètres m et n du perceptron, les inconnues du problème étant les matrices de poids M^c et M^s. Nous définissons ainsi un problème de minimisation basé sur la fonction de perte quadratique définie par

$$E((M^c_{ij}), (M^s_{jk})) = \sum_{r=1}^{p} \left\| f(x^r) - s^r \right\|^2,$$

où $s^r = P(M^c, M^s, \phi_c, \phi_s)(x^r)$ et $\|.\|$ est la norme euclidienne.

Nous cherchons plus précisément les matrices M^c et M^s qui permettent de minimiser E. Pour ce faire, nous utilisons **l'algorithme de descente de gradient** pour résoudre le problème d'optimisation. Cet algorithme de descente de gradient est ensuite utilisé dans le célèbre **algorithme de rétropropagation du gradient** pour mettre à jour les poids du réseau de neurones, permettant ainsi de se rapprocher le plus possible de la fonction f. L'algorithme distingue deux phases : une phase de **propagation avant**, pour calculer la sortie du réseau de neurones, suivie d'une phase de **rétropropagation**, après calcul de l'erreur issue de la phase de propagation avant. L'erreur est ensuite propagée couche par couche vers l'arrière, afin de mettre à jour successivement les poids obtenus grâce à l'algorithme de descente de gradient.
Généralement, les hyperparamètres du modèle de réseau de neurones sont le nombre d'unités m dans la couche cachée et le nombre d'itérations utilisées dans l'algorithme de descente de gradient. Ces hyperparamètres peuvent faire l'objet d'une validation croisée afin d'obtenir la meilleure performance possible.

Les bases étant posées concernant le modèle de réseau de neurones, nous introduisons sa combinaison avec le modèle GAM.

4.3.2 Modèle Additif Neuronal

Le **Modèle Additif Neuronal** (NAM) combine la structure additive du modèle GAM avec des réseaux de neurones, un modèle de machine learning. Le principe général du modèle NAM consiste à entraîner un réseau de neurones pour chaque variable, ce qui permet à chaque variable d'apporter une contribution distincte à la prédiction finale. Cette approche rend chaque variable interprétable, les fonctions splines étant remplacées

par des réseaux de neurones dont les prédictions individuelles sont représentées par des courbes.

Les modèles NAM sont encore en développement, avec diverses versions proposées par différents chercheurs. Parmi celles-ci, on distingue le modèle NAM développé par l'équipe de "Google Research Team" (Agarwal et al., 2020), le modèle CANN (Combined Actuarial Neural Network) de Schelldorfer et Wüthrich (2019), et le modèle **IGANN** (Interpretable Generalized Additive Neural Networks) proposé en 2023 par Kraus et al. Nous utiliserons ici le modèle IGANN, qui est directement implémenté en Python via le package `igann`.

La forme mathématique du modèle IGANN est exprimée par l'équation suivante :

$$\hat{y} = \theta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

Cette forme générale découle de l'équation initiale suivante :

$$\hat{y} = \underbrace{\langle \theta, \mathbf{x} \rangle + \theta_0}_{\text{composante linéaire}} + \underbrace{\sum_{l=1}^{L} s_l g_l(\mathbf{x})}_{\text{composante non linéaire}} \tag{4.13}$$

où $\mathbf{x}$ représente l'observation à prédire, vecteur de p éléments, θ le vecteur des coefficients associés, et θ_0 l'intercept de la composante linéaire. Le modèle agrège les sorties de plusieurs réseaux de neurones, dont les fonctions sont notées g_l, avec un paramètre d'agrégation s_l.

La définition de la fonction g associée aux réseaux de neurones du modèle est cruciale pour comprendre la structure initiale du modèle. Dans le modèle IGANN, les réseaux de neurones sont agrégés, chaque réseau étant constitué d'une couche cachée. La particularité de ce réseau de neurones réside dans le fait que chaque élément du vecteur d'entrée est modélisé par un sous-réseau de neurones à une couche cachée. L'objectif est de capturer l'effet de chaque covariable sur la variable à prédire. Pour obtenir l'effet additif, la sortie du modèle est calculée comme la somme des sorties de chacun de ces sous-réseaux de neurones.

La fonction g, définie par un réseau de neurones à une couche cachée avec N unités, pour un modèle à p variables explicatives, est donnée par :

$$g_\beta(x) = \sum_{k=1}^{p} \sum_{j=1}^{N} \alpha_j^k \phi(x_k w_j^k)$$

où $\beta = (\alpha^1, \alpha^2, \ldots, \alpha^N, w^1, w^2, \ldots, w^p)$, avec $\alpha^k \in \mathbb{R}^N$ et $w^k \in \mathbb{R}^N$.

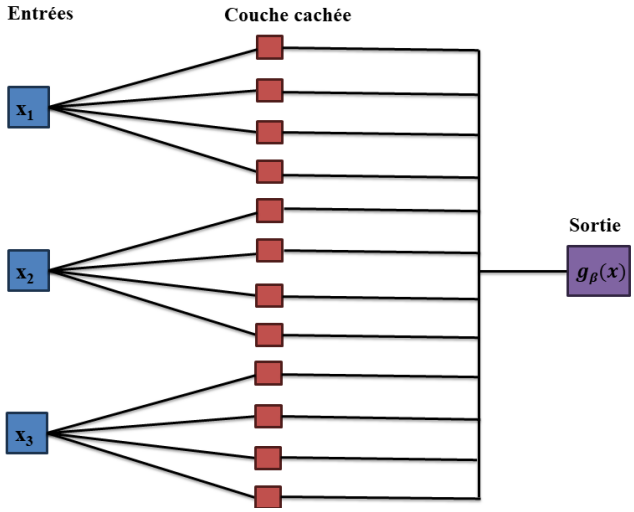

Figure 4.26: Illustration du réseau de neurones associé à g_β comportant trois entrées et 4 unités dans la couche cachée de chaque sous-réseau

L'agrégation des réseaux de neurones est réalisée successivement. En considérant une base de données de n observations (x^i, y^i), le choix du vecteur de poids β du réseau de neurones inclus dans g à l'instant t est déterminé par la fonction g_{β_t} qui minimise la fonction de perte pénalisée définie par :

$$\mathcal{L}^{(t)}(\beta_t) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\hat{y}^{i,t-1} + s_t g_{\beta_t}(x^i), y^i\right) + \frac{\lambda}{n}\|\beta_t\|^2.$$

où ℓ est la fonction de perte standard, donnée par $\ell(\hat{y}, y) = (y - \hat{y})^2$ pour un problème de régression, et $\ell = \ln(1 + \exp(-\hat{y}y))$ pour un problème de classification. Ici, $\hat{y}^{i,t-1} = \langle \theta, x^i \rangle + \theta_0 + \sum_{l=1}^{t-1} s_l g_{\beta_t}(x^i)$, et λ est le paramètre de régularisation.

L'agrégation des réseaux de neurones représentés par g_{β_t} est réalisée jusqu'à une condition d'arrêt (comme la stabilité d'une métrique) ou jusqu'à l'atteinte du nombre d'itérations T définies dans l'algorithme du modèle IGANN, avec le paramètre d'agrégation s_l supposé constant et égal à s. Les détails et spécificités de la résolution du problème de minimisation sont présentés dans l'article publié par Elsevier B.V sur le modèle IGANN.

Les poids des réseaux de neurones étant connus aux différents instants l d'agrégation, on peut alors définir la fonction f_k par :

$$f_k(x_k) = \theta_k x_k + s\sum_{l=1}^{L}\sum_{j=1}^{N} \alpha_j^{k,l}\phi(x_k w_j^{k,l})$$

En remplaçant d'abord les fonctions g_β et en retrouvant les fonctions f_k dans l'équation 4.13, nous retrouvons la forme mathématique générale du modèle donnée par :

$$\hat{y} = \theta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

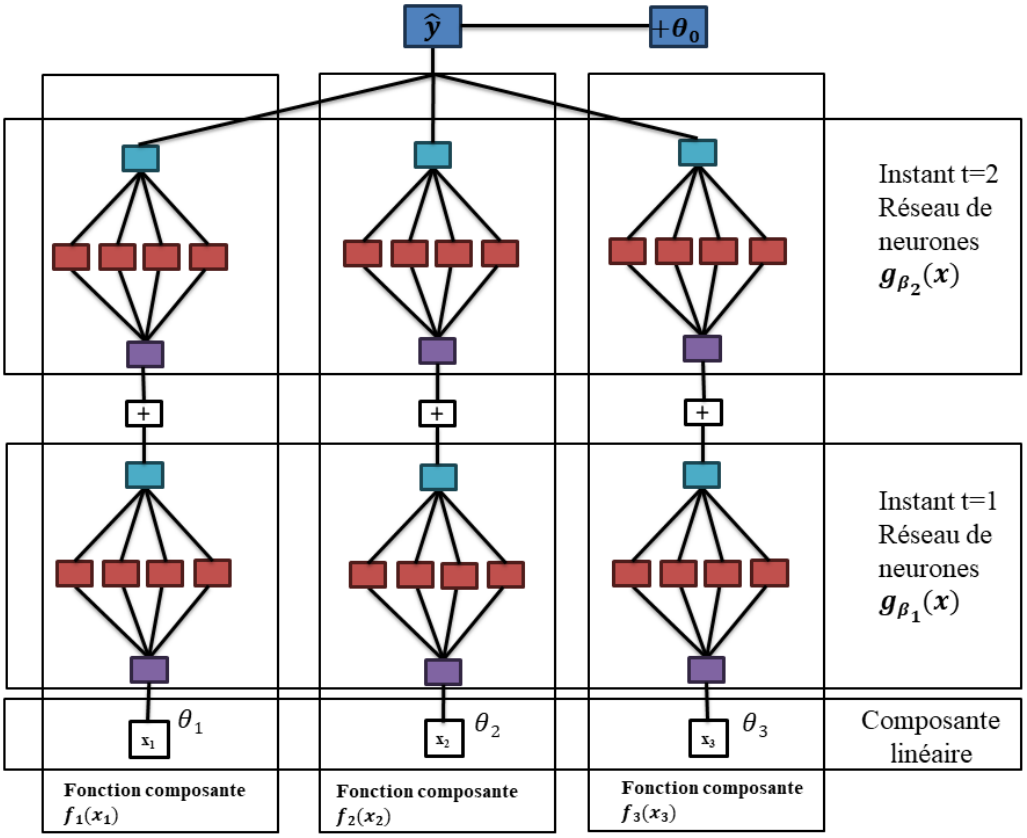

Figure 4.27: Illustration du modèle IGANN pour deux instants d'agrégation

4.4 Interprétation du Modèle IGANN

Le modèle IGANN propose trois types d'interprétation.

La première, déjà mentionnée à plusieurs reprises, est une interprétation globale basée sur l'analyse des courbes des fonctions composantes. Ces courbes permettent de comprendre l'influence des covariables sur la variable cible.

La deuxième interprétation repose sur une approche locale de l'interprétabilité. Ces courbes ne se limitent pas à représenter des relations, elles offrent **une description exacte de la façon dont le modèle IGANN calcule une prédiction**. Pour une observation donnée, il suffit de récupérer les valeurs correspondantes sur l'axe

des ordonnées de chaque courbe associée aux variables de l'observation, puis de les additionner, en incluant l'intercept θ_0 du modèle, pour obtenir la prédiction.

La troisième interprétation offerte par le modèle est une méthode d'évaluation globale de l'importance des variables. Bien que la théorie du calcul de cette importance ne soit pas présentée ici, le modèle IGANN a l'avantage de pouvoir afficher directement l'importance relative de chaque variable sur les courbes des fonctions composantes.

4.4.1 Modèle de coût moyen

Les graphiques suivants illustrent les courbes des fonctions composantes associées aux quatre premières variables les plus importantes du modèle IGANN ajusté sur le coût moyen.
Ces représentations montrent les relations entre les covariables et la variable cible. Le pourcentage d'importance de chaque variable est affiché en haut de chaque courbe.

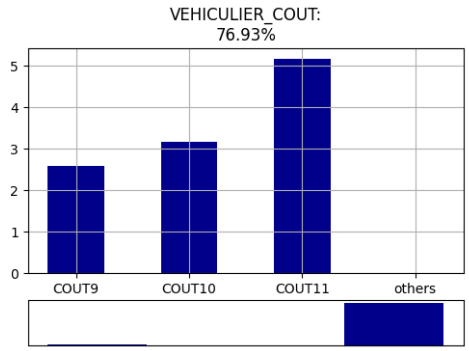

Figure 4.28: Véhiculier

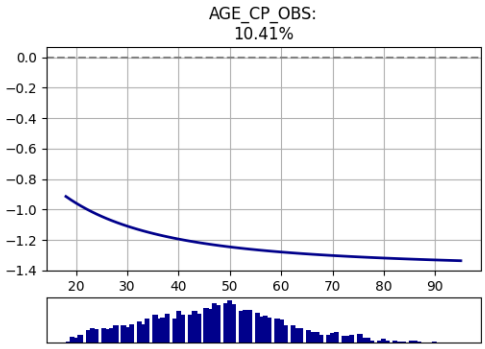

Figure 4.29: Âge du conducteur

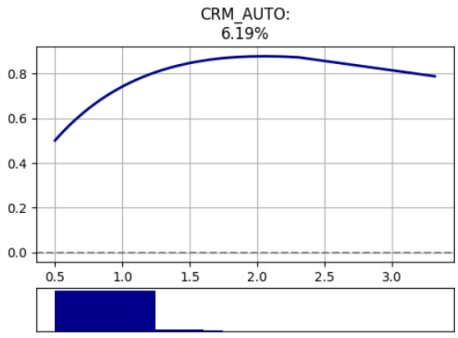

Figure 4.30: CRM

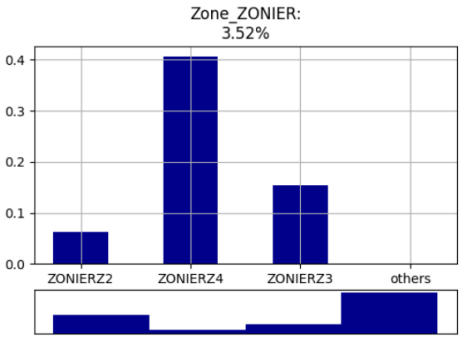

Figure 4.31: Zonier

Le véhiculier que nous avons construit joue un rôle crucial dans tous les modèles, en

particulier dans le modèle IGANN. Le véhiculier s'affiche comme une variable déterminante pour la modélisation du coût moyen, avec une importance évaluée à 76,93% dans le modèle, suivie par l'âge du conducteur (10,41%), le coefficient bonus-malus (6,19%), et le Zonier (3,52%).

Les relations observées entre ces covariables et la variable dépendante confirment les tendances déjà identifiées. Le graphique sur le véhiculier montre une relation croissante avec l'augmentation du risque de la classe en exerçant un impact positif sur le coût moyen.

L'âge du conducteur, quant à lui, a un impact négatif sur la prédiction du coût moyen. Comme la prédiction résulte de l'addition des contributions individuelles, le coût moyen diminue à mesure que l'âge du conducteur augmente.

Le coefficient bonus-malus, de son côté, augmente avec l'augmentation du CRM jusqu'à un seuil d'environ 2,4, puis diminue légèrement pour les valeurs de CRM plus élevées. Le CRM a un impact globalement positif sur la prédiction du coût moyen.

Enfin, le Zonier, désormais composé de quatre classes après regroupement, montre une relation croissante avec le coût moyen : ce dernier augmente avec l'augmentation du risque associé aux zones.

4.4.2 Modèle de fréquence

Nous présentons également les quatre variables les plus importantes du modèle IGANN ajusté sur la fréquence de sinistres.

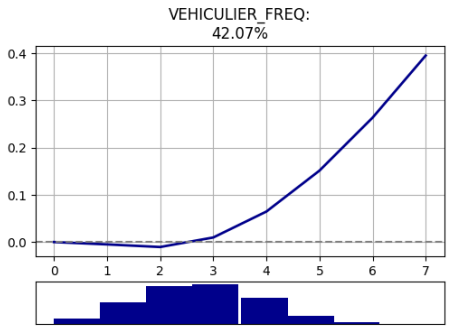

Figure 4.32: Véhiculier

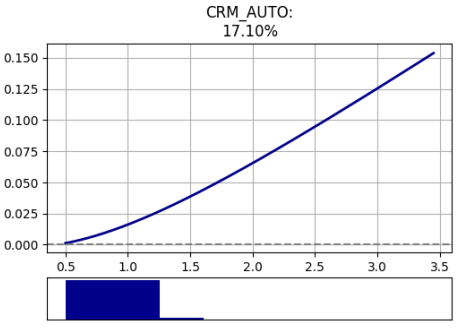

Figure 4.33: CRM

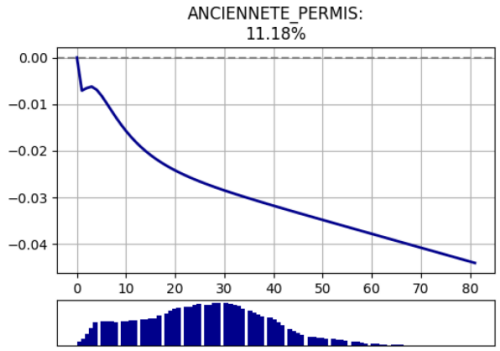

Figure 4.34: Ancienneté du permis

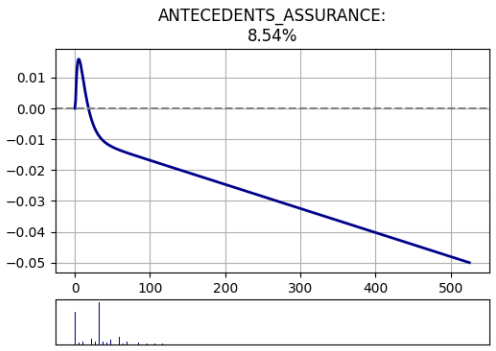

Figure 4.35: Antécédents d'assurance

Les figures illustrent globalement les relations attendues entre les covariables et la variable cible. Le véhiculier se distingue comme la variable la plus importante (42,07%), avec une relation croissante en fonction de l'augmentation de la classe de risque. On observe également que le modèle IGANN est capable de reconnaître les relations linéaires lorsque cela est nécessaire, comme en témoigne le graphique du CRM, qui est la deuxième variable la plus importante (17,10%) après le véhiculier. Les graphiques relatifs à l'ancienneté du permis et aux antécédents d'assurance montrent également des relations décroissantes linéaires à partir d'un certain seuil. Ces relations sont conformes aux attentes : la fréquence de sinistres diminue avec l'augmentation des antécédents d'assurance ou de l'ancienneté du permis.
Les histogrammes au bas des graphiques illustrent la distribution des variables dans la base de données.

4.5 Challenge du Modèle IGANN

Le modèle IGANN montre de meilleures performances lorsque la variable cible est standardisée. Nous avons donc appliqué une standardisation de la variable cible pour les deux modèles. La standardisation consiste à centrer et réduire la variable. La standardisation s'effectue comme suit :

$$y_s = \frac{y - \mu}{\sigma}$$

où y_s est la variable standardisée, y est la variable cible d'origine, et μ et σ sont respectivement la moyenne et l'écart-type empiriques de y.

L'avantage de combiner un modèle GAM avec un modèle de machine learning est qu'il est désormais possible d'effectuer une validation croisée sur les hyperparamètres du modèle. Comme dans le cas des modèles de machine learning, nous avons réalisé une optimisation des paramètres pour obtenir la meilleure performance possible. Nous avons utilisé la méthode Grid Search pour optimiser les paramètres du modèle, qui sont : la fonction d'activation ϕ (ici, les fonctions "ELU" et "RELU"), le nombre de neurones N dans la couche cachée des sous-réseaux de neurones, le nombre d'itérations T, le

paramètre d'agrégation s, et le paramètre de régularisation λ.

Alors que les modèles GLM et GAM utilisent la fonctionnalité *offset* pour prendre en compte l'exposition des observations dans la modélisation de la fréquence de sinistres, le modèle IGANN n'offre pas cette possibilité. Nous avons donc choisi de modéliser le nombre de sinistres en incluant l'exposition dans les variables explicatives.

4.6 Comparaison des modèles

Les modèles IGANN sont conçus pour offrir une meilleure interprétabilité par rapport aux réseaux de neurones artificiels. Chaque composante de ces modèles peut être analysée individuellement pour comprendre comment la variable correspondante influence les prédictions. De ce fait, le modèle IGANN est considéré comme un modèle "Glass Box", bien plus interprétable qu'un simple réseau de neurones, qualifié de "Black Box".

Le modèle IGANN conserve une structure additive, où la prédiction finale est obtenue en additionnant les contributions individuelles de chaque variable, telles que capturées par les fonctions composantes. Ces fonctions composantes, qui sont elles-mêmes des agrégations de réseaux de neurones, permettent de modéliser des relations complexes tout en maintenant l'aspect "Glass Box" du modèle.

Le modèle IGANN trouve donc une application précieuse dans des domaines où la prédiction et l'interprétabilité sont essentielles, comme dans le secteur bancaire pour la prédiction de la probabilité de défaut, ou en assurance. Dans ce dernier domaine, il permet non seulement d'obtenir de meilleures prédictions des indicateurs de risque pris en compte dans ce mémoire, mais aussi de comprendre comment ces prédictions sont réalisées.

La comparaison des modèles représentant la finalité de l'introduction du modèle IGANN, nous présenterons ces comparaisons de deux manières : la comparaison par les métriques et la comparaison par la visualisation des courbes des valeurs observées et prédites.
La première comparaison concerne les métriques. Le tableau 4.1 présente la comparaison des métriques du modèle de coût moyen.
Ce tableau compare les performances de quatre modèles de prédiction du coût moyen : GLM, GAM, IGANN, et XGBoost, en se basant sur trois indicateurs : le MAE (erreur absolue moyenne), le RMSE (racine carrée de l'erreur quadratique moyenne) et la variation du MAE du modèle GLM par rapport aux autres modèles.

Sur l'ensemble de test, le modèle XGBoost se distingue comme le modèle le plus performant, avec les erreurs les plus faibles, ce qui montre sa capacité à prédire avec précision les valeurs cibles. Le modèle IGANN suit de près, offrant également des

Modèles	Test			Train		
	MAE	**RMSE**	**VAR MAE**	**MAE**	**RMSE**	**VAR MAE**
GLM	1244,5	1752,47	100%	1341,15	1995,25	100%
GAM	1237,35	1745,68	-0,58%	1324,79	1984,51	-1,23%
IGANN	**1188,84**	**1711,87**	-4,68%	**1220,38**	**1792,16**	-9,90%
XGBoost	**1177,79**	**1681,60**	-5,66%	**1185,96**	**1674,93**	**-13,09%**

Table 4.1: Comparaison des performances des modèles de coût moyen

résultats satisfaisants, bien que légèrement en retrait par rapport à XGBoost. En revanche, les modèles GAM et GLM affichent des erreurs plus élevées, ce qui indique une moindre capacité prédictive que les modèles XGBoost et IGANN.

La variation du MAE du modèle GLM par rapport aux autres modèles est particulièrement révélatrice. Elle montre que les modèles XGBoost et IGANN réduisent mieux l'erreur MAE par rapport au modèle GLM, comparativement au modèle GAM qui n'apporte qu'une amélioration minime. Cela signifie que, bien que le modèle GAM soit une approche plus flexible que le modèle GLM, il n'apporte qu'un bénéfice limité en termes de réduction des erreurs **sur ce jeu de données particulier**.

Sur l'ensemble d'entraînement, le modèle XGBoost continue de dominer en termes de performance, ce qui démontre sa robustesse et son efficacité de modélisation. Les modèles de machine learning sont très susceptibles au surapprentissage, cependant, grâce à l'utilisation de conditions d'arrêt adaptées (condition d'arrêt marquant la fin des itérations dans le cas d'augmentation du RMSE sur la base de test), le risque de surapprentissage a été évité, assurant que les performances élevées observées sur l'ensemble d'entraînement se traduisent également par de bonnes performances sur l'ensemble de test. Le modèle IGANN montre également une bonne performance, confirmant sa capacité à généraliser efficacement, tandis que les modèles GAM et GLM offrent des performances moindres **sur ce jeu de données particulier**.

Le modèle XGBoost se démarque par sa capacité à réduire les erreurs par rapport au modèle GLM, sans pour autant tomber dans le piège du surapprentissage grâce à une gestion prudente de l'entraînement. Le modèle IGANN offre également un bon compromis entre précision et interprétabilité. Le modèle GAM, bien que légèrement meilleur que le modèle GLM, n'apporte qu'une amélioration modeste, indiquant que **sur ce jeu de données particulier**, les modèles plus avancés comme XGBoost ou IGANN sont préférables pour obtenir de meilleures prédictions du coût moyen de sinistres.

Le tableau suivant présente les mêmes indicateurs de comparaison sur le modèle de fréquence de sinistres.

Modèles	Test			Train		
	MAE	**RMSE**	**VAR MAE**	**MAE**	**RMSE**	**VAR MAE**
GLM	0,038	0,1403	100%	0,0505	0,1433	100%
GAM	0,037	0,1395	-2,70%	0,0503	0,1430	-0,40%
IGANN	**0,036**	**0,1395**	-5,56%	**0,0427**	**0,1414**	-18,27%
XGBoost	**0,036**	**0,1285**	-5,56%	**0,0317**	**0,1165**	**-59,01%**

Table 4.2: Comparaison des performances des modèles de fréquence de sinistres

L'analyse du tableau ci-dessus permet de tirer les mêmes conclusions que dans le cas de la modélisation du coût moyen. En effet, pour l'ensemble de test, le modèle XGBoost se démarque par ses erreurs minimales, indiquant une bonne précision de la fréquence de sinistres. Les modèles IGANN et GAM suivent de près, montrant également une bonne performance par rapport au modèle GLM. Concernant la variation du MAE par rapport au GLM sur l'ensemble de test, les modèles XGBoost et IGANN offrent de meilleures réductions du MAE que le modèle GAM.

Pour l'ensemble d'entraînement, le modèle XGBoost maintient sa supériorité avec une meilleure réduction des erreurs, tout en évitant le surapprentissage grâce à une gestion adaptée de l'entraînement. Le modèle IGANN montre également une bonne capacité de généralisation. Le modèle GAM améliore légèrement les résultats par rapport au modèle GLM, mais reste moins performant que les modèles IGANN et XGBoost sur ce jeu de données particulier.

Il est également à noter que les améliorations sont plus marquées sur la base d'entraînement pour les modèles IGANN et XGBoost, car ces modèles sont basés sur l'apprentissage automatique et sont donc fortement dépendants des données d'entraînement. En conséquence, ils affichent des métriques plus faibles par rapport aux modèles GAM et GLM, qui, eux, reposent sur des méthodes statistiques classiques.

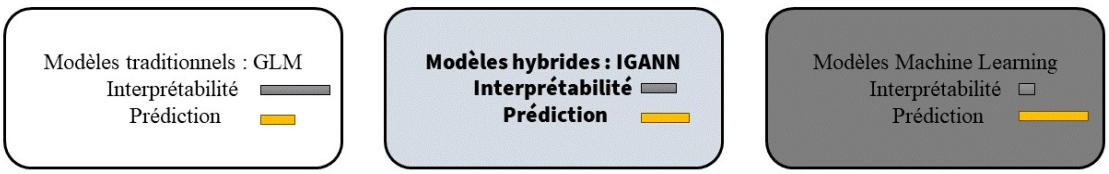

Figure 4.36: Positionnement des modèles

La deuxième comparaison repose sur les valeurs observées et prédites des modèles GLM, GAM, et IGANN. La figure ci-dessous montre les valeurs observées et prédites pour le coût moyen.

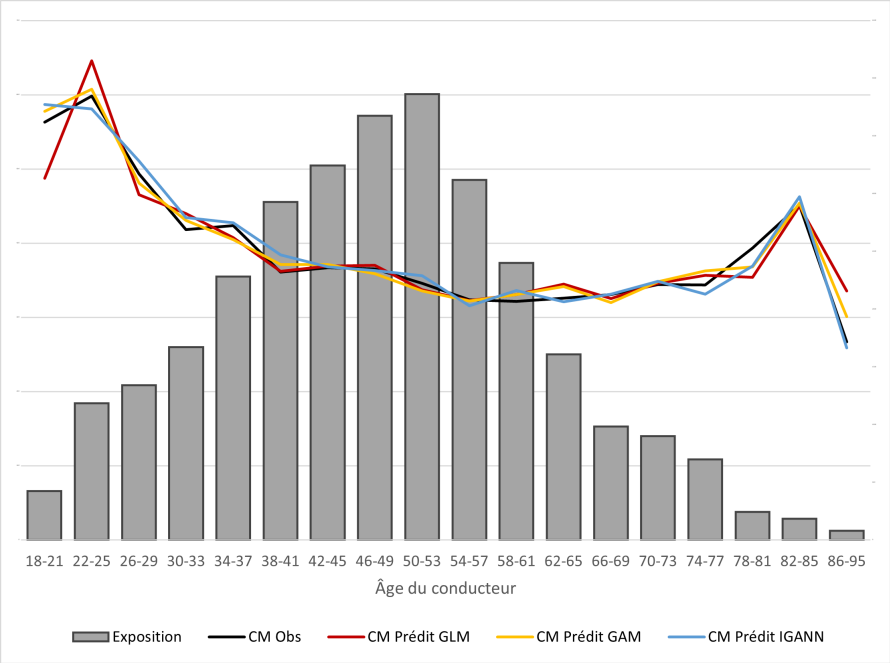

Figure 4.37: Valeurs observées vs valeurs prédites pour le modèle de coût moyen

La figure ci-dessus montre que les trois modèles comparés affichent dans l'ensemble un bon ajustement entre les valeurs observées et prédites, avec toutefois quelques remarques à retenir. Le modèle GLM peine à généraliser les prédictions pour les conducteurs âgés de 18 à 33 ans. Il sous-estime les prédictions pour la classe d'âge 18-21 ans et surestime celles de la classe 22-25 ans. L'analyse montre que les prédictions sont globalement meilleures pour les classes d'âge intermédiaires, avec une légère amélioration supplémentaire du modèle IGANN par rapport aux deux autres modèles pour ces classes. Une tendance similaire est observée pour les prédictions concernant les classes d'âge plus élevées.

Le graphique ci-dessous, qui montre la fréquence de sinistres, confirme la légère amélioration de l'ajustement du modèle IGANN par rapport aux deux autres modèles. En effet, la fréquence de sinistres pour les jeunes conducteurs est surestimée par l'ensemble des modèles, mais de manière moins prononcée par le modèle IGANN. De même, pour les classes d'âge intermédiaires, la courbe du modèle IGANN est plus proche de celle des fréquences observées que les autres modèles. La principale différence se trouve au niveau des deux dernières classes d'âge, où le modèle GLM généralise mieux. Le modèle GAM, quant à lui, peine à généraliser la fréquence de sinistres pour les classes d'âge les plus élevées en surestimant la fréquence de sinistres pour ces classes.

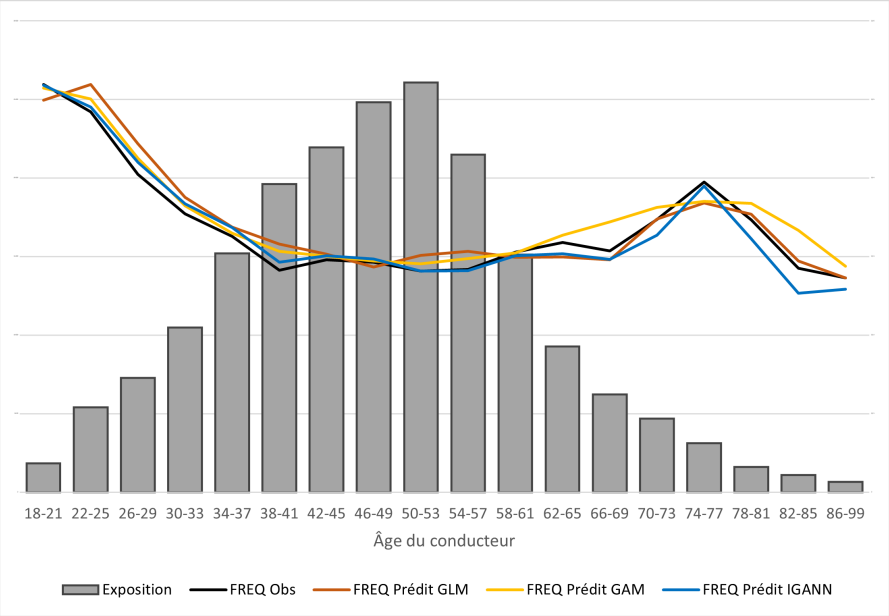

Figure 4.38: Valeur observées vs Valeurs prédites pour le modèle de fréquence de sinistres

Résumé du chapitre

Le dernier chapitre était destiné à l'introduction du modèle additif neuronal dérivé du modèle GAM dont l'étude a permis de découvrir une nouvelle manière de définir les relations entre la variable dépendante et les covariables dans un modèle additif, avec la flexibilité offerte par l'utilisation des fonctions splines des covariables. Les interprétations de ces fonctions permettent alors de comprendre les relations entre les covariables et la variable cible. Mais dans un concours de Kaggle, par exemple, où l'objectif est d'obtenir la meilleure performance, les modèles de machine learning s'imposent immédiatement comme les modèles au meilleur potentiel de prédiction mais moins interprétable. De nombreuses recherches sont alors intervenues pour proposer des méthodes d'interprétation telles que les méthodes PFI et SHAP. D'autres méthodes d'interprétation ont vu le jour, combinant les modèles GAM et les modèles d'apprentissage statistique, notamment le modèle IGANN, qui combine les réseaux de neurones et le modèle GAM. Ce modèle apporte un compromis entre prédiction et interprétabilité. Les résultats de ce chapitre ont montré que ce modèle, offrant trois types d'interprétabilité, se situe entre les modèles additifs (GAM et GLM) et le modèle XGBoost en termes de prédiction.

Conclusion

Ce mémoire a exploré de manière approfondie l'intégration du **véhiculier** et des modèles de Machine Learning interprétables, en particulier le Modèle Additif Neuronal (**IGANN**), dans le cadre de la tarification de la garantie **DTA en assurance automobile**. L'objectif principal était de proposer des méthodes permettant d'améliorer la précision des modèles prédictifs tout en préservant une interprétabilité suffisante pour les actuaires et les décideurs.

L'étude a montré que l'ajout du véhiculier, construit à partir des données du portefeuille de l'assureur, permet de mieux capturer les risques inhérents aux caractéristiques des véhicules. Ce véhiculier a démontré son efficacité en réduisant les erreurs de prédiction par rapport au modèle GLM sans cette variable, et même par rapport à l'utilisation du véhiculier standard fourni par la SRA, constitué du groupe SRA et de la classe SRA. Cette amélioration est particulièrement notable dans le modèle de coût moyen des sinistres.

Cette première solution a mis en lumière l'importance des modèles linéaires généralisés (GLM), qui permettent de modéliser des relations linéaires tout en étant dotés d'une structure interprétable. Bien que les modèles linéaires généralisés (GLM) soient largement utilisés pour leur simplicité, les modèles additifs généralisés (GAM) apportent une plus grande flexibilité en permettant l'intégration de fonctions splines pour mieux capturer des relations non linéaires entre les covariables et la variable cible tout en maintenant le caractère interprétable.

Bien que les modèles GAM aient apporté une flexibilité accrue dans la modélisation des relations non linéaires grâce à l'utilisation de fonctions splines, les modèles de Machine Learning sont plus performants en termes de capacité prédictive, repoussant ainsi les frontières de la prédiction des modèles additifs en capturant des interactions encore plus complexes entre les variables.

Cependant, malgré les gains en performance prédictive apportés par les modèles de Machine Learning, leur nature opaque a posé un défi majeur en termes d'interprétabilité. Au regard de ces limitations en matière d'interprétabilité des modèles de Machine Learning, le Modèle Additif Neuronal (IGANN) a été développé pour combiner la puissance

prédictive des réseaux neuronaux avec l'interprétabilité des modèles GAM, offrant ainsi une solution hybride qui permet de mieux comprendre les mécanismes sous-jacents tout en maintenant des performances élevées.

En appliquant ce modèle à la garantie DTA, il a été démontré que le modèle IGANN améliore significativement la capacité à prédire les indicateurs de risque tout en offrant une bonne compréhension de leur prédiction. Cette approche permet non seulement de réduire les erreurs de prédiction, mais aussi de conserver trois types d'interprétabilité (importance des variables, prédiction locale et relation entre les covariables et la variable dépendante) pour les actuaires, rendant les décisions tarifaires plus transparentes. La comparaison du modèle IGANN avec les autres modèles en termes de capacité prédictive a montré que le modèle IGANN se retrouve au milieu du modèle XGBoost avec les meilleures performances et les modèles classiques qui sont les modèles GAM et GLM. Le modèle IGANN est ainsi un bon modèle intermédiaire alliant prédiction et interprétabilité.

Il est essentiel de rappeler que l'objectif n'est pas de remplacer les modèles GLM, qui demeurent des outils fondamentaux en actuariat et ne sont pas près de disparaître. Ces modèles, bien qu'ils présentent certaines limites, restent indispensables pour leur simplicité et leur interprétabilité. Ce mémoire propose plutôt une ouverture vers d'autres approches de modélisation, telles que le modèle IGANN, qui peuvent être explorées en complément des GLM pour répondre aux défis actuels et futurs du secteur.
Cette recherche confirme la pertinence de l'intégration d'un véhiculier personnalisé dans les modèles de tarification, et met en lumière l'efficacité du modèle IGANN dans l'amélioration des prédictions tout en maintenant un niveau d'interprétabilité adapté aux exigences du secteur.

Les avancées réalisées dans ce mémoire ouvrent des perspectives prometteuses pour l'avenir de la tarification en assurance automobile. D'une part, elles soulignent l'importance de continuer à affiner la construction de variables issues de l'agrégation d'autres variables telles que le véhiculier ou le zonier, qui aurait également pu être construit et utilisé dans les modélisations. D'autre part, elles justifient la poursuite des recherches sur les méthodes d'interprétabilité des modèles de Machine Learning, afin de garantir que les outils les plus puissants sur le plan prédictif restent accessibles et utiles aux professionnels de l'assurance et d'autres secteurs comme la banque. Ces recherches contribueront à renforcer la compétitivité des compagnies d'assurance, tout en assurant une tarification plus juste et mieux alignée sur les risques réels des assurés.

Annexe

Présentation des variables

Variables	
Nom du distributeur	Formule d'assurance
Avenant	Version du véhicule
Cumul des avenants	Nombre de places du véhicule
Libellé de la police mère	Genre du véhicule
Clé du contrat	Type de boîte de vitesses du véhicule
Code de l'intermédiaire	Vitesse du véhicule
Code portefeuille	Puissance en CH du véhicule
Numéro associé du distributeur	Prix en euros du véhicule
Code SRA du véhicule	Groupe APSAD du véhicule
Carrosserie du véhicule	Classe APSAD du véhicule
Date de circulation du véhicule	Cylindrée du véhicule
Date d'acquisition du véhicule	Durée de détention du véhicule
Mode d'acquisition du véhicule	Type de garantie
Ville de stationnement	Résiliation Alcool
Code postal de stationnement	
Date de naissance du conducteur	
Interruption d'assurance	
Résiliation fausse déclaration	
Résiliation fréquence de sinistres	
Résiliation non-paiement	
Résiliation retrait de permis	
Résiliation suspension de permis	
Taux d'alcool	

Analyse univariée

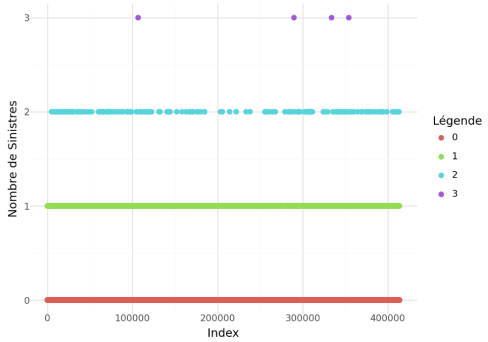

Figure 4.39: Nuage de points du nombre de sinistres

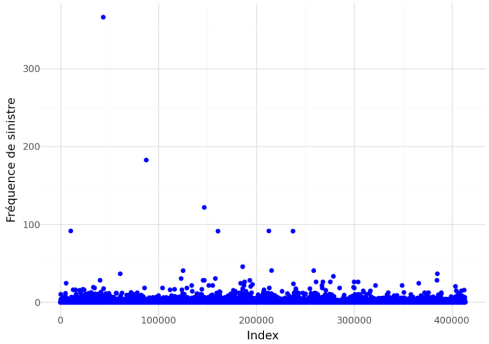

Figure 4.40: Nuage de points de la fréquence de sinistres

Analyse bivariée

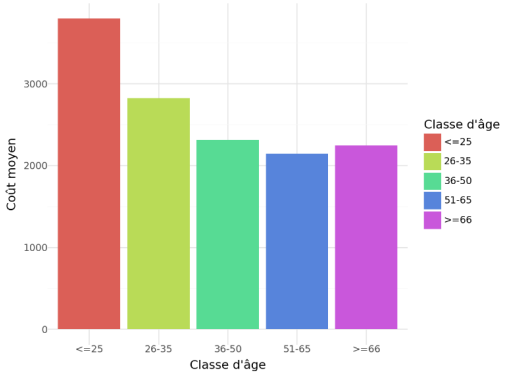

Figure 4.41: Coût moyen et Âge

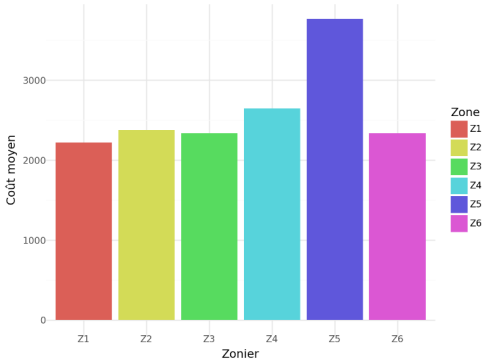

Figure 4.42: Coût moyen et Zonier

Bibliographie

[1] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] Junior Désiré ASSI. *Apport de données télématiques dans la modélisation du risque géographique en assurance automobile.* Mémoire d'actuariat. EURIA. Sept, 2022.

[3] François-Xavier CHAMOULAUD. *Réalisation d'un véhiculier à l'aide d'outils de Machine Learning.* Mémoire d'actuariat. ISFA. Novembre, 2019.

[4] C. Chouquet. *Modèles Linéaires.* Laboratoire de Statistique et Probabilités - Université Paul Sabatier - Toulouse - M1 IMAT, 2009 - 2010.

[5] Nexialog Consulting. *Techniques d'interprétabilité des modèles de machine learning, 2 mars.* Paris, France, 2022.

[6] Steven Côté. *Modèles additifs généralisés dans la modélisation de l'impact du kilométrage et de l'exposition au risque en assurance automobile.* Mémoire. Maîtrise en mathématiques. Université du Québec à Montréal, 2016.

[7] L. Bel et al. *Le Modèle Linéaire et ses Extensions.* 14 septembre, 2016.

[8] Khalil FADIL. *Optimisation de la tarification AUTO au travers du risque véhicule.* Mémoire d'actuariat. EURIA. Décembre, 2020.

[9] Leslie GNANSOUNOU. *Construction d'un véhiculier en assurance automobile à partir de méthodes de Machine Learning.* Mémoire d'actuariat. ENSAE. Mars, 2022.

[10] Pierre-Louis GONZALEZ. *L'analyse en composantes principales (A.C.P.).* 21 Mars, année d'édition inconnue.

[11] Pierre-Louis GONZALEZ. *Introduction à la théorie des valeurs extrêmes.* mois, et année d'édition inconnus.

[12] Christian Borel WAFO KANKEU. *Impact du montant de rente sur la longévité au sein d'un portefeuille de rentiers : Une approche par les modèles additifs généralisés.* Mémoire d'actuariat. EURIA. Septembre, 2023.

[13] Franklin FEUKAM KOUHOUE. *Interprétabilité des Modèles de Tarification en Actuariat : application*. Mémoire d'actuariat. ENSAE Paris. Novembre, 2023.

[14] Mathias Kraus, Daniel Tschernutter, Sven Weinzierl, and Patrick Zschech. Interpretable generalized additive neural networks. *European Journal of Operational Research*, 2023.

[15] Markéta KRÚPOVÁ. *Construction d'un modèle de Machine Learning interprétable pour la tarification en assurance non-vie*. Mémoire d'actuariat. Paris-Dauphine. Décembre, 2022.

[16] Belabbas Rachida. *Estimation par la méthode du maximum de vraisemblance*. Mémoire. Probabilités et Statistique. Université kasdi merbah, ouargla. Algérie, 2021.

[17] Franck Vermet. *Arbres de décision et méthodes ensemblistes*. EURIA Master 1 - Université de Bretagne Occidentale, 2021 - 2022.

[18] Franck Vermet. *Régression linéaire*. EURIA Licence 3 - Université de Bretagne Occidentale, 2021 - 2022.

[19] Franck Vermet. *Apprentissage statistique : une approche connexionniste*. EURIA Master 1 - Université de Bretagne Occidentale, 2023 - 2024.

[20] Simon N. Wood. *Copyright crc do not distribute*. Generalized Additive Models : an introduction with R, 2017.

[21] LI Siheng et Hu Chenyang Zhang Mudong. *Régression de Poisson*. 21 Mars, 2013.