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 6 Septembre 2023

Par : BENABDERRAHMAN Farah
Titre : Revue tarifaire de l'offre Propriétaire Non-Occupant (PNO) - Elaboration d'une démarche de modélisation

Confidentialité : Oui (2 ans)

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

Membre présent du jury de l'Institut des Actuaires :
LAÏLY Romain
STOCKSIEKER Samuel
GUELOU Sonia
Signature :

Entreprise :
MMA
Signature :

Membres présents du jury de l'EURIA :
RAINER Catherine

Directeur de mémoire en entreprise :
RIOULT Mathieu
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

Remerciements

Je souhaite exprimer ma sincère gratitude envers les personnes qui ont contribué à la réalisation de ce mémoire :

Tout d'abord, je tiens à remercier chaleureusement RIOULT Mathieu pour sa disponibilité, son encadrement rigoureux et son expertise éclairante qui ont été d'une valeur inestimable pour l'aboutissement de ce travail.

Mes remerciements vont également à WILCZYK Ewen, qui m'a offert l'opportunité de réaliser mon alternance au sein de l'équipe TSP de MMA. Ses conseils précieux et avisés ont grandement contribué à orienter mon travail de manière fructueuse.

Je remercie également NAHELOU Anthony pour son suivi durant mon alternance et pour ses suggestions pertinentes.

Un grand merci à tous les membres de l'équipe TSP de MMA pour leur soutien indéfectible et leur bonne humeur.

Je souhaite également exprimer ma reconnaissance envers le directeur de l'EURIA, Franck VERMET, ainsi qu'à l'ensemble de mes professeurs, pour la qualité de leurs enseignements au cours de ces trois années de formation.

Enfin, je tiens à exprimer ma profonde gratitude envers mes parents, ma grande sœur, mon petit frère ainsi que mes amis, dont le soutien inconditionnel et la présence constante ont été un pilier essentiel tout au long de ces années.

Résumé

Le marché de l'assurance habitation n'a cessé d'être pris aux entrailles de nombreux évènements. Entre multiplication d'événements climatiques et retombées inflationnistes des crises géopolitiques, les assureurs sont confrontés au défi d'ajuster les tarifs pour faire face aux coûts croissants résultant de ce contexte défavorable. De plus, le marché de l'assurance habitation est saturé, entraînant une forte concurrence entre les différents acteurs traditionnels et assurtechs, dans un contexte où les clients sont devenus plus exigeants et volatils en raison de changements législatifs en leur faveur.

Face à cette situation, se démarquer par des offres commerciales agressives pour attirer de nouveaux clients devient coûteux pour les assureurs. Ils cherchent alors à fidéliser leurs clients existants en utilisant des stratégies comme le multi-équipement.

Au sein de MMA, le produit Propriétaire Non-Occupant (PNO) s'affiche clairement comme un produit d'accompagnement de l'offre principale pour multi-équiper les clients notamment la clientèle patrimoniale. Cependant, ce produit est en déficit technique, avec un ratio combiné supérieur à 100%. Il est donc crucial de trouver des leviers à actionner pour s'approcher de l'équilibre technique.

Le présent mémoire s'inscrit dans ce cadre en proposant un levier d'action tarifaire, celui de challenger la tarification actuelle par une approche de modélisation de la prime pure. La dernière révision structurelle date de 2016 et seuls des renouvellements annuels à plat sont effectués. Il s'avère ainsi nécessaire de s'assurer de la cohérence du tarif vis à vis du risque encouru.

Mots clefs: Assurance Propriétaire Non-Occupant, Prime Pure, Garantie Dégâts des Eaux, Equilibre technique, GLM, Machine Learning, Random Forest, Gradient Boosting, SHAP, PDP, ICE

Abstract

The home insurance market has been continually affected by numerous events. Between the increasing occurrences of climate-related incidents and the inflationary aftermath of geopolitical crises, insurers are confronted with the challenge of adjusting rates to cope with the rising costs stemming from this unfavorable environment. Furthermore, the home insurance market is saturated, leading to intense competition among various traditional players and assurtechs, in a context where customers have become more demanding and volatile due to legislative changes in their favor.

Within MMA, the Non-Occupant Property Owner (PNO) product stands out as a supplementary offering to the main product, particularly tailored to serve the needs of clients with multiple properties, including high net worth clientele. However, this product is currently facing technical deficits, with a combined ratio exceeding 100%. As such, it is crucial to identify strategies to address this situation and move closer to achieving technical balance.

This thesis fits into this context by proposing a pricing lever, namely challenging the current pricing through a pure premium modeling approach. The last structural rate revision dates back to 2016, and only flat annual renewals are conducted. Thus, it is essential to ensure tariff coherence with the incurred risk.

Keywords: The non-occupant house insurance, Pure Premium, Water Damage Coverage, Technical Balance, GLM, Machine Learning, Random Forest, Gradient Boosting, SHAP, PDP, ICE

Note de synthèse

Contexte et objectif du mémoire

Dans un marché concurrentiel où la multi-détention est devenue une des clés stratégiques adoptées par les assureurs, le produit propriétaire non-occupant (PNO) de MMA se positionne comme un produit d'accompagnement de l'offre principale pour multi-équiper les clients notamment les assurés patrimoniaux. Ainsi, un surplus de rentabilité est attendu du produit. Néanmoins, sur ces dernières années, il est en perte de vitesse de production et est en déficit technique en affichant un ratio combiné supérieur à 100%.

Il est donc crucial d'actionner des leviers pour s'approcher de l'équilibre technique. Cette nécessité est d'autant plus accentuée par la dernière révision structurelle du tarif qui remonte à 2016, ce qui suscite des interrogations quant à la cohérence du tarif actuel face au risque encouru.

Les travaux réalisés dans ce mémoire s'inscrivent dans ce cadre en proposant un levier d'action tarifaire, celui de challenger la tarification actuelle par une approche de modélisation de la prime pure, afin de s'assurer de la cohérence du tarif au risque réellement encouru. L'étude s'est focalisée sur la garantie dégâts des eaux qui représente un poids significatif en termes de sinistralité et offre l'opportunité d'actionner des leviers tarifaires dans l'objectif d'atteindre l'équilibre technique du produit.

Pour ce faire, Les travaux ont débuté par la phase primordiale et la plus chronophage, celle de la construction de la base de données sur laquelle seront établis nos modèles.

1. Construction et traitement de la base de données

Les bases contenant les informations sur les contrats et leur sinistralité ont été fusionnées. L'historique retenu s'étend de 2011 jusqu'à 2022. Par la suite, cette base a été enrichie par les données clients et les zoniers.

Un travail de recherche ainsi qu'un traitement des données externes pertinentes ont été menés dans le but d'apporter des informations supplémentaires sur l'environnement dans lequel évoluent le client et le contrat. Ces données externes, à caractère sociodémographique et géographique, ont été raccordées à la base de données internes grâce au géocodage des adresses des biens assurés. La figure 1 présente le processus suivi à cet effet.

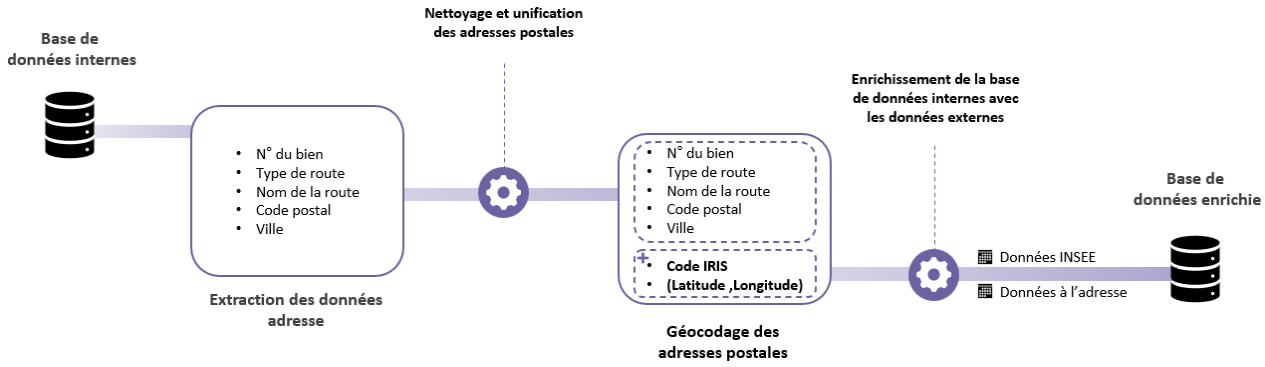


Figure 1 – Processus de raccordement des données externes à la base de données internes

Une attention particulière a été portée à la qualité des données (Données manquantes, aberrantes...) pour que les modèles construits soient fiables. Une fois que la base de données a été mise en place, des travaux préliminaires essentiels à la modélisation de la charge de sinistres ont été entrepris (écrêtement des sinistres, mise en « as if » de l'historique de sinistralité).

Par ailleurs, une analyse descriptive des données a permis d'avoir une meilleure connaissance de la structure du portefeuille et un aperçu sur les variables, à priori, segmentantes qui sont le plus susceptibles d'intervenir dans les modèles. De plus, l'étude de la corrélation entre les différentes variables a été effectuée, révélant les variables très fortement corrélées qui pourraient biaiser les modèles.

2. Modélisation de la garantie dégâts des eaux

La base de données étant construite et retraitée, elle a fait l'objet d'un échantillonnage aléatoire sans remise permettant de la séparer selon la répartition suivante : 70% pour la base d'apprentissage et 30% pour la base de test. La base d'apprentissage a servi au calibrage des modèles tandis que la base de test a permis d'évaluer leur pouvoir prédictif.

La phase suivante consiste à la mise en place du processus de modélisation de la prime pure de la garantie dégâts des eaux. Pour cela, une modélisation séparée de la fréquence et du coût moyen a été menée par l'approche classique des modèles linéaires généralisés (GLM). Cette approche présente l'avantage d'être lisible et interprétable. En outre, le choix s'est porté sur une modélisation par nature du bien pour chacun des modèles fréquence et coût moyen, en raison de la répartition hétérogène de la sinistralité entre les maisons et les appartements.

Une fois le cadre de la modélisation défini, la première étape consiste à déterminer les lois de distribution pour les modèles de fréquence et de coût moyen. Les lois les plus adaptées sont respectivement la loi Poisson pour le modèle de fréquence et la loi Gamma pour le coût moyen. S'ensuit l'étape de la sélection des variables significatives par la procédure stepwise selon le critère AIC. Le tableau ci-dessous résume les variables sélectionnées dans le modèle de fréquence x coût moyen pour les maisons :

Modèle Fréquence x Coût Moyen - Maisons							
Superficie	✓ ✓	Zonier DDE	✓ ✓	Part RP avec chauffage indiv		Part RP avec 5pièces ou +	✓
Nb logements	✓ ✓	Zonier inondation	✓ ✓	Part des RP avec Chauff Coll	✓	Part ménages en couple avec enf	
Niveau capital mobilier	✓	Zonier PNO	✓	Part RP-MAISON avant1919		Part ménages en couple sans enf	
Franchise	✓	Nombre de voisins	✓ ✓	Part RP-MAISON entre 1920-1945		Part pop entre 60-74ans	
Formule	✓	Mitoyenneté	✓	Part RP-MAISON entre 1946-1970		Part pop >= 75ans	
Nb sinistres sur 12 derniers mois	✓	Dépendance non-attenante		Part RP-MAISON entre 1971-1990		Part des artisans	
Nb sinistres sur 36 derniers mois	✓	Equipement	✓	Part RP avec spfc <30m²		Part des cadres	✓
Ancienneté	✓ ✓	Nb de contrats en cours	✓	Part RP avec SUP> 120m²		Part des agriculteurs	
Age	✓ ✓			Part Résidences Secondaires		Part des inactifs	
Groupe terme client	✓ ✓			Part logements vacants		Part des chômeurs	
				Part RP en location	✓	Part des prof. Inter	

Figure 2 – Variables sélectionnées dans le modèle fréquence x coût moyen - Maisons

Les variables sélectionnées pour le modèle de fréquence et le modèle de coût moyen sont de différents types (variables liées au bien, client, données à l'adresse, etc.) ce qui permet d'affiner la segmentation du risque.

Enfin, après avoir calibré les modèles, arrive l'étape de leur validation. Pour chacun d'eux, l'analyse des résidus a montré que les hypothèses des modèles sont vérifiées. Le pouvoir prédictif des modèles a été vérifié sur la base de test pour chacune des variables explicatives, et de manière globale également.

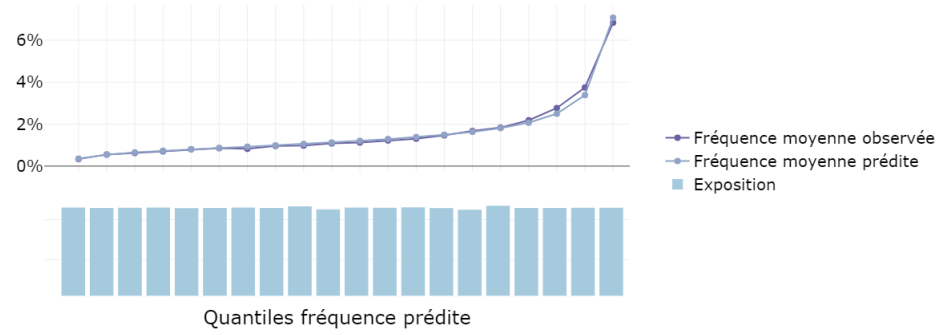


Figure 3 – Pouvoir prédictif du modèle de fréquence - base de test

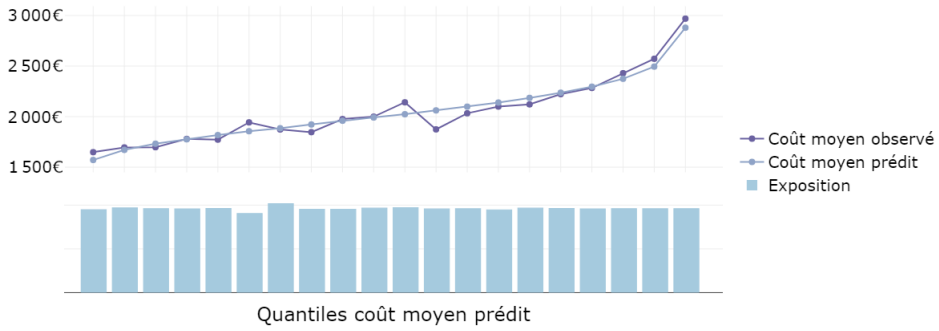


Figure 4 – Pouvoir prédictif du modèle de coût moyen - base de test

Il en ressort des différentes analyses que le modèle de fréquence présente un fort pouvoir prédictif tandis que le modèle de cout moyen se généralise moins bien sur la base de test. Cela peut être dû à la volumétrie réduite de la base de données utilisée pour la modélisation du coût moyen étant donné que le produit PNO manque de volume d'affaires comparé à l'offre principale. De plus, nous nous limitons aux sinistres de charge strictement positive.

Ainsi, afin d'améliorer les performances du modèle de coût moyen, une modélisation par les algorithmes de machine learning : Random Forest et Gradient Boosting a été explorée.

Après avoir procédé à l'optimisation des hyperparamètres des différents algorithmes, les modèles ont été comparés selon l'indicateur de performance RMSE. Il s'avère qu'aucun modèle de Machine Learning ne surpasse la modélisation par l'approche classique GLM.

	RMSE - Base d'apprentissage	RMSE - Base de test
GLM	2 626,21 €	2 648,10 €
Random Forest	2 607,23 €	2 660,21 €
Gradient Boosting	2 622,88 €	2 662,42 €

Figure 5 – Comparaison de la qualité de prédiction des modèles de coût moyen selon l'indicateur RMSE - Maisons

Toutefois, la modélisation par ces méthodes alternatives nous a conforté :

D'une part à travers l'analyse de l'importance des variables explicatives, sur la sélection des variables retenues pour le modèle GLM. En effet, dix variables du modèle GLM figurent dans le classement des quinze variables les plus importantes dans la construction des modèles de Machine Learning. Nous retrouvons des variables provenant de différents types :

- Des variables liées au bien assuré : la superficie, le niveau du capital mobilier et le nombre de logements.
- Des variables géographiques : les zoniers des garanties dégâts des eaux et inondation.
- Des variables relatives au client : le groupe terme client et l'ancienneté du contrat.
- Des variables externes : le nombre de voisins dans un rayon de 50 mètres (à l'adresse), la part des résidences principales en location et la part des résidences principales avec un chauffage collectif (Insee).

Le tableau 6 compare les variables explicatives des trois modèles de coût moyen.

Modèle GLM	Modèle Random Forest	Modèle Gradient Boosting
Superficie	Superficie	Superficie
Niveau du capital mobilier	Niveau du capital mobilier	Zonier de la garantie dégâts des eaux
Zonier de la garantie dégâts des eaux	Zonier de la garantie dégâts des eaux	Ancienneté du contrat
Groupe terme client	Ancienneté du contrat	Equipement
Nombre de voisins	Nombre de contrats en cours	Niveau du capital mobilier
Zonier de la garantie inondation	Age de l'assuré	Nombre de logements
Age de l'assuré	Part des résidences principales en location	Nombre de contrats en cours
Part des résidences principales en location	Part des résidences principales avec un chauffage collectif	Zonier de la garantie inondation
Part des résidences principales avec un chauffage collectif	Nombre de voisins	Part des résidences principales avec un chauffage collectif
Ancienneté du contrat	Zonier de la garantie inondation	Part des maisons en résidences principales construites entre 1920-1945
Nombre de logements	Nombre de logements	Part des maisons en résidences principales avec une superficie > 120m²
Sinistralité antérieure sur 12 mois	Equipement	Part des résidences secondaires
	Groupe terme client	Part des résidences principales en location
	Part des maisons en résidences principales avec une superficie > 120m²	Groupe terme client
	Part des maisons en résidences principales construites entre 1920-1945	Nombre de voisins

Variables retenues uniquement par le modèle GLM

Variables non retenues par le modèle GLM

Variables retenues* uniquement par le modèle Gradient Boosting

Variables non retenues* par le modèle Gradient Boosting

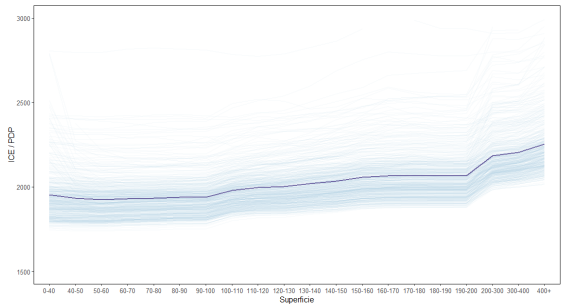
** Pour les méthodes de Machine Learning, le terme "retenues" est un abus de langage pour dire "figurent parmi les 15 variables les plus importantes"*

Figure 6 – Comparaison des variables retenues par les modèles de coût moyen - Maisons

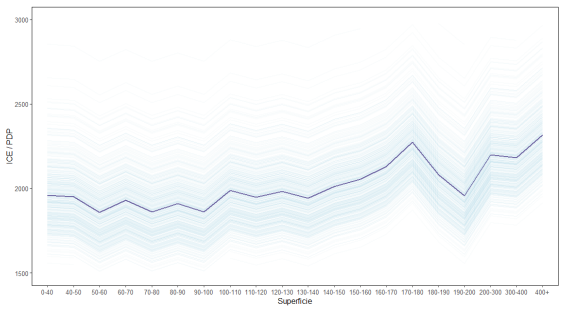
Et d'autre part, à travers les outils d'interprétabilité PDP/ICE et SHAP, sur l'effet et le sens d'influence des variables explicatives les plus importantes sur le coût moyen prédit.

Pour illustrer, prenons l'exemple de la variable la plus importante «Superficie» des maisons (en m^2). Dans le modèle GLM, les coefficients estimés pour les petites superficies sont inférieurs à ceux attribués aux grandes superficies.

Quant aux courbes ICE/PDP, à l'analyse des graphiques 7a et 7b, la courbe PDP du coût moyen prédit a une tendance haussière en fonction de la superficie. De plus, les courbes ICE présentent la même tendance haussière et sont globalement translatées les unes des autres ainsi que de la courbe PDP.



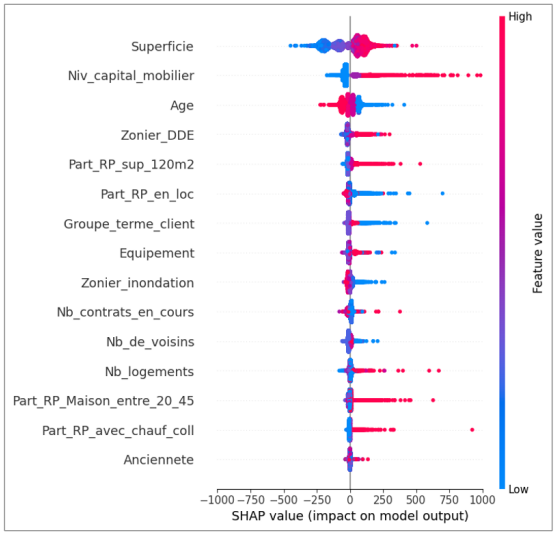
(a) Graphique de PDP et courbes ICE obtenus pour le modèle Random Forest associé à la variable «Superficie»



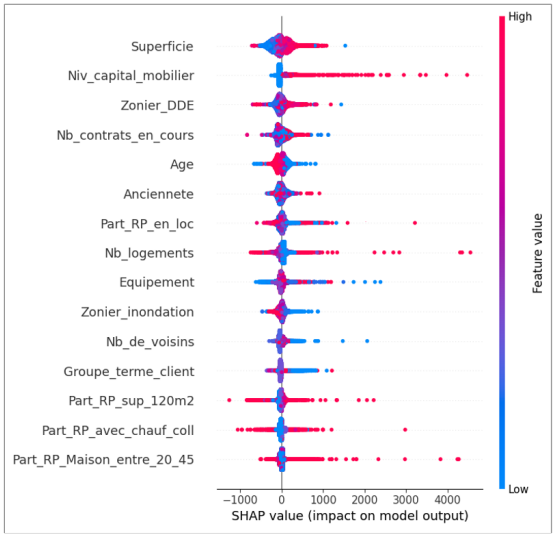
(b) Graphique de PDP et courbes ICE obtenus pour le modèle Gradient Boosting associé à la variable «Superficie»

Figure 7

En ce qui concerne la méthode SHAP, l'analyse des SHAP summary plot pour les modèles Random Forest (figure 8a) et Gradient Boosting (figure 8b) met en évidence que les grandes superficies ont des valeurs SHAP positives ce qui signifie qu'elles ont un impact positif sur le coût moyen des sinistres. Autrement dit, le coût moyen augmente avec de grandes superficies. Inversement, les petites superficies sont associées à des valeurs SHAP négatives, l'impact est négatif sur le coût moyen, En d'autres termes, le coût moyen diminue lorsque la superficie est petite.



(a) SHAP summary plot
pour le modèle Random Forest



(b) SHAP summary plot
pour le modèle Gradient Boosting

Figure 8

Les différents modèles s'accordent sur l'impact et le sens d'influence de la variable «Superficie» sur le coût moyen prédit.

3. Comparaison avec le tarif actuel

La modélisation de la fréquence et du coût moyen par l'approche GLM étant effectuée et validée. La prime pure est ensuite obtenue en faisant le produit des coefficients du coût moyen et de la fréquence des sinistres.

À la suite de cette étape, une étude comparative entre le modèle construit et le tarif actuel a pu être réalisée. L'objectif de cette étude est de dresser une liste de préconisations pour les prochains renouvellements tarifaires.

Pour chaque variable existante dans la structure tarifaire actuelle, l'écart tarifaire a été évalué en confrontant les coefficients résultants du modèle de la prime pure construit, à ceux issus du tarif actuel. Pour assurer la comparabilité de ces coefficients, les chargements et les taxes n'ont pas été pris en compte.

L'étude d'impact a révélé la nécessité d'améliorer l'adéquation du tarif au risque réellement encouru. À titre d'exemple, pour la variable «Superficie», la préconisation réalisée consiste en une baisse du coefficient tarifaire de l'ordre de 17% pour les maisons avec une superficie inférieure à 40 m^2, ainsi qu'une hausse de l'ordre de l'ordre de 9% pour les maisons avec une superficie

supérieure à 400 m^2.

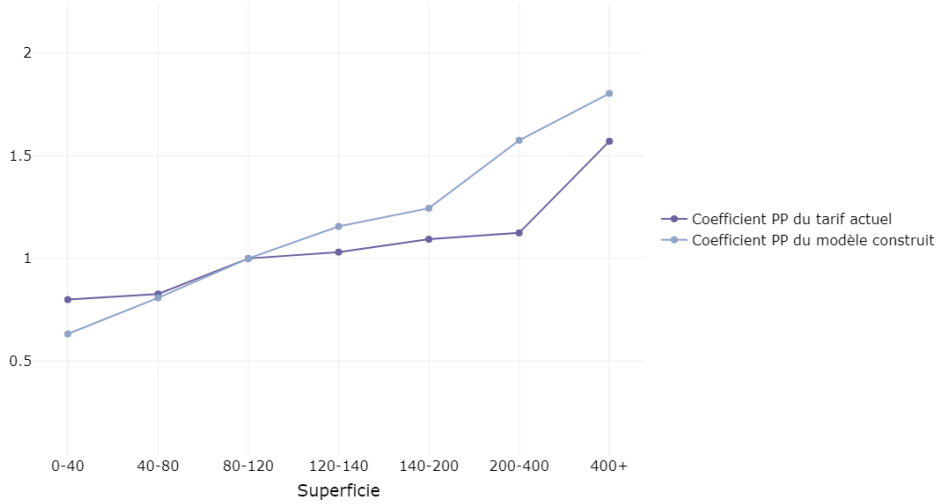


Figure 9 – Comparaison entre les coefficients de la prime pure du modèle et ceux du tarif actuel

Une fois les écarts tarifaires évalués pour chaque critère, un plan d'action a été dressé pour revaloriser à la hausse ou à la baisse les coefficients associés aux différents critères présents dans la structure tarifaire actuelle. De plus, des préconisations ont été formulées pour l'intégration dans le tarif commercial, de nouvelles variables significatives révélées par le modèle construit. Parmi ces variables figurent le zonier de la garantie dégâts des eaux (zonier de l'offre principal) et le groupe terme client. Ces variables exploitées sur l'offre principale s'avèrent pertinentes à intégrer dans la tarification de l'offre PNO pour avoir une meilleure segmentation du tarif.

L'étude menée dans le cadre de ce mémoire pour la garantie dégâts des eaux sera étendue à l'ensemble des garanties du produit PNO. Cette démarche permettrait d'apprécier au global l'équilibre technique du produit.

Executive summary

Context and objective of the thesis

In a competitive market where multi-possession has become one of the strategic keys adopted by insurers, MMA's non-occupant property owner (PNO) product positions itself as a complementary offering to the main portfolio, serving the purpose of equipping clients with multiple properties, especially those of substantial wealth. Consequently, a surplus of profitability is anticipated from the product. However, over the past few years, it has experienced a decline in production momentum and is facing technical deficits, evident in a combined ratio exceeding 100%.

Thus, it is imperative to activate strategies aimed at approaching technical equilibrium. This need is further accentuated by the fact that the last structural tariff revision dates back to 2016, raising questions about the current tariff's alignment with the incurred risk.

The work carried out in this thesis aligns with this framework by proposing a pricing lever, that is, challenging the current pricing through a modeling approach of pure premium, in order to ensure the alignment of the tariff with the actual risk incurred. The study focused on water damage coverage, which represents a significant portion of claims and provides an opportunity to implement pricing strategies with the goal of achieving the technical equilibrium of the product.

To achieve this, the work started with the crucial and time-consuming phase of constructing the database on which our models will be built.

1. Construction and Processing of the Database

The databases containing information about contracts and their claims were merged. The selected historical period covers from 2011 to 2022. Subsequently, this database was enriched with customer and zoning informations.

Research work and processing of relevant external data were conducted to provide additional insights into the environment in which the customer and the contract operate. These external data, of socio-demographic and geographical nature, were linked through geocoding of insured property addresses. The process followed for this purpose is illustrated in the figure 10.

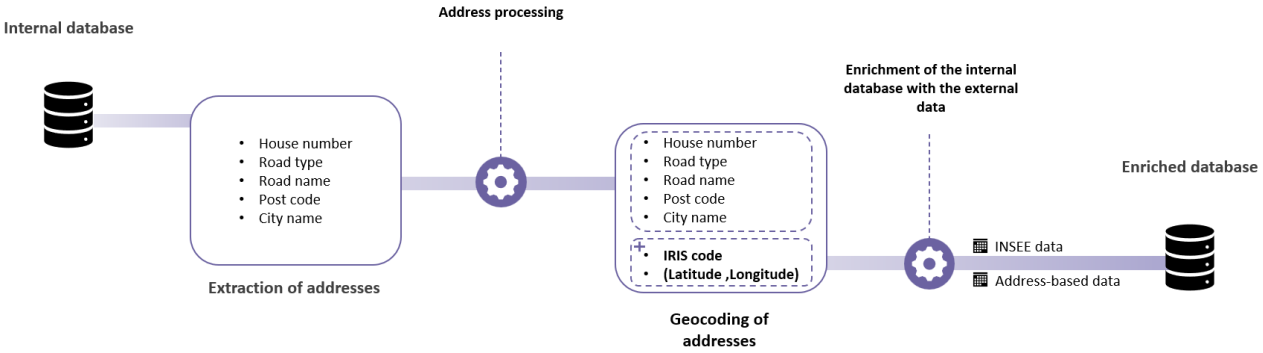


Figure 10 – Process of linking external data to internal database

Special attention was given to data quality (missing data, outliers, etc.) to ensure reliability of the constructed models. Once the database was established, essential preliminary work for modeling the claims load was undertaken (claims capping, "as if" claims handling).

Furthermore, a descriptive analysis of the data provided a better understanding of the portfolio structure and insights into variables that are potentially segmenting and likely to influence the models. Additionally, correlation analysis was performed among different variables, revealing highly correlated variables that might introduce bias to the models.

2. Modeling Water Damage Coverage

With the constructed and reprocessed database in place, random sampling without replacement was performed to split it into training (70%) and testing (30%) datasets. The training dataset was used for model calibration, while the testing dataset evaluated their predictive power.

The next phase involved setting up the modeling process for the pure premium of water damage coverage. This was achieved by separately modeling the frequency and average cost using the classical approach of Generalized Linear Models (GLM). This approach offers the advantage of being interpretable. Furthermore, separate modeling was chosen based on the nature of the property for each of the frequency and average cost models due to the heterogeneous distribution of claims between houses and apartments.

After defining the modeling framework, the first step is to determine the distribution laws for the frequency and average cost models. The most suitable laws are the Poisson distribution for the frequency model and the Gamma distribution for the average cost model. Subsequently, the stepwise variable selection procedure based on the AIC criterion was applied to choose the significant variables. The table below summarizes the selected variables in the frequency x average cost model for houses :

The selected variables for the frequency and average cost models are of different types (property-related, customer-related, address-related data, etc.), allowing for refined risk segmentation.

After calibrating the models, the validation step followed. For each model, analysis of residuals confirmed the model assumptions. The predictive power of the models was verified on the testing dataset for each explanatory variable and overall as well.

In conclusion, the frequency model demonstrated strong predictive power, while the average cost model generalized less effectively on the testing dataset. This could be attributed to the

Frequency x Average cost model - Houses							
Superficie	✓ ✓	Zonier DDE	✓ ✓	Part RP avec chauffage indiv		Part RP avec 5pièces ou +	✓
Nb logements	✓ ✓	Zonier inondation	✓ ✓	Part des RP avec Chauff Coll	✓	Part ménages en couple avec enf	
Niveau capital mobilier	✓	Zonier PNO	✓	Part RP-MAISON avant1919		Part ménages en couple sans enf	
Franchise	✓	Nombre de voisins	✓ ✓	Part RP-MAISON entre 1920-1945		Part pop entre 60-74ans	
Formule	✓	Mitoyenneté	✓	Part RP-MAISON entre 1946-1970		Part pop >= 75ans	
Nb sinistres sur 12 derniers mois	✓	Dépendance non-attenante		Part RP-MAISON entre 1971-1990		Part des artisans	
Nb sinistres sur 36 derniers mois	✓	Equipement	✓	Part RP avec spfc <30m²		Part des cadres	✓
Ancienneté	✓ ✓	Nb de contrats en cours	✓	Part RP avec SUP> 120m²		Part des agriculteurs	
Age	✓ ✓			Part Résidences Secondaires		Part des inactifs	
Groupe terme client	✓ ✓			Part logements vacants		Part des chômeurs	
				Part RP en location	✓	Part des prof. Inter	

Figure 11 – Selected variables in the frequency x average cost model - Houses

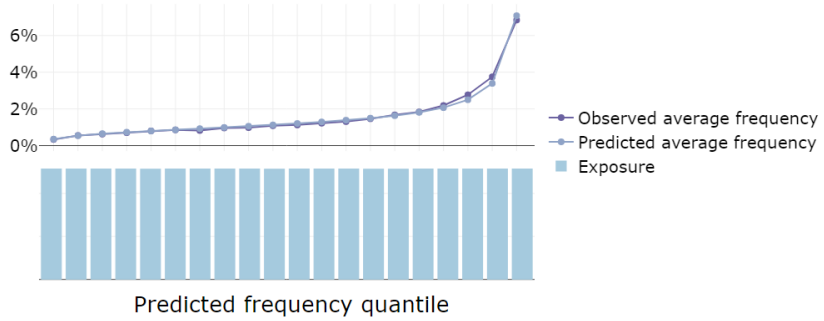


Figure 12 – Predictive power of the frequency model - testing dataset

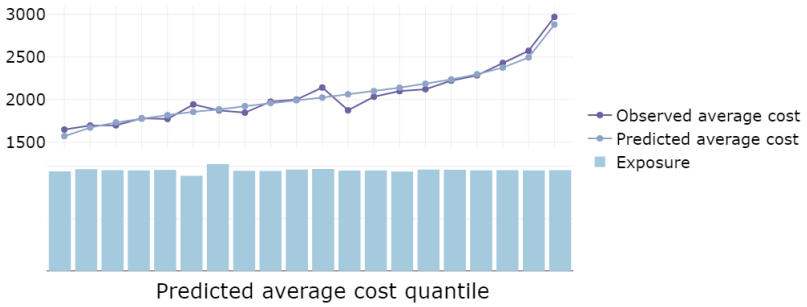


Figure 13 – Predictive power of the average cost model - testing dataset

limited volume of data available for modeling the average cost, given that the PNO product has a smaller business volume compared to the main offering. Additionally, we are limited to claims with strictly positive amounts.

To enhance the performance of the average cost model, modeling using machine learning algorithms—Random Forest and Gradient Boosting—was explored. After tuning the hyperparameters of the different algorithms, the models were compared using the RMSE performance indicator. It turns out that no Machine Learning model outperforms the traditional GLM approach.

	RMSE - Training data	RMSE - Testing data
GLM	2 626,21 €	2 648,10 €
Random Forest	2 607,23 €	2 660,21 €
Gradient Boosting	2 622,88 €	2 662,42 €

Figure 14 – Comparison of predictive quality of average cost models using the RMSE indicator - Houses

However, the modeling using these alternative methods has confirmed :

On one hand, through the analysis of the importance of explanatory variables in the GLM model's selection of retained variables. In fact, ten variables from the GLM model are among the top fifteen most important variables in the construction of the Machine Learning models. These variables span different types :

- Property-related variables : "superficie", "niveau du capital mobilier" and "nombre de logements"
- Geographic variables : "groupe terme client" and "ancienneté du contrat"
- Customer-related variables : "zonier DDE" and "zonier Inondation"
- External variables : "nombre de voisins dans un rayon de 50 mètres" (at the address), "part des résidences principales en location", "part des résidences principales avec un chauffage collectif" (INSEE data).

Table 15 compares the explanatory variables of the three average cost models.

GLM model	Random Forest model	Gradient Boosting model
Superficie	Superficie	Superficie
Niveau du capital mobilier	Niveau du capital mobilier	Zonier de la garantie dégâts des eaux
Zonier de la garantie dégâts des eaux	Zonier de la garantie dégâts des eaux	Ancienneté du contrat
Groupe terme client	Ancienneté du contrat	Equipement
Nombre de voisins	Nombre de contrats en cours	Niveau du capital mobilier
Zonier de la garantie inondation	Age de l'assuré	Nombre de logements
Age de l'assuré	Part des résidences principales en	Nombre de contrats en cours
Part des résidences principales en location	Part des résidences principales avec un chauffage collectif	Zonier de la garantie inondation
Part des résidences principales avec un chauffage collectif	Nombre de voisins	Part des résidences principales avec un chauffage collectif
Ancienneté du contrat	Zonier de la garantie inondation	Part des maisons en résidences principales construites entre 1920-1945
Nombre de logements	Nombre de logements	Part des maisons en résidences principales avec une superficie > 120m²
Sinistralité antérieure sur 12 mois	Equipement	Part des résidences secondaires
	Groupe terme client	Part des résidences principales en
	Part des maisons en résidences principales avec une superficie > 120m²	Groupe terme client
	Part des maisons en résidences principales construites entre 1920-1945	Nombre de voisins

Variables selected only by GLM model

Variables not selected by GLM model

Variables selected* only by Gradient Boosting model

Variables not selected* by Gradient Boosting model

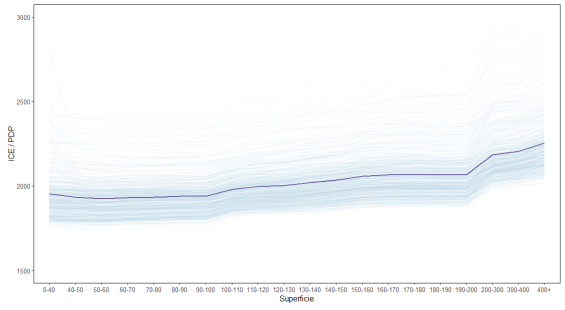
**The term "selected" erroneously refer to "among the 15 most important variables"*

Figure 15 – Comparison of selected variables in average cost models - Houses

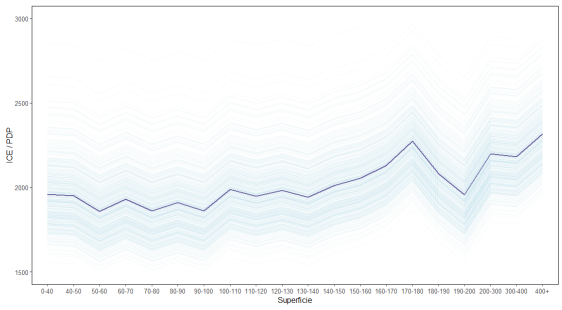
On the other hand, through interpretability tools PDP/ICE and SHAP, regarding the effect and direction of influence of the most important explanatory variables on the predicted average cost.

To illustrate, consider the most important variable "superficie", In the GLM model, the estimated coefficients for small areas are lower than those attributed to larger areas.

Regarding the ICE/PDP curves, analyzing graphs 16a and 16b, the PDP curve of the predicted average cost shows an increasing trend based on area. Additionally, the ICE curves exhibit the same upward trend and are generally shifted relative to each other as well as the PDP curve.



(a) PDP and ICE graphs obtained for the Random Forest model associated with the "Area" variable

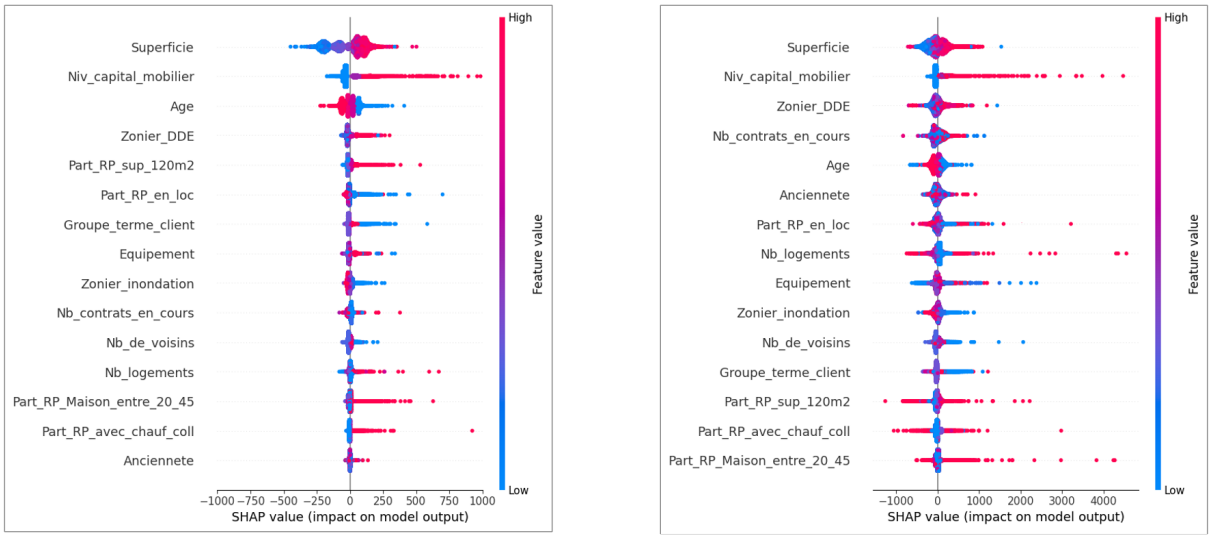


(b) PDP and ICE graphs obtained for the Gradient Boosting model associated with the "Area" variable

Figure 16

Regarding the SHAP method, analyzing the SHAP summary plots for the Random Forest

model (figure 17a) and Gradient Boosting model (figure 17b) highlights that larger areas have positive SHAP values, indicating a positive impact on the average cost of claims. In other words, the average cost increases with larger areas. Conversely, smaller areas are associated with negative SHAP values, indicating a negative impact on the average cost. In simpler terms, the average cost decreases with smaller areas.



(a) SHAP summary plot for the Random Forest model **(b)** SHAP summary plot for the Gradient Boosting model

Figure 17

The different models concur on the impact and direction of influence of the "Area" variable on the predicted average cost.

3. Comparison with the Current Tariff

The modeling of frequency and average cost using the GLM approach having been performed and validated, the pure premium is then obtained by multiplying the coefficients of average cost and claims frequency.

Following this step, a comparative study between the constructed model and the current tariff was conducted. The objective of this study is to establish a list of recommendations for the next tariff renewal.

For each variable existing in the current tariff structure, the tariff gap was evaluated by comparing the resulting coefficients of the constructed pure premium model to those derived from the current tariff. To ensure comparability of these coefficients, surcharges and taxes were not taken into account.

The impact study revealed the necessity to improve the adequacy of the tariff to the actual risk incurred. As an example, for the variable "superfice", the recommendation made consists of reducing the tariff coefficient by approximately 17% for houses with a surface area of less than 40 square meters, as well as increasing it by about 9% for houses with a surface area greater than 400 square meters.

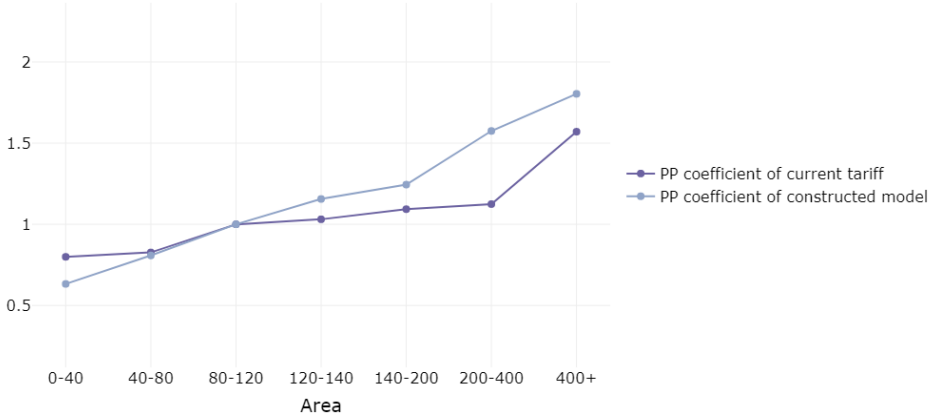


Figure 18 – Comparison of pure premium model coefficients vs. current pure premium

Once the tariff gaps were evaluated for each criterion, an action plan was developed to adjust the coefficients associated with the various criteria present in the current tariff structure, either upward or downward. Furthermore, recommendations were formulated for incorporating new significant variables revealed by the constructed model into the current tariff. Among these variables are "zonier DDE" and "groupe terme contrat". These variables utilized in Homeowners Insurance prove relevant for integration into the pricing of non-occupant property owner (PNO).

The study conducted within the scope of this thesis for water damage coverage will be extended to cover all coverages of the PNO product. This approach would allow for an overall assessment of the product's technical balance.

Table des matières

Introduction

Depuis les dix dernières années, le marché de l'assurance habitation évolue dans un contexte marqué par la hausse de la sinistralité causée par les dérèglements climatiques (grêle, inondations, sécheresse...). De plus, depuis 2021, les retombées inflationnistes des crises géopolitiques et économiques (augmentation des coûts des matériaux et des réparations ...) viennent dégrader la rentabilité des assureurs des dommages aux biens. Ces derniers sont ainsi confrontés au défi d'ajuster les primes pour faire face aux coûts croissants de ce contexte défavorable.

En outre, une saturation est constatée sur le marché de l'assurance habitation. Cela conduit à une concurrence féroce entre les nombreux acteurs pour conquérir les différentes parts de ce marché. Cette concurrence est accentuée davantage par l'arrivée des Assurtechs et par la législation en faveur des assurés. En effet, l'entrée en vigueur des lois Hamon et Chatel a renforcé le pouvoir des clients qui sont devenus plus exigeants et volatils.

Compte tenu de ces éléments, il est plus coûteux pour les assureurs de se démarquer de toutes les offres commerciales agressives pour conquérir de nouveaux clients. Ainsi, ils ont tout l'intérêt à se tourner vers leurs portefeuilles existants et à actionner des leviers pour fidéliser leurs clients. Pour ce faire, le multi-équipement se manifeste comme un outil stratégique pour maintenir un bon niveau de rentabilité.

Dans ce contexte, le produit Propriétaire Non-Occupant (PNO) s'affiche clairement comme un produit d'accompagnement de l'offre principale, notamment pour les clients patrimoniaux. Cependant, ce produit est en déficit technique, un ratio combiné [1] supérieur à 100%. Il est donc crucial de trouver des leviers à actionner pour s'approcher de l'équilibre technique.

Le présent mémoire s'inscrit dans ce cadre en proposant un levier d'action tarifaire, celui de challenger la tarification actuelle par une approche de modélisation de la prime pure. La dernière révision structurelle date de 2016 et seuls des renouvellements annuels à plat sont effectués. Il s'avère ainsi nécessaire de s'assurer de la cohérence du tarif vis à vis du risque encouru.

1. Il s'agit d'un indicateur de rentabilité. Il rapporte aux cotisations la charge des sinistres nette de réassurance, les frais de gestions et les commissions

Pour ce faire, ce mémoire s'articulera autour de cinq parties :

Nous débuterons par une présentation du produit propriétaire non-occupant dans le marché assurantiel français. Ses spécificités au sein de MMA seront présentées également.

Dans la deuxième partie, nous nous intéresserons à la construction de la base de données et à l'analyse des variables segmentantes. Une attention particulière sera accordée à l'utilisation d'un zonier géographique à une maille fine, de données clients qui contribuent à avoir une meilleure connaissance du profil des assurés et de données externes qui complètent les informations recueillies lors de la souscription.

Dans la troisième partie, nous testerons l'approche tarifaire traditionnelle couramment utilisée en non-vie. Nous modéliserons à cet effet la fréquence et le coût moyen des sinistres par les modèles linéaires généralisés (GLM).

Dans la quatrième partie, nous examinerons l'apport de deux méthodes de machine learning dans la tarification du produit PNO. Nous évaluerons dans ce cas, les performances de deux modèles non-paramétriques d'apprentissage statistique, à savoir le Random Forest et le Gradient Boosting. Une comparaison de ces différents modèles nous permettra ensuite de choisir le modèle le plus performant à retenir.

L'étude se conclura par la formulation de préconisations concrètes pour les prochains renouvellements à terme permettant ainsi d'aligner le tarif au risque réel encouru.

Première partie

Contexte

Chapitre 1

L'assurance Propriétaire Non Occupant (PNO)

1.1 Présentation de l'assurance Propriétaire Non Occupant

L'assurance **P**ropriétaires **N**on-**O**ccupants (PNO) est une assurance habitation qui s'adresse aussi bien aux personnes physiques qu'aux personnes morales (par exemple les Sociétés Civiles Immobilières SCI) qui sont propriétaires de biens immobiliers qu'ils n'occupent pas. Son objectif est de permettre à ces propriétaires de se prémunir contre les sinistres qui peuvent survenir à leurs biens. Elle est principalement destinée aux biens mis en location.

1.2 Les garanties assurables

Cette assurance couvre principalement les dommages causés à un logement en cas de :

- **Dégâts des eaux (DDE)** en prenant en charge les dégâts causés par des fuites d'eau, ruptures des conduites, pénétrations accidentelles, refoulements des égouts, etc. Cependant, cette garantie ne couvre pas les frais de réparation des toitures et terrasses à l'origine du sinistre ainsi que les dommages dus à un défaut d'entretien, un manque de réparation ou de l'usure des conduites et appareils ;
- **Incendie (INC)** en indemnisant les dommages matériels résultant accidentellement d'un incendie, d'une explosion même consécutif à un attentat, de la chute directe de la foudre, de surtension ou sous-tension, d'un court-circuit ou encore du dégagement de la fumée ;
- **Tempête, Grêle et Neige (TGN)** en couvrant les dommages matériels causés directement par le vent, la grêle sur les toitures et les murs du domicile, le poids de la neige ou de la glace accumulée ou encore la chute de pierres ou de rochers ;
- **Vol** en indemnisant les biens mobiliers en cas de perte ou de détérioration résultant de vols et de tentatives de vol ;
- **Vandalisme** en couvrant les dommages matériels (par exemple en cas de casse ou de graffitis), subis par les biens immobiliers assurés à la suite d'un acte de vandalisme causé à l'extérieur de ceux-ci. À noter que les dommages d'incendie, d'explosion, d'action de l'eau,

consécutif à un acte de vandalisme sont indemnisés au titre des garanties « Incendie et risques annexes » et « Dégâts des eaux » ;

- **Bris des glaces (BDG)** en prenant en charge le bris des vitres, y compris vitres d'inserts, glaces, miroirs, fenêtres, séparateurs de balcons, etc., ainsi que les dommages au mobilier assuré ;
- **Dommages électriques (DEL)** en complément de la garantie « Incendie ». Elle couvre les dommages électriques matériels subis par les appareils électriques à caractère mobilier (four, plaque de cuisson...) de moins de 10 ans résultant de la chute directe de la foudre, d'une surtension, sous-tension ou d'un court-circuit ;
- **Catastrophes naturelles (CATNAT)** qui complète les garanties précédemment citées qui ne s'appliquent pas en cas de catastrophe naturelle. Cette garantie ne peut être mise en jeu qu'après la publication au Journal Officiel de la République française d'un arrêté interministériel ayant constaté l'état de catastrophe naturelle ;
- **Responsabilité civile du propriétaire (RCP)** qui indemnise, à la place du propriétaire, les dommages causés à autrui et pouvant provenir d'un vice de construction ou d'un défaut d'entretien de l'habitation ;

1.3 La législation de l'assurance PNO

Le caractère obligatoire de l'assurance PNO

Contrairement à l'assurance habitation destinée aux locataires, la souscription d'une assurance PNO ne revêt pas un caractère obligatoire dans un certain nombre de cas. Cependant, l'article 58 de la **loi ALUR** du 26 mars 2014 établit une exception à cette règle, imposant la nécessité de souscrire à une assurance PNO dans une situation précise :

Article 9-1

« Chaque copropriétaire est tenu de s'assurer contre les risques de responsabilité civile dont il doit répondre en sa qualité soit de copropriétaire occupant, soit de copropriétaire non-occupant. Chaque syndicat de copropriétaires est tenu de s'assurer contre les risques de responsabilité civile dont il doit répondre. »

Cependant, il est important de souligner que l'assurance PNO en responsabilité civile, telle que requise par la loi ALUR, ne couvre que les dommages causés aux tiers. Elle n'indemnise en aucun cas les dommages matériels subis par le propriétaire lui-même.

Afin d'illustrer davantage, prenons l'exemple d'un propriétaire d'appartement qui a souscrit uniquement à une assurance PNO en responsabilité civile, conformément à la loi ALUR. Supposons qu'un dégât des eaux survienne et affecte à la fois ce propriétaire et son voisin. Dans cette situation, le voisin serait le seul à recevoir une indemnisation, laissant le propriétaire sans aucune compensation.

Il convient donc de noter que souscrire uniquement à une assurance PNO en responsabilité civile, bien que requise par la loi, peut laisser les propriétaires exposés à des risques financiers importants en cas de sinistre les affectant directement. De ce fait, il est souvent recommandé

aux propriétaires de souscrire à une assurance PNO plus complète qui couvre également les dommages matériels causés à leur propre bien, afin de bénéficier d'une protection adéquate dans de telles situations.

Convention d'indemnisation

La convention **CIDRE** est la Convention d'Indemnisation Directe et de Renonciation à Recours exclusivement en dégâts des Eaux. Créée en 2001 et mise en place en 2002, cette convention visait à simplifier et accélérer le règlement de la plupart des sinistres en dégâts des eaux dans les immeubles. Or, elle s'est vue remplacer le 1er juin 2018 par la convention **IRSI** (Convention d'Indemnisation et de Recours des Sinistres Immeuble) qui prend aussi en compte les incendies.

La convention IRSI vise une meilleure gestion du sinistre par rapport à la convention Cidre en fixant des règles claires des assureurs. En effet, la Convention :

- Désigne l'assureur de l'occupant du logement où le sinistre a eu lieu comme gestionnaire du sinistre. Dans le cas des copropriétaires non-occupants ou en cas de non-assurance/défaut de l'assurance de l'occupant, c'est l'assureur du (co)-propriétaire non-occupant qui interviendra. Celui-ci est chargé de vérification et l'évaluation des dommages ainsi que la recherche de fuite si nécessaire ;

- Comprend deux tranches de sinistre en fonction du montant des dommages matériels et des frais afférents :

  - la tranche 1 pour les sinistres inférieurs à 1600 € HT, où l'assureur gestionnaire évalue les dommages et prend en charge les dommages immobiliers et mobiliers ;

  - la tranche 2 pour les sinistres compris entre 1600 € et 5000 € HT, où une expertise est organisée pour toutes les parties et les assureurs prennent en charge les dommages en fonction de la propriété des biens. Les sinistres au-delà de 5000 € ne rentrent pas dans le cadre de cette convention ;

En résumé, la convention IRSI facilite l'indemnisation en permettant aux assureurs des différents intervenants de se répartir les responsabilités et les indemnisations. Cela signifie que les assureurs PNO peuvent être sollicités pour participer financièrement à l'indemnisation des dommages matériels causés à l'immeuble, même s'ils ne couvrent pas normalement ce type de risques, ce qui peut se traduire par une augmentation des coûts.

1.4 Le marché actuel de l'assurance PNO

Le marché de l'assurance habitation est en constante évolution, avec une augmentation continue du montant des chiffres d'affaires. Cette tendance est mise en évidence par une étude menée par la fédération France Assureurs (FA) dont les résultats sont représentés dans la figure 1.1.

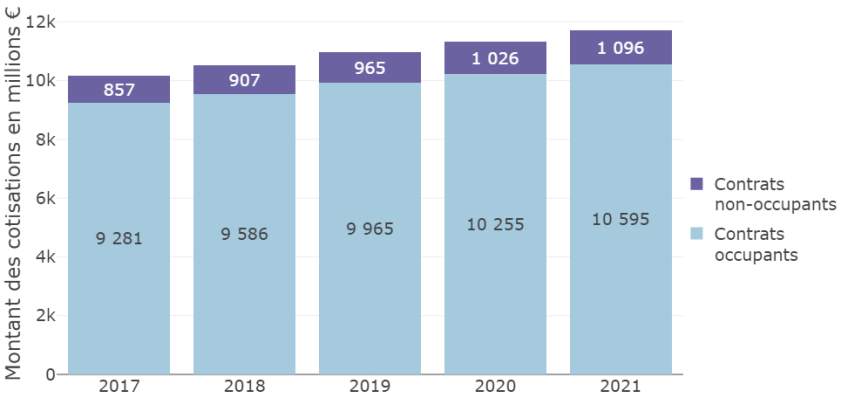


Figure 1.1 – Evolution des cotisations des contrats d'assurance habitation (en Md€) par type de contrats sur la période 2017-2021

Nous constatons une croissance significative du chiffre d'affaires des contrats d'assurance non-occupants au cours des dernières années. Entre 2017 et 2021, le chiffre d'affaires de ce type de contrat a augmenté de 27,8%. Une hausse notable de 6,8% a été enregistrée entre 2020 et 2021. En plus, cette croissance de ce marché s'avère plus rapide que celle observée sur le reste du marché de l'habitation, ce qui souligne l'intérêt manifeste d'investir dans les assurances pour les propriétaires non-occupants.

Cette évolution est principalement portée par deux facteurs. Premièrement, nous observons dans la figure 1.2 une augmentation des volumes des contrats non-occupants, avec une croissance de 3,8% entre 2020 et 2021. Deuxièmement, la prime moyenne a également connu une augmentation de 2,8% sur cette même période.

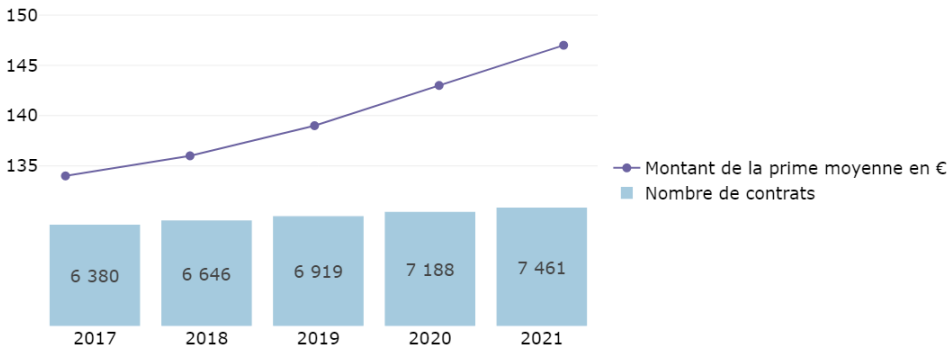


Figure 1.2 – Evolution du montant de la prime moyenne d'habitation (en €) sur la période 2017 - 2021

Cependant, il est important de noter que l'assureur ne peut pas continuellement répercuter la hausse des primes moyennes sur ses assurés. De plus, la législation française en vigueur favorise la concurrence, ce qui peut également limiter la capacité des assureurs à augmenter les tarifs de manière significative. Cette situation est encore amplifiée par l'arrivée des bancassureurs et des assurtechs qui renforcent la concurrence sur le marché.

En lien avec cette compétition croissante, la fédération FA a publié le classement, représenté ci-dessous, des assureurs en termes de chiffre d'affaires hors taxes réalisé en France en 2021 pour l'offre PNO. Or, il est important de noter que plusieurs assureurs n'ont pas communiqué les chiffres d'affaires du produit PNO, notamment les leaders en assurance MRH, à savoir Covéa,

Crédit Agricole Assurances et Groupama.

Rang	Société	Chiffre d'affaires 2021	Chiffre d'affaires 2020	Variation 2021-2020
1	Groupe Maif	105	97	8.2%
2	AXA	93	92	1.1%
3	Macif	79.3	77.9	1.8%
4	Groupe des assurances du Crédit Mutuel	57	52	9.6%
5	Matmut	39.7	38.3	3.7%
6	Abeille	33.3	32.3	3.4%
7	Allianz	22.4	21.6	3.6%
8	Société générale Assurances	20	18.8	6.4%
9	Generali	17	19	-10.5%
10	Suravenir Assurances	10.9	10.6	2.8%

Table 1.1 – Classement des assureurs en termes de chiffre d'affaires (HT) en M€ sur la période 2020-2021 pour l'offre PNO

MMA se positionne devant Matmut avec un chiffre d'affaires de 45.6 millions euros en 2021 avec son offre PNO. Dans la suite, un état des lieux de cette offre sera présenté.

1.5 Le produit PNO commercialisé par MMA

Présentation du produit PNO

Commercialisé depuis 2005, le produit Propriétaire Non-Occupant de MMA Assurances cible les propriétaires qui n'occupent pas leur logement. Sur la base des données de ces quatre dernières années, il représente 10% en nombre de contrats et 16% en chiffre d'affaires du portefeuille global d'assurance habitation de la marque. Sa distribution est principalement déléguée à des agents généraux avec une faible participation des courtiers. A sa création, ce produit comportait trois formules différentes :

- **La formule 1** est la formule minimale englobant les garanties responsabilité civile propriétaire, défense pénale et recours, incendie-explosion, dégâts des eaux, vandalisme extérieur, catastrophes naturelles et assistance en cas de sinistre ;

- **La formule 2** comprend, en plus des garanties de la première, les garanties bris de vitres, vol et tentative de vol et valeur à neuf sur mobilier ;
- **La formule 3** n'est plus commercialisée depuis 2013. Elle complétait la deuxième formule.

À l'instar de plusieurs assureurs, une des clés stratégiques de MMA Assurances est le multi-équipement en produits d'assurance de ses assurés par le développement des ventes croisées. Il s'agit d'équiper le client en plus du contrat initial d'un ou de plusieurs contrats additionnels chez le même assureur. D'une part, cette stratégie permet d'augmenter le volume des primes perçues tout en réduisant les frais d'acquisition liés à ces contrats. D'autre part, elle fidélise les clients et réduit par conséquent les taux de résiliation. Le produit PNO chez MMA s'inscrit dans ce cadre. En effet, il est considéré comme un produit d'accompagnement pour multi-équiper les clients notamment les clients à valeur. Néanmoins, les résultats techniques du produit PNO sont en dessous des objectifs fixés. Il est donc important pour MMA Assurance d'identifier les clients les moins risqués d'un point de vue nature et qualité du bien, aléa géographique et profil des clients.

Tarification actuelle du produit PNO

La modélisation actuelle du tarif du produit PNO de MMA repose sur l'estimation de la prime pure. Pour l'obtenir, une modélisation Fréquence x Coût moyen a été réalisée à travers des modèles généralisés de régression linéaire (GLM) [1] pour chaque type de logement. Les principales variables tarifaires qui entrent dans le calcul du tarif sont :

- La superficie pour les maisons et le nombre de pièces pour les appartements ;
- Le niveau du capital mobilier ;
- La franchise ;
- Un zonier au département qui s'applique à toutes les garanties confondues ;

La dernière révision structurelle date de 2016 sur une base des données d'un historique de six ans allant de l'année 2010 à l'année 2015. La base tarifaire actuelle semble donc ancienne. Des interrogations peuvent être soulevées quant à sa capacité à refléter les risques réels encourus dans le portefeuille actuel. De surcroît, comme énoncé précédemment, le produit PNO n'atteint pas le niveau de rentabilité escompté et sa vitesse de production est en baisse au fil de ces dix dernières années. En effet, les évolutions annuelles du portefeuille en termes de nombre de contrats sont aux alentours de 1% ces 3 dernières années contre 4% - 5% il y a dix ans.

Il s'avère ainsi nécessaire d'améliorer la connaissance de la structure du portefeuille avec des modèles actualisés qui prennent en compte de nouvelles variables tarifaires segmentantes et permettent de mieux s'adapter à la situation des assurés et aux risques auxquels ils font face tout en maintenant un bon niveau de rentabilité.

1. Les aspects théoriques de la modélisation par les modèles linéaires généralisés (GLM) sont exposés dans le chapitre 5

Deuxième partie

Périmètre de l'étude et préparation des données

Chapitre 2

La construction de la base de données et les retraitements effectués

Ce chapitre a pour objet de présenter la base de données utilisée dans le cadre du mémoire. Après délimitation du périmètre du portefeuille, quelques chiffres seront donnés afin de le cerner dans sa globalité. Des analyses univariées et bivariées seront également menées afin de développer une première intuition quant au caractère segmentant ou non de certaines variables. La corrélation des différentes variables fera également l'objet d'une attention particulière.

2.1 Périmètre de modélisation

Filtres

Dans le cadre de ce mémoire, nous nous focaliserons exclusivement sur les contrats d'assurance distribués par les agents, en excluant ainsi les contrats distribués par les courtiers.

Choix de la garantie à modéliser

Dans le cadre de ce mémoire, le choix a été fait de ne traiter qu'une seule garantie, celle qui représente une charge des sinistres importante tout en offrant la possibilité d'actionner des leviers pour atteindre l'équilibre technique du produit PNO, conformément aux besoins de l'entreprise.

En analysant la figure 2.1 qui illustre la répartition de la charge des sinistres en fonction des différentes garanties, les garanties « Dégâts des eaux » et « Incendie » ont des poids similaires et se démarquent de manière notable, représentant à elles seules 75% de la charge globale des sinistres.

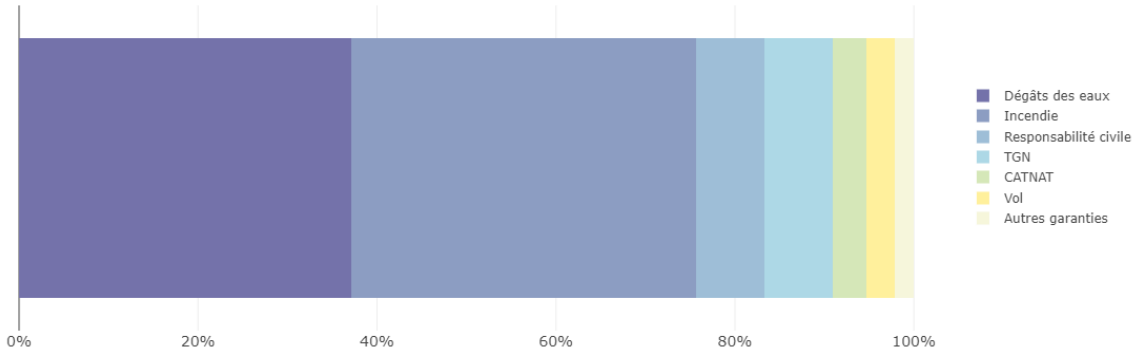


Figure 2.1 – Répartition de la charge totale des sinistres par garantie du produit PNO

La garantie incendie affiche des résultats dégradés qui sont fortement liés à des sinistres d'intensité forte mais avec une fréquence faible. Pour améliorer les résultats de cette garantie, les opportunités de mettre en place des leviers d'actions tarifaires se révèlent limitées conduisant à travailler davantage le périmètre de souscription du produit. Il s'agirait de limiter l'acceptation des grands risques susceptibles d'engendrer des sinistres incendie importants.

De surcroît, l'enjeu dans la garantie incendie est plus porté par les maisons alors que la garantie dégâts des eaux concerne autant les maisons que les appartements ce qui élargit le champ des leviers d'actions tarifaires à explorer. De ce fait, le choix s'est orienté vers **la modélisation de la garantie dégâts des eaux**.

Modélisation par type de logement

Nous avons décidé de créer deux modèles fréquence x coût moyen distincts pour la garantie dégâts des eaux par type de logement en raison de la répartition hétérogène de la sinistralité entre les maisons et les appartements.

À titre d'exemple, pour les appartements, le risque dégâts des eaux peut être grandement impacté par la présence de canalisations partagées, un facteur qui s'applique moins pour les maisons.

2.2 Construction de la base de données

Dans le cadre de cette étude, la base de données utilisée correspond à un portefeuille d'assurance propriétaire non-occupant sur la période 2011-2022. Contrairement au produit principal multirisques habitation, le produit PNO manque de volume d'affaires et par conséquent il comporte moins de sinistres. Ainsi, pour effectuer une modélisation de qualité, le choix d'un historique profond de 12 ans a été privilégié. Dans les parties suivantes, le processus de création de cette base de données sera explicité.

1. Extraction de la base de données contrats

Les données contrats sont disponibles dans des datamarts [1] avec un historique de 5ans et sont nommées DTM_MRH_YYYY. Afin d'avoir une base contrats PNO avec l'historique escompté, à savoir 12 ans, un filtre sur le produit PNO ainsi qu'une agrégation des 3 datamarts suivantes ont été effectués :

- DTM_MRH_2015 : pour récupérer les données contrats de l'année 2011 à l'année 2014 ;
- DTM_MRH_2018 : pour récupérer les données contrats de l'année 2015 à l'année 2018 ;
- DTM_MRH_2022 : pour récupérer les données contrats de l'année 2019 à l'année 2022.

Par ailleurs, la base obtenue est une base à la situation. Cette notion correspond à la période de temps durant laquelle le contrat ne subit aucune modification. Parmi les évènements qui peuvent changer la situation d'un contrat, nous retrouvons principalement :

- **Affaire Nouvelle (AN) :** qui fait référence à la souscription d'un nouveau contrat d'assurance ;
- **Avenant :** il s'agit d'un acte par lequel une modification du contrat est effectuée d'un point de vue risque ou de couverture des garanties.
- **Terme :** cette notion correspond à la date où le contrat d'assurance arrive à échéance ou se renouvelle automatiquement ;
- **Fin de l'année comptable :** En vertu du code de l'assurance, cette date correspond au 31 décembre.

Ainsi, en effectuant un découpage par situations, la table obtenue nous permet d'avoir une vision plus précise de l'état du contrat à chaque instant. Cette dernière est initialement composée de 500 variables et 5 millions de lignes.

2. Extraction de la base de données sinistres

Les données relatives aux sinistres sont accessibles via la table C_SINI_MENS qui regroupe l'ensemble des dossiers sinistres où une ou plusieurs garanties peuvent être ouvertes. Chaque ligne de cette table correspond à une garantie sinistrée.

De la même manière que pour la base des données contrats, nous avons extrait tous les sinistres survenus entre les années 2011 et 2022 vus à fin mars 2023 pour avoir la vision la plus récente sur l'état des sinistres. Ces données récupérées représentent un total de 1,5 million de lignes.

1. Un datamart est un regroupement de données traitant d'un même sujet. L'objectif principal d'un datamart est d'organiser les données de manière structurée et agrégée afin de répondre à des besoins spécifiques. Cela facilite l'utilisation des données au sein d'une entreprise.

3. Rapprochement des deux bases

Une fois la base de données contrats et la base de données sinistres extraites, nous avons procédé à la jointure de ces deux bases en utilisant comme clés de jointure les identifiants des contrats ainsi que leurs dates d'effet et de fin des situations. À cet effet, le sinistre devient un nouveau motif de découpage des situations. Les situations sinistrées sont alors dédoublées ce qui implique la création de nouvelles variables pour avoir l'historique complet de la situation et des montants de cotisations justes en fonction du nombre de jours de risque. Pour illustrer, prenons l'exemple des deux cas suivants :

A) Cas 1 : Plusieurs sinistres dans l'année

Dans le tableau ci-dessous, nous avons trois lignes de situations distinctes sur l'année 2019 :

- Situation n°1 (du 01/01 au 30/04) : pas de sinistres ;
- Situation n°2 (du 01/05 au 31/12) : survenance d'un sinistre « dommages électriques » - DEL le 04/06 ;
- Situation n°3 (du 01/05 au 31/12) : survenance d'un sinistre Responsabilité civile - RC le 17/08.

ID	Date début situ	Date fin situ	Date début situ recalculée	Date fin situ recalculée	Garantie sinistrée	Date survenance	Coût	NB jour risque	MT cotisation acquise HT
XYZ	01/05/2018	30/04/2019	01/05/2018	31/12/2018				245	135,21
XYZ	01/05/2018	30/04/2019	01/01/2019	30/04/2019				120	78,85
XYZ	01/05/2019	30/04/2020	01/05/2019	31/12/2019	DEL	04/06/2019	423,54	245	135,21
XYZ	01/05/2019	30/04/2020	01/05/2019	31/12/2019	RC	17/08/2019	818,78	245	135,21

Nous pouvons constater dans ce cas que le montant des cotisations acquises HT est doublé.

Ainsi, dans ce cas de figure, un dédoublement des lignes sinistrées sera effectué. La date de fin de la situation sera ajustée pour correspondre à la date de survenance du sinistre. Le nombre de jours de risque sera également modifié pour correspondre à la période entre les nouvelles dates de début et de fin des situations recalculées. Enfin, les montants des cotisations seront recalculés en fonction de la durée d'exposition.

ID	DT début situ recalculée	DT fin situ recalculée	DT début situ recalculée_2	DT fin situ recalculée_2	Garantie sinistrée	Date survenance	Coût	NB jour risque	MT cotisation acquise HT
XYZ	01/01/2019	30/04/2019	01/01/2019	30/04/2019				120	78,85
XYZ	01/05/2019	31/12/2019	01/05/2019	04/06/2019	DEL	04/06/2019	423,54	35	19,31
XYZ	01/05/2019	31/12/2019	05/06/2019	17/08/2019	RC	17/08/2019	818,78	74	40,84
XYZ	01/05/2019	31/12/2019	18/08/2019	31/12/2019				136	75,06

B) Cas 2 : Un dossier sinistre avec plusieurs garanties ouvrées

Prenons l'exemple suivant d'un seul sinistre qui a impacté deux différentes garanties. Nous avons alors trois lignes de situations distinctes sur l'année 2019 :

- Situation n°1 (du 01/01 au 31/07) : survenance d'un sinistre le 26/04 impactant la garantie RC ;
- Situation n°2 (du 01/05 au 31/12) : survenance d'un sinistre le 26/04 impactant la garantie DDE ;
- Situation n°3 (du 01/05 au 31/12) : pas de sinistres.

ID	Date début situ	Date fin situ	Garantie sinistrée	Date survenance	Coût	NB jour risque	MT cotisation acquise HT
UVW	01/08/2018	31/12/2018				153	88 ,52
UVW	01/01/2019	31/07/2019	RC	26/04/2019	818,78	212	120,85
UVW	01/01/2019	31/07/2019	DDE	26/04/2019	1800,54	212	120,85
UVW	01/08/2019	31/12/2019				153	88 ,52

ID	DT début situ	DT fin situ	DT début situ recalculée	DT fin situ recalculée	Garantie sinistrée	Date survenance	Coût	NB jour risque	MT cotisation acquise HT
UVW	01/08/2018	31/12/2018	01/08/2018	31/12/2018				153	88 ,52
UVW	01/01/2019	31/07/2019	01/01/2019	26/04/2019	RC	26/04/2019	818,78	116	66,12
UVW	01/01/2019	31/07/2019	01/01/2019	26/04/2019	DDE	26/04/2019	1800,54	116	66,12
UVW	01/08/2019	31/12/2019	27/04/2019	31/07/2019				96	54,72
UVW	01/08/2019	31/12/2019	01/08/2019	31/12/2019				153	88 ,52

De la même manière que le premier cas, nous pouvons constater que le montant des cotisations acquises HT est doublé. Ainsi, en reprenant la même démarche effectuée précédemment, nous retrouvons le tableau suivant :

À la différence du premier cas où nous avions affaire à deux sinistres distincts, nous avons ici un seul sinistre avec deux garanties impactées. Par conséquent, le nombre de jours de risque et le montant des cotisations sont doublés.

Pour remédier à ce problème, nous avons pris la décision de maintenir inchangée la première ligne (correspondant à la première garantie sinistrée) et nous avons fixé à zéro le nombre de jours de risque et le montant des cotisations de toutes les autres garanties sinistrées d'un même sinistre. De cette manière, nous évitons de compter deux fois le montant des cotisations acquises.

ID	DT début situ recalculée	DT fin situ recalculée	DT début situ recalculée	DT fin situ recalculée	Garantie sinistrée	Date survenance	Coût	NB jour risque	MT cotisation acquise HT
UVW	01/08/2018	31/12/2018	01/08/2018	31/12/2018				153	88 ,52
UVW	01/01/2019	26/04/2019	01/01/2019	26/04/2019	RC	26/04/2019	818,78	116	66,12
UVW	01/01/2019	26/04/2019	26/04/2019	26/04/2019	DDE	26/04/2019	1800,54	0	0
UVW	27/04/2019	31/07/2019	27/04/2019	31/07/2019				96	54,72
UVW	01/08/2019	31/12/2019	01/08/2019	31/12/2019				153	88 ,52

4. Enrichissement de la base d'étude

Le rapprochement et la mise en forme des bases de données étant accomplis, l'étape suivante consiste à enrichir la base de données par des variables pouvant contribuer à améliorer la connaissance du portefeuille et du risque encouru. Pour ce faire, les données relatives aux clients, les zoniers et des données externes ont été raccordés.

Les données relatives aux clients

Les données relatives aux clients fournissent des informations détaillées sur le profil de l'assuré telles que :

- Âge de l'assuré
- Ancienneté du contrat
- Equipement de l'assuré : variable qui indique si l'assuré est détenteur de contrat(s) PNO (mono équipé) ou bien s'il détient d'autres contrats en plus (MRH, Automobile, Epargne..)

Ces données sont disponibles dans deux bases de suivi du client. Elles sont raccordées grâce au numéro du dossier client.

Les zoniers

Les zoniers permettent de capturer la dimension géographique du risque. Leur construction se base sur un découpage du territoire en différentes zones de risques conformément à une maille préétablie.

L'ajout des zoniers semble pertinent étant donné que les facteurs géographiques jouent un rôle important dans la segmentation du risque habitation.

Dans le cadre de cette étude, nous avons raccordé :

- le zonier du produit PNO qui est commun à toutes les garanties. Il permet de distinguer six zones géographiques de sinistralité différentes selon la maille département, classées de « 1 » à « 6 » de manière croissante par rapport au risque.
- les zoniers du produit MRH pour les occupants spécifiques à chaque garantie (dégâts des eaux, incendie, vol...) qui sont à la maille IRIS [2]

L'ajout des données clients et des zoniers dans notre base de données construite précédemment, nous permet d'obtenir la base de données internes.

Les données externes

L'ajout de données externes en complément des variables explicatives internes dans les modèles de tarification non-vie surtout en habitation semble pertinent. Elles apportent des informations supplémentaires à celles communiquées par le client lors du parcours de souscription.

Dans le cadre de ce mémoire, nous avons retenu les données externes suivantes :

- **Données INSEE**
  - Données à la maille IRIS concernant des variables socio-démographiques (Nombre des ménages en couple avec enfants...) et socio-économiques (Nombre de cadres...).
  - Données à la maille IRIS concernant des variables liées aux logements (Nombre de résidences principales avec 5 pièces ou plus, nombre des résidences principales avec chauffage collectif...).
- **Données à l'adresse**
  - Données à la maille adresse concernant les caractéristiques de construction (Mitoyenneté...)
  - Données à la maille adresse concernant les caractéristiques de l'habitation (Présence d'une dépendance non attenante...)

2. Ilots Regroupés pour l'Information Statistique : Ce sont les plus petites unités spatiales statistiques, inférieures à l'échelon communal.

Ces données fournissent des informations supplémentaires susceptibles d'améliorer la connaissance de l'assuré, du bien assuré et de l'environnement l'entourant. Pour illustrer, prenons l'exemple de la variable indiquant la présence ou non de mitoyenneté entre deux habitations, cette information peut se révéler influente sur le risque de dégâts des eaux. En effet, lorsque deux habitations sont mitoyennes, c'est-à-dire qu'elles partagent la même clôture (mur mitoyen par exemple), la probabilité de survenance d'un sinistre dégâts des eaux est plus élevée que si elles sont indépendantes. Des problèmes tels que les fuites de tuyaux, de canalisations peuvent se propager plus facilement entre habitations mitoyennes.

Pour tirer pleinement parti de ces données externes, un processus essentiel concernant leur raccordement à la base de données internes a été réalisé. Il est schématisé dans la figure 2.2 et est décrit comme suit :

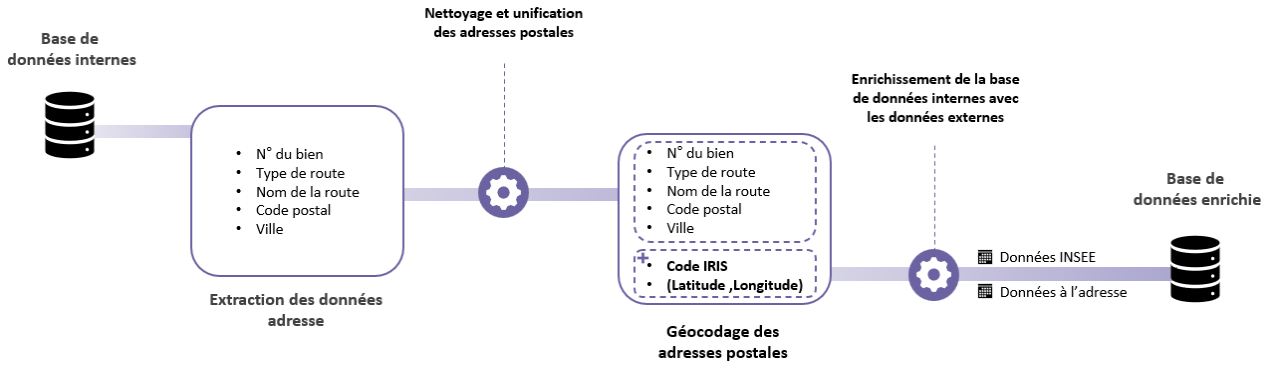


Figure 2.2 – Processus de raccordement des données externes à la base de données internes

➤ Traitement des adresses

Après l'extraction des différentes adresses de la base de données, seules celles comportant les informations suivantes ont été maintenues :

- Numéro du logement ;
- Type et nom de la route ;
- Nom de la ville ;
- Code postal.

Cette étape n'a pas empêché de garder la quasi-totalité des adresses extraites. Seules 3% ont été supprimées.

Une première tentative infructueuse visant à récupérer les données à l'adresse à leur état brut, a révélé l'importance d'effectuer des pré-retraitements des adresses postales. En effet, la qualité dégradée de certaines adresses a rendu leur exploitation en l'état impossible. Il était donc indispensable d'effectuer certaines modifications sur ces adresses :

- Suppression des caractères spéciaux (« ; », « ? », « ! », ...) ;
- Suppression des accents ;

- Correction manuelle des adresses comportant des erreurs de saisie en effectuant des recherches sur Google Maps.

➤ Géocodage des adresses

Une fois les adresses extraites et retraitées, l'étape suivante consiste à assigner à chaque adresse du portefeuille, un couple de coordonnées géographiques (latitude/longitude). Ce processus est connu sous le terme de géocodage [3]. En utilisant ces coordonnées géographiques, il sera possible de faciliter la géolocalisation des contrats et, en intégrant également le code IRIS, de les croiser avec des variables géographiques externes. Cela permettra in fine d'avoir la localisation précise de chaque sinistre et les facteurs géographiques contribuant à l'explication de sa fréquence de survenance ou bien de sa sévérité.

➤ Enrichissement avec les données externes

Grâce au géocodage, les diverses données externes présentées précédemment ont été raccordées à la base des données internes. Ce qui nous permet ainsi de construire notre base de données enrichie. Elle est constituée de :

- données liées aux contrats recueillies à la souscription ;
- données liées à la sinistralité historique observée ;
- données relatives aux clients ;
- données géographiques ;
- données externes.

Le tableau 2.3 résume l'ensemble des variables disponibles dans la base d'étude finale.

3. Le géocodage ne fera pas l'objet d'une description détaillée dans ce mémoire. Néanmoins, c'est une méthode très connue dans les sciences cartographiques qui bénéficie déjà d'une documentation étendue.

Information	Nom de la variable	Définition
Contrat	ID Contrat	Numéro du contrat, Numéro du dossier client
	DTDEFFSIT	Date d'effet de la situation
	DTFEFFSIT	Date de fin de la situation
	NU_DC	Numéro du dossier client
	Adresse du bien	Voie, Code postal, Département, IRIS, Longitude, Latitude ...
	Formule	La formule choisie du contrat d'assurance. Elle prend 3 modalités : F1, F2 et F3.
	Franchise	La franchise prévue dans le contrat d'assurance. Elle prend 3 modalités: 0€, 140€ et 280€
Sinistre	DT_survenance	Date de survenance du sinistre
	Gar_sinistree	La garantie sinistrée
	CD_CAUS_SINI	Cause du sinistre
	RA	Exposition au risque exprimée en jours
	Cout	Coût du sinistre
	NB_SIN	Nombre de sinistres
Bien assuré	CD_TYPE_HABI	Type de logement
	Superficie	Variable propre aux maisons . Elle correspond à la superficie d'une maison. Elle est exprimée en tranches
	Nb_logements	Variable propre aux maisons . Elle permet de distinguer une maison oridinaire d'un immeuble d'habitation appartenant en totalité à l'assuré
	Nb_pieces	Variable propre aux appartements. Elle correspond au nombre de pièces principales
	Niv_Capital_Mobilier	Niveau du capital mobilier exprimé en K€. Elle prend les modalités suivantes : 0, 6, 11, 50, 100, 200
Géographique	Zoniers	Zonier de l'offre PNO + Zonier de la garantie dégâts des eaux de l'offre principale (12 zones) + Zonier de la garantie Inondation de l'offre principale (6 zones)
Client	Anciennete	Ancienneté de l'assuré. Cette variable prend des valeurs de "1" à "30+"
	Age	L'âge de l'assuré. Cette variable prend les modalités suivantes: "0-18", "18-25", "26-35", "36-55", "56-65", "66+"
	Equipement	Type de contrats souscrits par le client PNO (PNO seul, Contrats MRH, Auto,Epargne, Santé, Pro, Autres ...)
	Groupe_terme_client	Classification commerciale du client vue en fin 2022. Elle prend les modalités suivantes: "Sinistrés", "Standard", "Privilège" et "VIP".
	NB_contrats_en_cours	Nombre de contrats en cours. Les valeurs vont de "1" à "15+"
	Sinant_X	Nombre de sinistres sur les X = 12, 24, 36 derniers mois
Externe	Données INSEE	Données socio-démographiques (Nombre des ménages en couple avec enfants...), socio-économiques (Nombre de cadres...), liées aux logements (Nombre des résidences secondaires...)
	Données à l'adresse	Nombre de voisins sur un rayon de 50mètres, Mitoyenneté, Présence d'une dépendance non-attenante, Présence d'une risque de DDE vertical ...

Figure 2.3 – Liste des variables de la base de données finale

2.3 Retraitements de la base de données

Traitement des valeurs aberrantes et des informations manquantes

Notre base de données finale étant construite, nous portons notre attention dans cette section au contrôle de la qualité et la fiabilité des données qu'elle contient. À cette fin, nous vérifions l'absence des valeurs aberrantes et nous nous assurons que les données soient complètes.

Pour le traitement des valeurs manquantes, plusieurs méthodes existent :

- Suppression des lignes avec des données manquantes ;
- Suppression des colonnes avec des données manquantes ;
- Maintien des valeurs non renseignées ;
- Imputation des données manquantes ;
- Création d'une modalité « Non Connu (NC) » pour les données manquantes.

L'étape de l'enrichissement de la base de données a engendré des valeurs manquantes sur les variables raccordées. Les variables présentant un taux de complétude inférieur à 90% ont été retirées. C'est le cas de quelques données à l'adresse comme les matériaux des murs ou le nombre

d'étage. Pour les autres variables, le choix de créer une modalité « NC » a été privilégié afin de préserver l'intégralité de l'information relative à ces variables.

Notre base de données comporte aussi quelques données aberrantes telles que des périodes d'exposition négatives ou des âges supérieurs à 150 ans. La plupart des anomalies ont été corrigées selon des approximations simples (l'âge peut être retrouvé avec la variable année de naissance, etc.). Pour les anomalies les plus complexes à approximer, vu qu'elles représentent moins de 2% de la base, il n'a pas été nécessaire d'utiliser un algorithme d'imputation. Dans le cadre de cette étude, nous avons décidé de les exclure.

Prise en compte de l'inflation annuelle

Plusieurs éléments peuvent être à l'origine de la hausse permanente des coûts des sinistres : des conditions météorologiques, des changements réglementaires et aussi des facteurs économiques. Ce dernier point est à prendre en considération car il peut introduire un biais important. En effet, étant donné l'historique de douze ans qui a été considéré dans notre base d'étude, il s'avère difficile de déterminer si l'évolution du coût d'un même sinistre est due à des caractéristiques qui lui sont propres ou si c'est à cause de l'inflation annuelle. Afin d'assurer la comparabilité d'une année à l'autre, les coûts des sinistres doivent être corrigés de l'inflation.

Pour ce faire, il existe deux principales méthodes :

- La prise en compte de l'année de survenance du sinistre au sein du modèle en tant que variable qualitative ;
- La mise en ***as-if*** des sinistres. Cela consiste à estimer le coût des sinistres passés comme s'ils survenaient à la date de vision (l'année 2022 dans le cadre de cette étude). L'estimation s'effectue en utilisant différents indices.
  - **ICC FFB** : Indice du Coût de la Construction établi par la Fédération Française du Bâtiment. Cet indice est calculé trimestriellement, en se basant sur le prix de revient d'un immeuble avec plusieurs étages à Paris, hors prix du terrain d'origine.
  - **IPEA** : Indice des Prix des travaux d'Entretien-Amélioration des bâtiments, émis trimestriellement par l'INSEE. Il mesure l'évolution des prix hors TVA pratiqués par les entreprises de la construction (les entreprises artisanales incluses) pour leurs travaux d'entretien et d'amélioration des bâtiments résidentiels et non résidentiels réalisés au cours du trimestre estimé.
  - **Indice BT01** : Indice national du bâtiment, publié mensuellement par l'INSEE. Il quantifie l'évolution générale des coûts de la construction en France.

En comparant l'évolution du coût moyen observé de notre base de données à celles des différents indices cités précédemment, l'indice FFB du coût de la construction avait l'évolution la plus proche. Ainsi, il a été décidé de le retenir pour redresser les sinistres. Cet indice est aussi utilisé au sein de la marque pour l'indexation des capitaux mobiliers garantis aux contrats.

Traitement des sinistres

➤ Les sinistres clos

Un sinistre clos est un sinistre survenu, pour lequel la charge finale est déjà connue.

Les sinistres clos sont pris en compte pour la modélisation de la fréquence et du coût moyen.

➤ Les sinistres en cours

Un sinistre en cours est un sinistre déclaré par l'assuré mais qui n'est toujours pas réglé par l'assureur. Ce type de sinistre a une incidence limitée sur les modèles de fréquence étant donné que le statut « en cours » n'altère pas la réalité de la survenance effective d'un sinistre. En revanche, ces sinistres peuvent introduire un biais dans les modèles de coût moyen résultant d'un manque de recul sur le montant de la charge du sinistre.

Dans le cadre du produit PNO, la liquidation des sinistres dégâts des eaux est rapide et leur charge n'est pas susceptible d'évoluer fortement. En effet, la figure 2.4 illustre l'évolution du coefficient d'ultimisation associé aux garanties dégâts des eaux, incendie et responsabilité civile. Le coefficient d'ultimisation des sinistres de la garantie dégâts des eaux est assez stable, augmentant uniquement de 5% sur une période de développement de six ans, contrairement aux deux autres garanties.

Dans ce contexte, nous avons décidé de ne pas projeter à l'ultime la charge des sinistres en cours. Nous disposons ainsi d'un volume suffisant de sinistres sans qu'un biais significatif soit présent dans la charge totale des sinistres.

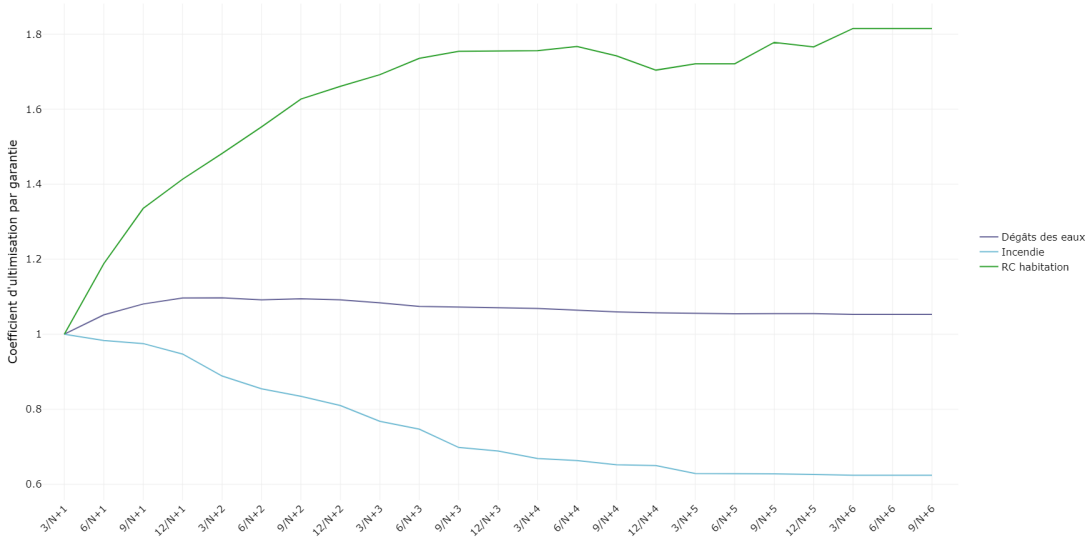


Figure 2.4 – Evolution des coefficients d'ultimisation associés aux garanties dégâts des eaux, incendie et responsabilité civile

➤ Les sinistres sans suite

Un sinistre sans suite fait référence à une situation où l'assuré déclare un sinistre qui n'aboutit pas à une indemnisation après l'évaluation de l'assureur. Par exemple, cela peut se produire lorsque les coûts de réparation des dommages d'un sinistre n'excèdent pas le montant de la franchise.

La base de données comporte 1% des sinistres sans suite. Ces sinistres sont retirés pour ne pas biaiser les modèles de fréquence en comptant des sinistres qui n'ont pas été considérés par la suite ou en surestimant la charge des sinistres dans les modèles de coût moyen.

➤ Les sinistres graves

Dans notre base de données, une homogénéité de la sinistralité est constatée sur les appartements. Cependant, le coût des sinistres sur les maisons varie considérablement allant de 0€ à 377 699 €. Cette forte hétérogénéité peut altérer la robustesse du modèle de coût moyen. En effet, il est important de modéliser distinctement les sinistres attritionnels (correspondant à des sinistres avec de forte fréquence et avec des charges de sinistres relativement faibles) et les sinistres graves (correspondant à des sinistres moins fréquents mais avec un coût important) puisqu'en général, ils ne suivent pas la même loi de distribution.

De ce fait, il est judicieux de déterminer un seuil au-dessus duquel la charge du sinistre est identifiée comme extrême. Pour ce faire, nous avons utilisé la méthode à dépassement de seuil (mean-excess plot) qui nous permet de visualiser l'allure de la queue d'une distribution et de définir ainsi le seuil à utiliser. Dans la suite, nous expliciterons le principe de cette méthode.

Soit X une variable aléatoire intégrable de fonction de répartition F et de point terminal x_F . Pour tout $u < x_F$, la fonction des excès moyens (Mean Excess) de X au-dessus du seuil u est définie pour tout x $\in$ [0,x$_F$ - u[par :

$$e(u) = \mathbb{E}[X - u | X > u], \quad u > 0$$

Remarques :

- Si $X \sim \mathcal{E}xp(\lambda)$ alors $e(u) = \lambda$ pour tout $u > 0$, la fonction des excès moyens est donc constante et aura une allure horizontale ;
- Si $X \sim GPD(\xi, \sigma)$ [4] alors $e(u) = \frac{\sigma}{1-\xi} + \frac{\xi}{1-\xi}u$ pour tout $u > 0$, la fonction des excès moyens est donc affine et aura une allure linéaire.

L'estimateur empirique de $e(u)$ est noté $\hat{e}_n(u)$. Il se définit de la manière suivante :

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n} (X_i - u)^+}{\sum_{i=1}^{n} \mathbb{I}_{X_i > u}}$$

Le numérateur représente la somme des excès, c'est à dire la partie des coûts de sinistres au-delà du seuil u et le dénominateur représente le nombre de sinistres qui dépassent le seuil u. L'outil mean excess plot trace cet estimateur et nous permet de déterminer la loi suivie par les excès en analysant son allure.

Le théorème de Pickands-Balkema-de Haan est de grande utilité quant à la détermination du seuil au-dessus duquel la charge du sinistre est identifiée comme extrême. Il stipule que pour un certain seuil u suffisamment grand, la loi des excès peut être approchée par la loi de Pareto Généralisée.

4. Distribution de Pareto généralisée (Generalized Pareto Distribution)

En nous basant sur ce résultat et sur les remarques citées précédemment, nous cherchons à identifier le seuil u qui correspond à la plus petite valeur au-delà de laquelle l'excès résiduel moyen est linéaire. Le graphique 2.5 montre que l'excès résiduel commence à être linéaire au-dessus de 30 000 €.

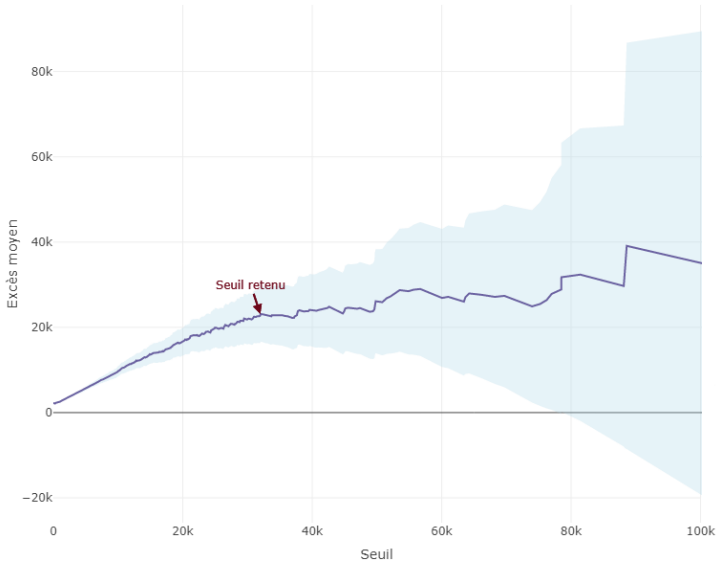


Figure 2.5 – Mean Excess des coûts de sinistres dégâts des eaux

Dans ce mémoire, seul le coût moyen des sinistres attritionnels de la garantie dégâts des eaux sera modélisé. Deux approches sont envisageables pour les sinistres identifiés comme graves :

- Les sinistres graves feront l'objet d'une troncature au seuil de 30 000 €, la partie de la charge du sinistre qui est inférieure à ce seuil sera conservée dans le modèle des sinistres attritionnels et la partie dépassant ce seuil sera traitée dans le modèle des sinistres graves ;
- Les sinistres graves ne feront pas l'objet d'une troncature. La charge totale du sinistre grave sera exclue du modèle des sinistres attritionnels et sera modélisée dans le modèle des sinistres graves.

Les sinistres dégâts des eaux qui excèdent le seuil de 30 000 € représentent à peine 2% de la charge totale des sinistres.

Le choix de la deuxième méthode a été privilégié dans cette étude. Ainsi, tout sinistre dépassant le seuil de 30 000 € est considéré comme grave et sera écarté du modèle de coût moyen.

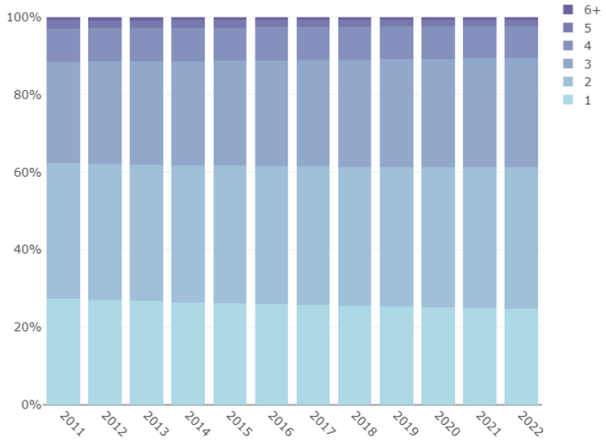
Chapitre 3

L'étude des variables

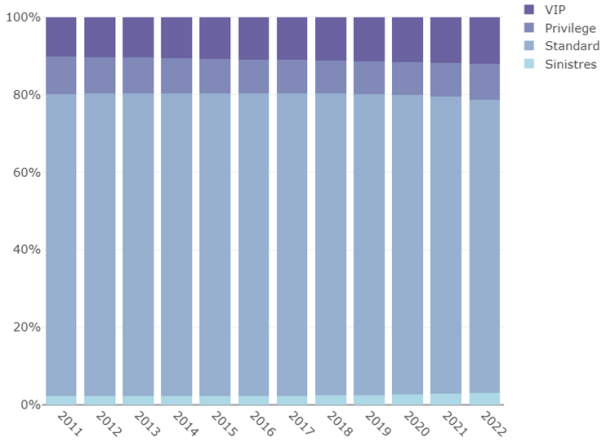
Après avoir effectué les différents retraitements, nous nous intéressons dans cette partie à l'étude de la stabilité du portefeuille dans le temps et à la réalisation des analyses descriptives des variables.

3.1 La stabilité du portefeuille

Il est important dans un premier temps de garantir la stabilité temporelle du portefeuille. Pour cela, nous avons analysé l'évolution de l'exposition annuelle des modalités relatives à chaque variable sur la période d'observation 2011-2022. Il en résulte que le portefeuille est globalement stable dans le temps. Pour ne pas alourdir le mémoire, nous exposons deux exemples représentatifs :



(a) Stabilité temporelle de la variable «Nombre de pièces»



(b) Stabilité temporelle de la variable «Groupe terme client»

3.2 Analyses descriptives des variables

La stabilité du portefeuille étant vérifiée, une analyse des données a été menée dans le but d'améliorer notre connaissance du portefeuille et d'obtenir une première impression sur les variables potentiellement influentes pour la garantie dégâts des eaux. Par souci de parcimonie, seules les analyses sur les maisons seront présentées.

Analyses uni-variées en fréquence

➤ **Exemple de variables liées au bien**

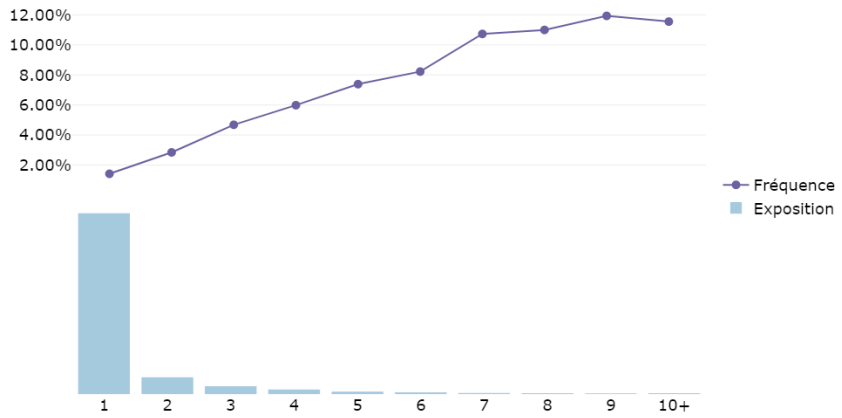


Figure 3.1 – Fréquence observée selon le « Nombre de logements »

Notons qu'une maison comportant deux logements ou plus correspond à un immeuble d'habitation appartenant en totalité à l'assuré.

La fréquence des sinistres dégâts des eaux est croissante avec le nombre de logements.

En cas de sinistre dégâts des eaux, dû à titre d'exemple à des fuites des tuyaux ou de canalisations, survenu à l'intérieur de l'immeuble, l'eau peut ruisseler le long des murs ou des plafonds. Ainsi, les dégâts des eaux peuvent se propager verticalement au sein de l'immeuble, ce qui peut expliquer la croissance de la fréquence.

➤ **Exemple de variables client**

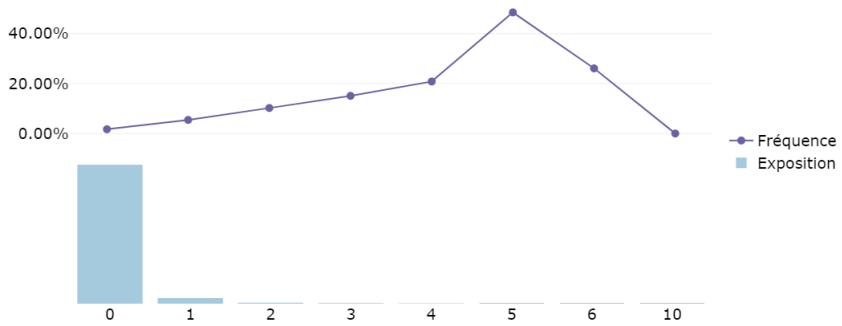


Figure 3.2 – Fréquence observée selon le « Nombre de sinistres antérieurs sur 36 mois »

Le nombre de sinistres antérieurs sur 36 mois présente un très fort impact discriminant sur la fréquence de sinistres lorsque les expositions sont significatives.

➤ **Exemple de variables géographiques**

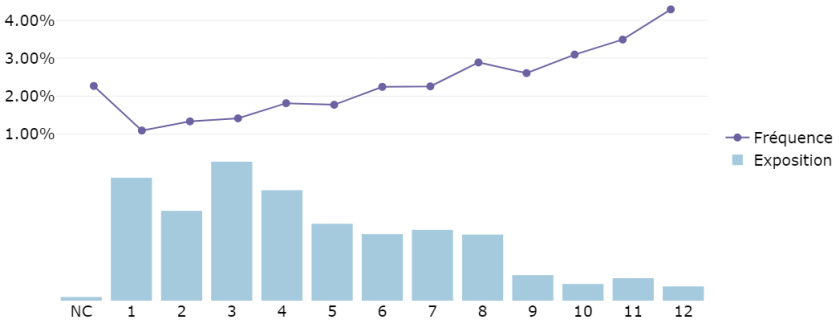


Figure 3.3 – Fréquence observée selon le « Zonier DDE »

Le zonier de la garantie dégâts des eaux permet de distinguer 12 zones géographiques de sinistralité différentes, classées de « 1 » à « 12 » de manière croissante par rapport au risque.

Une nette tendance haussière est constatée dans l'évolution de la fréquence des sinistres selon les différentes zones de risque. La sinistralité de la zone « 12 » est quatre fois plus élevée que celle de la zone « 1 ». Cette variable apparaît comme discriminante sur la fréquence des sinistres.

➤ **Exemple de variables externes**

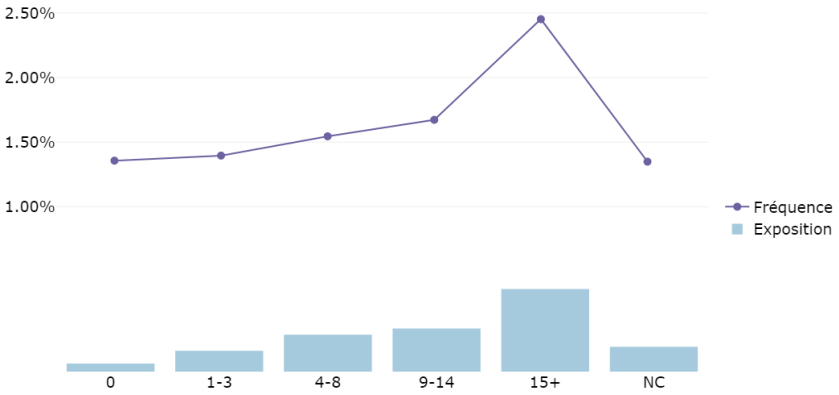


Figure 3.4 – Fréquence observée selon le « Nombre de voisins sur un rayon de 50 mètres »

La fréquence augmente avec le nombre de voisins dans un rayon de 50 mètres. Plus le nombre de voisins est important, plus la fréquence des sinistres est élevée.

Analyses uni-variées en coût moyen

➤ **Exemple de variables liées au bien**

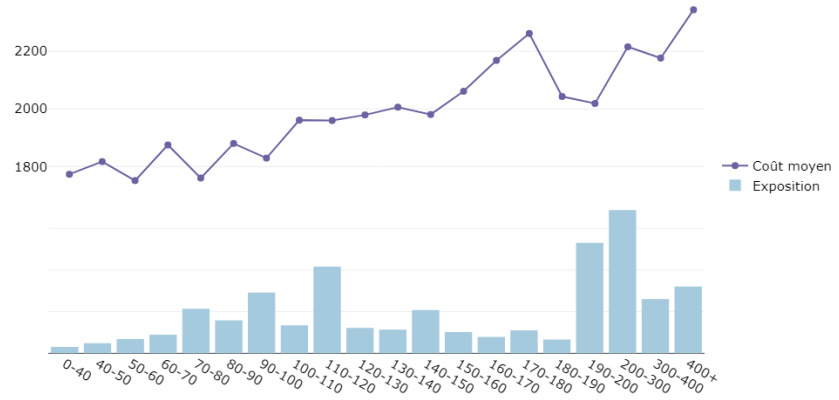


Figure 3.5 – Coût moyen observé (en €) selon la « Superficie (en m^2)»

Une tendance haussière se dégage de l'évolution du coût moyen des sinistres avec la superficie des maisons. L'évolution est cohérente, plus la superficie est grande, plus le coût du risque est élevé.

➤ **Exemple de variables client**

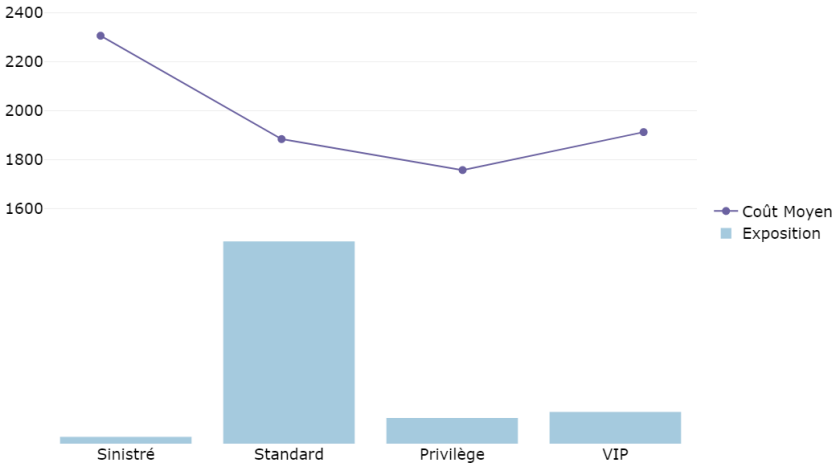


Figure 3.6 – Coût moyen observé (en €) selon le « Groupe terme client »

Cette variable est conçue par la marque à partir de plusieurs indicateurs dont l'équipement, le montant des cotisations et la sinistralité. Elle permet d'avoir la vision des profils des clients à l'échéance. Elle classe en conséquence les assurés en quatre catégories.

L'évolution du coût moyen selon les différentes catégories est cohérente avec la nature inhérente de la variable. La catégorie « privilège », étant construite pour regrouper les assurés avec le meilleur comportement selon les indicateurs cités précédemment, se distingue par le coût moyen le plus faible. À contrario, la catégorie « sinistré » présente le coût moyen le plus élevé étant donné qu'elle regroupe les assurés avec une sinistralité significative.

➤ **Exemple de variables géographiques**

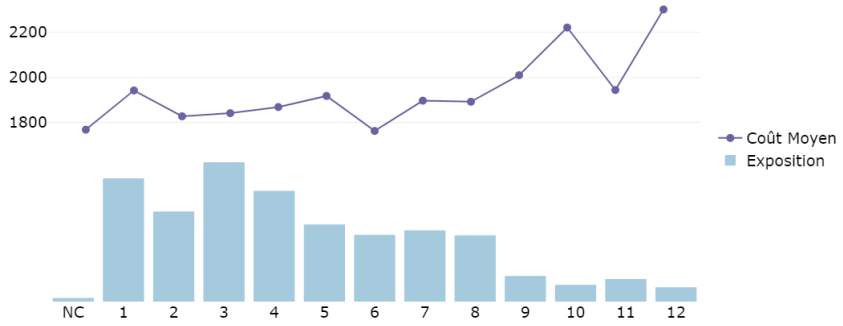


Figure 3.7 – Coût moyen observé (en €) selon le « Zonier DDE »

La tendance haussière constatée dans l'évolution du coût moyen en fonction des différentes zones de risque est moins marquée que celle observée dans l'évolution de la fréquence. Cette variable sera néanmoins conservée pour la modélisation du coût moyen de la garantie dégâts des eaux.

➤ **Exemple de variables externes**

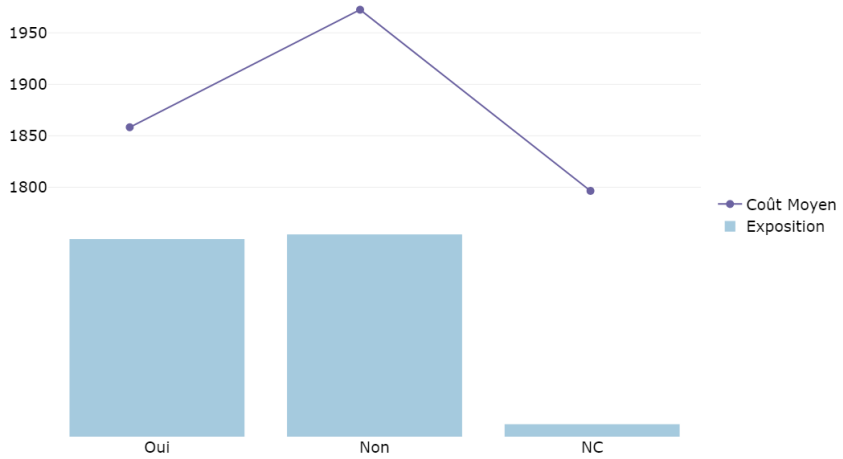


Figure 3.8 – Coût moyen observé (en €) selon la « Mitoyenneté »

Les modalités de la variable « mitoyenneté » sont représentées de la même manière en termes d'exposition dans le portefeuille. Par ailleurs, le coût des sinistres est plus faible en cas de mitoyenneté. En cas de survenance d'un sinistre dégâts des eaux et en présence de mitoyenneté, le coût du sinistre peut éventuellement être partagé entre les assureurs de chaque partie concernée.

3.3 Étude des corrélations entre les différentes variables

Une fois les analyses univariées effectuées, il est important de s'intéresser aux interactions entre les variables candidates. L'objectif est de détecter les dépendances entre les variables explicatives et d'évaluer la force de ces dépendances afin d'éviter de biaiser les modèles GLM.

Pour ce faire, différentes mesures de dépendance existent : le coefficient de Pearson, le T de Tschuprow, les corrélations de rang de Spearmann ou encore le V de Cramer. Dans ce mémoire, nous avons décidé d'utiliser le V de Cramer car il présente des avantages tels que l'invariance à la taille de l'échantillon et au nombre d'observations, ainsi que la capacité à traiter des variables catégorielles. Afin d'utiliser cette méthode, toutes les variables quantitatives ont été discrétisées en classes. Cela permet de les transformer en variables qualitatives.

Dans la suite, nous allons rappeler brièvement le cadre du test du χ^2 qui établit la dépendance ou l'indépendance de deux variables sans la quantifier.

Test d'indépendance du χ^2

Soient deux variables qualitatives X et Y à respectivement de r et h modalités avec r,h ≥ 2 et décrivant les caractéristiques de n individus. Ces variables sont mesurées simultanément sur N unités statistiques. Nous souhaitons étudier le croisement de ces deux variables, nous construisons alors un tableau de contingence, c'est-à-dire un tableau dénombrant les modalités croisées des deux caractères X et Y.

	y_1	y_2	...	y_j	...	y_h
x_1	n_{11}	n_{12}		n_{1j}		n_{1h}
x_2	n_{21}	n_{22}		n_{2j}		n_{2h}
...						
x_i	n_{i1}	n_{i2}		n_{ij}		n_{ih}
...						
x_r	n_{r1}	n_{r2}		n_{rj}		n_{rh}

où les $(n_{ij})_{ij}$ sont les effectifs empiriques entre X et Y. Nous notons donc $n_{i,} = \sum_j n_{ij}$ et $n_{,j} = \sum_i n_{ij}$

Afin de tester l'indépendance des variables X et Y, nous formulons deux hypothèses :

- L'hypothèse nulle H_0 : X et Y sont indépendantes ;
- L'hypothèse alternative H_1 : X et Y ne sont pas indépendantes.

Afin de réaliser ce test, nous introduisons la statistique du χ^2 qui quantifie la distance entre les effectifs observées et les effectifs empiriques n_{ij} sous l'hypothèse d'indépendance des variables, à savoir les effectifs donnés par $n^*_{ij} = \frac{n_{i,}n_{,j}}{n}$.

Celle-ci s'exprime comme suit :

$$\chi^2 = \sum_i \sum_j \frac{n_{ij} - n^*_{ij}}{n^*_{ij}}$$

Cette statistique varie de 0 à $+\infty$. Elle est égale à 0 lorsque nous sommes face à une situation d'indépendance. Cependant, celle-ci peut prendre une valeur strictement positive si deux variables X et Y ne sont pas strictement indépendantes. Cela signifie que cette statistique ne permet pas de conclure automatiquement à une dépendance entre les variables, car des fluctuations d'échantillonnage pourraient expliquer ces résultats.

Pour pallier à ce problème, il faut introduire une mesure normalisée dérivée de la statistique de χ^2, dont nous connaissons la valeur maximale.

V de Cramer

Le V de Cramer constitue une solution pertinente à ce problème, car il permet de compenser à la fois l'effet du nombre d'observations et celui du nombre de modalités. Il est défini comme suit :

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

Cette mesure est échelonnée entre 0 et 1. Une valeur proche de 1 indique une forte dépendance entre les variables candidates, ce qui signifie qu'elles sont fortement associées. Inversement, une valeur proche de 0 suggère une faible dépendance entre les variables.

Calcul des corrélations entre les variables

Après étude des corrélations entre les différentes variables présentes dans la figure 3.9, nous avons procédé à la suppression d'une des deux variables qui présentent une corrélation très forte (valeur du V de Cramer > 0.6). La présence de deux variables fortement corrélées dans le même modèle risque de le biaiser. Cela est notamment observé entre certaines variables internes telles que le nombre de gammes détenues par l'assuré et le nombre de contrats en cours, ainsi qu'entre des variables externes comme le nombre de voisins et la variable indiquant si l'habitation se situe en quartier prioritaire.

Pour les variables moyennement corrélées (valeur du V de Cramer entre 0.4 et 0.6), à dire d'expert, il a été décidé de ne pas les exclure immédiatement. Nous préférons plutôt évaluer leur significativité dans le modèle. C'est le cas notamment des zoniers des garanties dégâts des eaux, inondation et celui du produit PNO.

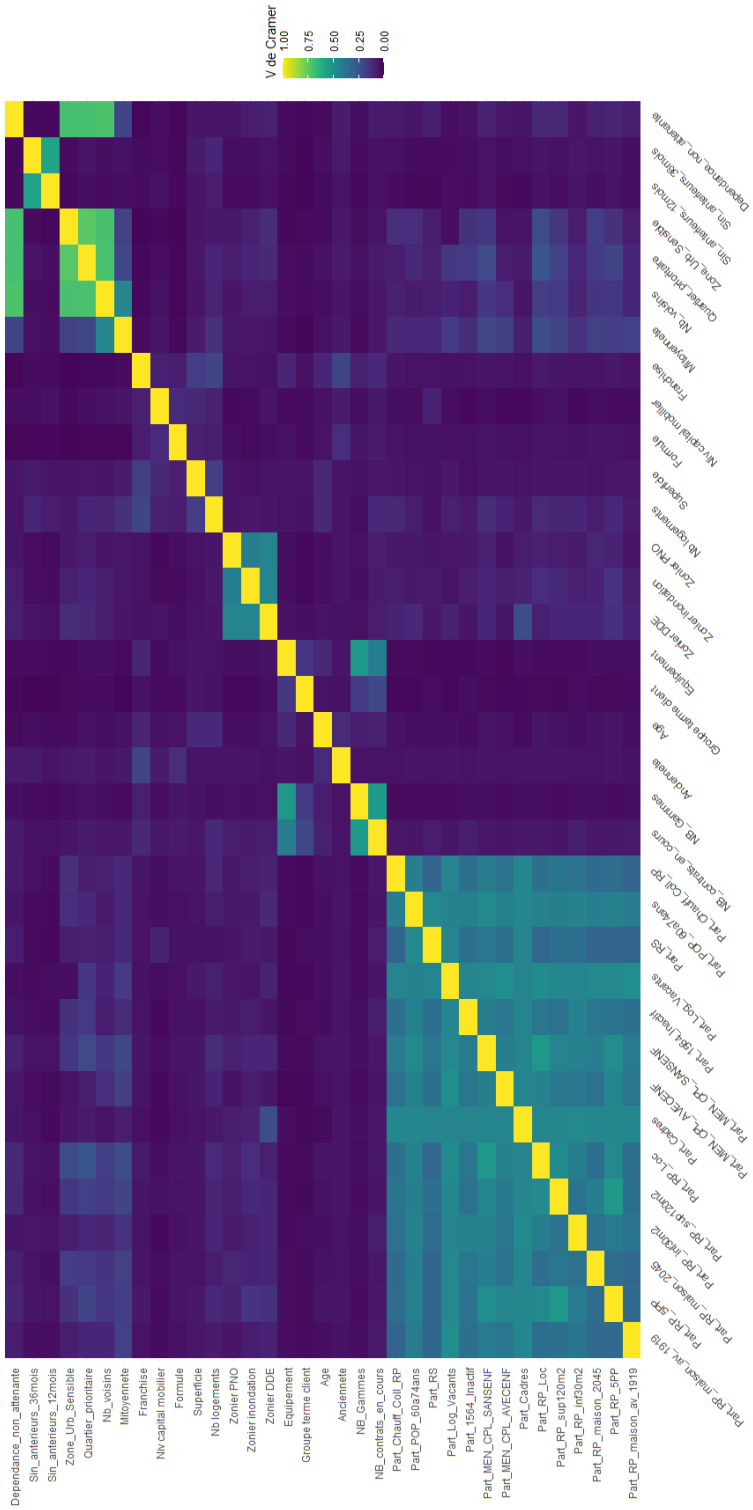


Figure 3.9 – Matrice de corrélation de V de Cramer

Troisième partie

Les aspects théoriques

Chapitre 4

La tarification dans l'assurance non-vie

4.1 Les enjeux de la tarification en assurance habitation

Les principes de la tarification en assurance habitation

L'activité de l'assurance non-vie est caractérisée par l'inversion du cycle de production. L'assureur ignore l'échéancier et le montant de versement des indemnisations en cas de réalisation du risque couvert (survenance des sinistres). Il est ainsi primordial pour l'assureur d'estimer le tarif adapté aux risques encourus et au portefeuille assuré.

Ce tarif doit prendre en compte à la fois le coût du risque pur, appelé également la prime pure, et les chargements nécessaires pour couvrir les frais occasionnés par l'assureur. Toutefois, l'évaluation précise du tarif se heurte à un enjeu majeur celui de l'asymétrie de l'information.

L'asymétrie d'information se caractérise par l'inégalité de connaissance entre l'assureur et l'assuré sur le risque couvert. L'assuré a une meilleure connaissance de son risque. Cette asymétrie peut engendrer deux types de risques :

L'antisélection

L'antisélection survient lors de la souscription d'un contrat d'assurance où l'assuré dispose de plus d'informations que l'assureur. Par conséquent, l'assureur peut faire face à une souscription massive de mauvais risques qui se font passer pour de bons risques.

L'aléa moral

L'aléa moral se manifeste après la souscription ou la modification d'un contrat d'assurance. Il se traduit par un changement du comportement de l'assuré, ce qui a pour effet de rendre le risque plus important pour l'assureur. En effet, lorsqu'une personne est assurée, son degré de précaution et de vigilance peut être altéré par la couverture d'assurance, induisant une augmentation de l'exposition au risque. À titre d'exemple, un individu ayant souscrit à une assurance habitation avec une garantie incendie pourrait être moins enclin à investir dans des dispositifs de prévention tels qu'un système de détection d'incendie.

La segmentation et la mutualisation des risques

La mutualisation des risques constitue le pilier central de l'assurance. Elle consiste à regrouper un ensemble d'assurés exposés aux mêmes risques pour que des compensations puissent s'opérer entre eux. Les primes versées par les assurés servent à indemniser les sinistres survenus à quelques-uns d'entre eux. De cette manière, les risques sont solidarisés et répartis entre eux.

Cette mutualisation est pertinente uniquement si les assurés sont regroupés dans des classes de « risques homogènes ». Autrement dit, si les assurés ont les mêmes profils de risque. En effet, si les regroupements sont hétérogènes, l'assureur peut se retrouver face aux problématiques de sélection adverse ou d'antisélection explicitées plus tôt. Faire payer la même prime à tous les assurés paraît inéquitable. Les assurés considérés comme « bons risques » et ayant un faible historique de sinistralité sont désavantagés puisqu'ils payent une prime supérieure au risque auquel ils sont exposés ce qui pourrait les inciter à résilier. En revanche, les assurés considérés comme « mauvais risques » ayant un historique de sinistralité important sont favorisés puisqu'ils payent une prime inférieure au risque auquel ils s'exposent. Ainsi, l'assureur est confronté à la nécessité de segmenter son portefeuille.

La segmentation du portefeuille consiste à le découper en classes de risques homogènes dans l'objectif de proposer le juste prix à payer à chaque classe. A cet effet, les assureurs ont recours aux données collectées lors de la souscription liées aux assurés, aux biens assurés et tendent même aujourd'hui à l'utilisation de données externes pour affiner davantage leurs segmentations. Voire même certains d'entre eux, s'orientent vers la personnalisation des tarifs.

Confronté à ce dilemme, L'assureur doit trouver un compromis entre segmentation et mutualisation. Une segmentation très fine (hyper-segmentation) risque de complexifier le tarif et va à l'encontre du principe fondamental de la mutualisation dont se base l'assurance. Inversement, une forte mutualisation appliquée à l'ensemble du portefeuille de l'assureur, risque de le retrouver avec un portefeuille composé principalement de mauvais risques et pourrait conduire à la dégradation de sa rentabilité.

4.2 Le modèle fréquence x coût moyen en tarification non vie

L'élaboration d'une tarification en assurance non vie (multirisque habitation, auto...) repose traditionnellement sur la modélisation de la prime pure dans le cadre général des modèles fréquence x coût moyen. Dans ces modèles, la fréquence moyenne des sinistres et leurs coûts sont modélisés séparément. Cette approche permet d'intégrer les variables tarifaires les plus pertinentes qui expliquent la variable modélisée en l'occurrence la fréquence ou le coût moyen. Dans le cadre de ce mémoire, c'est cette approche qui sera utilisée.

Cadre mathématique du modèle

Notations

Pour un portefeuille de risque donné :

- $N \in \mathbb{N}$ variable aléatoire représentant le nombre total de sinistres survenus au cours d'une période donné (une année en général) ;
- $X_i \in \mathbb{R}^+$: variable aléatoire représentant le montant du $i^{\text{ème}}$ sinistre ;

- $S = \sum_{i=1}^{N} X_i$ variable aléatoire représentant la charge totale des sinistres d'un risque au cours d'une année.

Hypothèses du modèle

- Les $(X)_i, \quad 1 \leq i \leq N$ sont indépendantes et identiquement distribuées (iid) ;
- Indépendance entre la fréquence et le coût des sinistres : càd $\forall i$ entre 1 et N $X_i \perp \mathbb{N}$

Les hypothèses étant posées, il est possible de déterminer l'espérance de la charge totale des sinistres $\mathbb{E}[S]$ comme suit :

$$\begin{aligned}
E[S] &= \sum_{k=0}^{\infty} E[S|N=k] \cdot P(N=k) \\
&= \sum_{k=0}^{\infty} E\left[\sum_{i=1}^{N} X_i | N=k\right] \cdot P(N=k) \\
&= \sum_{k=0}^{\infty} E\left[\sum_{i=1}^{k} X_i | N=k\right] \cdot P(N=k) \\
&= \sum_{k=0}^{\infty} E\left[\sum_{i=1}^{k} X_i\right] \cdot P(N=k) \quad \text{puisque } X_i \perp N \\
&= \sum_{k=0}^{\infty} \sum_{i=1}^{k} E[X_i] \cdot P(N=k) \quad \text{puisque } X_i \quad iid \\
&= \sum_{k=0}^{\infty} kE[X_1] \cdot P(N=k) \\
&= E[X_1] \sum_{k=0}^{\infty} kP(N=k) \\
&= E[X_1] \cdot E[N]
\end{aligned}$$

$$E[S] = E[X_1] \times E[N] = \text{coût moyen} \times \text{nombre moyen de sinistres.}$$

La prime pure selon le modèle « fréquence x coût moyen» est obtenue en rapportant l'espérance de la charge totale des sinistres à l'exposition [1] du portefeuille.

$$\text{Prime Pure} = \frac{E[S]}{\text{Exposition}} = \frac{E[N]}{\text{Exposition}} \times E[X_1] = \text{fréquence moyenne} \times \text{coût moyen.}$$

Habituellement, la modélisation du coût moyen et de la fréquence de sinistres est effectuée par les modèles linéaires généralises (GLM). Les aspects théoriques de cette modélisation seront abordés plus en détail dans le chapitre suivant.

1. L'exposition du portefeuille correspond à la somme des expositions de chaque assuré. Par exemple : l'exposition d'un assuré ayant souscrit à un contrat PNO prenant effet le 01/01/2023 et arrivant à échéance le 30/03/2023 est de 0,25 pour l'année 2023.

Chapitre 5

Le modèle linéaire généralisé

5.1 Le modèle

Les modèles linéaires généralisés (GLM) sont utilisés en tarification Non-Vie dans le domaine de l'assurance. Ils ont été introduits par Nelder et Wedderburn en 1972 et sont considérés comme la référence de la tarification Non-Vie en raison de leur interprétabilité et de leur facilité d'application dans les plans tarifaires.

Le principe des GLM est de déterminer la loi de probabilité de la variable à expliquer Y (nombre de sinistres ou coût de sinistres) en fonction des variables explicatives

$$X_1, ..., X_{p-1}$$

, en spécifiant leur relation à travers trois composantes :

1. **Composante aléatoire :** Elle permet d'identifier la distribution de probabilités de la variable à expliquer Y. Nous supposons que l'échantillon statistique est composé de n variables aléatoires indépendantes $(Y_1, ..., Y_n)^t$ qui appartiennent à la famille exponentielle, ce qui signifie que la densité f de ces variables est de la forme suivante :

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

où :

- $\theta \in \mathbb{R}$ est le paramètre dit canonique ou naturel de la loi de distribution ;
- $\phi \in \mathbb{R}$ est le paramètre de dispersion de la loi de distribution ;
- a est une fonction définie sur $\mathbb{R}$ et est non nulle ;
- b est une fonction connue définie sur $\mathbb{R}$, deux fois dérivable et sa dérivée b' est inversible ;
- c est une fonction définie sur $\mathbb{R}$, connue et dérivable.

Les deux principales propriétés qui caractérisent la famille exponentielle sont les suivantes :
Pour $i = 1, ..., n$

- $\mathbb{E}(Y_i) = b'(\theta_i)$
- $\text{Var}(Y_i) = b''(\theta) * a(\phi)$

Cette formulation étend les modèles de régression linéaire aux cas où Y ne suit pas une loi normale mais une loi de la famille exponentielle, qui inclut la loi normale comme cas particulier.

2. **Composante déterministe :** Les réalisations des variables explicatives $X_{1,i}, ..., X_{p-1,i}$ utilisées comme prédicteurs dans le modèle définissent sous forme d'une combinaison linéaire $(X\beta)$ la composante déterministe.

3. **Fonction de lien :** une fonction g déterministe, bijective, strictement monotone et définie sur $\mathbb{R}$, qui relie l'espérance mathématique de la variable de réponse Y à la composante déterministe. Nous avons alors :

$$g(\mathbb{E}(Y)) = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1}$$

Le tableau ci-dessous présente des exemples de fonctions liens usuelles :

Loi	Nom du Lien	Fonction de Lien
Normale	Lien identité	$g(\mathbb{E}(Y)) = \mathbb{E}(Y)$
Poisson	Lien log	$g(\mathbb{E}(Y)) = \ln(\mathbb{E}(Y))$
Gamma	Lien inverse	$g(\mathbb{E}(Y)) = \frac{1}{\mathbb{E}(Y)}$
Binomiale	Lien logit	$g(\mathbb{E}(Y)) = \ln\left(\frac{\mathbb{E}(Y)}{1-\mathbb{E}(Y)}\right)$

5.2 Estimation des paramètres

L'estimation des paramètres $\beta_0, \beta_1, ..., \beta_{p-1}$ des GLM se fait par la méthode du maximum de vraisemblance décrite ci-dessous.

Soit Y_i la variable réponse, indépendante et appartenant à une famille exponentielle. La vraisemblance s'écrit sous la forme suivante :

$$L(y_1, \ldots, y_n; \theta, \phi) = \exp\left(\sum_{i=1}^{n} y_i\theta_i - b(\theta_i)a_i(\phi) + c(y_i, \phi)\right)$$

L'expression de la log-vraisemblance est obtenue comme suit :

$$log(L) = \sum_{i=1}^{n} y_i\theta_i - b(\theta_i)a_i(\phi) + c(y_i, \phi) \text{ avec } L = L(y_1, \ldots, y_n; \theta, \phi)$$

Cette dernière expression est à maximiser, en calculant les dérivées partielles

$$\frac{\partial \log(L)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \left(\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right)$$

Ainsi, les équations de vraisemblance à résoudre sont :

$$\frac{\partial \log(L)}{\partial \beta_j} = 0 \quad \forall j = 0, \ldots, p-1$$

Pour résoudre ces équations, des méthodes numériques sont nécessaires telle que la méthode de Newton-Raphson.

5.3 Significativité des variables

Pour améliorer la qualité d'un modèle linéaire généralisé, on peut réduire l'espace des variables explicatives aux variables significatives en utilisant des tests statistiques de significativité tels que le test de Wald exposé ci-dessous.

Soit le test suivant :

$$H_0 : \forall j, \beta_j = 0 \quad \text{contre} \quad H_1 : \exists j, \beta_j \neq 0$$

La statistique de Wald s'écrit alors :

$$W = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$$

Sous H_0, la statistique du test suit approximativement une loi Normale $N(0, 1)$.

Le test de Wald est défini par :

$$Z = W^2 = \left(\frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)} \right)^2$$

Sous H_0, cette statistique suit une loi du khi-deux à un degré de liberté.

5.4 Sélection des variables

L'enjeu des modèles de régression est de parvenir à expliquer au mieux la variable réponse tout en minimisant le nombre de variables explicatives. Différentes méthodes permettent de sélectionner les variables les plus influentes. Parmi elles :

- **la méthode Forward :** aussi connue sous le nom de méthode ascendante. Cet algorithme part d'un modèle réduit avec tous les coefficients nuls à l'exception de la constante β_0. Les variables sont alors ajoutées successivement en incluant d'abord la variable la plus significative jusqu'à ce qu'aucune des variables restantes n'apporte d'amélioration au modèle lorsqu'elle y ajoutée ;

- **la méthode Backward :** aussi connue sous le nom de méthode descendante. Cet algorithme démarre du modèle complet. À chaque étape, il procède à l'élimination de la variable la moins significative. Il s'arrête dès lors qu'aucune des variables restantes ne contribue à l'amélioration du modèle lorsqu'elle est supprimée ;
- **la méthode Stepwise :** fusionne les méthodes forward et backward. L'algorithme par d'un modèle nul et, à chaque itération de la méthode ascendante, une étape d'élimination en arrière est effectuée. Il s'arrête lorsque plus aucune variable ne peut améliorer le modèle, que ce soit par ajout ou par élimination. C'est la méthode que nous préconisons dans cette étude.

Différents critères peuvent être utilisés pour départager les variables à ajouter ou à supprimer. Les critères AIC et BIC sont les plus répandus.

- **Akaike Information Criterion (AIC) :** $\mathrm{AIC} = -2\log(L) + 2n_{\mathrm{par}}$

  où L est la fonction de vraisemblance du modèle et n_{par} représente le nombre de paramètres inconnus.
- **Bayesian Information Criterion (BIC) :** $\mathrm{BIC} = -2\log(L) + \log(n) \cdot n_{\mathrm{par}}$

  où L est la fonction de vraisemblance du modèle, n représente le nombre d'individus et n_{par} représente le nombre de paramètres inconnus.

Dans le cadre de cette étude, le choix du critère AIC a été privilégié pour la sélection des variables.

5.5 Validation du modèle

Mesure de la qualité prédictive d'un modèle

Afin d'évaluer la qualité prédictive d'un modèle, la méthode « testset validation » est souvent employée. Elle assure la comparaison de la qualité prédictive des modèles sur un échantillon de données n'ayant pas servi à l'apprentissage du modèle. Elle repose sur la division du jeu de données en deux sous-ensembles distincts :

- L'ensemble d'apprentissage est utilisé pour estimer les différents paramètres du modèle. Il est obtenu en choisissant aléatoirement un certain pourcentage des observations (dans le cadre de cette étude, ce pourcentage est égal à 70% de la base).
- L'ensemble de test est constitué des observations restantes et sert à évaluer la qualité du modèle en calculant l'erreur de prédiction.

Pour une comparaison équitable, la graine aléatoire est, dans un premier temps, fixée pour la construction des différentes modèles. Pour chaque variable explicative, une analyse préliminaire a été effectuée pour garantir que les bases d'apprentissage et de test soient représentatives de la base globale.

Une fois les modèles construits, une analyse complémentaire a été entreprise pour vérifier la robustesse des résultats en faisant varier la graine aléatoire.

En ce qui concerne le calcul de l'erreur de prédiction sur la base de test, le critère couramment utilisé est l'erreur quadratique moyenne (Root Mean Square Error) qui est défini par :

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

où $\hat{y}_i$ est la valeur prédite par le modèle ajusté sur l'ensemble d'apprentissage, y_i est la valeur observée, et n représente le nombre d'observations dans l'ensemble de test.

Analyse des résidus

Dans le cadre de la validation du modèle GLM, les résidus permettent de vérifier les hypothèses du modèle (la loi paramétrique et la fonction lien choisie).

Les résidus sont basés sur la distance entre l'observation réelle y_i et sa valeur prédite $\hat{\mu}_i$ par le modèle.

Les résidus linéaires empiriques sont définis comme suit :

$$\hat{r}_i = y_i - \hat{\mu}_i$$

Plusieurs manières existent pour calculer les résidus. Parmi elles, nous retrouvons principalement :

- Les résidus de Pearson :
$$r_{p_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$
- Les résidus de déviance :
$$r_{d_i} = \text{signe}(y_i - \hat{\mu}_i)\sqrt{d_i}$$
où d_i est la contribution de l'observation i à la déviance $D = \sum_{i=1}^{n} d_i$

Chapitre 6

Les méthodes d'apprentissage statistique

6.1 Généralités

Avec la croissance exponentielle des données numériques, l'utilisation de nouvelles méthodes d'analyse est devenue essentielle. Les méthodes d'apprentissage statistique ont connu un réel essor puisqu'elles correspondent à ce besoin.

L'apprentissage statistique est une approche permettant à un algorithme d'analyser un ensemble de données statistiques afin d'apprendre à fournir des prédictions ou des classifications précises sans avoir été explicitement programmé à cet effet au préalable.

Nous distinguons deux types d'apprentissage statistique :

Apprentissage supervisé

Il vise à prédire une variable de sortie appelée variable à expliquer à partir d'un ensemble de variables d'entrée appelées variables explicatives. Ainsi, dans ce type d'apprentissage, le modèle cherche à apprendre les différents liens de cause à effet entre les variables explicatives et la variable à expliquer.

Nous considérons l'échantillon d'apprentissage E_n de taille $n \in \mathbb{N}$ suivant :

$(X_1, Y_1), ...(X_n, Y_n)$. C'est une suite de vecteurs aléatoires *iid* et d'une même loi de probabilité $\mathbb{P}$ inconnue. X_i et Y_i appartiennent respectivement aux espaces mesurables $\mathbb{X}$ et $\mathbb{Y}$.

Soit (X, Y) une nouvelle réalisation de la loi $\mathbb{P}$ non présente dans l'échantillon d'apprentissage, le but de l'apprentissage supervisé est d'apprendre dans un premier temps la loi inconnue $\mathbb{P}$ à partir de l'échantillon d'apprentissage E_n et de prédire par la suite la sortie $\hat{Y} \in \mathbb{Y}$. Elle doit être la meilleure estimation de la vraie valeur Y associée à X, ce qui revient à minimiser la différence entre la valeur prédite $\hat{Y}$ et la valeur prévue correspondante Y.

Nous distinguons deux catégories d'apprentissage supervisé. Les algorithmes de régression dans lesquels la variable à expliquer Y est quantitative. $Y \in \mathbb{Y}$ un sous ensemble de $\mathbb{R}$. Les algorithmes de classification dans lesquels la variable à expliquer Y est qualitative. Elle appartient à $\mathbb{Y}$ un ensemble fini.

Apprentissage non supervisé

L'apprentissage non supervisé est caractérisé par l'absence d'une variable à expliquer. Le modèle cherche à apprendre des structures ou des groupements intrinsèques dans la base de données. L'objectif principal est de regrouper les données dans des groupes homogènes. La méthode de clustering est la plus utilisée dans ce type d'apprentissage.

Nous nous intéressons dans ce mémoire à la catégorie d'apprentissage supervisé. Notre objectif est d'estimer une variable à expliquer (fréquence et/ou coût moyen), à partir de nos variables explicatives. Les méthodes d'apprentissage supervisé semblent donc être appropriées pour notre problématique. Nous présenterons dans cette partie le fonctionnement ainsi que les spécificités des algorithmes utilisés : l'arbre de décision CART (Classification And Regression Trees), le Random Forest et le Gradient Boosting Machine.

6.2 Algorithme des arbres de décision

Les arbres de décision font partie des méthodes d'apprentissage statistique. Ils ont été introduits à l'origine par James N. Morgan et John A. Sonquist [9] en 1963. Ils sont largement utilisés en data Science en raison de leur interprétabilité, étant donné qu'ils peuvent être représentés sous forme d'arborescence avec des décisions basées sur des règles lisibles et compréhensibles. Par ailleurs, ils peuvent s'adapter à différents problèmes de régression ou de classification.

Il existe une variété d'algorithmes d'arbres de décision et de classification parmi lesquels nous pouvons citer : CART (Classification And Regression Trees), CHAID (Chi-squared Automatic Interaction Detector), les algorithmes ID3, C4.5 et C5.0.

Les arbres de décision ont ouvert la voie à d'autres techniques d'apprentissage statistique plus avancées basées sur les arbres de décision, telles que le Random Forest et le Gradient Boosting Machine que nous aurons l'occasion de décrire en détail par la suite.

Fonctionnement

Dans le cadre de ce mémoire, nous nous concentrons sur la description du fonctionnement des arbres de décision en nous intéressant spécifiquement au cas des arbres de régression. Nous nous penchons particulièrement sur l'algorithme le plus couramment utilisé, à savoir CART. C'est une méthode d'apprentissage statistique non paramétrique et récursive, qui construit des modèles de régression en se basant exclusivement sur les données mises à disposition. Elle permet de réaliser des prédictions sans avoir besoin d'imposer de contraintes ou émettre des hypothèses sur la fonction qui relie les variables explicatives et la variable cible.

Pour construire l'arbre, l'algorithme commence par sélectionner la meilleure variable explicative X_j parmi les M variables explicatives existantes et déterminer une valeur du seuil associée s_j pour séparer les données en deux sous-ensembles $\mathcal{C}_g$ et $\mathcal{C}_d$ plus homogènes (création du nœud).

$$\mathcal{C}_g(j, s_j) = \{X_j \leq s_j\} \text{ et } \mathcal{C}_d(j, s_j) = \{X_j > s_j\}$$

Pour chaque séparation, l'algorithme maximise l'homogénéité des deux nœuds résultants. Cela équivaut à minimiser leur impureté.

Dans le cadre de la régression, la minimisation de l'impureté se traduit mathématiquement par :

$$\arg\min_{j,s_j} |\mathcal{C}_g(j; s_j)| Imp_{\mathcal{C}_g}(j; s_j) + |\mathcal{C}_d(j; s_j)| Imp_{\mathcal{C}_d}(j; s_j)$$

où $|\mathcal{C}_d(j; s_j)|$ est le cardinal de $\mathcal{C}_d$ $|\mathcal{C}_g(j; s_j)|$ est le cardinal de $\mathcal{C}_g$.

$$Imp_{\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{i:X_i \in \mathcal{C}} (Y_i - \hat{Y}_{\mathcal{C}})^2 - \frac{1}{|\mathcal{C}|} \sum_{i:X_i \in \mathcal{C}} (Y_i - \hat{Y}_{\mathcal{C}})^2$$

où $|\mathcal{C}|$ est le cardinal du nœud $\mathcal{C}$ et $\hat{Y}_{\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{i:X_i \in \mathcal{C}} Y_i$.

Le processus est répété récursivement pour chaque sous-ensemble jusqu'à atteindre l'arbre maximal. Autrement dit, le processus s'arrête lorsqu'il ait divisé les données en sous-ensembles les plus homogènes possibles. L'arbre CART continue de se développer jusqu'à ce que chaque feuille (nœud final) contienne une seule observation ou qu'aucune amélioration supplémentaire des performances ne soit obtenue.

L'arbre est ainsi construit en reliant tous les nœuds et toutes les feuilles aux valeurs de sortie. Pour prédire la valeur de la variable à expliquer pour un nouvel échantillon appartenant à une feuille donnée, nous utilisons tout simplement la moyenne des valeurs cibles de l'ensemble d'apprentissage appartenant à cette même feuille.

L'arbre obtenu est maximal avec une grande profondeur, c'est-à-dire avec de nombreuses étapes de la racine aux feuilles. Cela peut entraîner un risque significatif de surapprentissage (overfitting) où l'arbre s'ajuste parfaitement aux données d'apprentissage et perd sa capacité de généralisation sur de nouvelles données.

Afin d'éviter ce phénomène et produire un modèle plus généralisable et robuste, l'utilisation des règles d'arrêt appropriées est recommandée.

Optimisation des paramètres

Parmi ces critères d'arrêt, il est possible de limiter la profondeur maximale de l'arbre (notée **Max.depth** pour la suite) afin de trouver un juste équilibre entre deux extrêmes. D'un côté, un arbre très complexe qui risque de surapprendre les données d'apprentissage en mémorisant les spécificités de celles-ci, ce qui pourrait se traduire par une mauvaise généralisation sur de nouvelles données. D'un autre côté, un arbre trop simplifié qui pourrait sous-représenter la complexité des liens entre les variables explicatives et la variable à expliquer et donc sous apprendre les données d'apprentissage.

Trouver la profondeur de l'arbre optimale permet au modèle d'être suffisamment profond pour capter les tendances significatives des données tout en évitant d'être excessivement complexe. Par conséquent, le modèle sera généralisable et robuste.

Limites des arbres de décisions

Les arbres de régression sont très prisés pour leur facilité d'interprétation. Cependant, ils peuvent présenter une certaine instabilité. Elle est due au processus de construction des arbres de régression, où les décisions sont prises en fonction des caractéristiques des données d'apprentissage. De légères modifications dans les données peuvent entraîner des changements significatifs dans les séparations des nœuds de l'arbre et par conséquent, dans la structure de l'arbre résultant et donc dans la prédiction des valeurs.

Cette limite peut être atténuée en utilisant des techniques d'ensemble telles que le Random Forest et le Gradient Boosting Machine. Ces méthodes d'ensemble sont des approches puissantes en apprentissage statistique qui combinent les prédictions de plusieurs modèles individuels appelés "apprenants faibles afin de former un modèle global plus performant et robuste.

6.3 Algorithme du Random Forest

Fonctionnement

Le Random forest ou la forêt aléatoire [4] est un algorithme de Machine Learning introduit par Leo Breiman en 2001. C'est un cas particulier des méthodes de bagging appliquées aux arbres de décision de type CART décrits dans la section 6.2.

Le principe du fonctionnement du Random forest repose sur les étapes suivantes :

1. **La génération de plusieurs échantillons Bootstrap à partir de la base de données d'apprentissage**

   Un échantillon Bootstrap est un échantillon de même taille que la base de données d'origine mais conçu avec remplacement. Cela signifie que certaines observations peuvent être choisies plusieurs fois dans l'échantillon Bootstrap, tandis que d'autres peuvent être absentes. Ainsi, chacun des échantillons générés exclut une partie des données, qui est appelée données «Out-Of-Bag» (OOB). Autrement dit, ce sont les données qui n'ont pas été choisies pour être dans l'échantillon Bootstrap qui servira pour entraîner un arbre de décision spécifique.

2. **La construction de chaque arbre de décision en utilisant un échantillon Bootstrap différent**

   Cela permet d'introduire la variabilité dans les arbres. Cette variabilité est une composante clé des forêts aléatoires. Elle conduit à des différences dans la structure des arbres individuels, permettant ainsi d'exploiter la diversité des échantillons Bootstrap pour construire des arbres de décision variés.

3. **L'agrégation des prédictions**

   Une fois que tous les arbres sont construits, chacun d'eux fournit sa propre prédiction pour chaque observation. Les prédictions des arbres sont ensuite combinées pour obtenir la prédiction finale - la prédiction d'ensemble correspond à la moyenne des prédictions dans le cas où la variable à expliquer est quantitative - qui a une variance plus petite que celle prédite par un seul arbre CART. En tenant compte de la diversité de ces modèles individuels, le modèle final devient plus fiable et moins sensible aux fluctuations des données d'apprentissage. Cette approche permet donc d'améliorer la robustesse et la précision globale du modèle.

Une particularité importante du Random Forest est **le tirage aléatoire d'un échantillon aléatoire de variables parmi l'ensemble des variables explicatives existantes.** Cela signifie qu'au lieu de choisir la division à partir de toutes les variables, seulement un sous-ensemble aléatoire des variables explicatives est sélectionné pour la construction de chaque arbre.

Au premier abord, cette approche peut sembler contre-intuitive. En effet, lorsqu'on construit un arbre de décision classique, on cherche à diviser les données en fonction de la variable la plus importante. Cela risque d'entraîner une forte dépendance entre les arbres, car certaines variables peuvent dominer les autres à chaque division, et cela peut conduire au phénomène du surapprentissage.

Ainsi, si à chaque division seulement un sous-ensemble de variables est disponible pour être choisi, les variables moins importantes ont la possibilité d'être sélectionnées également. La taille de ce sous-ensemble est égale par défaut à $\frac{M}{3}$ dans le cadre de la régression à M variables et à $\sqrt{M}$ dans le cas de la classification. Cela permet à un très grand ensemble de variables explicatives d'être analysé. Par conséquent, les arbres obtenues seront différents les uns des autres, ce qui évite le problème de corrélation des arbres et de se retrouver avec des arbres semblables. Nous retenons par la suite la notation **Mtry** pour désigner le nombre de variables explicatives échantillonnées aléatoirement à chaque division d'un arbre de décision.

Importance des variables

À l'instar de la régression linéaire, la détermination des variables explicatives qui ont une influence sur le modèle construit est d'une importance capitale. Dans le cadre du Random Forest, une méthode existe pour évaluer la contribution de chaque variable à la construction du modèle.

L'idée derrière cette méthode est qu'une variable est considérée comme étant plus importante, lorsque les permutations aléatoires de ses valeurs dans les échantillons Out Of Bag (OOB) engendrent une forte augmentation de l'erreur de prédiction du modèle. Cette augmentation d'erreur reflète la sensibilité du modèle à cette variable, et par conséquent, sa pertinence.

La démarche de calcul de l'importance de chaque variable explicative X_j suit les étapes suivantes :

Pour chaque échantillon Bootstrap k parmi les R échantillons :

1. Détermination de son échantillon Out of Bag (OOB_k) associé (Il s'agit de l'ensemble des observations n'appartenant pas à l'échantillon Bootstrap k) ;
2. Calcul de l'erreur sur l'échantillon OOB_k ($Erreur_{OOB_k}(X_j)$) de l'arbre k construit ;
3. Permutation aléatoire des valeurs de la variable X_j dans l'échantillon OOB_k. Nous obtenons alors un nouvel échantillon dit perturbé ($\widetilde{OOB_k}$) ;
4. Calcul de l'erreur sur l'échantillon perturbé ($Erreur_{\widetilde{OOB_k}}(X_j)$).

L'importance de la variable X_j est obtenue comme suit :

$$Importance(X_j) = \frac{1}{R}\sum_{k=1}^{R} Erreur_{\widetilde{OOB_k}}(X_j) - Erreur_{OOB_k}(X_j))$$

L'erreur sur l'échantillon OOB_k est égale à :

$$Erreur_{OOB_k} = \frac{1}{|OOB_k|} \sum_{h \in OOB_k} (\hat{y}_h - y_h)^2$$

$|OOB_k|$ correspond au nombre d'observations dans l'échantillon OOB_k.

Optimisation des hyperparamètres

Nous avons introduit précédemment l'existence d'un certain nombre de paramètres (appelés hyperparamètres) qui sont utilisés dans le modèle Random Forest. Nous les récapitulons ci-après :

- Le nombre d'arbres de décision construits maximal **Num.trees** ;
- La profondeur maximale des arbres construits **Max.depth** ;
- Le nombre de variables explicatives échantillonnées aléatoirement à chaque division d'un arbre de décision **Mtry**.

En choisissant judicieusement les valeurs des hyperparamètres, nous pouvons maximiser l'efficacité de l'algorithme et par conséquent d'augmenter la qualité de prédiction du modèle.

Il existe plusieurs méthodes permettant d'optimiser ces paramètres. Nous nous contentons de présenter la méthode Grid Search par validation croisée.

Grid Search par validation croisée

La méthode **Grid search** ou bien la recherche par quadrillage consiste à définir dans un premier temps les hyperparamètres à optimiser ainsi que l'ensemble des valeurs à tester pour chacun d'eux. Ensuite, pour chaque combinaison possible des hyperparamètres spécifiés, un modèle est construit et ses performances sont évaluées par validation croisée selon une métrique choisie (RMSE par exemple). Le modèle qui obtient les meilleures performances selon cette métrique est sélectionné comme le modèle ayant les hyperparamètres optimaux

6.4 Algorithme du Gradient Boosting Machine

Fonctionnement

Le Gradient Boosting Machine (GBM) est une méthode d'ensemble qui a été développée par Jerome H Friedman en 1999 [10]. Il est fondé sur le même principe que le Random Forest : il combine plusieurs arbres de décision pour former un modèle plus performant. Cependant, le processus de construction des arbres diffère.

Dans le cas du GBM, il est effectué de manière progressive en introduisant à chaque fois un arbre supplémentaire qui vise à corriger les erreurs résiduelles des arbres précédents. En revanche, pour le modèle Random Forest, les arbres sont construits de manière indépendante.

Une autre différence entre le Random Forest et le GBM réside dans la manière dont ils combinent les résultats des arbres. Dans le Random Forest, les résultats sont agrégés à la fin du processus

(en calculant la moyenne dans le cas de la régression.) tandis que pour le GBM, les résultats sont combinés graduellement.

Il convient de noter qu'à la différence du Random Forest, le GBM cherche à prédire à chaque itération les résidus et non pas les données elles- mêmes.

Nous détaillons ci-dessous le fonctionnement de l'algorithme :

Soient le vecteur de variables explicatives $x = (x_1, x_2, ..., x_p)^t$ et y la variable à expliquer. Considérons l'ensemble d'apprentissage : $(x_1, y_1), ..., (x_n, y_n)$, c'est-à-dire : $x = (x_{i,j})_{1 \leq i \leq n, 1 \leq i \leq p}$ et $(y_i)_{1 \leq i \leq n}$ et F la fonction qui relie y à x.

L'algorithme GBM vise à trouver l'estimation optimale $\hat{F}$ de la fonction F qui minimise l'espérance d'une fonction de perte L :

$$\hat{F} = \arg\min_{F} \mathbb{E}_{x,y}[L(y, F(x))]$$

Pour ce faire, l'algorithme GBM suit la démarche décrite ci-dessous :

1. Initialisation du modèle en prenant une fonction constante :

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

2. Pour k allant de 1 à M :

   (a) Calcul du gradient négatif qui correspond aux résidus avec une fonction de perte L :

$$r_{i,k} = - \left. \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right|_{F(x)=F_{k-1}(x)}$$

   (b) Ajustement d'un arbre de décision A_k sur ces résidus

   (c) Estimation de γ_k (pas de descente) en résolvant l'équation :

$$\gamma_k = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{k-1}(x_i) + \gamma A_k(x_i))$$

   (d) Mise à jour du modèle avec :

$$F_k(x) = F_{k-1}(x) + \gamma_k A_k(x)$$

3. Itération des étapes (a), (b), (c) et (d) jusqu'à l'obtention de l'estimation

$$\hat{F}(x) = F_M(x) = \sum_{k=1}^{M} \gamma_k A_k(x)$$

Optimisation des hyperparamètres

Comme pour l'algorithme Random Forest, il est possible d'améliorer les performances des modèles construits par l'algorithme Gradient Boosting en optimisant ses hyperparamètres. Parmi ces hyperparamètres, nous pouvons citer :

- Le nombre d'arbres à ajuster **n.trees**.
- La profondeur maximale de chaque arbre construit **Interaction depth** ;
- Le taux d'apprentissage (ou coefficient de rétrécissement) **Shrinkage**

  Ce dernier paramètre régule la contribution de chaque arbre en contrôlant la vitesse avec laquelle l'algorithme va parcourir la descente du gradient. Compris entre 0 et 1, ce coefficient ν intervient dans l'étape 2.(d) de l'algorithme en substituant la formule par :

  $$F_k(x) = F_{k-1}(x) + \nu\gamma_k A_k(x)$$

  Une valeur faible ($\nu < 0.1$) améliore considérablement la qualité de prévision. En contrepartie, elle augmente le nombre d'arbres.

6.5 Interprétabilité des algorithmes de Machine Learning

De nombreux modèles complexes de Machine Learning sont généralement qualifiés de « boites noires ». Cela signifie qu'il est difficile de comprendre leur processus décisionnel étant donné qu'ils capturent des relations complexes non linéaires dans les données. Toutefois, le règlement général sur la protection des données (RGPD) met en avant les principes de « responsabilité algorithmes » et du « droit à l'explication » afin de promouvoir la transparence et la responsabilité dans l'utilisation des algorithmes.

Pour répondre à ces exigences, diverses techniques sont mises en œuvre afin de mieux apprendre le fonctionnement des modèles et de justifier leurs décisions. Dans le cadre de ce mémoire nous nous intéressons à l'utilisation des graphiques de dépendance partielle (Partial Dépendance Plots- PDP) ainsi que l'algorithme SHAP comme outils d'interprétabilité pour nos deux modèles de Machine Learning

Partial Dependence Plot (PDP)

Les graphiques de dépendance partielle (PDP) introduits par J. H. Friedman en 2001 dans l'article [11], sont considérés comme une méthode agnostique d'interprétabilité cela signifie qu'ils peuvent être appliqués à différents types de modèle de Machine Learning quel que soit leur algorithme sous-jacent. Ils permettent de visualiser l'effet marginal d'une ou de deux variables sur la prédiction finale faite par un modèle. Ils montrent comment les prédictions du modèle varient en fonction de la variable choisie, en tenant compte des autres variables à leurs valeurs moyenne.

En notant $\mathbb{X}_A$ l'ensemble des variables d'intérêt (généralement pas plus de 2 variables) et $\mathbb{X}_B$ les variables explicatives restantes. Avec $x_A \in \mathbb{X}_A$ qui représente la variable pour laquelle la fonction de dépendance partielle doit être tracée et ${x_B}^i \in X_B$ qui représente les valeurs des variables de l'observation i, hormis la variable x_A.

La fonction de dépendance partielle est définie alors comme suit :

$$\hat{f}_{\mathbb{X}_A}(x_A) = \mathbb{E}_{\mathbb{X}_B}[\hat{f}(x_A, x_B)] = \int \hat{f}(x_A, x_B)\, d\mathbb{P}(x_B)$$

Cette fonction, prise en une valeur de $\mathbb{X}_A$ représente la prédiction moyenne obtenue si l'on forçait toutes les observations à avoir cette valeur de $\mathbb{X}_A$.

Nous estimons généralement cette fonction par la méthode de Monte-Carlo en nous appuyant sur les n observations de notre échantillon :

$$\hat{f}_{\mathbb{X}_A}(x_A) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_A, {x_B}^i)$$

Individual Conditional Expectation (ICE)

Les graphiques PDP présentent l'avantage d'être facilement interprétables. Ils fournissent une visualisation intuitive de la relation entre la (les) variable(s) d'intérêt et les prédictions du modèle. Néanmoins, la représentation de l'effet moyen de la variable choisie sur les prédictions peut dissimuler les effets hétérogènes entre les variables explicatives.

Les Courbes ICE introduites par Goldstein et al. en 2017 dans l'article [12], permettent de contourner cette limite, en présentant pour chaque observation individuelle i, comment les prédictions du modèle évoluent en fonction de la variable choisie.

$$ICE_{(A,i)}(x_A) = \hat{f}(x_A, {x_B}^i)$$

La courbe PDP est obtenue en prenant la moyenne des courbes de profiles ICE. L'avantage de cette méthode réside dans sa capacité à révéler l'existence potentielle d'interaction entre la variable choisie et les autres variables lorsque les courbes de profils ICE diffèrent entre elles. Cependant, il convient de souligner que cette méthode ne permet pas d'identifier précisément quelle(s) variable(s) est/sont impliquées dans cette interaction.

Notons que les méthodes PDP/ICE reposent sur une hypothèse très forte, celle de l'indépendance des variables explicatives. Or, en pratique, il est fréquent que les variables interagissent les unes avec les autres, et leurs effets sur la variable à expliquer puissent ne pas être complètement indépendants. En outre, ces méthodes ne prennent pas en considération les distributions réelles des variables explicatives.

Algorithme SHAP

L'algorithme SHAP repose sur le calcul des valeurs de *Shapley* pour l'ensemble des variables explicatives. La notion de valeur de *Shapley* a été introduite dans le contexte de la théorie des jeux coopératifs. Son but est d'assurer une répartition équitable des gains parmi un groupe de joueurs qui ont collaboré, en prenant en compte leur contribution respective aux résultats obtenus.

Plus concrètement, les valeurs de *Shapley* sont calculées en considérant les écarts entre les valeurs prédites pour toutes les combinaisons possibles de modalités/valeurs des variables et la valeur moyenne.

L'expression générale de la valeur de *Shapley* ϕ_i est donnée par :

$$\phi_i = \sum_{S \in \mathcal{S}(\{1,\ldots,p\})_i} \frac{|S|!(p-|S|-1)!}{p!} \cdot (f_x(S \cup \{i\}) - f_x(S)),$$

où :

- i représente la ième variable ;
- p est le nombre de variables ;
- $\mathcal{S}(\{1,...,p\})$ est l'ensemble des sous-ensembles de $\{1,...,p\}$
- f_x est la fonction de prédiction en x, $f_x(S)$ est l'espérance de $f(x)$ sachant S.

La prédiction peut être décomposée en une somme des effets individuels des variables en plus d'une valeur de base ϕ_0, qui est la moyenne de toutes les prédictions :

$$f(x) = y_{pred} = \phi_0 + \sum_{i=1}^{p} \phi_i \cdot {z'}_i,$$

où :

- y_{pred} est la valeur prédite du modèle ;
- ϕ_0 est la valeur de base du modèle ;
- z'_i prend la valeur 1 lorsque la prédiction inclut la variable i et 0 sinon.

Cette expression permet d'assimiler l'effet de chaque variable sur une observation donnée. Si ϕ_i est strictement positif, la variable i augmente la prédiction, si ϕ_i est strictement négatif, la variable i diminue la prédiction, sinon ϕ_i n'a pas d'impact sur la prédiction. De plus, l'ampleur de ϕ_i indique l'intensité de l'effet de la variable i sur la prédiction. En moyennant les valeurs absolues des valeurs de *Shapley* pour chaque variable, l'importance globale des variables dans le modèle est obtenue.

Un lecteur désireux d'acquérir une compréhension plus approfondie de l'algorithme SHAP pourra se référer au mémoire [7] dédié à l'interprétabilité des modèles.

Quatrième partie

La mise en place des méthodes de tarification

Chapitre 7

La modélisation de la garantie dégâts des eaux par l'approche GLM

Dans cette partie, nous nous intéressons à la modélisation de la prime pure de la garantie dégâts des eaux du produit PNO. À cette fin, nous avons modélisé séparément la fréquence et le coût moyen en utilisant l'approche classique des modèles linéaires généralisés (GLM) dont les fondements théoriques ont été présentés dans les chapitres 4 et 5. Rappelons que la modélisation est réalisée par nature du bien (maisons et appartements).

Pour ne pas alourdir le mémoire, seuls les résultats propres aux maisons seront présentés dans le corps du mémoire. Ceux relatifs aux appartements seront disponibles en annexe. Le processus de modélisation étant totalement transposable au cas des appartements.

7.1 Choix de la loi de la fréquence et du coût moyen des sinistres

Loi de la fréquence des sinistres

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

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

Les lois Poisson et Binomiale Négative figurent parmi les lois les plus utilisées pour modéliser la fréquence en assurance. Il est important de choisir la loi qui s'ajuste aux mieux à nos données. Pour ce faire, nous avons comparé les AIC et BIC obtenus pour chacune de ces deux lois. Le tableau 7.1 récapitule les résultats obtenus :

Loi paramétrique	AIC	BIC
Poisson	198 427	200 725
Binomiale Négative	203 031	207 361

Table 7.1 – Choix de loi pour le modèle de fréquence selon les critères AIC et BIC

Les valeurs des critères AIC et BIC sont plus faibles pour la loi Poisson. Une analyse supplémentaire de la dispersion des données est entreprise. En effet, la loi Binomiale Négative est à privilégier lorsque les données sont sur-dispersées, c'est à dire lorsque $\mathbb{E}[N] < \mathbb{VAR}[N]$. Dans notre cas, nous avons une espérance de 0,01595 et une variance de 0,01591.

Nous décidons donc de retenir une modélisation de la fréquence des sinistres par la loi poisson et nous utilisons la fonction logarithmique afin d'avoir un modèle multiplicatif.

Loi du coût moyen des sinistres

Dans le modèle de coût moyen, nous cherchons à expliquer la variable suivante :

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

Les lois usuelles pour modéliser le coût moyen des sinistres sont les lois Gamma, Log-Normale et Weibull. Comme pour le modèle de fréquence, nous avons comparé les critères AIC et BIC obtenus pour chacune de ces trois lois. Le tableau 7.2 récapitule les résultats obtenus :

Loi paramétrique	AIC	BIC
Gamma	330 749	330 765
Log-Normale	333 711	333 987
Weibull	336 183	336 199

Table 7.2 – Choix de loi pour le modèle de coût moyen selon les critères AIC et BIC

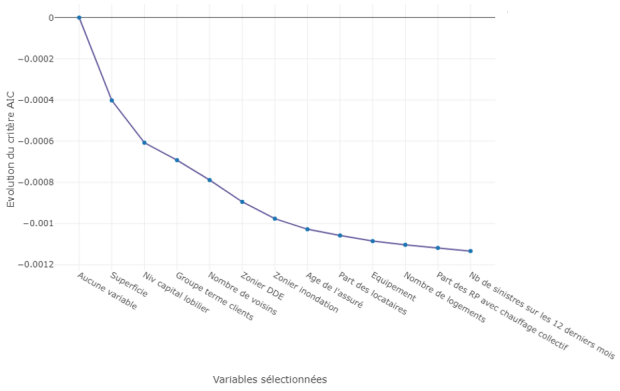
Les valeurs des critères AIC et BIC sont plus faibles quand nous supposons que le coût moyen des sinistres suit la loi Gamma. Celle-ci sera alors retenue pour notre modèle de coût moyen des sinistres.

7.2 Sélection des variables

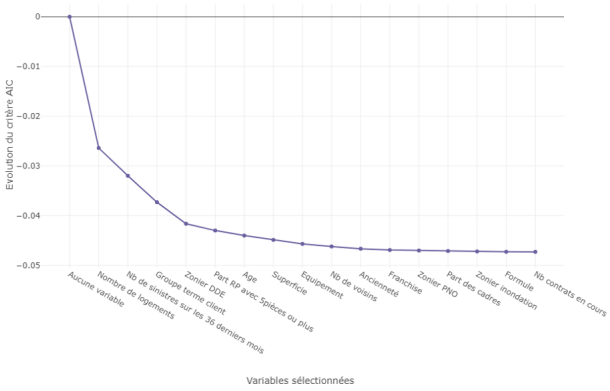
La sélection des variables est une étape cruciale de la modélisation car elle nous permet d'une part d'identifier les facteurs de risque les plus impactants et d'autre part de simplifier le modèle en ne conservant que les variables les plus significatives pour expliquer la sinistralité.

Nous avons commencé par effectuer une présélection des variables en nous basant sur les analyses descriptives, les corrélations entre les variables et l'avis d'expert. Ce faisant, 22 variables sont disponibles pour la sélection automatisée.

Par la suite, nous avons effectué une sélection de variables en utilisant une approche pas à pas (stepwise). Cette méthode nous a permis de choisir les variables les plus pertinentes pour nos modèles en éliminant progressivement celles qui n'apportent pas de contribution significative. Les graphiques 7.1a et 7.1b illustrent, respectivement, l'amélioration du critère AIC au fur et à mesure de la sélection des variables dans les modèles de coût moyen et de fréquence :



(a) Evolution du critère AIC à chaque ajout d'une variable explicative dans le modèle de coût moyen



(b) Evolution du critère AIC à chaque ajout d'une variable explicative dans le modèle de fréquence

Figure 7.1

Le tableau 7.2 récapitule les variables les plus importantes qui ont contribué à la construction des modèles de fréquence (en vert) et de coût moyen (en jaune) après la sélection stepwise selon le critère AIC.

Nous constatons que, tant pour le modèle de fréquence que pour le modèle de coût moyen, les variables sélectionnées sont de différents types (variables liées au bien, client, données à l'adresse, etc.) ce qui permet d'affiner la segmentation du risque.

Modèle Fréquence x Coût Moyen - Maisons

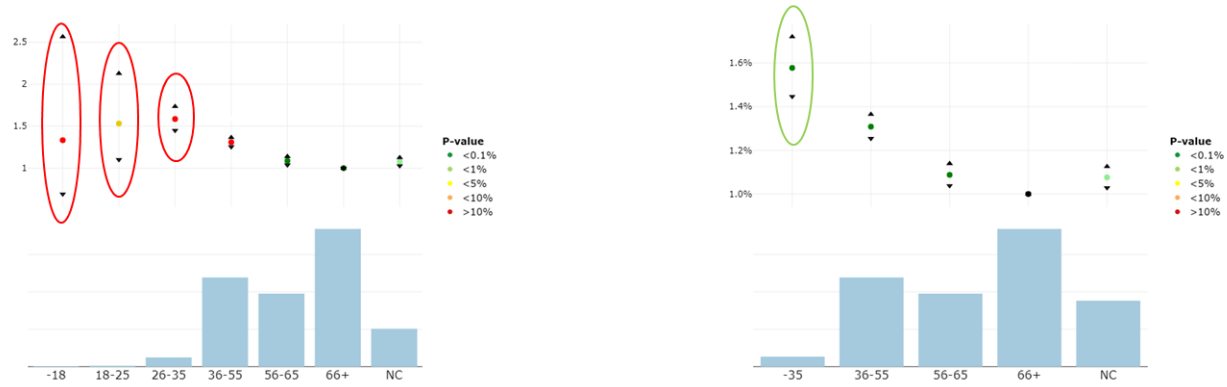
Superficie	✓ ✓	Zonier DDE	✓ ✓	Part RP avec chauffage indiv		Part RP avec 5pièces ou +	✓
Nb logements	✓ ✓	Zonier inondation	✓ ✓	Part des RP avec Chauff Coll	✓	Part ménages en couple avec enf	
Niveau capital mobilier	✓	Zonier PNO	✓	Part RP-MAISON avant1919		Part ménages en couple sans enf	
Franchise	✓	Nombre de voisins	✓ ✓	Part RP-MAISON entre 1920-1945		Part pop entre 60-74ans	
Formule	✓	Mitoyenneté	✓	Part RP-MAISON entre 1946-1970		Part pop >= 75ans	
Nb sinistres sur 12 derniers mois	✓	Dépendance non-attenante		Part RP-MAISON entre 1971-1990		Part des artisans	
Nb sinistres sur 36 derniers mois	✓	Equipement	✓	Part RP avec spfc <30m²		Part des cadres	✓
Ancienneté	✓ ✓	Nb de contrats en cours	✓	Part RP avec SUP> 120m²		Part des agriculteurs	
Age	✓ ✓			Part Résidences Secondaires		Part des inactifs	
Groupe terme client	✓ ✓			Part logements vacants		Part des chômeurs	
				Part RP en location	✓	Part des prof. Inter	

Figure 7.2 – Variables sélectionnées dans le modèle fréquence x coût moyen - Maisons

7.3 Regroupements des modalités

Afin d'améliorer le pouvoir prédictif des modèles et de travailler leur robustesse, nous avons étudié la pertinence de regrouper certaines modalités de variables. Dans notre étude, nous avons identifié trois facteurs qui peuvent être à l'origine de tels regroupements : Expositions assez faibles de certaines modalités, impact similaire de deux modalités sur la variable à expliquer ou encore à dire d'expert en se basant sur son expertise pour évaluer la pertinence des regroupements. Nous itérons cette procédure pour chaque variable. Nous vérifions ensuite à nouveau la significativité de toutes les modalités. Si certaines modalités restent non significatives, nous poursuivons le processus de regroupement en les combinant avec d'autres modalités similaires.

Prenons l'exemple de la variable explicative « âge de l'assuré » dans le modèle de fréquence :



(a) Variable explicative « Age de l'assuré » avant regroupement

(b) Variable explicative « Age de l'assuré » après regroupement

Figure 7.3

Dans le graphe 7.3a, les trois premières modalités présentent des expositions faibles qui biaisent l'évolution de la fréquence des sinistres. En regroupant ces modalités, la volatilité des coefficients a été réduite et leur évolution est conforme à l'attendu.

7.4 Significativité des variables et étude de leur qualité

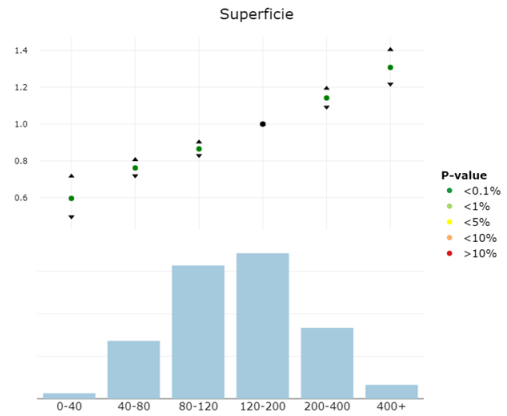
Pour chaque variable, nous présenterons à chaque fois l'évolution des coefficients obtenus à partir du modèle de fréquence et de coût moyen et les intervalles de confiance associés sur la base d'apprentissage. Dans ce cas, nous analysons l'effet unique de la variable explicative sur la variable à expliquer.

De plus, afin d'évaluer le pouvoir prédictif du modèle, une comparaison entre les valeurs observées et les valeurs prédites sur la base de test sera présentée en utilisant les coefficients estimés sur la base d'apprentissage. Dans ce cas, l'effet de la variable explicative est combiné aux effets des autres variables du modèle.

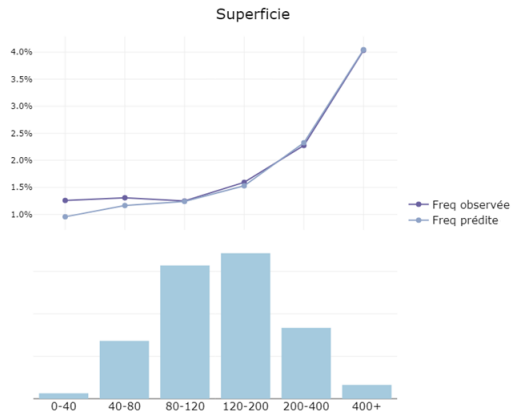
Par souci de parcimonie, nous nous limitons uniquement aux variables les plus pertinentes bien que cette analyse a été effectuée pour l'ensemble des variables sélectionnées par les modèles.

Modèle de fréquence

➤ Superficie (en m²) des maisons



(a) Coefficients et intervalles de confiance pour la variable «Superficie» - base d'apprentissage



(b) Comparaison entre la fréquence observée et la fréquence prédite pour la variable «Superficie» - base de test

FIGURE 7.4

Le graphe 7.4a montre que les coefficients liés à la variable superficie sont tous significatifs et évoluent de manière «linéaire». Nous avons une segmentation marquée entre les petites et les grandes maisons en termes de superficie. Les grandes maisons d'une superficie de 400m² et plus se voient attribuer un coefficient deux fois plus important que les petites maisons avec une superficie inférieure ou égale à 40m².

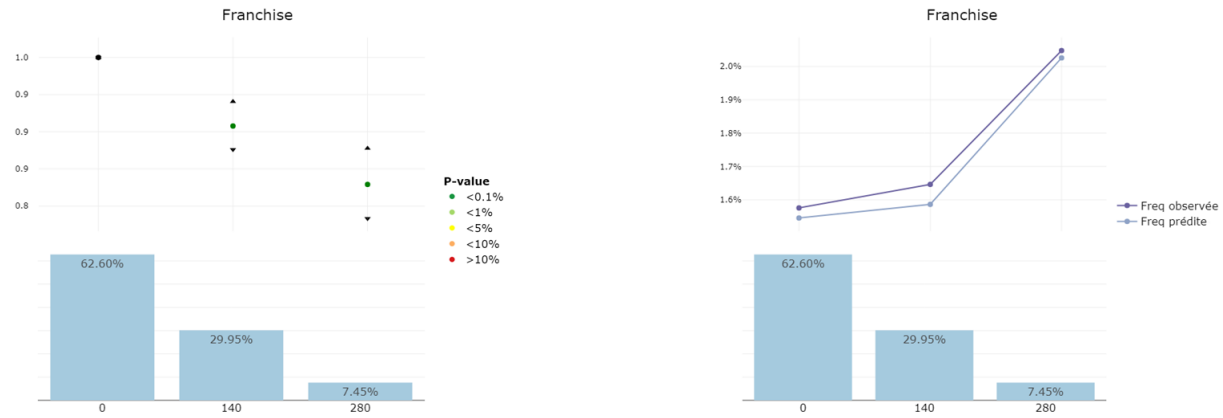
En contraste, la fréquence observée évolue de manière «exponentielle» avec la superficie. Un rapport de 4 est constaté entre les plus grandes maisons ($>$ 400m²) et les plus petites ($<$40m²) tant sur la base d'apprentissage que sur la base de test (figure 7.4b).

La différence d'allure des courbes s'expliquent par le fait que dans le modèle GLM seuls les effets individuels de la superficie sont captés, en maintenant les autres variables constantes, tandis qu'en univarié, tous les effets des variables sont combinés.

Ainsi en se référant uniquement à l'analyse univariée pour déterminer l'impact de la superficie sur la fréquence, tel qu'il est actuellement pratiqué dans le cadre des renouvellements tarifaires du produit PNO pour ajuster les coefficients au risque encouru, la variable superficie se révélerait très segmentante. Or, le modèle montre que son effet est plus modéré. Ce constat est appuyé par la contribution relativement faible de cette variable dans la diminution du critère AIC comme illustré dans la figure 7.1b.

Par ailleurs, en analysant le graphe 7.4b, nous constatons que la courbe relative à la fréquence observée et celle à la fréquence prédite sont presque confondues. Nous concluons que le modèle prédit correctement la fréquence pour la variable superficie.

➤ Franchise



(a) Coefficients et intervalles de confiance pour la variable «Franchise» - base d'apprentissage

(b) Comparaison entre la fréquence observée et la fréquence prédite pour la variable «Franchise» - base de test

Figure 7.5

Il ressort du graphique 7.5a que les coefficients de la variable « Franchise » sont tous significatifs. Leur évolution est logique étant donné qu'une absence de franchise pourrait rendre les assurés plus enclins à déclarer les sinistres aussi mineurs soient-ils. L'application d'une franchise de 280 € se traduit par une baisse du coefficient de 15% par rapport à la situation où aucune franchise n'est prévue dans le contrat d'assurance (modalité de référence). Cette évolution est inversée pour la fréquence observée (figure 7.5b) ce qui suggère que l'effet d'une ou d'autres variables semble l'emporter sur l'effet de la franchise.

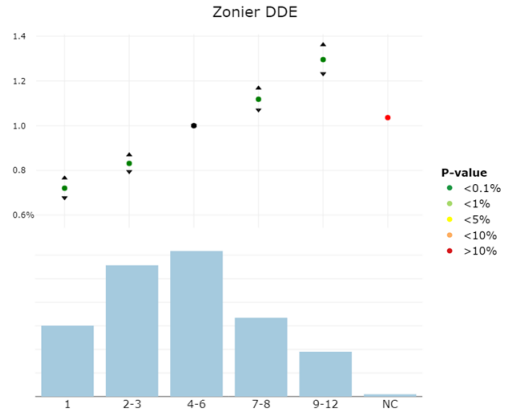
Par ailleurs, en analysant le graphe 7.5b, nous constatons que la courbe relative à la fréquence observée et celle à la fréquence prédite sont proches. Nous concluons que le modèle prédit correctement la fréquence pour la variable franchise.

➤ Zonier de la garantie dégâts des eaux

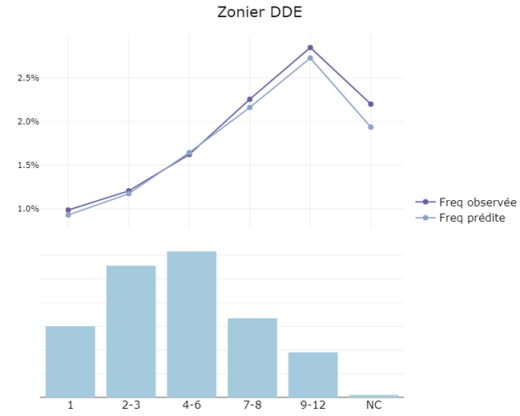
Rappelons que le zonier de la garantie dégâts des eaux permet d'effectuer un découpage de la France en 12 zones de sinistralité différentes classées de manière croissante par rapport au risque.

Le graphe 7.6a montre que les coefficients liés aux modalités de ce zonier sont tous significatifs et évoluent de manière cohérente. La zone la plus risquée (9-12) se voit attribuer un coefficient deux fois plus grand par rapport à celui de la zone la moins risquée (1).

L'évolution de la fréquence observée ne semble pas être impactée pas les effets d'autres variables. Son allure est similaire à celle du graphique 7.6a où l'impact individuel du zonier sur la fréquence est présenté. De plus, le même rapport de deux est constaté entre la zone la moins risquée et la zone la plus risquée. Ainsi, le modèle confirme la tendance univariée. Nous concluons donc que la variable zonier dégâts des eaux est discriminante. Sa sélection parmi les premières variables par la méthode stepwise vient conforter cette conclusion.



(a) Coefficients et intervalles de confiance pour la variable «Zonier DDE» - base d'apprentissage

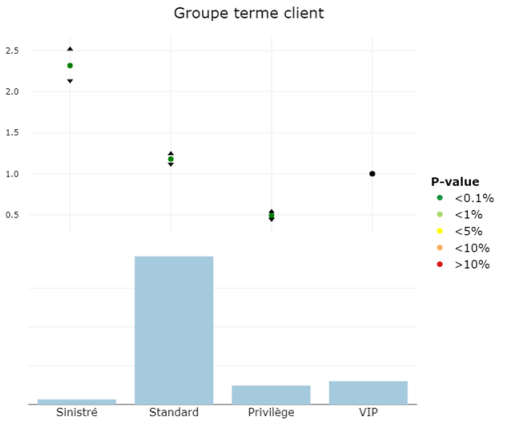


(b) Comparaison entre la fréquence observée et la fréquence prédite pour la variable «Zonier DDE» - base de test

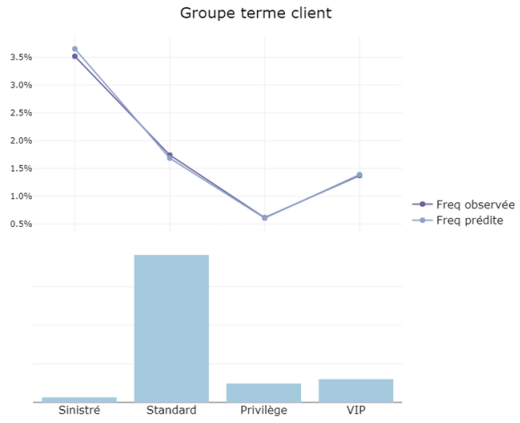
FIGURE 7.6

Par ailleurs, en analysant le graphe 7.6b, nous constatons que la fréquence prédite converge bien vers la fréquence observée.

➤ Groupe terme client



(a) Coefficients et intervalles de confiance pour la variable «Groupe terme client» - base d'apprentissage



(b) Comparaison entre la fréquence observée et la fréquence prédite pour la variable «Groupe terme client» - base de test

FIGURE 7.7

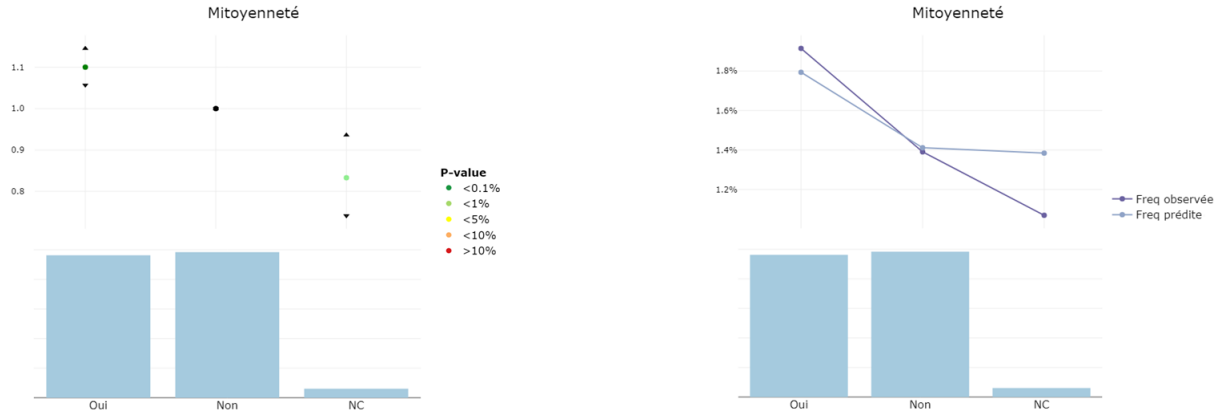
Rappelons que la variable «groupe terme client» permet d'avoir une vision du profil du client à l'échéance en tenant compte de plusieurs indicateurs liés à son équipement, le montant de ses cotisations et sa sinistralité.

Les coefficients associés aux modalités de cette variable sont tous significatifs. Le modèle retranscrit l'évolution de la fréquence observée des sinistres dégâts des eaux pour les différents profils de clients. La probabilité d'avoir un sinistre dégâts des eaux est plus élevée pour les

profils avec un historique de sinistralité important comme c'est le cas pour la classe « sinistré ». Inversement, cette probabilité diminue. C'est notamment valable pour les classes « Standard » et « privilège » qui, par construction, regroupent les clients avec une sinistralité assez faible voire inexistante.

De plus, à l'analyse du graphe 7.7b, nous constatons que la courbe relative à la fréquence prédite et celle relative à la fréquence observée sont superposées et presque confondues.

➤ **Mitoyenneté des maisons (donnée externe)**



(a) Coefficients et intervalles de confiance pour la variable «Mitoyenneté» - base d'apprentissage

(b) Comparaison entre la fréquence observée et la fréquence prédite pour la variable «Mitoyenneté» - base de test

Figure 7.8

Concernant la mitoyenneté des maisons, les coefficients sont significatifs. Nous constatons que le coefficient augmente d'un point en cas de mitoyenneté de clôture ou de mur entre deux maisons alors que la fréquence observée augmente de cinq points. Ainsi la mitoyenneté n'est pas considérée comme discriminante par le modèle.

Quant au pouvoir prédictif de cette variable, comme le montre la figure 7.8b, il est globalement satisfaisant. Notons tout de même une sous-estimation du risque pour les maisons mitoyennes. Nous décidons de garder, à dire d'expert, cette variable dans notre modèle.

$\Longrightarrow$ Après l'analyse des coefficients de chaque variable et l'étude de leur qualité sur la base de test, le modèle de fréquence apparaît comme étant globalement robuste.

Parmi les variables retenues dans le modèle de fréquence de sinistres, nous identifions les variables significatives suivantes qui quantifient le risque géographique :

- Les zoniers spécifiques aux garanties dégâts des eaux et inondation (de l'offre principal) et le zonier de l'offre PNO qui est construit sur la base de la sinistralité de toutes garanties.
- Les données INSEE : la part des résidences principales avec 5 pièces et plus et la part des cadres dans une zone géographique donnée (données fournies à la maille IRIS)

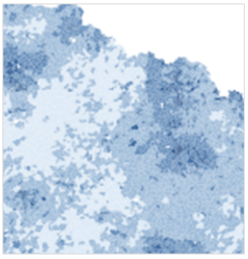
Intuitivement, nous aurions pensé que le zonier de la garantie dégâts des eaux est capable de capter toute l'information géographique étant donné que nous modélisons la fréquence des si-

nistres dégâts des eaux. Cependant, la présence de ces quatre variables en plus du zonier de la garantie dégâts des eaux est surprenante et soulève des interrogations quant à son pouvoir discriminant d'expliquer tout l'effet géographique. Afin de mesurer l'apport de ces variables, une analyse cartographique a été réalisée.

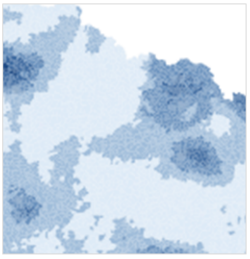
➤ Analyse cartographique

Nous avons cherché dans un premier temps à expliquer uniquement l'effet géographique sur la sinistralité. Dans cette optique, nous nous intéressons uniquement à l'impact de nos cinq variables citées précédemment sur la fréquence des sinistres. Pour ce faire, nous neutralisons l'effet des autres variables en leur attribuant la modalité de référence. Dans un second temps, les fréquences obtenues ont été projetées sur la carte de France.

En comparant la carte produite à celle du zonier de la garantie dégâts des eaux, nous constatons une forte similitude en termes de zones de risque. Pour des soucis de confidentialité, nous ne dévoilerons qu'une partie de ces cartes.



(a) Cartographie de l'effet géographique du modèle de fréquence



(b) Cartographie du zonier de la garantie dégâts des eaux

Figure 7.9

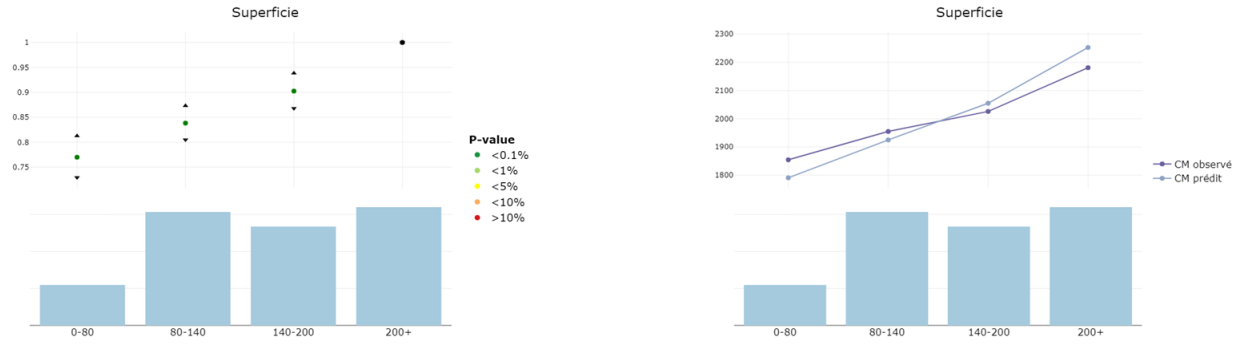
L'intensité de la couleur bleue présente dans les cartes 7.9a et 7.9b traduit le degré de risque allant d'une zone géographique à faible risque à une zone géographique à risque élevé.

Comme nous pouvons le constater sur l'extrait présenté, notre cartographie 7.9a a une allure très similaire à celle du zonier dégâts des eaux 7.9b. Nous tirons les conclusions suivantes :

- Cette forte ressemblance suggère que le zonier de la garantie dégâts des eaux explique la majeure partie de l'information géographique. Les autres variables géographiques ont donc un faible impact. Toutefois, nous décidons de combiner l'effet géographique de ces variables à celui du zonier de la garantie DDE pour expliquer l'effet géographique total.
- Ces résultats nous ont permis de nous assurer de la pertinence du zonier de la garantie dégâts des eaux comme variable tarifaire pour expliquer l'effet géographique sur la sinistralité du produit PNO.

Modèle de coût moyen

➤ Superficie (en m²) des maisons



(a) Coefficients et intervalles de confiance pour la variable «Superficie» - base d'apprentissage

(b) Comparaison entre le coût moyen observé et le coût moyen prédit pour la variable «Superficie» - base de test

Figure 7.10

Il ressort du graphique 7.10a que les coefficients liés à la variable superficie sont tous significatifs. Un rapport de 1.3 est constaté entre les plus grandes maisons ($> 200m^2$) et les plus petites ($<80m^2$) tant sur la base d'apprentissage que sur la base de test (figure 7.10b). La même évolution est constatée sur le coût moyen observé des sinistres en fonction de la superficie. Ainsi, le modèle valide la tendance univariée.

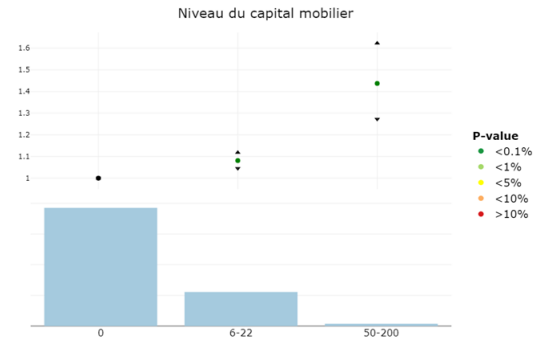
Le modèle retranscrit l'évolution du coût moyen observé des sinistres en fonction de la superficie. Le même rapport de 1.3 entre les plus grandes maisons ($> 200m^2$) et les plus petites ($<80m^2$) est constaté dans l'évolution du coût moyen observé des sinistres. Le modèle valide donc la tendance univariée.

Quant au graphe 7.10b, les prédictions du coût moyen des sinistres dégâts des eaux en fonction de la superficie sont assez proches des coûts moyens observés.

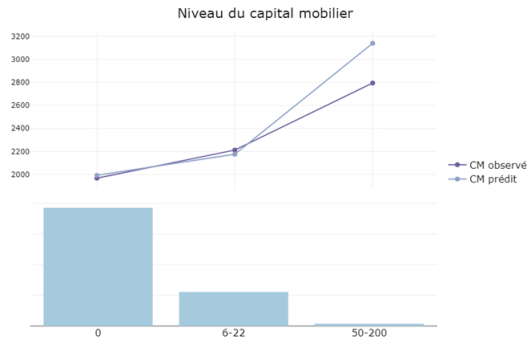
➤ Niveau du capital mobilier (en K€)

Du graphique 7.11a, nous pouvons constater que les coefficients des modalités du niveau du capital mobilier sont tous significatifs. La tendance haussière est conforme à l'attendu. Un niveau du capital mobilier assuré élevé suggère la présence d'un nombre important de biens ou d'objets de valeur. Il en découle que le montant de l'indemnisation sera plus important. Un rapport de 1.4 est observable entre un niveau de capital mobilier nul et un niveau supérieur à 50K €. Le modèle retranscrit l'évolution du coût moyen observé des sinistres en fonction du niveau de capital mobilier. Le même rapport de 1.4 entre un niveau de capital mobilier nul et un niveau de capital mobilier supérieur à 50K € est constaté dans l'évolution du coût moyen observé des sinistres. Le modèle valide donc la tendance univariée.

Du graphique 7.11b, il ressort que les prédictions du coût moyen des sinistres dégâts des eaux sont très proches des observées.



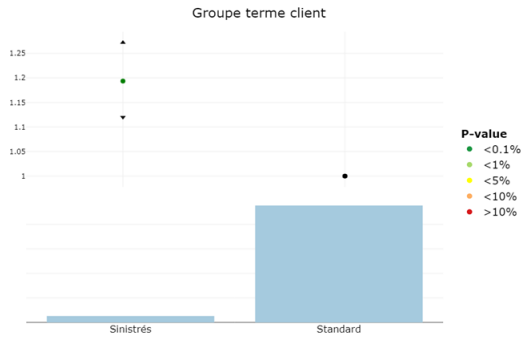
(a) Coefficients et intervalles de confiance pour la variable «Niveau du capital mobilier» - base d'apprentissage



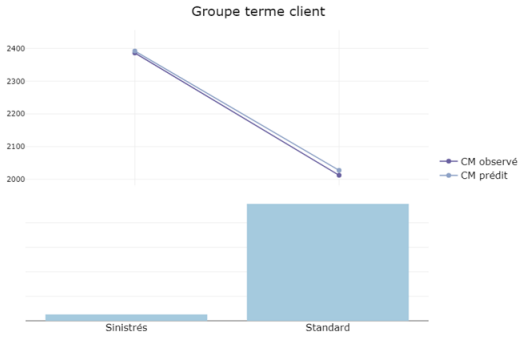
(b) Comparaison entre le coût moyen observé et le coût moyen prédit pour la variable «Niveau du capital mobilier» - base de test

Figure 7.11

➤ Groupe terme client



(a) Coefficients et intervalles de confiance pour la variable «Groupe terme client» - base d'apprentissage



(b) Comparaison entre le coût moyen observé et le coût moyen prédit pour la variable «Groupe terme client» - base de test

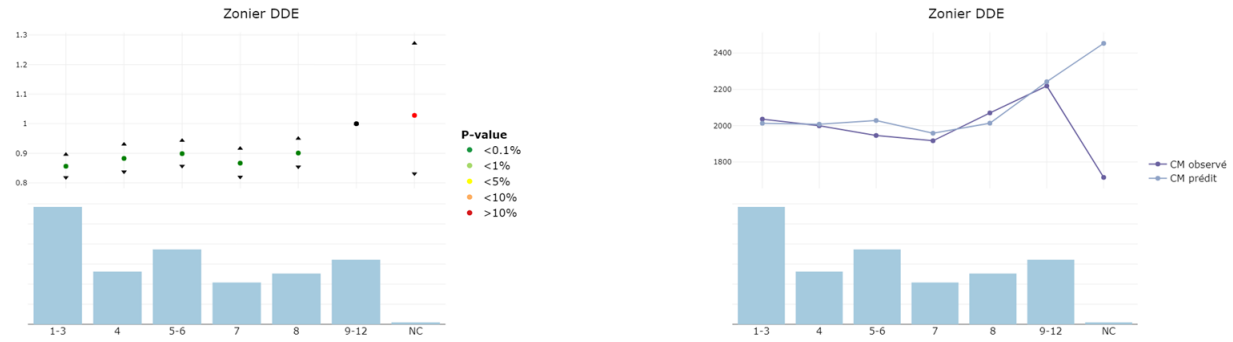
Figure 7.12

Le modèle du coût suggère de ne pas isoler les clients «VIP» et «Privilège». En effet, leurs impacts sur le coût moyen sont similaires. Ces deux classes ont été combinées avec la catégorie « Standard » qui s'oppose à la classe des clients sinistrés. Ce faisant, les coefficients de cette variable sont significatifs. Le modèle retranscrit l'évolution du coût moyen observé des sinistres en fonction du groupe terme client. Le même rapport de 1.2 entre la classe « Sinistré » et la classe « Standard » est constaté dans l'évolution du coût moyen observé des sinistres. Le modèle valide donc la tendance univariée.

Quant au graphe 7.12b, nous constatons que les courbes relatives aux coûts moyens observés et prédites se superposent et sont presque confondues. Nous concluons que le modèle prédit correctement la sévérité pour la variable «Groupe terme client».

➤ Zonier de la garantie dégâts des eaux

Compte tenu des résultats du graphique 7.13a, les coefficients liés aux douze zones du zonier de la garantie dégâts des eaux sont tous significatifs. Cependant, la segmentation est



(a) Coefficients et intervalles de confiance pour la variable «Zonier DDE» - base d'apprentissage

(b) Comparaison entre le coût moyen observé et le coût moyen prédit pour la variable «Zonier DDE» - base de test

Figure 7.13

moins évidente ce qui n'est pas très surprenant. En effet, par construction, l'effet géographique capté par le «zonier DDE» a été estimé sur la fréquence de survenance des sinistres. Ainsi il est plus discriminant pour le modèle de fréquence. Par ailleurs, le modèle retranscrit l'évolution du coût moyen observé des sinistres en fonction du zonier dégâts des eaux.

Quant au graphique 7.13b, L'adéquation est satisfaisante entre les coûts moyens prédits et observés lorsque l'exposition est significative. Les deux courbes sont proches et ont la même évolution à la hausse.

$\Longrightarrow$ Au global, en analysant l'ensemble des variables sélectionnées, le modèle est robuste. Notons tout de même que plusieurs regroupements ont été nécessaires pour améliorer la significativité des variables, ce qui requiert beaucoup de temps.

7.5 Validation des modèles fréquence x coût moyen pour la garantie dégâts des eaux

Analyse du pouvoir prédictif global des modèles

Il est également possible d'effectuer une analyse globale du pouvoir prédictif des modèles de fréquence et du coût moyen.

Une comparaison de la prédiction globale du modèle aux observations a été effectuée en séparant les bases d'apprentissage et de test en 20 échantillons correspondant aux 20 quantiles de fréquence prédite. Ces derniers sont triés de telle façon à avoir sur le premier échantillon, les 5% de la base ayant la fréquence moyenne la plus faible et au dernier échantillon, les 5% de la base avec la fréquence moyenne la plus élevée. Nous comparons ensuite les courbes des prédictions aux observations sur chacune de ces deux bases (d'apprentissage et de test).

La même démarche a été suivie pour le modèle de coût moyen.

Les graphiques 7.14 et 7.15 illustrent cette analyse sur la base de test.

La courbe relative à la fréquence prédite et celle relative à la fréquence observée sont superposées et presque confondues. Nous concluons que les prédictions du modèle de fréquence

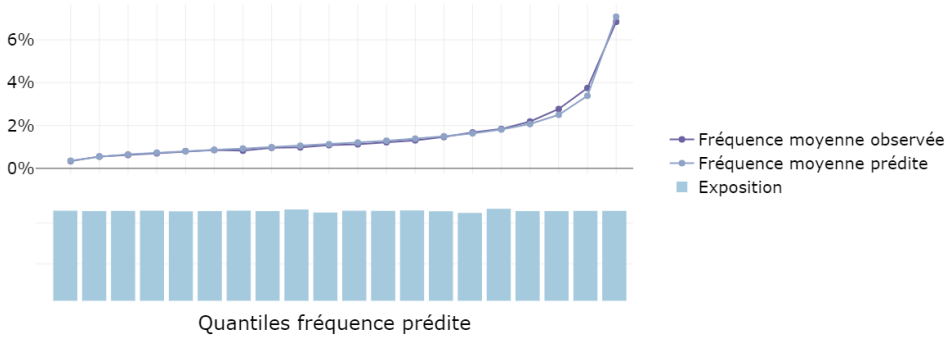


FIGURE 7.14 – Pouvoir prédictif du modèle de fréquence - base de test

sont conformes à la réalité.

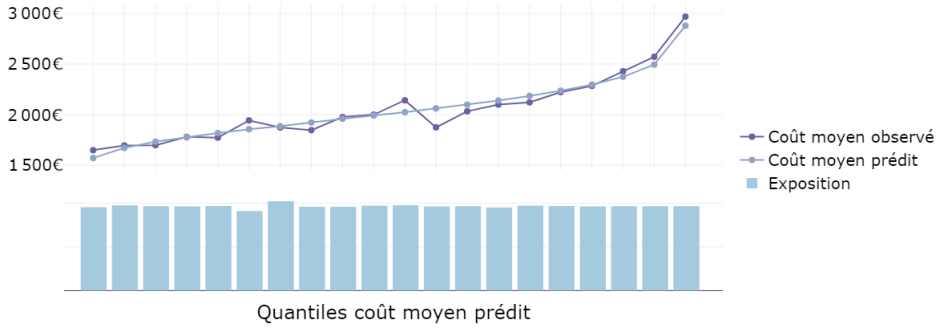


FIGURE 7.15 – Pouvoir prédictif du modèle de coût moyen - base de test

La courbe relative au coût moyen prédit et celle relative au coût moyen observé sont proches. Nous concluons que les prédictions du modèle de coût moyen sont globalement conformes à la réalité.

La fréquence moyenne prédite est très proche de la fréquence observée. Ce constat est également vrai pour le coût moyen. Le tableau 7.16 présente l'évaluation de la qualité de prédiction de notre modèle selon le critère de la racine de l'erreur quadratique moyenne (RMSE).

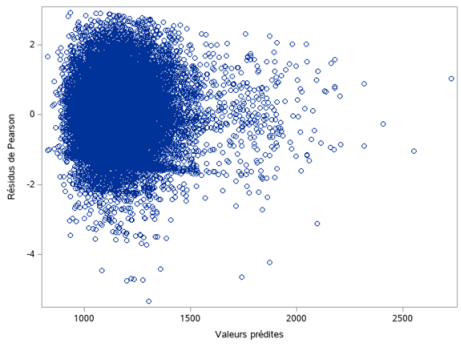
	Base d'apprentissage		**Base de test**	
	Modèle de fréquence	**Modèle de coût moyen**	**Modèle de fréquence**	**Modèle de coût moyen**
RMSE	9,33%	2 626,21 €	9,44%	2 648,10 €

FIGURE 7.16 – Evaluation de la qualité de prédiction du modèle Fréquence x Coût moyen selon l'indicateur RMSE - Maisons

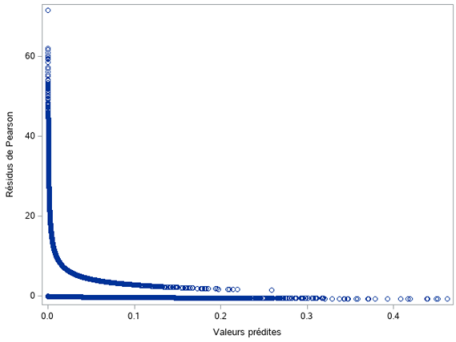
Après avoir effectué ces différents tests de validation, nous concluons que notre modèle de fréquence x coût moyen est globalement stable et semble assez robuste pour prédire les données.

Analyse des résidus

Une fois les modèles de fréquences et de coût moyen construits, il convient de s'assurer du caractère aléatoire des résidus du modèle. Ainsi nous allons les analyser en fonction des prédictions de chaque observation.



(a) Résidus standardisés de Pearson du modèle du coût moyen



(b) Résidus standardisés de Pearson du modèle de fréquence

Figure 7.17

Le graphe 7.17a montre que les résidus du modèle du coût moyen ont une répartition aléatoire autour du zéro et leur répartition est indépendante des valeurs prédites. Nous pouvons conclure que l'hypothèse de linéarité sur laquelle se base notre modèle est vérifiée.

Dans le graphe 7.17b, nous observons deux nuages de points distincts. Bien que cette allure des résidus paraisse atypique au premier abord, elle demeure classique pour le modèle de fréquence. Nous pouvons déduire que l'ajustement du modèle de fréquence est pertinent.

Conclusion intermédiaire

L'étape de la validation des modèles de fréquence et de coût moyen a mis en évidence le fort pouvoir prédictif du modèle de fréquence. **Ainsi, le modèle GLM est retenu pour la modélisation de la fréquence des sinistres.**
En revanche, le modèle de coût moyen semble moins performant. Il se généralise moins bien sur la base de test. Cela peut être dû à la volumétrie réduite de la base de données utilisée pour la modélisation du coût moyen étant donné que le produit PNO manque de volume d'affaires comparé à l'offre principale. De plus, nous nous limitons aux sinistres de charges strictement positives.
Dans le prochain chapitre, nous allons évaluer la pertinence d'investir dans les méthodes de Machine Learning pour améliorer la performance du modèle du coût moyen des sinistres.

Chapitre 8

La modélisation de la garantie dégâts des eaux par des méthodes alternatives

8.1 Modélisation par l'algorithme Random Forest

Modèle sans hyperparamétrage

Nous avons lancé un premier modèle de coût moyen par l'algorithme Random Forest avec les hyperparamètres par défaut suivants :

- **Num.trees** : Le nombre d'arbres construits est fixé à 500 arbres,
- **Mtry** : Le nombre de variables choisi aléatoirement à chaque division d'un arbre est fixé à 14 variables. Cela correspond au paramétrage par défaut en cas de régression (p/3 avec p le nombre de variables explicatives),
- **Max.depth** : La profondeur maximale de chaque arbre est illimitée (i.e. les nœuds de chaque arbre sont étendus jusqu'à ce que toutes les feuilles deviennent homogènes ou bien jusqu'à ce que le nombre d'observations dans chaque feuille soit inférieur à un seuil minimum prédéfini).

Afin d'éviter le sur-apprentissage ou le sous-apprentissage du modèle sur les données d'entraînement et améliorer ses performances, il est judicieux de procéder à l'optimisation des hyperparamètres cités précédemment.

Optimisation des hyperparamètres

Num.trees optimal

Il est recommandé d'optimiser le nombre d'arbres construits. Plus le nombre d'arbres est élevé, plus les performances du modèle sont améliorées. Cependant, cela n'est pas sans conséquences sur le temps de calcul. Ainsi, pour définir le nombre d'arbres optimal, nous avons fait varier le nombre d'arbres dans un intervalle allant de 1 à 1000 et nous avons déterminé le nombre d'arbres minimal qui permet de stabiliser la RMSE.

Le graphique 8.1 montre une stabilisation de l'erreur au-delà de 400. Nous décidons alors de fixer le nombre d'arbres construits par le modèle à 500.

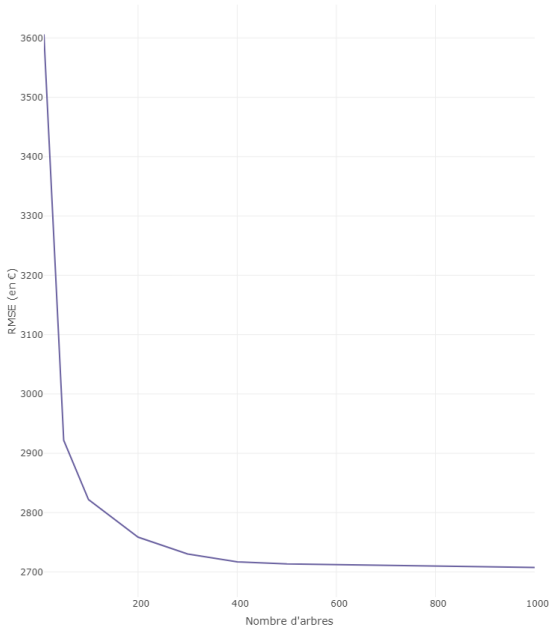


Figure 8.1 – Evolution de la RMSE en fonction du nombre d'arbres pour le modèle de coût moyen par l'algorithme Random Forest

Mtry et Max.depth optimaux

Mtry et Max.depth ont été optimisés par la méthode Grid Search dont le principe est explicité dans la section 6.3. Le tableau 8.1 présente les valeurs testées et retenues pour ces hyperparamètres.

Paramètre	Valeurs testées	Valeur retenue
Mtry	[1,2,....,18]	7
Max.depth	[1,2,....,15]	10

Table 8.1 – Optimisation des hyperparamètres du modèle de coût moyen par l'algorithme Random Forest

Comme le montre la figure 8.2, cette étape d'hyperparamétrage nous a permis, en nous basant sur le critère RMSE, d'améliorer la performance du modèle.

	Modèle non-optimisé	Modèle optimisé
RMSE	2 702,18 €	2 660,21 €

Figure 8.2 – Comparaison sur la base de test de la performance des modèles de coût moyen par l'algorithme Random Forest avant et après hyperparamétrage selon le critère RMSE

Importance des variables

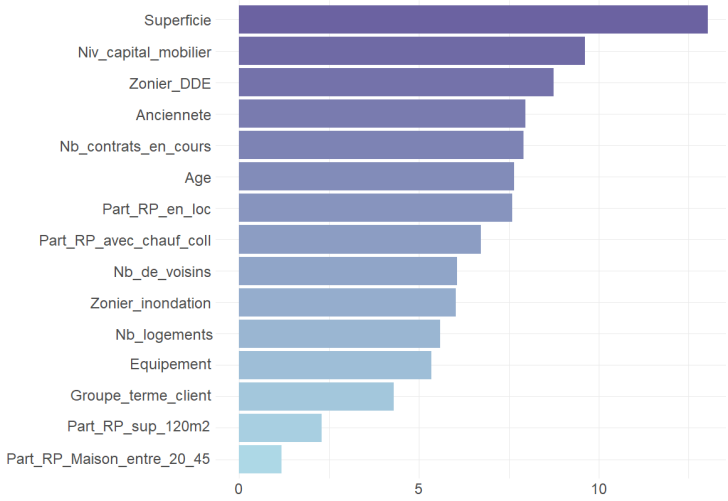


FIGURE 8.3 – Importance des variables dans le modèle de coût moyen des maisons par l'algorithme Random Forest

Le graphique 8.3 représente les 15 variables les plus importantes qui ont contribué à la construction du modèle Random Forest. L'importance des variables permet d'identifier les variables les plus utilisées dans l'algorithme. Néanmoins, elle ne permettra pas de quantifier l'impact d'une variable sur la prédiction du coût moyen des sinistres.

Il en ressort de l'analyse de l'importance des variables, que la variable superficie est la variable la plus discriminante pour la construction du modèle Random Forest. En effet, la superficie influe sur le montant indemnisé par l'assureur. Plus la superficie du bien est importante plus l'étendue des dégâts pourrait être considérable. Dès lors, le coût de sinistre serait éventuellement conséquent. Cette variable est également retenue dans le modèle GLM du coût moyen.

S'ensuit le niveau du capital mobilier, le zonier de la garantie dégâts des eaux, l'ancienneté du contrat, l'âge de l'assuré et également les variables INSEE à savoir la part des résidences principales ayant un chauffage collectif ou encore la part des résidences principales en location. Ces variables sont aussi présentes dans le GLM du coût moyen. En somme, parmi les 15 variables importantes dans la construction du Random Forest, 11 variables retenues par le GLM du coût moyen ont été jugées comme importantes dans le modèle Random Forest.

D'autres variables sont considérées comme importantes mais qui ne sont pas retenues par le GLM du coût moyen. Elles pourraient :

- compléter les informations présentes dans le GLM. C'est le cas notamment de la part des résidences principales dont la superficie est supérieure à 120 m^2. Une corrélation linéaire de 47% est constatée avec la part des résidences principales en location,

- ou apporter de nouvelles informations tarifaires non captées par le GLM. C'est le cas de la variable équipement qui reflète le statut de l'assuré s'il est mono ou multi-équipé et donc fournit une information supplémentaire sur le profil du client ou encore la part des maisons construites entre 1920 et 1945 qui livrent une information sur la période de construction du logement.

8.2 Modélisation par l'algorithme Gradient Boosting

Modèle sans hyperparamétrage

Nous avons lancé un premier modèle de coût moyen par l'algorithme Gradient Boosting avec les hyperparamètres par défaut suivants :

- **N.trees** : Le nombre d'arbres construits est fixé à 100 arbres,
- **Shrinkage** : Le taux d'apprentissage est fixé à 0.1,
- **Interaction.depth** : La profondeur des interactions entre les variables est fixée à 1.

De manière similaire au modèle Random Forest, il est judicieux de procéder à l'optimisation de ces hyperparamètres.

Optimisation des hyperparamères

Nous avons eu recours à la méthode grid search pour optimiser les hyperparamètres du modèle. Dans le tableau 8.2, nous représentons les valeurs testées ainsi que la combinaison optimale retenue qui minimise la RMSE.

Paramètre	Valeurs testées	Valeur retenue
N.trees	[50,100,....,1 000]	400
Shrinkage	[0.01,0.02,....,0.05]	0.03
Interaction.depth	[2,3,....,10]	7

Table 8.2 – Optimisation des hyperparamètres du modèle de coût moyen par l'algorithme Gradient Boosting

Comme le montre la figure 8.4, cette étape d'hyperparamétrage nous a permis, en nous basant sur le critère RMSE, d'améliorer la performance du modèle.

	Modèle non-optimisé	Modèle optimisé
RMSE	2 673,91 €	2 662,42 €

Figure 8.4 – Comparaison sur la base de test de la performance des modèles du modèle de coût moyen par l'algorithme Gradient Boosting avant et après hyperparamétrage selon le critère RMSE

Importance des variables

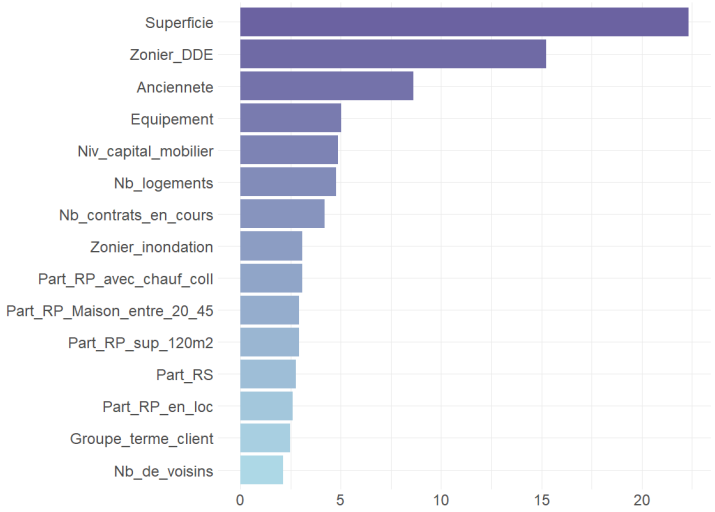


Figure 8.5 – Importance des variables dans le modèle de coût moyen des maisons par l'algorithme Gradient Boosting

Le graphique 8.5 représente les 15 variables les plus importantes qui ont contribué à la construction du modèle Gradient Boosting.

Nous retrouvons à nouveau la superficie comme la première variable la plus influente dans la construction du modèle Random Forest. Ce résultat conforte les constats faits précédemment. La variable « Superficie » est également considérée comme importante dans la construction du modèle Gradient Boosting. Au total, parmi les 15 variables les plus importantes pour le Gradient Boosting, 10 sont retenues par le GLM.

Les deux modèles Random Forest et Gradient Boosting possèdent globalement les mêmes variables importantes, En effet, parmi les 15 variables les plus influentes, les deux modèles ont 14 variables en commun. Toutefois, l'ordre de ces dernières diffère. Ces constats mettent à l'évidence que les 3 modèles retiennent pratiquement les mêmes variables explicatives.

8.3 Interprétabilité des modèles alternatifs

Cette partie s'inscrit dans la continuité de ce qui a été conduit précédemment à savoir l'analyse de l'importance des variables qui nous a permis d'identifier les variables les plus importantes qui ont contribué dans la construction des modèles. Cependant elle ne nous permet pas de savoir le sens d'influence de la variable. On ignore comment la variable explicative impacte la variable à expliquer.

À cet effet, nous nous intéressons dans cette partie à compléter cette analyse par l'étude de l'impact des variables tarifaires sur la prédiction du coût moyen. Nous voulons vérifier la pertinence et la cohérence des prédictions des variables dans les modèles.

Pour ce faire, nous avons eu recours, dans un premier temps, aux méthodes d'interprétabilité **Partial Dependance Plot (PDP)** et **Individual Conditional Expectation (ICE)** dont les principes sont présentés dans la section 6.5.

Notons qu'il est important d'interpréter ces graphiques avec prudence étant donné que l'effet marginal de certaines variables -fortement corrélées à d'autres- risque d'être biaisé.

Pour ne pas surcharger le mémoire, nous nous limitons à la présentation des résultats des deux variables les plus importantes.

Graphiques PDP/ICE

Interprétabilité du modèle de coût moyen par l'algorithme Random Forest

➤ Variable «Superficie» des maisons (en m²)

L'analyse du graphique PDP (figure 8.6) de la variable «Superficie», pour le modèle Random Forest montre que la courbe du coût moyen prédit présente une évolution haussière continue en fonction de la superficie, ce qui est conforme à l'attendu. De surcroit, cette tendance à la hausse est cohérente avec la modèle de coût moyen en GLM.

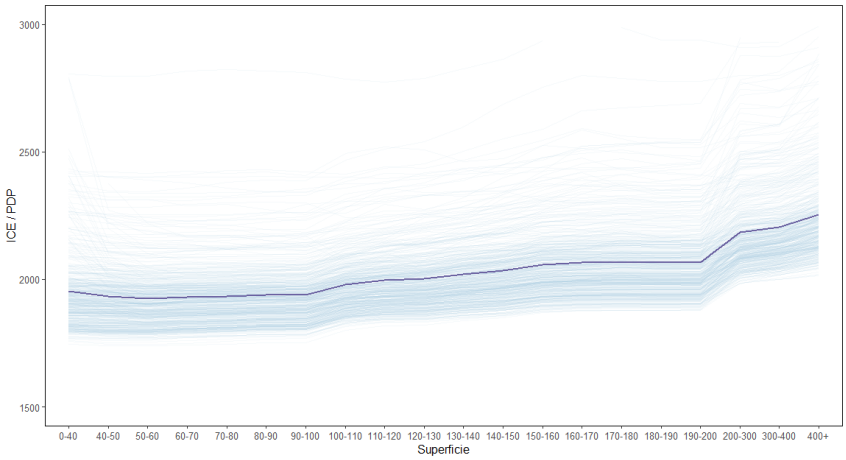


Figure 8.6 – Graphique de PDP et courbes ICE obtenus pour le modèle Random Forest associé à la variable «Superficie»

Quant aux courbes des profils ICE, nous constatons que dans la majorité des cas, elles présentent une tendance haussière avec la superficie et sont translatées des autres courbes. Elles sont également translatées du graphique PDP qui est une moyenne de toutes les courbes ICE. Cependant, il convient de remarquer la présence de certaines variations dans les formes des représentations graphiques des profils ICE qui diffèrent de la tendance haussière globale, suggérant l'existence d'une éventuelle interaction entre la superficie et une ou plusieurs variables explicatives. Néanmoins, le graphique ICE ne nous permet pas de déterminer avec quelle(s) variable(s) l'interaction se produit.

➤ Variable «Niveau du capital mobilier» (en K€)

L'analyse du graphique PDP (figure 8.7) de la deuxième variable la plus influente dans la construction du Random Forest révèle que la tendance haussière est bien observée sur le coût moyen prédit : Au fur et à mesure que le niveau de capital mobilier augmente, le coût de sinistre augmente de manière concomitante. La même tendance haussière et linéaire est constatée dans le modèle GLM.

Quant aux différentes courbes des profils ICE, dans l'ensemble, elles ont la même tendance

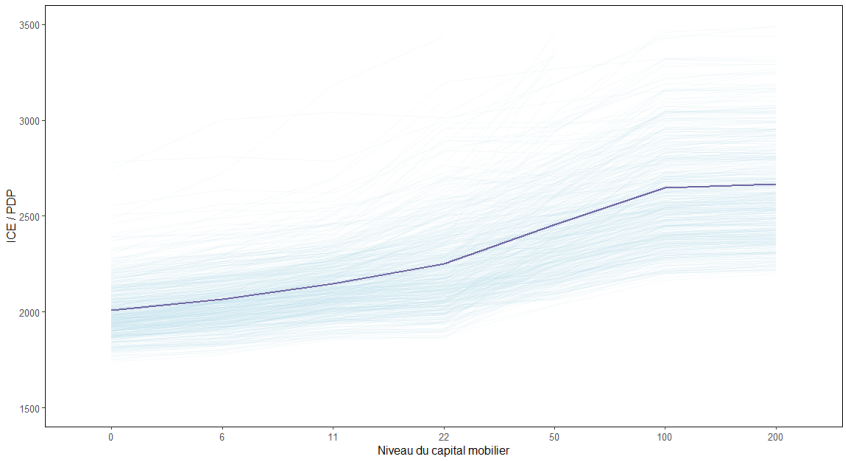


Figure 8.7 – Graphique de PDP et courbes ICE obtenus pour le modèle Random Forest associé à la variable «Niveau du capital mobilier»

haussière. Toutefois, il convient de remarquer la présence de certaines courbes non translatées. Cela peut être un signe d'hétérogénéité dans les effets de la variable sur les prédictions des coûts moyens.

Interprétabilité du modèle de coût moyen par l'algorithme Gradient Boosting

➤ Variable «Superficie» des maisons (en m²)

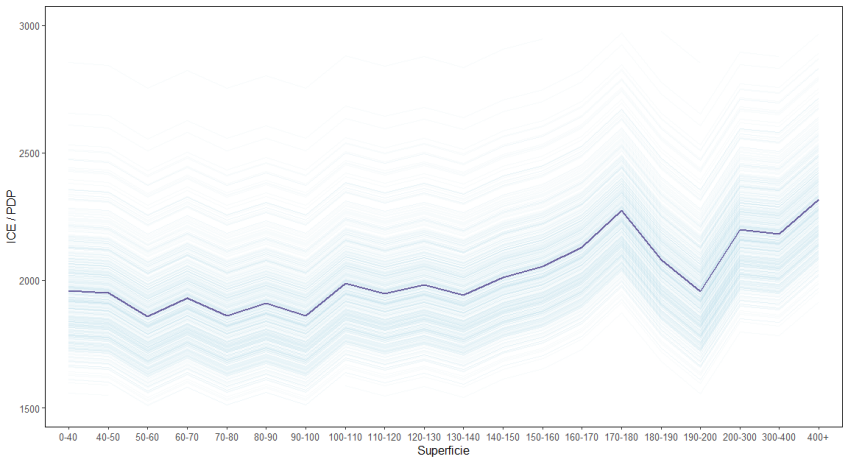


Figure 8.8 – Graphique de PDP et courbes ICE obtenus pour le modèle Gradient Boosting associé à la variable «Superficie»

L'analyse des graphique PDP/ICE (figure 8.8) de la variable «Superficie» pour le modèle Gradient Boosting montre Les courbe PDP et ICE présentent une tendance haussière du coût moyen en fonction de la superficie, très proche de celle observée en univarié (cf. Analyse univariée de la variable «Superficie», figure 3.5). Cela sème des doutes quant à un possible surapprentissage du modèle Gradient Boosting. Par ailleurs, cette tendance haussière est moins lissée comparée à celle du modèle Random Forest.

➤ Variable «Zonier dégâts des eaux»

L'analyse du graphique PDP (figure 8.9) permet d'observer que la logique de hausse de la

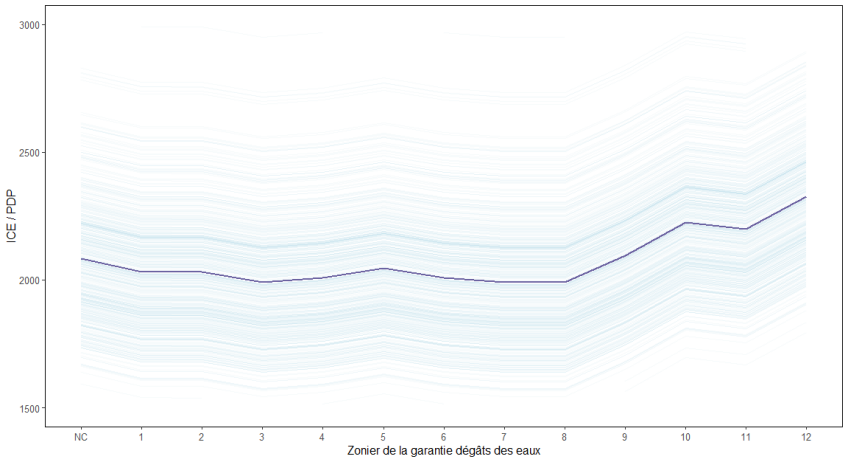


Figure 8.9 – Graphique de PDP et courbes ICE obtenus pour le modèle Gradient Boosting associé à la variable «Zonier DDE»

sinistralité avec l'augmentation du risque géographique est vérifiée. Cette tendance est également vérifiée avec celle du coût moyen prédit par le modèle GLM. Toutefois, la segmentation est moins marquée entre les zones « 1 » à « 8 ».

Quant aux courbes ICE, elles présentent globalement la tendance haussière attendue. Toutes les courbes sont translatées entre elles et sont également translatées de la courbe PDP.

Nous pouvons également utiliser les valeurs de *Shapley* pour interpréter les deux modèles Random Forest et Gradient Boosting. Cette méthode présente l'avantage d'être basée sur une théorie mathématique solide.

SHAP

Le **SHAP Summary Plot** combine l'importance des variables avec leurs effets. Afin de faciliter l'analyse de ces graphiques, il convient de noter :

- L'axe des ordonnées représente les variables explicatives classées selon leur degré d'importance. [1]
- L'axe des abscisses représente les valeurs SHAP permettant d'indiquer le sens et la force de variation des prédictions du coût moyen du modèle. Une valeur SHAP nulle se trouve à l'origine est correspond à la moyenne des prédictions du coût moyen.
- Pour chaque variable explicative, un ensemble de points est représenté sur le graphique. La position de chaque point par rapport à l'origine reflète l'écart entre le coût moyen prédit et la moyenne des prédictions du coût moyen.
- La couleur du gradient dépend des valeurs des variables explicatives (les fortes valeurs sont en rouge et les faibles valeurs sont en bleu).

1. L'importance d'une variable calculée au sens des valeurs SHAP correspond à la somme des valeurs absolues de toutes ses valeurs de Shapley sur chaque prédiction. Ce calcul diffère de celui présenté dans les graphiques 8.3 et 8.5 qui se base plutôt sur la diminution des performances du modèle.

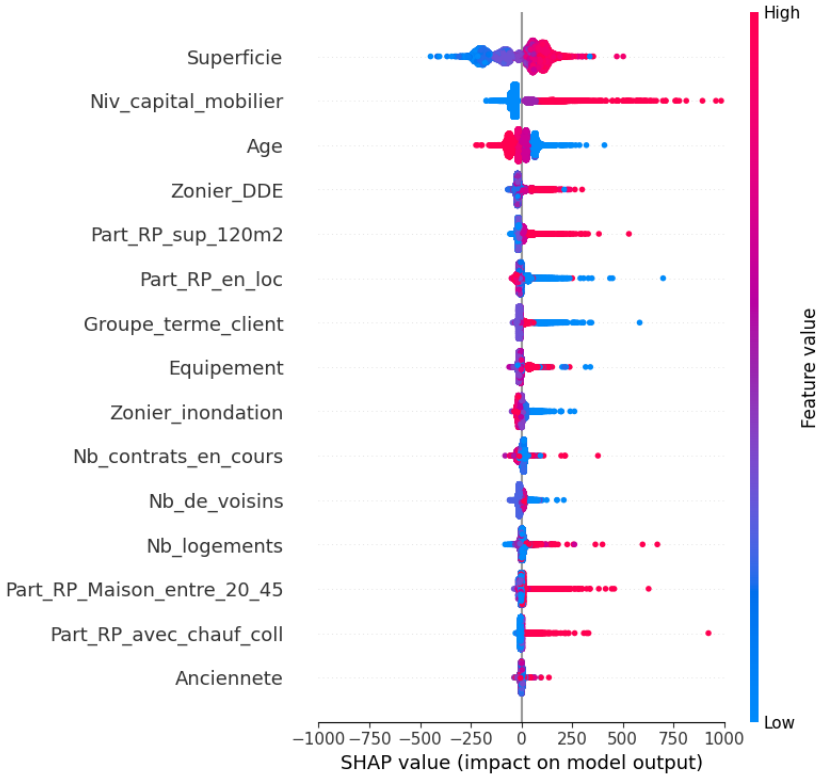


Figure 8.10 – SHAP summary plot pour le modèle Random Forest

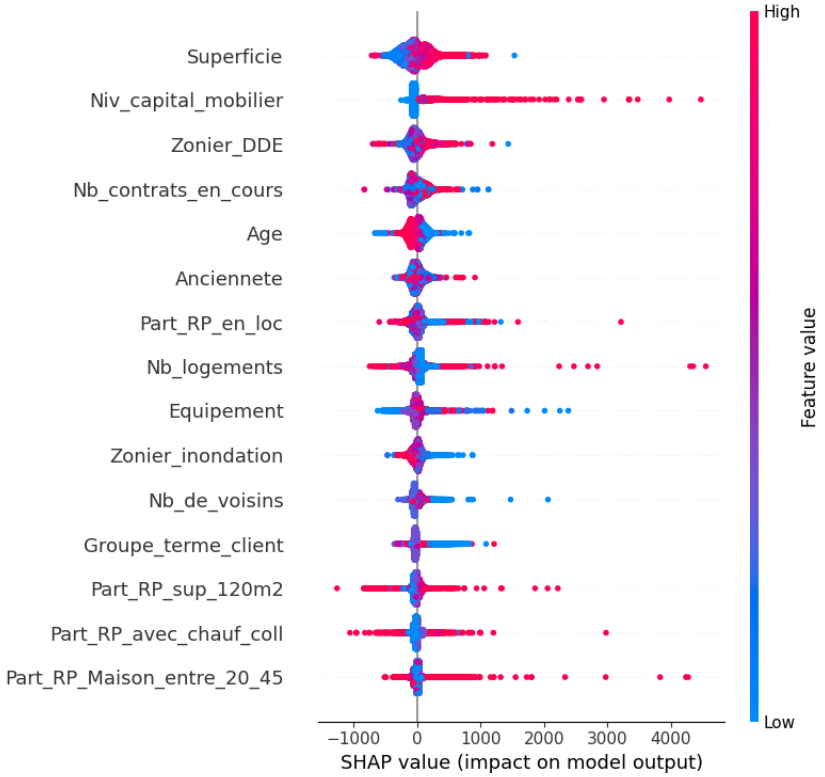


Figure 8.11 – SHAP summary plot pour le modèle Gradient Boosting

Dans les graphiques summary plot sont présentées les 15 variables les plus importantes. Pour les analyses nous nous focalisons sur les 3 premières variables plus influentes.

➤ **Analyse de l'impact de la superficie sur le coût moyen**

Tant le modèle Random forest que le modèle Gradient boosting convergent pour considérer la superficie comme la variable ayant contribué le plus à la prédiction du coût moyen des sinistres. Les grandes superficies (en rouge) ont des valeurs SHAP positives, on parle d'impact positif sur le coût moyen des sinistres. Cela signifie que le coût moyen augmente quand les superficies sont grandes. Inversement, les petites superficies (en bleu) sont associées à des valeurs SHAP négatives. On parle d'un impact négatif, ce qui implique une diminution du coût moyen.

➤ **Analyse de l'impact du niveau du capital mobilier sur le coût moyen**

Les modèles Random Forest et Gradient Boosting s'accordent pour considérer le niveau du capital mobilier comme la deuxième variable importante en termes de contribution à la prédiction du coût moyen des sinistres.

Un niveau du capital mobilier élevé a un fort impact positif sur le coût moyen des sinistres dégâts des eaux. À contrario, les valeurs faibles du capital mobilier entraînent une diminution du coût moyen des sinistres.

➤ **Analyse de l'impact de l'âge de l'assuré sur le coût moyen**

L'âge de l'assuré est la troisième variable la plus importante du modèle Random Forest.

Les assurés moins âgés ont un impact positif sur le coût moyen des sinistres. Ainsi, il apparaît que les jeunes propriétaires non occupants ont tendance à engendrer des charges de sinistres plus importantes. Une explication plausible pour expliquer ce résultat est que ces assurés manquent d'expérience dans la gestion et l'entretien des maisons qu'ils mettent en location. Inversement, les assurés plus âgés ont un impact négatif sur le coût des sinistres.

L'âge de l'assuré est la cinquième variable la plus importante dans le classement du modèle Gradient Boosting. Globalement, les analyses réalisées pour le modèle Random Forest demeurent valables pour le modèle Gradient Boosting.

➤ **Analyse de l'impact du zonier de la garantie dégâts des eaux sur le coût moyen**

Le zonier de la garantie dégâts des eaux est la quatrième (resp. troisième) variable la plus importante dans le modèle Random Forest (resp. Gradient Boosting).

Pour les deux modèles, les zones à risque élevé du zonier dégâts des eaux ont un impact positif sur le coût moyen des sinistres. Ceci indique que les maisons qui se trouvent dans des zones de risque élevé sont associées à des coûts moyens plus conséquents.

En revanche, nous pouvons remarquer que pour les zones de risque moins élevé, les points de différentes couleurs se chevauchent autour de l'origine, ce qui suggère une absence d'impact significatif sur les prédictions du coût moyen.

Chapitre 9

La comparaison des performances des modèles

9.1 Comparaison des variables retenues par les modèles

Modèle GLM	Modèle Random Forest	Modèle Gradient Boosting
Superficie	Superficie	Superficie
Niveau du capital mobilier	Niveau du capital mobilier	Zonier de la garantie dégâts des eaux
Zonier de la garantie dégâts des eaux	Zonier de la garantie dégâts des eaux	Ancienneté du contrat
Groupe terme client	Ancienneté du contrat	Equipement
Nombre de voisins	Nombre de contrats en cours	Niveau du capital mobilier
Zonier de la garantie inondation	Age de l'assuré	Nombre de logements
Age de l'assuré	Part des résidences principales en location	Nombre de contrats en cours
Part des résidences principales en location	Part des résidences principales avec un chauffage collectif	Zonier de la garantie inondation
Part des résidences principales avec un chauffage collectif	Nombre de voisins	Part des résidences principales avec un chauffage collectif
Ancienneté du contrat	Zonier de la garantie inondation	Part des maisons en résidences principales construites entre 1920-1945
Nombre de logements	Nombre de logements	Part des maisons en résidences principales avec une superficie > 120m²
Sinistralité antérieure sur 12 mois	Equipement	Part des résidences secondaires
	Groupe terme client	Part des résidences principales en location
	Part des maisons en résidences principales avec une superficie > 120m²	Groupe terme client
	Part des maisons en résidences principales construites entre 1920-1945	Nombre de voisins

Variables retenues uniquement par le modèle GLM

Variables non retenues par le modèle GLM

Variables retenues* uniquement par le modèle Gradient Boosting

Variables non retenues* par le modèle Gradient Boosting

** Pour les méthodes de Machine Learning, le terme "retenues" est un abus de langage pour dire "figurent parmi les 15 variables les plus importantes"*

Figure 9.1 – Comparaison des variables retenues par les modèles de coût moyen - Maisons

Les variables retenues par les différents modèles construits sont récapitulées dans le tableau 9.1. Parmi les douze variables sélectionnées dans le modèle GLM, dix figurent dans le classement des quinze variables les plus importantes dans la construction des modèles de Machine Learning. Nous retrouvons des variables provenant de différents types.

- Des variables liées au bien assuré : la superficie, le niveau du capital mobilier et le nombre de logements.
- Des variables géographiques : les zoniers des garanties dégâts des eaux et inondation.
- Des variables relatives au client : le groupe terme client et l'ancienneté du contrat.
- Des variables externes : le nombre de voisins dans un rayon de 50 mètres (à l'adresse) , la part des résidences principales en location et la part des résidences principales avec un chauffage collectif (INSEE).

D'autres variables se révèlent importantes dans les modèles de Random Forest et Gradient Boosting mais qui ne sont pas captées par le modèle GLM. C'est le cas des données relatives au client (le nombre de contrats en cours et l'équipement) et des données INSEE (Part des maisons en résidences principales construites entre 1920-1945 et Part des maisons en résidences principales avec une superficie $> 120m^2$). Ces variables permettent de compléter les informations présentes dans le GLM et/ou d'apporter de nouvelles informations sur l'environnement dans lequel évoluent le bien et le contrat.

Au-delà de la concordance entre les modèles GLM, Random Forest et Gradient Boosting sur plusieurs variables explicatives du coût moyen, les constats relevés lors de l'interprétabilité des modèles de Machine Learning sur les variables les plus importantes par des méthodes PDP et SHAP, montrent que l'effet et le sens de l'impact de chacune d'entre elles sur le coût moyen vont dans le même sens que ceux du modèle GLM.

9.2 Comparaison du pouvoir prédictif des modèles

	RMSE - Base d'apprentissage	RMSE - Base de test
GLM	2 626,21 €	2 648,10 €
Random Forest	2 607,23 €	2 660,21 €
Gradient Boosting	2 622,88 €	2 662,42 €

Figure 9.2 – Comparaison de la qualité de prédiction des modèles de coût moyen selon l'indicateur RMSE - Maisons

L'analyse des performances des modèles par le biais de l'indicateur RMSE révèle que :

- Sur la base d'apprentissage : les modèles Random Forest et Gradient Boosting semblent avoir le meilleur pouvoir prédictif puisqu'ils induisent des erreurs plus faibles comparées à celle du modèle GLM.
- Sur la base de test : il n'existe pas de modèle qui l'emporte nettement sur les autres. Les RMSE sont assez proches. Toutefois, le modèle GLM présente une amélioration marginale en termes de performances de prédiction par rapport aux autres modèles en engendrant la plus faible erreur.

Conclusion intermédiaire

À l'aune de ces résultats, les modèles Random Forest et Gradient Boosting ont renforcé la validité des variables sélectionnées par le modèle GLM. De plus, ces modèles ont confirmé l'effet et le sens d'influence de ces variables sur le coût moyen à travers l'utilisation des méthodes d'interprétabilité PDP, ICE et SHAP. Toutefois, ils ne surpassent pas le modèle GLM en termes de performances de prédiction. De ce fait, **le choix s'est porté sur le modèle GLM pour la modélisation du coût moyen.**

Chapitre 10

La comparaison avec le tarif actuel

À présent, nous sommes en mesure de déterminer le modèle final de prime pure de la garantie dégâts des eaux en combinant les deux modèles de fréquence et de coût moyen. La prochaine étape consiste à formuler des préconisations en préparation des prochains renouvellements tarifaires. À cet effet, une étude comparative a été réalisée entre le modèle construit et le tarif actuel.

Pour chaque variable existante dans la structure tarifaire actuelle, l'écart tarifaire a été évalué en confrontant les coefficients résultants du modèle de la prime pure construit à ceux issus du tarif actuel. A noter que dans la séquence tarifaire actuelle, nous sommes en mesure d'identifier les chargements et les taxes. Dans le cadre de ces comparaisons, ces derniers éléments n'ont pas été retenus.

Les résultats ont mis à l'évidence la nécessité d'améliorer l'adéquation du tarif au risque réellement encouru. Ainsi, par une approche technico-commerciale, un plan d'action a été dressé pour revaloriser à la hausse ou à la baisse les coefficients issus du tarif actuel. Pour des raisons de confidentialité, seul l'écart tarifaire sur la variable «Superficie» des maisons (en m^2) sera révélé.

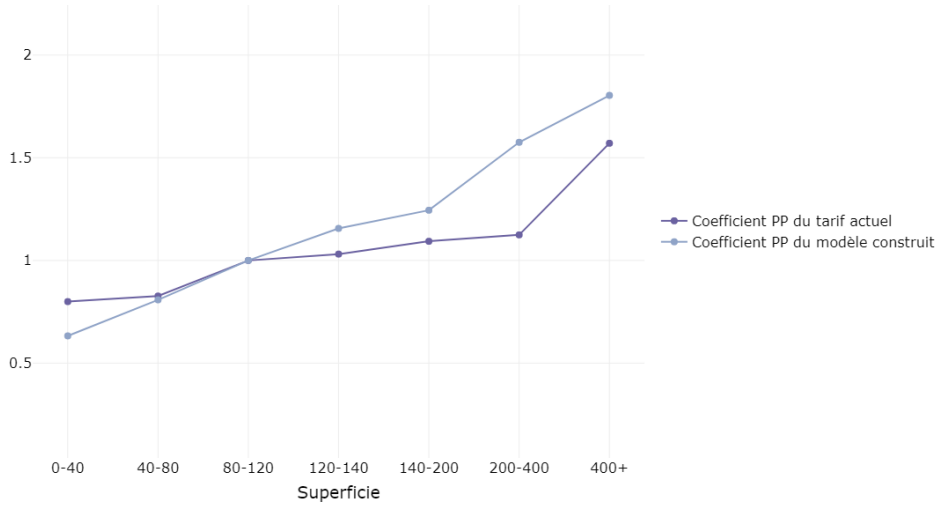


Figure 10.1 – Comparaison entre les coefficients de la prime pure du modèle et ceux du tarif actuel

Pour le critère «Superficie», comme le montre la figure 10.1, la préconisation réalisée consiste en une baisse du coefficient tarifaire de l'ordre de 17% pour les maisons avec une superficie inférieure à 40 m^2 , ainsi qu'une hausse de l'ordre de l'ordre de 9% pour les maisons avec

une superficie supérieure à 400 m^2.

Les ajustements tarifaires déterminés pour chaque critère seront par la suite lissés dans le temps. Un changement radical du tarif pourrait modifier complètement la structure du portefeuille et conduire à une situation d'antisélection et/ou rendre le tarif moins compétitif.

Par ailleurs, le modèle construit montre la pertinence de considérer de nouvelles variables tarifaires qui ne sont pas prises en compte dans la structure tarifaire actuelle. Deux variables ont retenu notre attention étant donné qu'elles sont présentes dans la structure tarifaire de l'offre principale et permettent donc de segmenter plus finement le tarif.

- L'étude menée dans ce mémoire a mis en évidence l'intérêt d'intégrer **le zonier de la garantie dégâts des eaux** dans la structure tarifaire en vigueur. C'est un facteur discriminant pour la fréquence des sinistres dégâts des eaux. Il figure également parmi les quatre variables qui contribuent le plus dans la prédiction du coût moyen.

  Afin de capter le risque géographique, le zonier PNO, construit à partir des fréquences de sinistres de toutes les garanties, est le critère utilisé dans la tarification actuelle. Le zonier de la garantie dégâts des eaux permettra de le remplacer, garantissant ainsi une meilleure segmentation du tarif.

  Afin d'intégrer ce zonier dans la séquence tarifaire actuelle du produit PNO, une révision s'avère nécessaire pour actualiser ce zonier dont la construction date de 2008. L'analyse cartographique réalisée dans l'étude du modèle GLM de fréquence appuie la nécessité de réviser ce zonier afin de l'adapter au risque encouru dans le produit PNO.

- Le **groupe terme client** ressort également comme une variable avec un impact significatif sur le tarif qui offre la possibilité d'identifier les clients dont les comportements pourraient justifier des ajustements à la hausse ou à la baisse des tarifs. En intégrant cette variable dans la structure tarifaire actuelle, il est possible de favoriser le tarif des bons risques et de faire payer le juste prix aux mauvais risques.

Il a été préconisé également de prendre en compte ces deux variables. Pour cela une expression de besoin a été formulée à la direction informatique pour changer la structure tarifaire actuelle.

Ces mesures appliquées à l'ensemble des garanties du produit PNO pourront potentiellement conduire à améliorer l'adéquation du tarif au risque, et par conséquent, de s'approcher de l'équilibre technique du produit PNO.

Conclusion

Dans un marché assurantiel saturé et hautement concurrentiel entre les différents acteurs classiques et néo-assureurs, la rétention des clients existants est devenue une priorité. Les assureurs adoptent des stratégies telles que le multi-équipement pour maintenir leur position sur le marché. Dans ce contexte, Le produit propriétaire non-occupant (PNO) de MMA se révèle comme un produit d'accompagnement pour multi-équiper les clients. Cependant, ce produit est en déficit technique. Il était donc impératif pour l'équipe de tarification de MMA de s'assurer de la cohérence du tarif au risque réellement encouru.

Le but du présent mémoire était de challenger la tarification actuelle du produit PNO -dont la dernière révision structurelle remonte à 2016- par une approche de modélisation de la prime pure. Cette étude s'est focalisée sur la garantie dégâts des eaux qui représente un poids significatif en termes de sinistralité et offre l'opportunité d'actionner des leviers tarifaires afin d'atteindre l'équilibre technique.

Après la construction de la base de données avec un historique suffisant et la réalisation de différentes étapes de retraitements, nous avons entrepris une modélisation GLM pour la fréquence et le coût moyen des sinistres. Cette démarche a été appliquée séparément aux maisons et aux appartements en raison de la répartition hétérogène de la sinistralité entre ces deux types de logements.

Les modèles GLM étant construits, une étude de leurs performances a été menée. À l'aune des résultats obtenus, les modèles de fréquence ont fait preuve d'un fort pouvoir prédictif. Cependant, les modèles de coût moyen ont montré des performances relativement moindres. Afin d'améliorer ces performances, il a été décidé de challenger ce modèle en ayant recours à deux algorithmes de Machine Learning à savoir le Random forest et le Gradient Boosting.

Malgré les étapes d'optimisation, une comparaison des performances des modèles selon l'indicateur RMSE a révélé que les deux modèles Random Forest et Gradient Boosting ne surpassent pas le modèle GLM. Toutefois, ils ont contribué à nous conforter quant à la significativité des variables sélectionnées par le modèle GLM, étant donné qu'une grande majorité d'entre elles se sont révélées importantes dans la construction des modèles de Machine Learning. De surcroît, ces modèles ont confirmé l'effet et le sens d'influence de l'impact de ces variables sur le coût moyen par le biais des méthodes d'interprétabilité PDP, ICE et SHAP. Nous avons donc décidé de conserver les modèles fréquence x coût moyen par l'approche GLM et de les combiner pour obtenir finalement le modèle de prime pure.

Une étude comparative avec le tarif actuel a été menée. À l'issue de cette comparaison, des préconisations ont été formulés en vue à d'améliorer la cohérence du tarif au risque encouru. Ces mesures concernent la revalorisation à la hausse ou à la baisse des coefficients associés aux différents critères présents dans la structure tarifaire actuelle. De plus, il est envisagé d'intégrer dans le tarif commercial, de nouvelles variables significatives révélées par le modèle construit. Parmi ces variables figurent le zonier de la garantie dégâts des eaux (zonier de l'offre principal)

et le groupe terme client. Ces variables exploitées sur l'offre principale s'avèrent pertinentes à intégrer dans la tarification de l'offre PNO pour avoir une meilleure segmentation du tarif.

Des travaux ultérieurs seront axés sur l'extension de cette étude aux autres garanties du produit PNO, dans le but d'apprécier l'équilibre technique global du produit.

Annexe

Analyses uni-variées pour les appartements

Fréquence

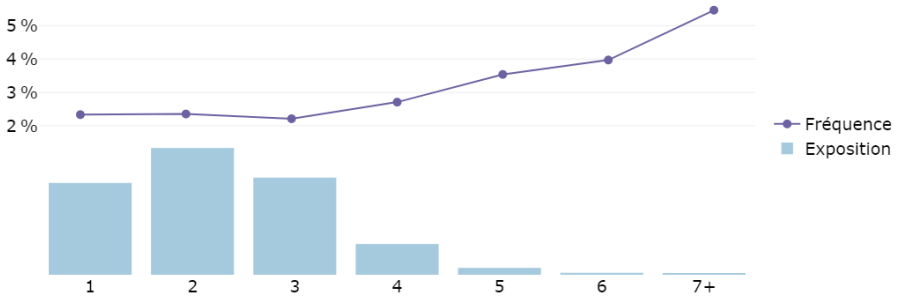


Figure 10.2 – Fréquence observée selon le « Nombre de pièces »

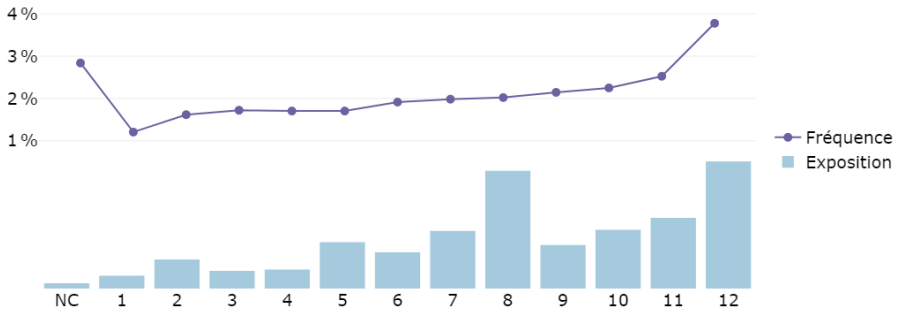


Figure 10.3 – Fréquence observée selon le « Zonier DDE »

Coût moyen

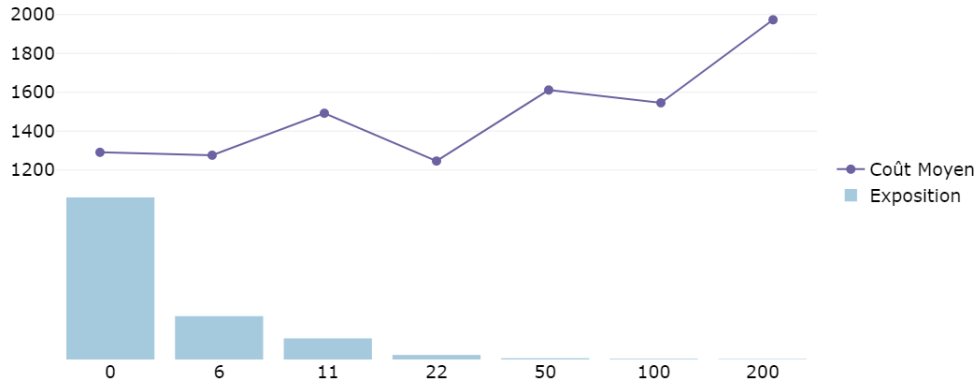


Figure 10.4 – Coût moyen observé selon le « Niveau du capital mobilier »

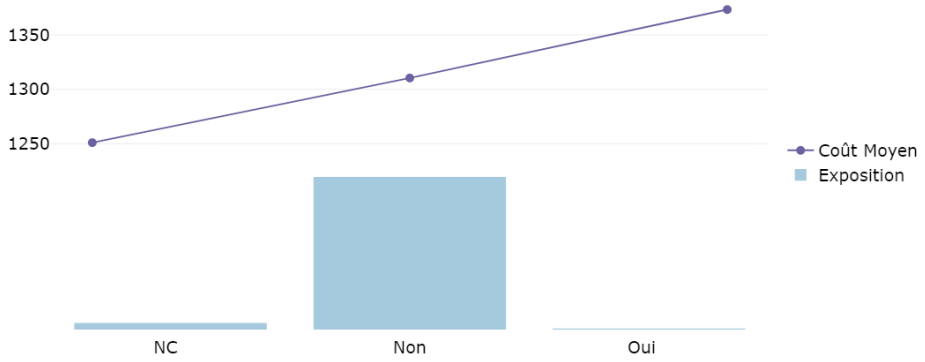


Figure 10.5 – Coût moyen observé selon la variable externe « DDE vertical »

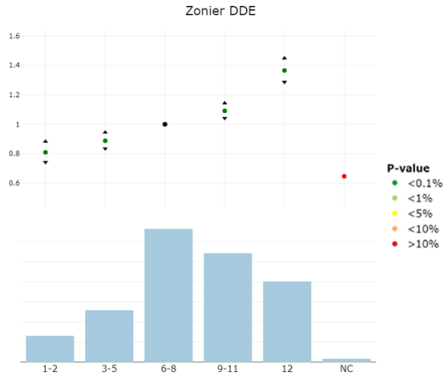
Modélisation de la garantie dégâts des eaux pour les appartements

Modèle Fréquence x Coût moyen – Appartements

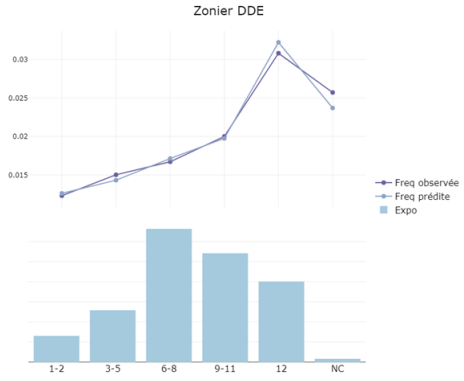
Nombre de pièces	✓ ✓	Zonier DDE	✓ ✓	Part RP avec chauffage indiv		Part RP avec 5pièces ou +	✓
Niveau capital mobilier	✓	Zonier inondation	✓ ✓	Part des RP avec Chauff Coll	✓	Part ménages en couple avec enf	✓ ✓
Franchise	✓	Zonier PNO	✓	Part RP-MAISON avant1919	✓	Part ménages en couple sans enf	
Formule	✓ ✓	DDE vertical	✓	Part RP-MAISON entre 1920-1945		Part pop entre 60-74ans	
Nb sinistres sur 12 derniers mois	✓ ✓	Equipement	✓ ✓	Part RP-MAISON entre 1946-1970		Part pop >= 75ans	
Nb sinistres sur 36 derniers mois	✓	Nb de contrats en cours	✓ ✓	Part RP-MAISON entre 1971-1990	✓	Part des artisans	
Ancienneté	✓ ✓			Part RP avec spfc <30m²	✓	Part des cadres	✓
Age	✓ ✓			Part RP avec SUP> 120m²		Part des agriculteurs	
Groupe terme client	✓ ✓			Part RS	✓	Part des inactifs	✓
				Part logements vacants	✓	Part des chômeurs	
				Part RP en location		Part des prof. Inter	

Figure 10.6 – Variables sélectionnées dans le modèle fréquence x coût moyen - Appartements

Fréquence

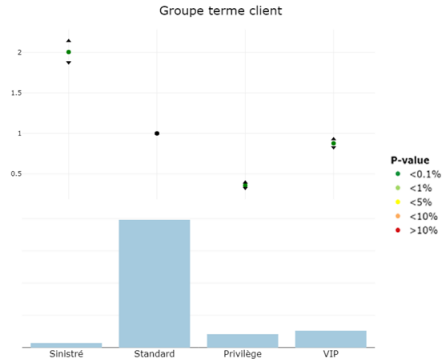


(a) Coefficients et intervalles de confiance pour la variable «Zonier DDE» - base d'apprentissage

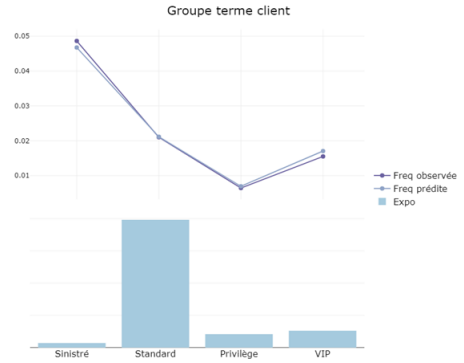


(b) Comparaison entre la fréquence observée et la fréquence prédite pour la variable «Zonier DDE» - base de test

Figure 10.7

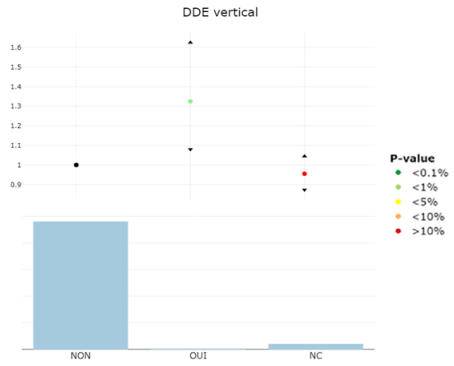


(a) Coefficients et intervalles de confiance pour la variable «Groupe terme client» - base d'apprentissage

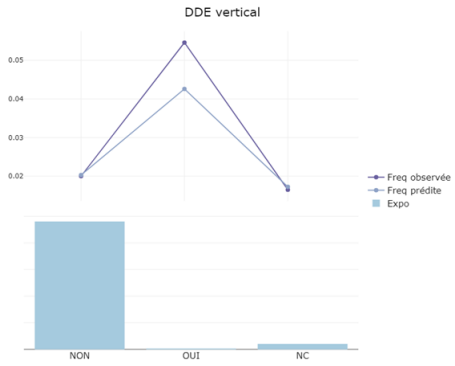


(b) Comparaison entre la fréquence observée et la fréquence prédite pour la variable «Groupe terme client» - base de test

Figure 10.8



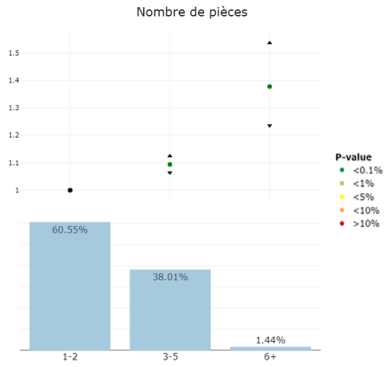
(a) Coefficients et intervalles de confiance pour la variable «DDE vertical» - base d'apprentissage



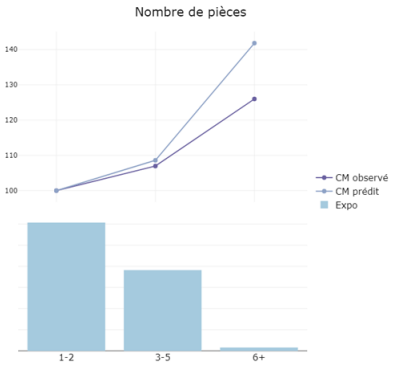
(b) Comparaison entre la fréquence observée et la fréquence prédite pour la variable «DDE vertical» - base de test

Figure 10.9

Coût moyen

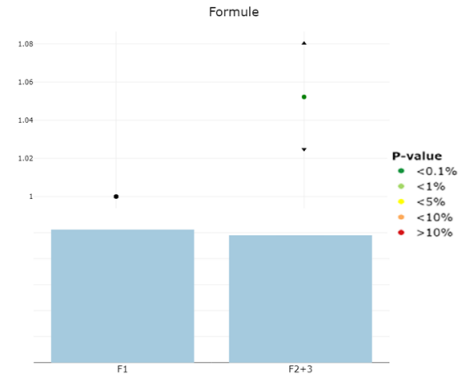


(a) Coefficients et intervalles de confiance pour la variable « Nombre de pièces» - base d'apprentissage

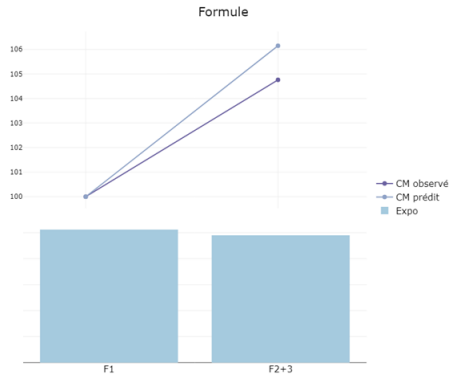


(b) Comparaison entre coût moyen observé et coût moyen prédit pour la variable «Nombre de pièces» - base de test

Figure 10.10

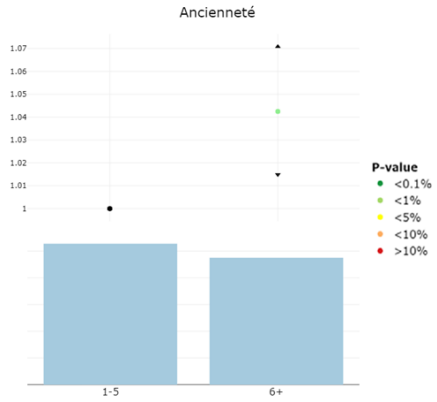


(a) Coefficients et intervalles de confiance pour la variable « Formule» - base d'apprentissage

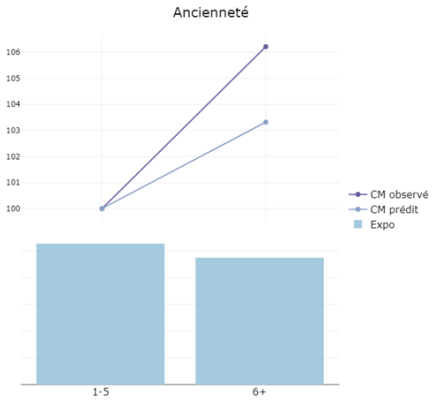


(b) Comparaison entre coût moyen observé et coût moyen prédit pour la variable «Formule» - base de test

Figure 10.11



(a) Coefficients et intervalles de confiance pour la variable « Ancienneté» - base d'apprentissage



(b) Comparaison entre coût moyen observé et coût moyen prédit pour la variable «Ancienneté» - base de test

Figure 10.12

Bibliographie

[1] P. Aillot. *Modèle linéaire généralisé*. 2022.

[2] K.J. Arrow. *Uncertainty and the welfare economics of medical care*. 1963.

[3] France Assureurs. *L'assurance habitation en 2021*. 2022.

[4] L. Breiman. *Random forests*. Machine learning 45.1, 2001.

[5] S. Bucci. *Étude et implémentation de techniques d'analyse de sensibilité dans les modèles de tarification Non-Vie. Application à la tarification à l'adresse*. 2021.

[6] A. Charpentier, M. Denuit, R. Elie. *Segmentation et mutualisation, les deux faces d'une même pièce*. 2015.

[7] D. Delcaillau. *Contrôle et Transparence des modèles complexes en actuariat*. 2019.

[8] G. Eldin. *Construction d'un indicateur de Valeur Client et optimisation tarifaire en assurance non-vie*. 2018.

[9] J. N. Morgan et J.A. Sonquist. *Problems in the analysis of survey data, and a proposal*. Journal of the American Statistical Association, 58(302) :415–434, 1963.

[10] J. H. Friedman. *Greedy Function Approximation : A Gradient Boosting Machine*. 1999.

[11] J. H. Friedman. *Greedy Function Approximation : A Gradient Boosting Machine. The Annals of Statistics*. 2001.

[12] A. Goldstein. *Peeking Inside the Black Box : Visualizing Statistical Learning with Plots of Individual Conditional Expectation*. 2015.

[13] N. Khemliche. *Tarification du produit propriétaire non-occupant*. 2017.

[14] R. Laïly. *Tarification non-vie sur R*. 2022.

[15] S. Navarro. *L'Open Data au service de la tarification à l'adresse*. 2017.

[16] N. Raillard. *Modélisation statistique des valeurs extrêmes*. 2021.

[17] F. Vermet. *Apprentissage statistique : une approche connexionniste*. 2022.