Mémoire présenté le 11 janvier 2021
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : **Alexis Bernanose**

Titre du mémoire : Modélisation du risque incendie en assurance Habitation

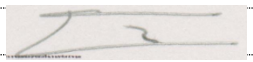
Confidentialité : ☐ NON ☒OUI (Durée : ☐ 1 an ☒2 ans)

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

Membres présents du jury de l'Institut des Actuaires signature

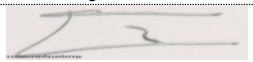
Entreprise : AXA FRANCE

Nom : LUU François

Signature : 

Directeur de mémoire en entreprise :

Nom : LUU François

Signature : 

Membres présents du jury de la filière

Invité :

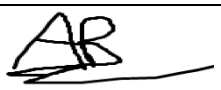
Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat



Résumé

Le marché de l'assurance Multirisque Habitation est en perpétuelle évolution et soumis à une très forte concurrence. Par conséquent, les assureurs doivent aujourd'hui proposer un tarif de plus en plus fin et adapté au plus près du risque des assurés. Dans ce cadre, Axa France propose un produit avec une gestion des risques bien contrôlée.

Le produit Multirisque Habitation est composé de plusieurs garanties, dont l'incendie, qui est l'une des garanties principales en terme de charges. Cette garantie est particulière car la majeure partie de la charge est causée par très peu de sinistres. Ce risque est alors défini comme un risque d'intensité. Ainsi, sa nature rend son traitement spécial, et implique de lui porter une attention particulière.

La segmentation tarifaire de ce risque n'a pas été revue depuis plus de deux ans. Or, les profils de risque des assurés inhérents à notre portefeuille évoluent avec le temps, et une étude de rentabilité récente a montré certaines lacunes concernant la segmentation de cette garantie. Ce défaut de segmentation implique l'augmentation de l'anti-sélection des risques, et empêche d'obtenir un tarif plus compétitif. L'enjeu de ce mémoire est de proposer une modélisation fine du risque incendie permettant de fournir une meilleure segmentation.

Tout d'abord, l'étude détaille l'ensemble du processus de construction de la base de modélisation. Ensuite, la modélisation du risque incendie est décomposée en trois parties : la modélisation de la fréquence, de la propension puis du coût moyen. Dans ce cadre, une attention particulière sera portée sur la sinistralité grave. Enfin, les trois modélisations sont combinées afin d'aboutir à la prime pure. Cette dernière correspond au montant du sinistre moyen auquel devra faire face l'assureur pour le risque. Ainsi, une évaluation de la performance de cette prime pure est effectuée.

Mots clés

Multirisque Habitation, Concurrence, Tarif, Gestion des risques, Incendie, Risque d'intensité, Segmentation, Rentabilité, Anti-sélection, Modélisation, Fréquence, Propension, Coût moyen, Sinistralité grave, Prime pure.

Note de synthèse

Cette étude a été effectuée au sein de la Direction Marché IARD des particuliers d'Axa France, dans l'équipe Pricing MRH, et concerne donc l'assurance Multirisque Habitation. L'objectif de ce mémoire est d'effectuer une refonte de la modélisation du risque incendie tout en proposant un processus générique de tarification. Ainsi, elle présente le processus de construction de la base de modélisation, le retraitement de plusieurs variables et la création de modèles de fréquence, de propension et de coût moyen.

Cadre et objectifs de l'étude

L'assurance IARD (Incendie, Accidents et Risques Divers) est une famille d'assurances couvrant les biens des assurés en cas de dommages. Ce mémoire se concentre sur l'assurance Multirisque Habitation, qui est l'une de principales assurances pour les particuliers sur le marché IARD. Effectivement, les cotisations perçues sur le produit Multirisque habitation représentent 18,7% des cotisations de ce marché.

Grâce aux nouvelles plateformes digitales, comme les réseaux sociaux, ou encore les comparateurs en ligne, les futurs preneurs d'assurances ont la possibilité de comparer un très grand nombre d'offres. Le marché de l'assurance Habitation est donc aujourd'hui soumis à une très forte concurrence. Dans ce cadre concurrentiel, Axa France propose des produits avec une gestion des risques bien contrôlée.

Le produit d'assurance Multirisque Habitation proposé par Axa France est structuré en deux parties. D'une part, des garanties de bases épurées vendues à un prix très compétitif, lesquelles forment le socle de l'offre. D'autre part, des garanties optionnelles que l'assuré peut choisir à la carte. Dans le cadre de ce mémoire, nous étudions en détail la garantie incendie, faisant partie intégrante du socle. C'est une garantie autant spéciale que fondamentale couvrant les dommages causés entre autres par les évènements tels que l'incendie, l'explosion, l'implosion, la chute de la foudre ou encore le choc d'un appareil aérien ou spatial.

Les données du portefeuille

La modélisation du risque associé à une garantie se fait par le biais de la création d'une prime pure. Dans un cadre classique, un modèle de fréquence et de coût moyen suffit pour créer cette prime. Cependant, parce que la garantie incendie possède un certain nombre de sinistres ayant un coût très élevé, nous intégrons un modèle de propension permettant de distinguer la sinistralité dite attritionnelle de la sinistralité dite grave. Dans ce mémoire, les informations permettant de créer cette prime ne prennent pas en compte les données géographiques.

Nous présentons dans ce chapitre le processus de construction de la base de données sur laquelle se base la modélisation. Ainsi, une base par image de risque de 2016 à 2018 est créée, par le biais de la jointure des données contrats et sinistres. Cette étape doit être effectuée soigneusement car la qualité des données est un enjeu primordial pour la modélisation.

Afin que la modélisation ne soit pas perturbée par des valeurs aberrantes et que les observations ayant un coût extrême soient étudiées séparément, un seuil (représentant un coût en €) doit être déterminé. Dans ce cadre, la théorie des valeurs extrêmes est utilisée et permet d'en définir deux : 100 000 € pour les appartements et 300 000 € pour les maisons.

La base sur laquelle la modélisation est effectuée est composée de données récentes, toutes comprises entre 2016 et 2018. De ce fait, le nombre et la charge de sinistres récupérés initialement ne sont pas définitifs. Certains sinistres survenus n'ont pas encore été déclarés à la date de l'étude, et d'autres nécessitent plusieurs évaluations du dédommagement. Or, l'objectif est d'obtenir une prime pure en accord avec la sinistralité observée, et par conséquent, nous avons besoin d'un nombre et d'une charge de sinistre définitive. Ainsi, pour développer nos sinistres, nous utilisons la méthode de Chain Ladder.

Les statistiques descriptives montrent que les propriétaires de maisons sont les plus exposés à voir survenir un sinistre incendie. Les locataires de maisons ont quant à eux le coût moyen le plus élevé. De plus, le risque augmente lorsque le nombre de pièces et lorsque l'habitation possède un insert (cheminée ou poêle à bois).

Les profils de risques au sein de notre portefeuille évoluent au cours du temps. Nous effectuons donc un audit du dernier modèle mis en production, afin d'observer s'il ajuste correctement les données actuelles. Nous remarquons alors qu'il sur-estime la sinistralité réelle, impliquant une sur-tarification, et de ce fait une exposition à l'anti-sélection.

Retraitement des variables

Une fois que les données nécessaires à la modélisation ont été rassemblées, certains retraitements de variables ont été effectués dans le but de modéliser dans les meilleures conditions.

Tout d'abord, les valeurs manquantes ont été traitées. Ces dernières peuvent introduire un biais, rendre le traitement des données plus laborieux et réduire l'efficacité des modèles. Nous les imputons donc dans cette étude par les plus proches voisins.

Les variables quantitatives sont discrétisées par arbre de régression. En d'autres termes, nous transformons les variables quantitatives en qualitatives, afin d'identifier plus facilement des sous-portefeuilles de risques au sein de notre portefeuille et pour prendre en compte certains effets non-linéaires.

Les variables explicatives utilisées pour les modèles linéaires généralisés nécessitent d'être non-corrélées. Par conséquent, une étude des corrélations au sein de nos variables est effectuée par un V de Cramer. Nous remarquons que deux variables sont extrêmement corrélées, et par conséquent, apportent globalement la même information. Nous décidons donc d'en supprimer une.

Aspects théoriques

Ce chapitre a pour ambition d'introduire théoriquement les concepts et méthodes utilisés par la suite. Dans un premier temps, nous présentons les modèles linéaires généralisés, certaines méthodes d'apprentissage supervisées et l'optimisation de leurs paramètres. Effectivement, ces différents modèles sont utilisés et comparés tout au long de l'étude. Enfin, le sur-échantillonnage synthétique qui est utilisé pour la gestion de la quantité de données sur la sinistralité grave est introduit.

Les modèles linéaires généralisés, notés GLMs, sont des méthodes classiques de tarification d'assurances non-vie. Ils sont structurés de telle manière que leur composante aléatoire et leur composante déterministe soient reliées par une relation fonctionnelle, appelée aussi fonction de lien. Nous introduisons donc l'estimation de leurs paramètres dans le cadre de la régression Lasso, Ridge et Elastic Net.

Nous introduisons ensuite les arbres de régression et de classification simples. Ces derniers segmentent l'espace des variables explicatives, et prédisent la moyenne ou la classe ayant la majorité par vote dans la région de l'espace associée à l'observation. Les forêts aléatoires et eXtreme Gradient Boosting sont des méthodes d'agrégation d'arbres simples permettant d'augmenter fortement la précision et la robustesse des prédictions.

Les paramètres de ces modèles doivent être optimisés. Le processus d'optimisation est effectué dans cette étude par validation croisée et par recherche sur grille des différents paramètres. La grille de recherche teste les modèles pour chaque combinaison de paramètres, tandis que la validation croisée permet d'obtenir une évaluation moyennée des modèles pour chacune de ces combinaisons.

La Synthetic Minority Over-sampling Technique (algorithme SMOTE) est une méthode de sur-échantillonnage synthétique d'une classe minoritaire, correspondant dans notre étude aux observations ayant eu un sinistre grave. Cet algorithme consiste à créer de nouveaux individus synthétiques ressemblant très fortement aux individus réels de la classe minoritaire.

Modèles de fréquence

Dans ce chapitre, nous modélisons la fréquence de sinistres en distinguant les appartements et les maisons par le biais des GLMs pénalisés et des méthodes d'agrégation d'arbres que sont les forêts aléatoires et le XGBoost. La distinction de la modélisation est justifiée par le fait que chaque type d'habitation répond à des problématiques de modélisation différentes.

Concernant la modélisation GLM, nous testons plusieurs GLMs Poisson pénalisés. Ainsi, les variables importantes pour les appartements sont le nombre de pièces, la possession d'un insert ou encore la surface des dépendances. Pour la modélisation du risque maison, la qualité de l'occupant et la surface des dépendances influent fortement sur le risque de survenance d'un sinistre.

La forêt aléatoire et le XGBoost ont des résultats ressemblants. En effet, l'importance des variables et les métriques d'évaluation sont assez proches. Ainsi, nous retenons que le nombre de pièces, le capital assuré, et la surface des dépendances jouent un rôle important sur le risque d'occurrence d'un sinistre incendie.

Au vu des résultats, il semble plus judicieux de retenir les GLMs pénalisés pour modéliser la fréquence de sinistres. Effectivement, face aux méthodes d'agrégation d'arbres, les GLMs ont l'avantage d'être plus simples d'interprétation par le biais de l'analyse des coefficients. De plus, dans notre étude, ces dernières possèdent une précision et une qualité de segmentation semblables aux méthodes d'apprentissage supervisés.

Modèles de propension

Ce chapitre a pour ambition de modéliser la probabilité d'occurrence d'un sinistre grave. Avec 1,1% de sinistres graves parmi l'ensemble des sinistres, nous avons très peu d'informations sur les profils des assurés ayant ce genre de sinistres. Ainsi, les modèles de propension ne distinguent pas la sinistralité des appartements de la sinistralité des maisons. Néanmoins, à l'instar de la modélisation de la fréquence, les GLMs pénalisés et les méthodes d'agrégation d'arbres sont comparés.

Un GLM Binomial (appelé aussi régression logistique) est effectué pour modéliser de façon classique cette probabilité. Nous remarquons que le nombre de pièces, l'étage, ou encore la possession d'un jardin augmentent fortement la probabilité d'occurrence d'un sinistre grave. De plus, malgré une quantité d'informations très faible sur cette sinistralité extrême, les GLMs arrivent à détecter près de 70% des sinistres graves sur une base test.

La forêt aléatoire créée obtient des résultats moins satisfaisants. L'importance des variables n'est pas similaire à celle de la régression logistique et nous notons un sur-apprentissage des données. Ce sur-apprentissage est expliqué par le fait que les arbres de décision laissent une complète liberté sur la forme du lien entre la variable réponse et les variables explicatives.

Comme pour le modèle de fréquence, nous conservons le GLM pénalisé. Dans le cadre du modèle de propension, les performances de la régression logistique surpassent celles de la forêt aléatoire.

Modèles de coût moyen

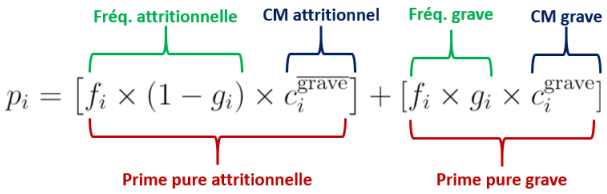
Pour compléter le calcul de la prime pure, une modélisation du coût moyen doit nécessairement être effectuée. Dans ce chapitre, nous séparons la modélisation du coût moyen attritionnel de celle du coût moyen grave. Pour le coût moyen attritionnel, tout en séparant l'étude par type de d'habitation (appartement et maison), nous comparons les GLMs avec les méthodes d'agrégation d'arbres. Ensuite, nous modélisons le coût moyen grave qui répond à une problématique bien différente. Les données sont très peu nombreuses et empêchent d'effectuer une modélisation classique. Ainsi, nous comparons une approche simple qu'est l'affectation de la moyenne par arbre de régression, avec une modélisation suite au sur-échantillonnage par la méthode SMOTE.

Plusieurs GLMs Gamma sont effectués pour modéliser le coût moyen attritionnel. Pour les deux types d'habitation, le nombre de pièces, la surface des dépendances ainsi que l'ancienneté du logement influent sur le coût lié à un sinistre. Cependant, la manière dont influent ces variables est différente pour chaque type d'habitation. De plus, les métriques d'évaluation de ces modèles sont satisfaisantes. Les résultats sont très similaires pour le XGBoost effectué. Seuls les aménagements extérieurs sont intégrés en plus comme variables influant sur le risque. De plus, sa qualité d'ajustement et de segmentation est très proche de celle des GLMs pénalisés. Pour des raisons d'efficacité et de simplicité, le GLM Gamma pénalisé est alors conservé.

La moyenne par arbre de régression fournit des résultats satisfaisants pour le coût moyen grave maisons. Néanmoins, concernant les appartements, les résultats observés sont à prendre avec plus de précaution. Effectivement, les résultats obtenus sont plus difficilement interprétables bien que les métriques d'évaluation montrent un bon ajustement aux données du modèle. L'algorithme SMOTE permet d'arriver à une conclusion : l'occurrence d'un sinistre grave pour les appartements est un évènement si rare qu'il relève plus de l'aléa que d'une explication statistique. Toutefois, les résultats de la modélisation grâce à l'algorithme SMOTE renvoient de bons résultats pour les maisons. Nous conservons l'approche simple qui a de bons résultats tout en évitant le sur-apprentissage des données.

Synthèse de modélisation

La prime pure s'obtient en combinant l'ensemble des modèles effectués précédemment. Pour obtenir la prime pure de chaque assuré i, notée p_i, nous appliquons la formule suivante :



Avec, pour l'assuré i :
— f_i : sa fréquence de sinistres ;
— g_i : sa probabilité d'occurrence d'un sinistre grave ;
— c_i^{grave} : son coût moyen grave ;
— $\overline{c_i^{\text{grave}}}$: son coût moyen attritionnel.

Les différents modèles permettent de créer une prime pure proche de la sinistralité réelle tout en identifiant un nombre conséquent de sous-profils de risque au sein de notre population. Effectivement, nous sur-estimons la charge de sinistres totale d'une grande base test de 3,6% tout en obtenant un indice de Gini proche de 40%. Ainsi, la modélisation du risque incendie est à présent à jour, et permet d'ajuster les données actuelles avec une meilleure qualité. Toutefois, la modélisation présentée dans cette étude ne prend pas en compte les informations géographiques. De ce fait, il reste donc la création d'un zonier de fréquence et de coût moyen pour finaliser la prime pure incendie.

Remerciements

Je tiens à remercier dans un premier temps l'ensemble de la Direction Marché IARD des particuliers d'Axa France et particulièrement l'équipe « Tarification Habitation » pour leur accueil et leur sympathie tout au long de mon alternance.

J'adresse notamment ma reconnaissance à Thomas Gauthron, Julien Durand, Anne-Laure Le Gallo et François Luu pour la confiance qu'ils m'ont accordée dès mon arrivée dans l'équipe.

Je remercie également ma tutrice de mémoire, Chae In Kim, pour son encadrement tout au long de ce travail de recherche.

Aussi, je souhaite remercier particulièrement Sebastien Perrin, Issam Mezrag et Mohamed Halimi pour les riches échanges qui ont nourri ce mémoire.

Mes remerciements à mes collègues Romain Toesca, Maud Vandekerchove, Hugo Hammerer, Gaelle Gouineau, Charles Partington et Chiu Wai Wong de l'équipe « Actuariat MRH » pour leur temps et les précieux conseils qu'ils m'ont fournis.

Enfin, je souhaite remercier les enseignants de l'ISUP qui ont assuré la partie théorique de ma formation et en particulier Maud Thomas, qui m'a suivi tout au long de ce mémoire.

Table des matières

Table des figures

Liste des tableaux

Introduction

Ce mémoire a été réalisé au sein de l'équipe Actuariat Tarification Habitation dans le cadre d'une refonte tarifaire du produit d'assurance Multirisque Habitation (dite MRH). D'une part, cette assurance permet de protéger le logement et les biens mobiliers qui s'y trouvent contre les dommages. De l'autre, elle permet de prendre en charge les préjudices causés aux tiers par le souscripteur, ses ayant-droits, voire ses animaux de compagnies par le biais de la garantie responsabilité civile. Les garanties les plus courantes sont le dégâts des eaux, l'incendie, le bris de glace, le vol et le vandalisme, et les catastrophes naturelles.

Le marché de l'assurance Habitation est aujourd'hui soumis à une très forte concurrence. De ce fait, la situation actuelle impacte fortement la dynamique de production du produit MRH. Dans ce contexte, le tarif du produit proposé par Axa France doit être revu régulièrement via une mise à jour des données de modélisation, et de l'application de nouvelles techniques actuarielles de tarification et de segmentation. Cette segmentation consiste à considérer que le risque diffère selon le client, et implique une prime différente. Cette notion permet alors de lutter efficacement contre l'anti-sélection. La tarification du produit Habitation étudié dans ce mémoire est basée sur une approche de modélisation de prime pure (prime minimale requise pour faire face à la sinistralité du portefeuille).

La segmentation du risque incendie n'a pas été revue depuis plus de deux ans. Ainsi, l'objectif de ce mémoire est d'effectuer une refonte tarifaire de ce risque. L'enjeu est alors de proposer une modélisation fine permettant de fournir une meilleure segmentation qu'auparavant. Dans un premier temps, nous présentons les caractéristiques de l'assurance Multirisque Habitation, quelques principes de tarification ainsi que la garantie incendie. Ensuite, une description détaillée du processus de construction de la base de modélisation sera effectuée. Nous aborderons notamment la gestion des sinistres graves, le développement des sinistres et le retraitement de certaines variables.

Dans une seconde partie, le risque incendie sera modélisé en prenant en compte les caractéristiques du logement et de l'assuré. Dans ce cadre, nous comparerons deux approches. La première est classique et consiste à utiliser les modèles linéaires généralisés. La seconde est innovante et consiste à utiliser des méthodes d'agrégation d'arbres que sont le Random Forest et le XGBoost. Ces dernières représentent une bonne alternative car elles permettent de faire face à la quantité massive de variables disponibles pour réaliser l'étude.

Chapitre 1

Cadre et objectifs de l'étude

1.1 L'Assurance Multirisque Habitation

1.1.1 Présentation du marché IARD

L'assurance **IARD**, signifiant "**I**ncendie, **A**ccidents et **R**isques **D**ivers", est une famille d'assurances couvrant les biens des assurés au quotidien en cas de dommages. Ces biens pouvant être endommagés par des évènements tels que l'incendie, le vol, le dégât des eaux ou encore les catastrophes naturelles. Plusieurs catégories d'assurance sont regroupées, dont les plus connues sont l'automobile, l'habitation ou encore les multirisques professionnelles. Ce groupement d'assurances se ditingue des assurances de personnes (regroupant l'assurance vie et de dommages corporels), qui, par exemple, contiennent des assurances santé, décès ou invalidité. Nous pouvons voir la distinction entre les différents risques ci-dessous :

<table>
<tr><th>Type d'assurance</th><th>Exemples</th><th colspan="2">Risque</th></tr>
<tr><td><u>Assurance vie</u>
Assurance dont l'aléa dépend de la durée de vie humaine</td><td>Temporaire décès</td><td rowspan="2">Assurance de personnes</td><td>Assurance Vie</td></tr>
<tr><td><u>Assurance de dommages corporels</u></td><td>Dépendance</td><td rowspan="2">Assurance non-vie</td></tr>
<tr><td><u>Assurance d'autres risques</u></td><td>Multirisque Habitation</td><td>Assurance IARD</td></tr>
</table>

FIGURE 1.1 – Distinction entre les différents risques d'assurance

Comme les assurances IARD ne protègent pas les assurés et leur personne, elles sont souvent complétées par des garanties afin de couvrir ces dommages (par exemple, la garantie responsabilité civile afin de couvrir les dommages causés à autrui).

Dans le cadre de ce mémoire, nous nous concentrons uniquement sur l'assurance **Multirisque Habitation**, qui, pour les particuliers, est l'une des principales assurances sur le marché **IARD**. Les cotisations perçues en France sur ce marché sont présentées sur la figure 1.2.

Sur ces diagrammes empilés, nous remarquons une tendance à la hausse du montant des cotisations perçues sur le marché **IARD** et **MRH**. Concernant le marché **MRH**, les cotisations sont passées de 9,1 milliards d'euros en 2014 à 10,5 milliards d'euros en 2018 (+15,4%). Également, en 2018, l'assurance **IARD** représentait un cumul de 56,1 milliards d'euros de cotisations. En d'autres termes, sur la même année, l'assurance **Multirisque Habitation** représentait 18,7% des cotisations perçues en assurance **IARD**.

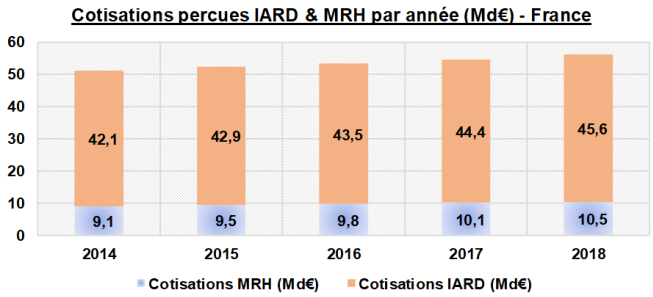


FIGURE 1.2 – Cotisations perçues sur le marché IARD et MRH par année en France

1.1.2 Axa France sur le marché Multirisque Habitation

Environnement concurrentiel et gestion des risques

L'environnement concurrentiel dans lequel évoluent les assureurs est en perpétuelle évolution. En effet, nous pouvons illustrer la situation avec la loi consommation, dite loi Hamon, entrée en vigueur le 1er janvier 2015. Cette loi donne la possibilité aux assurés de résilier leurs contrats auto, moto et **habitation** à la date de leur choix, passé un an de contrat. De ce fait, cette loi est vecteur de compétitivité et donc permet aux assurés de faire jouer la concurrence. Ainsi, les assureurs doivent constamment étudier leur tarification afin de conserver leurs parts de marché et stimuler leurs résultats.

Par ailleurs, les assureurs doivent également s'adapter aux comportements des clients, qui, aujourd'hui, sont davantage présents sur Internet et les réseaux sociaux. Ces nouveaux moyens de communication leurs permettent de comparer un grand nombre d'offres, et les rendent ainsi beaucoup plus sensibles aux tarifs et aux personnalisations proposés.

Dans ce cadre, Axa France propose des offres avec une gestion des risques bien contrôlée. En effet, ce contrôle est mis en place notamment par le biais d'un questionnaire de souscription complet, qui permet de proposer au prospect un tarif au plus proche de son profil de risque. De plus, certaines habitations peuvent subir un contrôle spécifique si leur profil de risque en est jugé ainsi. Il s'agit par exemple d'habitations contenant des chiens de catégorie 1 à 3, des équidés, ou encore du matériel professionnel. Une gestion des risques contrôlée pour un assureur est primordiale. En effet, une mauvaise gestion des risques impliquerait un fort impact négatif sur les objectifs, le patrimoine et la performance de l'entreprise.

Quelques chiffres

Malgré un chiffre d'affaires relativement stable autour de 900 millions d'euros par an entre 2016 et 2019, le volume du portefeuille **MRH** d'AXA France a légèrement diminué durant cette période. Nous pouvons le voir sur le graphique suivant :

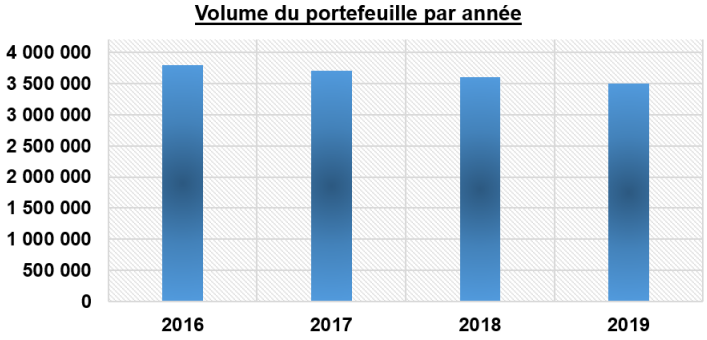


FIGURE 1.3 – Volume du portefeuille d'Axa France par année

Sur ces diagrammes, nous remarquons que le volume du portefeuille est passé de 3,8 millions en 2016 à 3,5 millions en 2019 (-7,9%). L'explication vient du fait que l'apport net (nombre d'affaires nouvelles soustrait du nombre de résiliations) est négatif durant ces années. Nous pouvons le constater sur le graphique suivant :

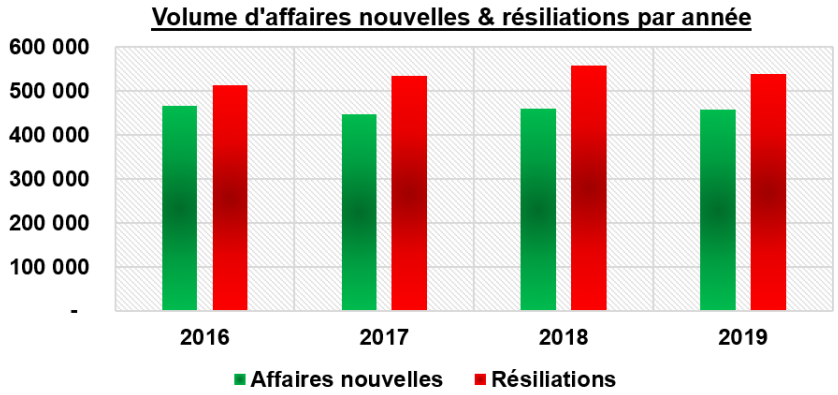


FIGURE 1.4 – Volume d'affaires nouvelles et de résiliations par année

L'offre MRH actuelle d'Axa France

Axa France possède deux principaux produits pour assurer les habitations de leurs clients :

— Le produit **Confort** : ancien produit qui ne peut être souscrit aujourd'hui que pour les risques "grandes demeures", "étudiant" et "propriétaire non occupant".

— Le produit **Ma Maison** : nouveau produit innovant lancé à partir de mai 2017, avec l'ambition d'être rentable sur la durée et de révolutionner le modèle économique MRH. C'est le principal produit MRH d'AXA France à ce jour, le produit **Confort** étant amené à disparaître progressivement. Néanmoins, le produit **Ma Maison** ne peut pas être souscrit aujourd'hui pour les risques "grandes demeures" et "propriétaire non occupant".

Ainsi, il sera proposé au prospect le produit Confort ou Ma Maison en fonction de son risque. Seul le produit "étudiant" peut encore être souscrit pour les deux produits. Nous résumons les différents risques couverts par chaque produit à travers la figure ci-dessous :

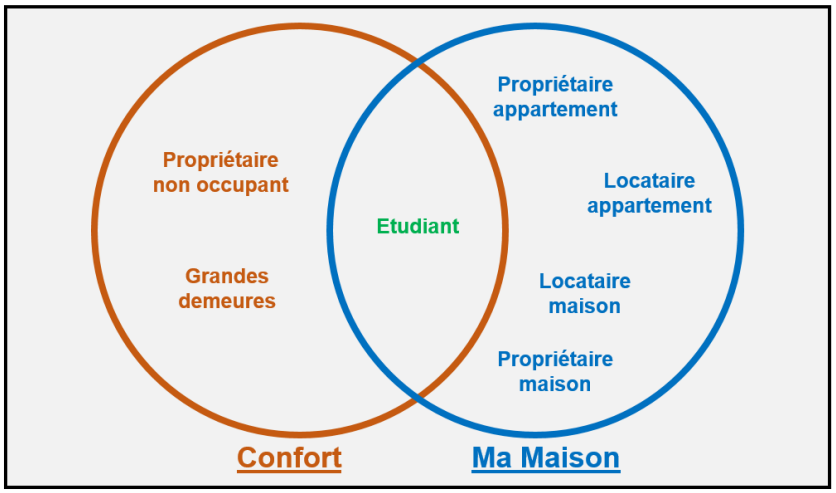


FIGURE 1.5 – Risques couverts : Confort et Ma Maison

Le produit **Ma Maison** est structuré en deux parties : tout d'abord avec des **garanties de base** épurées (appelé socle) ayant un prix très compétitif, et avec des **garanties optionnelles** que l'assuré peut choisir à la carte. Voici une brève description des principales garanties que contient le socle du produit **Ma Maison** :

- **l'incendie** : couvrant les dommages causés par l'incendie, l'explosion, l'implosion, la fumée, la foudre, le choc de véhicules terrestres à moteur et la chute d'appareils de navigation aérienne ;
- **les évènements climatiques** : couvrant les dommages causés par la tempête, la neige, la grêle et l'inondation ;
- **le dégât des eaux** : couvrant les dommages causés par une fuite d'eau ;
- **les catastrophes naturelles** : couvrant les dommages causés par une catastrophe naturelle. La garantie étant mise en jeu après publication au Journal Officiel de la République Française de la décision de l'autorité administrative ayant constaté l'état de catastrophe naturelle ;
- **les catastrophes technologiques** : couvrant les dommages causés par une catastrophe technologique. La garantie étant mise en jeu après publication au Journal Officiel de la République Française de la décision de l'autorité administrative ayant constaté l'état de catastrophe technologique ;
- **les attentats et les actes de terrorisme** : couvrant les dommages causés par un attentat ou un acte de terrorisme, tel que défini aux articles 421-1 et 421-2 du Code pénal ;
- **la responsabilité civile** : couvrant les dommages causés par l'assuré, lorsqu'il agit en qualité de simple particulier, dans le cadre de sa vie privée, y compris lors de la pratique de sports ou de loisirs.

Au sein des garanties optionnelles, les principales sont les suivantes : le vol au domicile, le vol à l'extérieur du domicile, le bris de vitres, le capital sécurité ou encore la protection juridique.

De plus, différents niveaux d'indemnisation sont mis en place. Concernant la valeur de remplacement des biens, trois choix sont envisageables :

- L'indemnisation correspondant à la valeur de reconstruction **vétusté déduite** ou encore appelée valeur d'usage. Cette indemnisation prend en compte l'usure du bien. De ce fait, elle sera inférieure à la valeur de reconstruction ;
- L'indemnisation correspondant à la **valeur à neuf 10 ans**. Cette indemnisation est égale à la valeur à neuf du bien endommagé pendant 10 ans. Au delà de cette période, l'indemnisation est inférieure à la valeur à neuf du bien ;
- L'indemnisation correspondant à la **valeur à neuf à vie**. Cette indemnisation est égale à la valeur à neuf du bien endommagé à vie.

En outre, une franchise peut être mise en place sur chaque contrat. La franchise étant la somme restant à la charge de l'assuré (donc non indemnisée par l'assureur) dans le cas où survient un sinistre. Intégrer une franchise permet à l'assureur de répondre à trois objectifs :

- Éliminer les dommages de fréquence qui ne sont pas du domaine des assurances et qui engendrent des frais de gestion importants ;
- Réduire l'aléa moral en incitant l'assuré à augmenter son niveau de prévention et de protection des risques ;
- Obliger l'assuré à dévoiler des informations précises sur son risque afin qu'il ait un niveau de prime en adéquation avec son risque.

Enfin, un traité de réassurance a été mis en place dans le but de se couvrir contre les sinistres les plus importants. Voici les différents types de réassurance auxquels Axa France a souscrit pour le périmètre MRH :

- Un **excédent de sinistres par risque**.
- Un **excédent de sinistres par évènements** de catastrophes naturelles.

1.2 Principes d'assurance

1.2.1 Principe de mutualisation et mesure du risque

L'assurance est un mécanisme de partage des risques, de sorte qu'ils se compensent entre eux. C'est ce que l'on appelle le **principe de la mutualisation des risques**. Il consiste à répartir le coût d'un sinistre entre les membres d'un groupe soumis potentiellement au même risque. Néanmoins, ce principe repose sur le fait que les sinistres survenus aux différents assurés soient **indépendants et identiquement distribués**, ce qui est approximativement le cas. Ainsi, ce principe permet d'obtenir, pour qu'une compagnie d'assurances soit rentable, l'inégalité suivante :

$$\sum_i P_i > \sum_i \mathbb{E}(X_i) + F_i \tag{1.1}$$

avec P_i la prime versée par l'assuré i, X_i le coût d'un sinistre pour l'assuré i et F_i les frais de l'assureur liés à l'assuré i.

Voici une illustration du principe de mutualisation en assurance MRH :

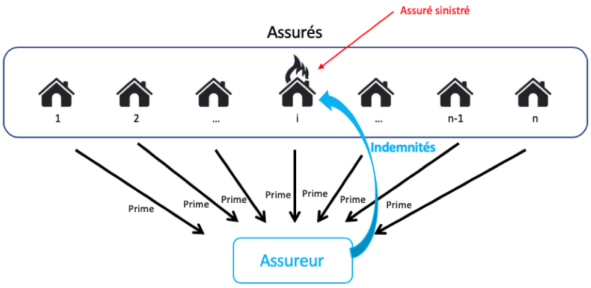


FIGURE 1.6 – Principe de mutualisation en assurance

Afin que l'effet de mutualisation puisse être effectif, l'assureur doit vendre un grand nombre de contrats. L'objectif étant d'obtenir un ratio de sinistralité sur prime favorable. Cependant, le niveau de primes doit être calculé en amont. La détermination de ce niveau de primes passe par une mesure du risque précise et robuste, afin de ne pas mettre en péril l'activité de la compagnie.

1.2.2 Construction de la prime

La prime d'assurance est le prix que l'assuré doit payer pour pouvoir bénéficier de la couverture d'assurance en cas de sinistre. Cette prime se décompose en plusieurs parties :

- **La prime pure** : correspond au montant du sinistre moyen auquel devra faire face l'assureur pour le risque. Mathématiquement, elle est égale à l'espérance des pertes.
- **Les chargements de gestion et d'acquisition** : chargements permettant à l'assureur de financer les coûts d'acquisition, la gestion des sinistres et la rémunération des intermédiaires (agents et courtiers).
- **Le chargement de sécurité** : chargement permettant à l'assureur de pouvoir résister à la volatilité naturelle des sinistres.

La prime ainsi définie est une prime entièrement technique. Finalement, cette prime est modifiée selon la politique commerciale de la compagnie d'assurances, pour aboutir à la **prime commerciale**. Seule **la prime pure** sera étudiée en détail dans ce mémoire. Ainsi, le passage de la prime pure à la prime technique et commerciale ne sera pas abordé.

1.3 Présentation du mémoire

Ce mémoire présente une étude effectuée au sein de la Direction Marché IARD des particuliers d'Axa France, dans l'équipe **Pricing MRH**. Le coeur de métier de cette équipe est la tarification. De ce fait, bien que les concepts évoqués lors de cette étude puissent être repris pour traiter d'autres sujets tels que le provisionnement ou encore l'analyse de la sinistralité, le mémoire est concentré sur la tarification.

En 2016, les modèles de primes pures ont été mis à jour pour plusieurs garanties dont **l'incendie**. Cependant, un constat a été fait : de nettes améliorations doivent être apportées, notamment au niveau de :

— la **segmentation** de la population. En effet, dans le cas d'une mauvaise segmentation, l'assureur est amené à sur-tarifer les "bon risques", tandis que les "mauvais risques" sont sous-tarifés. Cela est dû à la création de groupes de risques non homogènes. Ainsi, une anti-sélection a lieu et, de ce fait, l'assureur court à sa faillite ;

— la **précision** du modèle. Une mauvaise précision du modèle implique la mise en place d'un tarif qui n'est pas en adéquation avec le profil de risque de l'assuré. Ainsi, l'assureur s'expose également à l'anti-sélection.

L'objectif de ce mémoire est donc d'effectuer une refonte de **la modélisation du risque incendie**, et, de ce fait, d'obtenir une vision plus juste de ce risque au sein de notre portefeuille. Cette vision plus juste nécessite une meilleure segmentation de la population et une meilleure précision de nos modèles. Ces améliorations ont été apportées par le biais d'une étude approfondie et moderne. En effet, une comparaison détaillée des modèles créés par des méthodes dites "classiques" et d'apprentissage automatique a été effectuée. De plus, la création de ces modèles a été faite grâce à des données plus récentes et une vision plus précise du risque.

En outre, l'aspect recherche est une des composantes de ce mémoire. En effet, en plus de répondre à la problématique de modélisation du risque incendie, il fournit une méthode générique de tarification en assurance non-vie. De ce fait, cette méthode peut être utilisée pour modéliser le risque sur d'autres garanties Habitation ou d'autres périmètres tels que l'auto.

Les chapitres 2 et 3 traitent de la préparation des données, celles-ci étant l'exploitation des informations contrats et sinistres. C'est une étape très importante puisque les données sont la raison d'être des modèles. Effectivement, sans données préalablement bien préparées, les modèles ne peuvent être ni robustes ni correctement prédictifs. Par conséquent, la donnée a été sélectionnée de manière à ce qu'elle soit la plus exacte, récente et exhaustive possible. En effet, nous cherchons la vision du risque la plus juste.

Une fois les données préparées, l'objectif est de créer un modèle de prime pure sans données géographiques (la notion de prime pure est entièrement détaillée dans le chapitre 2). Effectivement, en amont de la création d'un zonier de fréquence et de coût moyen, de premiers modèles doivent être créés afin d'expliquer de manière robuste le risque. Les informations géographiques permettent dans un second temps d'expliquer les résidus de ces premiers modèles. Ainsi, dans le cadre de cette étude, des modèles de fréquence, de coût moyen et de propension ont été étudiés. Ces différents modèles sont issus de deux approches différentes :

— une approche dite **"classique"**. Dans ce cas, le modèle de prime pure est issu de modèles linéaires généralisés. Ces modèles statistiques supposent que les variables sont indépendantes, et se préoccupent avant tout de déterminer des paramètres.

— une approche par **apprentissage automatique**. Dans ce cas, le modèle de prime pure est issu de modèles de machine learning. Ces différents modèles, ayant pour but ultime la prédiction, sont rapides à exécuter, efficaces et trouvent seuls les interactions à inclure.

1.4 Présentation de la garantie Incendie

C'est à partir du XVIIe siècle que l'assurance va véritablement prendre les allures qu'on lui connaît aujourd'hui, et c'est un événement tragique qui va accélérer cette mutation. En effet, en 1666, à la suite du terrible incendie qui ravage plus de 13 000 bâtiments londoniens, Nicholas Barbon, un économiste britannique, invente le principe d'assurance incendie, ancêtre de l'assurance Habitation. En France, ce n'est que 37 ans plus tard qu'apparaît la première assurance couvrant les incendies, nommée le "Bureau des incendies".

Les 4 principales causes d'un incendie dans une habitation sont :
— les installations électriques et de chauffage ;
— les conduits de fumées ;
— la cuisine ;
— les bougies, cigarettes et allumettes.

La **garantie incendie** fait partie intégrante du socle du produit MRH d'Axa France. C'est une garantie fondamentale qui couvre les dommages causés par les évènements suivants :
— l'incendie : défini comme la combustion avec flammes en dehors d'un foyer normal ;
— l'explosion : définie comme une augmentation subite et violente de la pression ou de la dépression de gaz ou de vapeur ;
— l'implosion : définie comme une irruption très brutale d'un fluide, d'un gaz dont la pression est beaucoup plus faible que la pression extérieure ;
— l'enfumage : défini comme l'émission soudaine de fumées provenant du fonctionnement défectueux d'un appareil, ou de l'incendie d'un appartement ou d'un bâtiment voisin ;
— la chute de la foudre ;
— le choc de véhicules terrestres à moteur ;
— le choc d'un appareil aérien ou spatial ou des objets tombant de ceux-ci ;

Le nom de la garantie peut être trompeur. Effectivement, beaucoup d'évènements autres que l'incendie sont couverts par cette garantie tels que le choc de véhicules terrestres à moteur ou encore le choc d'un appareil aérien ou spatial.

Également, l'assuré a le choix entre plusieurs niveaux d'indemnisation modulaires :
— La **valeur de remplacement** : l'habitation et les biens sont assurés en valeur de reconstruction, en laissant le choix à l'assuré de déduire la vétusté ou de considérer la valeur à neuf ;
— La **franchise** : définie comme la somme restant à la charge de l'assuré (donc non indemnisée par l'assureur) dans le cas où survient un sinistre.
— Les **capitaux garantis** : de 6K€ à 500K€, correspondant à la valeur de l'ensemble des biens assurés.

La déclaration des capitaux garantis lors de la souscription du contrat est très importante. En effet, lorsque les capitaux déclarés à l'assureur se révèlent inférieurs à la valeur constatée au moment du sinistre, l'indemnisation ne sera pas totale. C'est la règle proportionnelle des capitaux qui s'applique, qui a comme conséquence une indemnisation non totale du dommage causé par le sinistre.

Concernant les exclusions, seules celles communes à toutes les garanties s'appliquent pour l'incendie. Néanmoins, l'assuré est fortement conseillé et informé sur la prévention de ce risque.

Enfin, le risque incendie en habitation du particulier est défini comme un risque d'intensité. Effectivement, le risque que la garantie incendie s'active est faible, ce qui implique de petites fréquences de sinistres. Cependant, dès lors qu'elle s'active, le sinistre a un coût généralement très élevé.

1.5 Problématiques soulevées

Nous venons de voir dans la section précédente que le risque incendie est défini comme un risque d'intensité. De ce fait, la quantité de données relative à la sinistralité incendie est faible, ce qui peut s'avérer problématique pour obtenir une modélisation du risque robuste. Un compromis doit nécessairement être effectué entre quantité de données, et données correspondantes au risque actuel. Effectivement, les données sont obtenues par le biais d'un historique de sinistralité. De ce fait, s'il est trop ancien, la majorité des données ne reflétera pas le risque actuel bien qu'elles soient en grand nombre.

Dans le but d'obtenir une quantification du risque précise, les modélisations des chapitres 5, 6 et 7 sont effectuées par groupes de risques homogènes. Dans le cadre de cette étude, nous identifions deux profils de risques bien distincts face à la sinistralité incendie : les appartements et les maisons. Cependant, ces deux profils de risques ne sont pas nécessairement ceux qui sont les plus distincts en terme de sinistralité. Néanmoins, ils sont conservés pour plusieurs raisons. Dans un premier temps, cette distinction est justifiée par le fait que ces deux types de sinistres répondent effectivement à des problématiques de modélisation bien différentes. Ensuite, elle permet d'être en accord avec les études d'autres garanties telles que le dégât des eaux ou encore le vol, et permet donc de comparer les résultats obtenus facilement. Enfin, sa simplicité permet une présentation des résultats claire et une compréhension facile pour les collaborateurs n'ayant pas forcément de bagage actuariel.

Enfin, un compromis doit être effectué entre l'interprétabilité des modèles et la qualité d'ajustement. Effectivement, la précision de ces derniers n'est pas l'unique objectif visé lors de la tarification d'une garantie en assurance non-vie. Ils doivent être interprétables afin d'identifier les sous-groupes de risques inhérents à notre population. De ce fait, une bonne interprétabilité des modèles permet une compréhension approfondie de notre portefeuille. Elle est souvent simplifiée en prenant un nombre de variables réduit lors des modélisations. Toutefois, bien qu'elle soit importante pour les modèles, nous rappelons qu'une grande attention est portée sur la précision de ce dernier.

1.6 Résultats obtenus

Tout d'abord, la préparation des données ainsi que leur analyse exploratoire ont permis d'avoir connaissance du risque actuel, et de ce fait d'obtenir une vision améliorée de notre portefeuille.

Ensuite, les résultats obtenus permettent de conclure que les modèles effectués lors de cette étude ajustent bien mieux les données actuelles que les anciens modèles. En effet, la précision de l'ajustement ainsi que la qualité de segmentation sont nettement supérieures. Une refonte de la modélisation du risque incendie était donc nécessaire.

Par ailleurs, la modélisation du risque incendie proposée dans ce mémoire suggère une approche différente dans le traitement de la sinistralité dite grave (cf. section 2.1.4). Effectivement, l'utilisation d'un modèle de propension permet de scinder la prime pure en une composante attritionnelle et une composante grave, et de ce fait permet d'obtenir une étude plus fine du risque.

Chapitre 2

Les données du portefeuille

2.1 Création de la base de données

2.1.1 Périmètre de modélisation

2.1.1.1 La prime pure

Soit X_i le coût total des sinistres survenus au cours de l'exercice considéré pour l'assuré i (avec $X_i = 0$ si aucun sinistre n'est survenu). Si la compagnie d'assurance a n assurés, et si on suppose que les $X_i, i = 1, ..., n$ sont indépendants et identiquement distribués, par la loi des grands nombres

$$\frac{1}{n}\sum_{i=1}^{n} X_i \quad \underset{n\to+\infty}{\xrightarrow{p.s}} \quad \mathbb{E}(X_1). \tag{2.1}$$

Ainsi, si n est suffisamment grand, le coût pour l'assureur est approximativement $n\mathbb{E}(X_1)$. Donc, si chaque assuré paye individuellement une prime $\pi = \mathbb{E}(X_1)$, appelée **prime pure**, la compagnie d'assurance devrait pouvoir dédommager ses assurés.

Un des grands principes de l'activité d'assurance est **l'inversion du cycle de production**. En effet, l'assuré paye une prime et ne reçoit la prestation qu'à posteriori. La prestation, correspondant aux dédommagements des sinistres, n'est pas connue lors du paiement de la prime. La création de modèles de primes pures à partir de méthodes statistiques permet alors d'évaluer cette prestation.

2.1.1.2 Modèle de fréquence et de coût moyen

La détermination de la prime pure par l'approche des modèles de fréquence et de coût moyen consiste à estimer l'espérance de la valeur totale des sinistres survenus au cours de l'exercice considéré pour l'assuré i, notée $\mathbb{E}(X_i)$. La variable aléatoire X_i peut alors être décomposée de la manière suivante :

$$X_i = \sum_{j=1}^{N} C_j \tag{2.2}$$

Avec :
— N : une loi de probabilité à valeur dans $\mathbb{N}$, correspondant au nombre de sinistres survenus au cours de l'exercice considéré ;
— C_j : un ensemble de lois de probabilité à support dans $\mathbb{R}_+$, correspondant au coût des N sinistres au cours de l'exercice considéré.

Néanmoins, deux hypothèses très fortes et en toute rigueur fausses sont à considérer. En effet, il est supposé que sachant les informations liées au profil du client dont disposent les assureurs :
— les C_j sont indépendants et identiquement distribués ;
— les C_j sont indépendants de N_i.

Sous ces hypothèses, $\mathbb{E}(X_i)$ peut s'écrire de la manière suivante :

$$\mathbb{E}(X_i) = \mathbb{E}[\mathbb{E}(X_i|N)] = \mathbb{E}[\sum_{j=1}^{N} \mathbb{E}(C_j|N)] = \mathbb{E}[N \times \mathbb{E}(C_j)] = \mathbb{E}(N) \times \mathbb{E}(C_j) \tag{2.3}$$

Ainsi, on observe que l'espérance de la charge totale de sinistres est égale au produit des espérances du nombre et du coût. Finalement, pour estimer $\mathbb{E}(X_i)$, la fréquence de sinistres et le coût moyen sont définis.

Fréquence de sinistres

La fréquence de sinistres de l'assuré i correspond à :

$$f_i = \frac{\text{Nombre de sinistres}}{\text{Durée d'exposition}} \tag{2.4}$$

Remarque : La durée d'exposition correspond au nombre d'années de couverture du client. Par exemple, la durée d'exposition sera égale à 0,5 si l'assuré est couvert 6 mois sur une année.

Coût moyen

Le coût moyen de l'assuré i correspond à :

$$c_i = \frac{\text{Charge totale des sinistres}}{\text{Nombre de sinistres}} \tag{2.5}$$

De ce fait, la prime annuelle du client sera alors :

$$p_i = \frac{\text{Charge totale des sinistres}}{\text{Durée d'exposition}} = \text{Fréquence de sinistres} \times \text{Coût moyen} = f_i \times c_i \tag{2.6}$$

La modélisation de la fréquence et du coût moyen sera donc abordée dans ce mémoire. Comme énoncé précédemment, seule la garantie **incendie** a été étudiée. Néanmoins, la méthode reste générique et de ce fait applicable à d'autres garanties telles que le vol ou encore le dégât des eaux.

2.1.1.3 La prise en compte des sinistres graves

Il est courant, notamment en assurance non-vie, de distinguer les sinistres graves des sinistres attritionnels pour la modélisation. Les sinistres attritionnels ayant une forte fréquence et des montants faibles, les sinistres graves ayant une faible fréquence et des montants très élévés. Cette distinction permet d'éviter que les sinistres ayant une charge très élevée influencent trop le calcul des coefficients et indicateurs renvoyés par le modèle. Cette partie est étudiée en détail dans la section 2.1.4 Cette section décrit deux approches appréhendant différemment la sinistralité grave, pour calculer la prime pure d'un assuré i.

L'écrêtement des sinistres

L'écrêtement des sinistres permet d'exclure les sinistres dits "graves" de l'étude. En effet, la **sur-crête**, correspondant à la charge de sinistre au dessus d'un seuil (seuil distinguant les sinistres attritionnels des sinistres graves), est **mutualisée uniformément** sur l'ensemble des contrats sinistrés pour la garantie concernée (l'incendie dans le cadre de ce mémoire). La charge de chaque sinistre attritionnel est donc modifiée :

$$\text{Charge}_{\text{mutualisée}} = \text{Charge}_{\text{écrêtée}} \left[1 + \frac{\text{Sur-crête totale}}{\text{Sous-crête totale}}\right] \tag{2.7}$$

Ainsi, une fois la sur-crête mutualisée, il n'y a plus de distinction entre les sinistres. La prime pure est alors calculée de la manière décrite en (2.6). Cependant, cette approche fausse la charge réelle des sinistres, et de ce fait, **ne sera pas retenue dans ce mémoire**. Afin d'avoir des modèles prédictifs plus précis et robustes, une distinction entre les sinistres est conseillée.

L'utilisation d'un modèle de propension

La distinction de la sinistralité attritionnelle et grave est rendue possible par le biais de la théorie des valeurs extrêmes (explicitée dans la section 2.1.4) et d'un modèle de propension (explicité dans le chapitre 6). Effectivement, la théorie des valeurs extrêmes permet de déterminer un **seuil**, tandis qu'un modèle de propension permet d'estimer la **probabilité d'occurrence** d'un sinistre grave pour un assuré i, notée g_i. La prime pure est alors calculée de la manière suivante :

$$p_i = f_i \times [g_i \times c_i^{\text{grave}} + (1 - g_i) \times c_i^{\overline{\text{grave}}}] \tag{2.8}$$

Avec :
— c_i^{grave} : le coût moyen d'un sinistre grave pour l'assuré i ;
— $c_i^{\overline{\text{grave}}}$: le coût moyen d'un sinistre attritionnel pour l'assuré i.

On peut alors interpréter cette prime pure comme le produit de **l'espérance du nombre de sinistres** et de **l'espérance conditionnelle du coût sachant le type du sinistre** pour l'assuré i. Le type de sinistre étant attritionnel ou grave. C'est cette approche qui est **retenue dans ce mémoire**.

2.1.2 Base par image de risque

Avant d'entamer une quelconque modélisation, il est nécessaire d'obtenir une base de données de qualité qui servira de support à cette dernière. La qualité des données est un enjeu essentiel pour la modélisation car elle constitue sa matière première. Cette section détaille la méthodologie de la construction de la base **par image de risque**. Cette base compte autant de lignes que d'images de risques durant la période considérée, et non autant de lignes que de numéros de contrats.

La vie d'un contrat n'est pas monotone. En effet, ses caractéristiques, et donc le risque assuré peuvent être modifiées au cours de l'exercice considéré. C'est en effet le cas lors de **remplacements** où l'assureur ou l'assuré peuvent par exemple proposer des modifications relatives au contrat, notamment sur :
— les garanties couvertes ;
— les plafonds de garanties ou de franchises ;
— la réévaluation des capitaux assurés.

Un contrat peut bien entendu subir plusieurs remplacements. Il aura donc, pour chacun de ses remplacements, le même numéro mais pas le même risque. Voici ci-dessous une illustration du principe de changement de risque d'un contrat d'assurance Habitation :

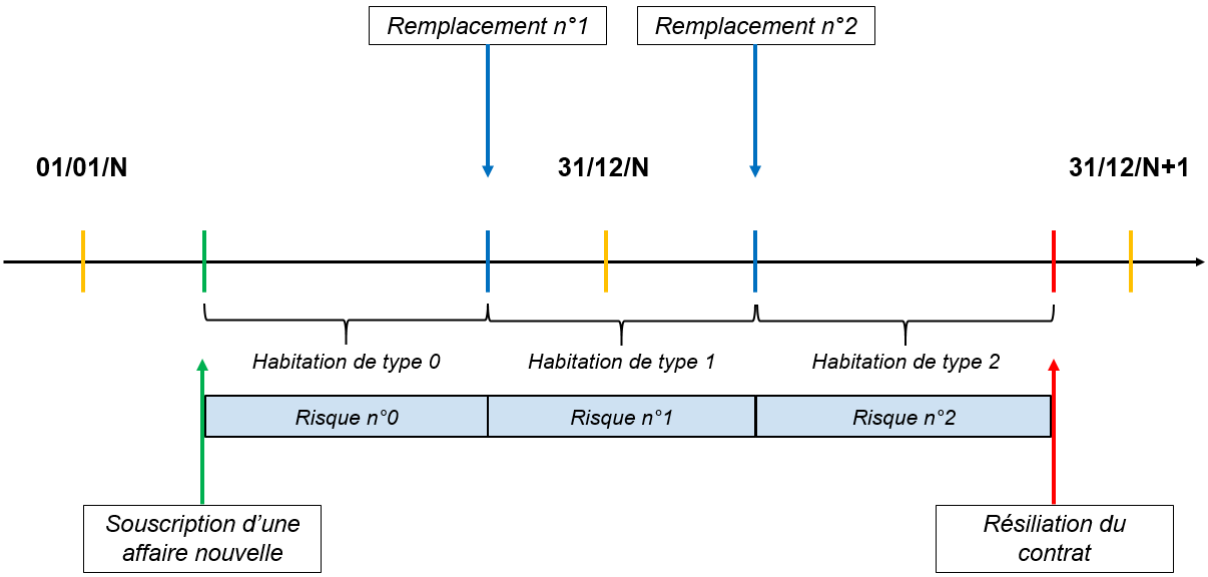


FIGURE 2.1 – Principe du changement de risque d'un contrat d'assurance

Ainsi, la base de données construite doit nécessairement contenir autant de lignes qu'il y a d'images de risques (et donc de remplacements) pour les contrats du portefeuille lors de l'exercice considéré.

Périmètre des contrats

Pour assurer une exhaustivité des données tout en faisant en sorte qu'elles représentent bien le risque actuel, notre périmètre contrats est défini de la manière suivante :

- **plage temporelle** : trois années d'observations, du 01/01/2016 au 31/12/2018 ;
- **portefeuille** : portefeuille MRH hors propriétaire non occupant, étudiant, mobil-home et grandes demeures. Les produits cités précédemment ont été exclus pour s'assurer que la base de données soit la plus homogène possible.

Présentons à présent la construction de cette base pour une année d'exercice donnée N. Dans un premier temps, nous récupérons les contrats du portefeuille **vus à chaque fin de mois de l'année N**. Nous obtenons alors 12 visions mensuelles. Cependant, à ce stade, chaque ligne correspond à la vision mensuelle d'un contrat et non à une image de risque.

Par souci de simplicité, voici un extrait de cette table pour les visions de fin de mois de novembre et décembre de l'année N :

Numéro de contrat	Date d'effet de l'affaire nouvelle	Date d'effet du remplacement	Date d'effet de résiliation	Qualité de l'occupant	Type d'habitation	Nombre de pièces	Vision
1	23/01/N			Locataire	Appartement	3	31/11/N
1	23/01/N			Locataire	Appartement	3	31/12/N
2	01/06/N			Propriétaire	Appartement	3	31/11/N
2	01/06/N	15/12/N		Propriétaire	Maison	5	31/12/N
3	10/03/N	10/11/N		Locataire	Maison	4	31/11/N
3	10/03/N	10/11/N	20/12/N	Locataire	Maison	4	31/12/N

FIGURE 2.2 – Contrats du portefeuille vus à la fin des mois de novembre et décembre de l'année N

Un contrat ayant été remplacé au moins une fois permet d'identifier au moins deux risques différents. Le risque postérieur à chaque remplacement étant identifié par la ligne contenant sa date d'effet.

Afin d'obtenir une base par image de risque, la méthode suivante a été suivie :

— **Tri ascendant** de la table par numéro de contrat, date d'effet du remplacement et date d'effet de résiliation ;

— **Dédoublonnage** de la table par le biais de la concaténation du numéro de contrat et de la date d'effet de remplacement. Ce dédoublonnage a pour effet d'identifier chaque version du contrat, et de ce fait, de conserver uniquement une ligne par image de risque ;

— Détermination de la **date de début et de fin de risque** par le biais de ces formules générales :

$$\text{Date de début de risque} = \max[01/01/\text{N}, \text{DTFAN}, \text{DTFRP}] \tag{2.9}$$

Avec **DTFAN** correspondant à la date d'effet d'affaire nouvelle, et **DTFRP** à la date d'effet de remplacement, s'il existe.

$$\text{Date de fin de risque} = \min[31/12/\text{N}, \text{DTFRS}, \text{DTFRP}^{\text{suivant}}] \tag{2.10}$$

Avec **DTFRS** correspondant à la date d'effet de résiliation, et **DTFRP$^{\text{suivant}}$** à la date d'effet du prochain remplacement, s'il existe.

— **Calcul de l'exposition** par la formule générale suivante :

$$\text{Exposition} = \frac{\text{Nombre de jours de couverture du risque durant l'année N}}{365} \tag{2.11}$$

Afin d'illustrer les résultats de la méthode explicitée ci-dessus, voici la transformation de la table affichée en figure 2.2 :

Numéro de contrat	Qualité de l'occupant	Type d'habitation	Nombre de pièces	Version	Date de début de risque	Date de fin de risque	Exposition
1	Locataire	Appartement	3	1	23/01/N	31/12/N	0,94
2	Propriétaire	Appartement	3	1	01/06/N	15/12/N	0,54
2	Propriétaire	Maison	5	2	15/12/N	31/12/N	0,05
3	Locataire	Maison	4	1	10/11/N	20/12/N	0,11

FIGURE 2.3 – Extrait de la base par image de risque

Nous obtenons ainsi une unique table telle que pour un exercice donné, nous disposons d'une seule image par risque. Néanmoins, il subsiste une limite quant à cette méthode de construction de base. Cette limite vient du fait que pour chaque contrat, seule une vision mensuelle est récupérée, et par conséquent, plusieurs images de risques sur un même mois ne peuvent être obtenues. Ainsi, les différentes images de risques des assurés ayant effectué plus d'un remplacement au cours du même mois ne sont pas distinguées dans ce mémoire. Cependant, ce cas de figure est rare et ne vient donc pas troubler notre étude.

Dans la suite, nous chercherons à compléter cette base en rattachant à chaque risque ses informations relatives à la sinistralité.

2.1.3 Jointure à la base sinistres

En marge de la base par image de risque, nous disposons d'une base "sinistres" dans laquelle sont renseignées toutes les caractéristiques des sinistres engendrés par les contrats du portefeuille.

Périmètre des sinistres

Afin d'être cohérent avec le périmètre de la base "contrats", le périmètre sinistres est défini de la manière suivante :

- **plage temporelle** : sélection des sinistres survenus entre le 01/01/2016 et 31/12/2018 ;
- **portefeuille** : portefeuille MRH hors propriétaire non occupant, étudiant, mobil-home et grandes demeures.

Également, les sinistres considérés sont tous évalués au **31/12/2019**. Premièrement, ce choix est justifié par la facilité de réconcilier cette base avec celles d'autres équipes, notamment les équipes travaillant directement avec **la comptabilité**. Effectivement, les chiffres tels que le nombre et la charge de sinistres obtenus par ces équipes sont considérés comme des références. Dans le but, d'assurer une exactitude des données sinistres obtenues, un travail de **réconciliation** des bases sinistres a donc été effectué. Évaluer l'ensemble des sinistres au 31/12/2019 a en effet facilité ce travail de réconciliation, étant donné que cette date correspond à la date de dernière évaluation des sinistres effectuée par ces équipes. Deuxièmement, ce choix de date d'évaluation des sinistres permet de maximiser le nombre de sinistres ayant une **charge définitive**, c'est à dire une charge qui ne fluctue plus dans le temps. Maximiser le nombre de sinistres **clos** permet alors d'être au plus proche de la charge totale réelle, étant donné qu'il y aura un minimum de sinistres à développer par la suite. La notion de développement de sinistres est étudiée en détail dans la section 2.1.5.

La base sinistres doit nécessairement être réconciliée et modelée afin d'obtenir la date de survenance des sinistres et leurs charges par garantie. Une fois ce travail effectué, nous joignons à la base par image de risque les sinistres par image de risque en utilisant le numéro de contrat comme clé de jointure, et en prenant en compte la date de survenance des différents sinistres. Nous obtenons alors une unique base pour laquelle nous disposons d'une seule ligne par risque avec la sinistralité (nombre et charge de sinistres par garantie) associée à ce dernier. Nous illustrons, ci-dessous, le détail de notre propos.

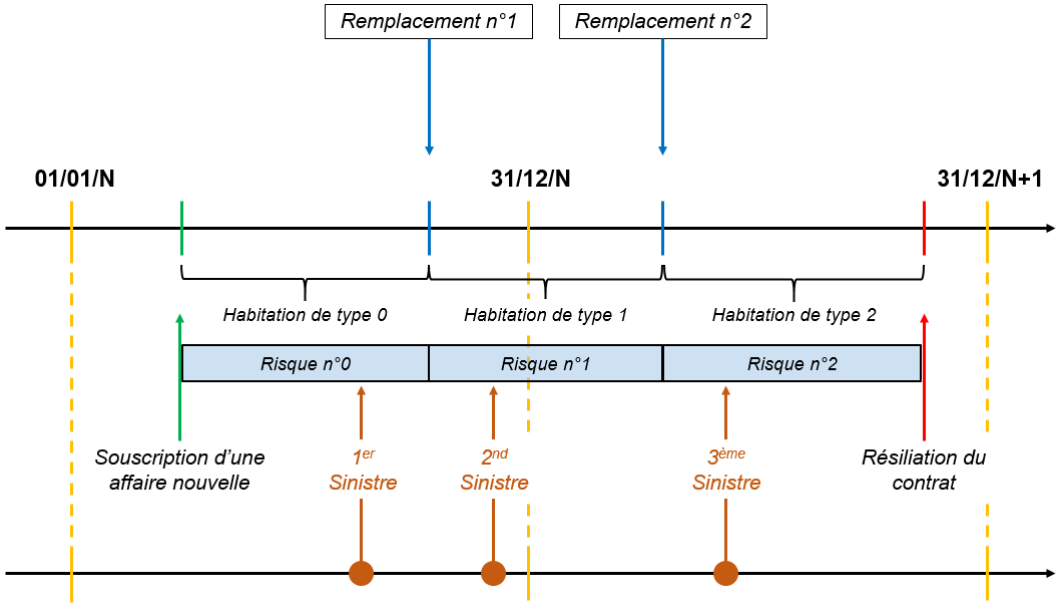


FIGURE 2.4 – Jointure des sinistres par image de risque

2.1.4 Détermination du seuil des sinistres graves

2.1.4.1 Contexte et objectifs

En assurance, un évènement défini comme extrême peut conduire à des pertes financières très importantes. Par exemple, en France, suite aux tempêtes *Lothar et Martin* de décembre 1999, près de 6,9 milliards d'euros ont été indemnisés par les assureurs, un record absolu. Cependant, si un tel évènement se produit, comment estimer les pertes financières qu'il implique ? La difficulté consiste à modéliser ces évènements très rares. Ces derniers se prêtent difficilement à l'application de la statistique classique puisque les données sont peu nombreuses voire inexistantes. Néanmoins, **la théorie des valeurs extrêmes** fournit le cadre mathématique probabiliste rigoureux pour répondre à cette problématique.

La théorie des valeurs extrêmes a pour but d'étudier et de caractériser le comportement des valeurs extrêmes d'un échantillon de variables aléatoires. Elle permet alors d'isoler ces valeurs jugées extrêmes et de les modéliser séparément du reste de la distribution. De ce fait, elle améliore leur prédiction tout comme celle des valeurs non extrêmes (qui ne sont plus perturbées par ces valeurs aberrantes). Cette distinction nécessite la définition d'un seuil à partir duquel une valeur est considérée comme extrême. Pour définir ce dernier, nous devons arbitrer entre :

— choisir un seuil **assez élevé** afin que les distributions asymptotiques soient valables ;
— choisir un seuil **le plus faible possible** afin de disposer d'un nombre d'observations suffisant.

Dans cette section, nous considérons uniquement les sinistres clos et survenus entre le 01/01/2016 et le 31/12/2018. Ainsi, nous déterminons le seuil de sinistralité grave par le biais de charges de sinistres définitives. En d'autres termes, les sinistres utilisés n'ont pas besoin d'être développés.

Lors de la dernière étude, le seuil pour la garantie **incendie** a été fixé à 188 000 €. Toutefois, la précision de ce dernier peut être nettement améliorée. En ce sens, une distinction entre la sinistralité appartements et la sinistralité maisons est effectuée dans cette section. De ce fait, l'objectif est de définir pour chaque type d'habitation le seuil qui sépare les sinistres fréquents et peu coûteux, appelés **sinistres attritionnels**, des sinistres rares et très coûteux appelés **sinistres graves**.

Cette distinction a été choisie afin d'être en accord avec les études des autres garanties. En effet, pour le dégâts des eaux par exemple, la sinistralité appartements et la sinistralité maisons ont été étudiées séparément. Nous avons conscience que cette distinction n'est pas nécessairement la plus optimale, néanmoins, ces deux types de sinistres répondent bien à des problématiques de modélisation différentes. Nous pouvons le concevoir aisément car, en pratique, le risque de sinistre au titre de la garantie incendie pour un appartement et pour une maison n'a rien de similaire.

Dans un premier temps, le cadre mathématique autour de la théorie des valeurs extrêmes est introduit. Ensuite, différentes méthodes de détermination de seuil sont présentées et appliquées sur nos données. Enfin, ces méthodes seront analysées et comparées pour décider du seuil adéquat.

2.1.4.2 Cadre mathématique

Tout d'abord, afin d'appréhender la notion de valeurs extrêmes, il est utile d'étudier la loi du maximum de la distribution. Soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées, $M_n = \max[X_1, ..., X_n]$ et $F_X(x) = \mathbb{P}(X \leq x)$. Il vient alors que la loi de M_n est :

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, ..., X_n \leq x) = \mathbb{P}(X_1 \leq x)...\mathbb{P}(X_n \leq x) = [F_X(x)]^n$$

Cependant, F n'est pas connue, et de ce fait, cette relation n'est pas utilisable. Rappelons que l'intérêt est d'étudier le comportement asymptotique de la distribution. Notons x^F, le point extrêmal de F défini par :

$$x^F = \sup\{x \in \mathbb{R} : F_X(x) < 1\}$$

Le support de F peut être borné ($x^F < \infty$) ou infini ($x^F = \infty$). $M_n \xrightarrow[n\to\infty]{\mathbb{P}} x^F$, autrement dit, la distribution asymptotique de M_n est dégénérée. Néanmoins, le **théorème de Fisher-Tippet** permet de trouver une loi non dégénérée.

Théorème 2.1 (Fisher-Tippet) *Soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées de loi F et $M_n = \max[X_1, ..., X_n]$. S'il existe :*
— *deux suites réelles $a_n > 0$ et b_n ;*
— *une fonction non dégénérée G telle que $\lim_{n\to+\infty} \mathbb{P}(\frac{M_n - b_n}{a_n} \leq x) = G,\ \forall x \in \mathbb{R}$;*

alors G est du type **GEV** *(Distribution des extrêmes généralisée).*

La loi d'extremum des extrêmes généralisée $G_{\mu,\sigma,\xi}$ est définie par la fonction de répartition

$$G_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\left(-[1+\xi(\frac{x-\mu}{\sigma})]_+^{1/\xi}\right) \text{ si } \xi \neq 0 \\ \\ \exp\left(-\exp[-(\frac{x-\mu}{\sigma})]\right) \text{ si } \xi = 0. \end{cases}$$

Avec ξ le paramètre de forme (aussi appelé paramètre de queue), μ le paramètre de position, et σ le paramètre d'échelle. Plus ξ est grand, plus le poids des extrêmes dans la distribution est important. La fonction GEV est dans le domaine d'attraction de :
— **Fréchet** si la queue de distribution est épaisse ($\xi > 0$) ;
— **Gumbel** si la queue de distribution est fine ($\xi = 0$) ;
— **Weibull** si la queue de distribution est finie à droite ($\xi < 0$).

Dans cette section, nous nous intéressons aux données X_i dépassant un seuil u défini, c'est à dire les données telles que $(X_i - u)$ soient strictement positives. Ces données sont caractérisées par des lois **GPD** (Distribution de Pareto Généralisée) définies par :

$$G^p_{\beta,\xi}(x) = \begin{cases} 1 - \left(1+\xi\frac{x}{\beta}\right)^{-1/\xi} \text{ si } \xi \neq 0 \\ \\ 1 - \exp(-\frac{x}{\beta}) \text{ si } \xi = 0. \end{cases}$$

où

$$x \geq 0 \text{ si } \xi \geq 0,$$
$$0 \leq x \leq \frac{-\beta}{\xi} \text{ si } \xi < 0.$$

Il existe une relation entre les lois GEV et GPD. En effet, les propositions suivantes sont équivalentes :
— Il existe a_n et b_n telles que $F(a_n x + b_n)^n \to G_{\mu,\sigma,\xi}(x)$;
— Il existe une fonction $a(.)$ telle que et b_n telle que :

$$\lim_{u\to x^F} \frac{\overline{F}(u + xa(u))}{\overline{F}(u)} = \begin{cases} [1+\xi x]^{-1/\xi} \text{ si } \xi \neq 0 \\ \exp(-x) \text{ si } \xi = 0. \end{cases}$$

Également, il est utile de savoir que la distribution GPD est une distribution seuil stable. Effectivement, soit $X \sim GPD(\beta, \xi)$ et u le seuil, alors

$$\mathbb{P}(X - u > x | X > x) = \frac{(1+\xi\frac{x+u}{\beta})^{-1/\xi}}{(1+\xi\frac{x}{\beta})^{-1/\xi}} = 1 + \frac{1+\xi u}{\beta + \xi x}$$

Donc pour $X - u|X \sim GPD(\beta + \xi x, \xi)$,le paramètre ξ est le même pour tout u.

2.1.4.3 Méthodes de détermination de seuil

Pour déterminer un seuil, nous exploiterons certaines propriétés des distributions de Pareto généralisée. Nous étudierons non seulement la fonction de dépassement moyen des excès, mais aussi la stabilité des paramètres d'une GPD, ainsi que l'estimateur de Hill. Toutes ces méthodes sont présentées dans cette section.

Fonction de dépassement moyen des excès

Cette méthode développée pour les distributions de Pareto généralisée par Davidson et Smith (1990) permet de sélectionner un seuil optimal. Supposons qu'une $GPD(\beta, \xi)$ soit un modèle approprié pour les charges de sinistres excédant un certain seuil donné u, alors la fonction de dépassement moyen des excès est définie par :

$$e(v) = \mathbb{E}[X - v | X > v] = \frac{\beta + \xi(v-u)}{1-\xi}, \text{pour } v > u \text{ et } \xi < 1.$$

Pour $\xi \geq 1$, la fonction est infinie. L'estimateur de cette fonction est quant à lui défini par :

$$e_n(v) = \frac{1}{N_v} \sum_{i=1}^{N_v} (X_i - v)_+$$

où N_v est le nombre de données supérieures à v.

Pour la distribution des données au dessus d'un certain seuil u, une distribution de Pareto généralisée est une approximation robuste. Ainsi, le seuil à partir duquel les sinistres peuvent être considérés comme graves correspond au montant de sinistres à partir duquel le graphique est linéaire en v (avec une pente $\frac{\xi}{1-\xi}$). En effet, cela signifierait que les montants correspondant aux sinistres graves suivent une loi de Pareto Généralisée.

Ainsi, le seuil optimal est choisi de telle sorte qu'au delà de ce seuil, la régression linéaire soit la "meilleure" possible. Pour juger de la qualité d'ajustement de cette régression, plusieurs indicateurs ont été considérés :

— Le R^2 où $R^2 = 1 - \frac{SCR}{SCT}$. Avec $SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ $SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2$.

— Le BIC où $BIC = -2\ln(L) + k\ln(n)$. Avec L la vraisemblance du modèle estimé, n le nombre d'observations dans l'échantillon et k le nombre de paramètres du modèle. Le modèle qui sera sélectionné est celui qui minimise le BIC.

A titre d'exemples, les graphiques maximisant le R^2 sont présentés pour respectivement, les appartements et les maisons :

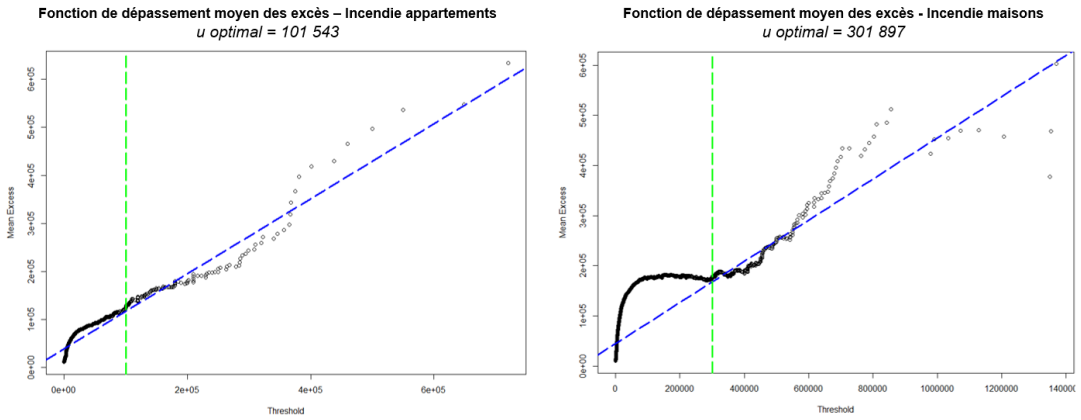


FIGURE 2.5 – Mean excess plot selon le critère du R^2 - appartements et maisons

Tout d'abord, nous remarquons aisément la différence des fonctions de dépassement moyen des excès pour les sinistres appartements et maisons. Déterminer deux seuils différents est de ce fait

nécessaire pour une étude précise. La droite verte correspond au seuil maximisant le critère du R^2, et est de ce fait, le point de départ de la régression linéaire du nuage de points, tracée en bleu. Le seuil retenu est très différents selon le type d'habitation, approximativement 100 000 € pour les appartements, et 300 000 € pour les maisons. Néanmoins, cette différence est facilement concevable, car en pratique, un incendie dans une maison a une sévérité plus forte qu'en appartement. Nous résumons les résultats obtenus dans le tableau ci-dessous :

Critère	**Appartements**	**Maisons**
Seuil R^2	101 543 €	301 897 €
% de données au delà du seuil	2,3%	0,8%
Sur-crête $\equiv$ % de charges au delà du seuil	47,8%	37%
Seuil BIC	99 608 €	299 765 €
% de données au delà du seuil	2,5%	0,8%
Sur-crête $\equiv$ % de charges au delà du seuil	48,1%	37,7%

TABLE 2.1 – Résultats obtenus par la méthode de la mean excess function

La méthode de l'étude de la fonction de dépassement moyen des excès est la plus utilisée en pratique. Cependant, il est parfois difficile de trouver un seuil précis. De plus, le graphique est perturbé par les grandes valeurs. C'est pourquoi nous comparerons cette méthode en en étudiant d'autres.

Stabilité des paramètres

Les GPD sont des distributions dites "seuil stable". En effet, cette propriété stipule que si les dépassements $X - u$ au delà d'un seuil u suivent une $GPD(\beta_u, \xi)$, alors pour tout seuil $v \geq u$, les dépassements $X - v$ suivent également une $GPD(\beta_v, \xi)$. Notons que ξ (paramètre de forme) ne dépend pas du seuil u et que le paramètre d'échelle est une fonction linéaire du seuil : $\beta_v = \beta_u + \xi(v - u)$.

Le choix du seuil se fait également de manière graphique, en analysant la stabilité du paramètre d'échelle (le paramètre de forme n'étant pas très explicite dans l'interprétation). L'estimation de ce paramètre est donc représentée pour plusieurs seuils, en incluant les intervalles de confiance à 95%. L'objectif est de retenir le plus petit seuil tel que la stabilité du paramètre d'échelle soit assurée en tenant compte de l'amplitude des intervalles de confiance pour se rendre compte de l'incertitude de mesure. Les graphiques sont représentés ci-dessous :

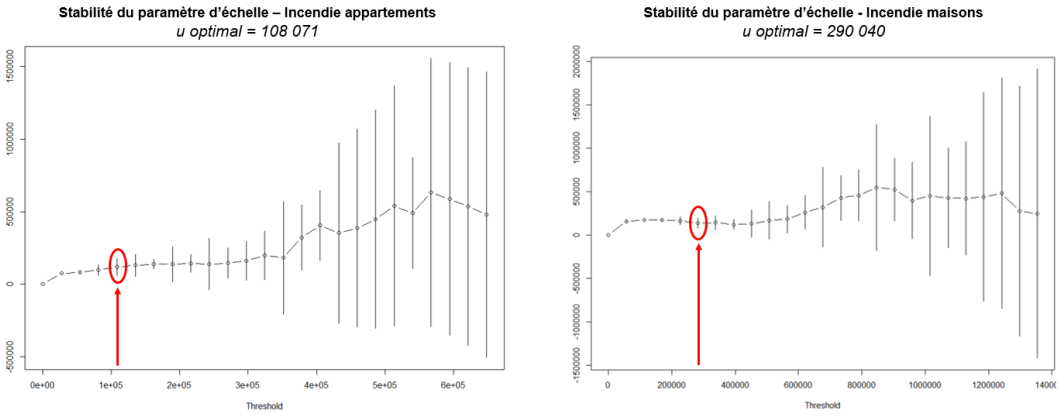


FIGURE 2.6 – Stabilité du paramètre d'échelle - appartements et maisons

Encore une fois, les seuils suggérés par cette méthode sont très différents selon le type d'habitation. Effectivement, le seuil pour les appartements suggéré est de 108 071 € tandis que le seuil

pour les maisons est de 290 040 €. Au delà de ces seuils, chacun des paramètres d'échelle devient trop instable au vu de l'étendue des intervalles de confiance.

L'estimateur de Hill

L'estimateur de Hill est le plus fréquemment utilisé lorsque $\xi > 0$. De plus, il assure un bon équilibre biais-variance. Il est défini de la façon suivante :

$$\hat{\xi}^{H}_{X,k,n} = \frac{1}{k}\sum_{j=1}^{k} \log(X_{n-j+1,n}) - \log(X_{n-k,n})$$

Avec $X_{1,n}, ..., X_{n,n}$ les statistiques d'ordre associées à l'échantillon $X_1, ..., X_n$, n le nombre d'observations, et k un entier inférieur ou égal à n.

Le graphique de l'estimateur de Hill représente la valeur de l'estimateur en fonction de l'indice k de la statistique d'ordre, soit l'estimateur construit à partir des observations supérieures ou égales à $X_{k,n}$. Chaque statistique d'ordre peut être alors reliée à un seuil. L'objectif est de trouver une zone, appelée plateau, où l'estimateur semble robuste. Le plus petit seuil u appartenant à cette zone est défini comme optimal. La méthode est illustrée par le biais des graphiques ci-dessous :

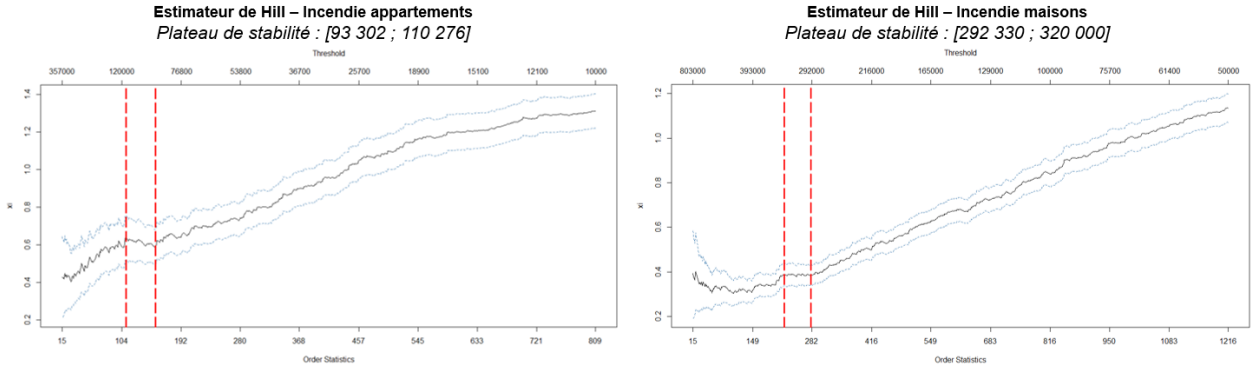


FIGURE 2.7 – Graphe de l'estimateur de Hill - appartements et maisons

Pour chaque graphique, nous observons que l'estimateur de Hill croît, puis se stabilise, avant de croître à nouveau. Comme énoncé précédemment, l'objectif est d'identifier la zone de stabilité du paramètre. De ce fait, nous identifions une zone de stabilité autour de 100 000 € pour les appartements, et autour de 300 000 € pour les maisons. Zoomons alors sur ces zones afin d'être plus précis :

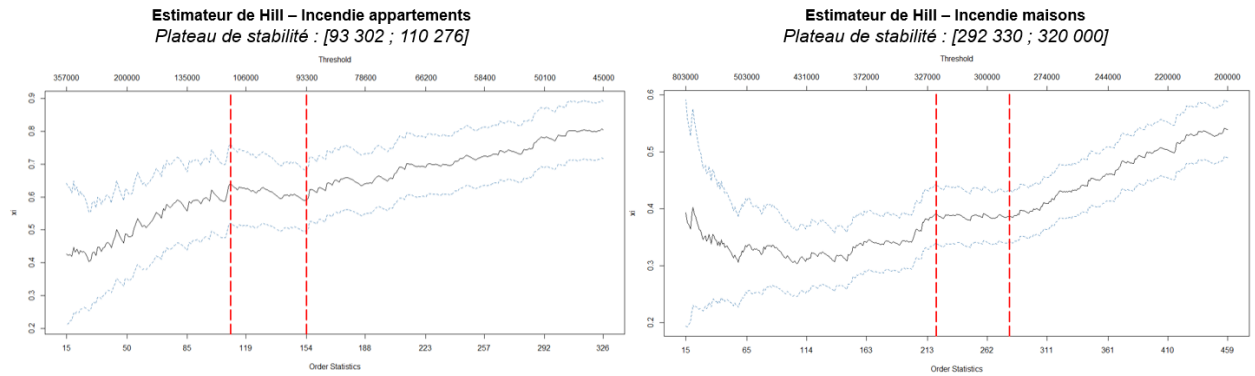


FIGURE 2.8 – Graphe de l'estimateur de Hill - Zoom - appartements et maisons

Encore une fois, les seuils suggérés par cette méthode sont très différents selon le type d'habitation. Nous résumons les résultats obtenus dans le tableau ci-dessous :

Méthode de l'estimateur de Hill	**Appartements**	**Maisons**
Seuil	93 302 €	292 330 €
% de données au delà du seuil	2,6%	0,8%
Sur-crête $\equiv$ % de charges au delà du seuil	50,7%	38,5%

TABLE 2.2 – Résultats obtenus par la méthode de l'estimateur de Hill

2.1.4.4 Résultats

Finalement, chacune des trois méthodes appliquées nous suggère des seuils relativement proches. Nous définissons, pour chaque type d'habitation, les catégories de sinistres suivantes :

- Les **sinistres attritionnels appartements** : regroupant les sinistres appartements ayant une charge inférieure à **100 000 €** ;
- Les **sinistres graves appartements** : regroupant les sinistres appartements ayant une charge supérieure à **100 000 €** ;
- Les **sinistres attritionnels maisons** : regroupant les sinistres maisons ayant une charge inférieure à **300 000 €** ;
- Les **sinistres graves maisons** : regroupant les sinistres maisons ayant une charge supérieure à **300 000 €**.

Pour s'assurer de la fiabilité des seuils sélectionnés, une étape de vérification d'adéquation à une GPD est nécessaire. En effet, pour chaque seuil u sélectionné, nous estimons par la méthode du maximum de vraisemblance une estimation de β et ξ. Ensuite, pour apprécier la qualité d'ajustement de nos excès à une GPD, un *Quantile-Plot* est analysé. Ce graphique est un nuage de points ayant pour abscisse les quantiles théoriques d'une GPD (calculés à partir des paramètres estimés), et pour ordonnée les quantiles empiriques de la distribution de nos excès. Ainsi, ce nuage de point devrait être proche de la première bissectrice si l'ajustement est de bonne qualité. La figure ci-dessous témoigne d'un très bon ajustement :

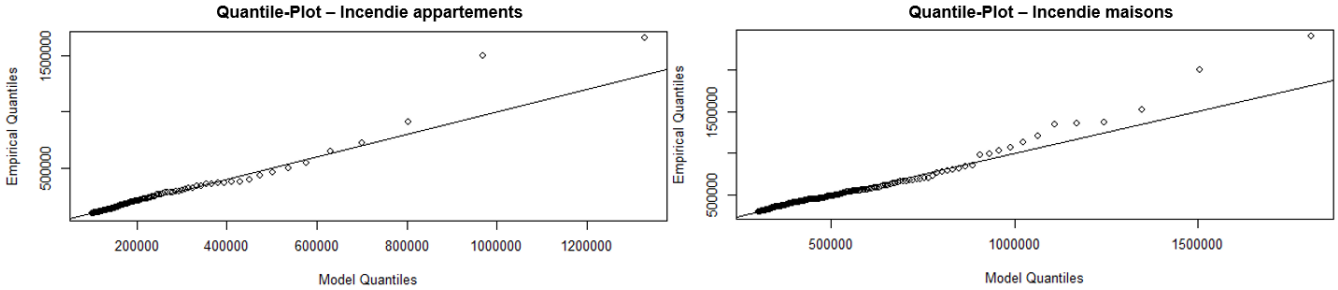


FIGURE 2.9 – Quantile-Plot - appartements et maisons

2.1.5 Développement des sinistres

2.1.5.1 Contexte et objectifs

Par définition, la prime pure correspond au montant du sinistre moyen auquel devra faire face l'assureur pour couvrir le risque. Cette espérance se calcule nécessairement sur des **nombres** et **charges** de sinistres qui doivent être définitifs. Or, ces deux caractéristiques fluctuent dans le temps. Par exemple, pour certains sinistres, notamment pour ceux avec de gros montants (tels que ceux provenant d'incendies), plusieurs expertises sont nécessaires pour évaluer le montant réel de dédommagement. Ainsi, la charge définitive que va dédommager l'assureur peut parfois être évaluée sur plusieurs années. Dans d'autres cas, l'assuré peut avoir subi un sinistre sans pour autant l'avoir encore déclaré (sinistres IBNR, Incurred but not Reported). Ce comportement implique une non prise en compte immédiate de la totalité des sinistres. Ainsi, un sinistre peut avoir plusieurs statuts :

— **En cours** : correspondant aux sinistres non clos ;
— **Clos sans suite** : correspondant aux sinistres pour lesquels il n'y a pas eu de dédommagement ;
— **Clos avec suite** : correspondant aux sinistres pour lesquels le dédommagement a été non nul.

La totalité des sinistres considérés sont survenus entre le 01/01/2016 et le 31/12/2018. De plus, nous rappelons qu'ils sont tous évalués au 31/12/2019 pour minimiser le nombre de sinistres non clos, et de ce fait maximiser le nombre de sinistres ayant une charge définitive.

L'objectif de cette section est d'évaluer pour la garantie **incendie**, la charge et le nombre définitifs des sinistres considérés pour le modèle de prime pure. La plage temporelle des sinistres considérée étant nécessairement identique à celle considérée lors de la détermination du seuil. De plus, afin de développer nos sinistres de la manière la plus précise possible, nous scindons l'étude en quatre parties :

— Le développement des sinistres attritionnels appartements ;
— Le développement des sinistres graves appartements ;
— Le développement des sinistres attritionnels maisons ;
— Le développement des sinistres graves maisons.

Comme énoncé précédemment, le temps d'évaluation peut fortement dépendre de la sévérité du sinistre. Il est alors judicieux de séparer l'étude en ces quatre parties. Cette distinction, non effectuée lors de la dernière étude, amène une précision non négligeable.

Pour ce faire, plusieurs méthodes peuvent être utilisées. Ces méthodes peuvent être aussi bien déterministes (Chain Ladder, London chain ou London pivot) que stochastiques (Mack). Nous utilisons dans cette section la méthode de Chain Ladder car elle est facile à mettre en oeuvre tout en étant fiable.

Tout d'abord, la méthode de Chain Ladder est donc présentée. Ensuite, nous l'appliquons sur nos données et déterminons les coefficients de développements associés aux quatre types de sinistres.

2.1.5.2 Chain Ladder : une méthode déterministe

La méthode de Chain-Ladder est une méthode déterministe fréquemment utilisée car elle est facile à mettre en oeuvre. Elle s'applique aussi bien aux triangles d'évaluation de la charge cumulée qu'aux triangles du nombre de sinistres cumulés. Bien que la méthode soit générique, nous la présentons à titre d'exemple uniquement pour l'évaluation de la charge de sinistres cumulés. Dans un premier temps, définissons ce triangle. On note :

— i l'année de survenance, soit l'année où le sinistre est survenu ;

— j l'année de développement, c'est à dire le nombre d'années après l'année de survenance où la charge est évaluée. Si i est l'année de survenance et j l'année de développement, alors l'évaluation est effectuée l'année calendaire $i+j$;
— $C_{i,j}$ l'évaluation de la charge de sinistres effectuée j année(s) après l'année i, pour tous les sinistres survenus l'année i.

Ce triangle d'évaluation de la charge cumulée se présente sous la forme suivante :

Année de survenance	**Devpt 0**	**Devpt 1**	**...**	**Devpt j**	**...**	**Devpt** $n-i$	**...**	**Devpt** $n-1$	**Devpt** n
Surv 0	$C_{0,0}$	$C_{0,1}$	...	$C_{0,j}$	...	$C_{0,n-i}$	...	$C_{0,n-1}$	$C_{0,n}$
Surv 1	$C_{1,0}$	$C_{1,1}$	...	$C_{1,j}$	...	$C_{1,n-i}$	...	$C_{1,n-1}$	**?**
...	...	...	...	...	...	...	...	**?**	**?**
Surv i	$C_{i,0}$	$C_{i,1}$	...	$C_{i,j}$	...	$C_{i,n-i}$	**?**	**?**	**?**
...	...	...	...	...	...	**?**	**?**	**?**	**?**
Surv $(n-1)$	$C_{n-1,0}$	$C_{n-1,1}$	**?**	**?**	**?**	**?**	**?**	**?**	**?**
Surv n	$C_{n,0}$	**?**	**?**	**?**	**?**	**?**	**?**	**?**	**?**

TABLE 2.3 – Triangle d'évaluation de la charge cumulée

La valeur de chaque cellule du tableau correspond à $C_{i,j}$ donc à l'évaluation de la charge de sinistres cumulée l'année j, pour tous les sinistres survenus l'année i. Les " ?" représentent les évaluations des charges de sinistres cumulées que nous devons estimer. Le but du développement des sinistres est donc de trouver une méthode pour estimer ces charges inconnues. Dans cette section, la méthode est celle de Chain Ladder.

Les rapports $\frac{C_{i,j+1}}{C_{i,j}}$ sont appelés facteurs de développement. La méthode de Chain Ladder repose sur une hypothèse forte. En effet, les facteurs de développement sont supposés indépendants de l'année d'origine i, c'est à dire pour $j = 0, ..., n-1$ fixé

$$\frac{C_{i,j+1}}{C_{i,j}} := f_j, \text{ pour tout } i.$$

Pour valider cette hypothèse, on vérifie si approximativement

$$\frac{C_{0,j+1}}{C_{0,j}} = \frac{C_{1,j+1}}{C_{1,j}} = ... = \frac{C_{n-j-1,j+1}}{C_{n-j-1,j}} = \text{constante}.$$

Dans ce cas, ces rapports devraient approximativement être égaux à $\frac{\sum_{i=0}^{n-j-1} C_{i,j+1}}{\sum_{i=0}^{n-j-1} C_{i,j}}$. Ainsi, pour estimer les facteurs de développement nous procédons de la manière suivante :

1. Calculer pour tout j
$$\hat{f}_j = \frac{\sum_{i=0}^{n-j-1} C_{i,j+1}}{\sum_{i=0}^{n-j-1} C_{i,j}}$$
2. En déduire une estimation de l'évaluation de la charge de sinistres cumulée, pour tout $j = 0, ..., n$
$$\hat{C}_{i,j} = \hat{f}_{j-1}\hat{f}_{j-2}...\hat{f}_{j-i}C_{i,j-i}$$
En effet, supposons que l'année en cours soit l'année n, alors pour l'année i, on connaît les évaluations pour $n-i$ années après. Donc
$$\hat{C}_{i,j} = \hat{f}_{j-1}\hat{C}_{i,j-1} = \hat{f}_{j-1}\hat{f}_{j-2}\hat{C}_{i,j-2} = ... = \hat{f}_{j-1}\hat{f}_{j-2}...\hat{f}_{j-i}C_{i,j-i}$$

À présent, appliquons cette méthode à nos données afin d'obtenir les coefficients de développement associés à chaque type de sinistres.

2.1.5.3 Application

En premier lieu, le développement des sinistres doit être basé sur un historique de sinistralité suffisamment profond. Cette démarche est nécessaire pour s'assurer de leur bon développement. En effet, sans cela, leurs charges et nombres définitifs ne peuvent pas être estimés de façon optimale. Nous avons donc pris en compte les sinistres survenus entre 2005 et 2019, ce qui représente 14 années d'historique. On note :

— i l'année de survenance $\in \{2005, 2006, ..., 2019\}$;
— j l'année de développement $\in \{0, 1, ..., 14\}$;
— $C_{i,j}$ l'évaluation de la charge de sinistres effectuée j année(s) après l'année i, pour tous les sinistres survenus l'année i.

Comme évoqué précédemment, l'étude a été scindée en quatre parties :
— Le développement des sinistres attritionnels appartements ;
— Le développement des sinistres graves appartements ;
— Le développement des sinistres attritionnels maisons ;
— Le développement des sinistres graves maisons.

Cette distinction est d'autant plus importante que le temps d'évaluation de la charge du sinistre dépend fortement de sa sévérité. Ainsi, les seuils déterminés dans la section 2.1.4 ont été appliqués en ce sens. La méthode de détermination des charges et nombres de sinistres définitifs étant générique, seul le développement de l'évaluation de la charge des sinistres attritionnels maisons est présenté dans cette section. Néanmoins, la totalité des résultats est communiquée.

Calcul des facteurs de développement

Les facteurs de développement sont calculés grâce au triangle d'évaluation de la charge cumulée. Nous présentons les résultats :

Triangle de charges cumulées - Atrritionnels Maisons - Incendie (M€)									
Surv / Devpt	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**...**	**14**
2005	28,98	50,46	50,75	51,06	51,32	51,39	51,42	...	51,18
2006	34,31	51,32	52,68	52,76	52,52	52,63	52,67	...	*52,47*
2007	32,18	54,11	54,30	53,95	53,78	53,92	53,99	...	*53,74*
2008	37,19	59,49	61,66	62,01	61,51	62,01	61,88	...	*61,51*
...	...	...	...	...	...	...	...	...	...
2016	40,04	66,07	66,82	**66,99**	*66,98*	*67,16*	*67,17*	...	*67,00*
2017	45,32	68,29	**67,58**	*67,73*	*67,72*	*67,91*	*67,92*	...	*67,75*
2018	39,28	**65,10**	*66,06*	*66,22*	*66,21*	*66,39*	*66,40*	...	*66,23*
2019	40,22	*65,09*	*66,06*	*66,21*	*66,20*	*66,38*	*66,39*	...	*66,22*

Facteurs de développement							
0/1	**1/2**	**2/3**	**3/4**	**4/5**	**5/6**	**...**	**13/14**
1,62	1,01	1,00	1,00	1,00	1,00	...	1,00

FIGURE 2.10 – Facteurs de développement

Grâce à la méthode de Chain Ladder, des facteurs de développement ont pu être calculés, et de ce fait, une charge à l'ultime pour chaque année de survenance a pu être estimée. De plus, les facteurs de développement sont très proches de 1 après 2 ans de développement. En pratique, cela signifie que la charge de sinistres est correctement évaluée après 2 ans. Les cases "fluos" correspondent aux montants des sinistres considérés pour le modèle de prime pure. Nous voyons qu'ils sont tous évalués au 31/12/2019.

Proportion de la charge à l'ultime

Ensuite, pour chaque année de développement, la proportion de la charge à l'ultime doit nécessairement être évaluée. Ainsi, un tableau est construit de telle sorte que chacune de ces cellules soit définie de la manière suivante :

$$\text{Prop}_{i,j} = \frac{C_{i,j}}{C_{i,n}}$$

En d'autres termes, la proportion de la charge à l'ultime pour l'année de survenance i évaluée après j année(s) est le rapport entre : la charge pour l'année de survenance i évaluée après j année(s) et la charge pour l'année de survenance i évaluée après n années. Nous synthétisons ces propos à travers la figure ci-dessous :

Proportion de la charge à l'ultime - Attritionnels Maisons									
Surv / Devpt	0	1	2	3	4	5	6	...	14
2005	57%	99%	99%	100%	100%	100%	100%	...	100%
2006	65%	98%	100%	101%	100%	100%	100%	...	
2007	60%	101%	101%	100%	100%	100%	100%	...	
2008	60%	97%	100%	101%	100%	101%	101%	...	
...	...	...	...	...	...	...	...	...	
2016	60%	99%	100%	**100%**					
2017	67%	101%	**100%**						
2018	59%	**98%**							
2019	61%								

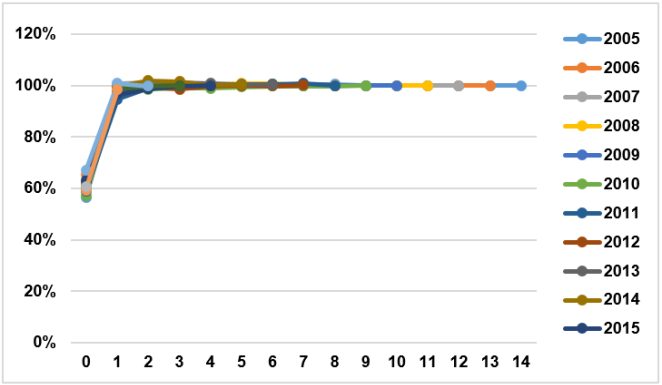


FIGURE 2.11 – Proportion de la charge à l'ultime par année de développement

En liaison avec les facteurs de développement, le constat est identique. La charge définitive est évaluée pratiquement à 100% après 2 années de développement. En effet, sur le graphique ci-dessus, nous remarquons que quelle que soit l'année de survenance, la proportion de la charge à l'ultime se stabilise autour de 100% après 2 années. L'importance du périmètre sinistres de départ ici est évident. Plus nous considérons les sinistres avec une date d'évaluation avancée, moins les sinistres sont à développer. De ce fait, nous sommes plus proches de la charge réellement dédommagée. De plus, la charge des sinistres survenus en 2018 évaluée le 31 décembre 2019 est légèrement sous-estimée. Les coefficients de développement viendront en toute logique augmenter ce montant.

Calcul des coefficients de développement

Pour rappel, l'objectif est d'évaluer la charge définitive des sinistres considérés pour le modèle de prime pure. Ainsi, nous calculons uniquement les coefficients de développement pour les sinistres survenus en 2016, 2017 et 2018, tous évalués au 31/12/2019. Voici la définition afin d'obtenir les coefficients de vieillissement pour chacune des années de survenance $i \in \{2016,\ 2017,\ 2018\}$:

$$\text{Coeff}_{i,2019} = \frac{\hat{C}_{i,14}}{C_{i,2019-i}} = \frac{1}{\text{Prop}_{i,2019-i}}$$

Voici les résultats obtenus pour les sinistres attritionnels maisons :

Année de survenance	Charge vue au 31/12/2019 (M€)	Charge développée (M€)	Coeff. dev.
2016	67,00	67,00	**1,00**
2017	67,58	67,75	**1,00**
2018	65,10	66,23	**1,02**

TABLE 2.4 – Coefficients de développement - Attritionnels maisons

Nous retrouvons bien des coefficients de développement supérieurs ou égaux à 1, qui viennent augmenter les charges vues au 31/12/2019 sous-estimées.

Synthèses des résultats

Voici les résultats obtenus en charges et en nombres pour chaque type de sinistre :

Année de survenance	**Charge au 31/12/2019**		**Charge vieillie**		**Coeff. dev.**	
Appartements	Attri.	Graves	Attri.	Graves	Attri.	Graves
2016	12,03	8,11	12,18	9,42	**1,01**	**1,16**
2017	11,08	5,15	11,40	6,75	**1,03**	**1,31**
2018	9,48	5,37	10,36	9,97	**1,09**	**1,86**
Maisons	Attri.	Graves	Attri.	Graves	Attri.	Graves
2016	67,00	25,47	67,00	26,72	**1,00**	**1,05**
2017	67,58	32,28	67,75	36,61	**1,00**	**1,13**
2018	65,10	18,38	66,23	29,56	**1,02**	**1,61**

TABLE 2.5 – Coefficients de développement - Résultats charges

Année de survenance	**Nombre au 31/12/2019**		**Nombre vieilli**		**Coeff. dev.**	
Appartements	Attri.	Graves	Attri.	Graves	Attri.	Graves
2016	2 014	38	2 009	40	**1,00**	**1,07**
2017	1 965	29	1 957	33	**1,00**	**1,14**
2018	1 985	20	1 963	29	**0,99**	**1,47**
Maisons	Attri.	Graves	Attri.	Graves	Attri.	Graves
2016	10 811	52	10 786	54	**1,00**	**1,03**
2017	10 471	69	10 428	75	**1,00**	**1,09**
2018	11 508	45	11 394	67	**0,99**	**1,48**

TABLE 2.6 – Coefficients de développement - Résultats nombres

Nous nous apercevons que quel que soit le type de sinistres, les charges non vieillies sont sous-estimées. Cette tendance est d'autant plus marquée pour les sinistres graves. Cette sous-estimation implique un coefficient de développement supérieur à 1. Concernant le nombre de sinistres, nous nous apercevons que quel que soit le type d'habitation, il est globalement sous-estimé. Du fait de la différence entre les coefficients, nous voyons ici tout l'intérêt de scinder l'étude par typologie de sinistres. Sans cette distinction, le développement des sinistres aurait été moins précis.

Finalement, les coefficients obtenus sont affectés aux sinistres **non clos** considérés pour le modèle de prime pure, autrement dit les sinistres non clos survenus en 2016, 2017 et 2018. Ainsi, nous avons à présent les charges et nombres de sinistres à l'ultime.

2.2 Statistiques descriptives

Une fois les étapes détaillées précédemment effectuées, nous obtenons une base où une ligne correspond à une image de risque. Ce risque est défini grâce aux informations qui concernent :

- **le contrat d'assurance** : date d'effet d'affaire nouvelle, réseau de distribution, prime, exposition... ;
- **l'habitation** : adresse, nombre de pièces, qualité de l'occupant, type d'habitation, ancienneté du logement, montant de capital assuré... ;
- **l'assuré** : genre, âge, ancienneté, catégorie socioprofessionnelle... ;
- **les sinistres** : nombre et charge totale des sinistres incendie durant la période du risque.

Étant donné que l'objectif de ce mémoire est la création d'un modèle de prime pure pour la garantie incendie, nous nous intéressons dans cette section à la fréquence et au coût moyen observés des sinistres considérés de l'étude. Pour rappel, ces deux indicateurs sont définis dans la section 2.1.1.2. De plus, cette analyse statistique permet de visualiser la répartition du portefeuille tout en quantifiant son risque. Pour ce faire, nous étudions la fréquence et le coût moyen selon certaines variables de la base de données construite.

Année d'exercice

Les trois années d'exercice correspondent aux trois années considérées de l'étude. Tout d'abord, nous justifions à travers ces graphiques que le risque incendie est un risque d'intensité. En effet, la fréquence de sinistre aux alentours de **0,4%** traduit une fréquence faible, tandis que le coût moyen aux alentours de **10 000 €** traduit un dédommagement fort. Ensuite, l'idée est d'observer si notre portefeuille est plus ou moins risqué selon l'année. En 2018, la fréquence augmente légèrement (**0,44%**) tandis que le coût moyen diminue faiblement (**9 600 €**). Le profil de risque du portefeuille a donc évolué depuis la dernière modélisation. C'est d'ailleurs une des raisons pour lesquelles nous avons effectué une refonte des modèles de primes pures. Également, nous observons une tendance à la baisse de l'exposition, traduisant un apport net négatif (cf. 1.4).

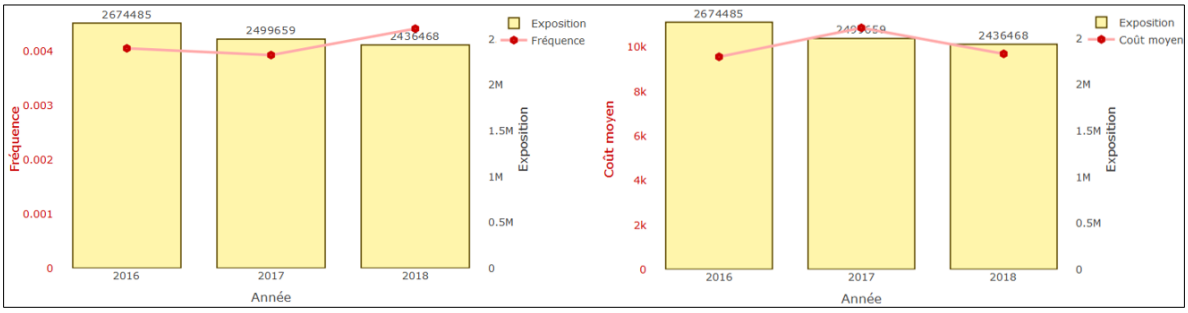


FIGURE 2.12 – Fréquence et coût moyen observés par année

Qualité de l'occupant et type d'habitation

La qualité de l'occupant indique le statut de l'assuré, qui peut être propriétaire ou locataire, alors que le type d'habitation indique si l'habitation est un appartement ou une maison. La variable ainsi étudiée est le croisement des deux précédemment définies. Tout d'abord, nous observons que les propriétaires de maisons composent majoritairement notre portefeuille, suivi des locataires d'appartements, des propriétaires d'appartements et des locataires de maisons. Globalement, la fréquence de sinistres est plus élevée pour les maisons que pour les appartements. En outre, les propriétaires de maisons ont une fréquence 3,5 fois plus élevée que celle des locataires de maisons, faisant d'eux le segment ayant la plus forte fréquence (**0,69%**). Les locataires de maisons possèdent quant à eux le coût moyen le plus élevé à **15 900 €**, tandis que les autres segments sont aux alentours de 10 000 €.

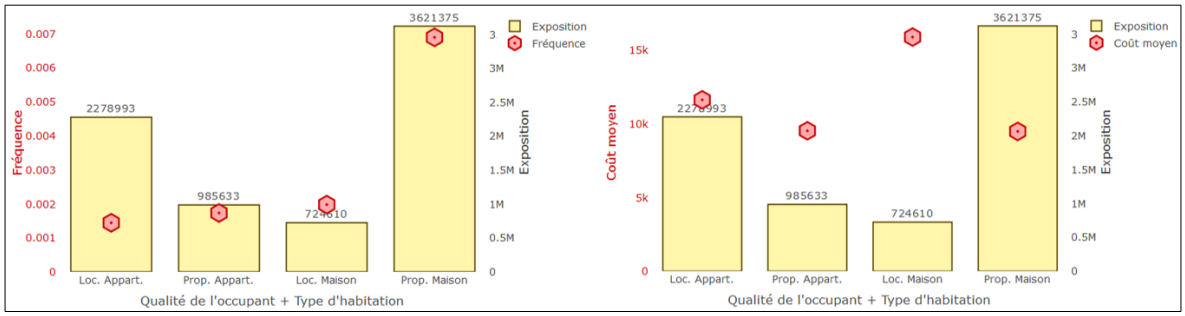


FIGURE 2.13 – Fréquence et coût moyen observés par qualité de l'occupant et type d'habitation

Nombre de pièces

Nous pouvons voir que la fréquence de sinistre croît avec le nombre de pièces de l'habitation, ce qui à priori semble cohérent dans cette étude marginale de la sinistralité. En effet, l'idée est que plus une habitation possède de pièces, plus elle est grande, et de ce fait, plus elle a de chances de voir se déclencher un incendie. Nous notons une fréquence de **1,1%** pour les habitations de 9 pièces et plus. Concernant le coût moyen, nous observons un point d'inflexion en 6 pièces, ce qui à priori peut être contre intuitif.

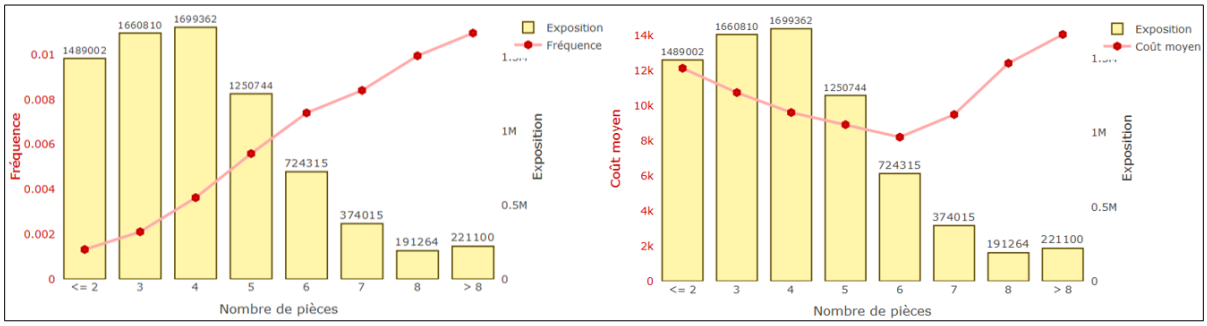


FIGURE 2.14 – Fréquence et coût moyen observés par nombre de pièces

Présence d'un insert

Un insert peut être défini comme un générateur de chaleur ajouté à l'habitation. Les inserts les plus fréquents sont les cheminées et les poêles à bois. L'idée est d'observer si l'installation de ce matériel influe sur la sinistralité incendie. Logiquement, nous constatons que la fréquence et le coût moyen sont plus élevés lorsqu'un insert est présent. La fréquence est de **0,76%** lorsque l'habitation possède un insert, ce qui est 2,2 fois plus élevé que la fréquence des habitations n'en possédant pas. Concernant le coût moyen, il est 1,25 fois plus élevé lorsque l'habitation comporte un insert (**11 600 €**).

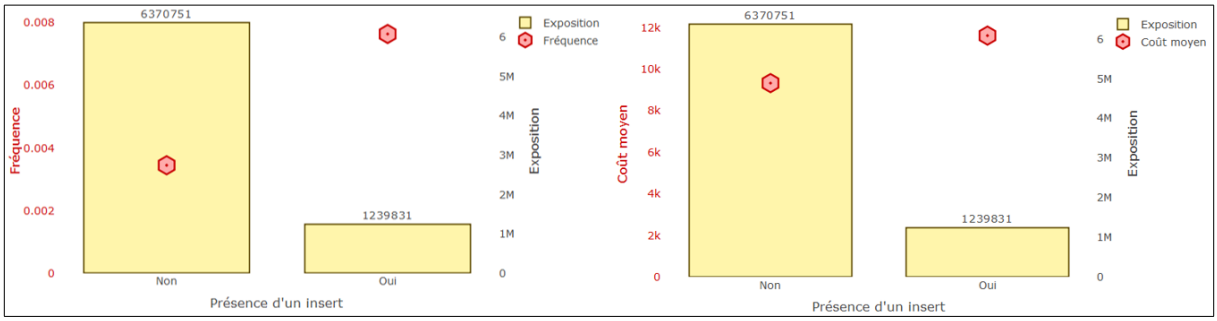


FIGURE 2.15 – Fréquence et coût moyen observés avec et sans insert

Cette brève analyse statistique montre qu'au sein de notre portefeuille, il existe des groupes de risques très différents. Effectivement, le nombre de pièces, le type d'habitation, la qualité de

l'occupant ou encore la présence d'un insert influent fortement sur la sinistralité incendie. Il est donc nécessaire de ventiler le portefeuille en groupe de risques homogènes, afin de créer un modèle de prime pure le plus robuste possible. Sans cela, la segmentation de la population serait mauvaise et impliquerait une tarification non précise.

2.3 Audit du modèle actuel

Comme nous avons pu le voir dans la section 2.2, le profil de risque du portefeuille évolue dans le temps, ce qui modifie la fréquence et le coût moyen au cours des années. Par conséquent, le modèle de prime de pure doit être ajusté régulièrement afin de quantifier le risque du portefeuille actuel. Effectivement, le portefeuille actuel n'a pas le même profil de risque que le portefeuille de 2014 sur lequel le dernier modèle de prime pure a été effectué.

Dans cette section, un audit du dernier modèle produit est réalisé. Cette étape est importante dans le sens où nous pourrons apprécier sa qualité en l'appliquant sur des données récentes, tout en permettant de juger de l'importance de la refonte du modèle de prime pure pour la garantie incendie. Cet audit consiste en l'étude de deux indicateurs permettant de juger de la qualité d'un modèle de tarification :

1. **L'erreur totale**, qui permet d'évaluer **la qualité d'ajustement** du modèle :

$$\text{Erreur totale} = \frac{\sum \text{Valeurs prédites} - \sum \text{Valeurs réelles}}{\sum \text{Valeurs réelles}}$$

   Cet indicateur permet de constater si le tarif calculé aurait permis à la compagnie de faire face à ses engagements en considérant l'ensemble du portefeuille 2016, 2017 et 2018.

2. **L'indice de Gini**, qui permet d'évaluer **la capacité de segmentation** du modèle :

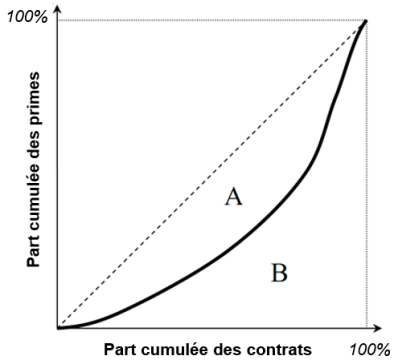


FIGURE 2.16 – Courbe de Lorenz et indice de Gini

$$\text{Indice de Gini} = \frac{\text{Aire A}}{\text{Aire A} + \text{Aire B}}$$

   Calculé à partir de la courbe de Lorenz, l'indice de Gini est généralement utilisé pour déterminer le niveau de répartition des richesses au sein d'une population. Néanmoins, il peut être utilisé dans un autre contexte comme nous le faisons ici. Effectivement, une tarification uniforme possède un indice de Gini de 0 étant donné que la courbe de Lorenz correspondrait à la bissectrice. Ainsi, plus l'indice de Gini est proche de 1, plus le tarif est segmenté.

Nous présentons les résultats de l'audit :

Profil	Erreur totale	Indice de Gini
Global	**21,04%**	**33,71%**
Segment		
Loc. Appart.	31,61%	31,74%
Loc. Maison	86,18%	32,88%
Prop. Appart.	36,46%	26,60%
Prop. Maison	12,13%	31,43%

TABLE 2.7 – Synthèse - Audit du modèle

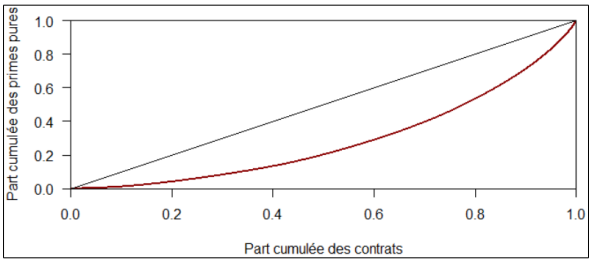


FIGURE 2.17 – Gini - Portefeuille global

Le tableau de synthèse renseigne les indicateurs obtenus lorsque le dernier modèle de prime pure est appliqué sur le portefeuille global, ainsi que sur chacun de ses segments qui sont les locataires d'appartements, les locataires de maisons, les propriétaires d'appartements et les propriétaires de maisons.

Tout d'abord, l'erreur totale de 21,04% sur le portefeuille global traduit d'une surestimation de la sinistralité. Or, par définition, la prime pure est l'espérance des pertes auxquelles devra faire face l'assureur, et de ce fait, l'erreur totale devrait en toute théorie être nulle. De plus, l'indice de Gini de 33,71% montre que le tarif pourrait être segmenté davantage. La figure 2.17 représente graphiquement cet indice avec pour abscisse la part cumulée de contrats et pour ordonnée la part cumulée des primes pures. Nous remarquons que la courbe de Lorenz est assez distante du triangle inférieur, ce qui révèle un tarif trop uniforme.

Les conclusions sont identiques quelque soit le segment. Effectivement, les erreurs totales sont toutes positives, montrant une surestimation de la sinistralité. Notons que les locataires de maisons ont la surestimation la plus élevée à 86,18%. De plus, les différents indices de Gini indiquent un tarif encore trop uniforme.

Pour conclure, les erreurs totales présentées traduisent d'une imprécision dans le tarif tandis que les indices de Gini montrent une capacité de segmentation qui peut être améliorée. L'interprétation de ces indicateurs est très importante car elle permet d'apprécier la robustesse du modèle ainsi que sa capacité à identifier les différents profils de risques au sein de notre portefeuille. Par conséquent, le dernier modèle de prime pure pour la garantie incendie n'optimise pas le tarif. Une refonte du modèle est de ce fait nécessaire afin d'améliorer les aspects détaillés précédemment.

Chapitre 3

Retraitements des variables

3.1 Traitement des valeurs manquantes

3.1.1 Introduction aux valeurs manquantes

Avant de commencer la phase de modélisation, il est essentiel de vérifier la qualité des données. L'obtention de données précises et exhaustives permettra d'obtenir une tarification au plus proche du risque auquel l'assurance est exposée, tandis que des données approximatives impliqueraient une incertitude au niveau du risque modélisé. Or, le traitement des valeurs manquantes est au coeur de cette problématique. Effectivement, ces dernières peuvent, introduire une quantité importante de biais, rendre le traitement et l'analyse des données plus laborieux, et réduire l'efficacité des modèles. De ce fait, la manière dont elles sont traitées est cruciale pour l'analyse et les conclusions qui seront tirées. Les valeurs manquantes peuvent être de trois types :

1. Missing Completely At Random **(MCAR)** : correspond aux valeurs manquantes d'origine complètement aléatoire. C'est le cas lorsque le fait qu'une donnée soit manquante pour une variable Y est complètement indépendant de la variable Y ou des autres variables X de la base. Ce type de valeurs manquantes n'introduit pas de biais dans l'analyse, mais dans la pratique, elles sont très rares ;
2. Missing At Random **(MAR)** : correspond aux valeurs manquantes d'origine non aléatoire et étant liées à une variable X qui contient une information. Ce type de données manquantes peut introduire un biais dans l'analyse, surtout si elles déséquilibrent notre base en raison de nombreuses valeurs manquantes pour une certaine catégorie ;
3. Missing Not At Random **(MNAR)** : correspondant aux valeurs manquantes qui ne sont ni MCAR ni MAR. C'est le cas lorsque le fait qu'une donnée soit manquante dépend de la variable Y elle même ou de variables non observées.

Plusieurs approches permettant d'imputer ces trois types de valeurs manquantes ont été appliquées, puis comparées dans cette étude telles que :

— L'imputation par les plus proches voisins **(KNN Imputation)** ;
— L'imputation par la moyenne et la médiane pour les variables quantitatives ;
— L'imputation par la classe la plus représentée pour les variables qualitatives ;

Après avoir considéré les avantages et les inconvénients de chacune des approches, et parce qu'elle a obtenu les résultats les plus statistiquement fiables, l'imputation par les plus proches voisins a été retenue. Elle est donc présentée et appliquée dans cette section.

3.1.2 L'imputation par les plus proches voisins

Pourquoi utiliser les plus proches voisins ?

La *KNN Imputation* est une méthode non paramétrique très profitable permettant d'estimer la valeur manquante d'une observation en trouvant ses K plus proches voisins dans l'espace des variables. Cette méthode peut être utilisée pour des données continues, discrètes, catégorielles, et ordinales, ce qui la rend particulièrement pratique pour traiter tous types de données.

L'idée sous-jacente à l'utilisation de la *KNN Imputation* est qu'une valeur manquante d'une observation peut être approchée par les valeurs des observations qui lui sont les plus proches, en fonction d'autres variables. Dans le cas d'une variable catégorielle, la valeur manquante sera imputée par la classe majoritaire parmi les K plus proches voisins. Dans le cas d'une variable numérique, la valeur manquante sera imputée par la médiane parmi les K plus proches voisins.

Comment mesurer la distance entre les observations dans l'espace des variables ?

Avant d'estimer la valeur manquante d'une observation, il faut trouver ses K plus proches voisins. De ce fait, plusieurs distances sont calculées afin de définir les observations proches de l'observation contenant la valeur manquante. Pour les variables numériques, il est commun d'utiliser la distance euclidienne (correspondant à la norme l^2). Cependant, dans notre cas, l'information de variables catégorielles doit également être pris en compte. Par conséquent, la distance euclidienne seule ne suffit pas pour calculer la distance entre deux observations. Néanmoins, la **distance de Gower** permet de tenir compte à la fois des variables numériques et catégorielles pour calculer une distance. Elle a donc été retenue dans cette étude. Soit un espace à p variables, la distance de Gower pour deux observations i et j est alors définie par :

$$D_G^{i,j} = \frac{\sum_{k=1}^{p} W_k^{i,j} \cdot d_k^{i,j}}{\sum_{k=1}^{p} W_k^{i,j}}$$

Avec $d_k^{i,j}$ la distance entre 2 observations sur la variable k :
- Pour une variable catégorielle k, $d_k^{i,j} = 0$ si la valeur est la même, sinon $d_k^{i,j} = 1$.
- Pour une variable numérique k, $d_k^{i,j} = 0$ si la valeur est la même, sinon $d_k^{i,j} = \frac{|x_i - x_j|}{\max(x) - \min(x)}$. Le dénominateur permet de rapporter les distances à la même échelle.

La pondération $W_k^{i,j} = 1$ si les observations i et j contiennent des données pour la variable k, sinon, $W_k^{i,j} = 0$. Les K plus proches voisins de l'observation i sont alors les observations ayant les K plus faibles distances de Gower par rapport à l'observation i.

Application - âge du client

La méthode étant similaire quelque soit la variable, seul un exemple sur *l'âge du client* est présenté. Néanmoins, toutes les valeurs manquantes de la base ont été imputées par cette méthode. L'exemple est présenté ci-dessous :

Base initiale				
Observation	**Habitation**	**Qualité**	**Nb pièces**	**Age du client**
1	Appartement	Locataire	2	36
2	Maison	Propriétaire	4	54
3	Maison	Locataire	3	51
4	*Appartement*	*Locataire*	*3*	
5	Appartement	Propriétaire	4	45

FIGURE 3.1 – Base initiale

Distance de Gower - Observation 4	
$D_G^{4,1}$	**0,17**
$D_G^{4,2}$	0,83
$D_G^{4,3}$	0,33
$D_G^{4,5}$	0,5

TABLE 3.1 – Distance de Gower

Comme nous pouvons le constater, *l'âge du client* pour l'observation 4 est manquant. Le nombre de voisins étant fixé à $K = 1$ pour l'exemple, on assigne à cette observation l'âge du client de l'observation qui minimise la distance Gower. Le résultat est présenté dans la figure suivante :

Base imputée				
Observation	Habitation	Qualité	Nb pièces	Age du client
1	Appartement	Locataire	2	36
2	Maison	Propriétaire	4	54
3	Maison	Locataire	3	51
4	*Appartement*	*Locataire*	*3*	*36*
5	Appartement	Propriétaire	4	45

FIGURE 3.2 – Base imputée

En l'occurrence, on assigne un âge de **36** ans. Nous voyons dans ce simple exemple que l'imputation par les plus proches voisins est statistiquement plus fiable qu'une imputation par la moyenne, la médiane ou la classe la plus représentée au sein de la population totale.

Avant propos

Les propos des sections 3.2 et 3.3 concernent uniquement la base permettant de réaliser des modèles linéaires généralisés (ces derniers étant présentés en section 4.1). Effectivement, pour faciliter leur interprétabilité, il est préférable que la base sur laquelle ils sont créés contienne uniquement des variables explicatives catégorielles et non corrélées. Ainsi, ces aspects sont détaillés dans les sections suivantes. Toutefois, ces traitements de variables ne sont pas effectués sur la base permettant de réaliser des arbres de décisions. De fait, ces derniers assurent une bonne gestion de la corrélation et de la continuité des variables explicatives.

3.2 Discrétisation des variables

3.2.1 Justification de la discrétisation

Une fois que le processus de construction de la base de modélisation a été détaillé (section 2.1), les variables au sein de cette dernière sont étudiées. Effectivement, deux types de variables sont présentes : celles dites qualitatives et celles dites quantitatives. Dans cette section, nous cherchons à discrétiser l'ensemble des variables quantitatives. En d'autres termes, nous voulons uniquement des variables qualitatives dans la base de modélisation. L'idée sous-jacente à l'utilisation de variables qualitatives est de segmenter le portefeuille selon celles-ci, afin de constituer des sous-portefeuilles dans lesquels les risques peuvent être considérés comme indépendants et de même loi. De plus, regrouper les variables quantitatives par classes permet de prendre en compte certains effets non-linéaires.

Avant de présenter la discrétisation par arbres de régression, justifions cette méthode. Effectivement, nous étudions dans cette section l'impact de la discrétisation des variables quantitatives sur la sinistralité. Pour ce faire, une variable indicatrice est créée : $Y = \mathbb{1}_{\text{Nb sinistres} \geq 1}$.

Tout d'abord, nous réalisons une régression logistique (méthode présentée en annexes) de la variable Y sur une variable quantitative X donnée. Conjointement, nous calculons $\mathbb{P}(Y) = 1$ pour chacune des classes de la variable X. Pour le moment, la discrétisation de la variable X est issue d'une discrétisation par quantile. A titre d'exemple, la figure 3.3 confronte la régression logistique ainsi que la probabilité observée selon le *nombre de pièces*.

Tout d'abord, nous remarquons que lorsque la variable *nombre de pièces* est considérée comme catégorielle, un effet non linéaire est pris en compte. De plus, lorsque que le *nombre de pièces* est considérée comme une variable quantitative, la régression logistique surestime $\mathbb{P}(Y) = 1$ pour les habitations de 1 à 3 pièces. Au contraire, elle sous-estime la même probabilité pour les habitations de 4 à 8 pièces. Une discrétisation semble donc nécessaire.

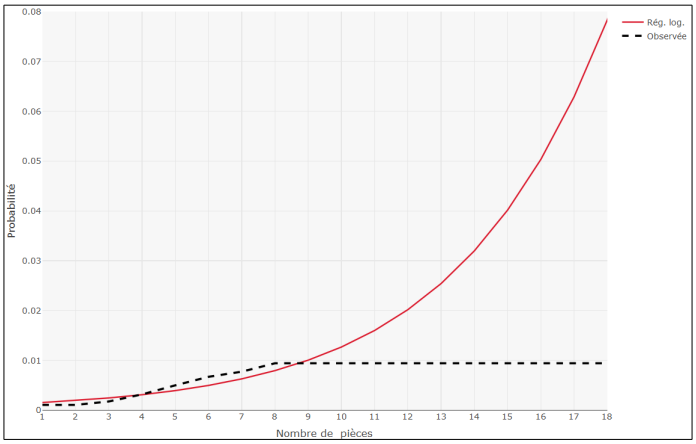


FIGURE 3.3 – Régression logisitique et probabilité observée

La discrétisation par quantile à l'avantage de fournir suffisamment d'exposition au sein de chaque classe créée. De ce fait, aucune classe n'est sous ou sur représentée. Néanmoins, cette discrétisation a le défaut de créer des classes de risques non homogènes en terme de sinistralité. C'est pour ces différentes raisons que nous utilisons les arbres de régression pour la discrétisation. Effectivement, cette méthode permet à la fois d'obtenir suffisamment d'exposition au sein de chaque classe, et de créer des classes de risques homogènes.

3.2.2 Arbre de régression

Comme explicité précédemment, afin de créer des classes de risques homogènes, nous cherchons à régresser par un arbre, la variable $Y = \mathbb{1}_{\text{Nb sinistres} \geq 1}$ sur chaque variable quantitative X. Le lecteur est invité à lire la section 4.2.1 qui présente théoriquement les arbres de régression.

Nous présentons la discrétisation de la variable *nombre de pièces* par un arbre de régression. Dans ce cadre, la dimension de l'espace des variables est égal à 1, car il n'y a qu'une seule variable explicative. De plus, le critère d'arrêt considéré est que chaque région finale doit nécessairement contenir au moins 10% des observations totales. Ce dernier a été choisi afin de ne pas obtenir une classe de risque sous représentée. Aussi, les moyennes de la variable réponse Y au sein des différentes régions définies permettent de quantifier le risque de chacune des classes.

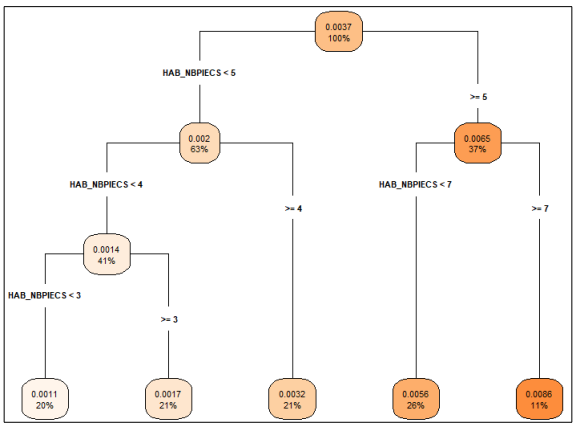


FIGURE 3.4 – Discrétisation du nombre de pièces par arbre de régression

Ainsi, nous remarquons que les habitations ayant 7 pièces ou plus ont une sinistralité presque 8 fois plus élevée que les habitations ayant deux pièces ou moins. L'intérêt de créer des classes avec une sinistralité homogène prend ici tout son sens. De plus, l'arbre de régression effectué permet d'obtenir des classes ayant une exposition suffisante. Dans la suite, la totalité des variables quantitatives (discrètes ou continues) ont été groupées par classe de risque homogènes de cette manière là.

3.3 Étude des corrélations

3.3.1 Justification de l'étude des corrélations

La corrélation fait référence à la situation dans laquelle deux variables sont liées. Lors de la création de modèles linéaires généralisés, la présence de corrélation au sein des variables explicatives peut poser des problèmes. Effectivement, il serait compliqué de séparer les effets individuels de chaque variable explicative sur la variable réponse. De plus, la précision des estimations des coefficients associés aux variables corrélées serait très fortement réduite. Cette étude est donc essentielle et permettra de sélectionner les variables à intégrer dans le modèle.

Il existe plusieurs mesures pour quantifier la corrélation entre deux variables telles que :
— Le ρ de Pearson : permettant de déterminer si deux variables **quantitatives** sont dépendantes l'une de l'autre ;
— Le V de Cramer : permettant de déterminer si deux variables **qualitatives** sont dépendantes l'une de l'autre ;

Pour rappel, toutes les variables explicatives de la base permettant de réaliser les GLMs ont été discrétisées. De ce fait, nous utiliserons dans cette section le V de Cramer.

3.3.2 Le V de Cramer

La mesure du V de Cramer se base sur le test d'indépendance du χ^2. Cependant, contrairement au χ^2, il reste stable si l'on augmente la taille de l'échantillon dans les mêmes proportions inter-modalités. Plus V est proche de zéro, plus il y a indépendance entre les deux variables étudiées. Au contraire, il vaut 1 en cas de complète dépendance. Il est alors défini de la manière suivante :

$$V = \sqrt{\frac{\chi^2}{n[\min(l,c)-1]}}$$

Où n est le nombre d'observations, et l et c sont respectivement le nombre de lignes et de colonnes du tableau de contingence.

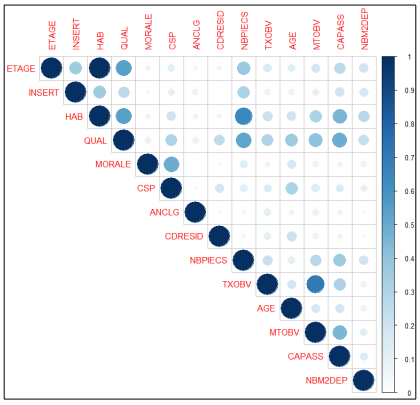


FIGURE 3.5 – V de Cramer

Lorsque nous sommes face à un problème de corrélation entre deux variables, deux solutions simples peuvent être choisies. La première est de supprimer l'une des variables corrélées. La seconde est de combiner les variables entre elles. Ainsi, suite à l'analyse du V de Cramer, nous supprimons la variable *TXOBV* (correspondant au taux d'objets de valeur parmi le capital assuré) car elle est très fortement corrélée avec la variable *MTOBV* (correspondant au montant d'objets de valeur). Aussi, nous décidons de combiner les variables *HAB* et *ETAGE* correspondant respectivement au type d'habitation (appartement/maison) et à l'étage de l'appartement.

Chapitre 4

Aspects théoriques

4.1 Les modèles linéaires généralisés

La méthode classique de tarification en assurance non-vie est l'utilisation de modèles linéaires généralisés, notés **GLMs**. Ainsi, au cours de cette étude, ces modèles sont créés et analysés afin de construire la prime pure de la garantie incendie. Nous les présentons donc théoriquement dans cette section.

4.1.1 Présentation générale

Des modèles linéaires aux modèles linéaires généralisés

Il convient de considérer la matrice $X_{n,p}$ contenant les n observations de p variables explicatives, et $Y_{n,1}$ la matrice réponse. L'objectif est de trouver la fonction μ telle que $\mathbb{E}[Y|X] = \mu(X)$. Les modèles linéaires considèrent que la fonction μ est linéaire et que le bruit est gaussien. Or, deux limites peuvent être identifiées :

- **La forme de la fonction μ :** considérer une forme linéaire est dans de nombreux cas trop restrictif ;
- **La loi que suit le bruit :** la loi gaussienne peut ne pas être adaptée aux données.

De plus, la tarification non-vie fait souvent l'objet d'une modélisation de coûts étant à valeurs dans $\mathbb{R}^+$ ou bien de nombres de sinistres étant à valeurs dans $\mathbb{N}$. Or, le support de Y dans le cadre des modèles linéaires est $\mathbb{R}$. De ce fait, pour pallier à ces différentes limites, les GLMs ont été introduits.

La famille de lois exponentielles

Un modèle statistique $(\Omega, \mathbb{F}, (\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi>0})$ est appelé famille exponentielle si les probabilités $\mathbb{P}_{\theta,\phi}$ admettent une densité f par rapport à une mesure dominante avec

$$f_{\theta,\phi} = c_\phi(y) \exp\left(\frac{y\theta - a(\theta)}{\phi}\right)$$

.

- θ s'appelle le paramètre canonique et ϕ le paramètre de dispersion, souvent considéré comme un paramètre de nuisance. Un paramètre de nuisance est un paramètre qui n'est pas d'intérêt immédiat mais qui doit être pris en compte dans l'analyse des paramètres d'intérêt ;
- $a(\theta)$ est de classe C^2 et convexe ;

— $c_\phi(y)$ ne dépend pas de θ.

4.1.2 Composantes d'un GLM

Trois composantes peuvent être identifiées au sein des GLMs :

1. **Une composante aléatoire :** caractérisée par la loi de probabilité de la variable Y à expliquer. Le principe est d'identifier cette loi, sous hypothèse que celle-ci appartienne à la famille de lois exponentielles ;
2. **Une composante déterministe :** caractérisée par la matrice des variables explicatives $X_{n,p}$, et un vecteur de coefficients noté β ;
3. **Une relation fonctionnelle :** établissant la relation entre la composante aléatoire et la composante déterministe. Cette relation est assurée par le biais d'une fonction de lien g, supposée monotone et différentiable au moins une fois :

$$g(\mathbb{E}[Y|X]) = X\beta$$

Si l'on choisit $g = log$, nous obtenons le modèle suivant pour l'individu i :

$$\mathbb{E}[Y_i|X_i] = \exp\left(X_i^T\beta\right) = \prod_{j=1}^{p} \exp\left(X_{i,j}\beta_j\right)$$

S'agissant d'un modèle multiplicatif et donc simple à mettre en place, le log est en pratique la fonction de lien la plus souvent utilisée.

En résumé, un modèle statistique peut être considéré comme un modèle linéaire généralisé s'il vérifie les hypothèses suivantes :

— $Y|X = x \sim \mathbb{P}_{\theta,\phi}$ appartient à la famille de lois exponentielles ;
— $g(\mu(X)) = g(\mathbb{E}[Y|X]) = X\beta$ pour une certaine fonction g bijective, appelée fonction de lien.

Avantages GLM	**Inconvénients GLM**
Ils permettent de conserver la simplicité des modèles linéaires tout en autorisant une forme plus générale.	La procédure d'estimation n'est efficace que si la vraie loi conditionnelle appartient à cette famille exponentielle.
Les coefficients sont estimés par maximisation d'une vraisemblance qui provient d'une famille de lois exponentielles (qui peuvent ne pas être gaussiennes).	Il y a une liberté sur le choix de la forme de $\mathbb{E}[Y]$ à travers la fonction de lien. Mais, ce choix est souvent imposé par un certain choix canonique correspondant à la famille exponentielle choisie.
	Utiliser une famille exponentielle impose de ne pas avoir de valeurs extrêmes.

TABLE 4.1 – Avantages et inconvénients des GLMs

4.1.3 Estimation des paramètres

L'estimation du vecteur des paramètres β se fait par la méthode du maximum de vraisemblance. Fixons les notations suivantes :

$$\begin{cases} \eta_i = X_i^T \beta \\ \mu_i = \mathbb{E}[Y_i|X_i] = g^{-1}(X_i^T\beta) = g^{-1}(\eta_i) \\ \theta_i = (a')^{-1}(g^{-1}(\eta_i)) \end{cases}$$

Dans le cas des modèles exponentiels, la log-vraisemblance est donnée par :

$$l(\beta) = \sum_{i=1}^{n} log\left[f(Y_i;\beta,\phi)\right] = \sum_{i=1}^{n} \left\{ log(c_\phi(Y_i)) + \frac{Y_i\theta_i - a(\theta_i)}{\phi} \right\}$$

En utilisant la dérivée d'une composée de fonction, nous obtenons $\frac{\partial \theta_i}{\partial \beta_j}$, et de ce fait :

$$\frac{\partial}{\partial \beta_j} l(\beta) = \frac{1}{\phi} \sum_{i=1}^{n} \frac{x_{i,j}(Y_i - \mu_i)}{g'(\mu_i)a''(\theta_i)}$$

Où $x_{i,j}$ est la j$^{\text{ème}}$ coordonnée de X_i. Soit D la matrice diagonale dont les coefficients sont égaux à $\frac{1}{g'(\mu_i)a''(\theta_i)}$, alors $\frac{\partial}{\partial \beta_j} = 0$ pour tout $j = 1, ..., p \iff X'D(Y-\mu) = 0$. Par conséquent, l'estimateur du maximum de vraisemblance est solution de

$$X'D(Y - g^{-1}(X\beta)) = 0$$

Les p équations (j allant de 1 à p) se résolvent numériquement, par exemple par une descente de gradient.

4.1.4 Pénalisations

Dans le cadre d'un modèle linéaire généralisé, nous venons de voir que l'estimation du vecteur des paramètres β se fait en maximisant la quantité suivante :

$$l(\beta) = \sum_{i=1}^{n} \left\{ log(c_\phi(Y_i)) + \frac{Y_i\theta_i - a(\theta_i)}{\phi} \right\}$$

Toutefois, pour estimer les coefficients, aucune pénalisation sur le nombre de variables présentes dans le modèle n'est prise en compte ici. Or, un nombre de variables trop élevé implique souvent une variance du modèle trop forte, et de ce fait, un sur-apprentissage des données. Ce sur-apprentissage peut être traduit comme le fait que le modèle ajuste très bien les données sur lesquelles il a été créé, et ajuste très mal dès lors que les données lui sont nouvelles. De plus, réduire le nombre de variables permet d'améliorer l'interprétation des modèles, qui dans un contexte de tarification, est très important afin d'identifier la typologie des assurés à risque.

De ce fait, pour répondre à ces problématiques, des méthodes permettant d'estimer les coefficients des GLMs tout en pénalisant le nombre de variables sont considérées dans cette étude. Nous en présentons trois dans cette section : la régression **Ridge**, la régression **LASSO** et **l'Elastic Net**.

La régression Ridge

La régression **Ridge** est très similaire à la méthode du maximum de vraisemblance, à l'exception que les coefficients sont estimés en maximisant une quantité légèrement différente. Le vecteur des paramètres de **Ridge** β^R est celui qui maximise la quantité suivante :

$$\sum_{i=1}^{n}\left\{log(c_\phi(Y_i)) + \frac{Y_i\theta_i - a(\theta_i)}{\phi}\right\} - \lambda\sum_{i=1}^{p}\beta_j^2 = l(\beta) - \lambda\left\|\beta\right\|_2$$

Où $\lambda \geq 0$ est un paramètre de réglage (ou tuning parameter en anglais), qui doit être déterminé séparément. Nous verrons dans la section 4.3.3 comment ce paramètre est déterminé de façon optimale. Le second terme de la quantité : $\lambda\left\|\beta\right\|_2$, fonctionne comme une pénalisation dans le sens où ce terme est faible lorsque les coefficients β_j sont proches de 0. Ainsi, la valeur de λ permet de contrôler l'impact de ce second terme sur l'estimation des coefficients.

La régression Ridge permet de conserver l'ensemble des variables explicatives du modèle, même si elles sont très proches de 0, contrairement aux méthodes *backward stepwise et forward stepwise* de sélection de variables. Cependant, bien que conserver l'ensemble des variables du modèle améliore sa prédiction, cela complexifie son interprétation dans le cas où il y a un nombre de variables trop élevé.

La régression LASSO

La régression **LASSO** est une alternative à Ridge qui permet justement de réduire le nombre de variables du modèle. Effectivement, cette méthode force certains coefficients à 0, lorsque les variables explicatives associées n'ont pas de lien significatif avec la variable réponse. Le vecteur des paramètres de **LASSO** β^L est celui qui maximise la quantité suivante :

$$\sum_{i=1}^{n}\left\{log(c_\phi(Y_i)) + \frac{Y_i\theta_i - a(\theta_i)}{\phi}\right\} - \lambda\sum_{i=1}^{p}|\beta_j| = l(\beta) - \lambda\left\|\beta\right\|_1$$

L'Elastic Net

De ces deux méthodes apparaît une nouvelle qui les combine : la méthode **Elastic Net**. Ici, le vecteur des paramètres de l'**Elastic Net** β^{EN} est celui qui maximise la quantité suivante :

$$\sum_{i=1}^{n}\left\{log(c_\phi(Y_i)) + \frac{Y_i\theta_i - a(\theta_i)}{\phi}\right\} - \lambda\left(\alpha\sum_{i=1}^{p}|\beta_j| + (1-\alpha)\sum_{i=1}^{p}\beta_j^2\right) = l(\beta) - \lambda[\alpha\left\|\beta\right\|_1 + (1-\alpha)\left\|\beta\right\|_2]$$

Avec $\alpha \in [0;1]$ qui contrôle la part distribuée entre la pénalité l_1 et l_2. Ainsi, nous pouvons voir que la régression **Ridge** et la régression **LASSO** sont des cas particuliers de **l'Elastic Net**, respectivement pour $\alpha = 0$ et $\alpha = 1$.

4.2 Les arbres de décision

Dans cette étude, les GLMs sont comparés à des méthodes innovantes de tarification que sont les arbres de décision. Effectivement, ces méthodes permettent également de construire une prime pure, tout en évitant de supposer des hypothèses fortes telles que la linéarité du lien entre les variables explicatives et la variable réponse. Nous les présentons donc théoriquement dans cette section.

4.2.1 Arbre de régression

Notons Y de la manière suivante : $Y = f(X) + \epsilon$, où ϵ est un bruit centré. Dans le cas de la régression, c'est à dire lorsque la variable Y est continue, nous avons $f(X) = \mathbb{E}[Y|X = x]$. L'intérêt de la modélisation est d'estimer f. Dans ce cadre, la régression par arbre est une méthode non paramétrique, dans le sens où elle ne fait pas d'hypothèse sur la forme de la fonction f.

Un arbre de régression segmente l'espace des variables explicatives en un nombre fini de régions. Pour ensuite prédire la valeur d'une observation donnée, la moyenne de la variable réponse Y des observations de la région est appliquée. Présentons à présent le processus de construction d'un arbre de régression. Voici les deux étapes majeures :

1. L'espace des variables est divisé en K régions distinctes et non chevauchantes : $R_1, R_2, ..., R_K$;
2. Pour toutes les observations étant comprises dans la région R_k, la même prédiction est effectuée : la moyenne de la variable Y des observations de la base, comprises dans la région R_k.

En théorie, les régions construites peuvent être de toutes formes. Dans le cas d'un espace à une dimension (une seule variable explicative) les régions correspondent à des segments. Ces dernières sont construites afin de minimiser la somme des carrés résiduels (notée RSS) donnée par :

$$RSS = \sum_{k=1}^{K} \sum_{i \in R_k} (y_i - \hat{y}_{R_k})^2$$

Où $\hat{y}_{R_k}$ est la moyenne de Y pour les observations au sein de la région k. Malheureusement, d'un point de vue informatique, il n'est pas possible de considérer toutes les partitions possibles de l'espace des variables. C'est pour cette raison qu'est utilisé un découpage de l'espace des variables dit binaire récursif.

Soit p le nombre de variables explicatives et $j \in \{1, ...p\}$. Dans le but d'effectuer ce découpage binaire récursif, on sélectionne dans un premier temps la variable X_j et le seuil s tel que le découpage de l'espace des variables en les régions $\{X|X_j < s\}$ et $\{X|X_j \geq s\}$ amène la plus grande réduction du RSS. En d'autres termes, nous considérons l'ensemble des variables explicatives $X_1, ..., X_p$ ainsi que l'ensemble des valeurs de leurs seuils, pour sélectionner la variable et le seuil créant l'arbre ayant le plus petit RSS. De ce fait, $\forall$ j et s, nous définissons la paire suivante :

$$R_1(j, s) = \{X|X_j < s\} \text{ et } R_2(j, s) = \{X|X_j \geq s\}$$

L'objectif est alors de trouver j et s qui minimisent l'équation suivante :

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

Où $\hat{y}_{R_1}$ est la moyenne de Y pour les observations comprises dans $R_1(j, s)$ et $\hat{y}_{R_2}$ la moyenne de Y pour les observations comprises dans $R_2(j, s)$. Ensuite, nous réitérons le processus qui est de segmenter l'espace en cherchant la variable explicative et le seuil associé minimisant le RSS au sein

de chacune des régions. En effet, cette fois-ci, au lieu de considérer un découpage sur l'ensemble de l'espace des variables, nous segmentons une des deux régions précédemment définies. A ce stade, trois régions sont ainsi créées. Par la suite, nous découpons l'une des trois régions définies afin de minimiser le RSS. Le processus continue jusqu'à ce qu'un critère d'arrêt soit atteint.

Une fois les régions $R_1, ..., R_J$ créées, la prédiction pour une observation donnée est la moyenne de Y pour les observations comprises dans la région à laquelle elle appartient.

4.2.2 Arbre de classification

Un arbre de classification est très similaire à un arbre de régression, à l'exception que la variable réponse est qualitative. Rappelons que dans le cadre d'un arbre de régression, la prédiction pour une observation est donnée par la moyenne de Y des observations appartenant à la région définie. Dans le cadre d'un arbre de classification, la prédiction pour une observation, sera la classe la plus représentée (appelée aussi le mode) de la région dans laquelle elle appartient.

La méthode de construction d'un arbre de classification est très proche de celle d'un arbre de régression. Effectivement, un découpage récursif binaire est également utilisé pour découper l'espace des variables. Néanmoins, dans le cadre de la classification, le RSS ne peut pas être utilisé comme critère pour découper l'espace. Pour ce faire, l'indice de Gini est utilisé. Il est défini de la manière suivante :

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Où $\hat{p}_{mk}$ représente la proportion d'observations de la région m appartenant à la classe k. On voit facilement que l'indice de Gini prend de petites valeurs lorsque $\hat{p}_{mk}$ est proche de 1 ou de 0. C'est pour cette raison que cet indice est défini comme une mesure de qualité de noeuds, car des petites valeurs indiquent que le noeud contient essentiellement des observations d'une seule classe.

Avantages des arbres simples	**Inconvénients des arbres simples**
Ces arbres sont plus simples à interpréter et à expliquer que les modèles linéaires généralisés. Ces arbres peuvent être affichés graphiquement, et facilement interprétables par un non-expert. Ces arbres gèrent facilement tous types de variables explicatives.	En cas de trop grosse profondeur, ces arbres souffrent d'une trop grande variance. En effet, un léger changement dans les données peut engendrer un grand changement sur l'arbre finalement estimé.

TABLE 4.2 – Avantages et inconvénients des arbres de décisions simples

Dans de nombreux cas, les arbres de décisions simples souffrent d'une trop grande variance. Cependant, en agrégeant plusieurs arbres de décisions, par des méthodes de *Bagging*, de forêts aléatoires (*Random Forests* en anglais), ou encore de *Boosting*, la qualité de prédiction des arbres peut être nettement améliorée. Nous introduisons ces concepts dans les sections suivantes.

4.2.3 Random Forest

Les forêts aléatoires permettent de créer des arbres décorrélés tout en générant un grand nombre d'arbres par la méthode de rééchantillonnage *Bootstrap* (le lecteur est invité à se référer aux annexes pour obtenir une présentation théorique du *Bootstrap*). De plus, cette méthode d'agrégation d'arbres répond à la fois aux problèmes de régression et de classification.

Lorsque les arbres sont construits, à chaque fois qu'un noeud est considéré, un sous-ensemble aléatoire m des p variables explicatives est choisi comme candidat pour le découpage de l'espace des variables. Généralement, m est choisi de telle sorte que $m = \sqrt{p}$ (nous verrons par la suite comment optimiser ce paramètre). En d'autres termes, lors de la construction d'une forêt aléatoire, à chaque noeud d'un arbre, l'ensemble des variables explicatives n'est pas considéré pour couper l'espace. Bien que contre-intuitif, il y a une explication rationnelle à cette méthode.

Effectivement, supposons qu'il y ait dans notre base de modélisation, une variable explicative expliquant très fortement notre variable réponse, notée p_1. Par conséquent, la majorité des arbres construits utiliseront p_1 pour découper l'espace dans le premier noeud. Ainsi, la plupart des arbres construits se ressembleront, et les prédictions de ces derniers seront fortement corrélées. C'est pourquoi les forêts aléatoires permettent de contourner ce problème en forçant chaque noeud à considérer seulement un sous ensemble de variables explicatives.

La figure 4.1 illustre le processus de prédiction d'une forêt aléatoire dans le cadre d'un problème de classification. Effectivement, à chaque arbre de décision construit, la classe majoritaire de la région dans laquelle l'observation appartient est associée. Ensuite, un ultime vote par majorité est effectué dans le but de combiner les résultats de l'ensemble des arbres. Ainsi, la prédiction sera la classe majoritaire parmi l'ensemble des classes associées à chaque arbre de décision. Dans le cadre d'un problème de régression, le processus est le même, à l'exception que la classe majoritaire est remplacée par la moyenne des observations.

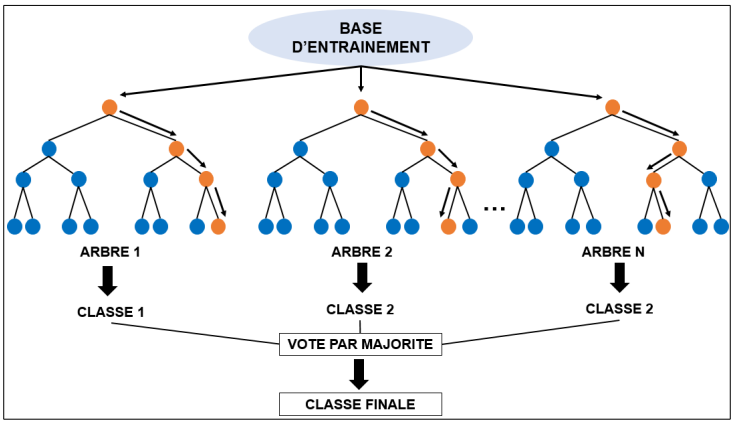


FIGURE 4.1 – Prédictions des forêts aléatoires

4.2.4 eXtreme Gradient Boosting

XGBoost (e**X**treme **G**radient **Boost**ing), est une méthode d'apprentissage supervisé qui agrège un ensemble d'arbres de décision, tout comme le Random Forest. Individuellement, ces arbres ont une faible qualité de prédiction, mais lorsqu'ils sont regroupés, ils peuvent être très performants.

La différence entre XGBoost et Random Forest réside dans la façon dont les arbres de décision sont construits et combinés. Effectivement, XGBoost construit des arbres très courts et simples de façon itérative. Ainsi, chaque arbre apprend faiblement des données. Pour commencer, XGBoost crée un premier arbre simple qui a très probablement de faibles performances. Ensuite, il construit un autre

arbre qui est entraîné à prédire ce que le premier arbre n'a pas pu prédire, lui-même étant un arbre qui apprend peu des données. L'algorithme continue ainsi en construisant séquentiellement un certain nombre d'arbres de décision, chacun corrigeant l'arbre précédent jusqu'à ce qu'une condition d'arrêt soit atteinte, telle que le nombre d'arbres à construire. La figure 4.2 illustre les propos précédents :

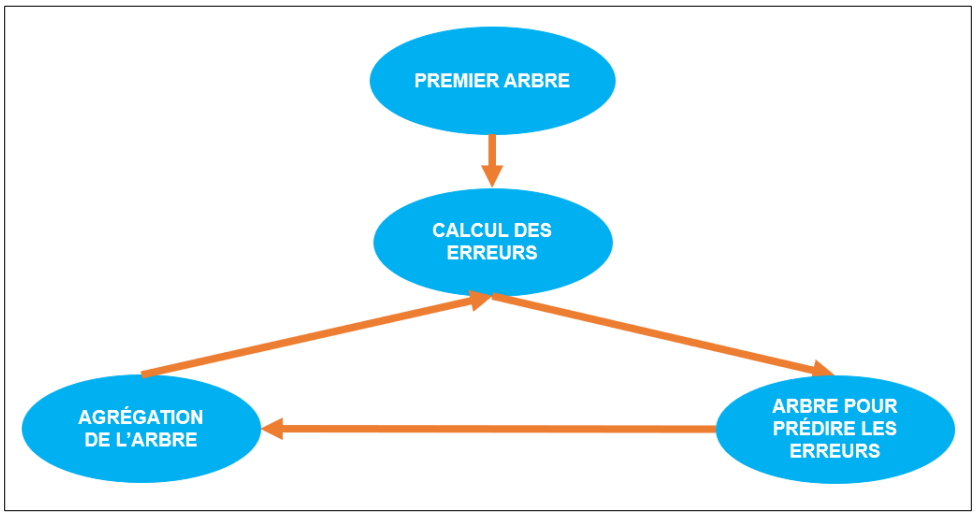


FIGURE 4.2 – eXtreme Gradient Boosting

Cette méthode d'agrégation d'arbres fonctionne aussi bien pour les problèmes de régression que de classification. De plus, une fois la totalité des arbres agrégés, la prédiction s'effectue similairement à un arbre de décision simple.

Afin de gérer la construction des différents arbres, plusieurs paramètres sont à considérer (nous verrons comment les optimiser dans la section 4.3.3). Ici, nous donnons une liste non exhaustive de ceux que nous considérons comme essentiels :

- Le nombre d'arbres : correspond au nombre d'itérations lors du processus d'agrégation des arbres ;
- Le taux d'apprentissage $\in [0, 1]$: ce paramètre contrôle la vitesse d'apprentissage des arbres. Plus il est faible, plus le nombre d'arbres doit être élevé afin d'obtenir de bonnes performances ;
- La profondeur maximale : correspond au nombre maximal de noeuds pour chaque arbre. Plus il est élevé, plus le risque de sur-apprentissage est grand ;
- Le nombre de variables considérées aléatoirement pour construire un arbre. Plus ce paramètre est élevé, plus le risque de corrélation entre les arbres est fort ;
- Le nombre d'observations minimal de chaque feuille de l'arbre ;

Mesure de l'importance des variables

Bien que les méthodes d'agrégation d'arbres améliorent nettement la qualité de prédiction, il peut être difficile d'interpréter leurs résultats. Effectivement, lorsque l'on agrège un grand nombre d'arbres, il devient impossible d'afficher les résultats comme pour un arbre de décision simple, et il est plus difficile d'interpréter l'importance des variables. Ainsi, ces méthodes améliorent la prédiction au détriment de l'interprétabilité. Néanmoins, nous pouvons obtenir une vue d'ensemble de l'importance des variables par le biais du RSS (régression) ou de l'indice de Gini (classification). Dans le cadre de la régression, nous pouvons enregistrer pour une variable explicative donnée, la diminution totale du RSS grâce à l'ensemble des noeuds dans laquelle elle intervient, moyennée sur l'ensemble des arbres. Ainsi, une grande valeur traduit d'une variable importante. De manière similaire, dans le cadre de la classification, nous remplaçons le RSS par l'indice de Gini.

4.3 Optimisation et sélection des modèles

4.3.1 Division de la base de modélisation

La robustesse d'un modèle est sa capacité à apprendre sur l'ensemble des données qui lui sont fournies, de sorte que lorsqu'on lui en présente de nouvelles, il les ajuste avec précision. Afin de remplir cet objectif, une méthode est de diviser la base de modélisation en plusieurs sous bases. Cette approche a ainsi été retenue dans cette étude, et de ce fait, notre base de modélisation a été divisée en trois parties : **une base d'entraînement**, de **validation** et de **test**. La figure ci-dessous présente cette séparation :

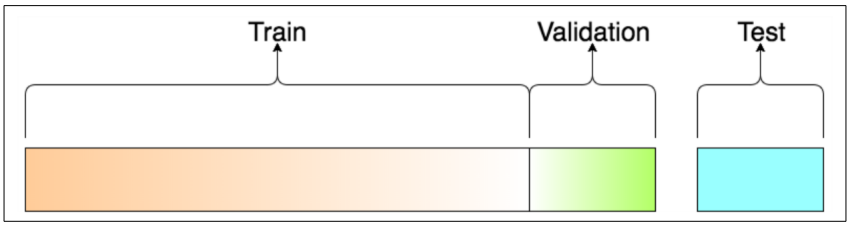


FIGURE 4.3 – Division de la base de modélisation

Chacune de ces bases a une utilité bien différente lors du processus de modélisation. C'est pourquoi nous les présentons dans cette section.

La base d'entraînement

La base **d'entraînement** est l'échantillon de données sur lequel le modèle est ajusté. C'est sur celui-ci que les paramètres d'un GLM sont estimés, ou que les régions d'un arbre sont définies. Dans cette étude, cette base contient 70% des observations de la base de modélisation, afin de permettre au modèle d'apprendre sur un nombre de données suffisant.

La base de validation

La base de **validation** est l'échantillon de données utilisé pour :
— fournir une évaluation non biaisée d'un modèle ajusté sur la base d'entraînement. En effet, cette base est utilisée pour évaluer un modèle donné, mais il s'agit d'une évaluation fréquente.
— optimiser les paramètres d'un modèle.

Ainsi, le modèle voit parfois les données de la base de modélisation, mais n'apprend jamais à partir de celles-ci. Cette base contient 20% des observations de la base de modélisation.

La base de test

La base de **test** est l'échantillon de données utilisé pour fournir une évaluation non biaisée d'un modèle final, adapté à l'ensemble des données de la base d'entraînement. En effet, le modèle n'ayant jamais été entraîné sur cette base, elle permet une évaluation finale et fiable des modèles. De plus, elle permet de comparer différents modèles de manière non biaisée. Cette base contient 10% des observations de la base de modélisation.

4.3.2 La validation croisée

L'évaluation d'un modèle s'effectue à partir d'indicateurs calculés sur une base test, tels que **l'erreur quadratique moyenne**, noté MSE pour Mean Squared Error, ou encore la **déviance** (ces indicateurs étant définis dans la section 4.3.3). Dans ce cadre, la validation croisée permet de rendre plus solide cette évaluation, en effectuant une première sélection des modèles par le biais d'une évaluation de la qualité de prédiction sur une base de validation.

Lors de de cette étude, la méthode *k-fold Cross Validation* est considérée. Celle-ci consiste à diviser l'ensemble de la base d'entraînement en k groupes, ou plis, de taille approximativement égale. Le premier pli est traité comme une base de validation, et le modèle est entraîné sur les $k-1$ plis restants. Ainsi, un premier indicateur d'évaluation (souvent le MSE ou la déviance), noté Ind_1, est calculé sur les observations de la base de validation créée. Cette procédure est répétée k fois, car à chaque fois, un groupe différent d'observations est traité comme base de validation. Ce processus aboutit à k calculs d'indicateurs : $Ind_1, ..., Ind_k$. Ainsi, l'indicateur d'évaluation final sur la base de validation est défini comme étant la moyenne des Ind, soit $\frac{1}{k}\sum_{i=1}^{k} Ind_k$. La figure ci-dessous illustre la méthode pour $k = 5$:

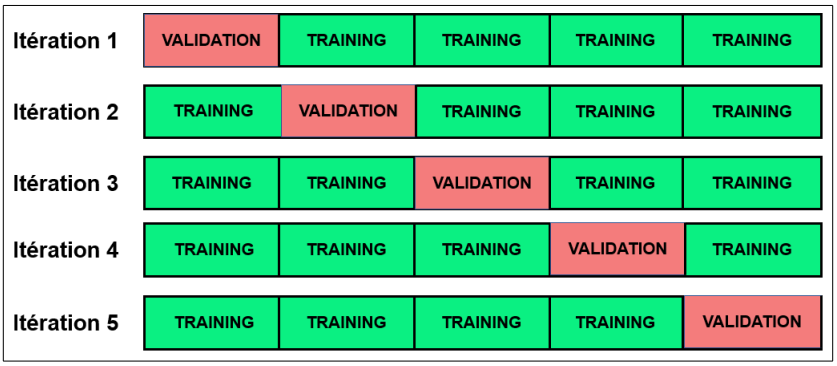


FIGURE 4.4 – Illustration de la 5-fold validation croisée

De ce fait, l'optimisation des paramètres d'un modèle peut être faite de manière robuste par le biais de cette méthode, puisque pour chaque paramètre, une évaluation moyennée du modèle correspondant est effectuée.

4.3.3 Optimisation des modèles

Comme énoncé précédemment, les paramètres des différents modèles doivent être optimisés afin d'obtenir la prédiction la plus précise et éviter le sur-apprentissage. Bien que les paramètres à optimiser soient différents pour les GLMs et les arbres de décision, le processus d'optimisation reste similaire. Nous le présentons donc dans cette section.

L'optimisation de ces différents paramètres s'effectue par le biais d'une grille de recherche (grid search en anglais). C'est une méthode d'optimisation qui va permettre de tester une série de paramètres et de comparer les performances pour en déduire le meilleur paramétrage. Ainsi, pour chaque paramètre, on détermine un ensemble de valeurs que l'on souhaite tester. Pour les GLMs pénalisés, une grille de paramètres pour λ et α sera définie. Concernant les arbres de décisions, le nombre d'arbres ou encore le nombre de variables considérées aléatoirement à chaque noeud sera testé. Ainsi, le Grid Search croise simplement chacune de ces valeurs et va créer un modèle pour chaque combinaison. Enfin, les modèles sont évalués efficacement grâce à la validation croisée.

Détaillons à présent le processus d'optimisation des paramètres pour les GLMs pénalisés (dans le cadre d'un modèle de régression) par le biais de la figure 4.5. Ainsi, pour une valeur de α fixée, une grille de valeurs de λ est définie. Dans cet exemple, nous avons n valeurs pour λ. Pour chaque valeur de λ, la base d'entraînement est découpée de manière aléatoire en k plis. Par validation croisée, une moyenne d'évaluation de la prédiction des modèles créés pour chaque valeur de λ peut être calculée. La valeur de λ créant le modèle ayant la meilleure prédiction sur la base de validation peut alors être sélectionnée : le λ qui minimise les indicateurs moyennés. Ainsi, nous obtenons pour un α fixé, le λ qui lui est optimal. Pour finir, nous obtenons une évaluation non biaisée de la qualité du modèle sélectionné en utilisant la base test. En effet, nous calculons, avec les données de la base test, le MSE ainsi que l'indice de Gini associé pour juger de la qualité de prédiction, mais également de la qualité de segmentation du modèle.

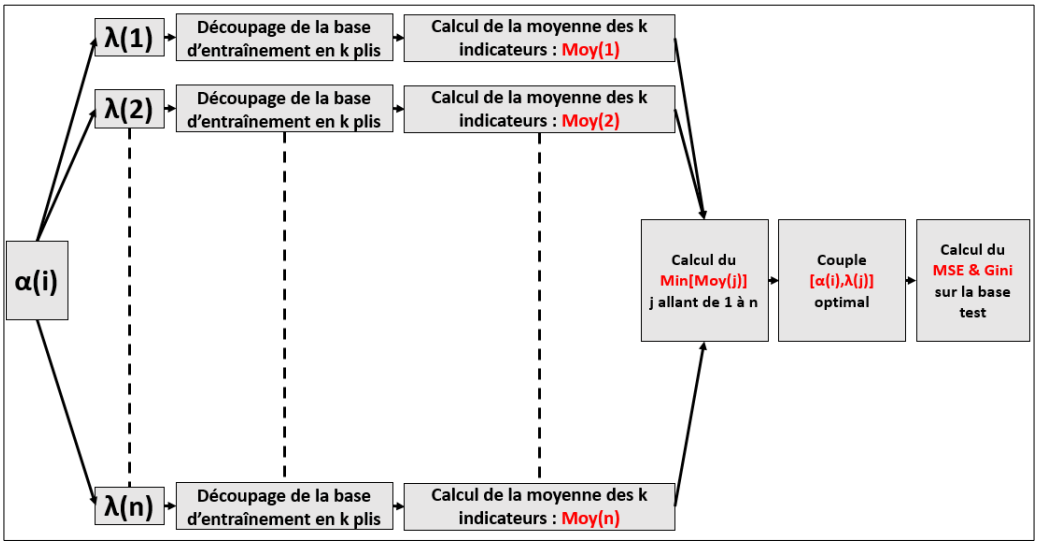


FIGURE 4.5 – Processus d'optimisation des paramètres pour les GLMs

Pour rappel, $\alpha \in [0;1]$ et contrôle la part distribuée entre la pénalité Ridge et LASSO. De ce fait, nous effectuons le processus détaillé précédemment pour une grille de valeurs de α. La grille retenue pour α dans cette étude correspond à 101 valeurs de α, allant de 0 à 1, chacune incrémentée par un pas de 0.01. Ainsi, pour chaque valeur de α, nous pouvons associer un λ optimal et évaluer la qualité du modèle sur la base test. Pour finir, trois GLMs sont conservés afin de les comparer aux arbres de décisions :

— Le GLM avec la pénalisation LASSO : correspondant au modèle ayant un λ optimal et $\alpha = 0$;
— Le GLM avec la pénalisation Ridge : correspondant au modèle ayant un λ optimal et $\alpha = 1$;
— Le GLM avec la pénalisation Elastic Net : correspondant au modèle ayant un couple (λ,α) optimal, avec un $\alpha \in [0,01;0,99]$. Le couple (λ,α) optimal est sélectionné après étude de certains indicateurs sur la base test tels que le R^2 ajusté, le BIC (défini en section 2.1.4.3), l'indice de Gini, le MSE, la déviance ou encore l'erreur totale.

L'erreur totale ainsi que l'indice de Gini sont définis dans la section 2.3. De ce fait, définissons les indicateurs restants :

— L'erreur moyenne quadratique, notée **MSE** :

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

Le modèle sélectionné sera celui qui minimise ce critère.

— La **déviance**, permettant de comparer les GLMs :

$$\text{Déviance} = 2log(\tilde{l} - \hat{l})$$

Où $\tilde{l}$ est la log-vraisemblance du modèle saturé, et $\hat{l}$ celle du modèle estimé. Le modèle saturé est défini comme le modèle possédant autant de paramètres que d'observations, et estimant donc exactement les données. Le modèle sélectionné sera celui qui minimise ce critère.

Comme énoncé précédemment, le processus d'optimisation des arbres de décision est similaire à celui des GLMs, excepté que les paramètres sont différents. En effet, une grille ainsi que la validation croisée sont utilisées dans le but de sélectionner les paramètres minimisant un critère. Le critère utilisé pour la régression est le MSE, tandis que pour la classification nous utiliserons l'aire en dessous de la courbe ROC, la spécificité et la sensibilité (définies en section 6.1.2).

4.3.4 Sélection des modèles

La comparaison des modèles est facilitée par le biais de certaines métriques d'évaluation. Toutes ces dernières sont calculées sur la base test, et diffèrent selon le type de problématique : de régression ou de classification. Ainsi, les métriques d'évaluation retenues pour les problèmes de régression sont le MSE, l'indice de Gini et l'erreur totale. Quant aux problèmes de classification, l'aire en dessous de la courbe ROC, la spécificité et la sensibilité sont retenues. Le choix de toutes ces métriques a été effectué dans le but d'obtenir l'analyse la plus complète de chaque modèle, et ainsi, d'obtenir la prime pure la plus fidèle au risque réellement observé.

4.4 Rééquilibrage des données

Avec 1,1% de sinistres graves parmi l'ensemble des sinistres, notre jeu de données est très déséquilibré. De ce fait, la modélisation de la classe minoritaire (correspondant aux assurés ayant eu un sinistre grave) peut être problématique. En effet, cette situation peut rendre l'estimation des paramètres d'un modèle par maximum de vraisemblance hasardeuse, et peut limiter la performance prédictive des arbres. Ainsi, pour améliorer les performances des modélisations, des techniques de rééquilibrage de données peuvent être effectuées comme :

— Le sur-échantillonnage : consistant à ajouter à la population des individus sinistrés ;
— Le sous-échantillonnage : consistant à supprimer à la population des individus non sinistrés ;
— L'échantillonnage mixte : consistant à combiner les deux types d'échantillonnage précédents.

Dans cette section, nous présentons la *Synthetic Minority Over-sampling Technique*, plus connue sous le nom **SMOTE**. Cette méthode développée par *Chawla et al* est une méthode de sur-échantillonnage synthétique de la classe minoritaire. L'intérêt de cette méthode est d'éviter le sur-apprentissage qui viendrait d'une simple réplication des individus de la classe minoritaire.

4.4.1 Suréchantillonnage synthétique

Le **SMOTE** consiste à créer de nouveaux individus synthétiques ressemblant très fortement aux individus réels de la classe minoritaire. Pour ce faire, les nouveaux individus sont situés aléatoirement sur des segments liants les individus de la classe minoritaire, dans l'espace des variables explicatives. L'algorithme **SMOTE** fonctionne de la manière suivante :

1. Sélectionner aléatoirement un individu i de la classe minoritaire ;
2. Identifier les k plus proches voisins de i. Pour ce faire, la distance de Gower explicitée en section 3.1.2 a été utilisée ;
3. Sélectionner aléatoirement un individu j parmi les k plus proches voisins de i ;
4. Calculer la distance entre i et j ;
5. Multiplier cette distance par un nombre aléatoire α compris entre 0 et 1 ;
6. Considérer ce nouvel individu synthétique comme membre de la classe minoritaire.

La figure 4.6 illustre une itération de la méthode pour 2 variables quantitatives, et un nombre de voisins égal à 3. Ainsi, cette itération est répétée autant de fois que nécessaire jusqu'à obtenir une base de modélisation équilibrée.

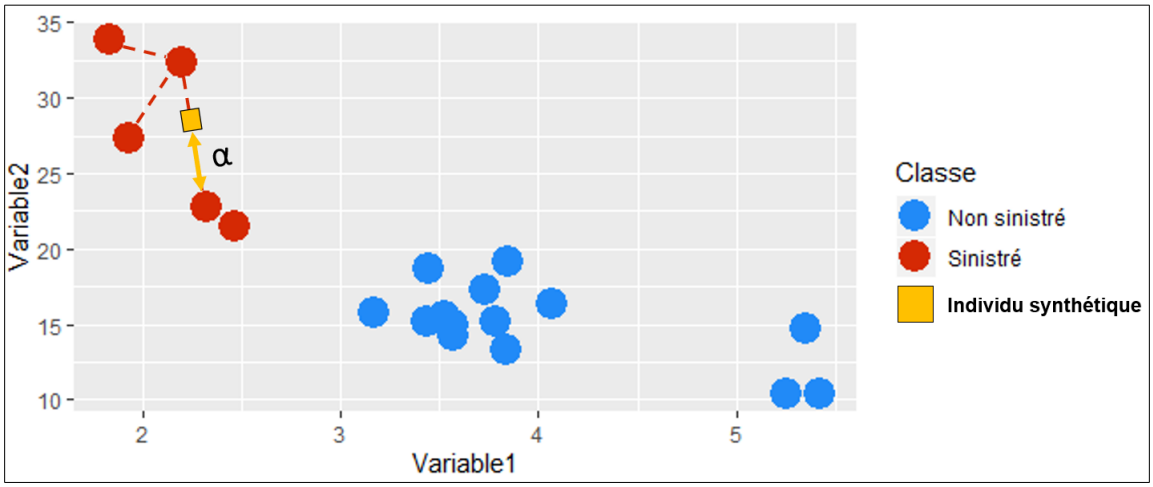


FIGURE 4.6 – Exemple de création d'un individu synthétique par la méthode SMOTE

4.4.2 Effet sur la modélisation

Effet sur les arbres de décision

Le sur-échantillonnage par réplication permet uniquement un apprentissage sur les régions très spécifiques des individus observés, et de ce fait, n'améliore pas significativement la capacité à prédire les observations minoritaires. En revanche, le sur-échantillonnage synthétique permet l'apprentissage sur des zones plus larges parmi la région des observations minoritaires. Essentiellement utilisé pour les arbres de classification, la méthode **SMOTE** permet de construire des arbres générant bien plus de vrais positifs sur la base test, étant donné qu'ils ont appris la réelle région occupée par les observations minoritaires. Toutefois, cette amélioration n'est possible qu'au prix d'une augmentation du nombre de faux positifs. Néanmoins, l'effet de cette augmentation est négligeable lorsqu'elle est rapportée au nombre total d'observations négatives. Par conséquent, la performance du modèle augmente.

Effet sur les modèles linéaires généralisés

Faisant l'hypothèse d'une relation linéaire entre la variable réponse et les variables explicatives, les GLMs ne sont pas réellement touchés lorsqu'il faut apprendre sur des régions très spécifiques de l'espace. Néanmoins, lorsque les données minoritaires sont trop peu nombreuses, la maximisation de la vraisemblance peut conduire à une estimation trop générale des coefficients. Dans ce cas, la prédiction est quasi identique pour tous les individus. Or, en densifiant la région des observations minoritaires par la méthode **SMOTE**, les GLMs parviennent à apprendre sur nos données. Ainsi, le sur-échantillonnage synthétique sera retenu sous conditions que les modèles aient de meilleurs résultats que ceux issus de la base de modélisation initiale.

4.4.3 Effet de l'espace sur la méthode SMOTE

La gestion de l'espace des variables explicatives à considérer pour la méthode **SMOTE** est très importante. Il est fortement conseillé de considérer uniquement des variables ayant un lien significatif avec le risque (le risque étant le fait d'avoir un sinistre dans notre cas). En effet, si des variables n'ayant pas de lien avec le risque sont considérées lors du sur-échantillonnage, des individus synthétiques (qui de fait appartiennent à la classe minoritaire) peuvent être créés sans pour autant avoir le profil de risque de la classe minoritaire. Ainsi, cette situation réduirait fortement la performance des modèles, qui apprendraient sur des régions faussées. Le sur-échantillonnage peut donc être effectué correctement après une étude du lien entre les variables et le risque considéré.

Chapitre 5

Modèles de fréquence

5.1 Cadre et objectifs

Tout d'abord, rappelons la formule de la prime pure que nous cherchons à calculer (cf. section 2.1.1.3) pour un assuré i du portefeuille :

$$p_i = f_i \times [g_i \times c_i^{\text{grave}} + (1 - g_i) \times c_i^{\overline{\text{grave}}}]$$

Avec $f_i = \frac{\text{Nombre de sinistres}}{\text{Durée d'exposition}}$.

Ce chapitre traite de la modélisation de f_i par différentes méthodes : les GLMs pénalisés et l'apprentissage supervisé. D'un côté, les GLMs pénalisés sont des méthodes classiques de tarification des contrats d'assurance non-vie. Effectivement, ces modèles sont établis à partir de fortes hypothèses telle que la forme du lien entre la variable réponse et les variables explicatives, et sont faciles à interpréter. De l'autre, l'apprentissage supervisé est un ensemble de méthodes innovantes pouvant également être utilisées à des fins de tarification. De plus, ces dernières ne font aucune hypothèse sur la forme du lien entre la variable réponse et les variables explicatives.

Dans un premier temps, la modélisation de la fréquence de sinistres par les GLMs pénalisés est présentée. Pour ce faire, l'optimisation par grille de recherche et validation croisée ainsi que les résultats des régressions LASSO, Ridge et Elastic sont exposés. Ensuite, la modélisation par l'apprentissage supervisé est détaillée. Deux méthodes d'agrégation d'arbres ont ainsi été retenues : le Random Forest et le XGBoost. De manière similaire aux GLMs pénalisés, l'optimisation des paramètres de ces modèles ainsi que leurs résultats sont exposés. Enfin, une synthèse est effectuée afin de résumer l'efficacité de chacun de modèles, et ainsi retenir le plus performant en terme de précision et d'interprétation.

Lors de la dernière étude du risque incendie, une unique modélisation a été réalisée, dans le sens où cette dernière a été faite sur l'ensemble de la population. Or, l'objectif de cette étude est d'obtenir une tarification similaire pour les sous-groupes de risques homogènes de notre échantillon. Ainsi, pour réactualiser et optimiser les derniers modèles, une distinction entre la sinistralité appartements et maisons a été effectuée. Effectivement, ces deux types de sinistres répondent bien à des problématiques de modélisation différentes. De plus, elle permet d'obtenir une interprétation simple tout en étant en accord avec la modélisation des autres garanties telle que le dégât des eaux.

5.2 Régression Poissonnienne pénalisée

Dans cette section, nous présentons les résultats de trois GLMs pénalisés : la régression LASSO, Ridge et l'Elastic Net. La variable réponse Y étant à valeurs dans $\mathbb{N}$, nous considérons un GLM Poisson avec sa fonction de lien canonique ($g = log$). Ainsi, pour une observation i donnée, nous avons :

— $Y_i|X_i = x_i$ qui suit une loi de Poisson. La loi de Poisson faisant partie de la famille de lois exponentielles ;
— $log(\mu(X_i)) = log(\mathbb{E}[Y_i|X_i]) = X_i'\beta$.

Néanmoins, pour modéliser une fréquence de sinistres, l'exposition (définie en section 2.1.1.2), notée ω, doit nécessairement être prise en compte. Nous supposons alors pour une observation i que $\mu(X_i) = \mathbb{E}\left(\frac{Y_i}{\omega_i}|X_i\right)$. Le modèle est donc légèrement modifié : $log\left(\frac{\mathbb{E}[Y_i|X_i]}{\omega_i}\right) = X_i'\beta$, et peut se réécrire : $log(\mathbb{E}[Y_i|X_i]) = log(\omega_i) + X_i'\beta$. L'offset (correspondant à $log(\omega_i)$) peut être vu comme une nouvelle variable du modèle avec un coefficient β constant égal à 1.

L'estimation des paramètres est effectuée en maximisant la vraisemblance pénalisée par la méthode Elastic Net :

$$\max_{\beta_0,\beta} \frac{1}{N}\sum_{i=1}^{N}\left\{y_i(\beta_0+\beta' x_i) - e^{\beta_0+\beta' x_i}\right\} - \lambda[\alpha\left\|\beta\right\|_1 + (1-\alpha)\left\|\beta\right\|_2]$$

Pour rappel, les régressions LASSO et Ridge peuvent être vues comme des cas particuliers de l'Elastic Net avec respectivement $\alpha = 1$ et $\alpha = 0$.

5.2.1 La modélisation du risque appartement

Dans cette section, nous étudions la fréquence de sinistres des assurés ayant un appartement. Pour ce faire, la quantification de la qualité de prédiction ainsi que l'interprétation des coefficients de chacun des modèles est effectuée. Comme énoncé dans la section 4.3.3, trois GLMs pénalisés sont conservés et seront comparés aux arbres de décision : les GLMs avec les pénalisations LASSO, Ridge et Elastic Net optimales.

Toutes les variables potentiellement explicatives (décrites dans l'interprétation des coefficients) étant catégorielles, les GLMs pénalisés considèrent chaque modalité possible comme une variable binaire.

Choix du λ optimal

Pour chacun des GLMs, le choix du λ optimal se fait par validation croisée. En outre, le lecteur est invité à se référer à la section 4.3.3 pour obtenir plus de détails sur l'optimisation des paramètres d'un GLM. La métrique d'évaluation permettant de sélectionner le λ optimal par validation croisée est la déviance de Poisson, qui est définie comme :

$$D_{poisson} = -2\sum_{i=1}^{N}\{y_i log(y_i/\hat{y}_i) - (y_i - \hat{y}_i)\}$$

Analysons la figure 5.1. Tout d'abord, pour la régression LASSO, $log(\lambda) = -10,6 \iff \lambda = 2,6 \times 10^{-5}$. Les nombres au dessus de chaque graphique correspondent au nombre de modalités des variables explicatives conservées dans le modèle pour chaque valeur de lambda. Ainsi, la régression LASSO n'en conserve que 21. On voit alors l'effet de la norme 1 dans la pénalisation, qui permet de retirer du modèle des variables n'ayant pas de lien significatif avec la fréquence de sinistres.

Ensuite, pour la régression Ridge, $log(\lambda) = -7,8 \iff \lambda = 3,9 \times 10^{-4}$. Ainsi, la régression Ridge conserve l'ensemble des modalités de chaque variable explicative quelque soit la valeur de λ.

Enfin, pour l'Elastic Net, le couple optimal obtenu est $(\alpha, \lambda) = (0,2 \ ; \ 1,2 \times 10^{-4})$. Cette pénalisation conserve 24 modalités des variables explicatives. Ainsi, nous voyons que l'Elastic Net est un compromis entre le LASSO et le Ridge.

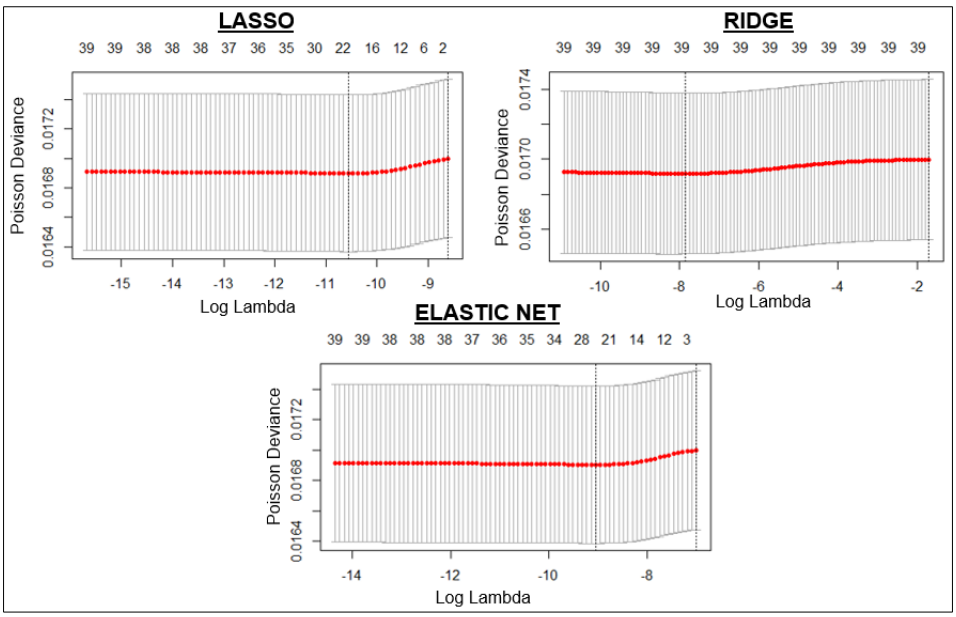


FIGURE 5.1 – Déviance de Poisson en fonction de $log(\lambda)$

Valeur des coefficients en fonction de λ

Les graphiques ci-dessous ne sont pas facilement lisibles compte tenu du grand nombre de variables. Voici comment ces derniers se lisent. Pour chaque GLM, lorsque $\lambda = 0$, l'ensemble des variables est conservé. Au fur et à mesure que λ augmente, la pénalisation s'accentue. Ainsi, la pénalisation LASSO force certains coefficients à 0, tandis que celle de Ridge leurs donne moins d'importance en les diminuant, sans jamais les forcer à 0. L'Elastic Net est facilement lisible comme un compromis des deux autres pénalisations.

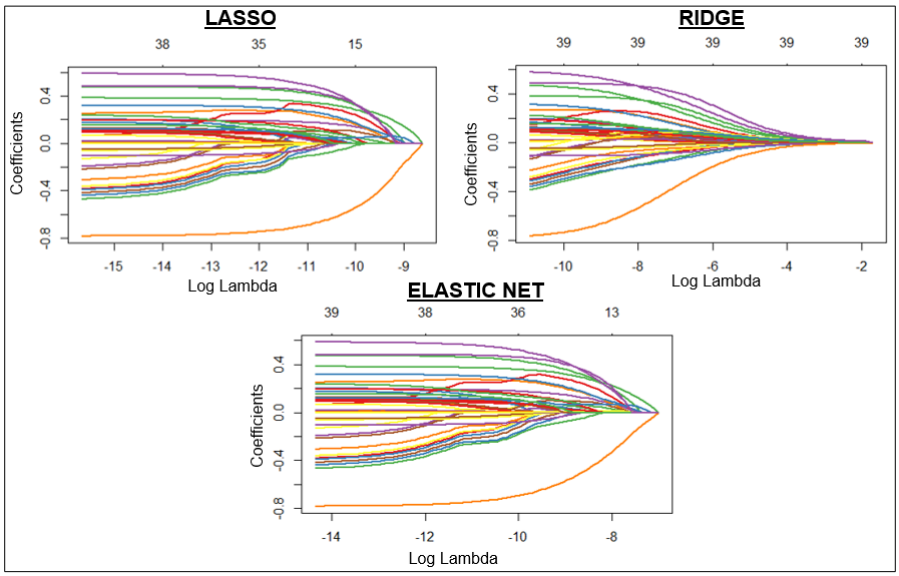


FIGURE 5.2 – Valeurs des coefficients en fonction de $log(\lambda)$

Comparaison des coefficients

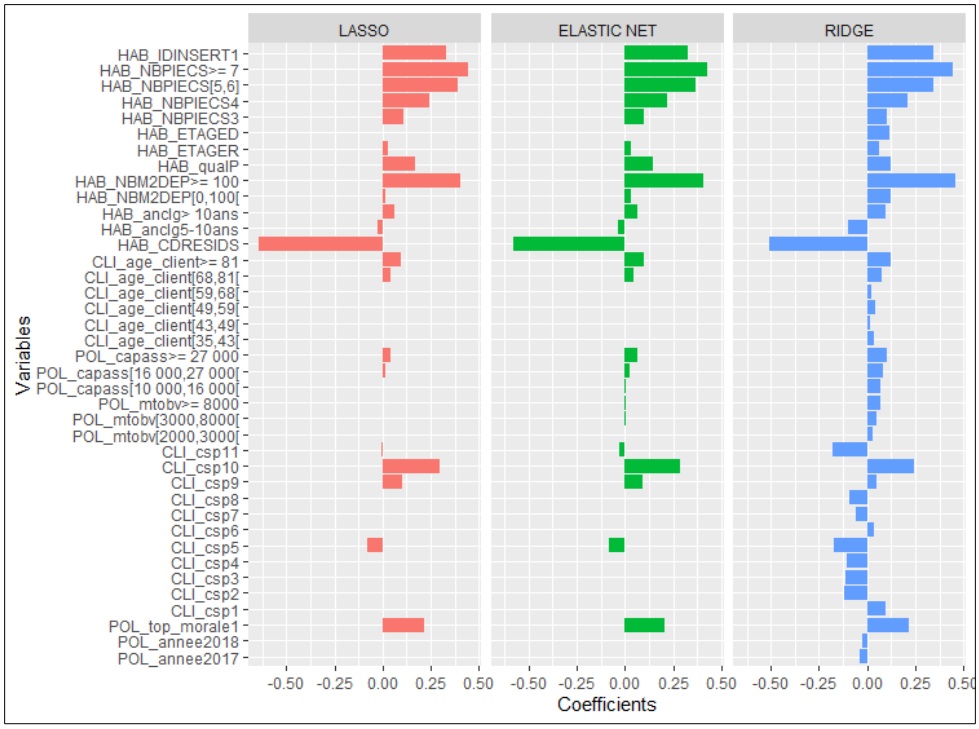


FIGURE 5.3 – Coefficients GLMs - Fréquence appartements

La figure ci-dessus permet de comparer la valeur des coefficients pour chacune des pénalisations. Ces dernières ont chacune leurs avantages et inconvénients. En conservant l'ensemble des variables, la pénalisation Ridge permet de maximiser la précision, mais rend l'interprétation des modèles plus complexe et est fortement exposée au sur-apprentissage. La pénalisation LASSO crée des modèles plus facilement interprétables, mais parfois trop simples pour être robustes.

Dans un premier temps, nous voyons facilement que lorsque les coefficients sont non nuls, la tendance est la même quelque soit la pénalisation. Ainsi, la présence d'un insert augmente logiquement le risque de sinistre incendie en appartement. De manière similaire, ce dernier augmente avec le nombre de pièces. Également, le coefficient associé aux résidences secondaires est négatif, ce qui signifie que le risque incendie pour une résidence secondaire est moins élevé que pour une résidence principale. Par ailleurs, nous remarquons que les propriétaires ont un risque plus élevé. Les dépendances jouent également un rôle sur ce risque. Effectivement, ce dernier augmente avec la surface de celles-ci.

Ensuite, l'étage, l'ancienneté du logement, le capital assuré, l'âge du client, le montant d'objets de valeurs, ainsi que l'année de survenance n'influent pas sur le risque ou que très peu. Concernant les CSP, seules les 9 et 10 (correspondant respectivement aux retraités et aux inactifs) jouent sur le risque de manière significative. De plus les personnes morales sont plus exposées au risque.

Métriques d'évaluation des modèles

Le tableau 5.1 nous fournit quelques métriques d'évaluation afin de juger de la qualité des différents modèles. Tout d'abord, nous remarquons que les modèles sont robustes au vu des résultats sur la base test. De plus, ils ne diffèrent que très peu en terme de qualité de prédiction et de segmentation. Effectivement, l'arrondi des métriques au centième près nous donne des résultats identiques. De ce fait, à précision quasi identique, il faudrait conserver le modèle ayant l'interprétation la plus simple et le moins variant, et de ce fait, étant le moins exposé au sur-apprentissage. Pour ce faire, nous ajoutons les analyses graphiques précédentes à cette étude d'indicateurs. Ainsi, en retirant près de 18 variables du modèle, le GLM LASSO devrait être favorisé.

Modèle	Couple $(\alpha;\lambda)$	BIC	MSE Train	MSE Test	Erreur globale Train	Erreur globale Test	Gini Train	Gini Test
LASSO	$(1;2,6\times10^{-5})$	$-2,1\times10^{4}$	$1,3\times10^{-3}$	$1,3\times10^{-3}$	0%	3%	27%	27%
ELASTIC NET	$(0,2;1,2\times10^{-3})$	$-2,1\times10^{4}$	$1,3\times10^{-3}$	$1,3\times10^{-3}$	0%	3%	27%	27%
RIDGE	$(0;3,9\times10^{-4})$	$-2,1\times10^{4}$	$1,3\times10^{-3}$	$1,3\times10^{-3}$	0%	3%	27%	27%

TABLE 5.1 – Métriques d'évaluation GLMs - Fréquence appartements

5.2.2 La modélisation du risque maison

Après avoir modélisé la fréquence de sinistres pour les assurés ayant un appartement, nous étudions ce risque pour les assurés ayant une maison. Le processus de modélisation étant similaire, seuls les graphiques les plus importants seront présents et analysés dans cette section.

Comparaison des coefficients

Pour chaque couple (α, λ) optimal, la figure 5.4 compare la valeurs des coefficients. Tout d'abord, nous voyons aisément la différence du risque incendie entre les appartements et les maisons, car l'influence de chaque modalité pour chaque type d'habitation est bien distincte.

Ensuite, nous remarquons que lorsque les coefficients sont non nuls pour chacune des pénalisations, la tendance est la même. Ainsi, le fait d'avoir un insert et une grande maison (maison avec 7 pièces ou plus) influe positivement sur le risque. A contrario, les résidences secondaires sont moins susceptibles de voir se déclencher un incendie que les résidences principales. Aussi, les propriétaires sont plus à risque que les locataires, et la surface des dépendances augmente légèrement le risque.

De plus, l'année de survenance, la CSP, le montant d'objets de valeur, le capital assuré, l'ancienneté du logement et l'âge du client n'influent pas sur le risque ou que très peu.

Finalement, nous remarquons que le GLM LASSO fournit un modèle beaucoup plus épuré que le Ridge ou l'Elastic Net en forçant la plupart de coefficients à 0. De ce fait, la pénalisation LASSO rend le modèle plus interprétable et moins variant. Néanmoins, avant de faire un choix de modèle, il reste à vérifier la qualité de leurs prédictions.

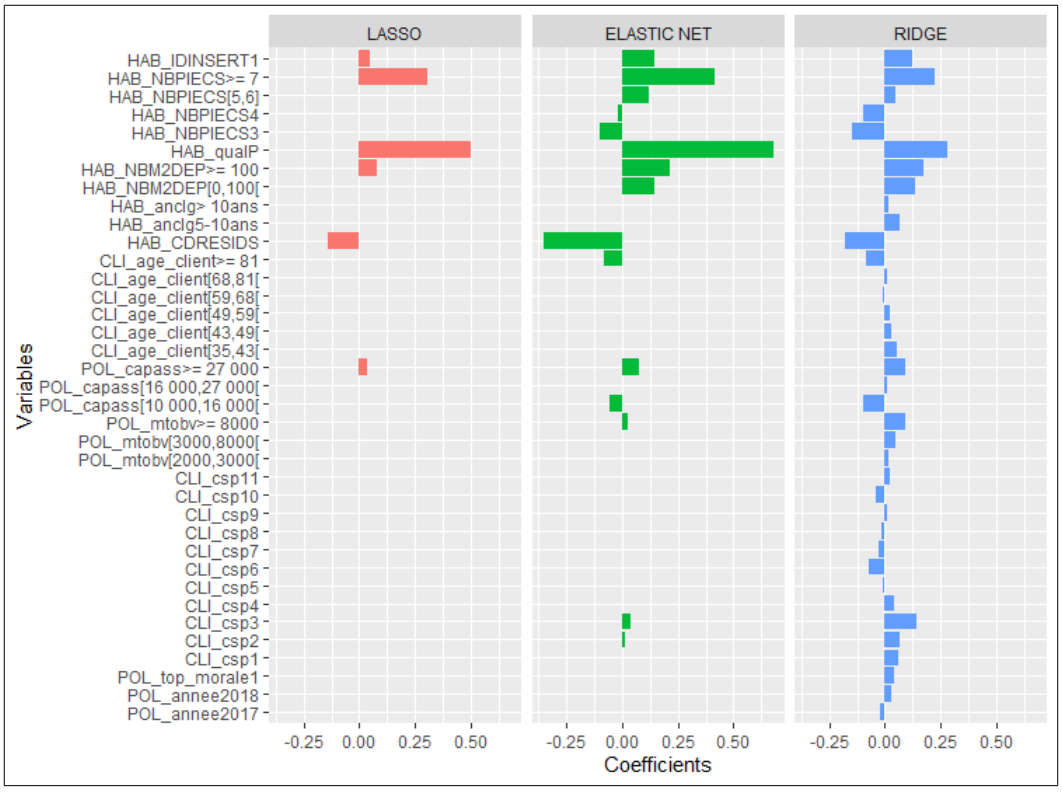


FIGURE 5.4 – Coefficients GLMs - Fréquence Maisons

Métriques d'évaluation des modèles

Le tableau ci-dessous nous fournit quelques métriques d'évaluation, afin de juger de la qualité des différents GLMs sur la fréquence de sinistres des maisons. De manière similaire aux appartements, les modèles sont robustes au vu des résultats sur la base test, et ne diffèrent que très peu en terme de qualité de prédiction. Néanmoins, le GLM Elastic Net possède l'indice de Gini le plus élevé, qui traduit une meilleure segmentation de notre portefeuille. De ce fait, il serait judicieux de conserver ce dernier qui est plus simple d'interprétation et moins exposé au sur-apprentissage que le GLM Ridge, tout en segmentant mieux le portefeuille que le GLM LASSO.

Modèle	Couple $(\alpha;\lambda)$	BIC	MSE Train	MSE Test	Erreur globale Train	Erreur globale Test	Gini Train	Gini Test
LASSO	$(1;6,4\times10^{-4})$	$-9,3\times10^{4}$	$5,6\times10^{-3}$	$5,6\times10^{-3}$	0%	0%	23%	23%
ELASTIC NET	$(0,34;6,2\times10^{-4})$	$-9,3\times10^{4}$	$5,6\times10^{-3}$	$5,6\times10^{-3}$	0%	0%	29%	29%
RIDGE	$(0;7,4\times10^{-3})$	$-9,3\times10^{4}$	$5,6\times10^{-3}$	$5,6\times10^{-3}$	0%	0%	25%	25%

TABLE 5.2 – Métriques d'évaluation GLMs - Fréquence maisons

5.3 Arbres de régression

Dans cette section, nous modélisons la fréquence de sinistres des assurés de notre portefeuille par le biais de méthodes d'agrégation d'arbres : Random Forest et XGBoost. Le lecteur est invité à se référer aux sections 4.2.3 et 4.2.4 pour obtenir une présentation théorique de ces modèles.

Tout d'abord, nous allons présenter, pour chaque méthode, le processus d'optimisation des paramètres. Ensuite, nous analyserons l'importance des variables de chaque modèle. Enfin, les métriques d'évaluations obtenues seront commentées. Le processus d'optimisation des paramètres étant similaire pour chaque type d'habitation (appartement et maison), il ne sera détaillé qu'une seule fois.

5.3.1 Random Forest

5.3.1.1 La modélisation du risque appartement

Optimisation des paramètres

Comme énoncé dans la section 4.2.3, Random Forest a deux principaux paramètres à optimiser :
— Le nombre d'arbres, noté T ;
— Le nombre de variables explicatives considérées aléatoirement à chaque noeud, noté m.

Pour cette méthode d'agrégation d'arbres, le risque de sur-apprentissage n'augmente pas avec le nombre d'arbres considéré. Ainsi, seul m sera optimisé. Néanmoins, le nombre d'arbres à été choisi de telle sorte que ce paramètre génère des résultats robustes et précis. Par conséquent, 500 arbres ont été retenus.

La modélisation comprenant 32 variables, nous considérons la grille de valeurs entières $[\![1, 15]\!]$ pour le paramètre m. Ainsi, avec un maximum de 15 variables considérées, les chances de corrélation entre les arbres sont nettement réduites. Comme énoncé, le paramètre est optimisé par la méthode de validation croisée k-folds. Effectivement, cette méthode robuste permet d'affecter à chaque valeur de la grille un indicateur, qui est dans le cas de la régression le $RMSE$. Pour rappel, le $RMSE$ est simplement la racine carrée du MSE défini en section 4.3.3. De ce fait, nous cherchons la valeur de m, qui pour 500 arbres, minimise ce critère.

La figure 5.5 permet d'illustrer l'optimisation de ce paramètre. En effet, elle associe pour chaque valeur m considérée, le $RMSE$ calculé par validation croisée k-folds. Ainsi, nous voyons facilement que le $RMSE$ est maximal pour $m = 1$, atteint un mimimum en $m = 3$, puis croît jusqu'à $m = 15$.

Par conséquent, pour 500 arbres considérés, le paramètre m optimal est 3, soit 10% des variables explicatives totales.

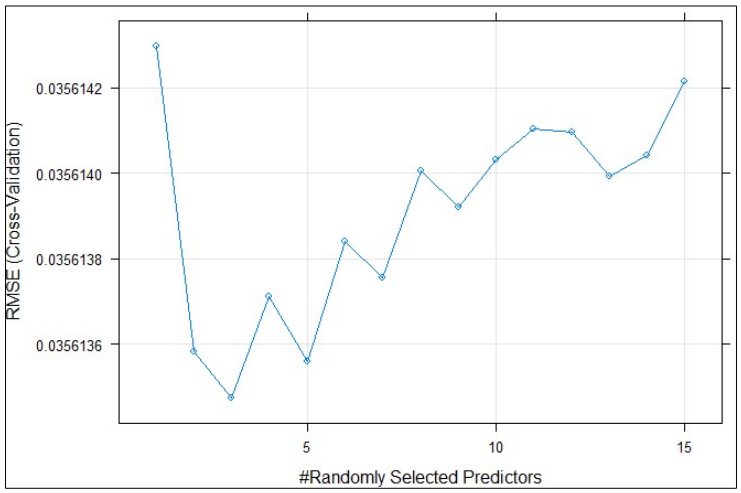


FIGURE 5.5 – Optimisation du paramètre m

Importance des variables

Après avoir optimisé le paramètre m et fixé $T = 500$, nous pouvons étudier l'importance des variables du modèle obtenu. Pour rappel, la section 4.2.4 indique comment l'importance des variables est mesurée pour les méthodes d'agrégation d'arbres. La figure 5.6 présente uniquement les 15 variables les plus importantes du modèle. Ainsi, nous remarquons que le montant de capital assuré (*POL_capass*) et la surface des dépendances (*HAB_NBM2DEP*) jouent un rôle très important dans la segmentation de l'espace, et de ce fait influent sur la fréquence de sinistres des assurés. Ici, le montant de capital assuré correspond au capital mobilier, et la surface des dépendances à la surface d'un ou des locaux placés à l'extérieur de l'habitation. Ensuite, vient le montant d'objets de valeur, puis le nombre de pièces et l'âge du client. Enfin, le reste des variables a une influence moins significative telles que la qualité de l'occupant (propriétaire ou locataire), l'étage, l'ancienneté du client, la présence d'un insert ou encore la catégorie socioprofessionnelle.

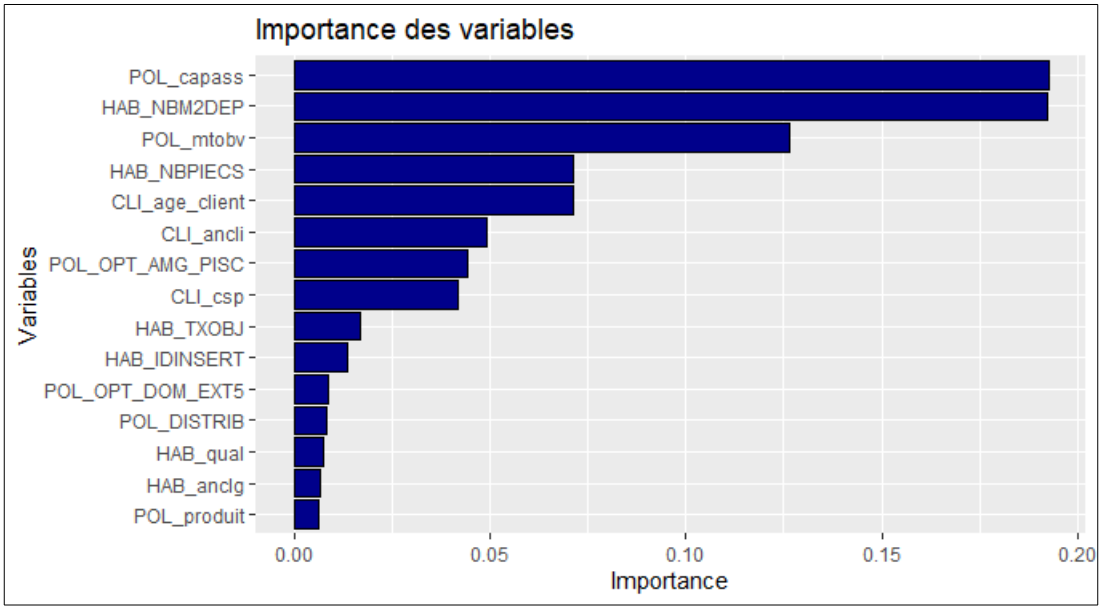


FIGURE 5.6 – Importance des variables Random Forest - Fréquence Appartements

Métriques d'évaluation

Concernant la qualité de prédiction, nous remarquons que le Random Forest ajuste plutôt bien les données. Effectivement, que ce soit sur la base d'entraînement ou la base de test, les MSE ainsi que l'erreur globale sont très faibles. Néanmoins, avec un indice de Gini plutôt faible, nous nous apercevons que la qualité de segmentation n'est pas optimale, et est bien moins bonne que celle obtenue avec les GLMs pénalisés (27%). De ce fait, nous pouvons conclure que pour les appartements, cette méthode d'agrégation d'arbres ajuste bien les données, mais qu'elle n'identifie pas de façon optimale les sous-groupes de risques homogènes du portefeuille.

Métriques Random Forest - Fréquence appartements			
Base	MSE	Erreur globale	Gini
Train	$1,3 \times 10^{-3}$	0	14%
Test	$1,3 \times 10^{-3}$	-1%	14%

TABLE 5.3 – Métriques d'évaluation Random Forest - Fréquence appartements

5.3.1.2 La modélisation du risque maison

A présent, nous modélisons la fréquence de sinistres des assurés ayant une maison. L'optimisation des paramètres m et T étant similaire à la modélisation du risque appartement, seules l'importance des variables et les métriques d'évaluation du modèle optimal sont détaillées dans cette section.

Importance des variables

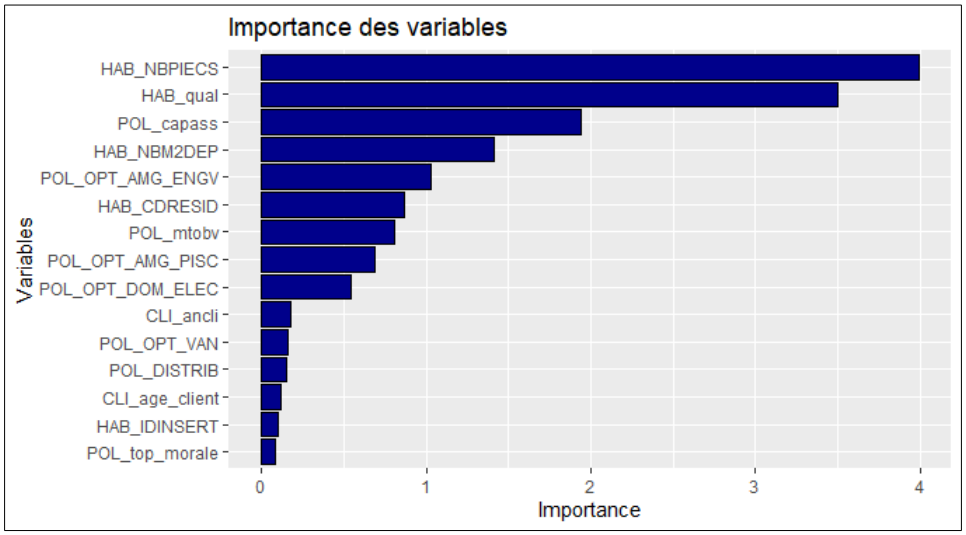


FIGURE 5.7 – Importance des variables Random Forest - Fréquence Maisons

Nous remarquons que le nombre de pièces ainsi que la qualité de l'occupant (propriétaire ou locataire) ont la plus grosse influence sur la fréquence de sinistres. Ensuite, viennent le montant de capital assuré et la surface des dépendances. De plus, la possession de biens assurés au titre de l'option « Énergies renouvelables » (*POL_OPT_AMG_ENGV*) joue sur le risque d'occurrence de sinistres. Pour information, les pompes à chaleur, les panneaux solaires ou encore les éoliennes sont des biens pouvant être assurés au titre de cette option. Le reste des variables a une influence moins importante. Toutefois, nous pouvons noter la différence entre le risque appartement et le risque maison. Effectivement, pour chaque type d'habitation, le poids de chacune des variables sur la fréquence de sinistres est différent, et de ce fait, justifie bien que ces deux risques répondent à des problématiques de modélisation distinctes.

Métriques d'évaluation

Le MSE nous indique que le modèle ajuste bien les données. Néanmoins, nous observons une erreur globale sur la base test de -6%. Cet indicateur traduit simplement le fait que le modèle sous estime la fréquence de sinistres de la base test de 6%. Aussi, l'indice de Gini de 27% traduit d'une qualité de segmentation meilleure que les GLMs LASSO et Ridge.

Métriques Random Forest - Fréquence maisons			
Base	MSE	Erreur globale	Gini
Train	$5,6 \times 10^{-3}$	0	27%
Test	$5,9 \times 10^{-3}$	-6%	27%

TABLE 5.4 – Métriques d'évaluation Random Forest - Fréquence maisons

5.3.2 eXtreme Gradient Boosting

5.3.2.1 Modélisation du risque appartement

Optimisation des paramètres

Dans la section 4.2.4, nous avons listé et défini les principaux paramètres à optimiser pour le XGBoost. Néanmoins, le processus étant identique et pour éviter la répétition, nous détaillons dans cette section uniquement l'optimisation des paramètres suivants :

— Le nombre d'arbres, noté n ;
— Le taux d'apprentissage $\in [0, 1]$, noté η.

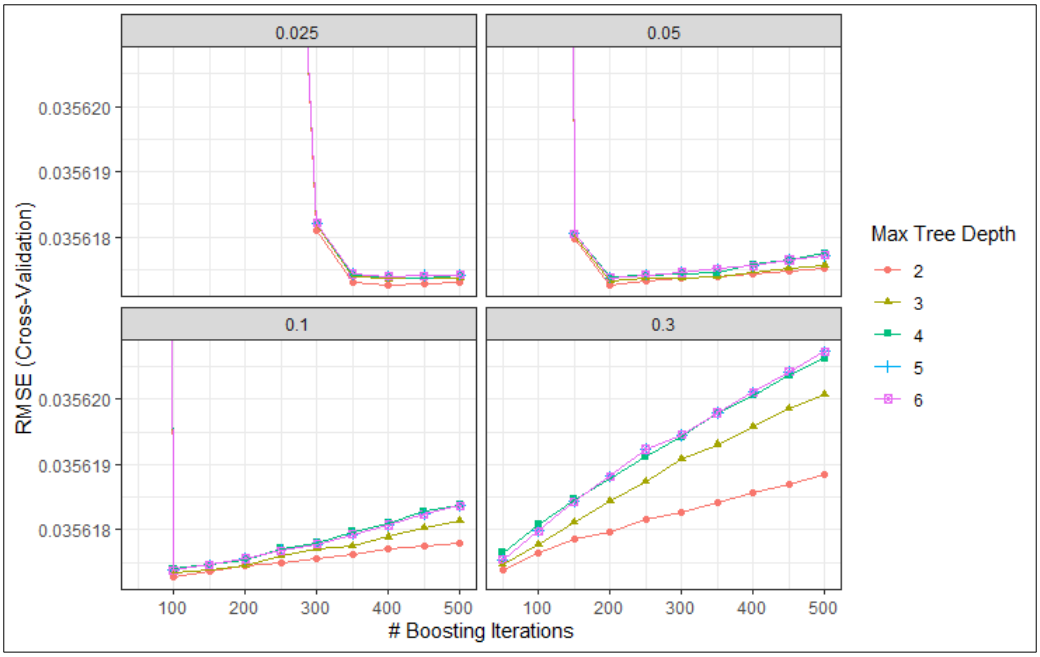


FIGURE 5.8 – Optimisation de n et η - Partie 1

Comme le témoigne la figure 5.8, l'optimisation des différents paramètres est effectuée par le biais d'une recherche sur grille. En effet, nous voyons la grille suivante définie :

— $n \in \{100, 150, 200, 250, 300, 350, 400, 450, 500\}$;
— $\eta \in \{0.025, 0.05, 0.1, 0.3\}$;
— La profondeur maximale $\in \{2, 3, 4, 5, 6\}$.

Ainsi, pour chaque combinaison de ces paramètres, le $RMSE$ est calculé par validation croisée k-folds. Nous remarquons alors que pour $\eta \in \{0.05, 0.1, 0.3\}$, le $RMSE$ n'est pas stable, et croît avec le nombre d'itérations. Cette situation traduit d'un sur-apprentissage des données lorsque le taux d'apprentissage et le nombre d'itérations sont trop élevés. Pour $\eta = 0.025$, nous remarquons que le $RMSE$ se stabilise à partir de 350 itérations. De ce fait, nous pouvons noter qu'avec un taux d'apprentissage plus faible et un nombre suffisant d'itérations, le modèle ajuste bien les données tout en ne sur-apprenant pas. Ainsi, nous retenons dans un premier temps $n = 350$ et $\eta = 0.025$. L'optimisation détaillée de ces paramètres est illustrée par le biais de la figure 5.9.

Après avoir obtenu de possibles valeurs de n et η optimales, nous avons optimisé la profondeur maximale, le nombre d'observations minimal de chaque feuille de l'arbre, ainsi que le nombre de variables considérées aléatoirement à chaque itération. Effectivement, nous avons retenu les paramètres minimisant le $RMSE$ par validation croisée k-folds pour $n = 350$ et $\eta = 0.025$. A présent, nous pouvons optimiser plus en détail ces deux derniers paramètres. Nous définissons alors une grille de recherche plus fine :

— $n \in \{100, 200, 300, ..., 2000\}$;
— $\eta \in \{0.0025, 0.00375, 0.00625, 0.025\}$.

Nous remarquons logiquement que plus η est faible, plus n doit être grand pour que le modèle apprenne correctement des données. Aussi, il est important de noter que cette grille de paramètres permet de créer des modèles ne sur-apprenant pas, et dont le $RMSE$ converge vers une unique faible valeur. Finalement, nous décidons de retenir le modèle avec les paramètres $\eta = 0.025$ et $n = 500$, car sa précision est de qualité tout en ayant un temps d'exécution raisonnable.

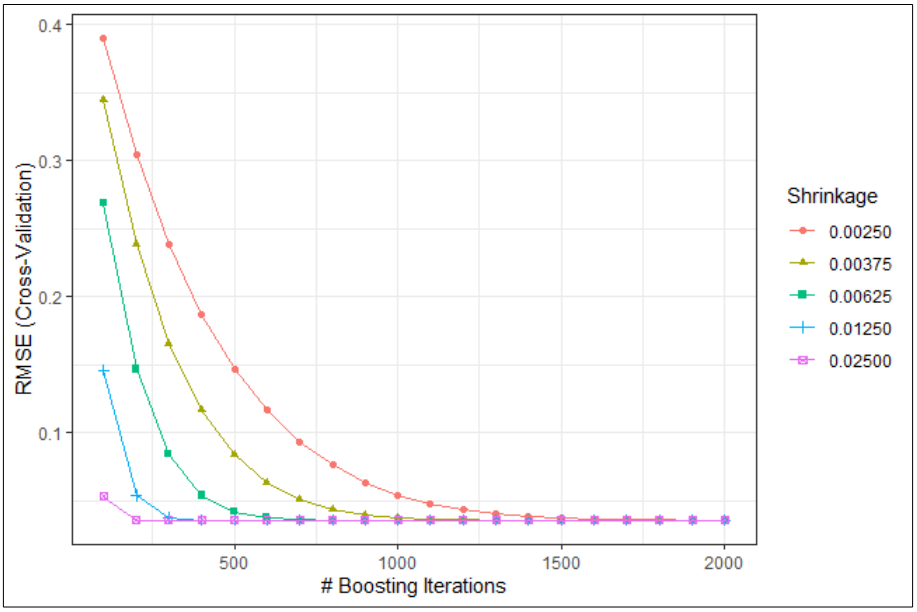


FIGURE 5.9 – Optimisation de n et η - Partie 2

Importance des variables

Tout comme pour le Random Forest, seules les 15 variables les plus importantes sont affichées ici. Ainsi, nous remarquons que le nombre de pièces et l'âge du client segmentent fortement l'espace, et par conséquent influent sur la fréquence de sinistres des assurés ayant un appartement. Ensuite, le montant de capital assuré, l'ancienneté du client et la catégorie socioprofessionnelle « inactifs » jouent de façon modérée sur le risque. Pour finir, les habitants de rez-de-chaussée, le taux d'objets de valeur ou encore la qualité de l'occupant influent faiblement sur la fréquence de sinistres.

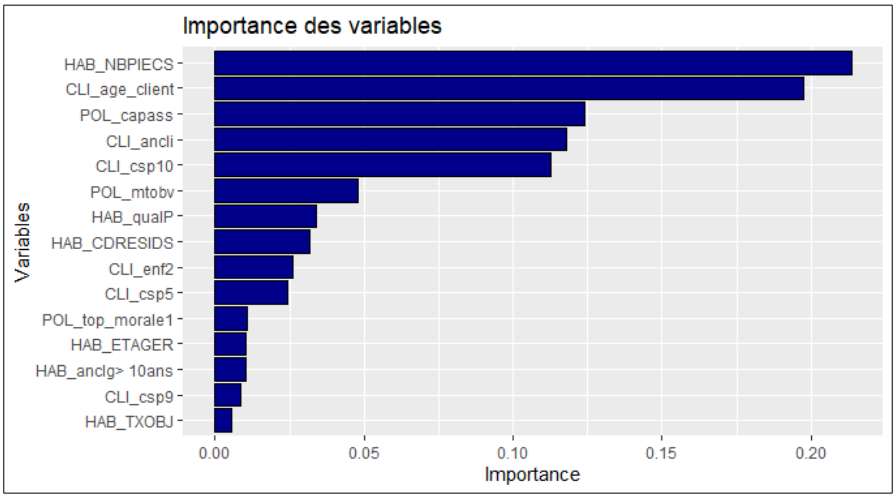


FIGURE 5.10 – Importance des variables XGBoost - Fréquence appartements

Métriques d'évaluation

Le tableau de métriques d'évaluation ci-dessous montre que le modèle ajuste plutôt bien les données. En effet, le MSE ainsi que l'erreur globale justifient ces propos. Néanmoins, l'indice de Gini de 19% témoigne d'une qualité de segmentation du modèle moins bonne que celle des GLMs pénalisés.

Métriques XGBoost - Fréquence appartement			
Base	MSE	Erreur globale	Gini
Train	$1,3 \times 10^{-3}$	2%	19%
Test	$1,3 \times 10^{-3}$	0%	19%

TABLE 5.5 – Métriques d'évaluation XGBoost - Fréquence appartements

5.3.2.2 Modélisation du risque maison

Le processus d'optimisation des paramètres du XGBoost est similaire quelque soit le type d'habitation. Ainsi, identiquement aux autres méthodes, nous présentons dans cette section uniquement l'importance des variables et les métriques d'évaluation du modèle optimal.

Importance des variables

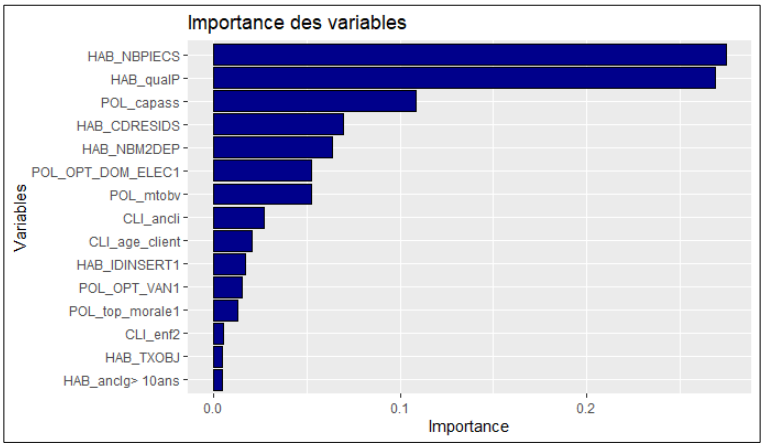


FIGURE 5.11 – Importance des variables XGBoost - Fréquence maisons

Ainsi, nous voyons que le nombre de pièces et que la qualité de l'occupant influent fortement sur la fréquence de sinistres. Ensuite, viennent le montant de capital assuré, le type de résidence (principale ou secondaire) et la surface des dépendances. Le reste des variables telles que la possession d'un insert ou le taux d'objets de valeur a une influence moindre sur le risque.

Métriques d'évaluation

Le MSE nous indique que le modèle ajuste bien les données. Aussi, XGBoost améliore la qualité de segmentation par rapport aux GLMs pénalisés (entre 23% et 29%). Néanmoins, lorsque l'on observe l'erreur globale sur la base test, nous nous apercevons que le modèle sous-estime la fréquence de sinistres de 5% sur cette base.

Métriques XGBoost - Fréquence maisons			
Base	MSE	Erreur globale	Gini
Train	$5,6 \times 10^{-3}$	0%	30%
Test	$5,9 \times 10^{-3}$	-5%	30%

TABLE 5.6 – Métriques d'évaluation XGBoost - Fréquence maisons

5.4 Synthèse des résultats

Dans cette section, nous fournissons une synthèse des résultats obtenus par le biais des GLMs pénalisés et des méthodes d'agrégation d'arbres. Pour cela nous partageons les plus importantes variables ainsi que les métriques d'évaluation pour chaque modèle.

Classement des 5 variables les plus importantes par modèle

Modèle	**Top 1**	**Top 2**	**Top 3**	**Top 4**	**Top 5**
Appartements					
GLM	Type résid.	Nb pièces	Surf. dép.	Insert	Csp inactif
Random Forest	Cap. ass.	Surf. dép.	Mt. objets val.	Nb pièces	Âge client
XGBoost	Nb pièces	Âge client	Cap. ass.	Anc. client	Csp inactif
Maisons					
GLM	Qualité occ.	Nb pièces	Type résid.	Insert	Surf. dép.
Random Forest	Nb pièces	Qualité occ.	Cap. ass.	Surf. dép.	Énergies renouv.
XGBoost	Nb pièces	Qualité occ.	Cap. ass.	Type résid.	Surf. dép.

TABLE 5.7 – Classement des 5 variables les plus importantes par modèle - Fréquence

Métriques d'évaluation

Modèle	**MSE Train**	**MSE Test**	**Erreur glob. Train**	**Erreur glob. Test**	**Gini Train**	**Gini Test**
Appartements						
GLM	$1,3 \times 10^{-3}$	$1,3 \times 10^{-3}$	0%	0%	27%	27%
Random Forest	$1,3 \times 10^{-3}$	$1,3 \times 10^{-3}$	0%	-1%	14%	14%
XGBoost	$1,3 \times 10^{-3}$	$1,3 \times 10^{-3}$	2%	0%	19%	19%
Maisons						
GLM	$5,6 \times 10^{-3}$	$5,6 \times 10^{-3}$	0%	0%	29%	29%
Random Forest	$5,6 \times 10^{-3}$	$5,9 \times 10^{-3}$	0%	-6%	27%	27%
XGBoost	$5,6 \times 10^{-3}$	$5,9 \times 10^{-3}$	0%	-5%	30%	30%

TABLE 5.8 – Métriques d'évaluation globales - Fréquence

Chapitre 6

Modèles de propension

6.1 Cadre et objectifs

6.1.1 Introduction

Après avoir modélisé la fréquence de sinistres pour les assurés de notre portefeuille, nous cherchons à présent, à travers ce chapitre, à modéliser la probabilité d'occurrence d'un sinistre grave, lorsque les assurés ont subi un sinistre. Nous rappelons l'équation définie en section 2.1.1.3 permettant de calculer la prime pure p_i pour chaque assuré i :

$$p_i = f_i \times [g_i \times c_i^{\text{grave}} + (1 - g_i) \times c_i^{\overline{\text{grave}}}]$$

Avec g_i la probabilité d'occurrence d'un sinistre grave pour l'assuré i, et de ce fait, $1 - g_i$ la probabilité d'occurrence d'un sinistre attritionnel pour ce même assuré. Le lecteur est invité à se référer à la section 2.1.4 pour obtenir une étude détaillée de la distinction entre sinistres attritionnels et graves. Ainsi, nous cherchons dans ce chapitre à estimer g pour chaque assuré i.

Les modèles de propension répondent à une problématique différente de celle des modèles de fréquence. Effectivement, ils cherchent à déterminer, pour chaque assuré i, la classe à laquelle il appartient, ou plus finement, la probabilité d'appartenance à chacune des classes. Les assurés de notre portefeuille ayant eu au maximum un sinistre grave dans l'année, nous répondons dans cette étude à un problème binaire :

$$Y = \begin{cases} 1 \text{ Si l'assuré a eu un sinistre grave au cours de l'année} \\ 0 \text{ sinon} \end{cases}$$

Avec 1,1% de sinistres graves parmi l'ensemble des sinistres, nous avons très peu d'informations sur les profils des assurés ayant des sinistres graves. Ainsi, contrairement aux modèles de fréquence et de coût moyen, les modèles de propension étudiés dans ce chapitre ne distingueront pas la sinistralité des appartements de la sinistralité des maisons. En effet, nous considérons ici que la distinction n'est pas nécessaire étant donné du faible nombre de données.

Les métriques d'évaluation des modèles de classification sont différentes de celles des modèles de régression. Par conséquent, nous les définissons dans la section suivante. Ensuite, nous présentons, pour la régression logistique pénalisée ainsi que le Random Forest, l'optimisation des paramètres, l'importance des variables et les métriques d'évaluation. Enfin, une synthèse des résultats obtenus est effectuée.

6.1.2 Métriques d'évaluation d'un classifieur binaire

Afin d'affecter à chaque assuré ayant eu au moins un sinistre une classe (assuré ayant un sinistre grave au cours de l'année ou non), il est nécessaire de définir un seuil s. Effectivement, pour une probabilité supérieure à ce seuil s, on prédit $\hat{Y}_i = 1$, sinon, on prédit $\hat{Y}_i = 0$. Ainsi, une fois le seuil s déterminé, il est possible de quantifier la performance du classifieur à partir de sa matrice de confusion :

	Négatifs prédits	**Positifs prédits**
Négatifs observés	*TN*	*FP*
Positifs observés	*FN*	*TP*

FIGURE 6.1 – Matrice de confusion

Avec TN (True Negatives) le nombre d'observations négatives ($Y = 0$) correctement prédites, FP (False Positives) le nombre d'observations positives ($Y = 1$) incorrectement prédites, FN (False Negatives) le nombre d'observations négatives ($Y = 0$) incorrectement prédites, TP (True Positives) le nombre d'observations positives ($Y = 1$) correctement prédites.

Plusieurs mesures peuvent être construites à partir de cette matrice de confusion. La plus globale est l'exactitude de prédiction (Accuracy) :

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Et le taux d'erreur défini comme $1 - \text{Accuracy}$. Aussi, nous pouvons définir le taux de prédiction correcte des observations positives (sensibilité) et négatives (spécificité) :

$$\text{Sensibilité} = \frac{TP}{TP + FN}$$

$$\text{Spécificité} = \frac{TN}{TN + FP}$$

Dans une situation de jeu de données déséquilibré, les observations identifiées comme négatives ($TN + FP$) dominent largement les observations identifiées comme positives ($TP + FN$). De ce fait, l'exactitude de précision (Accuracy) n'est pas une mesure adaptée car le meilleur classifieur sera celui qui prédit correctement toutes les observations négatives, et un tel classifieur est sans intérêt. En revanche, la sensibilité et la spécificité sont insensibles à ce déséquilibre et seront donc utilisées.

La courbe ROC est une représentation graphique traduisant la relation existante entre la sensibilité et la spécificité d'un classifieur binaire. De cette dernière, découle l'AUC (Area Under Curve), qui est une mesure globale de la performance d'un modèle. Elle varie entre 0 (performance moindre) et 1 (performance parfaite). Ainsi, plus l'AUC d'un modèle est élevée, plus la classification est de qualité.

6.2 Régression logistique pénalisée

Dans cette section, nous présentons les résultats de trois GLMs pénalisés : le GLM LASSO, Ridge et Elastic Net. La variable réponse Y étant à valeurs dans $\{0\ ;1\}$, nous considérons un GLM Binomial, appelé aussi régression logistique. Cette régression modélise la probabilité qu'une observation appartienne à une catégorie donnée (par exemple $\mathbb{P}(Y = 1|X = x)$). Dans notre étude, nous avons deux catégories parmi les assurés sinistrés : les assurés ayant eu un sinistre grave et les assurés n'ayant pas eu de sinistres graves. La fonction de lien canonique de la régression logistique est la fonction logit, aussi connu sous le nom de log odds :

$$log\left(\frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=1|X=x)}\right)$$

Son inverse est la fonction logistique, qui prend un nombre réel quelconque et le projette sur l'intervalle $[0, 1]$ comme souhaité pour modéliser la probabilité d'appartenir à une classe. Ainsi, pour chaque assuré sinistré i, le modèle peut s'écrire :

$$\hat{Y_i} = log\left(\frac{\mathbb{P}(Y_i=1|X_i=x_i)}{1-\mathbb{P}(Y_i=1|X_i=x_i)}\right) = X_i'\beta$$

L'estimation des paramètres est effectuée en maximisant la vraisemblance pénalisée par la méthode Elastic Net :

$$\max_{\beta_0,\beta} \frac{1}{N}\sum_{i=1}^{N}\left\{y_i(x_i^T\beta+\beta_0) - log(1+e^{x_i^T\beta+\beta_0})\right\} - \lambda[\alpha\left\|\beta\right\|_1 + (1-\alpha)\left\|\beta\right\|_2]$$

Pour rappel, les régressions LASSO et Ridge peuvent être vues comme des cas particuliers de l'Elastic Net avec respectivement $\alpha = 1$ et $\alpha = 0$.

Aussi, la déviance correspondante est définie de la manière suivante :

$$D_{binomial} = -2\sum_{i=1}^{N}\{y_i log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i)\}$$

6.2.1 La modélisation de la probabilité d'occurrence d'un sinistre grave

Dans cette section, nous modélisons la probabilité d'occurrence d'un sinistre grave par le biais de plusieurs régressions logistiques pénalisées. Pour ce faire, nous présentons tout d'abord l'optimisation des paramètres des GLMs pénalisés que sont λ et α (cf. 4.1.4). Ensuite, nous analysons les coefficients de certaines variables que nous jugeons intéressantes pour interpréter les modèles. Enfin, nous exposons quelques métriques d'évaluation des modèles, nous permettant de quantifier leurs qualités. Les différentes étapes citées précédemment étant similaires pour chaque GLM, seules celles du GLM Elastic Net sont présentées en détail. Toutefois, l'ensemble des résultats est exposé afin de fournir au lecteur une synthèse.

Optimisation des paramètres

Comme pour la modélisation de la fréquence, le choix des paramètres optimaux se fait par validation croisée. Ainsi, pour les trois GLMs pénalisés, les paramètres maximisant l'AUC ont été retenus. A présent, considérons un $\alpha_i \in [0, 1]$. Pour cet α_i fixé, nous cherchons le λ optimal. En d'autres termes, nous cherchons le λ qui maximise l'AUC. Nous observons facilement sur la figure 6.2 que l'AUC croît jusqu'à atteindre un maximum en $\lambda = 8 \times 10^{-4}$, puis décroît progressivement jusqu'à atteindre un minimum ($AUC = 0,52$). Par conséquent, pour cet α_i, le λ optimal est $\lambda = 8 \times 10^{-4}$. Cette étape est effectuée autant de fois qu'il y a de $\alpha \in [0, 1]$ dans notre grille de

recherche. Ainsi, après un certain nombre d'itérations, le couple optimal obtenu pour l'Elastic Net est $(\alpha = 0,6\ ;\ \lambda = 8 \times 10^{-4})$. Pour rappel les régressions LASSO et Ridge sont des cas particuliers de l'Elastic avec respectivement $\alpha = 1$ et $\alpha = 0$.

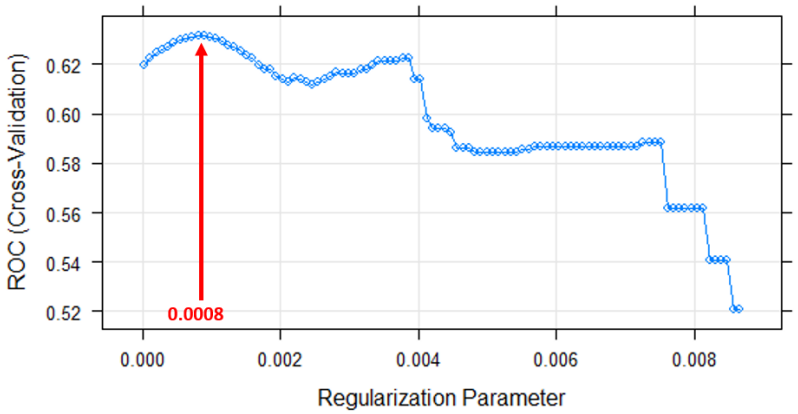


FIGURE 6.2 – Régression logistique - Optimisation de λ

Analyse des coefficients

Malgré un paramètre de régularisation faible ($\lambda = 8 \times 10^{-4}$), l'Elastic Net épure efficacement le modèle en forçant près de la moitié des coefficients à 0. De ce fait, la pénalisation permet de conserver la qualité d'ajustement du modèle tout en simplifiant son interprétation. La figure 6.3 nous permet d'interpréter la valeur des coefficients associés aux variables : nombre de pièces, étage, surface des dépendances, possession d'une piscine, de biens au titre de l'option « Énergies renouvelables », et d'un jardin. Les modalités de références permettant la modélisation par GLM pour chacune de ces variables sont respectivement : deux pièces ou moins, rez-de-chaussée, pas de dépendances et la possession de chacun des aménagements extérieurs. Concernant le nombre de pièces, nous voyons que les grandes habitations (7 pièces ou plus) sont moins exposées au risque de sinistres graves que les petites habitations. La variable étage nous informe que les maisons (*HAB_ETAGEM*) ont plus de chances de voir surgir un sinistre dit grave que les appartements. Cette variable nous confirme une fois de plus que pour la garantie incendie, la sinistralité des appartements et des maisons est bien différente. La possession de biens au titre de l'option « Énergies renouvelables » augmente significativement le risque, tandis que la possession d'une piscine ou d'un jardin l'augmente dans une moindre mesure. Aussi, la possession de grandes dépendances (supérieures ou égales à $100m^2$) diminue ce risque, tandis que de plus petites dépendances l'augmente.

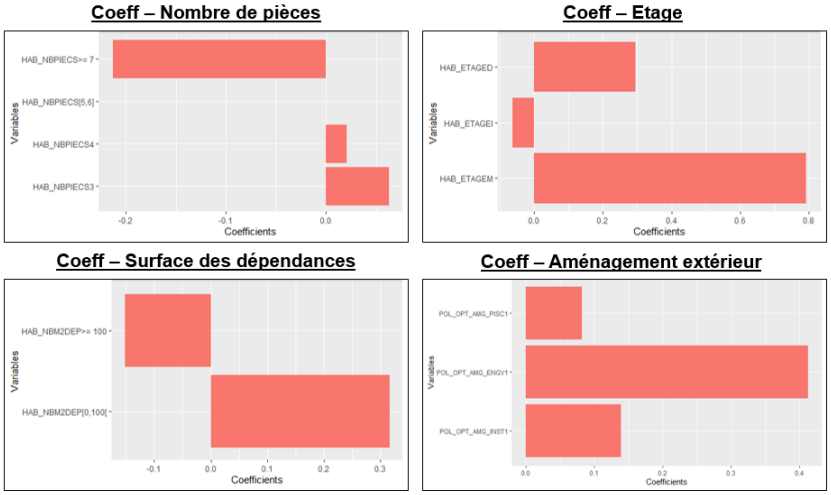


FIGURE 6.3 – Régression logistique - Analyse des coefficients

En amont de la présentation des métriques d'évaluation des modèles, montrons comment le seuil s présenté en section 6.1.2 a été sélectionné de façon optimale. Pour chaque modèle construit, nous pouvons associer une courbe ROC calculée sur la base de test :

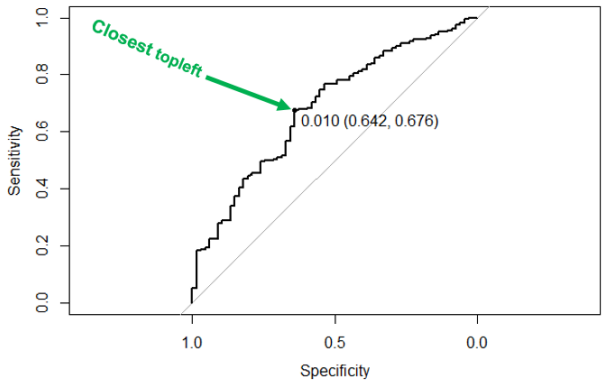


FIGURE 6.4 – Courbe ROC et seuil optimal

Ensuite, à la courbe ROC est associé un seuil optimal. Ce dernier est définit comme étant le point le plus proche du coin supérieur gauche (*closest topleft* en anglais), et donc le point ayant le meilleur couple sensibilité et spécificité. Mathématiquement, ce seuil doit satisfaire le critère suivant :

$$\text{Seuil optimal} = \min[(1 - sensibilité)^2 + (1 - spécificité)^2]$$

Une fois le seuil optimal défini, une sensibilité et une spécificité optimales peuvent y être associées. Ainsi, pour l'Elastic Net, nous pouvons voir que pour le seuil optimal de 0.01, la sensibilité est de 67,6% et la spécificité de 64,2%. Présentons à présent les métriques d'évaluation pour tous les GLMs testés.

Métriques d'évaluation des modèles

Modèle	Couple $(\alpha;\lambda)$	Seuil Test	Sens. Train	Sens. Test	Spec. Train	Spec. Test	AUC Train	AUC Test
LASSO	$(1; 7,5 \times 10^{-4})$	$1,0\%$	$67,6\%$	$70,1\%$	$61,5\%$	$58,2\%$	$68,3\%$	$67,3\%$
ELASTIC NET	$(0,6; 8,4 \times 10^{-4})$	$1,0\%$	$64,6\%$	$67,6\%$	$68,1\%$	$64,2\%$	$69,7\%$	$68,5\%$
RIDGE	$(0; 1,1 \times 10^{-1})$	$1,0\%$	$59,8\%$	$63,5\%$	$67,0\%$	$61,2\%$	$67,4\%$	$65,4\%$

TABLE 6.1 – Métriques d'évaluation GLMs - Régression logistique

Nous constatons que les métriques d'évaluation sont assez proches quel que soit le GLM pénalisé. Toutefois, le GLM Elastic Net est retenu pour plusieurs raisons. Tout d'abord, cette pénalisation obtient les meilleurs résultats en terme d'AUC (68,5%) tout en étant robuste. Effectivement, la variation d'AUC entre la base d'entraînement et la base test est de 1,72%. De plus, parmi les 3 GLMs, elle possède le meilleur couple sensibilité et spécificité. De fait, le GLM Elastic Net arrive à identifier correctement 67,6% des sinistres graves et 64,2% des sinistres non graves. Bien que ces métriques peuvent paraître peu élevées, elles restent raisonnables étant donné le très faible nombre de sinistres graves.

6.3 Modélisation par Random Forest

Dans cette section, nous modélisons la probabilité d'occurrence d'un sinistre grave par le biais d'un Random Forest. Comme pour les GLMs pénalisés, nous présentons tout d'abord l'optimisation des paramètres afin d'obtenir le modèle le plus robuste possible. Ensuite nous analysons l'importance des variables de ce modèle. Enfin, nous commentons les métriques d'évaluation associées aux modèles de classification.

Optimisation des paramètres

Comme énoncé en section 4.2.3, le Random Forest a plusieurs paramètres tels que :
— Le nombre d'arbres, noté T ;
— Le nombre de variables explicatives considérées aléatoirement à chaque noeud, noté m ;
— Le nombre minimal d'observations comprises dans les dernières feuilles, noté f, qui correspond au critère d'arrêt de chaque arbre.

Comme dans la section 5.3.1, le nombre d'arbres T ne sera pas optimisé, car le risque de sur-apprentissage n'augmente pas avec le nombre d'arbres considéré. Toutefois, T à été choisi de telle sorte que ce paramètre génère des résultats robustes et précis ($T = 500$). Ainsi, nous détaillons ici l'optimisation de m et de f. La grille de valeur pour m est similaire à celle utilisée lors de la section 5.3.1. Effectivement, nous avons retenu la grille de valeurs entières $[\![1, 15]\!]$. De plus, plusieurs critères d'arrêt f ont été testés dans le but de construire des arbres plus ou moins variants à partir de la base d'entraînement. La grille suivante a donc été retenue pour f : 1%, 2,5%, 5% et 10% du nombre d'observations de la base d'entraînement. Enfin, tout comme pour les régressions logistiques pénalisées, les paramètres sont optimisés par le biais d'une validation croisée et d'un critère d'évaluation : l'AUC.

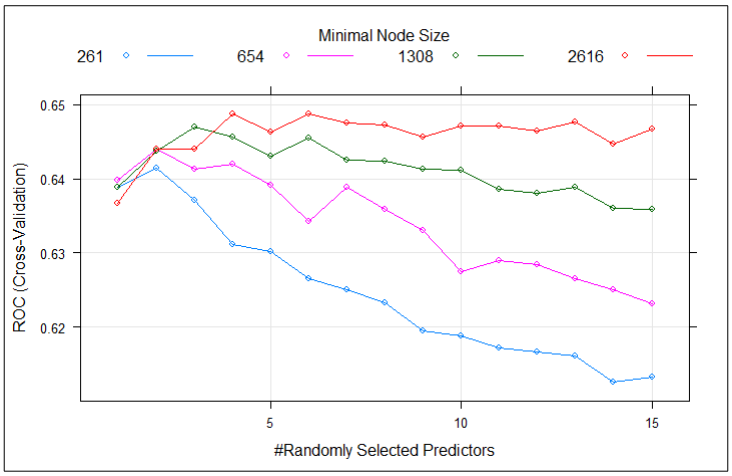


FIGURE 6.5 – Optimisation Random Forest - AUC

Sur la figure 6.5, nous remarquons que lorsque $f = 2616$, soit 10% du nombre d'observations de la base d'entraînement, l'AUC est nettement supérieure. De plus, le fait d'opter pour un f plus élevé implique un risque plus faible de sur-apprentissage. Concernant le nombre de variables explicatives m, nous voyons que l'AUC se stabilise à partir de $m = 4$ (pour $f = 2616$). Ainsi, pour minimiser le risque de corrélation entre les arbres et le temps d'exécution des modèles, nous optons pour $m = 4$.

Importance des variables

Après avoir optimisé les paramètres m et f, et fixé $T = 500$, nous pouvons étudier l'importance des variables du modèle obtenu. La figure 6.6 présente uniquement les 10 variables les plus importantes du modèle. Ainsi, comme dans les GLMs pénalisés, nous retrouvons le nombre de pièces (*HAB_NBPIECS*), la surface des dépendances (*HAB_NBM2DEP*), ainsi que l'étage (*HAB_ETAGE*) parmi les variables les plus importantes. Également, nous pouvons noter que le montant d'objets de valeur (*POL_mtobv*) et le montant de capital assuré (*POL_capass*) influent très fortement sur le risque d'avoir un sinistre considéré comme grave. Enfin, le reste des variables a une influence moins significative telle que la qualité de l'occupant (propriétaire ou locataire).

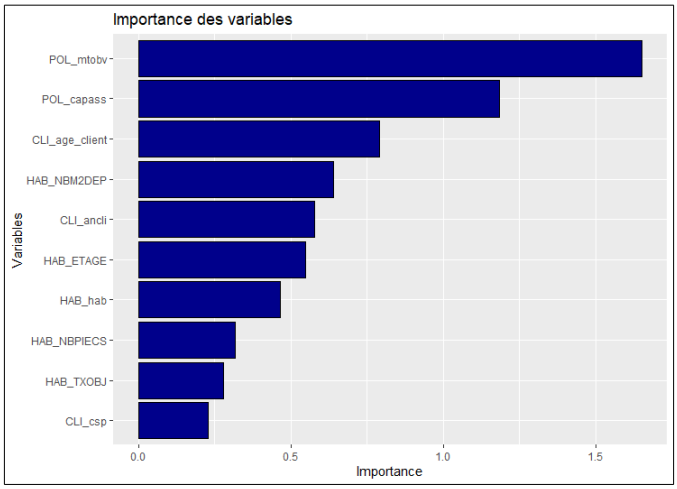


FIGURE 6.6 – Importance des variables - Random Forest propension

Métriques d'évaluation

A partir de la table de métriques d'évaluation ci-dessous nous voyons que le modèle obtient de bons résultats sur la base d'entraînement. Néanmoins, nous remarquons un sur-apprentissage de nos données. Cela s'explique car le Random Forest laisse une complète liberté sur la forme du lien entre la variable réponse et les variables explicatives. Par conséquent, le modèle créé est trop variant. Les résultats obtenus sur la base test restent cependant comparables à ceux obtenus par les GLMs. Effectivement, le Random Forest permet d'identifier 54,7% des sinistres graves et 64,5% des sinistres non graves.

Métriques Random Forest - Propension					
Base	Seuil	Accuracy	Sens.	Spec.	AUC
Train	$1,00\%$	$82,9\%$	$71,5\%$	$83,0\%$	$84,0\%$
Test	$7,95 \times 10^{-3}$	$64,4\%$	$54,7\%$	$64,5\%$	$61,4\%$

TABLE 6.2 – Métriques Random Forest - Modèle de propension

6.4 Synthèse des résultats

Dans cette section, nous fournissons une synthèse des résultats obtenus par le biais des GLMs pénalisés et du Random Forest. Pour cela nous partageons les plus importantes variables ainsi que les métriques d'évaluation pour chaque modèle.

Modèle	**Top 1**	**Top 2**	**Top 3**	**Top 4**	**Top 5**
GLM	Étage	Aménagement ext.	Surf. dép.	Anc. log.	Nb. pièces
Random Forest	Mt. objets val.	Capital assuré	Âge client	Surf. dép.	Anc. client

TABLE 6.3 – Classement des 5 variables les plus importantes par modèle - Propension

Modèle	Seuil Test	Sens. Train	Sens. Test	Spec. Train	Spec. Test	AUC Train	AUC Test
GLM	$1,00\%$	$64,6\%$	$67,6\%$	$68,1\%$	$64,2\%$	$69,7\%$	$68,5\%$
Random Forest	$7,95 \times 10^{-3}$	$71,5\%$	$54,7\%$	$83,0\%$	$64,5\%$	$84,0\%$	$61,4\%$

TABLE 6.4 – Métriques d'évaluation globales - Propension

Chapitre 7

Modèles de coût moyen

7.1 Cadre et objectifs

Après avoir modélisé la fréquence de sinistres (f_i) et la probabilité d'occurrence d'un sinistre grave (g_i), nous modélisons, à travers ce chapitre, le coût moyen. Nous rappelons la formule de la prime pure présentée en section 2.1.1.3 :

$$p_i = f_i \times [g_i \times c_i^{\text{grave}} + (1 - g_i) \times c_i^{\overline{\text{grave}}}]$$

Avec $c_i = \frac{\text{Charge totale des sinistres}}{\text{Nombre de sinistres}}$.

Le risque incendie en assurance Habitation est défini comme un risque d'intensité. En d'autres termes, la probabilité d'occurrence de ce type de sinistres est très faible, mais lorsque qu'il se produit, le coût est souvent très élevé. De plus, nous distinguons dans ce mémoire l'étude des sinistres dits attritionnels et des sinistres dits graves (cf. 2.1.4). En effet, cette distinction est nécessaire pour la garantie incendie car la sévérité de ces deux types de sinistres est très différente. Elle est moins nécessaire pour les garanties telles que le bris de glace ou le dégât des eaux. Nous pouvons aisément le concevoir en comparant les coûts de dédommagement de la combustion d'une gazinière (qui peuvent être à hauteur de quelques milliers d'euros) à ceux de la combustion d'un immeuble tout entier (qui peuvent être à hauteur de plusieurs millions d'euros). C'est pourquoi, dans ce chapitre, nous séparons la modélisation du coût moyen attritionnel, noté $c_i^{\overline{\text{grave}}}$, de celle du coût moyen grave, noté c_i^{grave}.

Nous commençons par la modélisation du coût moyen attritionnel qui est comparable avec celles effectuées lors des chapitres 5 et 6. Effectivement, tout en séparant l'étude par type de d'habitation (appartement et maison), nous comparons les GLMs, qui sont des méthodes classiques de tarification avec l'apprentissage automatique qui comprend des méthodes plus innovantes.

Par la suite, nous modélisons le coût moyen grave qui répond à une problématique bien différente. La rareté de ce type de sinistres implique un nombre de données très réduit et ne permet pas d'effectuer une modélisation standard. Nous comparons alors une approche simple qui est l'affectation de la moyenne par arbre de régression, avec la modélisation suite au sur-échantillonnage par la méthode SMOTE présentée en section 4.4.1.

7.2 Coût moyen attritionnel

7.2.1 GLM Gamma pénalisés

Dans cette section, nous modélisons le coût moyen attritionnel, noté $c_i^{\overline{\text{grave}}}$ par le biais de plusieurs GLMs pénalisés : le GLM Lasso, Ridge et Elastic Net. La variable réponse Y représente le coût du dédommagement moyen, et de ce fait, est comprise dans $\mathbb{R}_+^*$. Ainsi, nous considérons un GLM Gamma avec la fonction de lien $g = log$, ce qui signifie que pour chaque observation i :

— $Y_i|X_i = x_i$ suit une loi gamma ;

— $log(\mu(X_i)) = log(\mathbb{E}[Y_i|X_i = x_i]) = X_i^T\beta$.

Néanmoins, pour modéliser le coût moyen attritionnel, le nombre de sinistres attritionnels, noté ω, doit nécessairement être pris en compte. Nous supposons alors pour une observation i que $\mu(X_i) = \mathbb{E}\left(\frac{Y_i}{\omega_i}|X_i\right)$. Le modèle est donc légèrement modifié : $log\left(\frac{\mathbb{E}[Y_i|X_i]}{\omega_i}\right) = X_i^T\beta$, et peut se réécrire : $log(\mathbb{E}[Y_i|X_i]) = log(\omega_i) + X_i^T\beta$. L'offset (correspondant à $log(\omega_i)$) peut être vu comme une nouvelle variable du modèle avec un coefficient β constant égal à 1.

L'estimation des paramètres est effectuée en maximisant la vraisemblance pénalisée par la méthode Elastic Net :

$$\max_{\beta_0,\beta} \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{y_i}{x_i^T\beta+\beta_0} + log(x_i^T\beta+\beta_0)\right\} - \lambda[\alpha\left\|\beta\right\|_1 + (1-\alpha)\left\|\beta\right\|_2]$$

Pour rappel, les régressions LASSO et Ridge peuvent être vues comme des cas particuliers de l'Elastic Net avec respectivement $\alpha = 1$ et $\alpha = 0$.

Aussi, la déviance correspondante est définie de la manière suivante :

$$D_{gamma} = 2\sum_{i=1}^{N}\left\{-log\left(\frac{y_i}{\hat{y}_i}\right) + \frac{(y_i-\hat{y}_i)}{\hat{y}_i}\right\}$$

7.2.1.1 La modélisation du risque appartement

Nous étudions tout d'abord le coût moyen attritionnel des assurés ayant un appartement. Comme dans les chapitres précédents, nous analysons l'optimisation des paramètres λ et α, les coefficients des GLMs optimisés ainsi que leurs métriques d'évaluation. Pour la même raison que celle des chapitres 5 et 6, nous présentons les étapes citées précédemment uniquement pour le GLM Elastic Net. Toutefois, l'ensemble des résultats est exposé afin de fournir une synthèse au lecteur.

Optimisation des paramètres

Le processus d'optimisation des paramètres étant similaire à celui du chapitre 5, il n'est pas détaillé. Effectivement, λ et α sont également optimisés par k-fold validation croisée, et doivent minimiser un critère qui est la déviance définie auparavant.

Modèle	λ	α
LASSO	$8,2 \times 10^{-2}$	1
ELASTIC NET	$1,8 \times 10^{-1}$	0.1
RIDGE	$1,75$	0

TABLE 7.1 – Paramètres optimisés GLM Gamma - Appartement

Analyse des coefficients

Analysons à présent les coefficients issus du GLM Elastic Net. Ici, nous avons décidé de représenter les variables les plus importantes et les plus facilement interprétables. De plus, la pénalisation Elastic Net permet de forcer un nombre important de coefficients à 0, ce qui permet de simplifier l'interprétation du modèle. Ainsi, la figure 7.1 permet d'interpréter la valeur des coefficients des variables suivantes : nombre de pièces, étage, surface des dépendances et ancienneté du logement. Nous remarquons que le coût moyen croît avec le nombre de pièces, mise à part pour les appartements de 3 pièces (ce qui peut paraître contre-intuitif). Ainsi, lorsqu'un incendie survient, les grands appartements ont en moyenne un coût de dédommagement plus élevé. Les appartements situés au dernier étage ont des coûts moyens plus élevés que les appartement situés au rez-de-chaussée ou aux étages intermédiaires. Concernant les dépendances, nous remarquons que le coût moyen attritionnel décroît lorsque la surface croît. Pour finir, l'ancienneté du logement influe sur le coût du sinistre incendie. Effectivement, plus l'appartement est ancien, plus le coût est élevé.

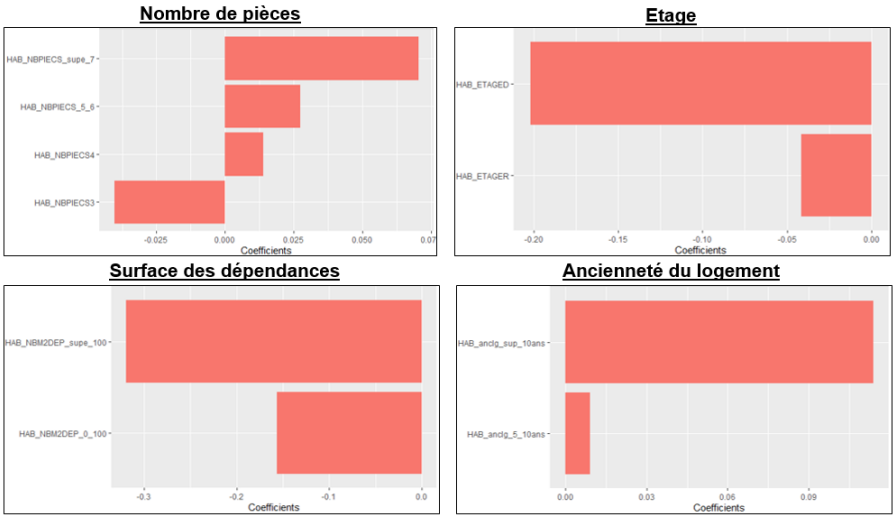


FIGURE 7.1 – GLM Gamma Appartements - Analyse des coefficients

Métriques d'évaluation

La figure 7.2 présente les métriques d'évaluation pour les trois GLMs pénalisés. Tout d'abord, nous remarquons grâce à l'erreur globale que les modèles obtenus sont moins robustes que ceux modélisant la fréquence du chapitre 5. Cela peut s'expliquer par la quantité de données. En effet, les bases permettant la modélisation du coût moyen comprennent uniquement les individus ayant subi au moins un sinistre, et de ce fait sont beaucoup plus petites que les bases permettant la modélisation de la fréquence qui comprennent la totalité des individus. Néanmoins, bien que le modèle issu de la pénalisation Elastic Net n'obtienne pas les meilleurs résultats en terme de prédiction, il reste le modèle segmentant le mieux la population avec un indice de Gini sur la base test de 21%. De plus, il simplifie efficacement le modèle en forçant certains coefficients à 0. Nous privilégions donc cette pénalisation.

Modèle	Couple $(\alpha;\lambda)$	RMSE Train	RMSE Test	Err. glob. Train	Err. glob. Test	Gini Train	Gini Test	Déviance Test
LASSO	$(1;8,2\times10^{-2})$	12958	11850	-1%	9%	10%	10%	1075
ELASTIC NET	$(0,1;1,8\times10^{-1})$	12801	12072	-4%	9%	21%	21%	1107
RIDGE	$(0;1,75)$	12862	11845	-5%	5%	14%	13%	1070

TABLE 7.2 – Métriques d'évaluation GLMs - Coût moyen attritionnel appartements

7.2.1.2 La modélisation du risque maison

Nous étudions maintenant le coût moyen attritionnel des assurés ayant une maison. Comme précédemment, seuls les résultats du GLM Elastic Net sont présentés en détail.

Optimisation des paramètres

Le tableau 7.3 présente les paramètres optimaux pour chaque GLM pénalisé.

Modèle	λ	α
LASSO	$1,5 \times 10^{-1}$	1
ELASTIC NET	$3,4 \times 10^{-2}$	0.7
RIDGE	$13,4$	0

TABLE 7.3 – Paramètres optimisés GLM Gamma - Maison

Analyse des coefficients

Comme précédemment, nous analysons ici les coefficients de certaines variables du GLM Elastic Net. Ainsi, la figure 7.2 permet d'interpréter la valeur des coefficients des variables suivantes : nombre de pièces, l'ancienneté du logement, la surface des dépendances, possession d'une piscine, de biens au titre de l'option « Énergies renouvelables », et d'un jardin. Nous remarquons que les petites maisons (entre 1 et 4 pièces) ont un coût moyen plus élevé que les grandes maisons (5 pièces ou plus). Aussi, comme pour les appartements, plus le logement est ancien, plus le coût moyen est élevé. Les petites dépendances (moins de 100 m^2) sont moins risquées en terme de coût que les grandes dépendances (plus de 100 m^2). Pour finir, la possession d'une piscine ou d'un jardin diminue le risque, tandis que la possession de biens au titre de l'option « Énergies renouvelables » l'accroît.

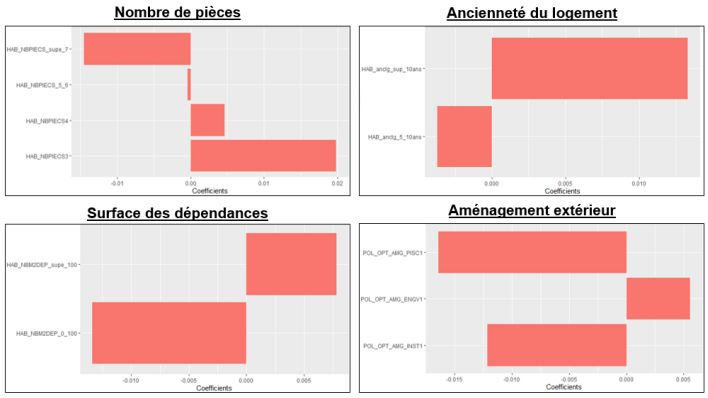


FIGURE 7.2 – GLM Gamma Maisons - Analyse des coefficients

Métriques d'évaluation

La figure 7.2 présente les métriques d'évaluation pour les trois GLMs pénalisés. Tout d'abord, nous remarquons grâce à l'erreur globale que les modèles obtenus sont plus robustes que ceux modélisant le coût moyen appartement. Une des raisons peut être le nombre supérieur d'observations pour la sinistralité des maisons. Nous privilégions ici la pénalisation Elastic Net qui obtient les meilleurs résultats en terme de segmentation de la population (Indice de Gini à 19%).

Modèle	Couple $(\alpha; \lambda)$	RMSE Train	RMSE Test	Err. glob. Train	Err. glob. Test	Gini Train	Gini Test	Déviance Test
LASSO	$(1; 1,5 \times 10^{-1})$	25423	25199	1%	0%	4%	4%	97021
ELASTIC NET	$(0,7; 3,4 \times 10^{-2})$	25296	25116	−3%	−3%	19%	19%	95075
RIDGE	$(0; 13,4)$	25427	25211	1%	−1%	4%	4%	9752

TABLE 7.4 – Métriques d'évaluation GLMs - Coût moyen attritionnel maisons

7.2.2 eXtreme Gradient Boosting

Dans cette section, nous modélisons le coût moyen attritionnel par le biais d'un XGBoost. Le processus d'optimisation des paramètres étant identique à celui des chapitres 5 et 6, nous ne le détaillons pas dans cette section. Néanmoins, comme pour les GLMs pénalisés, nous présentons pour chaque type d'habitation, l'importance des variables et les métriques d'évaluation associées aux modèles de régression.

7.2.2.1 La modélisation du risque appartement

Importance des variables

Après avoir optimisé les paramètres du XGBoost par k-fold validation croisée, nous analysons l'importance des variables du modèle. Tout d'abord, la variable ayant le plus d'influence sur le risque est l'âge du client. Cependant, la figure 7.3 ne nous renseigne pas sur la manière dont la variable influe sur le coût moyen des appartements. Néanmoins, en regardant la segmentation de l'espace des variables explicatives, nous pouvons conclure que le coût moyen augmente avec cette variable. La qualité de l'occupant se place à la deuxième position. Effectivement, comme le témoigne la figure 2.13, les locataires ont un coût moyen plus élevé que les propriétaires. Ensuite, le montant de capital assuré, l'ancienneté du logement, le nombre de pièces ainsi que le montant d'objets de valeur influent fortement sur le risque. Par ailleurs, ces dernières variables ont également un rôle important dans les GLMs pénalisés. Enfin, le reste des variables telles que la CSP ou l'étage n'influent que très peu.

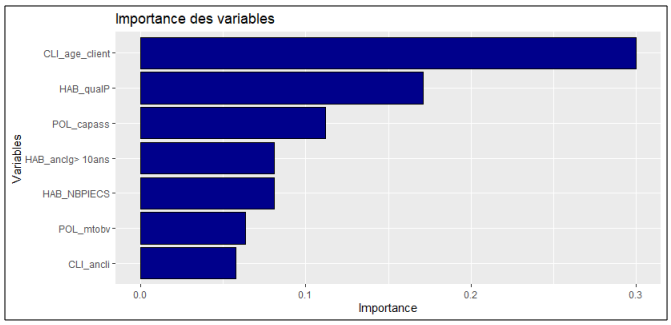


FIGURE 7.3 – Importance des variables XGBoost - Coût moyen appartements

Métriques d'évaluation

Les métriques d'évaluation ci-dessous traduisent d'une précision meilleure que celle des GLMs pénalisés. Néanmoins, la qualité de segmentation de la population est moins bonne.

Base	RMSE	Erreur globale	Gini
Train	12636	-6%	12%
Test	12636	-6%	12%

TABLE 7.5 – Métriques d'évaluation XGBoost - Coût moyen appartements

7.2.2.2 La modélisation du risque maison

Importance des variables

La figure 7.4 montre que les trois variables les plus influentes que le coût moyen des sinistres maisons sont respectivement le montant de capital assuré, la qualité de l'occupant ainsi que l'ancienneté du client. Effectivement, les assurés logeant dans une maison avec un capital assuré élevé, propriétaires et âgés possèdent le coût moyen le plus élevé. Ensuite, viennent le nombre de pièces ou le montant d'objets de valeur. Le reste des variables a une influence moindre.

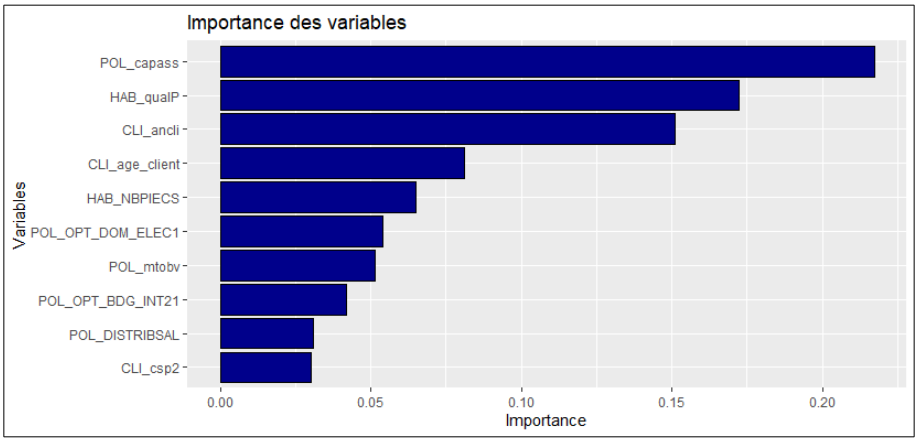


FIGURE 7.4 – Importance des variables XGBoost - Coût moyen maisons

Métriques d'évaluation

Les métriques d'évaluation ci-dessous traduisent d'un ajustement légèrement moins bon que celui issu des GLMs. De plus, la qualité de segmentation de la population est moins bonne.

Base	RMSE	Erreur globale	Gini
Train	25836	0%	17%
Test	22969	−4%	17%

TABLE 7.6 – Métriques d'évaluation XGBoost - Coût moyen maisons

7.2.3 Synthèse des résultats

Dans cette section, nous fournissons une synthèse des résultats obtenus sur le coût moyen attrtionnel par le biais des GLMs pénalisés et du XGBoost. Pour cela nous partageons les plus importantes variables ainsi que les métriques d'évaluation pour chaque modèle.

Modèle	**Top 1**	**Top 2**	**Top 3**	**Top 4**	**Top 5**
Appartements					
GLM	Surf. dép.	Anc. logement	Qualité occ.	Âge client	Nb pièces
XGBoost	Âge client	Qualité occ.	Cap. ass.	Anc. logement	Nb pièces
Maisons					
GLM	Qualité occ.	Cap. ass.	Nb pièces	Aménagement ext.	Surf. dép.
XGBoost	Cap. ass.	Qualité occ.	Anc. client	Âge client	Nb pièces

TABLE 7.7 – Classement des 5 variables les plus importantes par modèle - Coût moyen attritionnel

Modèle	**RMSE Train**	**RMSE Test**	**Err. glob. Train**	**Err. glob. Test**	**Gini Train**	**Gini Test**
Appartements						
GLM	12801	12072	−4%	9%	21%	21%
XGBoost	12636	12636	−6%	−6%	12%	12%
Maisons						
GLM	25296	25116	−3%	−3%	19%	19%
XGBoost	25836	22969	0%	4%	17%	17%

TABLE 7.8 – Métriques d'évaluation globales - Coût moyen attritionnel

7.3 Coût moyen grave

Dans le cadre de ce mémoire, la prime pure incendie est calculée en distinguant les sinistres dits attritionnels et les sinistres dits graves. Cette distinction est importante pour plusieurs raisons : construire un modèle de prime pure plus précis, analyser les facteurs de la sinistralité dite grave pour développer la prévention, et identifier les profils de haut risque.

L'étude des valeurs extrêmes détaillée en section 2.1.4 permet de définir les sinistres graves comme des sinistres ayant une charge supérieure à 100 000 € pour les appartements, et 300 000 € pour les maisons. Il est important de noter que les sinistres graves sont très minoritaires au sein de notre base de modélisation. Effectivement, ils ne représente que 1,1% de nos observations. Ainsi, dans ce cadre, nos données sont de faible volume et ne permettent pas de modéliser ce risque de manière classique. Pour pallier ce problème, nous utilisons un algorithme de suréchantillonnage synthétique, le SMOTE qui est présenté en section 4.4.1. Effectivement, cet algorithme permet de générer des individus synthétiques de la classe minoritaire (dans notre cas les sinistrés graves), dans l'espace des variables explicatives, localisés entre deux sinistres graves réels.

Dans cette section, plusieurs méthodes sont testées afin de modéliser le coût moyen grave. Tout d'abord, nous appliquons et analysons une méthode simple permettant d'obtenir rapidement des résultats facilement interprétables : l'affection de la moyenne par segment défini par arbre de régression simple. Ensuite, nous utilisons l'algorithme SMOTE afin de générer un modèle linéaire simple à partir d'une base comprenant des sinistres graves fictifs. Enfin, les deux méthodes citées précédemment sont comparées grâce aux métriques d'évaluation des modèles de régression.

7.3.1 Moyenne par arbre de régression simple

7.3.1.1 La segmentation de l'espace

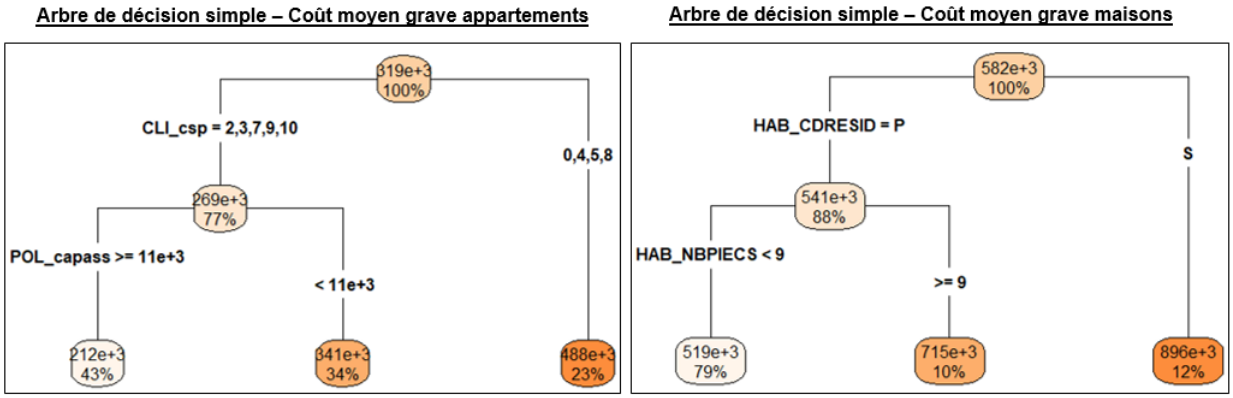


FIGURE 7.5 – Arbres de décision simples - Coût moyen grave

Dans un premier temps, nous avons décidé de segmenter l'espace des variables explicatives pour les deux types d'habitations. Comme le montre la figure 7.5, les arbres sont peu profonds. Étant donné le faible nombre de données, il a été décidé de créer des arbres variant au minimum. En effet, le risque de sur-apprentissage augmente considérablement lorsque les données sont peu nombreuses. Concernant les appartements, nous observons que la variable segmentant le plus l'espace des variables explicatives est la CSP. Ensuite, vient le montant de capital assuré. Toutefois, les informations fournies par cet arbre sont à prendre avec beaucoup de précaution. Pour les maisons, les résultats sont plus facilement interprétables. Tout d'abord, les résidences secondaires ont les coûts moyens graves les plus élevés. Ensuite, les grandes maisons sont plus sujettes à avoir un coût de sinistre incendie élevé.

7.3.1.2 Métriques d'évaluation

Modèle	RMSE Train	RMSE Test	Erreur globale Train	Erreur globale Test
Moyenne appartements	394279	194516	−1%	5%
Moyenne maisons	296327	208853	−1%	2%

TABLE 7.9 – Métriques d'évaluation Moyenne - Coût moyen grave

Le tableau ci-dessus présente les métriques d'évaluation pour le modèle présenté précédemment. Malgré un nombre très faible de données, nous remarquons qu'un modèle simplifié et peu variant prédit, au global, correctement. Effectivement, les erreurs globales sur la base test montrent que le modèle sur-estime pour les appartements et les maisons, respectivement de 5% et de 2% la charge totale grave. Les avantages de ce modèle sont sa simplicité d'interprétation et la qualité de ses résultats.

7.3.2 Application du sur-échantillonnage synthétique

Dans cette section, nous effectuons un modèle linéaire simple afin de modéliser le coût moyen grave. Cette approche permet de comparer les résultats obtenus précédemment tout en utilisant un algorithme innovant qu'est le SMOTE. Par ailleurs, le lecteur est invité à se référer à la section 4.4.1 pour obtenir une présentation théorique de cet algorithme.

La figure 7.5 permet d'identifier les variables influant majoritairement sur le coût moyen grave. Effectivement, pour les appartements nous retenons les variables CSP et le montant de capital assuré, tandis que pour les maisons nous retenons le type de résidence et le nombre de pièces. De ce fait, pour garder la simplicité de modélisation, ne pas rajouter des variables n'ayant pas de lien significatif avec la variable réponse, et obtenir un modèle peu variant, nous décidons de conserver uniquement les variables citées précédemment pour entraîner le modèle. Pour rappel, nous sur-échantillonnons la base d'entraînement afin d'entraîner le modèle avec suffisamment de données.

7.3.2.1 Analyse des coefficients

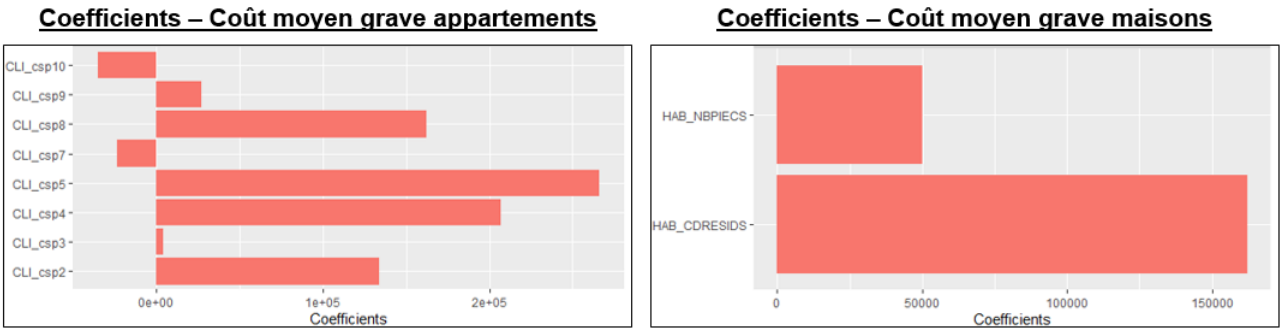


FIGURE 7.6 – Analyse des coefficients - Coût moyen grave

La figure ci-dessus fournie les coefficients de chaque modalité des variables sélectionnées pour la modélisation que sont : la CSP pour les appartements, le type de résidence et le nombre de pièces pour les maisons. Tout d'abord, malgré la modification de la base d'entraînement par l'algorithme SMOTE, nous retrouvons la même interprétation que celle donnée par les arbres de régression simples. Effectivement, concernant les appartements, les CSP 4, 5 et 8 correspondant respectivement aux professions libérales, cadres supérieurs et ouvriers influent fortement sur le coût moyen grave. Concernant les maisons, la variable la plus influente reste le type de résidence, suivi par le nombre de pièces.

7.3.2.2 Métriques d'évaluation

Modèle	RMSE Train	RMSE Test	Erreur globale Train	Erreur globale Test
Modèle linéaire appartements	271483	196971	0%	46%
Modèle linéaire maisons	202544	224338	0%	−2%

TABLE 7.10 – Métriques d'évaluation modèle linéaire - Coût moyen grave

Le tableau ci-dessus nous fournit des informations sur la qualité des modèles créés grâce à l'algorithme SMOTE. Tout d'abord, nous nous apercevons que la modélisation de la sinistralité grave appartement obtient de très mauvais résultats. Effectivement, le modèle sur-apprend très fortement : 46% de la charge réelle est sur-estimée sur la base test. Un modèle de cette qualité ne peut donc pas être conservé. Ainsi, nous pouvons constater qu'un simple modèle linéaire n'arrive pas à ajuster correctement les données. Cela peut venir du fait que l'occurrence d'un sinistre grave est un évènement si rare qu'il relève plus de l'aléa que d'une explication statistique. Concernant les maisons, les résultats du modèle sont satisfaisants et très proches de l'arbre de régression simple. Effectivement, le modèle ne sous-estime la charge réelle sur la base test que de 2%. Nous pouvons donc être plus confiants sur cette modélisation.

7.3.3 Synthèse des résultats

Dans cette section, nous fournissons une synthèse des résultats obtenus par le biais des arbres de régressions simples et des modèles linéaires simples. Pour cela, nous partageons les variables les plus importantes ainsi que les métriques d'évaluation pour chaque modèle.

Modèle	Top 1	Top 2
Appartements	CSP	Montant de capital assuré
Maisons	Type de résidence	Nombre de pièces

TABLE 7.11 – Classement des variables les plus importantes par modèle - Coût moyen grave

Modèle	RMSE Train	RMSE Test	Erreur globale Train	Erreur globale Test
Appartements				
Arbre de régression	394279	194516	−1%	5%
Modèle linéaire	271483	196971	0%	46%
Maisons				
Arbre de régression	296327	208853	−1%	2%
Modèle linéaire	202544	224338	0%	−2%

TABLE 7.12 – Métriques d'évaluation globales - Coût moyen grave

Synthèse de modélisation

Les chapitres 5, 6 et 7 nous ont permis de modéliser chacune des composantes de la prime pure définie par la formule 2.8. Lors de ces chapitres, chaque modèle a été analysé afin d'évaluer leur qualité d'ajustement et de segmentation. Cependant, la prime pure incendie s'obtient en combinant l'ensemble de ces modèles. De ce fait, nous appliquons dans cette section la formule 2.8 permettant d'obtenir la prime pour chaque observation. Nous présentons ensuite la qualité de l'ensemble de la modélisation.

Dans le but d'analyser plus en détail la qualité de l'ensemble de la modélisation, nous avons décomposé la prime pure de la manière suivante :

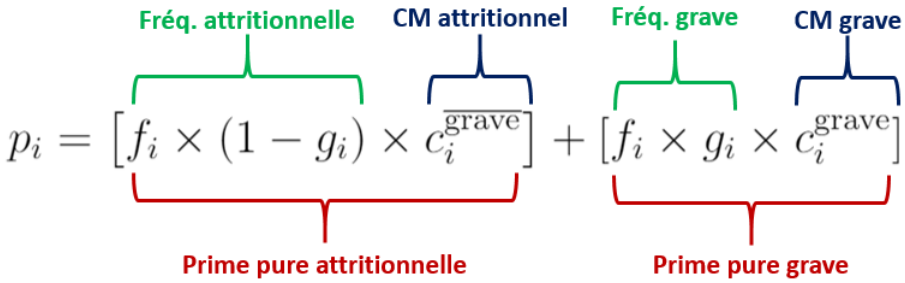


FIGURE 7.7 – Décomposition de la prime pure

Le modèle de propension permet alors de distinguer la modélisation des sinistres attritionnels et des sinistres graves. Effectivement, la prime pure de ces différents types de sinistres est pondérée par leurs probabilités d'occurrence modélisées dans le chapitre 6. Ainsi, nous évaluons la qualité d'ajustement et de segmentation de la fréquence, du coût moyen et de la prime pure pour ces différents types de sinistres, et pour la totalité. Enfin, dans le but d'obtenir une évaluation rigoureuse et précise, cette dernière est effectuée en appliquant nos différents modèles sur 2 millions de données non connues par le modèle. En d'autres termes, elle est effectuée sur une grande base de validation.

Dans ce cadre, les modèles conservés sont les GLMs pénalisés pour la fréquence, la propension et le coût moyen attritionnel. Ce choix est justifié par le fait que les résultats des GLMs pénalisés et des arbres de décision diffèrent peu, et que les GLMs sont plus simples à interpréter et à mettre en production. Concernant le coût moyen grave, nous retenons la moyenne par arbre de décision simple présentée en section 7.3.1 car elle obtient les meilleurs résultats.

Concernant la modélisation attritionnelle, les résultats sont satisfaisants. En effet, la fréquence réelle de ces sinistres est sur-estimée de seulement 1,3%, et l'indice de Gini de 33,5% traduit une bonne qualité de segmentation. La qualité d'ajustement du coût moyen est elle aussi satisfaisante (la somme du coût moyen réelle est sur-estimée de 2,2%). Cependant, pour cette même composante, nous notons une faible segmentation de la population (indice de Gini de 12,5%). Cela est justifié

par le fait que les modèles sont entraînés sur un nombre de données assez faible. Effectivement, la modélisation du coût moyen attritionnel est effectuée uniquement à partir des observations ayant eu un sinistre attritionnel. Les résultats de la prime pure attritionnelle sont également corrects. La somme des sinistres attritionnels réelle est sur-estimée de seulement 3,7%, et la segmentation de la population est bonne (indice de Gini à 35,7%).

Étant donné la très faible quantité de données permettant d'effectuer la modélisation des sinistres graves, les résultats sont satisfaisants. La prime pure correspondante sous-estime de 8,1% la somme des sinistres graves réelle, et la qualité de segmentation reste correcte grâce au modèle de fréquence (indice de Gini à 28,1%).

Pour la totalité des sinistres, nous remarquons que les résultats sont très proches de la modélisation attritionnelle. Effectivement, le modèle de propension assigne une probabilité g_i très faible, impliquant un faible poids de la modélisation des sinistres graves dans la modélisation totale. De ce fait, les résultats globaux sont satisfaisants, car les différents modèles permettent de créer une prime pure proche de la sinistralité réelle tout en identifiant un nombre conséquent de sous-profils de risque au sein de notre population. En effet, nous obtenons une erreur globale de 3,6% et un indice de Gini à 37,4%.

Métriques d'évaluation - Prime pure Incendie		
Modèle	**Erreur globale**	**Gini**
Attritionnel		
Fréquence	$1,3\%$	$33,5\%$
Coût moyen	$2,2\%$	$12,5\%$
Prime pure	$3,7\%$	$35,7\%$
Grave		
Fréquence	$-2,8\%$	$27,3\%$
Coût moyen	$-4,3\%$	$2,3\%$
Prime pure	$-8,1\%$	$28,1\%$
Total		
Fréquence	$1,3\%$	$36,3\%$
Coût moyen	$1,9\%$	$13,6\%$
Prime pure	$3,6\%$	$37,4\%$

TABLE 7.13 – Synthèse de modélisation

Nous précisons également que cette refonte permet de conserver des modèles contenant moins de variables que les modèles précédents. Ainsi, elle implique à la fois une amélioration de l'évaluation du risque et une simplification de l'interprétation des modèles.

Pour conclure, nous rappelons qu'il reste la création d'un zonier de fréquence et de coût moyen avant de finaliser la prime pure incendie. La modélisation présentée au cours de cette étude ne prends pas en compte les informations géographiques des différents contrats. Ces dernières permettent de créer un zonier de fréquence et de coût moyen à partir des résidus de la première modélisation effectuée. De ce fait, les modèles intégrant les informations géographiques possèdent une qualité d'ajustement supérieure aux modèles ne les intégrant pas. Ainsi, au vu des résultats de notre modélisation, nous pouvons être confiants sur les futurs résultats intégrant les zoniers de fréquence et de coût moyen.

Conclusion

Dans le but d'effectuer une refonte tarifaire de la garantie incendie, ce mémoire détaille un processus de tarification par le biais de méthodes classiques et innovantes. Aussi, il a pour ambition de fournir une méthode générique de tarification d'assurance non-vie, via le produit Multirisque Habitation.

Une étude de rentabilité a montré certaines lacunes concernant l'ancienne segmentation de cette garantie. L'enjeu de ce mémoire est alors de proposer une modélisation fine du risque incendie permettant de fournir une meilleure segmentation.

Nous avons dans un premier temps construit la base de données sur laquelle la modélisation s'est appuyée. Une fois construite, cette dernière comporte pour chaque image de risque, une quantité importante d'informations liées au logement, à l'assuré et à la sinistralité. Cette première étape doit être effectuée soigneusement car la qualité des données est un enjeu primordial pour la modélisation. Par ailleurs, une première limite fut constatée : plusieurs images de risque sur un même mois pour un même contrat ne peuvent pas être obtenues. Aussi, cette étape a révélé plusieurs axes d'amélioration possibles. Effectivement, la méthode de développement des sinistres utilisée est celle de Chain-Ladder pour des raisons de simplicité et de fiabilité. Toutefois, des méthodes stochastiques telles que celle de Mack auraient pu être appliquées. Aussi, le traitement des valeurs manquantes ainsi que la discrétisation des variables auraient pu être effectués autrement.

Dans un second temps, nous avons modélisé le risque incendie par le biais de trois modèles : un modèle de fréquence, de propension et de coût moyen. Le modèle de propension a permis de distinguer la sinistralité attritionnelle de la sinistralité grave. Dans ce cadre, une attention particulière a été portée sur la sinistralité grave. En effet, au vu de la très faible quantité de données, le coût moyen associé a été estimé par une méthode de moyennes par arbre de régression, et par une modélisation suite à un sur-échantillonnage synthétique.

Chaque composante de la prime pure a été modélisée en comparant les résultats obtenus via des modèles linéaires généralisés et des méthodes d'agrégation d'arbres que sont les forêts aléatoires et le XGBoost. Aussi, nous avons distingué la modélisation de la sinistralité appartement de celle de la sinistralité maison en raison de la différence de ces risques. Cette étape satisfait plusieurs conclusions. Tout d'abord, les méthodes d'agrégation d'arbres améliorent trop peu les modèles en termes de précision et de segmentation par rapport à la complexité d'interprétation qu'ils apportent. Les modèles linéaires généralisés obtiennent des résultats satisfaisants et sont donc conservés pour refondre la prime pure. Ainsi, une nette amélioration de la segmentation du risque incendie est apportée par le biais de ce mémoire.

Cependant, bien que le risque appartement et le risque maison répondent à des problématiques de modélisation différentes, cette distinction peut être remise en cause. Effectivement, rien n'assure que cette segmentation crée deux sous-groupes de risques totalement homogènes, ni qu'elle apporte les meilleurs résultats. Elle peut être revue par des méthodes d'apprentissage non supervisées telles que les K-means ou encore la Classification Ascendante Hiérarchique (CAH). Cette étape pourrait ainsi faire l'objet d'une étude annexe à l'avenir.

Ce mémoire ne prétend en aucun cas fournir de conclusions générales concernant les méthodes de modélisation étudiées. Les méthodes retenues sont reliées à la base de données utilisée dans le cadre d'une modélisation du risque incendie sur le produit Multirisque Habitation. Néanmoins, l'étude réalisée a permis de fournir un ensemble de programmes automatisés mis à la disposition du groupe Axa. Ainsi, les méthodes utilisées dans cette étude peuvent être reprises sur d'autres bases de modélisation, d'autres garanties Habitation, et d'autres lignes de métier telles que l'assurance automobile.

Bibliographie

[1] Arthur Charpentier et Michel Denuit. *Mathématiques de l'assurance non-vie : Tome 2, Tarification et provisionnement.* Economica, 2005.

[2] Gareth James, Daniela Witten, Trevor Hastie et Robert Tibshirani. *An Introduction to Statistical Learning. With Applications in R.* Springer, 2013.

[3] Trevor Hastie, Robert Tibshirani et Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer, 2009.

[4] Robert Tibshirani. *Regression Shrinkage and Selection via the Lasso.* Series B, 1996.

[5] Peng Zhao et Bin Yu. *On Model Selection Consistency of Lasso.* David Madigan, 2006.

[6] Claire Boyer. *Machine Learning. Cours ISUP.* 2019.

[7] Maud Thomas. *Économétrie de l'assurance non-vie. Cours ISUP.* 2019.

[8] Marie Kratz. *Extreme Value Theory. Theory and Application to Risk Managment. Cours ISUP.* 2019.

[9] Elena Di Bernardino. *Théorie des valeurs extrêmes. Cours ISUP.* 2018.

[10] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall et W. Philip Kegelmeyer. *SMOTE : Synthetic Minority Over-sampling Technique.* 2002.

[11] Jennifer Pariente. *Modélisation du risque géographique en assurance habitation. Mémoire d'actuariat.* Université Paris Dauphine, 2017.

[12] Mohamed Halimi. *Réactualisation des méthodes classiques de tarification IARD. Mémoire d'actuariat.* ENSAE, 2017.

[13] Charles Tremblay. *Prédire les sinistres graves en assurance : les apports de l'apprentissage statistique aux modèles linéaires. Mémoire d'actuariat.* ENSAE, 2017.

[14] Axa France. *Condition générales Ma Maison* 2019.

Annexes

Régression logistique

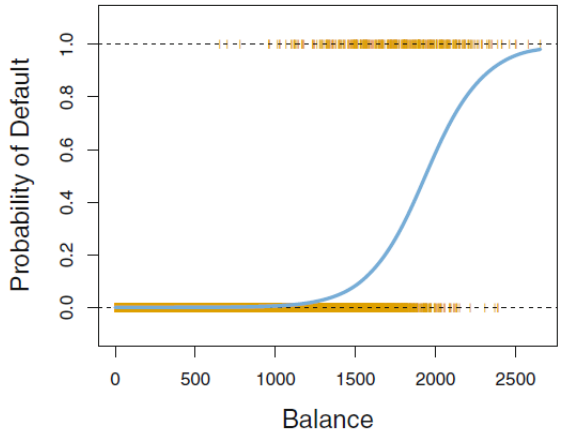
Dans plusieurs situations, la variable réponse est qualitative. Cette dernière est alors définie comme catégorielle. Prédire une réponse catégorielle pour une observation donnée revient ainsi à effectuer une classification, car elle implique l'attribution d'une classe à cette dernière. D'autre part, les méthodes utilisées pour la classification prédisent souvent d'abord la probabilité, pour chaque observation, d'appartenir à chacune des classes.

Il existe de nombreuses techniques de classification, ou classificateurs, que l'on peut utiliser pour prédire une réponse qualitative. Nous abordons ici un des classificateurs les plus utilisés : la régression logistique.

Soit X la matrice contenant l'ensemble des observations des p variables explicatives, Y la matrice réponse, et $(0, 1)$ deux catégories. Plutôt que de modéliser directement la réponse Y, la régression logistique modélise la probabilité que Y appartienne à une catégorie particulière. Mais comment peut-on connaître la relation entre $p(X) = \mathbb{P}(Y = 1|X)$ et X ? Pour ce faire, la fonction logistique est utilisée :

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}$$

Pour ajuster le modèle, la méthode utilisée est le maximum de vraisemblance. La figure ci-dessous présente l'ajustement de la régression logistique permettant d'estimer la probabilité de défaut en fonction de la balance du compte bancaire :



La fonction logistique produira toujours une courbe en forme de S de cette forme, et donc quelque soit la valeur de X, nous obtiendrons une prédiction raisonnable.

Après une légère manipulation de la fonction logistique, nous obtenons :

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}$$

Cette quantité est appelée *odds*, et peut prendre n'importe quelle valeur entre 0 et ∞. En appliquant la fonction logarithme, nous obtenons :

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

Cette quantité est appelée le *log-odds* ou *logit*. Nous constatons alors que le modèle de régression logistique a un *logit* linéaire en X.

Les coefficients β sont inconnus et doivent être estimés grâce à la base d'entraînement. Pour ce faire, nous utilisons la méthode du maximum de vraisemblance. Les coefficients β sont choisis de telle sorte qu'ils maximisent la fonction de vraisemblance. Cette dernière est définie de la manière suivante :

$$l(\beta_0, \beta_1, ..., \beta_p) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1-p(x_i))$$

Une fois les coefficients estimés, il est très simple d'estimer la probabilité que Y appartienne à la classe 1. Nous le faisons de la manière suivante :

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+...+\hat{\beta}_1 X_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1+...+\hat{\beta}_1 X_p}}$$

Un seuil à partir duquel l'observation appartient à chaque classe doit ensuite être défini. Une des méthodes est présentée en section 6.2.1.

Bootstrap

Les méthodes de rééchantillonnage sont des outils indispensables dans les statistiques modernes. Elles impliquent de prélever à plusieurs reprises des échantillons d'une base d'entraînement et de réajuster le modèle sur chacun de ces échantillons afin d'obtenir des informations complémentaires sur le modèle ajusté. Les méthodes de rééchantillonnage peuvent être coûteuses en termes de calcul, car elles impliquent l'ajustement de la même méthode statistique plusieurs fois en utilisant différents sous-ensembles de données sur la base d'entraînement. Toutefois, en raison des récents progrès de la puissance des calculs, les méthodes de rééchantillonnage en général peuvent être exécutées avec un temps raisonnable. Nous abordons ici une des plus courantes méthodes : le Bootstrap.

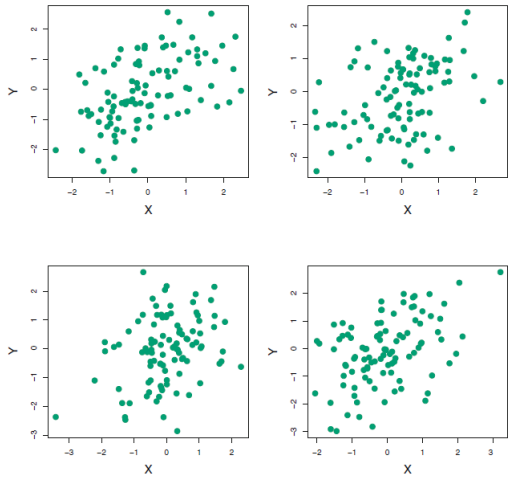
Le Bootstrap est un outil statistique très souvent appliqué et extrêmement puissant qui peut être utilisé pour quantifier l'incertitude associée à un estimateur donné ou à une méthode d'apprentissage statistique. A titre d'exemple simple, le Bootstrap peut être utilisé pour estimer les écarts types des coefficients à partir d'une ajustement par régression.

Dans cette section, nous illustrons le Bootstrap par un jeu dans lequel nous souhaitons déterminer la meilleure allocation d'investissement dans le cadre d'un modèle simple. Supposons que nous souhaitions investir une somme d'argent fixe dans deux actifs financiers dont les rendements sont respectivement de X et Y, où X et Y sont des quantités aléatoires. Nous investissons une fraction α de notre argent dans X, et nous investissons la fraction $1-\alpha$ restante dans Y. Comme il existe une variabilité associée aux rendements de ces deux actifs, nous souhaitons choisir α pour minimiser le risque total, ou la variance, de notre investissement. En d'autres termes, nous voulons minimiser $Var(\alpha X + (1-\alpha Y))$. On peut montrer que la valeur qui minimise le risque est donnée par :

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}}$$

Où $\sigma_X^2 = Var(X)$, $\sigma_Y^2 = Var(Y)$, et $\sigma_{XY} = Cov(X, Y)$. En réalité, ces quantités sont inconnues. Elles peuvent être estimées en utilisant une base de données qui contient des mesures antérieures pour X et Y.

La figure ci-dessous illustre cette approche pour l'estimation de α sur un ensemble de données. Pour chaque panel, 100 paires de rendements pour les investissements X et Y sont simulées. Grâce à ces rendements, α peut être estimé. Ainsi, la valeur de $\hat{\alpha}$ résultant de chaque ensemble de données simulé varie de 0,532 à 0,657.



Il est naturel de vouloir quantifier la précision de notre estimation de α. Pour estimer l'écart type de $\hat{\alpha}$, le processus de simulation d'observations est répété 1000 fois. Nous avons ainsi obtenu 1000 estimations de α : $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_{1000}$. Supposons que nous connaissons la réelle valeur de α, ce qui est en pratique impossible. La valeur moyenne sur l'ensemble des 1000 estimations pour α est de :

$$\overline{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0,5996$$

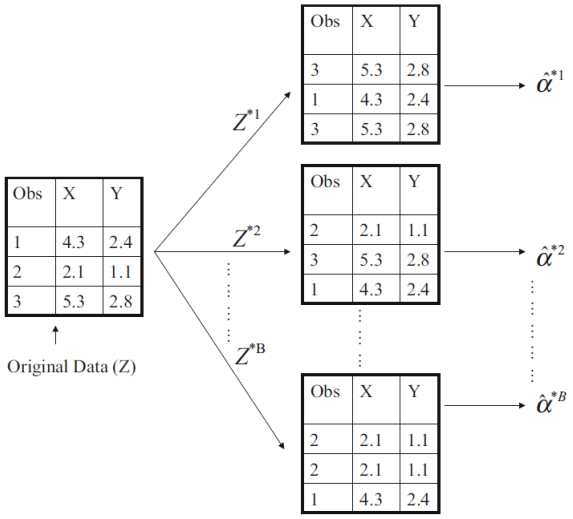
La valeur estimée est donc très proche de la valeur réelle. L'écart type des estimations est le suivant :

$$\sqrt{\frac{1}{1000 - 1} \sum_{r=1}^{1000} (\hat{\alpha}_r - \overline{\alpha})^2} = 0,083$$

Cela nous donne une très bonne idée de la précision de $\hat{\alpha}$. Cependant, en pratique, la procédure d'estimation ci-dessus ne peut pas être appliquée, car pour des données réelles, nous ne pouvons pas générer de nouveaux échantillons à partir de la population d'origine. Toutefois, le Bootstrap nous permet de le faire, de telle sorte que nous pouvons estimer la variabilité de $\hat{\alpha}$ sans générer de nouveaux échantillons. Plutôt que d'obtenir de manière répétée des ensembles de données indépendants de la population, nous obtenons des ensembles de données distincts en échantillonnant de manière répétée les observations de l'ensemble de données original.

Cette approche peut être illustrée par la figure ci-dessous, sur un ensemble de données simples, que nous appelons ici Z, et qui ne contient que 3 observations. Nous sélectionnons aléatoirement n observations de l'ensemble de données afin de produire un ensemble de données Bootstrap, noté Z^{*1}. L'échantillonnage est effectué avec remplacement, ce qui signifie que la même observation peut

apparaître plusieurs fois dans le même ensemble de données Bootstrap. Ainsi, nous pouvons utiliser Z^{*1} pour obtenir une nouvelle estimation Bootstrap de α, que nous appelons $\hat{\alpha}^{*1}$. Cette procédure est répétée B fois pour une valeur importante de B, dans le but de produire B différents ensemble de données Bootstrap, Z^{*1},Z^{*2},...,Z^{*B}, et B estimation Bootstrap de α, $\hat{\alpha}^{*1}$,$\hat{\alpha}^{*2}$,...,$\hat{\alpha}^{*B}$.



Ainsi, nous pouvons obtenir l'écart type de ces estimations par la formule :

$$\mathrm{SE_B}(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}$$

Il s'agit alors d'une estimation de l'écart type $\hat{\alpha}$ estimé à partir de l'ensemble des données d'origine.