Mémoire présenté le : 26 mai 2020

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

Par : Bastien FOUQUET

Titre Détection des sinistres incendie graves des contrats multirisques 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

- B. Potentier
- A. Hassler
- M. Toledano

Membres présents du jury de l'ISFA

Entreprise :

Nom : **Thélem Assurances**

Signature :

Directeur de mémoire en entreprise :

Nom : **PICHARD Céline**

Signature :

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

DETECTION DES SINISTRES INCENDIE GRAVES DES CONTRATS MULTIRISQUE HABITATION

MASTER 2 ACTUARIAT

Bastien FOUQUET

Ce mémoire est à caractère confidentiel

Tuteur Entreprise : Céline PICHARD

Tuteur Universitaire : Stéphane LOISEL

Résumé

L'assurance multirisque habitation propose différentes garanties dont la garantie incendie. Cette garantie permet de se couvrir contre les risques d'incendie se déclarant dans le logement de l'assuré. Lorsque le sinistre est total, c'est-à-dire lorsque tout ou la quasi-totalité du logement est détruit, le coût du sinistre engendré peut vite atteindre une somme élevée.

Aujourd'hui, lors de la tarification du produit habitation, les sinistres attritionnels (dont la fréquence est élevée mais la charge sinistre est faible) et les sinistres graves (dont la fréquence est faible mais la charge sinistre est élevée) sont étudiés séparément. Des modèles statistiques sont utilisés pour modéliser la charge des sinistres attritionnels afin de répartir cette charge de la manière la plus juste. Cependant, la charge des sinistres graves est répartie de manière uniforme sur toutes les garanties. Le but de cette étude est d'estimer la charge des sinistres graves à l'aide de modèles d'apprentissage statistique (modèle logit, arbres de décision, descente de gradient) afin de la répartir de manière non-homogène sur les critères de tarification.

Enfin, à l'aide des résultats des modèles nous pourrons analyser les profils des clients les plus à risque (c'est-à-dire dont la probabilité de déclarer un sinistre incendie dans l'année est élevée). Cela permettra à l'assureur de mandater un expert chez les assurés identifiés afin d'effectuer une visite de risque de prévention.

Mots clés : Assurance multirisque habitation, incendie, sinistre attritionnel, sinistre grave, modèle d'apprentissage statistique, régression logit, arbres de décision, descente de gradient, profil de risque, matrice de confusion, prévention, visite de risque.

Abstract

The household insurance offers various guarantees including the fire guarantee. This one covers the fire risks occurring in the insured's house. When the claim is total, that is when all or almost all the insured's accommodation is destroyed, the cost of the claim can quickly reach a high amount of money.

Today, when insurers pricing household insurance, attritional claims (with high frequency but low loss) and large claims (with low frequency but high loss) are separately studied. Statistical models are used to modeling the cost of attritional claims in order to distribute the fairest the cost. However, large claim's cost is evenly distributed all over the guarantees. The aim of this study is to estimate the large claim's cost with machine learning models (logit regression, decision trees, gradient boosting) in order to don't evenly distribute the cost all over the criterions.

Finally, with the models' results, we will be able to analyze the profiles of risky insureds (that is when the probability to claim a large fire disaster during the year). The insurer can appoint an expert to identify the possible risks in the insured's house and to prevent against these risks.

Key words: household insurance, fire claim, attritional claim, large claim, machine learning model, logit regression, decision trees, gradient boosting, risk profile, confusion matrix, risk visit.

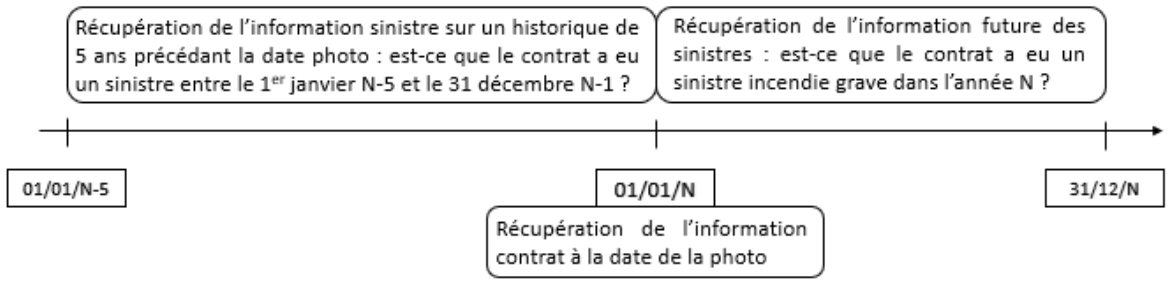
Note de synthèse

Contexte

La tarification du produit multirisque habitation se décompose en plusieurs phases. Dans un premier temps il faut séparer les sinistres attritionnels des sinistres graves. Les sinistres dont la fréquence est élevée mais dont le coût est relativement faible sont caractérisés comme attritionnels. Au contraire, les sinistres dont la fréquence est faible (voire très faible) mais avec une charge beaucoup plus importante sont considérés comme graves. Pour faire la distinction, il faut fixer un seuil à partir duquel on considère le sinistre comme grave. Ce seuil a été choisi à l'aide de plusieurs études effectuées au moment de la création de la dernière génération du produit habitation. Pour ce mémoire, le seuil a été placé à 30 000€ afin de ne pas avoir un nombre de sinistre grave trop faible pour la modélisation.
Lorsque la séparation des sinistres est effectuée, les sinistres attritionnels sont modélisés à l'aide d'un modèle coût/fréquence afin de répartir au plus juste la charge en fonction des garanties. La charge des sinistres graves est ensuite répartie de manière homogène sur toutes les garanties. L'objectif des modèles mis en place dans ce mémoire est de pouvoir répartir différemment la charge des sinistres graves afin que ce ne soit plus homogène.

Base de données

La création de cette base a été faite en plusieurs extractions. L'idée générale est de construire une base regroupant des « photos du portefeuille » à plusieurs dates données. On appelle « photo du portefeuille » le fait de figer à une date choisie les informations de tous les contrats actifs, des clients rattachés aux contrats et de leur(s) sinistre(s). Les photos ont été prises à 7 dates différentes : tous les 1ers janviers entre 2012 et 2018. Pour chaque contrat multirisque habitation présent en portefeuille au moment de la photo, les informations suivantes sont extraites : les sinistres que l'assuré a déclaré sur les 5 dernières années, les sinistres incendies de l'assuré sur les 12 mois suivant la photo, les équipements du client (c'est-à-dire si le client possède d'autres contrats comme par exemple un contrat automobile, santé, …) et enfin les différentes garanties du contrat souscrit. Nous pouvons schématiser les photos comme suit :



Toutes les données sont ensuite agrégées en une seule base avant d'être retravaillées. En effet, du fait des différentes générations du produit habitation, plusieurs variables ont été modifiées dans le temps, comme par exemple les niveaux de franchises, et il est nécessaire de tout harmoniser. De plus, certaines variables catégorielles sont retraitées afin d'avoir moins de catégories et des variables continues sont transformées en variables catégorielles comme

par exemple la surface des dépendances. Lorsque les retraitements sont effectués, toutes les lignes possédant des données manquantes ont été supprimées. Bien que cela puisse fausser les résultats, la proportion de ces individus était négligeable.

Modélisation

La modélisation a été construite en comparant les résultats de plusieurs modèles d'apprentissage statistique dans le but de choisir le meilleur. Les modèles de régression logit, et les arbres de décisions ont été utilisés. De plus, afin d'améliorer le modèle des arbres de décision, la descente de gradient a été appliquée. Dans une logique similaire d'amélioration des modèles, des techniques de rééchantillonnage (sous-échantillonnage et validation croisée) ont été appliquées aux données.
Afin de comparer tous ces modèles, des critères de performance ont été mis en place :

- La matrice de confusion qui recense dans un tableau si les individus sont correctement classés ou non ;
- La table de classification permet de placer le seuil de probabilité qui distinguera des individus risqués des individus non-risqués (selon les modèles) ;
- La courbe ROC (*Receiver Operating Characteristic*) et l'AUC (*Area Under the Curve*). La courbe ROC décrit l'évolution du taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (1-spécificité) et peut être représenté dans un graphique. On peut alors ensuite calculer l'aire sous la courbe, c'est ce qui nous donne l'AUC. Plus l'AUC est proche de 1 plus le modèle est performant.

Les résultats de l'AUC des modèles se situent dans l'intervalle [0,62 ; 0,68], ce qui n'est pas très satisfaisant. En effet, une AUC à 0,5 signifie que les résultats sont donnés de manière aléatoire. Dans notre cas, les modèles sont tout juste meilleurs que l'aléatoire.
Enfin, une fonction de coût a été introduite afin d'observer la différence entre la perte de cotisation si l'assureur décidait de résilier tous les contrats à risque par rapport aux potentiels sinistres évités. Cette fonction permet de voir le problème d'un point de vue économique et permet de réfléchir sur la perte de cotisation par rapport aux coûts des sinistres évités.

Application métier

Malgré des résultats de modèles faibles, nous voulions présenter la démarche de l'application métier qui découle de la modélisation. Pour ce faire, le modèle logit présente des facilités d'explicabilité des coefficients (sens de variation, intensité de l'impact des variables) et des performances proches du meilleur modèle, c'est pourquoi il a été choisi pour la fin de l'étude. Ce modèle permet d'ordonner les contrats en fonction de la probabilité d'avoir un sinistre incendie. Si l'on isole les 500 contrats les plus risqués selon ce modèle, nous pouvons comparer la répartition de la population totale par rapport aux individus à risque. On peut alors remarquer que les individus possédant des surfaces de dépendances supérieur à 900m² ou encore les individus ayant plus de 9 pièces principales sont majoritaires dans la base risquée. Ces critères permettent alors d'effectuer des filtres sur les contrats actifs du portefeuille à la date d'aujourd'hui afin de sélectionner les clients à expertiser.

De plus, les coefficients du modèle donnent des indications pour distribuer la charge sinistre des graves de manière non-homogène. En effet, pour les variables vues précédemment, la répartition peut être effectuée en fonction de l'importance du coefficient.

Conclusion

La fréquence des sinistres étant trop faible, les modèles d'apprentissage ne sont pas performant malgré un seuil de 30 000€ relativement bas. Cependant, tous les modèles convergent vers des variables qui pourraient expliquer la survenance des sinistres incendie. C'est pourquoi une étude plus approfondie sur les clients présentant les caractéristiques convergentes (par exemple, les clients possédant des surfaces de dépendances supérieures à 900m² et avec plus de 9 pièces principales). Cette étude pourra être menée en lien avec le service souscription qui étudiera de façon plus précise les clients sélectionnés. Si ces études sont concluantes (c'est-à-dire, si le risque d'avoir un sinistre incendie grave est avéré pour les individus examinés) alors il pourra être mis en place des contrôles récurrents sur ces profils de clients.

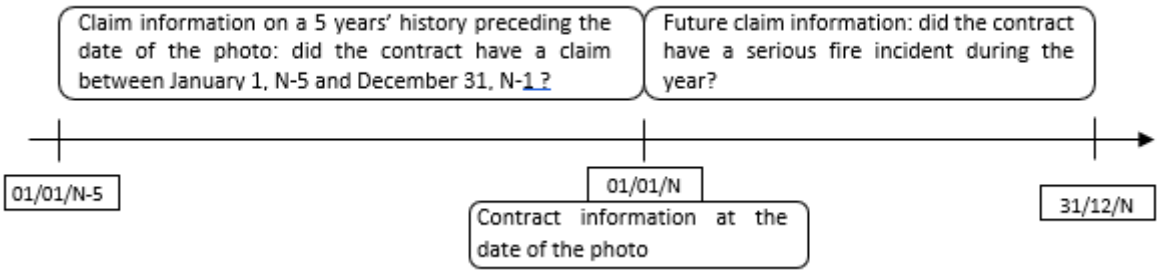
Synthesis

Context

Pricing a household product is decomposed into several steps. First, the insurer separate attritional claims to large claims. A claim with high frequency but low loss is named attritional. In contrary, a claim with low frequency but high loss is named large. To do distinction between the two categories, a threshold must be set. This threshold was chosen using several studies carried out when the last generation of the household product was created. For this thesis, the threshold is set at 30,000€ to have enough large claims for modeling.
Then, when the distinction is made, attritional claims are modeled with cost/frequency model in order to distribute the fairest the cost between guarantees. Large claims' cost is evenly distributed all over the guarantees. The aim of models in this thesis is to distribute the large claims' cost differently (no longer homogenous distribution).

Data base

The creation of this database was made with several extractions. The global idea is to join several "portfolio photos" in a database at different moment. Freezing the information of all the active contracts, of the clients attached to the contracts and of their claims on a chosen date is called "photo of the portfolio". Photos were taken on 7 dates: all of 1st January between 2012 and 2018. For each contract in the portfolio at the photo's date, the following information is extracted: the claims that the insured has declared out of the 5 last years, the fire claims of the insured over the 12 months following the photo, the client's equipment (that is to say if the client has other contracts such as for example a car, health contract, etc.) and finally the various guarantees of the contract subscribed. We can plot photos as follows:



All data is then aggregated into a single database before reworked. Indeed, due to the different generation of this product, several variables have been changed over time, for example the level of the franchise, and it is necessary to harmonize everything. In addition, several categorical variables were reworked to have less categories and several continuous variables were reworked into categorical variables, like the are of dependencies. When restatements are made, all rows with missing data have been deleted. Although this could skew the results, the proportion of these lines was insignificant.

Modeling

Modeling was built by comparing the results of different machine learning models in order to choose the best. Logit regression model and decision trees model were used. In addition, in order to improve decision trees model, a gradient boosting was applied. In the same logic, sampling methods (sub-sampling and cross-validation) were applied to the data.
In order to compare all these models, performance criterions have been put:

- Confusion matrix to see if an insured is correctly classified.
- Classification table to set the probability threshold which separate risky insured to non-risky insured.
- ROC curve (Receiver Operating Characteristic) and AUC (Area Under the Curve). ROC curve explains the evolution of right positive rate (sensibility) on the false positive rate (1-specificity) and is represented on a plot. We can calculate the area under the curve (AUC). The closer the AUC is to 1, the more efficient the model.

AUC's results for all the models are between 0.62 and 0.68. It is not satisfying. Indeed, an AUC at 0.5 is like random. In our case, models are a little bit better than random.
Finally, a cost function was introduced in order to see the difference between the loss of premium if the insurer decided to cancel all contracts with a high risk of fire disaster and the gain of all fire disaster avoided.

Application

Despite weak model results, we wanted to present the approach of the business application that results from modeling. We keep the logit model because it is easy to explain the coefficients (direction of variation, intensity of variables) and the performance is close to the best model.
With this model, we can order all contracts by the probability to have a fire claim. When we keep the 500 contracts whose are riskier, we can compare the distribution of the global population to the risky population. We can note that insured with dependencies areas more than 900m² or insured with more than 9 principal rooms are the most in the risky database. These criterions permit to do filter on present portfolio contracts in order to select insured to appointing an expert.
Moreover, the coefficients of the model give indications to distribute non-homogenously the large claims' loss. Indeed, for the variables we have seen before, the distribution can be made according to the importance of the coefficient.

Conclusion

The frequency of these claims is to low, machine learning models can't perform despite a large claim's threshold at 30,000€. However, the models converge towards variables which can explain fire disaster. Therefore, a depth study on the insured which have these characteristics should be interesting to continue (like people who have more than 900m² of dependency area). This study should be led by the subscription service which studied in a depth way the

selected insureds. If the study is approved (if the risk to have a fire disaster is proved for each insured studied) so recurrent controls could be done for these profiles.

Remerciements

Je tiens à remercier en premier lieu Thélem Assurances de m'avoir accueilli durant cette année d'alternance et particulièrement à Céline Pichard de m'avoir fait confiance sur ce projet et qui m'a permis de présenter ce mémoire. Je remercie également toute l'équipe de l'offre qui m'a accompagné et qui était disponible pour répondre à mes questions.

Je remercie également l'ISFA et l'ensemble des professeurs qui permettent de mener à bien le suivi entre les cours et l'alternance. Ils ont su se montrer disponible tout au long de la scolarité.

Enfin je remercie tous mes proches, en particulier mes parents, ma conjointe, et mes amis de m'avoir soutenu du début jusqu'à la fin de ce mémoire.

Table des matières

Introduction

D'après la définition de la Fédération Française de l'Assurance (FFA), un « contrat d'assurance multirisque habitation offre à l'assuré des garanties complètes pour protéger son patrimoine familial contre les conséquences d'évènements affectant son domicile ou mettant en cause sa responsabilité ou celle des membres de sa famille ». En outre, l'assurance multirisque habitation est composée de diverses garanties telles que le dégât des eaux, le vol, les catastrophes naturelles, les attentats, le bris de glace, l'incendie, la responsabilité civile, etc, permettant ainsi de couvrir les dommages matériels subis par le logement, et matériels ou corporels causés à un tiers, que l'assuré soit propriétaire, locataire ou non-occupant.

En France, le marché de l'assurance multirisque habitation représente 10.5Md€ de cotisations acquises en 2017 contre 7Md€ de prestations versées, un montant en hausse dû aux catastrophes naturelles survenues durant l'année. De plus, d'après le rapport annuel de la FFA de 2017, 9 000 sinistres habitations surviennent par jour, dont 530 sinistres incendie. En 2018, Thélem Assurances a connu une augmentation de 3.5% de la fréquence des sinistres incendie des contrats Habitation par rapport à 2017, c'est pourquoi une étude approfondie sur ce type de sinistre semble intéressante puisque les sinistres incendie ont une faible fréquence mais une sévérité importante lors de sinistre total, la charge sinistre atteint rapidement des niveaux élevés.

Aujourd'hui, afin de prédire au mieux la charge sinistre, et donc proposer un tarif adapté au client, les assureurs font appel à des modèles d'apprentissage statistique. Ces modèles estiment le comportement des sinistres dont la fréquence est élevée mais dont le coût est faible. Cependant, il est plus compliqué de modéliser le comportement des sinistres graves. L'objectif de cette étude est de construire des modèles d'apprentissage statistique qui permettent de détecter le risque d'un assuré d'avoir un sinistre incendie grave dans les douze mois. Grâce aux prédictions, il sera possible de déceler des profils de clients ayant un fort risque de provoquer un sinistre incendie grave et ainsi ajuster au mieux la prime de la garantie incendie pour ces profils. Enfin, pour les individus dont le risque sera maximum, la compagnie pourra mandater un expert pour une visite de risque dont l'intention sera de vérifier si le logement présente un réel risque en l'état et si des mesures de prévention sont envisageables.

Dans ce mémoire, nous présenterons dans un premier temps l'objet et le contexte de l'étude. Ensuite, sera exposé le cheminement complet de l'extraction des données à la base finale. Les deux dernières parties seront consacrées à la présentation des modèles d'apprentissage utilisés ainsi que leurs performances et l'application métier des résultats. Des sujets de réflexions sur des thèmes complémentaire seront exposés en toute dernière partie de ce mémoire.

I. Mise en place du cadre d'étude et du périmètre.

Dans cette partie, une présentation du portefeuille habitation de Thélem assurances ainsi que la sinistralité de celui-ci sont détaillées.

1. Le produit habitation à Thélem Assurances :

Thélem Assurances est une société d'assurance mutuelle, proposant des produits en santé, accidents de la vie, épargne, prévoyance mais aussi en automobile, agricole et professionnels. En plus de ces produits, il y a le produit MultiRisque Habitation (MRH), dont trois générations différentes sont présentes actuellement dans le portefeuille. Nous avons la génération MRH, qui est le plus ancien des produits mais qui n'est plus commercialisé aujourd'hui, la génération Néologis, qui a été commercialisé à partir de 2012 et enfin la génération Néologis 2 qui est commercialisé depuis juin 2018. Actuellement, le portefeuille est composé de 27% de contrats MRH, 68% de Néologis et 5% de Néologis 2.

Selon les dispositions générales, ce qui est couvert par la garantie incendie sont les dommages matériels causés aux biens assurés par :
• un incendie (combustion avec flammes en dehors d'un foyer normal), une explosion, une implosion, un dégagement accidentel de fumées,
• la chute de la foudre,
• la chute d'appareils de navigation aérienne, ou spatiaux (ou des objets tombant de ceux-ci),
• le choc de véhicules, c'est-à-dire les dommages matériels autres que ceux d'incendie ou d'explosions, causés aux biens assurés par le choc d'un véhicule terrestre quelconque, dont vous n'avez ni la propriété, ni l'usage, ni la garde.

Chaque génération du produit habitation est décomposée en formules afin de couvrir les clients selon leurs besoins. Par exemple, le produit Néologis 2 est composé de 5 formules : Déclic, Tonic, Dynamic, Confort et Zen. Nous pouvons constituer le tableau suivant, détaillant les garanties acquises en fonction de la formule souscrite. De plus, il est possible d'ajouter des options couvrant des risques spécifiques non-assurés avec l'offre initiale.

	Déclic	Tonic	Dynamic	Confort	Zen
Dommages électriques					
Pertes denrées congélateurs caves à vins					
Emeutes - mouvements populaires					
Honoraires expert d'assuré					
PJ Bailleur					
RC Assistante Maternelle					
Responsabilité civile vie privée					
Responsabilité civile Prop. d'Immeuble					
Usages collaboratifs					
Incendie					
Dégâts des eaux, gel					
Tempêtes, grêle, neige sur toitures					
Catastrophes naturelles					
Catastrophes technologiques					
Attentats matériels					
Vol - Actes de vandalisme					
Bris de glaces					
Défense pénale recours suite à accident					
Inondation					
Vol des objets de valeur					
Assistance					
Multimédia nomades, objets de loisirs					
Biens en plein air et végétaux					
Canalisations extérieures					
Perte d'eau canalisations extérieures					
Energies renouvelables					
Pannes électroménager, audio et vidéo					
Piscine et spa					

Légende	Garantie non disponible	Garantie Obligatoire	Garantie Optionnelle

Tableau 1 : Tableau des garanties en fonction des formules pour le produit Néologis 2.

Par exemple, si un client souscrit la formule Tonic, il sera garanti pour la Responsabilité Civile vie privée, les usages collaboratifs, les incendies, les dégâts des eaux et gel, les tempêtes, grêle et neige sur toitures, les catastrophes naturelles, les attentats matériels, les vols et actes de vandalisme, le bris de glaces, la défense pénale et recours à la suite d'un accident, les inondations et l'assistance. Il pourra prendre en option la responsabilité civile assistance maternelle ou encore la garantie multimédia nomades et objets de loisir qui permet d'assurer des objets tels que des tablettes tactiles à la suite de chute, choc ou de vol par agression.

Au niveau de la répartition du portefeuille, à fin décembre 2018, le portefeuille était constitué principalement de propriétaire occupant de maison avec 4 pièces principales et des surfaces

de dépendances comprises entre 0 et 40m². Les contrats ont, pour 19% du portefeuille, entre 4 et 6 ans d'ancienneté.
De plus, Thélem Assurances est principalement implanté en Région Centre Val de Loire et particulièrement dans le Loiret puisque l'on retrouve 12% du portefeuille habitant dans le Loiret devant la Loire Atlantique (8%) et l'Indre (6%).

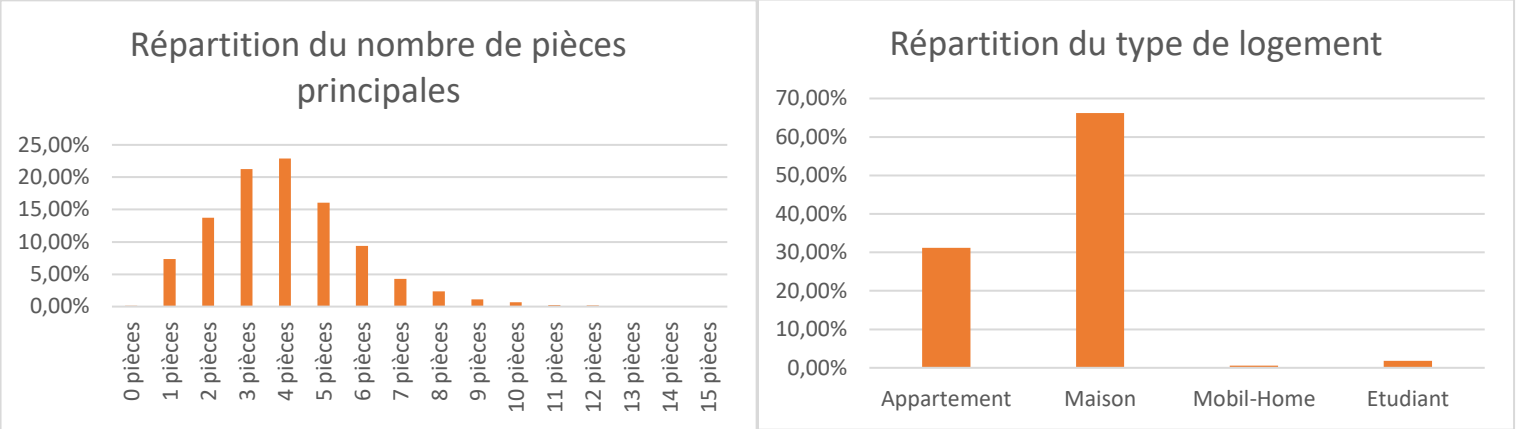


Figure 1 : Exemple de répartition des variables. A gauche, la répartition du nombre de pièces principales. A droite, la répartition du type de logement.

2. Etude de la Sinistralité du produit Habitation

Avant de débuter l'étude des sinistres du produits habitation, quelques rappels de formules sont exposés afin de mieux comprendre les résultats.

Rappels :

- L'exposition correspond au temps passé par un contrat dans le portefeuille. C'est-à-dire, si le contrat prend effet le 1er mars de l'année N et que le client résilie au 1er mars de l'année N+1, alors le contrat aura une exposition d'environ 0.83 (306 / 365 si l'année n'est pas bissextile) durant l'année N. L'exposition totale d'une année est la somme des expositions contrats de tout le portefeuille.
  L'intérêt de calculer l'exposition est de pouvoir pondérer les contrats du portefeuille. En effet, un contrat qui a passé la moitié de l'année dans le portefeuille et qui a eu un sinistre est plus risqué qu'un contrat qui est resté toute l'année et qui n'a eu qu'un sinistre. Cela permet de calculer la fréquence au plus juste.

- $Fréquence\ de\ l'année\ N\ = \frac{nombre\ de\ sinistre\ survenus\ l'année\ N}{exposition\ totale\ de\ l'année\ N}$

- $Coût\ Moyen = \frac{Montant\ total\ des\ sinistres}{Nombre\ total\ de\ sinistres}$;

- $Coût\ Moyen\ écrêté = \frac{Montant\ total\ écrêté\ des\ sinistres}{Nombre\ total\ de\ sinistres}$.

- Si on considère X le montant d'un sinistre. Le montant écrêté de ce sinistre est égal au minimum entre le montant du sinistre X et le seuil d'écrêtement. Le seuil d'écrêtement est le seuil qui délimite les sinistres attritionnels des sinistres graves. Dans notre cas, le seuil d'écrêtement est de 30 000€. C'est à dire si X est supérieur à 30 000€, le montant écrêté sera de 30 000€, sinon si c'est inférieur alors le montant écrêté sera égal à X. Le montant total écrêté correspond à la somme de tous les montants de sinistres écrêtés.
  On peut définir la surcrête comme le total des montants au-delà du seuil d'écrêtement.

Le tableau qui suit regroupe des informations sur la sinistralité du produit habitation et des informations sur les sinistres incendie.

	2012	2013	2014	2015	2016	2017	2018
Exposition	260 988	266 736	273 305	280 383	285 418	288 597	292 293
Nombre total de sinistre Habitation	23 199	25 471	24 445	19 655	24 109	23 084	22 158
Nombre total de sinistre incendie Habitation	1 941	2 552	2 300	2 375	2 532	2 469	2 572
Charge des sinistres totale Habitation	35 657 453	40 202 330	43 633 387	29 637 952	69 449 963	44 594 780	49 439 645
Charge des sinistres incendie Habitation	9 785 360	12 219 034	11 867 923	10 107 615	12 524 381	13 269 375	12 780 156
Charge des sinistres incendie graves Habitation (> 30 000€)	6 413 709	8 062 025	8 497 438	5 818 748	8 452 048	9 156 483	8 779 738
Coût Moyen des sinistres incendie Habitation	5 271	4 790	5 169	4 318	4 946	5 374	4 969
Coût Moyen Ecrêté des sinistres incendie Habitation	2670	2411	2 261	2 367	2 343	2 407	2 220

Tableau 2 : Tableau recensant les informations sur la sinistralité des sinistres Habitation et plus particulièrement les sinistres incendie

D'après le tableau, on remarque une constante progression de l'exposition du portefeuille, avec en moyenne 2% d'augmentation par an entre 2012 et 2018, ce qui se traduit par une augmentation du risque (plus on a de clients dans le portefeuille, plus il y a de risque de couvrir un sinistre). De plus, 16 741 sinistres Incendie sont survenus sur la période du 1er janvier 2012 au 31 décembre 2018, avec les années 2013 et 2016 atypiques à la suite des différents évènements climatiques tels que les orages, qui ont entrainés des sinistres incendie.
Si on se focalise sur les sinistres incendie en habitation, on remarque que le nombre de sinistres incendie représente environ 10% du nombre de sinistres totaux en Habitation. Cependant, la charge des sinistres incendie représente environ 30% de la charge totale. Les sinistres incendie ont donc des montants plus élevés par rapport aux autres sinistres.
De plus, la part de la charge des sinistres incendie graves représente entre 65 et 70% de la charge totale des sinistres incendie Habitation. Cela signifie que les sinistres graves sont prépondérants dans la garantie incendie des contrats Habitation.

Si l'on observe la répartition des montants de sinistres incendie en habitation, on obtient l'histogramme suivant :

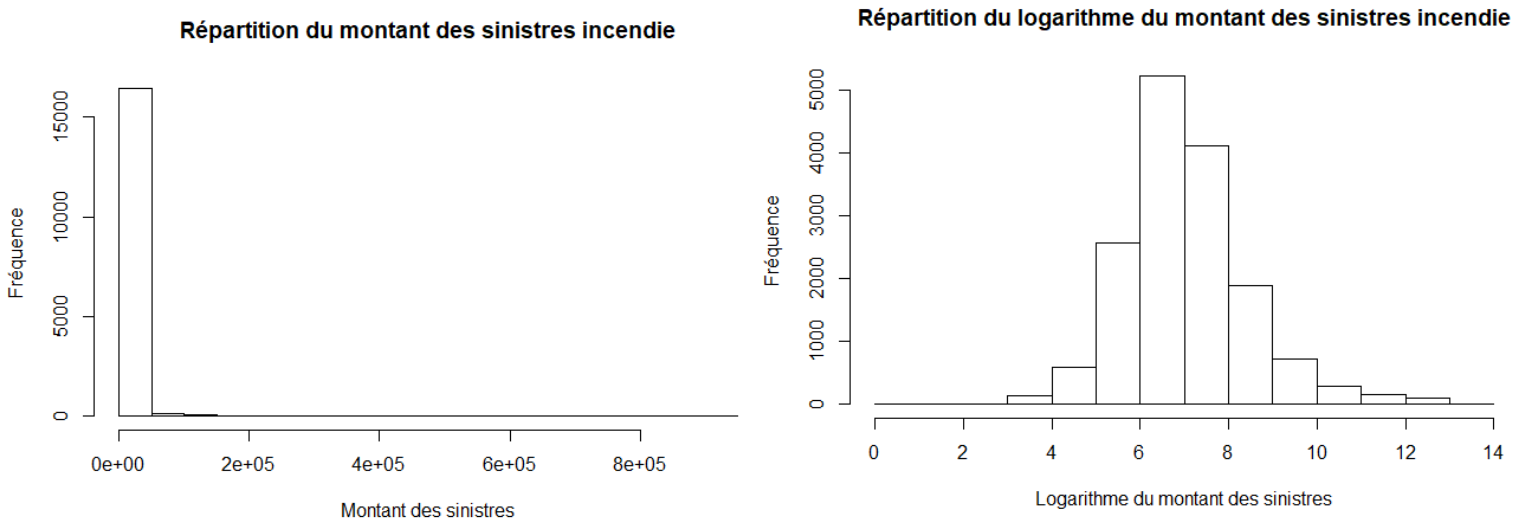


Figure 2: Histogrammes de la répartition des montants de sinistre incendie et de la répartition du logarithme des montants de sinistres incendie.

D'après l'histogramme sur la répartition des montants de sinistres, on remarque que la majorité des sinistres incendie sont très faibles. Afin de mieux observer la répartition des montants de sinistres, on considère le logarithme du montant des sinistres incendie.

3. Sinistres attritionnels et sinistres graves

Dans la section précédente, il est question de sinistres attritionnels et de sinistres graves. Nous allons dans cette partie exprimer la différence entre les deux notions.

On parle de sinistre attritionnel lorsque ce sont des sinistres à faible coût mais avec une fréquence relativement élevée. Au contraire, les sinistres graves correspondent à des sinistres dont le montant dépasse un seuil déterminé mais qui ont généralement une fréquence très faible. Les sinistres attritionnels ne suivent pas la même loi que les sinistres graves. Il est donc nécessaire d'effectuer une étude sur le montant des sinistres afin de déterminer le seuil qui différencie les sinistres graves des sinistres attritionnels. Pour ce faire, la théorie des valeurs extrêmes permet d'identifier le seuil à l'aide de différentes méthodes. Le choix du seuil est important car plus le seuil est élevé, moins il y aura de sinistres graves et alors il sera complexe de modéliser les montants de sinistres graves. Inversement, plus le seuil est bas et plus il y a de chances de ne pas modéliser correctement les sinistres attritionnels. Lors de la création du

produit Néologis 2, une étude a été effectuée à l'aide du logiciel **AddactisPrincing**. Les résultats sont présentés ci-dessous.

Cadre théorique :

On considère X_1, X_2, …, X_n des variables aléatoires indépendantes et identiquement distribuées représentant les montants des sinistres.
On pose $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$ la statistique d'ordre, avec $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$.
Soit $\mathrm{F}(\mathrm{x}) = \mathrm{P}(\mathrm{X} \leq \mathrm{x})$ la fonction de répartition des X_i. On a $\bar{\mathrm{F}}(\mathrm{x}) = 1 - \mathrm{F}(\mathrm{x})$ la fonction de survie de F.
Soit $M_n = max\ (X_1, \ldots, X_n)$.

Définition 1 : Distribution des extrêmes généralisés (GEV)
La distribution des extrêmes généralisés est définie par la fonction de répartition,

$$G(\mu, \sigma\, \xi) = \begin{cases} exp\ (-(1 + \ \xi \left(\frac{x - \mu}{\sigma}\right))_+^{-\frac{1}{\xi}}, & si\ \xi \neq 0\ ; \\ exp\left(-\left(exp\left(-\left(\frac{x - \mu}{\sigma}\right)\right)\right)\right), & sinon. \end{cases}$$

Où μ est le paramètre de position, σ le paramètre d'échelle et ξ le paramètre de forme ou d'épaisseur de la queue.

La définition 1 représente la distribution asymptotique du maximum parmi un ensemble de valeurs observées. Cependant, ce n'est pas le maximum qui nous intéresse mais les valeurs au-dessus d'un seuil. C'est pourquoi on étudie le comportement des excès des montants de sinistre au-delà d'un seuil s représentés par les variables aléatoires indépendantes et identiquement distribués $(Xi - s \mid Xi > s)_{i = 1 \text{ à } n}$.

Théorème 1 : Distribution des excès au-delà d'un seuil
Soit $\xi \in \mathbb{R}$
Il existe deux suites a_n et b_n telles que $\mathbb{P}\left(\frac{\mathrm{M_n} - \mathrm{b_n}}{\mathrm{a_n}} \leq \mathrm{x}\right)$ converge en $+\infty$ vers une distribution des extrêmes généralisée de paramètre ξ si et seulement si il existe une fonction a telle que :

$$\lim_{\mathrm{u} \to \mathrm{x^F}} \frac{\bar{\mathrm{F}}(\mathrm{s} + \mathrm{x}\mathrm{a}(\mathrm{s}))}{\bar{\mathrm{F}}(\mathrm{s})} = \begin{cases} (1 + \xi \mathrm{x})_+^{-\frac{1}{\xi}}, & \text{si } \xi \neq 0\ ; \\ \exp(-\mathrm{x}), & \text{sinon.} \end{cases}$$

Avec x^F le point extrêmal associé à la distribution de fonction de répartition F

Par le théorème 1, la loi des excès des montants de sinistre est une loi de Pareto Généralisé (GPD : *Generalized Pareto Distribution*) de paramètre de forme ξ et de paramètre d'échelle 1 avec :

$$G_{(\sigma,\xi)}^{GPD} = \begin{cases} 1 - \left(1 + \frac{\xi x}{\sigma}\right)_+^{-\frac{1}{\xi}}, & \text{si } \xi \neq 0 ; \\ 1 - \exp\left(\frac{-x}{\sigma}\right), & \text{sinon.} \end{cases}$$

On cherche donc à trouver un seuil s qui permet de dire que la distribution des montants des excès vérifie les propriétés d'une distribution Pareto Généralisée.

i. Fonction moyenne des excès

Soit *e* la fonction des excès.
On a : $e(s) = E[X - s \mid X > s]$ avec s>0.
La méthode de la fonction moyenne des excès consiste à calculer l'estimateur empirique de la fonction des excès puis de tracer les points de coordonnées $(X_{(i)}; \hat{e}_n(X_{(i)}))$ où $\hat{e}_n(X_{(i)})$ est l'estimateur de la fonction des excès pris au point $X_{(i)}$.

Définition 2 : Estimateur empirique de la fonction de dépassement moyen.
L'estimateur empirique de la fonction de dépassement moyen se définit par :

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

La définition 2 peut être vue comme le montant total des excès divisé par le nombre total de montants supérieurs au seuil.

On peut alors tracer le graphique suivant :

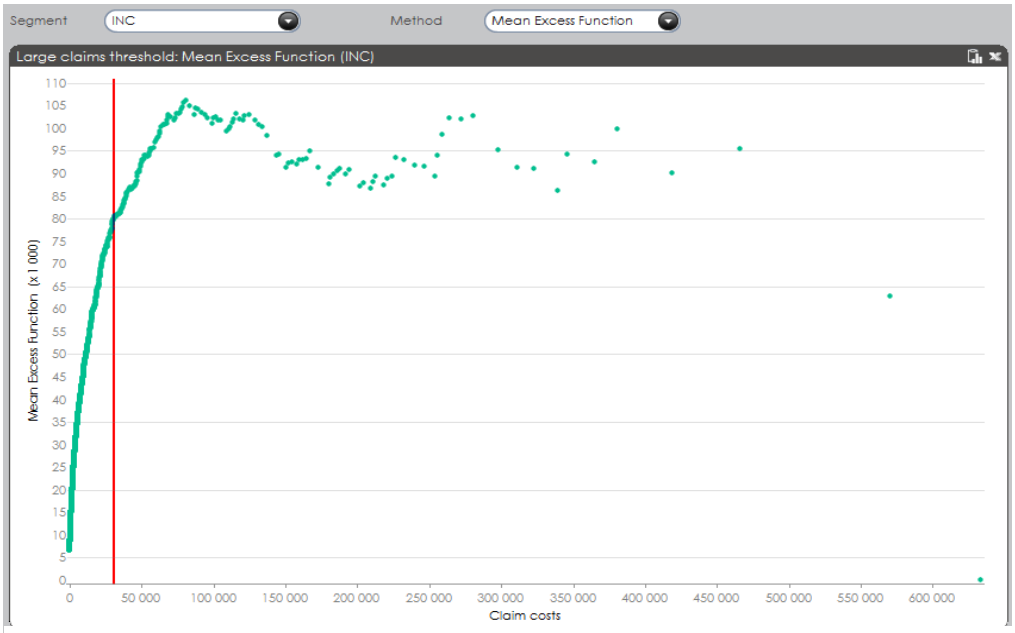


*Figure 3: Sortie du logiciel **AddactisPrincing** présentant la fonction des excès moyens.*

D'après ce graphique, nous remarquons un changement de tendance de la fonction d'excès moyen vers la valeur 30 000€ et devient instable ensuite. Par conséquent, on peut considérer à l'aide de cette méthode de fixer le seuil à 30 000€. De plus, on remarque que la fonction est croissante, cela signifie que la loi possède une queue épaisse.

ii. Estimateur de Hill

La variation des estimateurs en fonction du seuil choisit permet de sélectionner le seuil des graves. C'est une approche non paramétrique et elle ne nécessite pas d'hypothèse sur la distribution des montants de sinistre.

Définition 3 : Estimateur de Hill.
On définit l'estimateur de Hill comme l'estimateur du maximum de vraisemblance du paramètre de queue ξ :

$$\xi_{k,n}^{Hill} = \frac{1}{k}\sum_{i=1}^{k} \ln\left(X_{(i)}\right) - \ln\left(X_{(k+1)}\right)$$

D'après la définition 3 de l'estimateur de Hill, nous pouvons tracer le graphique suivant à partir des montants de sinistres triés par ordre décroissant et de la valeur estimée de l'estimateur de Hill afin de déterminer à partir de quel moment il y a une stabilisation.

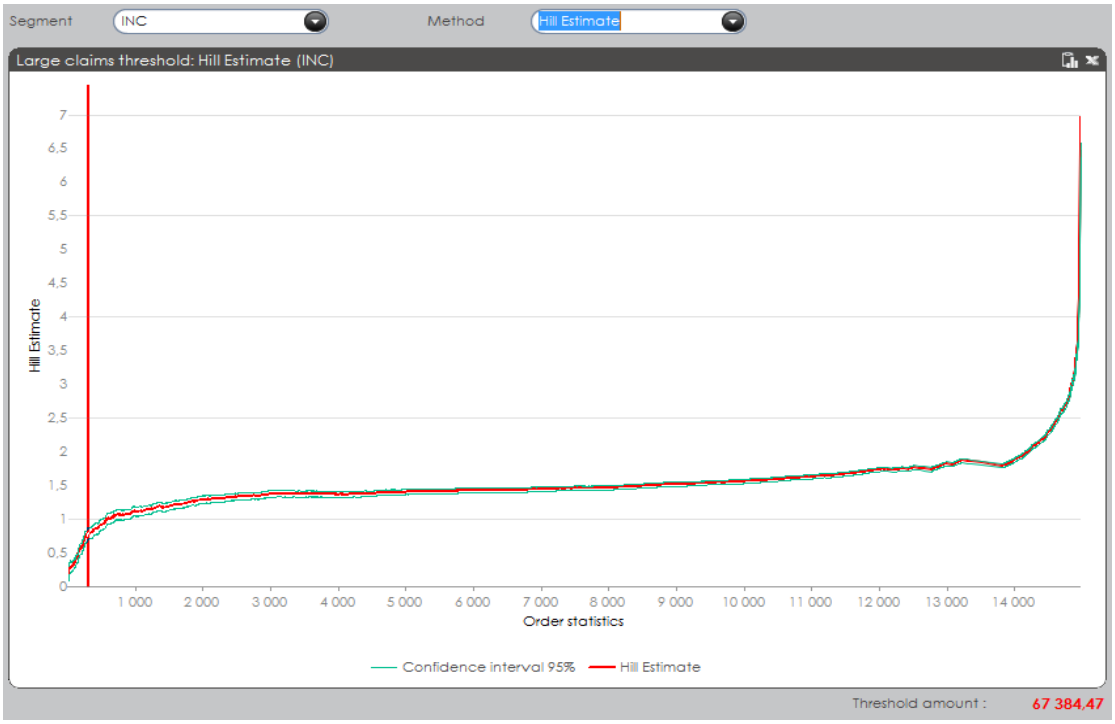


Figure 4: Sortie du logiciel AddactisPrincing présentant la représentation de l'estimateur de Hill.

D'après la sortie du logiciel **AddactisPricing**, le seuil est d'environ 67 000€. Ce seuil parait élevé pour la suite de la modélisation car il y aurait trop peu de sinistres graves.

Pour la suite de l'étude, nous fixerons le seuil à 30 000€. Ce seuil correspond aussi au seuil historique utilisé par Thélem Assurances.

4. Le principe de création du tarif Habitation

Un des objectifs des modèles de prédiction de ce mémoire est de pouvoir détecter des profils de risques de clients ayant un fort risque d'avoir un sinistre incendie grave. A l'aide de ces profils de risque, la charge sinistre des sinistres graves pourrait être répartie, non plus de manière uniforme, mais avec des poids différents selon les caractéristiques du bien et du client. Une brève explication du processus de tarification est exposée dans cette partie afin de mieux comprendre la construction et la répartition de la charge sinistre des graves.

La première étape de l'élaboration du tarif est la constitution d'une base de données regroupant les caractéristiques clients ainsi que leurs sinistralités antérieures mais aussi des

données socio-démographiques. Cette base est construite sur un historique de 5 ans pour ne pas avoir trop d'hétérogénéité entre les données due au temps.
Ensuite, lorsque la base est extraite, une étude du portefeuille est effectuée afin de déterminer les profils de risques qui ressortent mais aussi pour observer les variables pouvant être corrélées entre-elles. En effet, pour utiliser les modèles linéaires généralisés nous avons besoin d'une hypothèse d'indépendance des variables. Si cette condition n'est pas vérifiée, alors le modèle ne sera pas efficient. Les variables corrélées sont supprimées de la base de données, les autres seront les critères tarifant du produit. Les données sont retraitées (regroupement de catégories pour les variables catégorielles, traitement des valeurs manquantes, …). De plus, une étude sur la sinistralité est effectuée afin de séparer les sinistres attritionnels des sinistres graves. En effet, ce sont les sinistres attritionnels qui seront modélisés et la charge sinistre des graves (la surcrête) sera répartie de façon uniforme sur toutes les garanties.

C'est un modèle « Coût - Fréquence » qui est modélisé, c'est-à-dire que l'on utilise une loi de poisson pour modéliser la fréquence, et une loi Gamma pour modéliser le coût. Pour chaque garantie, un modèle de coût et un modèle de fréquence sont construits, donc s'il y a n garanties, alors il y a 2 * n modèles de créés.
Lorsque les modèles sont finalisés, les tarifs par garantie sont obtenus en multipliant l'estimation du coût par l'estimation de la fréquence. La surcrête est ensuite ajoutée de façon homogène. C'est à cette étape que les modèles statistiques interviendraient en pondérant chaque garantie par l'importance de la variable dans le modèle de détection d'un sinistre grave. La prime pure finale est la somme de toutes les primes pures par garanties.

Des chargements sont ajoutés afin de couvrir les frais de gestions, les frais de réassurance moins les produits financiers et enfin il faut ajouter la Taxe Spéciale sur les Conventions d'Assurances (TSCA).

II. Base de données.

Cette partie permet d'expliquer le cheminement qui a permis d'obtenir la base de données finale qui sera utilisée pour la modélisation.
Depuis 2003, Thélem Assurances a mis en place un Dataware permettant d'avoir des données structurées et les plus fiables possibles.

Idée générale de la construction de la base de données :
Nous avons décidé d'ordonner la base de données de la façon suivante. La base de données est constituée de « vues » à différentes dates (ces vues peuvent être comparées à des photos) regroupant les informations des contrats actifs à la date de la vue mais aussi des informations des sinistres des contrats sélectionnés. 7 vues ont été extraites : à chaque 1er janvier des années 2012 à 2018. Le fait de fixer des dates et de regarder si le contrat a eu un sinistre l'année suivante permet de répondre à la problématique initiale. En effet, nous cherchons à connaître à une date donnée le risque que le contrat ait un sinistre dans l'année suivant la date.

1. Partie Sinistres :

Les données sur les informations sinistres sont décomposées en deux parties. La première regroupant les informations des sinistres futurs. C'est-à-dire si le contrat a eu un sinistre grave dans l'année qui suit la photo. Par exemple, nous avons un contrat vu au premier janvier 2015. Si le contrat a eu un sinistre incendie grave entre le premier janvier 2015 et le 31 décembre 2015, alors l'information du sinistre est récupérée dans une variable binaire qui vaut 0 si l'individu n'a pas eu de sinistre grave, 1 sinon. Dans la modélisation ce sera la variable à expliquer, c'est-à-dire celle que nous chercherons à prédire. Plusieurs variables telles que l'identifiant sinistre, la date de survenance du sinistre ou bien les montants nets et brut de recours sont, dans un premier temps, retenus afin de garder un maximum de variables avant le retraitement de la base. Une seule base de données est extraite regroupant tous les sinistres incendie entre le 1er janvier 2012 et le 31 décembre 2018.

La deuxième partie des informations sinistres est composée de la sinistralité antérieure du contrat. Un historique des 5 dernières années est construit en regroupant plusieurs variables binaires, permettant de recueillir diverses informations sur la nature, l'antériorité (1 ans et 5 ans) et le nombre de sinistres : par exemple si le contrat a eu au moins un sinistre quelconque dans les 5 dernières années ou bien si le contrat a eu au moins un sinistre incendie l'année précédente ou encore le nombre de dommages électriques que le contrat a eu sur les 5 dernières années. Au total, 11 variables sont créées. Une seule base de données est extraite comme pour les sinistres futurs.

Une liste des variables sinistres est présentée dans l'annexe A.

Problème rencontré :

Nous souhaitions étudier les rapports d'expertises des sinistres graves afin d'observer s'il y avait des éléments en commun. Nous voulions utiliser la méthode de fouille de texte (*Text Mining*) afin de détecter la fréquence d'apparition de certains mots.
Après avoir analysé quelques rapports, nous nous sommes rendu compte que cela serait compliqué dans le temps imparti. En effet, le logiciel de traitement informatique des rapports d'expertise n'était pas encore en place sur la première partie de notre périmètre. Les rapports d'expertise ne sont pas disponibles sur cette période. De plus, d'une compagnie d'experts à une autre, les rapports ne sont pas structurés de la même façon et ne regroupent pas les mêmes informations. Et dans une même compagnie d'expertise, les rapports sont aussi structurés différemment en fonction de l'expert qui rédige le rapport. Nous avons alors décidé de ne pas ajouter ces informations.

2. Partie équipement clients

Ensuite, nous apportons l'information de l'équipement client. Nous entendons par équipement client, le fait qu'un assuré ait souscrit à d'autres contrats au sein du portefeuille Thélem Assurances.
L'extraction est faite au 1er janvier de 2012 à 2018 sur tous les clients actifs du portefeuille. 7 bases de données équipements sont créées avec des variables qui comptabilisent le nombre de contrat MRH, Automobile, Accident de la vie, protection juridique, santé et les autres contrats. Ces données peuvent apporter des informations sur le profil de l'assuré : si l'assuré possède plusieurs contrats alors il se pourrait qu'il soit averse au risque en se couvrant un maximum. De plus, afin de pouvoir souscrire plusieurs contrats il faut avoir des moyens financiers suffisants. Nous pouvons donc penser qu'un client équipé fasse le nécessaire pour éviter tout départ d'incendie.

Problème rencontré :

Nous souhaitions ajouter la segmentation commerciale du client. Cependant, elle n'est fiable que depuis 2015 et ne rentre pas dans le périmètre de notre étude.
De plus, en ce qui concerne l'information de l'équipement, nous aurions voulu aller au niveau foyer (« quels sont les autres contrats souscrits au niveau du foyer d'un assuré ? »), mais ces informations ne sont exploitables que depuis 2013 et ne rentrent pas non plus dans le périmètre. C'est pourquoi nous nous sommes restreints au niveau client.

3. Partie Contrats

Enfin, les données sur les informations contrats sont extraites. Les données contrats sont par exemple le nombre de pièces principales, la surface des dépendances, le type de l'occupation de l'habitation, … Comme pour l'information équipement, 7 bases sont extraites correspondant aux 7 années du périmètre.

Nous pouvons alors schématiser ces 3 parties comme le graphique suivant :

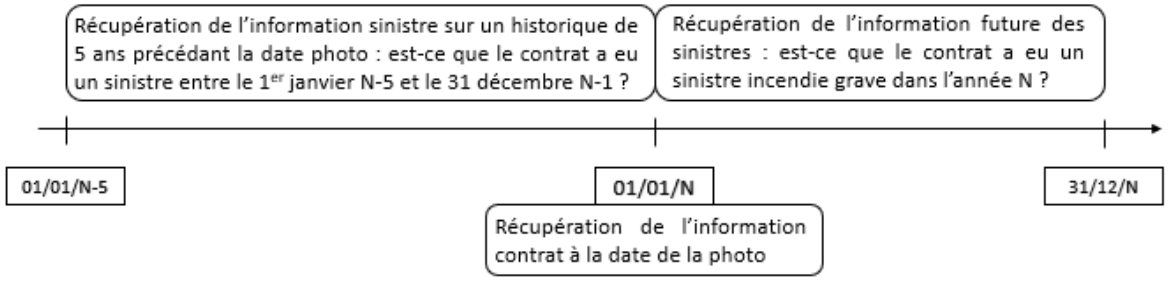


Figure 5 : Schéma explicatif de l'extraction des données pour une année N

Problème rencontré :

Les données sur les matériaux de construction du logement n'ont pas pu être intégrées dans la base de données car l'information n'est pas présente dans le modèle de données. De même sur l'année de construction qui peut paraître compliquée à intégrer puisque des logements peuvent être anciens mais entièrement rénovés. Il faudrait alors avoir la date de construction et potentiellement la date de rénovation. Mais ces questions ne sont pas demandées lors de la souscription du contrat. Il aurait fallu modifier le questionnaire de souscription. Pour des questions de temps et de simplification, les informations n'ont donc pas pu être intégrées.

4. Jointure

Première jointure :

Lorsque les premières bases sont extraites, une première jointure est nécessaire afin de regrouper, année par année, les données des contrats, des sinistres passés et futurs ainsi que les données sur l'équipement client.
Dans cette jointure, quelques lignes des sinistres futurs ne sont pas insérées car un contrat peut ne pas être actif au 1er janvier de l'année N mais avoir un sinistre incendie durant cette même année. La base sinistre est dotée de la ligne correspondante à ce sinistre mais la base contrat n'est pas alimentée de l'information puisque le contrat n'était pas actif au moment de la vue. La base jointe de la vue a autant de lignes que de contrats actifs à cette date-là.

Entre les données contrats et les données sinistres : pour chaque base contrats, la jointure est faite de la façon suivante : si le contrat j a eu un sinistre entre le 1er janvier et le 31 décembre de l'année N alors l'information est ajoutée dans la base.

Ensuite, la deuxième jointure est effectuée en regroupant les 7 bases ensemble.
Plusieurs jointures sont opérées : une première jointure est effectuée afin de stocker les informations de chaque année. 7 bases de données, représentant les 7 photos, sont ainsi créées. Ensuite une jointure est effectuée afin de regrouper toutes les photos. La base initiale sans retraitement est terminée et elle est constituée de 1933248 lignes et environ 98 variables.

5. Retraitement

Après avoir obtenu la base de données complète, il est nécessaire de vérifier la qualité des données mais aussi de retravailler les variables. Cette partie permet d'expliciter les différents retraitements opérés qui ont mené à la base de données qui servira à la modélisation.

Les variables présentes initialement ont servi lors de la jointure de toutes les bases, comme par exemple l'identifiant contrat ou les dates d'effet du contrat. Ces colonnes ne seront pas nécessaires à la modélisation et sont supprimées.

Ensuite, des variables servent à harmoniser les deux générations de produit. Par exemple, pour les formules, il existe 5 formules pour le produit Néologis qui ne correspondent pas exactement aux mêmes formules du produit MRH. Ce sont les préconisations qui ont été établies au moment du changement de génération qui ont permis de faire l'harmonisation afin de créer la variable « formule » regroupant les deux générations. Les variables intermédiaires sont supprimées. La même manipulation a été nécessaire pour les niveaux de franchises.

D'autres variables catégorielles sont retraitées afin de regrouper des catégories. Par exemple, le code qualité contient 13 catégories : Colocataire occupant partiel, Colocataire occupant unique, Propriétaire occupant total, Propriétaire occupant partiel, Copropriétaire occupant unique, Copropriétaire occupant partiel, Locataire occupant unique, Locataire occupant partiel, Usufruitier, Propriétaire non occupant, Copropriétaire non occupant, Locataire non occupant, Nu-propriétaire. Les classes présentent des disparités en termes de nombre d'individus et certaines classes peuvent être regroupées ensemble sans perdre d'information. Le retraitement ici est de passer de 13 classes à 5 : les non-occupants, les propriétaires occupants partiel, les locataires occupants partiel, les locataires occupants et les propriétaires occupants. Cela permet de réorganiser le nombre d'individus par classe.

Certaines variables continues sont transformées en variable catégorielle afin de segmenter au mieux le portefeuille. Par exemple le nombre de pièces principales et la surface des dépendances. Le regroupement a été fait suivant la répartition de la variable mais aussi en ajustant les groupes en homogénéisant la fréquence intra-groupe afin de regrouper des profils

similaires dans chaque classe. La variable « nombre de pièces principales » possède après retraitement 4 catégories et la variable « surface des dépendances » 6 catégories.

Une fois que tous les retraitements sont effectués, toutes les lignes avec des valeurs manquantes sont supprimées afin de ne pas avoir de problème lors de l'implémentation des modèles. De plus, nous ne souhaitions pas remplacer les valeurs manquantes de manière aléatoire, ni avec des valeurs proches ou par des moyennes afin de ne pas biaiser les données en apportant de la « fausse » information. Ce choix a été fait car la base de données reste conséquente et que le nombre de données manquantes ne représente que 3% de la base ce qui est très faible dans notre cas. Enfin, lors de la suppression des valeurs manquantes nous ne perdons pas de données sur les sinistres incendies graves.

Anciennes Catégories	Nouvelles Catégories
Copropriétaire non occupant	Non-occupant
Propriétaire non occupant	
Nu-propriétaire	
Locataire non occupant	
Copropriétaire occupant partiel	Propriétaire Occupant partiel
Propriétaire occupant partiel	
Locataire occupant partiel	Locataire Occupant Partiel
Colocataire occupant partiel	
Colocataire occupant unique	Locataire Occupant
Locataire occupant unique	
Usufruitier	
Copropriétaire occupant unique	Propriétaire Occupant
Propriétaire occupant total	

Tableau 3 : Exemple de regroupement pour la variable « Code Qualité »

6. Statistiques de la base de données.

Après retraitement total, la base de données est constituée de 1 871 446 lignes et 31 variables. Il y a 396 sinistres graves, soit une fréquence d'environ 0.021% ce qui est extrêmement faible puisque la fréquence d'un sinistre incendie est de 7.6% en 2018. Une première étude statistique sur les variables est nécessaire afin de vérifier si certaines variables catégorielles permettent de faire ressortir une fréquence de sinistre grave plus élevée.

Statistiques Variable Surface des dépendances	***1_ < 50m²***	***2_ 50-100***	***3_ 100-150***	***4_ 150-300***	***5_ 300-500***	***6_ 500 et +***
Répartition Totale (en %)	75,84	9,51	8,21	4,16	1,54	0,73

Répartition parmi les sinistres graves (en %)	59,81	11,48	15,31	8,13	2,63	2,63
Fréquence (en %)	0,0171	0,0261	0,0403	0,0422	0,0370	0,0775

Tableau 4 : Tableau regroupant les statistiques de la variables « Surface des dépendances ».

Le tableau ci-dessus présente les statistiques de la variable « Surface des dépendances ». On peut y trouver sur la première ligne la répartition de chaque classe de la variable. On peut remarquer que 9.5% de la base possède des dépendances ayant une surface comprise entre 50 et 100m². La seconde ligne permet d'observer la répartition des classes parmi les individus sinistrés. Cela permet d'observer si des catégories sont plus à risque que d'autres. Par exemple, les assurés sinistrés ayant des surfaces de dépendances supérieures à 500m² sont présent pour 2.63% alors que dans le portefeuille total ils ne représentent que 0.73%. Ce changement de proportion marque que cette catégorie peut être plus à risque que les autres. On le remarque aussi à l'aide de la 3e ligne, puisque la fréquence d'avoir un incendie grave de cette catégorie est plus élevée que les autres.

Lorsque l'on regarde les statistiques de toutes les variables, nous remarquons qu'une personne ayant eu un sinistre quelconque ou un dommage électrique dans les 5 ans aura une plus forte probabilité d'avoir un sinistre incendie dans l'année. Au contraire, une personne qui a eu un sinistre incendie l'année précédente aura une plus faible probabilité, ce qui semble être expliqué par le fait qu'après le sinistre, l'assuré sera plus attentionné sur les risques d'incendie ou aura procédé à des réparations ou améliorations. Nous observons aussi les cas suivants :

- Plus le taux de dégressivité[1] augmente, c'est-à-dire plus l'assuré n'a pas eu de sinistre dans les années antérieures, plus le risque de provoquer un incendie grave devient faible.
- Si l'individu déclare au moins un sinistre antérieur au moment de la souscription alors l'individu sera plus susceptible d'avoir un incendie grave l'année suivante.
- Plus l'assuré possède de pièces principales (plus de 9 pièces) plus le risque d'avoir un incendie grave dans l'année est fort.

Étude des variables en fonction des années :

Afin de vérifier la stabilité des variables dans le temps, il semblait important d'observer la répartition des classes année par année. Cette vérification permet de montrer si certaines variables peuvent influencer les résultats des modèles en ne faisant ressortir que les années atypiques de la base de données par rapport aux autres années. Les résultats suivants ressortent :

Années\Catégories	***1***	***2***	***3***	***4***

[1] Le taux de dégressivité est un taux d'abattement sur la franchise. Ce taux évolue dans le temps. Un nouveau client a un taux de dégressivité à 0. S'il n'a pas de sinistre dans l'année alors le taux augmente de 20 points. Au bout de 5 ans, si l'assuré n'a pas eu de sinistre, son taux de dégressivité sera de 100. Il n'aura alors pas de franchise à payer s'il devait déclarer un sinistre. Cependant, dès lors que l'assuré a un sinistre, son taux de dégressivité revient à 0.

2012	59,2011	39,4656	1,333292	0
2013	53,85063	43,13094	2,648735	0,369697
2014	47,86393	43,9307	4,506266	3,699107
2015	41,97011	42,53534	8,612149	6,882405
2016	36,53498	41,27271	12,594476	9,597835
2017	32,59004	41,04222	15,654242	10,7135
2018	29,37765	41,2038	17,99869	11,419865

Tableau 5 : Répartition des classes de la variable « Niveau de Franchise » en fonction des années.

On remarque que le niveau de franchise 4 est à 0 en 2012 et que le niveau 1 ne fait que diminuer dans le temps. Cela s'explique par le fait que le produit MRH ne comprenait que 3 niveaux de franchise et comme Néologis 1 a été lancé en 2012, la base ne peut pas avoir de niveau 4. C'est pourquoi on retrouve une répartition qui évolue dans le temps. Cela ne parait pas dérangeant pour la suite.
De même, nous remarquons que le taux de dégressivité présente un changement de répartition dans le temps. Le taux de dégressivité a été développé en 2012, c'est pourquoi on retrouve un taux d'environ 100% à 0. Au fur et à mesure, le taux évolue si l'assuré n'a pas eu de sinistre. La répartition évolue donc dans le temps.
Pour les autres variables rien de particulier n'est constaté.

Etude de la corrélation des variables :

Le graphique ci-dessous représente la corrélation entre les variables à l'aide du Rho de Spearman.

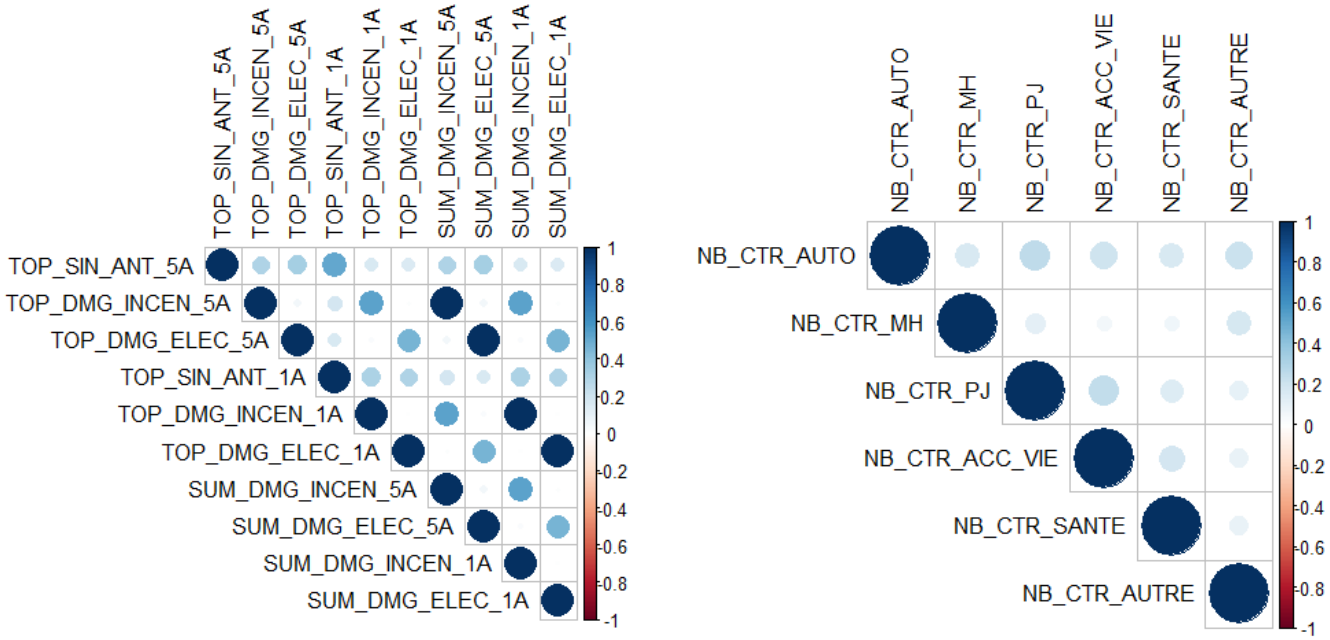


Figure 6: Représentation graphique de la corrélation entre les variables. Le graphique de gauche représente les variables sur les antécédents de sinistre. Le graphique de droite représente la corrélation des variables de l'équipement client.

Sur le graphique, plus les ronds sont gros et foncés plus la corrélation est forte entre les deux variables. De plus, si la couleur est bleue, alors les variables sont corrélées positivement. Au contraire, si la couleur est rouge, alors les deux variables sont corrélées négativement.

Cette matrice a été tracée pour toutes les variables de la base de données retraitées et après étude des corrélations, nous remarquons que certaines variables présentent des corrélations élevées. Comme par exemple le nombre de sinistres antérieurs. Le nombre sur 5 ans semble corrélé positivement au nombre sur 1 an. C'est-à-dire que si l'individu a eu au moins un sinistre l'année précédente alors en général il en a eu au moins autant, voire plus, sur 5 ans. De plus, les variables « top » sont très corrélées avec les variables « sum » et n'apportent pas autant d'informations que ces dernières. Aux vues du graphique, les variables de top vont être supprimées afin de ne garder que les variables « sum » afin d'éviter la redondance d'information.

De même, les nombres d'équipements semblent légèrement corrélés entre eux. Si l'assuré possède un contrat auto, il est possible qu'il ait aussi un contrat habitation, PJ, … Cependant, la corrélation ne semble pas assez élevée pour supprimer une des variables.

III. Modélisation et résultats

Dans cette partie, nous allons présenter le cadre théorique des différents modèles utilisés. Ensuite, les résultats seront comparés à l'aide d'indicateurs de performances permettant de juger le pouvoir prédictif. En effet, il est nécessaire de mettre en place des indicateurs de performances car ils permettent de voir si les modèles s'ajustent correctement aux données. Autrement dit, les indicateurs permettent de confronter les prédictions avec la réalité, et donc de voir si l'erreur est élevée ou non. De plus, nous chercherons les variables qui ont le plus d'importance et de significativité dans les différents modèles. Ces variables seront comparées aux statistiques de la partie II.6. Commençons par présenter les indicateurs de performance.

1. Critères de performance.

i. Matrice de confusion

Nous considérons que notre variable binaire à expliquer possède deux modalités : 0 ou 1, où le 0 correspond aux contrats n'ayant pas eu de sinistre incendie grave et le 1 correspond aux contrats ayant eu un sinistre incendie grave. On peut voir le 0 comme la classe des « négatifs » et les 1 comme la classe des « positifs ». Les prédictions sont sous forme de probabilité, il faut donc établir un seuil au-delà duquel on considère les individus comme « positif ».
A l'aide des prédictions et des observations nous pouvons construire le tableau suivant :

Prédictions/ Observations	**0**	**1**
0	Vrai Négatifs (VN)	Faux Négatifs (FN)
1	Faux Positifs (FP)	Vrais Positifs (VP)

Tableau 6 : Exemple de matrice de confusion

A partir de ce tableau, nous pouvons introduire des nouvelles notions :

- La sensibilité : c'est le taux de vrais positifs : $\text{sensibilité} = \frac{\text{VP}}{\text{VP} + \text{FN}}$;
- La spécificité : c'est le taux de vrais négatifs : $\text{spécificité} = \frac{\text{VN}}{\text{VN} + \text{FP}}$;
- Le taux de succès : c'est la probabilité de bon classement : $\text{taux de succès} = \frac{\text{VN} + \text{VP}}{\text{VN} + \text{FP} + \text{FN} + \text{VP}}$;
- Le taux d'erreur : c'est la probabilité d'avoir un mauvais classement : $\text{taux d'erreur} = \frac{\text{FN} + \text{FP}}{\text{VN} + \text{FP} + \text{FN} + \text{VP}} = 1 - \text{taux de succès}$.

Mais ce sont principalement la sensibilité ainsi que la spécificité que nous utiliserons dans ce mémoire car un modèle ayant un fort pouvoir prédictif doit avoir des valeurs élevées dans ces deux valeurs.

ii. Table de classification :

Nous considérons une table qui, parmi les individus triés dans l'ordre décroissant de leur score, détermine le nombre d'individus correctement classé. Ce tableau permet d'ajuster le seuil de probabilité qui déterminera ensuite si un individu est classé positif ou négatif. Le seuil est choisi afin d'avoir une sensibilité et une spécificité élevée. On remarque à l'aide de ce tableau le pouvoir prédictif dans un intervalle. Cela permet de s'arrêter à un seuil lorsque le modèle ne détecte plus assez d'individus positifs. Le tableau est construit comme suit :

	intervalle 1	nombre détecté	sensibilité	intervalle 2	nombre détecté	proportion trouvée intervalle
1	[0 : 5000]	5	6.329114	[0 : 5000]	5	6.329114
2	[0 : 10000]	8	10.126582	[5000 : 10000]	3	3.797468
3	[0 : 15000]	13	16.455696	[10000 : 15000]	5	6.329114
4	[0 : 20000]	13	16.455696	[15000 : 20000]	0	0.000000
5	[0 : 25000]	15	18.987342	[20000 : 25000]	2	2.531646
6	[0 : 30000]	20	25.316456	[25000 : 30000]	5	6.329114
7	[0 : 35000]	21	26.582278	[30000 : 35000]	1	1.265823

Figure 7: Exemple de table de classification utilisée dans ce mémoire

La première colonne répertorie le nombre d'individus observés (entre 0 et n). La deuxième colonne indique le nombre de sinistres graves détectés dans l'intervalle [0,n], la sensibilité associée au nombre de positifs détectés est dans la troisième colonne. Ensuite dans la partie droite du tableau est détaillée par intervalle de 5 000 individus, la quantité de sinistres graves détectés. Cela permet d'identifier plus précisément à quel niveau les sinistres sont découverts. Dans le tableau précédent, on remarque que dans l'intervalle [15 000, 20 000] le modèle ne détecte aucun sinistre grave, il est peut-être intéressant que de ne regarder que dans l'intervalle [0,15 000] et alors de garder comme seuil la probabilité du 15 000^{e} score.

iii. Courbes ROC et AUC

La courbe ROC (*Receiver Operating Characteristic*) est la courbe qui décrit l'évolution du taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (1-spécificité). A l'aide de cette courbe, nous pouvons calculer l'aire présente en dessous de la courbe, l'AUC (*Area Under the Curve*). Un modèle possède un fort pouvoir prédictif lorsque l'AUC est proche de 1. Cependant, si l'AUC est égal à 1, il est fort probable que le modèle est en sur-apprentissage ou qu'il y ait un problème. Au contraire, si l'AUC est en dessous de 0.5 alors cela signifie que le modèle est moins bon qu'une pièce de monnaie, ce qui est très mauvais.

2. Rééchantillonnage

i. Sous et Sur échantillonnage

La différence de proportion entre les individus positifs et les individus négatifs étant très élevée il est parfois intéressant de rééquilibrer les classes afin de permettre aux modèles de mieux repérer les individus positifs. Dans notre cas, le rééchantillonnage va modifier la fréquence d'apparition des sinistres incendie graves. Pour ce faire, deux solutions existent : soit créer artificiellement des individus de la classe la moins représentée à l'aide des lignes déjà existantes (sur-échantillonnage ou *over-sampling*), soit enlever un nombre défini de lignes tirées aléatoirement des individus de la classe sur-représentée (sous-échantillonnage ou *under-sampling*).

Dans notre étude, le nombre de cas positifs étant très faible, le fait d'en créer de nouveaux pourrait créer trop de biais dû à la carence d'information de ce groupe et le fait de créer aléatoirement des sinistres graves pourrait impliquer de changer le profil de risque des sinistrés, or c'est cette information qui nous intéresse le plus dans cette étude. De plus, nous avons beaucoup de lignes négatives, le fait d'en enlever une certaine proportion ne biaiserait pas le modèle ou ne modifierait que légèrement le profil des non-sinistrés, c'est pourquoi nous allons utiliser cette deuxième approche et tester différents pourcentages de rééchantillonnage afin de trouver le plus performant.

ii. Validation Croisée

La validation croisée est une méthode ensembliste permettant d'améliorer un modèle en séparant la base de données en plusieurs bases évitant ainsi le surapprentissage. Cette méthode se déroule comme suit :

- Partitionner ses données en k échantillons de taille homogène k/n, avec n le nombre de lignes totales de la base. Il se peut que le dernier échantillon ne soit pas de la même taille que les autres si n n'est pas un multiple de k ;
- Prendre k-1 échantillons pour l'apprentissage du modèle et le dernier servira de test ;
- Entrainer le modèle sur la base d'apprentissage et effectuer une prédiction sur la base de test ;
- Calculer l'erreur entre la prédiction et la réalité ;
- Répéter l'opération k fois (correspondant aux k classes) ;
- Faire la moyenne des erreurs de tous les modèles.

Deux exemples particuliers de la validation croisée :

- La méthode « *leave-one-out* ». Si l'on a un échantillon de 10 000 individus, alors on crée 10 000 modèles composés de 9 999 individus, ce qui correspond au cas k=(n-1). C'est une méthode très coûteuse en temps puisqu'il faut créer (n-1) modèles. Dans notre cas, ce serait extrêmement long puisque la base de données est conséquente.
- La méthode « *Testset Validation* » : correspond au cas où k=2, avec un pourcentage de 60% pour la base de modèle et 40% pour le test. C'est l'autre cas extrême de la validation croisée.

Il faut trouver le bon compromis entre les deux méthodes afin d'avoir des résultats améliorés mais un temps d'exécution convenable.

Posons maintenant le cadre théorique des modèles.
Soit X la matrice contenant toutes les caractéristiques des individus. Les caractéristiques sont aléatoires et supposées indépendantes. Soit m_i le montant brut de recours du sinistre i. On pose la variable binaire y_i qui nous donne l'indication sur la gravité du sinistre :

$$y_i = \begin{cases} 0, & m_i < 30\,000 \ € \\ 1, & m_i \geq 30\,000 \ € \end{cases}$$

On construit le vecteur M composé des montants bruts de recours de tous les m_i et Y le vecteur contenant tous les y_i. Y est une variable Binomiale.
Le but des modèles statistiques est de trouver une fonction f(X) qui permet de prédire Y grâce aux caractéristiques des assurés. On pose :

$$Y = f(X) + \epsilon$$

où f est la fonction qui permet de prédire la variable Y par rapport aux caractéristiques déterministes et ϵ est l'erreur associé au modèle. On a $\hat{Y} - Y = \epsilon$, où $\hat{Y} = f(X)$ la prédiction de Y.
La solution du problème est donc de trouver la fonction f définie par :

$$f(x) = E[Y|X = x]$$

Cette fonction représente l'espérance conditionnelle de Y sachant toutes les caractéristiques. Autrement dit, la meilleure approximation de Y en tout point $X = x$ est la moyenne conditionnelle.

3. Modèle Logit

Aux vues de la très faible quantité de sinistres incendie grave, il semble difficile d'avoir un modèle Logit performant. Cependant, nous souhaitons tout de même le mettre en place afin de pouvoir le comparer à l'efficacité des autres modèles.

i. Théorie

Le modèle Logit est un cas particulier des modèles linéaires généralisés (GLM). Il est donc important de rappeler les bases des GLM. Ce sont des modèles relativement anciens puisque l'algorithme a été conçu en 1972 par Nelder et Wedderburn. Les travaux sont exposés dans Nelder et Mc Cullagh (1983), Agresti (1990) ou Antoniadias et al. (1992). Ces modèles sont toujours très utilisés de nos jours par les compagnies d'assurance pour construire les modèles de tarification. Les modèles linéaires généralisés font partis des modèles paramétriques.

Les modèles linéaires généralisés sont de la forme :

$$f(X) = \beta_0 + \sum_{j=1}^{n} X_j \beta_j$$

Ils présentent 3 composantes :

- **Une composante aléatoire :** cette composante définit la distribution de Y dont la fonction de densité s'écrit :

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta}{u(\phi)} + \omega(y, \phi)\right)$$

où θ est le paramètre naturel et ϕ le paramètre de dispersion. On peut dire que Y fait partie de la famille exponentielle ;

- **Composante déterministe :**

$$\eta = \beta_0 + \sum_{j=1}^{n} X_j \beta_j,$$

Avec $\beta = (\beta_0, \ldots, \beta_p)$ représente le vecteur des coefficients à estimer et β_0 est une constante ;

- **Une fonction lien** : cette fonction, déterministe, définie sur iR et strictement monotone permet de relier l'espérance conditionnelle de Y à la composante déterministe :

$$f(E[Y|X = x]) = \eta = \beta_0 + \sum_{j=1}^{n} X_j \beta_j.$$

Dans notre cas, Y ne prend que deux valeurs : 0 ou 1. On décide de poser :

$$p_x = E[Y|X = x] = \mathbb{P}(Y = 1 \mid X = x) * 1 + \mathbb{P}(Y = 0 \mid X = x) * 0 = \mathbb{P}(Y = 1 \mid X = x).$$

La fonction de lien définie précédemment permet pour notre étude d'effectuer une classification à l'aide d'une régression. En effet, la régression logistique avec la fonction de lien Logit permet à l'aide de résultat de régression de calculer des probabilités d'appartenance à une classe. On obtient :

$$f(E[Y|X = x]) = f(p_x)$$

Définition 4 : la fonction Logit :
Soit x appartenant à]0,1[alors :

$$\text{Logit}(x) = \ln\left(\frac{x}{1-x}\right)$$

Et estimé en p_x :

$$\text{Logit}(p_x) = \ln\left(\frac{p_x}{1-p_x}\right) = X\beta = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$$

Ou encore :

$$p_x = \frac{e^{\beta_0 + \sum_{j=1}^{n} \beta_j x_j}}{1 - e^{\beta_0 + \sum_{j=1}^{n} \beta_j x_j}}$$

Avec β_0 la constante et $\beta_1, \ldots, \beta_p$ les coefficients associés aux variables du modèle

Avec la définition 4, nous obtenons le rapport $\frac{p_x}{(1-p_x)}$ qui peut être interprété comme le risque d'avoir un sinistre grave pour une classe par rapport à l'autre classe.

Le vecteur des paramètres est estimé par maximisation de la log-vraisemblance à l'aide de la méthode de Newton-Raphson par exemple.

$$\text{Ln}(\beta) = \prod_{i=1}^{n} p_{x_i}(\beta)^{y_i} \left(1 - p_{x_i}(\beta)\right)^{(1-y_i)}$$

On obtient le vecteur $\hat{\beta}$ qui contient tous les $\hat{\beta}_i$, on obtient alors l'estimation des probabilités par :

$$\widehat{p_x} = \frac{e^{\hat{\beta}_0 + \sum_{j=1}^{n} \hat{\beta}_j x^j}}{1 - e^{\hat{\beta}_0 + \sum_{j=1}^{n} \hat{\beta}_j x^j}}$$

ii. Implémentation sous R

Nous avons implémenté sous R le modèle Logit à l'aide de la fonction *glm* de la librairie *stats*. La formule se code sur R comme suit :

```
modele_logit <-glm(formula = TOP_GRAVE ~ .,data = base_entrainement ,family=binomial(link="logit"))
```

Les paramètres utilisés de la fonction sont :

- *formula* : ce paramètre permet de sélectionner les variables que l'on souhaite ajouter au modèle. Si l'on met un « . » cela signifie que l'on prend toutes les variables de la base de données ;
- *data* : définit la base de données à utiliser ;

- *family* : Permet de sélectionner la famille qui décrit l'erreur de distribution du modèle avec la fonction de lien associé. Dans notre cas, c'est un modèle binomial avec la fonction de lien Logit ;
- *weights* : ce paramètre permet d'affecter des poids sur les variables afin de donner des priorités de sélection. Nous ne l'utilisons pas dans notre cas.
- *Offset* : permet d'insérer une variable offset dans le modèle. Pour les modèles Logit, il n'y en a pas.

4. Modèle d'arbre de décision (CART)

i. Théorie

La méthode d'arbre de décision (*Classification and Regression Tree* , CART) a été inventée en 1984 par les statisticiens L.Breiman, J.H.Friedman, R.A.Olshen et C.J.Stone (Universités de Berkeley et de Stanford).
L'arbre de décision est une suite de décisions prises de manière séquentielle qui permet de diviser la base de données en plusieurs groupes. Le moment, où la décision est prise, est appelé un nœud. Le nœud fils est le nœud qui suit la décision actuelle. La base de données initiale est le tronc de l'arbre. La séparation des données qui permet d'obtenir l'arbre maximal est faite de la manière suivante :

- Une variable est sélectionnée afin de discriminer au maximum les données et d'obtenir deux sous classes qui maximisent le gain en information. On calcule le gain en information à l'aide de l'indice de Gini. L'indice de Gini est défini par la fonction :

$$\mathrm{i(n)} = 1 - \sum_{\mathrm{j=1}}^{2} \mathrm{P^2(j|n)}$$

avec $\mathrm{P(j|n)}$ la proportion d'individus dans la classe j au nœud n.

Et la fonction de gain en information est définie par :

$$\mathrm{Gain(n,x)} = \mathrm{i(n)} - \sum_{\mathrm{j=1}}^{2} \mathrm{P_j} * \mathrm{i(n_j)}$$

Avec n le nœud où on calcule le gain et n_j le nœud fils ; x la variable explicative ; P_j la proportion d'individus au nœud n qui vont dans le nœud fils ; $i(n)$ la fonction de l'indice de Gini pour mesurer le degré de désordre.

- Une fois que l'on a choisi la meilleure variable, on effectue la séparation des données et on recommence sur les deux sous classes établies.

- On recommence ce procédé plusieurs fois. On s'arrête lorsqu'il ne reste qu'un individu dans la classe ou que tous les individus du groupe sont homogènes.

Lorsque l'arbre maximal est construit, le modèle est en sur-apprentissage. C'est-à-dire que le modèle connait par cœur chaque individu mais lorsqu'on lui donne des nouvelles données, il

ne saura pas nécessairement les classer dans la bonne catégorie. Pour éviter ce problème, il faut élaguer l'arbre maximal. En d'autres termes, il faut couper l'arbre avant la fin permettant de gagner en prédiction même si l'on perd en précision.

ii. Implémentation sous R

Nous avons implémenté sous R les arbres de décisions à l'aide de la fonction *tree* de la librairie *tree*. La formule se code sous R comme suit :

```
modele_CART<-tree(formula = TOP_GRAVE~., data = data_train ,split = "gini",mincut = 10000)
```

Les paramètres utilisés de la fonction sont :

- *formula* : ce paramètre permet de sélectionner les variables que l'on souhaite ajouter au modèle. Si l'on met un « . » cela signifie que l'on prend toutes les variables de la base de données ;
- *data* : définit la base de données à utiliser ;
- *weights* : ce paramètre permet d'affecter des poids sur les variables afin de donner des priorités de sélection. Nous ne l'utilisons pas dans notre cas.
- *split* : définit le critère de séparation. Il y a le choix entre le critère de Gini (usilisé dans notre cas) ou la déviance ;
- *nobs :* définit le nombre d'observations à utiliser lors de la phase d'entrainement ;
- *mincut :* définit le nombre minimum d'observations à inclure dans chaque nœud fils ;
- *minsize :* définit la taille du plus petit nœud.

5. Modèle de Descente de Gradient.

i. Théorie

Le principe de cet algorithme est d'améliorer un modèle en construisant de nouveaux arbres de décision, à l'aide des erreurs du modèle précédent qui permettent de pondérer les observations. Plus l'erreur est élevée et plus l'observation sera pondérée.
Nous avons toujours notre variable Y à deux catégories (0 ou 1), la matrice X avec les caractéristiques des individus.

Soit la fonction de coût :

$$j(y_i, f(x_i)) = \sum_{k=1}^{2} \ln(p^k) 1_{\{Y_i=k\}})$$

Où $1_{\{Y_i=k\}}$ est la fonction indicatrice qui vaut 1 lorsque $Y_i = k$, 0 sinon ; p^k correspond à la probabilité conditionnelle d'appartenir à la modalité k.

Le gradient de cette fonction est égal à :

$$\nabla j(y_i, f(x_i)) = 1_{\{Y_i=k\}} - p^k$$

En d'autres termes, le gradient de la fonction j correspond à la différence entre 1 (ou 0) et la probabilité d'appartenir à la classe.

L'algorithme de *boosting* est le suivant :

- Créer un modèle initial ;
- Calculer le gradient de la fonction de coût et prendre le signe opposé ;
- Construire un nouvel arbre à l'aide des coûts calculés à l'étape précédente ;
- Recommencer jusqu'à ce qu'on atteigne le minimum de la fonction de coût global.

Le modèle est amélioré en construisant des arbres de décisions sur les erreurs des modèles qui le précède afin d'obtenir le modèle optimal.

ii. Implémentation sous R

Nous avons implémenté sous R le modèle Gradient Boosting à l'aide de la function *xgb.cv* de la librairie *xgboost*. Cette fonction permet d'effectuer une validation croisée permettant de construire le modèle de Gradient Boosting le plus efficace. La formule se code sur R comme suit :

```
xbg_cross_val <- xgb.cv(params = parameters, data = Ddata_train, nrounds = 600, nfold = 4, showsd = T,
                        print_every_n = 10, early_stopping_rounds = 40, maximize = T, metrics = "auc")
```

Les paramètres utilisés de la fonction sont :

- *params* : ce paramètre est à entrer sous forme de liste. Il permet d'initialiser l'objectif (si c'est une régression ou une classification), la taille de chaque échantillon de *boosting*, la profondeur maximale de l'arbre ou encore le nombre de branches utilisées ;
- *data* : définit la base de données à utiliser. Elle est sous forme de *Dmatrix,* et les variables explicatives et la variable à expliquer sont déjà scindées, c'est pourquoi il n'y a pas de formule dans les paramètres ;

- *nrounds :* le nombre maximum d'itération que l'on souhaite opérer ;
- *nfold :* le nombre de classe de sous-échantillons afin d'effectuer la validation croisée ;
- *showsd :* permet d'afficher en sortie l'écart type induite par la validation croisée ;
- *print_every_n :* affiche toutes les n itérations les résultats de l'entrainement ;
- *early_stopping_rounds :* à entrer sous forme d'un entier k. Ce paramètre permet de stopper les itérations si le modèle ne s'est pas amélioré au bout k itérations ;
- *maximize :* garde en mémoire la meilleure itération ;
- *metrics :* choix de l'objectif à maximiser. Dans notre cas, nous avons utilisé l'AUC.

6. Résultats des modélisations

Dans cette partie, les différentes étapes de modélisations seront exposées.

Partitionnement de la base Complète :

La base est découpée en 3 bases : la base d'entrainement, la base de test et la base de validation. La base d'entrainement permet de faire apprendre les modèles, c'est-à-dire que le modèle va créer ses critères de prédiction sur cette base. Ensuite, la base de test permet de tester la performance du modèle et d'ajuster les paramètres (si le modèle possède des paramètres) sur des nouvelles données afin d'obtenir les meilleurs résultats. Enfin, la base de validation permet de tester le modèle optimal sur des données qui n'ont jamais été prises en compte dans la modélisation, ce qui permet de valider les résultats et de voir si le modèle n'est pas en surapprentissage.

i. Résultats initiaux

La première modélisation a été faite à l'aide de la base d'entrainement complète, c'est-à-dire avec toutes les variables et toutes les lignes. Ces premiers modèles serviront d'initialisation et l'objectif sera d'améliorer les résultats à chaque étape.

Les premiers modèles élaborés ne semblent pas très performants. En effet, le modèle logistique présente une *AUC* à 0.666, à 0.623 pour le modèle CART et à 0.673 pour le Gradient Boosting. Si l'on compare les *AUC,* il semblerait que le modèle du Gradient Boosting soit le plus satisfaisant.

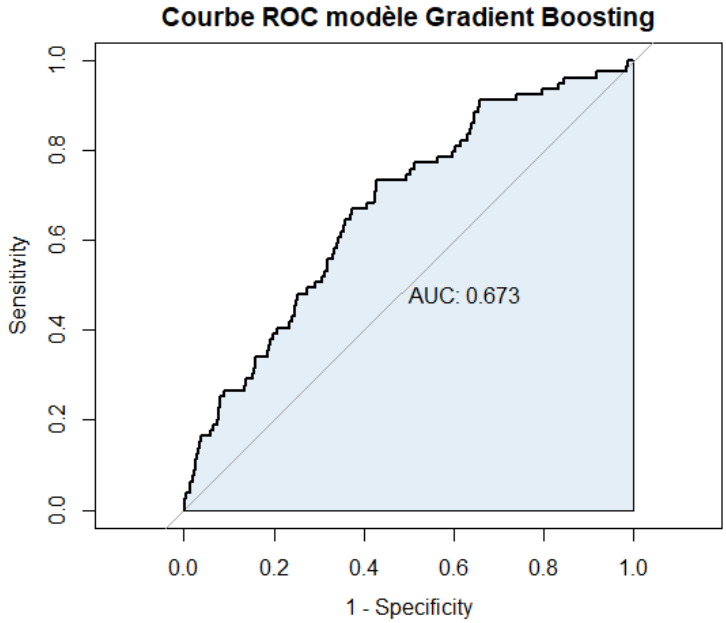


Figure 8 : Courbe ROC du modèle Gradient Boosting.

Maintenant, si nous nous concentrons sur la détection des sinistres incendie grave parmi les individus ayant le score le plus élevé, nous obtenons aussi que le modèle Gradient Boosting est le plus efficace. En effet, si l'on regarde les 5 000 contrats les plus scorés, nous trouvons 5 sinistres incendie grave pour le Gradient Boosting et pour les arbres de décisions contre 4 pour le modèle Logit. Si nous regardons sur les 30 000 contrats les plus scorés, nous obtenons 20 sinistres graves découverts pour le Gradient Boosting, 17 pour le modèle Logit et 14 pour le modèle CART. Le tableau de classification présente les résultats du Gradient Boosting.

	intervalle 1	nombre détecté	sensibilité	intervalle 2	nombre détecté	proportion trouvée	intervalle
1	[0 : 5000]	5	6.329114	[0 : 5000]	5		6.329114
2	[0 : 10000]	8	10.126582	[5000 : 10000]	3		3.797468
3	[0 : 15000]	13	16.455696	[10000 : 15000]	5		6.329114
4	[0 : 20000]	13	16.455696	[15000 : 20000]	0		0.000000
5	[0 : 25000]	15	18.987342	[20000 : 25000]	2		2.531646
6	[0 : 30000]	20	25.316456	[25000 : 30000]	5		6.329114
7	[0 : 35000]	21	26.582278	[30000 : 35000]	1		1.265823

Figure 9 : Nombre de sinistre découvert en fonction des contrats les plus scorés.

Si l'on place notre seuil de prédiction au niveau du score du 30 000e individu, nous pouvons alors tracer la matrice de confusion suivante pour le modèle de Gradient Boosting :

Prédictions/ Observations	**0**	**1**
0	344 289	59
1	29 980	20

Tableau 7 : Matrice de confusion du modèle Gradient Boosting.

D'après la matrice précédente, nous obtenons une sensibilité de 25.32% et une spécificité de 92%. Le taux de succès est à 91.9%. Même si la spécificité et le taux de succès semblent relativement bons, la sensibilité est mauvaise. De plus, le modèle considère 29 980 individus en faux positifs. Ce nombre est beaucoup trop élevé pour le si peu de vrais sinistres graves détectés. Le but est de trouver le meilleur compromis entre la sensibilité et la spécificité afin d'avoir le plus de sinistres graves détectés avec le moins de faux positifs possible. La première approche n'est pas concluante.

ii. Résultats des sous-échantillonnages

Nous allons maintenant essayer d'améliorer les résultats précédents à l'aide d'un sous échantillonnage. Afin de trouver la meilleure solution, nous avons testé plusieurs sous-échantillonnages sur les 3 modèles . Nous avons opéré de la façon suivante :

Pour chaque sous-échantillonnage de 10% à 90% :

Effectuer 10 fois les étapes suivantes afin de tester la variabilité du modèle :

1- Apprentissage du modèle avec la base d'entraînement sous-échantillonnée ;
2- Effectuer les prédictions sur la base de test ;
3- Calculer l'AUC.

Les résultats de cette procédure sont en annexes C.

On remarque que quel que soit le modèle utilisé et pour chaque sous-échantillonnage, la variance de l'AUC reste très faible. Le modèle Logit ne semble pas être impacté par le sous-échantillonnage. En effet, on remarque que pour n'importe quel pourcentage de rééchantillonnage, l'AUC moyen sur chaque test reste le même (il faut aller jusqu'à la 4e décimale pour comparer les valeurs). Pour la suite de l'étude, nous allons garder le modèle Logit avec la base complète afin de garder un maximum d'informations. Au contraire, les modèles d'arbres de décisions et de Gradient Boosting semblent eux impactés par ce rééchantillonnage. Pour le Gradient Boosting, le modèle qui ressort comme le plus performant est celui où les négatifs sont supprimés à 60% et 50% pour les arbres de décisions. Nous choisirons donc de continuer avec ces deux modèles rééchantillonnés.

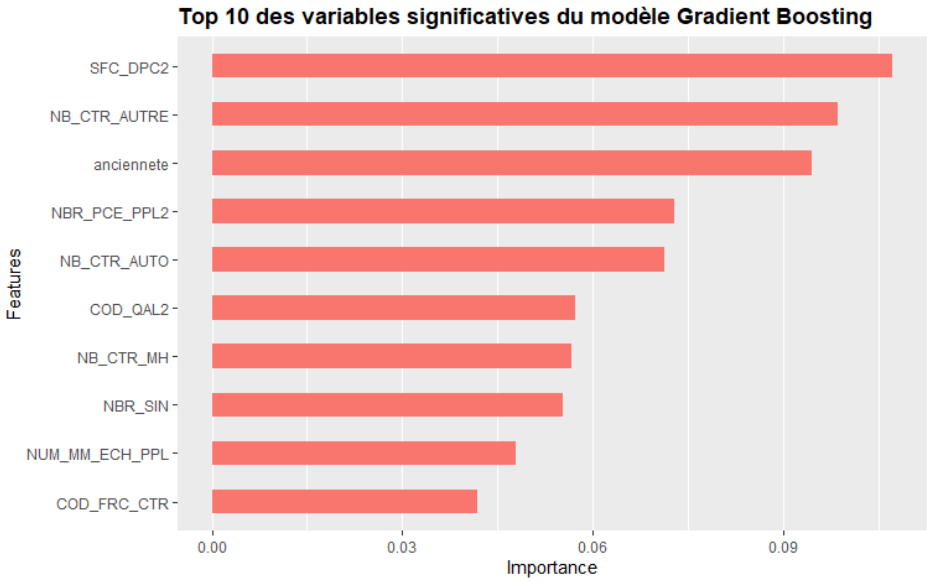


Figure 10: Représentation des variables les plus importantes pour le modèle XGBoost avec un sous-échantillonnage à 60%

Lorsque l'on analyse plus en profondeur ces 3 modèles, on remarque que certaines variables ressortent plus que d'autres. En effet, lorsque l'on observe l'estimation des coefficients des variables du modèle Logit, on remarque que les variables les plus influentes dans le modèle sont la surface de dépendance, le nombre de pièces principales ou encore les variables qui permettent de voir l'équipement de l'assuré. Ces variables induisent positivement la probabilité d'avoir un sinistre incendie grave. Au contraire, la variable du taux de dégressivité joue un rôle inverse sur la probabilité d'avoir un sinistre grave, cela permet de distinguer les individus ayant moins de risque d'avoir un sinistre incendie grave.

D'après la figure 7 représentant l'importance des variables du modèle XGBoost avec un sous-échantillonnage à 60%, nous remarquons que les variables les plus influentes sont la surface des dépendances, les équipements clients, ou encore le nombre de pièces principales. Enfin pour le modèle d'arbre de décisions, seulement la moitié des variables disponibles ont été utilisées. Parmi celles-ci on retrouve entre autres la surface des dépendances et le nombre de pièces principales.

D'après les modèles étudiés, nous retrouvons les mêmes variables importantes permettant d'identifier les individus à risque. Ce recoupement entre les modèles permet de penser que ce sont ces variables qui influent fortement sur la probabilité d'avoir un sinistre incendie grave et qu'il faut chercher dans quelles proportions elles agissent. On aurait pu penser que les variables sur la sinistralité antérieure auraient un rôle plus important dans les modèles, soit en indiquant que plus l'individu a eu de sinistres dans le passé alors plus sa probabilité d'en avoir l'année qui suit sera élevée. Mais nous retrouvons surtout des indications sur le logement ou sur l'équipement du client.

Les logements possédant de grandes dépendances possèdent un fort risque d'avoir un sinistre incendie grave. Cela peut s'expliquer par le fait que cela demande de l'entretien régulier et que les clients n'ont pas forcément les moyens financiers pour tout gérer ou qu'ils sont négligeant vis-à-vis de leurs résidences. Le fait qu'ils soient suréquipés peut nous indiquer que ces clients sont averses au risque puisque le fait de souscrire une assurance (en dehors des obligations légales) est généré par la conscience d'un risque, l'objectif étant de vouloir s'en prémunir. Ce suréquipement peut aussi induire un certain niveau de moyens financiers qui permet à l'assuré d'entretenir ses dépendances. Le nombre de pièces principales peut s'expliquer de la même manière que la variable de la surface des dépendances. En effet, si le logement possède un nombre élevé de pièces principales, cela peut impliquer une surface de logement élevé, ce qui demande aussi beaucoup d'entretien. Si l'assuré est négligent ou n'a pas assez de moyens financiers pour entretenir cette surface, le risque d'avoir un incendie grave sera plus fort.

Maintenant que l'on a observé ces premiers résultats nous allons les analyser d'un point de vue économique. Si on souhaite résilier tous les contrats qui présentent une probabilité de risque élevée (au-delà d'un certain seuil), nous pouvons tracer une courbe qui représente la différence entre la perte des cotisations du client s'il n'est plus dans le portefeuille et le gain du (ou des) sinistre(s) évité(s).
La fonction est construite comme suit :

$$\text{Gain(n)} = \text{Montant sinistres détectés(n)} - \text{Somme cotisations clients résiliés(n)}$$

Où n est le nombre de contrats observés (triés par ordre décroissant du score).

L'objectif de cette fonction est de trouver le seuil qui permet d'avoir un gain positif le plus élevé, c'est-à-dire que le modèle permet d'éviter plus de sinistre que ne fait perdre de cotisations. Par exemple, sur le modèle Logit, si l'on regarde les 5 000 contrats ayant le plus gros score (du plus risqué au moins risqué), nous obtenons le graphique suivant :

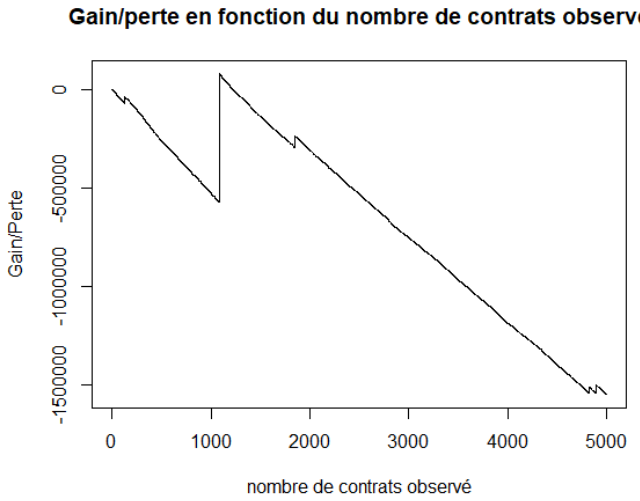


Figure 11: Représentation graphique de la fonction de perte du modèle Logit.

On remarque que sur les 5 000 contrats ayant le score le plus élevé, le modèle ne détecte que 5 sinistres dont un, avec un montant important aux alentours du 1000e score, qui fait passer la fonction de gain en positif. Cependant, si l'on se place du point de vue de l'assureur, c'est extrêmement risqué de résilier 1 000 contrats afin de ne détecter que quelques sinistres. De plus, cette fonction ne prend en compte que la cotisation client du contrat Multirisque Habitation, il se peut que l'assuré ou même le foyer (femmes, enfant, …) ait d'autres contrats au sein de l'assurance. Si l'assureur résilie le contrat habitation, le client partira avec tous les autres contrats, ce qui engendrera une perte plus importante pour l'assureur. Il serait donc intéressant que le modèle détecte plus de sinistres dans un intervalle plus petit, ce qui permettrait d'avoir un gain plus élevé.

iii. Approfondissement du modèle Logit

Dans cette partie, nous allons améliorer le modèle Logit afin d'affiner les coefficients du modèle et ainsi extraire des pondérations en fonction des variables. Le modèle Logit semble plus exploitable que les deux autres modèles car il permet de faire ressortir les coefficients de toutes les variables, permettant ainsi de voir le pouvoir prédictif de cette dernière. En effet, si le coefficient de la variable est positif alors cela engendrera une probabilité plus élevée d'avoir un sinistre incendie grave (au contraire pour un coefficient négatif).
Afin d'améliorer le modèle Logit et de sélectionner au mieux les variables du modèle, plusieurs méthodes existent : la méthode ascendante, la méthode descendante et la méthode mixte.

Avant d'expliciter les différentes méthodes, il est nécessaire d'introduire la notion de d'AIC (*Akaike Information Criterion*).
La valeur de l'AIC se calcule comme suit :

$$\mathrm{AIC} = -2\ln(\mathrm{L}) + 2\mathrm{p}$$

Avec L la vraisemblance du modèle et p le nombre de paramètres du modèle.
Ce critère permet de pénaliser les modèles en fonction du nombre de paramètres. Plus il y aura de variables non significatives dans le modèle, plus l'AIC sera élevé. L'objectif étant d'avoir un modèle avec l'AIC le plus faible, c'est-à-dire avec des variables les plus significatives.

La **méthode ascendante** consiste à construire une régression qui ne comporte qu'une seule variable (en théorie celle qui semble la plus significative). Ensuite, le but est d'ajouter pas à pas les autres variables au modèle. La variable qui est sélectionnée est la variable qui ajoute l'information statistique la plus significative (cet ajout d'information peut être calculée à l'aide de l'AIC). La variable sélectionnée est celle qui fera diminuer le plus l'AIC. Lorsque la variable est choisie, le processus est effectué avec les variables restantes. La procédure s'arrête lorsque plus aucune variable ne fait diminuer l'AIC et le modèle final sera celui avec le plus faible AIC et donc avec les variables les plus significatives au sens de l'AIC.

La **méthode descendante** est en quelque sorte l'inverse de la méthode ascendante puisque le point de départ est le modèle complet (avec toutes les variables) et on enlève pas à pas les variables qui sont non-significatives. Comme la méthode ascendante, on peut se servir de l'AIC comme critère. Le modèle final est celui avec l'AIC le plus faible.

La **méthode mixte**, comme son nom l'indique, va compiler les deux méthodes précédentes. A chaque étape, le modèle est testé avec l'ajout ou la suppression d'une variable. Le modèle final est le modèle le plus significatif au sens de l'AIC.

Pour appliquer cette amélioration, nous utilisons la fonction *step* de la librairie *stats* de R qui permet d'exécuter les méthodes précédentes. D'après la section précédente, il semblerait que ce soit la surface des dépendances qui soit la variable la plus significative dans les modèles. Nous débuterons l'algorithme avec cette variable. Lorsque la procédure est terminée, nous obtenons un modèle avec 7 variables : la surface des dépendances, le nombre de pièces principales, le nombre de sinistre déclaré à la souscription, l'ancienneté du contrat, le taux de dégressivité, le type du bien immobilier et le nombre de contrat « autre »[2]. Le modèle présente une AUC à 0.658 ce qui est moins bon que le modèle initial en terme d'AUC mais meilleur en terme d'AIC puisque l'AIC initial était à 5190 et pour le nouveau modèle l'AIC est à 5145.

Nous obtenons la courbe de gain suivante :

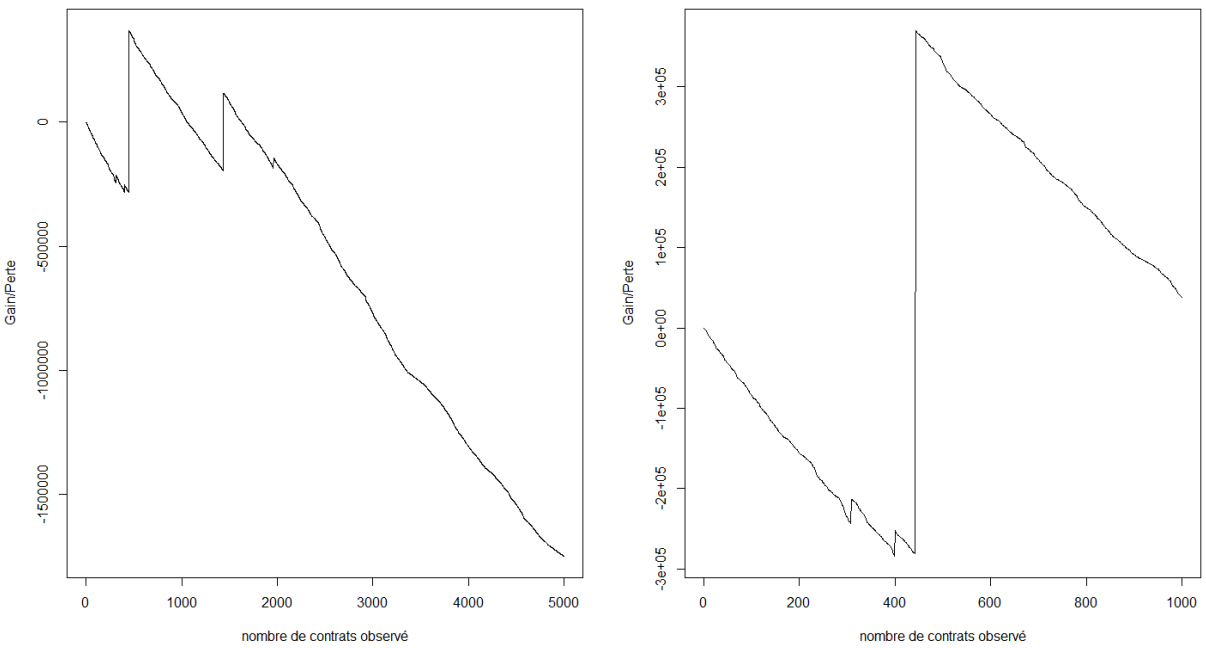


Figure 12 : fonction de gain pour les 5 000 plus scorés et zoom sur les 500 premiers

[2] Nous considérons dans « contrat autre » les produit qui ne sont pas automobile, multirisque habitation, protection juridique, accident vie privée ou encore santé.

Sur ce modèle amélioré, nous remarquons que dans l'intervalle [0 , 5 000], le modèle détecte 5 sinistres mais que 3 d'entre eux se situent dans les 500 premiers scores. On remarque que la courbe de gain atteint un gain assez élevé dû à un sinistre d'un montant de 652 200€. Cependant, si le montant de sinistre avait été plus faible, le gain n'aurait pas été aussi élevé.
Si nous regardons le tableau de classification, nous remarquons que ce modèle est particulièrement efficace dans les premiers contrats observés puis décroît vite, c'est pour cela que l'AUC est plus faible.

intervalle 1	nombre détecté	sensibilité	intervalle 2	nombre détecté	proportion trouvée intervalle
[0 : 5000]	5	6.329114	[0 : 5000]	5	6.329114
[0 : 10000]	6	7.594937	[5000 : 10000]	1	1.265823
[0 : 15000]	8	10.126582	[10000 : 15000]	2	2.531646
[0 : 20000]	10	12.658228	[15000 : 20000]	2	2.531646
[0 : 25000]	10	12.658228	[20000 : 25000]	0	0.000000
[0 : 30000]	14	17.721519	[25000 : 30000]	4	5.063291
[0 : 35000]	16	20.253165	[30000 : 35000]	2	2.531646

Figure 13: Table de classification du modèle Logit amélioré.

A l'aide du tableau précédent, nous positionnons le seuil de probabilité à 0,0005483481, qui correspond à la probabilité du 15 000^{e} individu. Ce seuil est choisi car il permet d'avoir une spécificité à 96% et une sensibilité à 10%.
Lorsque les prédictions sont effectuées sur l'ensemble de validation, nous obtenons la matrice de confusion suivante :

Prédictions / Observations	**0**	**1**
0	179757	38
1	7348	2

Tableau 8 : Matrice de confusion du modèle amélioré.

A la lecture de cette matrice de confusion, nous avons une sensibilité de 5% et une spécificité de 96% ce qui très proche des résultats de test même si la sensibilité est deux fois moins élevée. Si nous traçons la fonction de perte sur les 7350 individus classés « positif » nous observons les deux sinistres découverts :

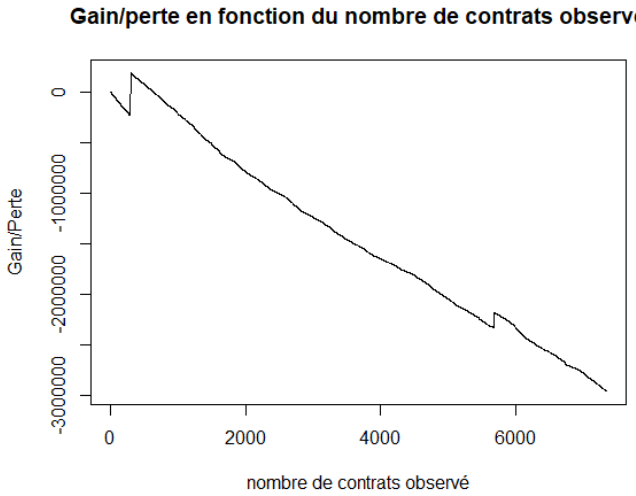


Figure 14: Représentation graphique de la fonction de perte de l'ensemble de validation du modèle Logit amélioré.

Ces résultats semblent confirmer que le modèle n'est pas en surapprentissage puisque les résultats sont similaires quelle que soit la base utilisée. De plus, nous obtenons une fonction de gain positive grâce à un sinistre découvert relativement tôt comme lors de la phase de test. Cependant, le nombre de sinistre détecté par rapport au nombre de contrats observés semble trop faible afin de pouvoir utiliser ce modèle.

Nous pouvons encore améliorer ce modèle au niveau de l'AIC car certaines variables de ce modèle présentent des catégories non significatives (comme par exemple la variable « surface des dépendances » ou « l'ancienneté du contrat »). Il faut alors effectuer des regroupements de catégories afin d'ajuster encore mieux le modèle.
Lorsque l'on opère tous les regroupements nécessaires pour que toutes les classes soient significatives, on remarque que l'AIC baisse. Cependant, au moment de tracer la fonction de gain/perte de ce nouveau modèle, nous remarquons que la prédiction est moins précise sur les sinistres les plus scorés mais plus précis sur les 30 000 plus scorés, c'est-à-dire que le modèle découvre les sinistres plus tardivement. Les contrats ayant une probabilité élevée ne sont pas des contrats sinistrés. Or, ce qui nous intéresse dans cette étude, c'est de détecter le plus de sinistres avec le moins de faux positifs, donc de détecter le plus rapidement possible les sinistres. Du fait de ce résultat, nous garderons alors le modèle sans les regroupements dans la suite de cette étude.

Bien que les résultats des modèles ne soient pas bons, nous allons tout de même explorer des pistes d'application métier. Pour ce faire, nous allons nous concentrer sur le modèle GLM amélioré sans les regroupements de variables, puisqu'il permet de détecter des sinistres rapidement. Nous utilisons ce modèle car, contrairement aux arbres de décisions ou Gradient Boosting qui sont des « boîtes noires » en ce qui concerne l'explicabilité des variables, le modèle Logit présente des coefficients qui permettent de détecter le sens de variation de la probabilité de survenance d'un sinistre incendie grave.

IV. Utilisation métier des résultats

Dans cette partie, nous allons présenter les utilisations métier que l'on peut effectuer à l'aide des résultats des modèles précédents. En effet, le but premier est de pouvoir détecter les caractéristiques d'individus qui sont mis en avant par le modèle afin d'avoir un premier filtre sur les éléments à risques. De plus, les coefficients du modèle GLM amélioré permettront d'ajuster la répartition de la surcrête au moment de la tarification.

Attention, nous sommes conscients que les modèles possèdent des résultats faibles et que les variables ne sont pas toutes significatives, les variations présentées ci-dessous ne sont qu'à titre d'indications afin de décrire la procédure.

1. Profilage des individus les plus à risques

Dans cette partie nous allons observer les individus ayant une probabilité élevée d'avoir un sinistre incendie grave afin de comparer la répartition de cette nouvelle population avec la population totale. Ce sont les résultats du modèle Logit amélioré qui ont été retenus pour toute cette analyse. La nouvelle population est composée des 500 assurés ayant le plus fort risque dans le but de rester cohérent avec les résultats du modèle puisque la courbe de gain passe positive sur les 500 premiers scores. La base contenant ces 500 individus est extraite afin d'observer la répartition des classes de chaque variable. Cette base sera appelée « Base risquée ». Les comparaisons sont présentées ci-dessous.

La première remarque que nous pouvons faire c'est que le modèle ne place dans le haut du classement que des individus ayant une maison. En effet, parmi les 500 premiers individus scorés par le modèle, 99,8% possèdent une maison alors que dans la base de données initiale il y avait 67.8% de maison, cela implique que, selon ce modèle, les maisons sont les plus exposées à la survenance d'un sinistre incendie grave.

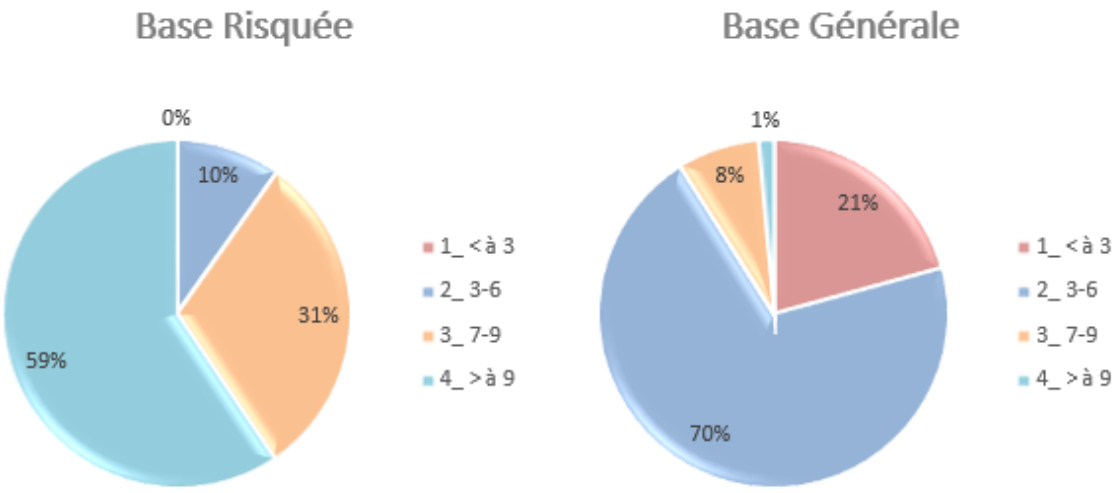


Figure 15: Comparaison de répartition du nombre de pièces principales.

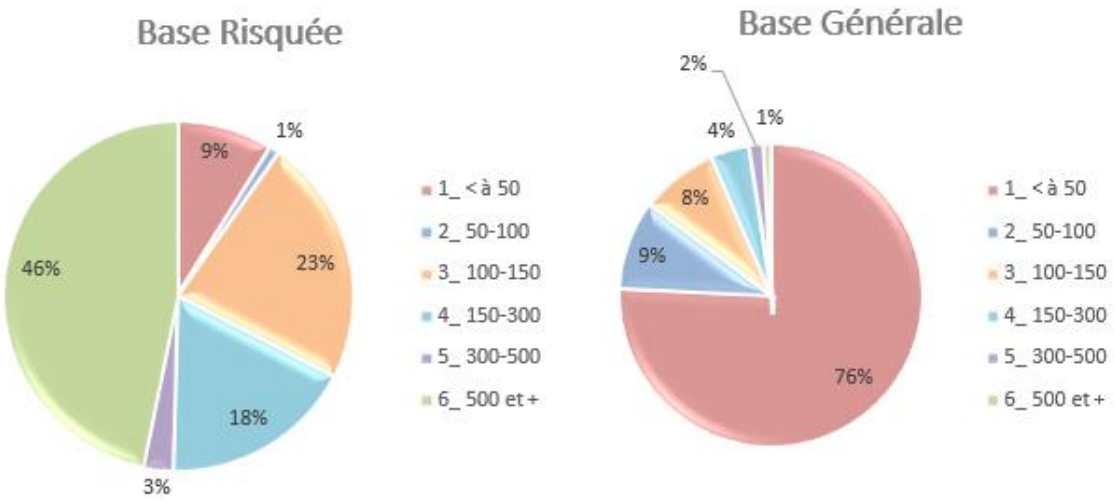


Figure 16: Comparaison de répartition de la surface des dépendances.

D'après les deux graphiques précédents, nous remarquons que les individus ayant des grandes dépendances ou ceux ayant un nombre de pièces principales important sont classés comme des personnes à risque. En effet, sur le graphique représentant la répartition du nombre de pièces principales, le modèle classe, parmi les 500 premiers plus risqués, des individus possédant plus de 6 pièces principales. De même pour la surface des dépendances puisque la proportion des individus ayant plus de 100m² de surface de dépendances augmente. Initialement, ces catégories ne représentent respectivement que 9% (8% + 1%) pour la répartition générale du nombre de pièces principales supérieur à 6 et 15% (8% + 4% + 2% + 1%) de la répartition générale des surfaces de dépendance alors que ces classes représentent 90% de la répartition de la base risquée pour les deux variables, c'est-à-dire que le modèle sélectionne principalement ces individus.

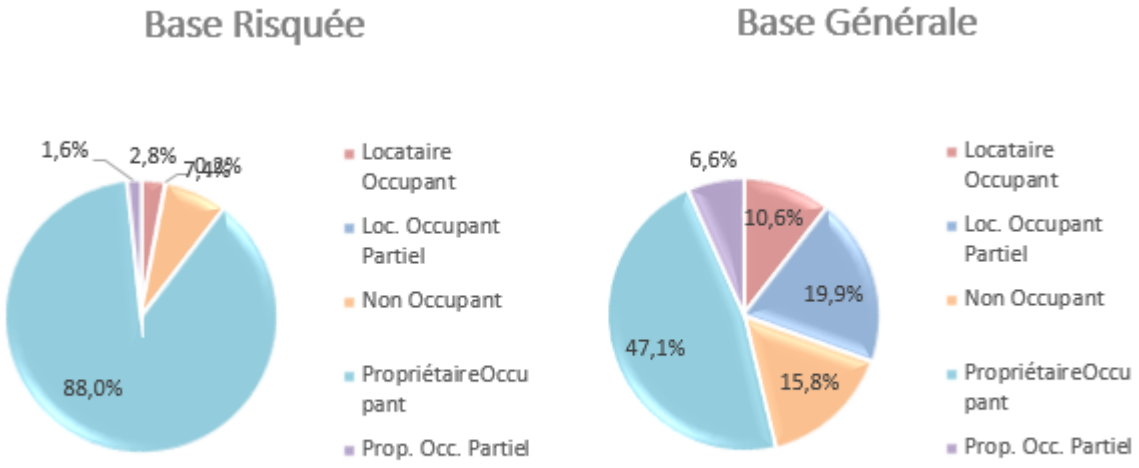


Figure 17: Comparaison de répartition de la qualité de l'occupant

Le modèle fait donc ressortir en particulier des profils d'assurés qui sont des propriétaires occupants de maison, ayant une surface de dépendance supérieure à 500m² et 9 pièces principales. Ce profil correspond au profil établi lors de l'analyse statistique de la base de données, c'est-à-dire que sur cette analyse le modèle ne nous donne pas d'information supplémentaire par rapport à l'analyse initiale. Cependant, des profils non identifiés initialement sont ressortis à l'aide du modèle. En effet, les personnes morales, très peu présentes dans la base générale, ont tendance à être plus représentées dans la base risquée. Nous considérons les personnes morales comme les Sociétés Civiles Immobilières (SCI), c'est-à-dire des personnes qui s'associent pour acheter un bien immobilier et le gérer ensemble.

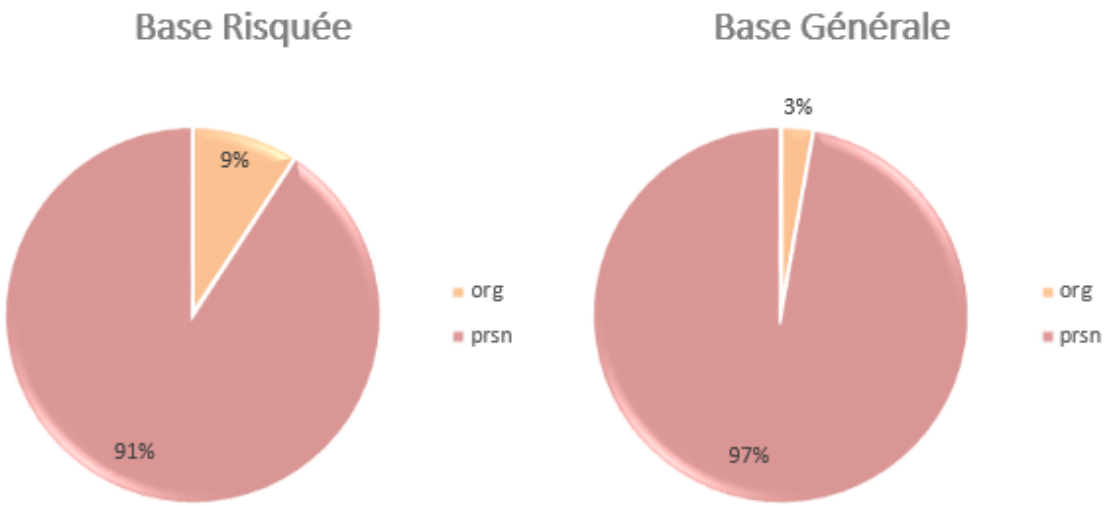


Figure 18: Comparaison de répartition du type de personne.

Le graphique indique que les personnes morales sont 3 fois plus représentées dans la base risquée que dans la base générale. On aurait pu penser que comme les mobilhomes qui sont très peu présents dans la base générale, ils ne soient pas du tout représentés dans la base risquée. On peut alors imaginer que les SCI peuvent présenter un risque non négligeable pour le risque d'un incendie grave.

2. Application sur le tarif : Répartition non-homogène de la surcrête

Nous allons maintenant chercher à trouver l'importance qu'une classe de variable apporte à la probabilité d'avoir un sinistre. Nous travaillons toujours avec le modèle Logit amélioré sans le regroupement des classes pour tenter de détecter l'effet des variables sur la probabilité. Les variables étudiées dans cette partie sont la surface de dépendance et le nombre de pièces principales car ce sont les variables qui semblent le plus être mises en avant dans le modèle comme nous l'avons vu dans l'analyse de la partie précédente. De plus, ces variables permettent de faire le lien avec le modèle de la souscrête. Cela permet de comparer la variation des coefficients et donc de pouvoir ajuster différemment la surcrête.

L'individu de référence du modèle est une personne ayant une surface de dépendance inférieur à 50m², qui a moins de 3 pièces principales, qui n'a pas déclaré de sinistre à la souscription, qui est dans le portefeuille depuis moins d'un an, qui a un taux de dégressivité à 0 et qui vit dans un appartement. Pour cet individu et selon le modèle Logit amélioré, la probabilité d'avoir un sinistre grave est de 0.0000508. Posons cet individu comme référence et comparons maintenant, toute chose égale par ailleurs, ce qu'engendre un changement de catégorie.

Par exemple, si l'individu possède maintenant, non plus moins de 3 pièces principales, mais entre 3 et 6 pièces. Sa probabilité sera de 0.0001213. Cet individu aura donc une probabilité plus élevée d'avoir un sinistre incendie grave dans l'année. Nous pouvons dans un premier temps observer l'évolution du coefficient de la variable « Nombre de pièces principales » dans le modèle. Nous obtenons le graphique suivant :

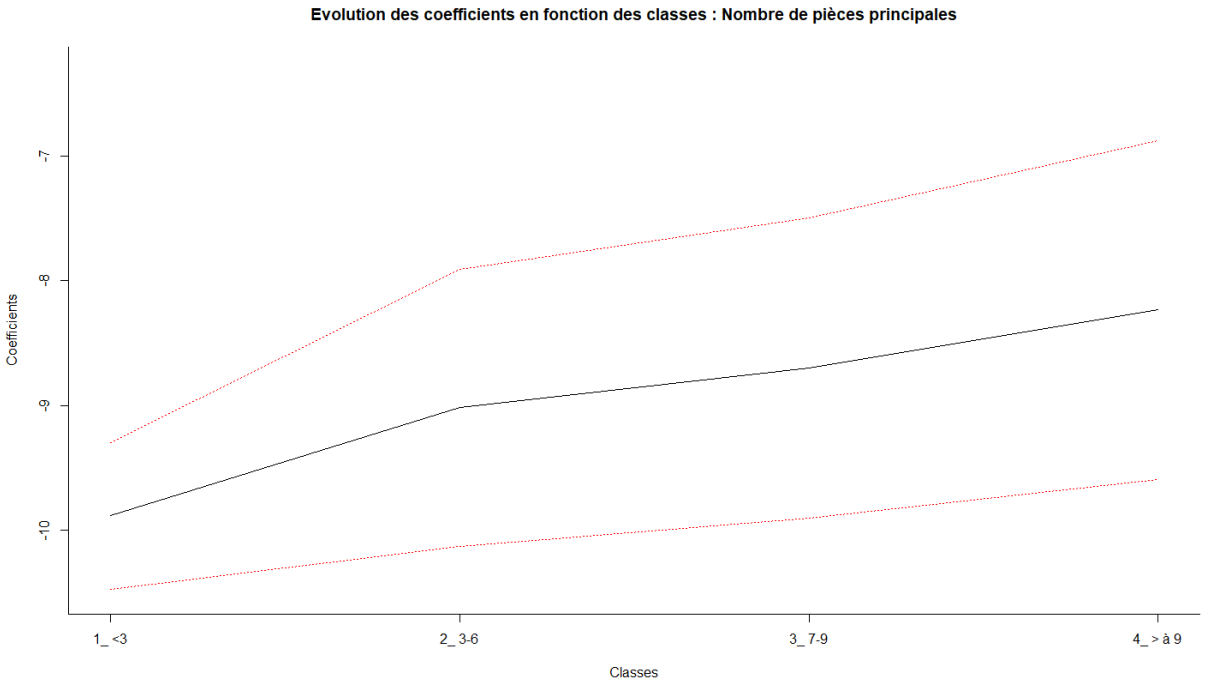


Figure 19 : Évolution des coefficients de la variable "nombre de pièces principales".

Nous remarquons que le coefficient croit en fonction des classes, ce qui indique que plus le nombre de pièces principales est élevé, alors plus la probabilité d'avoir un sinistre incendie grave sera important. On peut alors observer la variation de la probabilité avec le graphique suivant :

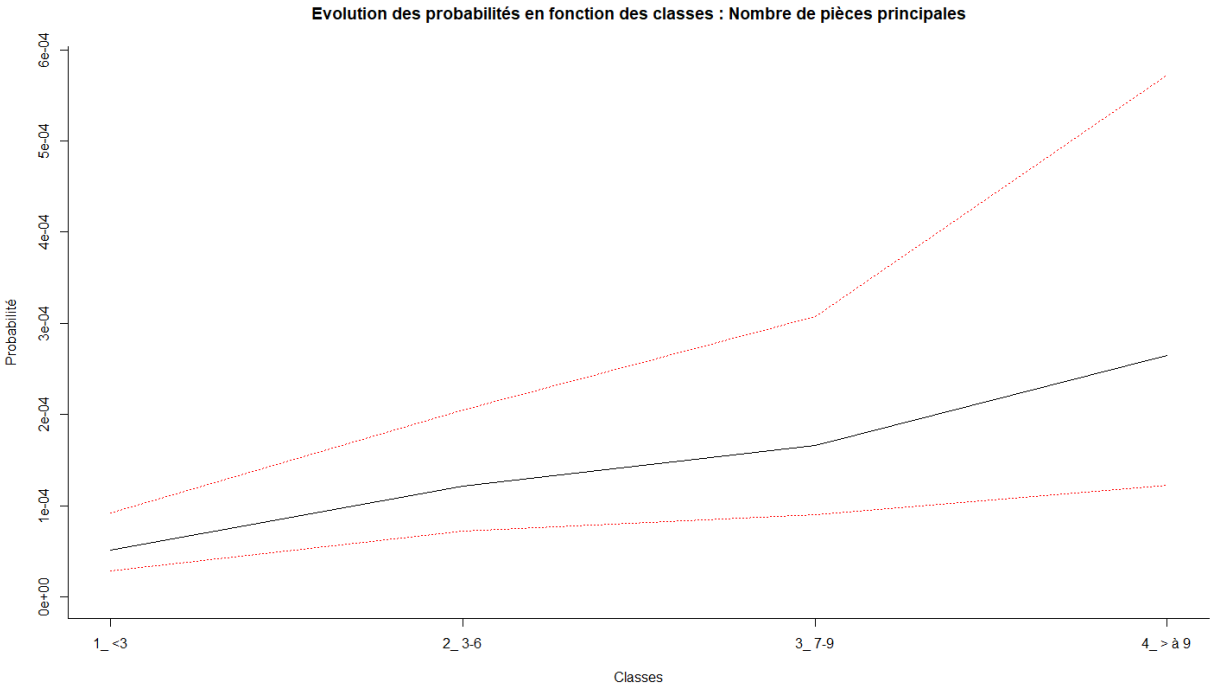


Figure 20 : Évolution de la probabilité d'avoir un sinistre grave en fonction du nombre de pièces principales du logement.

Nous remarquons que plus le nombre de pièces principales augmente, plus la probabilité est élevée, ce qui complète l'analyse du graphique des coefficients. Il faut tout de même faire attention puisque les intervalles de confiance à 95% semblent relativement grands et de moins en moins précis.
Si l'on considère la catégorie de référence comme la base 1, nous remarquons que la probabilité est 2.38 fois plus élevée pour les individus ayant 3 à 6 pièces principales, 3.3 fois pour les 7-9 pièces et de 5.3 pour ceux ayant plus de 9 pièces.

Pour la variable représentant les surfaces de dépendances, nous ne remarquons pas de réelle tendance, mis à part que pour les assurés qui ont plus de 500m² de dépendances, la probabilité croit assez fortement mais l'intervalle de confiance semble lui aussi très élevé.

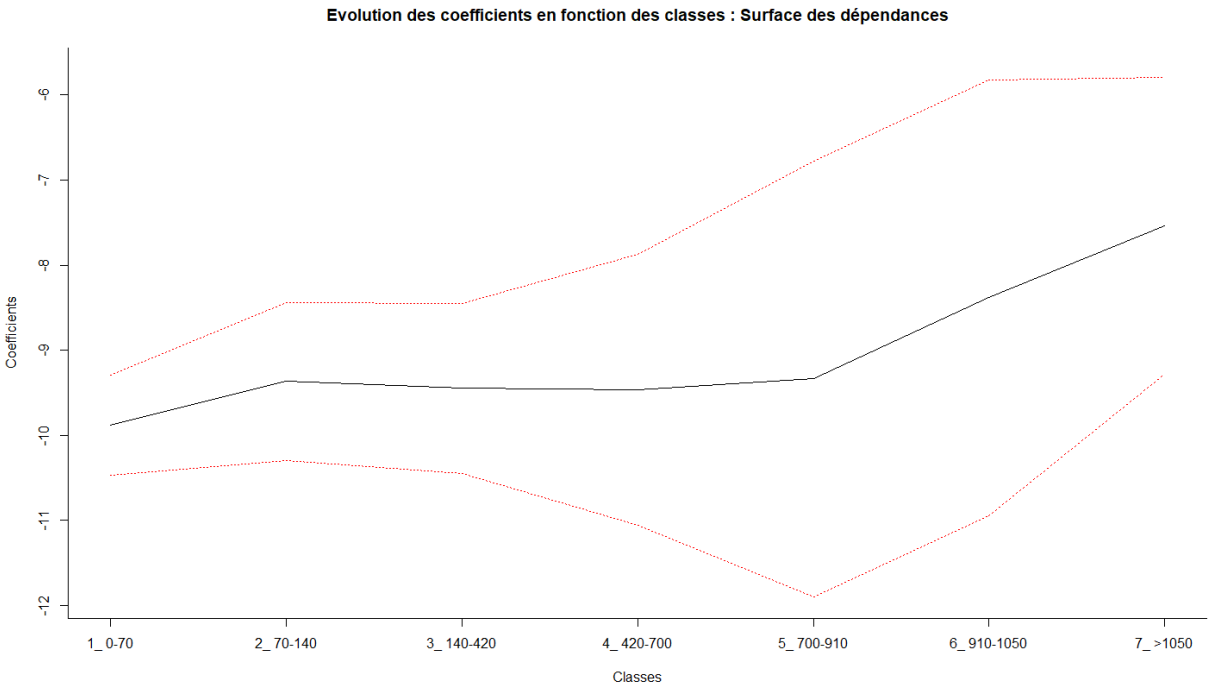


Figure 21 : Évolution des coefficients de la variable "Surface des dépendances".

Le graphique précédent représente l'évolution des coefficients de la variable « Surface des dépendances » dans le modèle Logit avec son intervalle de confiance. On remarque que le coefficient croit puis stagne jusqu'à 700m² et croit fortement au-delà. Cependant, comme nous l'avons précisé au début de cette partie, certaines classes ne sont pas significatives, et cela se remarque au fait que la taille de l'intervalle de confiance est très large et comprend la valeur 0. Il faut donc rester prudent avec l'exploitation de ces coefficients.

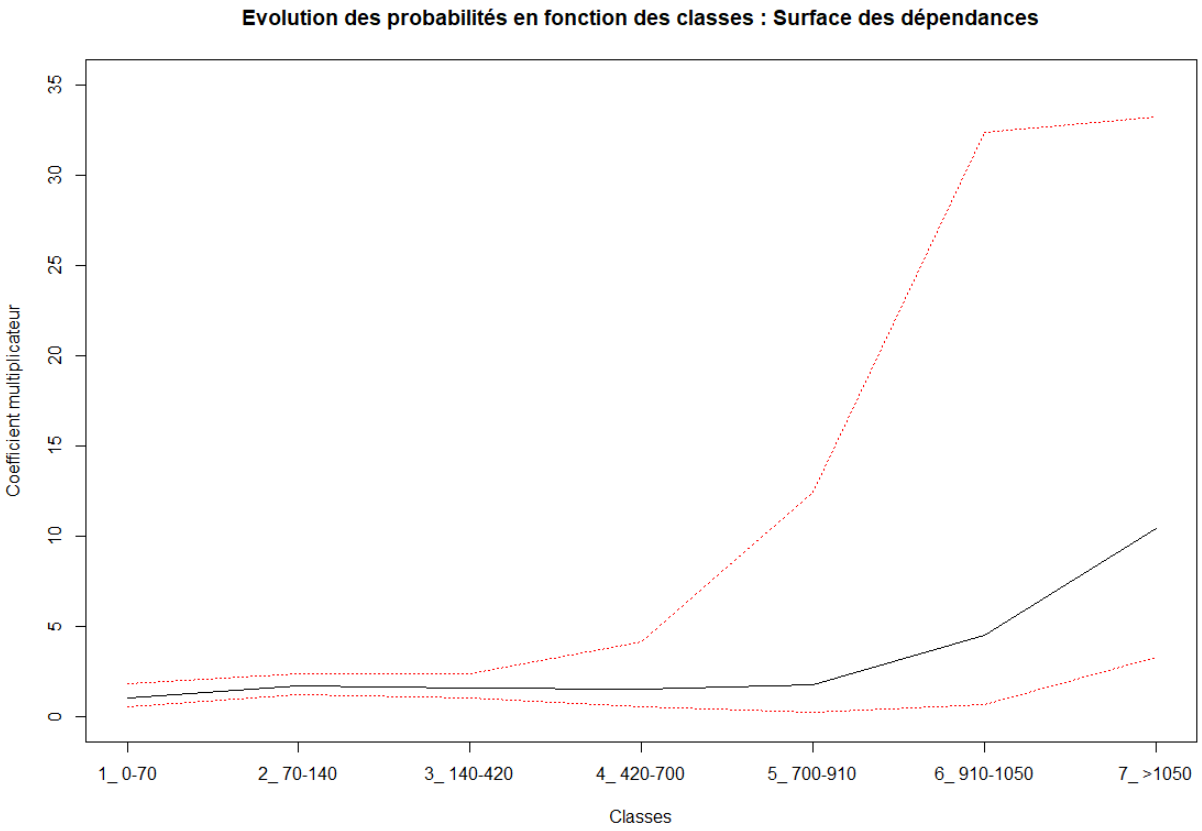


Figure 22: Évolution de la probabilité d'avoir un sinistre grave en fonction de la surface des dépendances.

Si le modèle était performant et que toutes les classes étaient significatives, il aurait été intéressant de pouvoir pondérer les classes lors de l'ajout de la surcrête dans la tarification. En effet, au lieu de la répartir de façon homogène sur toutes les classes les mêmes coefficients que la souscrête, les coefficients trouvés précédemment auraient pu servir de poids afin d'ajuster au mieux les primes pour le risque encouru.
Nous pouvons comparer les coefficients actuels de la souscrête avec les coefficients du modèle de la surcrête. Nous obtenons en variation les graphiques suivants :

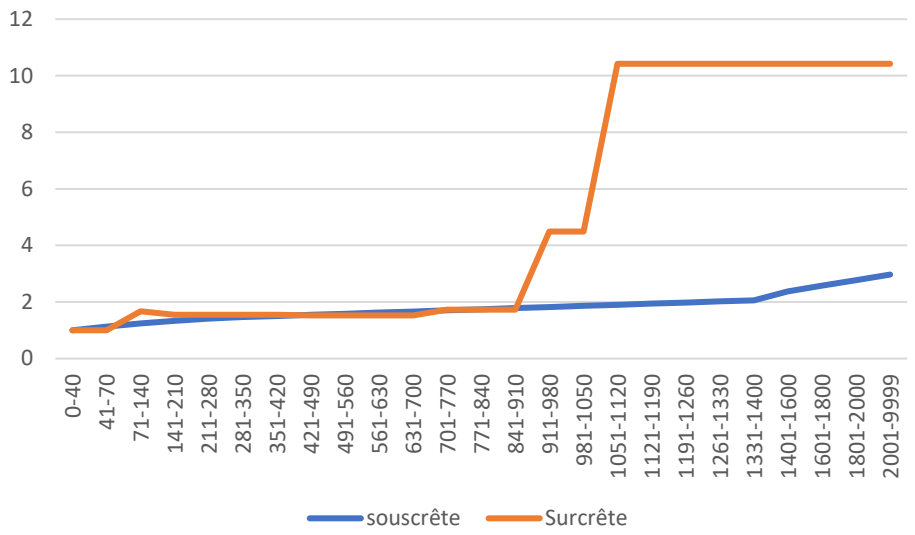


Figure 23 : Variation des coefficients pour la surface des dépendances par rapport à la classe de référence.

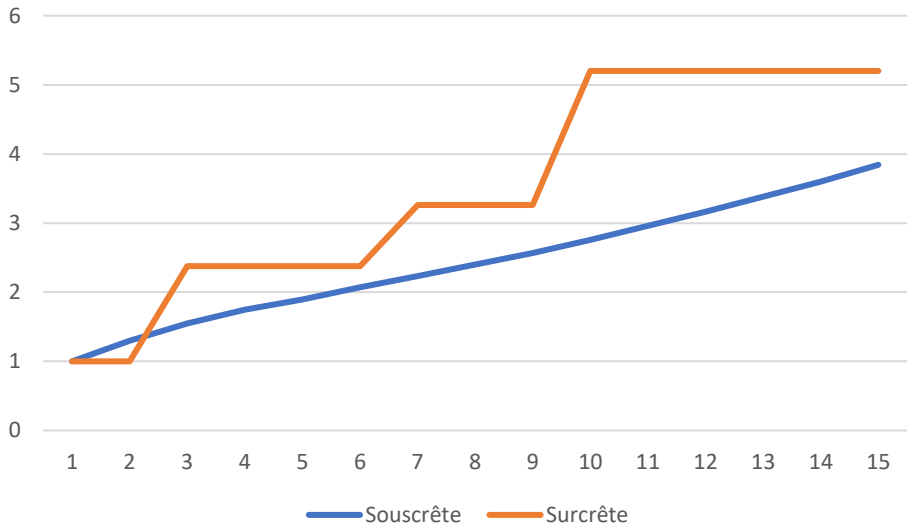


Figure 24 : Variation des coefficients du nombre de pièces principales par rapport à la classe de *référence*.

Pour la variable « Surface des dépendances », nous remarquons que jusqu'à 900m², les variations des coefficients sont similaires entre la souscrête et la surcrête, ce qui signifie qu'il n'y a pas besoin de répartir différemment sur ces classes. Cependant, au-delà de 900m² la variation des coefficients augmente très fortement, pour atteindre un niveau plus de 10 fois supérieur à la classe de référence. Il paraît important d'effectuer une analyse plus approfondie sur ces classes.
De plus, pour la variable « Nombre de pièces principales », nous remarquons que la surcrête est quasiment tout le temps au-dessus de la souscrête, ce qui signifie qu'il faudrait alors pondérer de manière différente la répartition de la surcrête.

Lorsque l'on regarde les individus ayant une surface de dépendance supérieure à 900m², nous remarquons que ces personnes ont aussi un nombre de pièces principales élevé comme nous pouvons le remarquer sur le graphique suivant. En effet, la classe des assurés ayant plus de 9 pièces principales est 7 fois supérieure à la répartition initiale de cette même classe. Au contraire, pour un nombre de pièces principales faible, tous les coefficients sont diminués. Si on effectue une majoration sur les grandes surfaces de dépendance et le nombre de pièces principales élevé, alors nous toucherons les mêmes individus et ils seront doublement impactés par cette hausse. Cette majoration aura un effet de faire résilier ces assurés, ce qui n'est pas l'objectif recherché.

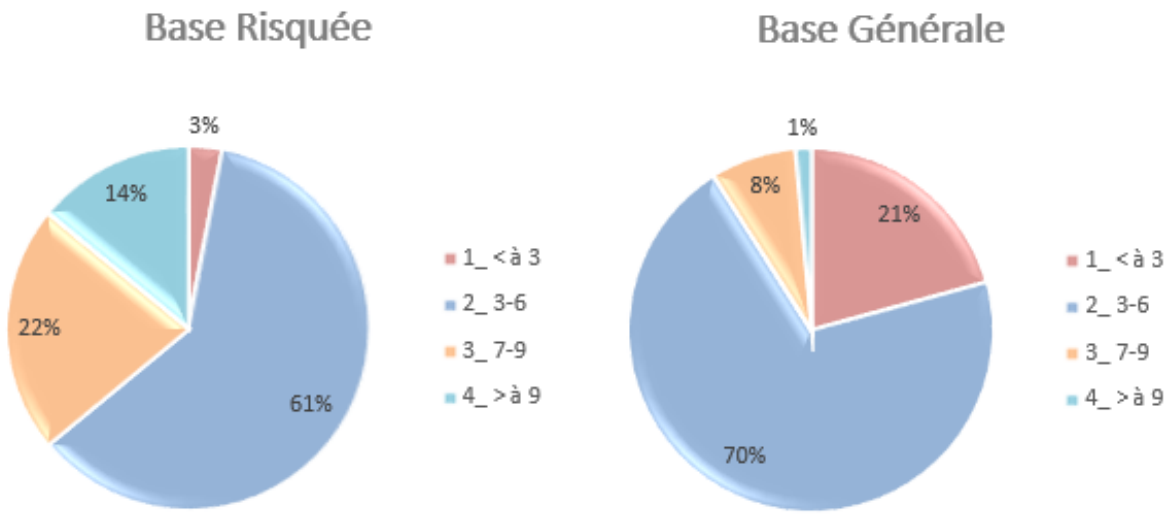


Figure 25 : Répartition de la variable « Nombre de pièces principales » des individus possédant plus de 900m² de surface de dépendance et de la base générale.

Une autre solution afin de réduire le risque de ces assurés est d'effectuer des études plus approfondies lors de la visite de risque de l'agent général pour les nouveaux assurés, ce qui permettrait de prendre des décisions au moment de la souscription. Il pourrait aussi être bien de mandater un expert, différent de l'agent général, afin d'avoir un œil différent sur le risque et de détecter des zones pouvant être synonyme d'incendie. L'objectif étant aussi de faire de la prévention, en indiquant à l'assuré d'effectuer des changements (réparation, nettoyage, …) pour que l'on accepte le risque.

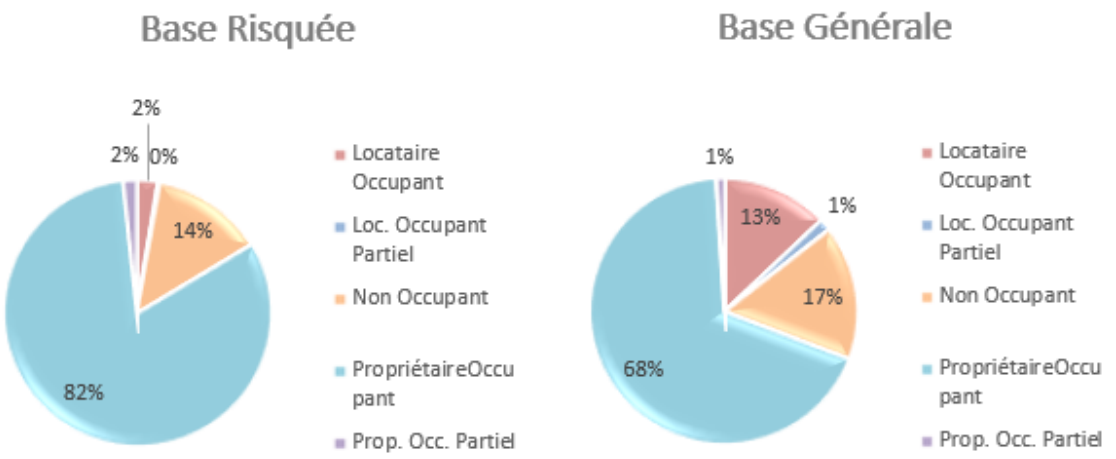


Figure 26 Répartition de la variable « Qualité de l'occupant » entre la base risquée et la base générale.

Nous remarquons, comme dans la partie précédente que parmi les assurés possédant plus de 900m² de surface de dépendance, la catégorie des non-occupants et les SCI ressortent. Il peut être intéressant d'étudier au cas par cas les contrats du portefeuille encore actif des personnes possédant plus de 900m² de dépendances et avec un nombre de pièces principales supérieur à 6, de même pour les contrats des SCI encore actifs à ce jour. Le but serait maintenant de prévenir le risque, en effectuant une visite de contrôle afin de s'assurer que le logement ne présente pas de danger imminent.

V. Sujets de réflexion

Dans ce chapitre, nous allons discuter des sujets non-traités jusqu'ici mais qui auraient pu être intégrés dans les différentes étapes de cette étude.
En effet, ce mémoire traite des sinistres incendie graves des contrats Multirisque Habitation et parmi les sinistres graves, certains nécessitent, pour la solvabilité de l'entreprise, de faire appel à un programme de réassurance du fait de leur montant potentiellement très élevé. Nous verrons ainsi ce qui est en place à Thélem Assurances en termes de réassurance ainsi qu'un risque particulier, le risque de conflagration.
Ensuite, un objectif de cette étude est la prévention du risque incendie, nous allons nous demander si le fait d'intégrer la domotique (ou maison connectée) dans la maison des assurés permettrait de prévenir ou de diminuer les risques d'incendie ou de fraude à l'assurance.
Enfin, la dernière partie traitera de trois variables non-intégrées dans la base de données et qui peuvent être en lien avec le coût et la gravité des sinistres incendies.

1. La réassurance et le risque de conflagration.

Dans cette partie, nous allons tout d'abord aborder ce qu'est la réassurance avec ce qui est en place à Thélem Assurances contre le risque incendie des contrats Multirisque Habitation. Ensuite, nous définirons le risque de conflagration et pourquoi il n'a pas été détaillé particulièrement dans la mise en place des modèles.

i. La réassurance à Thélem Assurances.

La réassurance permet à un assureur de s'assurer contre les risques les plus importants de son portefeuille et lui permet de protéger son bilan : l'assureur va pouvoir limiter sa charge sinistre en ne prenant en compte qu'une partie du sinistre, l'autre partie sera prise en charge par le réassureur, on dira alors que l'assureur va céder au réassureur une partie de sa volatilité. Plusieurs formes de réassurances existent : la réassurance obligatoire (les traités) et la réassurance facultative. Pour ces deux formes s'ajoutent deux catégories : la réassurance proportionnelle et la réassurance non-proportionnelle.

A Thélem Assurances, un traité est actif afin de couvrir les risques graves des contrats Multirisque Habitation. Parmi ces risques graves, nous retrouvons le risque incendie. C'est un traité de réassurance non-proportionnelle, appelé excédent de sinistre (ou XS signifiant *Excess of Loss*), qui permet de céder la partie du montant du sinistre au-delà d'un seuil fixé. Ce seuil est appelé la priorité. Pour certains cas, le réassureur ne prend en charge qu'un montant maximum au-delà de la priorité. Ce montant est appelé la portée. Donc si le montant du sinistre est supérieur à la priorité plus la portée, la partie restante est à la charge de l'assureur ou potentiellement, l'assureur fait appel à un autre réassureur pour prendre en charge le reste via différentes tranches de montant. En général, plusieurs tranches sont proposées pour un même traité permettant à différents réassureurs d'intervenir. Certains préférant les tranches « basses » (c'est-à-dire de faibles montants, elles surviennent en premier et la période de

retour est relativement faible) et d'autres préférant des tranches « hautes » (c'est-à-dire avec des montant de plus en plus élevés, elles surviennent plus rarement et les périodes de retour sont plus élevées). L'objectif de ces différentes tranches est de permettre à l'assureur de diversifier ses risques afin de réduire le risque de défaut du réassureur.

Pour un traité simple avec une tranche, si par exemple le contrat est composé d'une priorité à 500 000€, une portée de 1 000 000€ et qu'un sinistre de 2 000 000€ survient. L'assureur devra couvrir le risque entre 0 et 500 000€, puis l'assureur cédera au réassureur la partie du sinistre entre 500 000€ et 1 500 000€ et enfin la partie entre 1 500 000€ et 2 000 000€ sera à la charge de l'assureur. Nous pouvons schématiser cet exemple à l'aide de la figure 27.

Cependant, la portée et la priorité ne sont pas imposées par le réassureur. Dans un objectif d'être couvert au mieux par rapport aux risques encourus, Thélem Assurances a tout un processus d'optimisation afin de déterminer les deux valeurs qu'il demandera de coter au réassureur. En effet, chaque année afin de renouveler les traités de réassurance, Thélem Assurances réalise des modélisations qui permettent d'observer le gain ou la perte potentielle que pourrait engendrer un changement de priorité et de portée. Lorsque le choix de priorité est déterminé, des études complémentaires sont effectuées afin de vérifier si le comportement du ratio sinistres/cotisations et d'autres indicateurs sont toujours en accord avec les objectifs attendus. De plus, lorsque l'assureur a connu une trop forte sinistralité durant un exercice sur les dossiers réassurés, la prime de réassurance peut fortement augmenter l'année suivante. Dans ce cas, afin de faire diminuer la prime de réassurance, il est nécessaire d'augmenter la priorité.

De plus, dès lors qu'un sinistre survient, l'assureur doit 'reconstruire' sa protection en payant une nouvelle prime afin d'être couvert à nouveau si jamais un autre sinistre dont le montant est supérieur à la priorité survient. Ce processus est appelé reconstitution. Cette reconstitution peut être gratuite et il peut y en avoir plusieurs. Le nombre et le prix sont définis en amont dans le contrat.

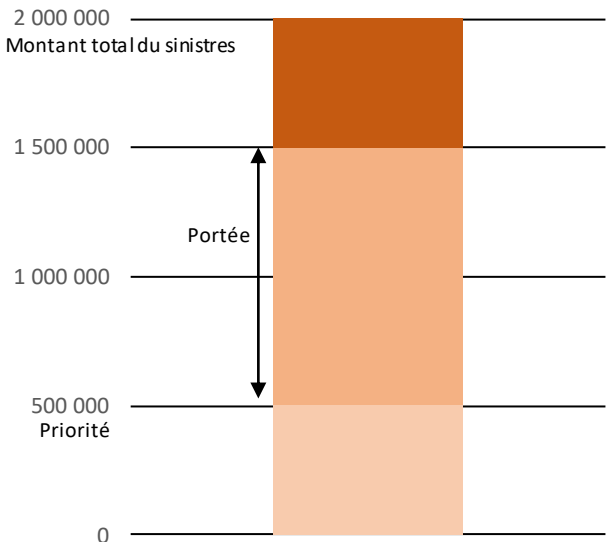


Figure 27 : Schéma de réassurance obligatoire excédent de sinistre.

Concernant la réassurance dans cette étude, comme nous l'avons vu à la fin de la *partie II : Base de données*, le nombre de sinistre incendie des contrats Multirisque Habitation supérieur à 30 000€ est très faible : 396 sinistres graves dans notre base. Il y en a alors encore moins faisant appel au traité de réassurance. Le service Réassurance de Thélem Assurances indique que depuis 2012, seulement 4 sinistres ont été supérieurs à la priorité pour le produit Multirisque Habitation et ces sinistres ne concernaient pas nécessairement la garantie incendie. C'est pourquoi dans les modèles mis en place dans cette étude, nous n'avons pas intégré d'étude spécifique sur la réassurance. En effet, l'étude présentée est principalement axée sur le coût total du sinistre incendie afin d'en avoir une vision globale sans prendre en compte ni la responsabilité du sinistre ni les sinistres frauduleux ni le recours à la réassurance.

ii. Le risque de conflagration.

Parmi les risques de réassurance en lien avec les sinistres incendie des contrats Multirisque habitation, nous pouvons retrouver le risque de conflagration. Selon l'APREF (l'Association des Professionnels de la REassurance en France), « *la notion de conflagration en réassurance tend à désigner un sinistre de caractère événementiel touchant plusieurs risques, mais qui n'est pas de nature événement naturel. On entend généralement par conflagration en réassurance un événement, soit de type Catastrophe Technologique au sens de la loi Bachelot, soit de type incendie/dommages aux biens toutes causes qu'il parait utile d'expliciter* ». Autrement dit, le risque de conflagration est un risque catastrophe d'origine humaine qui peut être le résultat de différents évènements et qui cause des dégâts dans une zone définie. Ce sont donc des risques peu fréquents mais dont l'impact peut être d'une grande ampleur. L'objectif du traité de réassurance couvrant la conflagration est de couvrir ce risque sur une période de 200 ans et les pertes maximales estimées.

En France, nous avons l'exemple de l'explosion de l'usine AZF à Toulouse en 2001. Cette catastrophe a causé 31 décès et 2 500 blessés et l'estimation des dégâts matériels est de 2 milliards d'Euros due au fait que l'explosion a atteint les infrastructures autour de l'usine (écoles, salles de spectacles, habitations …). C'est pourquoi, dans le cadre des exigences de la loi Solvabilité II sur la survenance d'évènement à période de récurrence de 200 ans et dans la prise en compte des risques de réassurance lors de la signature du traité, une étude d'impact et une étude sur les zones géographiques doivent être menées afin de prendre en compte le risque de conflagration du portefeuille de l'assureur.

Dans ce mémoire, il aurait pu être intéressant d'ajouter une dimension supplémentaire prenant en compte le risque géographique des contrats. En effet, les contrats situés dans une zone recensée dans les études d'impacts du risque de conflagration auraient pu avoir une pondération plus importante sur le fait d'avoir un sinistre incendie grave par rapport aux contrats non-présents dans une zone à risque. Mais comme nous l'avons vu dans la sous-partie précédente, les cas de sinistre catastrophe sont très peu présents et cette pondération n'aurait pas eu d'impact dans notre étude. Cependant, une étude géographique plus précise des sinistres incendies en fonction des régions (ou départements ou même avec une vision sur les communes) de France aurait permis d'ajouter une échelle de risque en fonction de la

zone géographique du logement. Il faudrait refaire tous les modèles et regarder si l'ajout de cette variable dans la base de données, permettrait d'améliorer les résultats.

2. Prévention des risques des contrats multirisque habitation à l'aide de la maison connectée.

Dans cette partie, nous allons dans un premier temps nous demander si l'ajout d'objets connectés dans la maison des assurés permettrait à l'assureur de prévenir le risque d'incendie puis dans un second temps le risque de fraude à l'assurance.

i. Prévention des sinistres incendies.

Un des objectifs de ce mémoire et des modèles mis en place est de prévenir les risques de sinistre incendie grave. Les modèles utilisés précédemment ont permis de détecter des profils types de clients afin d'établir des visites de risque plus approfondies pour ces logements lors de la souscription afin de déceler des risques avérés.

Mais, dans un environnement où tout est connecté, l'utilisation de la domotique à l'intérieur des maisons pourrait aussi servir à la prévention du risque incendie. En effet, placer dans le domicile de l'assuré une caméra reliée au téléphone ou encore un détecteur de fumée connecté directement aux services des pompiers permet d'avertir les secours et d'intervenir le plus rapidement possible sur le lieu du sinistre. Cela permettrait de limiter le coût du sinistre pour l'assureur puisque le logement serait pris en charge à temps et la propagation de l'incendie serait plus faible. De plus, les caméras installées et les objets supplémentaires (lampes connectées, thermostats connectés, …) nécessaires à la mise en place d'une maison connectée pourraient aussi permettre de déceler des défaillances électriques (surchauffe électrique, augmentation anormale de la consommation, …) avant même que le sinistre survienne en émettant une alerte à l'assuré et en lui indiquant d'effectuer une mise aux normes de ses installations le plus rapidement possible. D'après l'étude « *Insurance and Technology* » de Morgan Stanley menée en 2015, il apparait que les assureurs réduiraient le risque pour l'assurance Multirisque Habitation (tous risques confondus) de 40 à 60% à l'aide de ces nouveaux outils. Dans cette logique, si la maison connectée réduit le risque de l'assureur, alors les clients pourraient ainsi voir diminuer leur prime d'assurance.

Cependant, aujourd'hui aucune étude concrète ne confirme la baisse du nombre de sinistre incendie dans les foyers due à l'installation des objets connectés. Il peut alors être intéressant de trouver une autre utilisation de ces derniers. En effet, les objets connectés recueillent une masse d'information supplémentaire qui pourraient être utilisés dans des modèles de détection de sinistre. Les informations enregistrées par les objets connectés juste avant (ou quelques jours avant) un sinistre peuvent donner des indications sur les circonstances de l'incendie. Mais alors un autre problème peut se poser : l'utilisation de ces données. La période actuelle avec la Réglementation Générale sur la Protection des Données (RGPD)

impose un cadre juridique complexe sur l'utilisation des données personnelles et les clients ne seraient pas forcément d'accord de laisser l'assureur utiliser ces données.

Actuellement, Thélem Assurance est en collaboration avec IMA Protect qui est une société qui proposent des solutions de télésurveillance et d'objets connectés pour les logements. Cependant, leur objectif est principalement la prévention du risque de vol et d'incendie et l'utilisation des données récoltées reste restreinte aux propriétaires et à la société de protection. L'assureur ne peut alors en aucun cas récupérer les données pour ensuite les utiliser. C'est pour cela qu'il est difficile de définir le réel gain qu'apporte ces objets et les potentielles autres utilisations que pourraient avoir l'assureur.

Dans ce contexte, la maison connectée peut être un atout pour l'assureur dans la prévention et la réduction du risque incendie mais aujourd'hui l'efficacité n'a pas encore été totalement prouvée. De plus, l'utilisation des données enregistrées est encore assez compliquée à cause de la réglementation en vigueur. Nous allons essayer de voir si l'utilisation des objets connectés pourrait permettre de prévenir du risque de fraude à l'assurance.

ii. Prévention de la fraude à l'assurance.

Dans cette partie nous allons tout d'abord définir ce qu'est la fraude à l'assurance et présenter ce qui est en place à Thélem Assurances. Ensuite nous nous interrogerons sur la prévention du risque de fraude à l'assurance et en particulier la fraude lors d'un sinistre incendie à l'aide des outils des maisons connectées.

Toute d'abord, selon l'Agence de Lutte contre la Fraude à l'Assurance (ALFA), la fraude à l'assurance se définit comme « *un acte intentionnel, réalisé par une personne morale ou physique, afin d'obtenir indûment un profit du contrat d'assurance* ». Autrement dit, la fraude à l'assurance se caractérise par le fait que l'assuré récupère de l'argent de la part de son assureur à l'aide de mauvaises intentions. Cette fraude à l'assurance peut se diviser en trois parties : la fraude à la souscription (fausse identité ou renseignements erronés au moment de souscrire le contrat d'assurance), fraude à la déclaration du sinistre (déclaration exagérée, sinistre intentionnel) et la fraude à la multi-assurance (souscrire à plusieurs contrats chez différents assureurs et se faire indemniser plusieurs fois).

En 2018 en France, le montant de fraude à l'assurance IARD était de près d'un demi-milliard d'euros selon l'ALFA. Afin de limiter ce montant, Thélem Assurances possède depuis quelques années une cellule spéciale qui analyse les dossiers potentiellement frauduleux. Ces dossiers sont remontés soit par les apporteurs ou les experts soit à l'aide de critères spécifiques lors de la déclaration du sinistre. Conjointement avec cette cellule, des études ont été menées afin de trouver des profils types de fraudeurs. En effet, à l'aide des antécédents de cas avérés de fraude, il est possible de déterminer des variables représentatives qui permettent de distinguer les dossiers soupçonnés de fraude des dossiers normaux grâce à des critères précis. De plus, des « blocs-notes » sont renseignés lors de la déclaration du sinistre. Une étude sur

l'apparition de mots-clés dans ces textes libres a été effectuée afin de faire remonter des mots ou groupes de mots dont la fréquence d'apparition est la plus élevée. Ces études ont permis de faire remonter de plus en plus de dossiers frauduleux et ainsi réduire le coût engendré pour l'assureur. Cette efficacité est due au fait que la cellule fraude permet de lutter au quotidien sur tous les cas détectés, mais aussi que grâce aux modèles développés en parallèle, cela permet de donner un appui supplémentaire à la partie métier pour aller plus loin dans la sélection et l'étude des dossiers. Mais l'un ne marche pas sans l'autre : l'œil d'un expert reste indispensable dans l'efficacité de la détection de la fraude à l'assurance et la modélisation ne peut se faire que si elle répond bien à l'attente métier.

Cependant, cette méthode n'est pas infaillible et des cas de fraude à l'assurance pour les contrats Multirisque Habitation peuvent être plus compliqués à déceler. On peut donc se demander si le fait d'intégrer des objets connectés dans le logement de l'assuré permettrait de réduire les cas de fraude à l'assurance.

Tout d'abord, la maison connectée pourrait réduire les cas de fraude à la souscription. En effet, si par exemple l'assuré ne déclare pas de cheminée ou d'insert, mais que les caméras de surveillance détectent le contraire, nous avons une fausse déclaration de l'équipement de l'assuré. Les caméras permettraient d'évaluer plus réellement le risque encouru par l'assureur. De plus, les vidéos de surveillance peuvent filmer les différents objets dans les pièces concernées. Si lors du sinistre, l'assuré exagère sur le montant des pertes, les images capturées pourront alors faire preuve de fausse déclaration de sinistre. Ces exemples montrent que les maisons connectées peuvent permettre à l'assureur de réduire les cas de fraude lors de la souscription ou même lors de la déclaration de sinistre.

Mais, comme vu dans la sous-partie précédente, malgré la collaboration entre Thélem Assurances et la société IMA protect dont l'objectif reste de prévenir le risque d'incendie ou de vol mais pas de fraude, nous restons confrontés au même problème sur l'utilisation des données. Si l'assuré ne donne pas l'accord de pouvoir traiter les images alors ces objets ne sont pas utiles directement à l'assureur. Même si ces technologies peuvent permettre de réduire le risque en matière d'incendie ou même en matière de fraude, l'utilisation ne parait pas simple et surtout semble difficile à faire souscrire aux assurés.

3. Variables supplémentaires.

Dans cette partie, nous allons nous intéresser à trois variables qui n'ont pas été traitées dans ce mémoire du fait qu'elles ne sont pas présentes dans le modèle de données : une variable sur les matériaux de construction des bâtiments, une variable sur l'année de construction et une variable sur l'année de rénovation. Ces variables auraient pu être déterminantes dans la détection des sinistres incendie grave mais aussi sur la gravité. On se demandera alors comment ces données auraient pu être intégrées d'une manière différente à l'aide des sources de données externes.

Tout d'abord, une variable sur les matériaux de construction n'est pas intégrée aux modèles. En effet, la donnée n'est pas renseignée lors de la souscription du contrat. Cependant, c'est une donnée qui paraît pertinente pour la détection des sinistres incendie grave puisque par exemple, un bâtiment principalement construit à l'aide de matériaux inflammables (en bois) ou facilement destructibles (en tôle) aura plus tendance à brûler entièrement et à présenter un montant de sinistre élevé. Au contraire, un bâtiment construit avec des matériaux moins inflammables (comme par exemple en béton ou en brique) aura tendance à moins propager les flammes et à réduire la gravité du sinistre et donc le montant. C'est pourquoi, le montant du sinistre semble corrélé aux matériaux de construction et aurait pu apparaître comme une variable représentative dans les différents modèles présentés. Cependant, en France en 2016 selon l'étude de ForumConstruire.com, 86% des maisons individuelles du panel interrogé ont été construites soit à l'aide de parpaings ou de briques ce qui implique que les maisons récentes sont principalement construites avec des matériaux résistant aux incendies.

De même, la variable sur l'année de construction du logement n'est pas présente dans la base de données alors qu'elle pourrait donner des indications sur la vétusté des matériaux ou bien encore sur les installations du logement. Mais cette variable peut ne pas être totalement fiable, puisqu'un logement peut être entièrement rénové alors que l'année de construction du bâtiment est très ancienne. Cette différence peut conduire à un biais. Il faudrait ajouter l'année de rénovation en plus de l'année de construction. Il a été question d'ajouter cette information lors de la construction de la nouvelle offre de Thélm Assurances mais cela n'a pas été validé lors de la mise en production en raison de la multiplicité des cas possibles et de l'ignorance potentielle du client au moment de la souscription.

Ces trois variables, non-prises en compte dans les modèles, paraissent tout de même intéressantes à observer. Malheureusement dans cette étude, nous n'avons pas pu les intégrer du fait du manque d'information dans la base de données. Cependant, pour combler ce manque d'information, nous aurions pu intégrer des données externes grâce à des bases *Open Data*. En effet, ces données sont libres de droits et d'utilisation et sont accessibles directement sur internet. Ces données sont de plus en plus présentes et permettent aux acteurs du marché de l'assurance d'intégrer des informations nouvelles et de divers domaines.

Par exemple, l'INSEE propose des bases de données regroupant des informations sur l'année de construction du logement, le type de chauffage, la catégorie de construction et d'autres indicateurs. Mais toutes ces données sont à un niveau de code IRIS (« Ilots Regroupés pour l'Information Statistique »), c'est-à-dire que l'information observée correspond à une zone géographique (comme le code postal mais plus précis) et non à une adresse exacte. Il aurait été nécessaire de faire des catégories en fonction des zones géographiques observées et ensuite réintégrer l'information dans la base de données. Ces retraitements ainsi que la fiabilisation des données pour intégrer ces informations aux modèles auraient pris trop de temps pour être traiter dans le temps imparti et n'étaient pas la priorité de cette étude.

Les données externes sont des puits d'information pouvant être utilisées par les assureurs et d'autres types de données auraient pu être intégrées. Par exemple des données sur la météo

et faire des cas en fonction des zones les plus couramment touchées par de violents orages ou des températures extrêmes. Mais ces données peuvent faire le cas d'une étude complémentaire du fait de leur complexité et il faudrait les examiner avec d'autres experts métiers.

En outre, nous pouvons remarquer que les nouvelles technologies ou bien les sources de données externes telles que les *Open Data* peuvent apporter des éléments de prévention concernant le risque de sinistre incendie pour les contrats Multirisque Habitation. Une étude plus approfondie de ces éléments pourrait venir compléter les résultats de ce mémoire et ainsi permettre de mettre en place de nouvelles procédures de prévention.

Conclusion

Dans ce mémoire, il est question de détecter les sinistres incendie graves des contrats Multirisque Habitation. Nous nous sommes rendu compte qu'il était difficile d'obtenir un modèle performant car la fréquence de ces sinistres est extrêmement faible. Ce rapport présente la procédure qui permet de tester des modèles d'apprentissages statistiques et de les comparer à l'aide de critères de sélection tels que l'aire sous la courbe ROC ou bien encore les matrices de confusions. Ces modèles ne sont pas assez performants pour que d'un point de vue économique, l'assureur résilie tous les contrats scorés comme les plus à risque par les modèles ou que l'assureur organise une vérification du risque systématique.

Cependant, les modèles permettent de révéler des informations sur les profils de risques des assurés. En effet, lors de l'étude de chaque modèle, des variables significatives ressortent dans chacun d'eux. En plus d'être significatives, ces variables permettent de trouver des explications sur le risque d'avoir un sinistre incendie. Il parait alors intéressant d'essayer de comprendre le sens de variation de la probabilité en fonction des classes de ces variables. L'étude des variations permet de détecter des pondérations entre les classes permettant ainsi de répartir la charge des sinistres graves, non plus de façon homogène, mais avec des proportions différentes selon les probabilités trouvées. Les coefficients trouvés sont relativement proches des coefficients déjà existants pour la souscrête, seules des catégories spécifiques (surfaces de dépendance supérieures à 500m² ou bien plus de 6 pièces principales) présentent des coefficients élevés. Le problème qui se pose, c'est que si l'on augmente le tarif pour ces individus, ils seront doublement touchés car la surface des dépendances est étroitement liée au nombre de pièces principales et cette application risquerait de faire partir les assurés.

De plus, l'étude a permis de mettre en avant les propriétaires non-occupant ou les Sociétés Civiles Immobilières (SCI) qui présentent des caractéristiques particulières. En effet, ils ne sont pas nombreux dans la base initiale mais restent toujours aussi présents en proportion dans les bases risquées. Une étude sur les contrats en cours possédant des dépendances supérieures à 900m² et plus de 9 pièces principales et les contrats SCI sera organisée afin de vérifier si le risque est véritablement présent ou non. Cette étude se fera en deux étapes : la première étape sera menée par le service souscription qui fera une étude approfondie des dossiers sélectionnés ; la deuxième étape est de demander à l'agent général une visite de risque si des éléments objectifs apparaissent lors de l'étude du dossier.
Si cette étude est concluante, c'est-à-dire que les dossiers sélectionnés paraissent pertinents, il pourra être prévu d'organiser des visites de risque récurrentes et de modifier en conséquence les modalités de souscription lors de l'entrée de nouveaux risques.

Bibliographie

[1] Arthur CHARPENTIER, *Statistique de l'assurance*, 3e cycle, Université de Rennes 1 et Université de Montréal, 2010, pp.133. cel-00550583.

[2] Trevor HASTIE, Robert TIBSHIRANI, Jerome FRIEDMAN, *The elements of statistical learnings – Data Mining, Inference and Prediction.* Springer, 2e edition, 2008.

[3] Ricco RAKOTOMALALA, *Gradient Boosting, Technique ensembliste pour l'analyse prédictive. Introduction explicite d'une fonction de coût,* Université Lumière Lyon 2, 2016.

[4] Antoine PAGLIA, *Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique.* Master's Thesis, EURIA, 2011.

[5] Christian ROBERT, *Théorie des valeurs extrêmes,* Institut de Science Financière et d'Assurances, 2019.

[6] Fédération Française de l'Assurance (07/06/2019), *Les garanties du contrat multirisque habitation*, consulté sur : https://www.ffa-assurance.fr/infos-assures/les-garanties-du-contrat-multirisques-habitation.

[7] WikiStat, *Régression logistique ou modèle* binomial, Rapport Technique, 2017.

[8] Ricco RAKOTOMALALA, *Pratique de la régression logistique : Régression logistique binaire et polytomique*, Université Lumière Lyon 2, 2014.

[9] Fédération Française de l'Assurance, *Tableau de bord de l'assurance en 2017*, Rapport technique, 2018.

[10] Association des Professionnels de la Reassurance en France (APREF), *Note conflagration*, janvier 2010.

[11] Agence de Lutte contre la Fraude à l'Assurance (ALFA), *Fraude à l'assurance*, consulté sur : https://www.alfa.asso.fr/fraude-a-lassurance/#definition.

[12] Morgan Stanley, *Insurance and Technology, Insight: The Emerging Role of Ecosystems in Insurance,* 23 mars 2015.

Table des illustrations

Annexes

Annexe A : tableau des variables sinistres.

ID_SIN	Identifiant sinistre
DAT_EFF	Date effet
DAT_VLD	Date validation
ID_CTR	Identifiant contrat
DAT_SVN_SIN	Date survenance sinistre
MNT_NET_RCR_SIN	Montant net de recours
MNT_BRT_RCR_SIN	Montant brut de recours
MNT_RCR_ENC_SIN	Montant recours encaissé
COD_EVT_SIN	Code événement sinistre
COD_CAU_SIN	Code cause sinistre
COD_TYP_DMG	Code type dommage
COD_STT_SIN	Code statut sinistre
COD_MTF_CHG_STT_SIN	Code motif changement statut sinistre
TX_RSP	Taux responsabilité
COD_PRD	Code Produit

Annexe B : Extrait des variables « contrat ».

COD_PRD	Code produit (MRH, MHA)
ID_CTR	Identifiant du contrat
DAT_EFF_CTR	Date effet du contrat
DAT_FIN_EFF_CTR	Date fin effet du contrat
DAT_DEB_ETT	Date de début d'état du contrat
DAT_FIN_ETT	Date de fin d'état du contrat
COD_TEC_GAR_FRM_CMR	Formule
ID_APT	Identifiant apporteur
ID_RIO_SSC	Identifiant du souscripteur
COD_FRC_CTR	Code fractionnement contrat
NIV_FRN	Niveau de franchise
TX_DGR	Dégressivité de la franchise
COD_OPT_CTR	Code option
TOP_RNN_RCR_CNR_LOA	Disposition : Renonciation à recours contre locataire (O/N)
TOP_RNN_RCR_CNR_PPT	Disposition : Renonciation à recours contre propriétaire (O/N)
TOP_ASS_CPT_CMM	Disposition : Assurance pour compte commun (O/N)
TOP_EXS_RSQ_LOC	Disposition : Exclusion risques locatifs (O/N)
TOP_INH_PLUS_075_JJ	Disposition : Inhabitation > 75 jours (O/N)
TOP_RC_CDH	Disposition : RC chambre d'hôte (0/1/2/3/4)
TOP_VOL_RSD_SCD	Extension : Vol résidence secondaire
TOP_GAR_DMG_ELC	Option : Garantie Dommages électriques pour formules 1 ou 3 (O/N)
TOP_GAR_VOV_VDL_XTR_FRM_001	Option : Garantie VOV et vandalisme extérieur pour formule 1 (O/N)
TOP_PCK_PLN_001	Option : Pack Plein air I pour formules 1, 2 ou 4 (O/N)
TOP_PCK_PLN_002	Option : Pack Plein air II pour formules 1, 2 ou 4 (O/N)
TOP_PCK_NRG_RNV_001	Option : Pack Energies renouvelables I pour formules 1, 2 ou 4 (O/N)
TOP_PCK_NRG_RNV_002	Option : Pack Energies renouvelables II pour formules 1, 2 ou 4 (O/N)
TOP_PCK_MBL	Option : Pack mobilité pour formules 1 ou 2 si résidence principale (O/N)
TOP_PCK_PAN_FRM_002	Option : Pack Pannes pour formule 2 (O/N)
TOP_INH_PLUS_120_JJ	Disposition : Inhabitation >120 jours (O/N)
TOP_DPC_MAT_LGR	Disposition : Dépendances en matériaux légers
TOP_BTM_1ER_CAT	Disposition : Bâtiment de 1ère catégorie
TOP_RQP_ANF	Extension : Rééquipement à neuf
TOP_DBL_PFD	Disposition : doublement du plafond
NUM_JJ_ECH_PPL	Numéro jour échéance principale du contrat
NUM_MM_ECH_PPL	Numéro mois échéance principale du contrat
COD_STT_CTR	Statut du contrat

Annexe C : Tableaux récapitulatifs des résultats après sous-échantillonnages

Les trois tableaux suivants regroupent les résultats des tests de sous-échantillonnage pour les trois modèles utilisés. Les cellules en jaunes pour les modèles d'arbre de décision et Gradient Boosting représentent les maximums respectifs de la moyenne de l'AUC.

	Pourcentage d'individus négatifs sélectionnés (Modèle Logit)								
Itération	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
1	0,6653	0,6662	0,6663	0,6657	0,6660	0,6655	0,6659	0,6656	0,6657
2	0,6658	0,6670	0,6654	0,6660	0,6656	0,6658	0,6658	0,6658	0,6659
3	0,6658	0,6649	0,6655	0,6658	0,6657	0,6656	0,6656	0,6660	0,6660
4	0,6652	0,6662	0,6659	0,6649	0,6661	0,6660	0,6658	0,6656	0,6654
5	0,6653	0,6647	0,6654	0,6660	0,6651	0,6660	0,6660	0,6661	0,6659
6	0,6659	0,6655	0,6658	0,6655	0,6656	0,6658	0,6658	0,6657	0,6656
7	0,6650	0,6658	0,6659	0,6660	0,6656	0,6661	0,6658	0,6658	0,6659
8	0,6638	0,6654	0,6663	0,6653	0,6663	0,6656	0,6654	0,6657	0,6658
9	0,6648	0,6648	0,6660	0,6660	0,6660	0,6660	0,6660	0,6653	0,6658
10	0,6655	0,6662	0,6663	0,6653	0,6653	0,6655	0,6654	0,6660	0,6657
Moyenne	0,6652	0,6657	0,6659	0,6656	0,6657	0,6658	0,6657	0,6658	0,6658
Variance	0,00000040	0,00000053	0,00000014	0,00000016	0,00000013	0,00000005	0,00000005	0,00000005	0,00000003

	Pourcentage d'individus négatifs sélectionnés (Modèle Arbre de décision)								
Itérations	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
1	0,630	0,621	0,626	0,643	0,645	0,600	0,616	0,625	0,607
2	0,630	0,621	0,631	0,630	0,646	0,601	0,620	0,624	0,608
3	0,634	0,620	0,631	0,636	0,646	0,603	0,617	0,625	0,609
4	0,630	0,621	0,631	0,637	0,641	0,600	0,617	0,624	0,608
5	0,630	0,621	0,623	0,643	0,641	0,604	0,617	0,623	0,607
6	0,627	0,621	0,631	0,640	0,646	0,601	0,617	0,624	0,608
7	0,630	0,620	0,622	0,633	0,644	0,601	0,619	0,624	0,607
8	0,634	0,620	0,623	0,643	0,641	0,608	0,618	0,632	0,608
9	0,627	0,620	0,622	0,631	0,645	0,600	0,617	0,624	0,608
10	0,630	0,620	0,623	0,632	0,641	0,600	0,618	0,630	0,607
Moyenne	0,630	0,621	0,626	0,637	0,644	0,602	0,618	0,625	0,608
Variance	0,0000057	0,0000001	0,0000166	0,0000279	0,0000051	0,0000061	0,0000010	0,0000090	0,0000004

	Pourcentage d'individus négatifs sélectionnés (Modèle Gradient Boosting)								
Itération	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
1	0,661	0,681	0,671	0,678	0,682	0,670	0,676	0,682	0,676
2	0,671	0,670	0,678	0,680	0,678	0,670	0,664	0,679	0,675
3	0,673	0,671	0,674	0,674	0,673	0,679	0,671	0,675	0,677
4	0,669	0,681	0,677	0,670	0,670	0,683	0,682	0,684	0,674
5	0,663	0,676	0,666	0,676	0,686	0,683	0,678	0,680	0,678
6	0,681	0,678	0,675	0,673	0,681	0,671	0,679	0,666	0,663
7	0,672	0,665	0,675	0,677	0,679	0,682	0,670	0,668	0,693
8	0,668	0,684	0,679	0,678	0,674	0,678	0,680	0,668	0,676
9	0,667	0,673	0,695	0,673	0,675	0,681	0,672	0,675	0,667
10	0,667	0,680	0,627	0,661	0,662	0,683	0,647	0,676	0,673
moyenne	0,669	0,676	0,672	0,674	0,676	0,678	0,672	0,675	0,675
Variance	0,0000304	0,0000356	0,0002997	0,0000308	0,0000469	0,0000316	0,0001085	0,0000414	0,0000590