Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA

et l'admission à l'Institut des Actuaires

Par : PERINI Hugo

Titre : Modélisation de la tarification d'un contrat santé issu de la gamme « retraite »

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

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

Membre présents du jury de l'Institut des Actuaires signature

Entreprise : LA MUTUELLE VERTE

Signature :

Membres présents du jury de l'ISFA

Directeur de mémoire en entreprise :

Nom : GIRAUDO

Signature : Giraudo.

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

Marie Madeleine DUBEZ

Signature du candidat

RÉSUMÉ

Mots-clés :

Tarification, Modèles linéaires généralisés, GLM, Apprentissage statistique, Arbres, Boosting, Random Forest, Santé, XgBoost.

Problématique :

Le marché de l'assurance Santé en France est de plus en plus soumis à des réglementations qui modifient l'activité de l'assureur. Ces évolutions réglementaires en faveur des assurés, rendent le marché de l'assurance Santé très concurrentiel. Dans ce contexte, les assureurs doivent proposer des garanties adaptées à la sinistralité des individus et aux tarifs les plus fins.

La réalisation d'un tarif en assurance s'appuie généralement sur l'analyse de la prime pure dans le cadre d'un modèle « Fréquence x Coût moyen » dans lequel l'effet des variables explicatives sur le niveau du risque est modélisé par des modèles de régression de type Modèle Linéaire Généralisé (*GLM)*. Ces dernières années, l'amélioration des performances informatiques conduit à un intérêt pour des approches alternatives, non paramétriques ou semi-paramétriques. L'objet de notre étude est de présenter différentes méthodes de tarification en assurance santé et de les comparer sur la gamme « Retraite » de La Mutuelle Verte.

Prise de connaissance des données à disposition et leurs retraitements :

Les données disponibles proviennent de la gamme « Retraite » de La Mutuelle Verte pour les survenances 2019, 2020 et 2021. Cette gamme est composée de quatre garanties, commercialisée sur l'ensemble du territoire français. Deux bases de données sont utilisées :

- La base « effectif » contenant les données administratives des adhérents (âge du bénéficiaire, département de l'adhérent, …)
- La base « sinistre » répertoriant tous les actes de consommations réalisés par les adhérents ayant souscrit à l'une des garanties de la gamme.

Les bases de données ont ensuite été fiabilisées :

- Suppression de valeurs aberrantes,
- Vérification de valeurs manquantes,
- …

Pour compléter l'analyse, une variable « exposition » a été ajoutée. En effet, les contrats d'assurance sont annuels, cependant certaines données peuvent être censurées dans la base (radiation par exemple).

Par ailleurs, les sinistres ont également été retraités en « as if » afin de prendre en compte la dérive médicale.

Modélisations :

La méthodologie utilisée pour calculer la prime pure est l'approche « Fréquence x Coût moyen ». Nous avons opté pour une tarification globale et non acte par acte, afin de vérifier le gain de temps et de

précision permis par l'apport de nouvelles techniques numériques dans l'analyse des tarifs. Afin de comparer le pouvoir prédictif des variables, la base de données a été scindée en deux :

- La base d'apprentissage représentant 80% des données,
- La base de test avec les 20% restant.

Les modélisations ont été réalisées sur la base d'apprentissage. Quant à elle, la base de test a permis de vérifier la qualité des modèles et de comparer les prédictions entre les différentes modélisations effectuées.

1. Modélisation de la prime pure via les Modèles Linéaires Généralisés :

Dans un premier temps nous avons procédé à un calcul de la prime pure par modèle linéaire généralisé (*GLM*). Pour cela nous choisissons le « log » comme fonction de lien afin d'obtenir un modèle multiplicatif. De plus, une analyse de l'adéquation des données à une loi de probabilité a été réalisée pour le modèle Coût et pour le modèle Fréquence :

- Le modèle Fréquence fait apparaitre de la surdispersion. A cet effet, la loi *Binomiale Négative* sera utilisée,
- Le modèle Coût permet de mettre en évidence que la loi de *Pareto* est la plus appropriée, bien qu'elle ne permette pas l'utilisation d'un modèle linéaire généralisé dans la mesure où ladite loi n'appartient pas à la famille exponentielle. De ce fait, malgré ses lacunes dans l'approximation des sinistres ayant un coût plus élevé, nous utiliserons la loi *Gamma* qui décrit bien les petits sinistres.

En supprimant les sinistres du poste hospitalisation, l'adéquation de la loi Gamma est acceptée. Il faudrait donc séparer les sinistres dits « attritionnels » et les sinistres « graves » dans une étude classique du tarif. L'objectif de notre étude est donc double ; comparer différentes modélisations de la prime pure et vérifier la possibilité d'une réalisation de tarification globale par ces autres méthodes.

Une fois les lois choisies, les modèles sont implémentés en prenant soin de :

- Vérifier l'absence de corrélation entre les variables explicatives,
- Réaliser une sélection des variables explicatives,
- Optimiser les résultats des *GLM* en regroupant certaines variables non significatives.

2. Modélisation de la prime pure par *Random Forest* :

Les *Random Forest* consistent à utiliser plusieurs arbres de décision pour en faire des « forêts ». Le principe de l'algorithme est de chercher pour chaque scission, non pas la meilleure scission parmi toutes les variables explicatives (« n »), mais la meilleure scission pour « p » variables explicatives, tirée aléatoirement parmi « n ».

Une optimisation des paramètres est effectuée afin d'améliorer le modèle.

3. Modélisation de la prime pure avec l'algorithme *XGboost*

L'algorithme *XGBoost* est un algorithme ensembliste agrégeant des arbres de décision. A chaque itération, l'arbre construit apprend de l'erreur de son prédécesseur et la corrige dans le sens du gradient.

De la même manière, une optimisation des paramètres est réalisée afin d'améliorer le modèle. La méthode du *V-fold* est également utilisée afin d'éviter le phénomène de surapprentissage.

4. Comparaison des méthodes

La comparaison finale des modèles se fait sur l'échantillon de test, jusqu'à présent inutilisé. Cette étape a pour but de déterminer quel modèle est le meilleur pour répondre aux objectifs de l'étude.

Un premier point de comparaison repose sur le pouvoir prédictif associé à chaque modèle. Les modèles optimisés sont utilisés pour prédire la prime pure de l'échantillon de test. Les résultats obtenus sont présentés dans le tableau ci-dessous :

Type de modélisation	RMSE - Modèle Coût de sinistre	RMSE - Méthode Fréquence
GLM	24,14	57,82
Random Forest	23,62	57,91
XgBoost	23,52	57,77

Tableau 1 – Tableau récapitulatif des RMSE par modèles et par type de modélisation

Dans un premier temps, nous constatons que le pouvoir prédictif des modèles est assez similaire notamment sur la méthode Fréquence. Le modèle *GLM* de Coût est légèrement en deçà en raison du manque d'optimisation de la loi *Gamma*. Toutefois, le recours à l'utilisation d'autres méthodes ne permet pas d'améliorer de manière significative (moins de 1%) le modèle *GLM*. En outre, compte tenu du nombre d'adhérents sur la gamme « Retraite » le gain est marginal.

Un autre facteur important à prendre en compte est le temps de calcul. En raison de l'immédiateté de ses résultats le modèle GLM n'est pas concerné par cette problématique. En revanche, les algorithmes de *Random Forest* et de *XgBoost* nécessitent un *tuning* des hyperparamètres, une méthodologie qui s'avère longue (environ 8h par modèle *XgBoost*) et donc coûteuse pour l'entreprise.

Conclusion :

En réalisant cette étude nous constatons que la base de données nous fournit que peu d'informations sur les assurés. Il est donc nécessaire pour les compagnies d'assurance d'obtenir des informations pertinentes sur les assurés pouvant être utiles à la tarification d'un contrat. Enfin, dans l'optique de compléter leurs modèles, les assureurs peuvent avoir recours à l'ajout de variables exogènes.

Les résultats obtenus montrent que les algorithmes de *Machine Learning* peuvent améliorer les performances d'un *GLM* sans pour autant être significativement meilleurs (dans notre cas). De plus, compte tenu du temps de calcul et d'optimisation important que nécessitent les algorithmes de *Machine Learning*, nous estimons que, à ce niveau de segmentation, leur utilisation n'est pas justifiée. Néanmoins, dans le cas où nous disposerions d'une base de données plus conséquente, il est probable que les résultats obtenus à partir des algorithmes de *Machine Learning* soient bien meilleurs que les résultats obtenus par *GLM*. La précision de ces résultats justifierait alors un temps de calcul plus conséquent.

L'approche d'une tarification globale ne nous semble pas pertinente dans la mesure où nous gardons les modèles GLM. En effet, pour le modèle Coût Moyen nous avions constaté que la loi de Gamma n'était pas optimale en tenant compte du fait qu'elle sous-estime les sinistres graves. En conclusion, il est nécessaire de segmenter les modèles par famille d'acte afin d'être plus pertinent dans l'approche tarifaire.

ABSTRACT

Keywords :

Pricing, Generalized linear models, GLM, Statistical learning, Trees, Boosting, Random Forest, Health, XgBoost.

Problem :

The health insurance market in France is increasingly subject to regulations that modify the activity of the insurer. These regulatory changes in favor of policyholders make the health insurance market very competitive. In this context, insurers must offer guarantees adapted to the loss experience of individuals and at the lowest prices.
The realization of a pricing is classically based on the analysis of the pure premium within the framework of a frequency x cost model in which the effect of the explanatory variables on the level of risk is modeled by GLM regression models. The improvement in computer performance has led in recent years to an interest in alternative, non-parametric or semi-parametric approaches. The purpose of our study is to present different pricing methods in health insurance and to compare them on the "Retirement" range of La Mutuelle Verte.

Consideration of the data available and adjustments :

The data available comes from the "retirement" range of La Mutuelle Verte for occurrences in 2019, 2020 and 2021. This range is made up of four guarantees and marketed throughout France. Two databases are used:

- The "workforce" database containing the administrative data of the insured (age of the beneficiary, localization of the insured, etc.)
- The "claims" database listing all acts of consumption carried out by insured who have subscribed to one of the guarantees in the range.

The databases were then made more reliable:

- Removal of outliers,
- Missing value check,
- …

To complete the analysis, an exposure variable was added. Indeed, the insurance contracts are annual, but in the database, some data may be censored (deregistration for example).
In addition, the claims have also been restated in "as if" in order to take into account medical drift.

Modelings:

The methodology used to calculate the pure premium is the Frequency * Cost approach. We opted for global pricing and not an act by act pricing in order to verify whether the contribution of new digital techniques made it possible to gain in precision and time in the analysis of prices. In order to be able to compare the predictive power of the variables, the database was split into two:

- The train base representing 80% of the data.
- The test base with the remaining 20%.

The models were carried out on the train base and the test base made it possible to check the quality of the models and to compare the predictions between the different models carried out.

1. Pure premium modeling by Generalized Linear Models:

First, we proceeded with a calculation of the pure premium by generalized linear model (GLM). For this we choose the "log" as a link function to have a multiplicative model. In addition, an analysis of the adequacy of the data to a law of probability was carried out for the Cost model and for the Frequency model.

- The Frequency model makes an appearance of the overdispersion. For this purpose, the negative binomial law will be used,
- The Cost model shows that Pareto's law would be the most appropriate. However, this law does not allow the use of a generalized linear model insofar as the Pareto law does not belong to the exponential family. Therefore, we will use the Gamma law which describes small claims well, but which does not allow us to properly approximate claims with a higher cost.

For information, by removing claims from the hospitalization item, the adequacy of the Gamma law is accepted. So-called "attritional" claims and "serious" claims should therefore be separated in a classic tariff study. The objective of our study will be twofold, it will make it possible to compare different models of the pure premium but also to check whether we can achieve via these other methods a global pricing and not a care post (or per act).
Once the probabilities laws have been chosen, the models have been implemented taking care:

- To verify the absence of correlation between the explanatory variables
- Select explanatory variables.
- To optimize the results of the GLMs by grouping together some non-significant variables.

2. Pure premium modeling by Random Forest:

Random Forests consist of using several decision trees to make "forests" of them. The principle of the algorithm is to seek, for each split, not the best split among all explanatory variables ("n"), but the best split for "p" explanatory variables drawn randomly among "n". An optimization of the parameters is carried out to improve the model.

3. Pure premium modeling by XGboost algorithm

The XGBoost algorithm is a set algorithm aggregating decision trees. At each iteration, the constructed tree learns from the error of its predecessor and corrects it (in the sense of the gradient). In the same way, an optimization of the parameters has been carried out to improve the model. The V-fold method is also used (to avoid overfitting problem).

4. Comparison of methods

The final comparison of the models is done on the test sample, so far unused. This step aims to determine which model is the best to meet the objectives of the study.
A first point of comparison is the predictive power associated with each model. The optimized models are used to predict the pure premium of the test sample. The results obtained are presented in the table below:

Modeling type	RMSE - Average cost method	RMSE - Frequency method
GLM	24,14	57,82
Random Forest	23,62	57,91
XgBoost	23,52	57,77

Table 2 – Summary table of RMSE by model and by type of modeling

Initially, we note that the predictive power of the models is quite similar, especially for the frequency method. The Cost model GLM is slightly below due to the use of the Gamma distribution which is not optimal. However, the use of other methods does not significantly improve (less than 1%) the GLM model and given the number of insured in the "retirement" range, the gain is marginal.
Another factor to consider is the calculation time. The question does not arise for the GLM model because the results are immediate. However, the Random Forest and XgBoost algorithms require hyperparameter tuning. This methodology is long (about 8 hours for XgBoost models) and therefore expensive for the company.

Conclusion :

By carrying out this study, we realized that the database provided us with little information on the insured. It is necessary for insurance companies to obtain relevant information on the insured which could be useful for the pricing of a contract. To complete this point, insurers can use the addition of exogenous variables to complete their models.

The results obtained show that Machine Learning algorithms can improve the performance of a GLM without being significantly better (in our case). Moreover, given the significant calculation and optimization time required by Machine Learning algorithms, we believe that, at this level of segmentation, their use is not justified. Nevertheless, if we have a larger database, it is likely that the results obtained from Machine Learning algorithms will be much better than the results obtained by GLM. The accuracy of these results would then justify a more substantial calculation time.

The global pricing approach does not seem relevant to us insofar as we would keep the GLM models. However, for the Average Cost model, we found that the Gamma law was not optimal given the fact that it underestimates large claims. Thus, it is necessary to segment the models by family of act in order to be more relevant in the pricing approach.

REMERCIEMENTS

Tout d'abord, je tiens à remercier La Mutuelle Verte et plus particulièrement Madame REINAUD Valérie, Directrice du Développement ainsi que Madame GIRAUDO Caroline Responsable du Contrôle Interne et présente tutrice, pour la confiance et le soutien apportés tout au long de la réalisation de ce mémoire.

Je souhaite par ailleurs remercier Madame CHAMPAGNE DE LABRIOLLE Caroline, tutrice universitaire, pour le suivi assidu de mon mémoire et ses multiples conseils. J'en profite également pour saluer la qualité des enseignements dispensés par l'ensemble du corps enseignant de l'IFSA au cours de ces trois années de formation.

Enfin, j'adresse un grand merci à tous mes proches pour m'avoir soutenu dans ce projet, avec une pensée plus singulière pour ma femme et ses (nombreuses) relectures.

SOMMAIRE

INTRODUCTION

Le système de santé en France est en constante évolution et les organismes complémentaires d'Assurance Maladie doivent faire face aux changements successifs de réglementation. Pour être compétitifs et maîtriser leur activité, les organismes d'assurance doivent connaître de manière précise la charge future en soins d'une population assurée, qu'elle soit nouvelle ou déjà couverte par l'organisme. Pour cela, elle doit se doter d'outils de tarification performants. L'objet de notre étude est de présenter différentes méthodes de tarification en assurance santé et de les comparer.

A cette fin, tour à tour quatre grandes parties sont présentées. La première se consacre au système d'assurance santé français ainsi qu'à ses dernières reformes impactantes afin de mettre en lumière le caractère évolutif du monde de la santé. La deuxième partie détaille quant à elle les principes de tarification en assurance santé. En effet, ce cadre plus théorique est nécessaire à la compréhension des différentes méthodes qui seront implémentées par la suite. La réalisation d'un tarif en assurance s'appuie généralement sur l'analyse de la prime pure dans le cadre d'un modèle Fréquence x coût dans lequel l'effet des variables explicatives sur le niveau du risque est modélisé par des modèles de régression de type *GLM*. Ces dernières années, l'amélioration des performances informatiques ont conduit à un intérêt pour des approches alternatives, non paramétriques ou semi-paramétriques, présentées plus en détail ci-après. La troisième partie porte sur la description du portefeuille étudié : la gamme « Retraite » de La Mutuelle Verte. Enfin, au cours de la quatrième et dernière partie, une application des modèles précédemment présentés ainsi qu'une analyse comparative des résultats seront menées. Cette analyse aura pour objectif de démontrer la capacité prédictive des différents modèles implémentés.

I – L'ASSURANCE MALADIE COMPLÉMENTAIRE : ÉLEMENTS DE CONTEXTUALISATION

La France est caractérisée par un système de couverture santé universelle qui prône l'accès aux soins pour tous. Le financement du système de remboursement des frais médicaux s'appuie sur une structure à deux niveaux : l'Assurance Maladie Obligatoire et l'Assurance Maladie Complémentaire. La Sécurité Sociale constitue le premier niveau de protection sociale via l'assurance maladie (dit Assurance Maladie Obligatoire). Le second niveau est atteint grâce à l'Assurance Maladie Complémentaire soit via des contrats de complémentaire santé individuels ou collectifs. Différents acteurs se partagent ce marché : les sociétés d'assurance, les mutuelles et les institutions de prévoyance. Les enjeux (sanitaire, politique, …) sont importants, en conséquence ce marché est en constante évolution avec notamment la mise en place des contrats responsables et du 100% santé.

1. L'Assurance Maladie Obligatoire

L'assurance maladie est une composante de la Sécurité Sociale[1]. Cette branche assure la prise en charge des dépenses de santé des assurés et garantit l'accès aux soins. Toutefois, différents organismes interviennent pour le remboursement des frais de santé.

2.1. Organisation de l'Assurance Maladie Obligatoire

Divers organismes sont en charge des remboursements de frais de santé en fonction des régimes obligatoires des assurés :

- La Caisse Nationale d'Assurance Maladie (CNAM) gère les assurés du régime Général et du Régime Social des Indépendants (RSI),
- La Mutualité Sociale Agricole (MSA) pour les salariés du régime agricole. La MSA verse toutes les prestations (maladie, retraite, famille),
- Les salariés relevant des régimes spéciaux sont soit gérés par leur régime soit par la CNAM.

L'Union nationale des caisses d'assurance maladie (UNCAM), instance créée depuis 2004, regroupe les trois principaux régimes d'Assurance Maladie Obligatoire. Elle définit non seulement les prestations de santé admises au remboursement mais également leur taux de prise en charge. De manière plus globale, l'UNCAM se charge de la promotion des actions de santé publique et négocie avec les professionnels de santé afin de garantir la qualité des soins.

2.2. Le régime local d'Alsace-Moselle

Le régime local d'Alsace-Moselle est le produit de l'histoire mouvementée des départements de la Moselle, du Bas-Rhin et du Haut-Rhin. Ces territoires étaient rattachés à l'Empire Allemand entre 1871

[1] Créée en 1945, la Sécurité Sociale « *est la garantie donnée à chacun qu'en toutes circonstances il disposera des moyens nécessaires pour assurer sa subsistance et celle de sa famille dans des conditions décentes* » ordonnance du 4 octobre 1945, texte fondateur

et 1918. Ainsi, la population bénéficiait du fonctionnement de l'Assurance maladie mis en place par les lois de Bismarck. Ces lois permettaient aux assurés de jouir d'une grande prise en charge de leurs dépenses de soins. En 1946, à la suite de la Seconde Guerre mondiale, cette règlementation a été maintenue, dans l'attente d'un alignement avec le régime général de la France. Celui-ci n'a jamais été réalisé et la loi du 31 décembre 1991 a permis d'officialiser et de pérenniser le régime local d'Alsace-Moselle.

La principale différence entre les assurés du régime local d'Alsace-Moselle et du régime général réside dans le fait que les assurés du régime local payent des montants de cotisation plus élevés. Ces montants de cotisation sont contrebalancés par un meilleur taux de remboursement par rapport au Régime Général.

Voici un tableau récapitulatif des différences de prise en charge du régime local et du Régime Général:

Acte médical	Remboursement du régime local	Remboursement du régime général
Hospitalisation	100 %	80 %
Consultation médicale (spécialiste et généraliste)	90 %	70 %
Soins infirmiers, analyses médicales et kinésithérapie	90 %	60 %
Médicaments vignettes blanches	90 %	65 %
Médicaments vignettes bleues	80 %	35 %
Frais de transport	100 %	65 %

Tableau 3 - Détails des prises en charge du régime local et du régime général

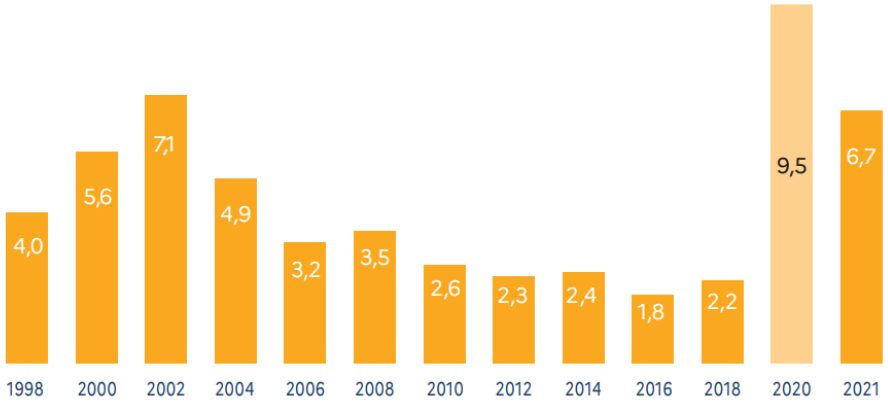
Ces meilleurs remboursements permettent aux assurés de la région Alsace-Moselle de souscrire à un contrat de complémentaire santé moins onéreux. En effet, l'organisme de complémentaire santé devra fournir un effort moindre pour compenser les remboursements d'un Régime Obligatoire plus avantageux

Le portefeuille étudié dans ce document ne présente pas d'adhérent affilié au régime local. Ainsi il ne sera pas réalisé d'analyse particulière sur ce régime.

2.3. Chiffres clés (publication juin 2021)

Pour rappel, la Caisse nationale d'assurance maladie (CNAM) gère la branche maladie du régime général de la Sécurité Sociale et pilote le réseau des caisses primaires d'assurance maladie (CPAM). La CNAM finance 92 % de l'ensemble des dépenses d'assurance maladie. Ainsi les données présentées dans le bilan annuel de la Sécurité Sociale concernent exclusivement ce périmètre.

Dans un premier temps, le rapport présente la progression annuelle des dépenses d'assurance maladie en pourcentage. L'évolution des dépenses annuelles était « maitrisée » jusqu'en 2020, date du début de la pandémie Covid 19. Cette pandémie a entrainé une hausse significative des dépenses de santé, (pouvant se traduire comme « la » preuve des actions menées par le gouvernement pendant cette période). Comparativement à 2020, la progression des dépenses a diminué en 2021 toutefois le niveau de dépense reste extrêmement élevé si on se réfère aux 15 dernières années.

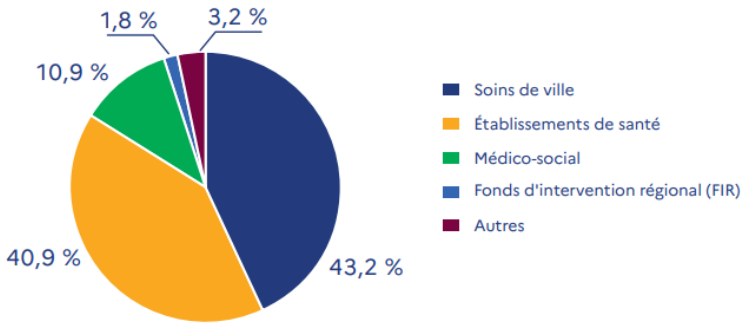


Pour 2020, le taux d'évolution est y compris impacts lié à la crise du Covid-19. Initialement le taux prévu était de 2,45 %. Pour 2021, le taux, en cohérence avec la CCSS de juin 2021 tient compte du Ségur de la santé et des surcoûts générés par la crise.

Source : Commission des comptes de la s écurité sociale, juin 2021.

Figure 1 - Progression annuelle des dépenses d'assurance maladie

Le graphique suivant présente la répartition des dépenses. Ainsi les soins de ville et les établissements de santé représentent 84,1% des dépenses du régime général d'Assurance Maladie. Par ailleurs, les dépenses de soins de ville regroupent les honoraires des professionnels de santé libéraux, les prestations en espèces (indemnités journalières), les dépenses ambulatoires de médicaments et de dispositifs médicaux, ainsi que les transports sanitaires.



Source : Commission des comptes de la sécurité sociale, juin 2021.

Figure 2 - Dépenses de santé financées par l'Assurance maladie (Ondam, estimation pour 2019)

Le rapport détaille également le déficit lié à la branche maladie du régime général (représentant 92 % de l'ensemble des dépenses d'Assurance Maladie Obligatoire).

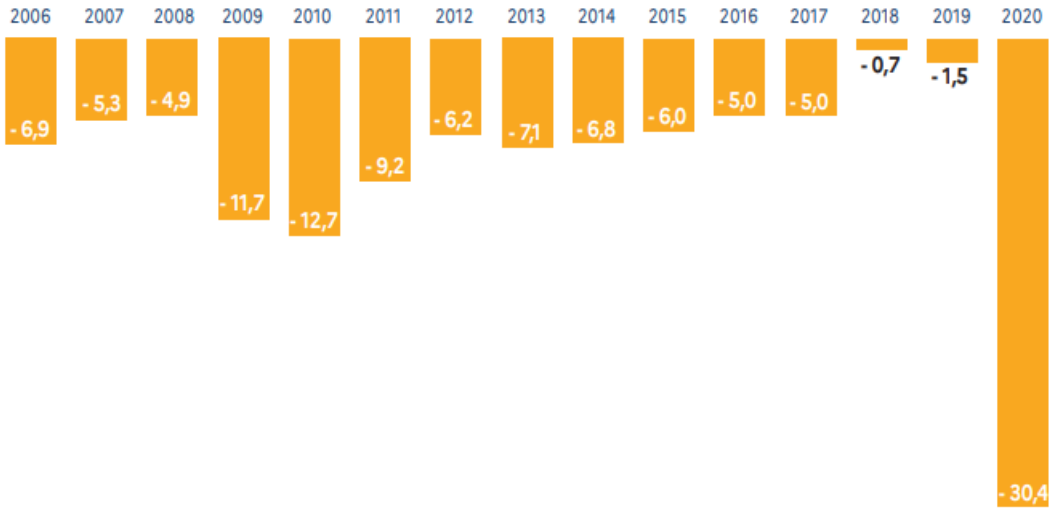


Figure 3 - Évolution du solde de la branche maladie en milliards d'euros

L'impact des actions du gouvernement en 2020, lié à la pandémie Covid-19, est particulièrement marqué dans l'équilibre de la branche maladie. Le régime Obligatoire est intervenu pour les remboursements de santé des Français à hauteur de 100% du coût de l'acte. Cette forte implication est visible sur le graphique, le déficit engrangé dépasse celui des dernières années.

Les frais de santé ne sont pas totalement pris en charge par les Régimes Obligatoires. De ce fait, le système de santé français permet l'utilisation d'Assurance Maladie Complémentaire afin d'augmenter la prise en charge des frais médicaux.

2. L'Assurance Maladie Complémentaire

Dès qu'un adhérent est affilié à un Régime Obligatoire il bénéficie d'un remboursement partiel de ses dépenses de santé. Cette prise en charge correspond à un taux du tarif conventionnel (barème fixé par l'Assurance maladie). Il lui incombe alors de régler la somme restante, appelée « ticket modérateur ». Pour certaines prestations (équipement optique, soins dentaires, consultations avec dépassement d'honoraires…), la facture peut vite s'avérer élevée. La complémentaire santé vient alors compléter les garanties de base. Elle prend en charge, partiellement ou en totalité, les actes non remboursés par l'Assurance maladie et ceux qui le sont très faiblement, pour assurer une couverture plus optimale. Différents acteurs se partagent ce secteur très concurrentiel.

2.1. Présentation du marché des complémentaires santé

2.1.1. Les sociétés d'assurance

Les sociétés d'assurance peuvent présenter l'une des deux formes juridiques suivantes :

- Les sociétés anonymes avec un statut de société commerciale,
- Les sociétés d'assurances mutuelles qui sont des sociétés à but non lucratif.

Les sociétés d'assurances anonymes par action ont vocation à réaliser des bénéfices et à les distribuer à leurs actionnaires. Les sociétés d'assurance commercialisent des assurances de biens, de responsabilité et de personnes soit par l'intermédiaire d'agents généraux ou courtiers, soit par vente directe. Elles interviennent dans tous les domaines de l'assurance : prévoyance, santé, assurance de biens ou encore retraite.

2.1.2. Les institutions de prévoyance

Les institutions de prévoyance sont des personnes morales de droit privé ayant un but non lucratif. Elles sont gérées paritairement et couvrent les risques de maladie, d'incapacité de travail et d'invalidité, de dépendance et de décès. Cette parité implique que les représentants des salariés et des employeurs gèrent, lors des conseils d'administration, la gestion et l'évolution des garanties proposées aux salariés de l'entreprise.
Les institutions de prévoyance sont régies par le code de la Sécurité Sociale, soumises aux mêmes règles techniques que chaque entreprise d'assurance. L'essentiel de leur activité relève de la Prévoyance collective (Prévoyance et Santé à 50/50). Elles ont une forte implication dans le secteur de la Prévoyance notamment grâce aux anciennes clauses de désignation présentes dans les Conventions Collectives.
Le nombre actuel d'institutions de prévoyance affilié au CTIP (centre technique des institutions de prévoyance), qui représente et défend les intérêts des institutions de prévoyance, s'élève à environ 38 organismes en 2021[2].

[2] Chiffre issue du *CTIP CAHIER STATISTIQUE 2021*

2.1.3. Les mutuelles

Les mutuelles sont des sociétés à but non lucratif et régies par le Code de la mutualité.
L'article L111-1, alinéa 1 du code de la mutualité les définit comme : « des personnes morales de droit privé à but non lucratif. Elles sont soumises aux dispositions du présent code à dater de leur immatriculation au registre national des mutuelles. Elles mènent notamment au moyen de cotisations versées par leurs membres, et dans l'intérêt de ces derniers et de leurs ayants droit, une action de prévoyance, de solidarité et d'entraide, dans les conditions prévues par leurs statuts afin de contribuer au développement culturel, moral, intellectuel et physique de leurs membres et à l'amélioration de leurs conditions de vie. ».

Elles sont reparties en trois catégories de livre :

- Livre III : les mutuelles gérant des réalisations sanitaires et sociales,
- Livre II : les mutuelles exerçant une activité d'assurance,
- Livre I : les mutuelles ne rentrant ni dans la catégorie II, ni dans la catégorie III.

Les mutuelles proposent principalement des assurances complémentaires santé, de la prévoyance et des assurances complémentaires retraite.
Une mutuelle se distingue d'une société d'assurance par son fonctionnement égalitaire : chaque adhérent possède une voix dans les délibérations. Il n'existe pas de capital social et donc pas d'actionnaire, il s'agit de la grande différence entre une société de personnes, dont font partie les mutuelles, et les sociétés en participation de capital.
La Fédération Nationale de la Mutualité Française dénombrait 462 mutuelles en 2021 contre 7 500 dans les années 1980. L'augmentation des exigences règlementaires, avec notamment la mise en place de la Directive Solvabilité II, a accélérer d'avantage la tendance de réduction du nombre de mutuelles.

2.2. Les différents contrats proposés

2.2.1. Contrat collectif adhésion obligatoire

Un contrat collectif est un contrat d'assurance souscrit par une personne morale ou une entreprise dans le but de faire adhérer des salariés pour les couvrir contre les risques de maladie, d'incapacité de travail et de décès.

Le contrat collectif peut résulter d'un accord de branche ou d'entreprise, d'un référendum ou d'une décision unilatérale de l'employeur.

Depuis le 1er janvier 2016, les entreprises du secteur privé ont l'obligation de proposer à tous leurs salariés une couverture santé complémentaire. Cette décision fait suite à l'Accord National Interprofessionnel du 14 juin 2013. Il est également prévu que l'employeur prenne à sa charge au minimum 50% de la cotisation. Selon les conditions du contrat, la complémentaire peut également bénéficier aux ayants droit du salarié. La mise en place d'une complémentaire obligatoire d'entreprise peut se faire de trois façons :

- Par accord collectif : en négociant avec les organisations syndicales représentées dans l'entreprise. Il peut également être décidé par la branche professionnelle,
- Par référendum au sein de l'entreprise,

- Par décision unilatérale de l'employeur.

Les salariés ont l'obligation de souscrire à cette complémentaire. Certains cas de dispense sont toutefois prévus (par exemple, si le salarié est déjà couvert par le contrat complémentaire obligatoire de son conjoint).

Le principal avantage du contrat collectif est de mutualiser des risques similaires ce qui va permettre aux salariés de bénéficier de tarifs compétitifs.

2.2.2. Contrats individuels

Ce type de contrat s'adresse majoritairement aux personnes n'étant pas couvertes par un contrat collectif. Comme par exemple les étudiants qui ne sont pas couverts par la complémentaire de leurs parents, les chômeurs, ... Il est également possible pour les salariés du privé, déjà couverts par la complémentaire obligatoire de leur entreprise, de souscrire à une surcomplémentaire individuelle.

2.2.3. Contrats Madelin (travailleurs non-salariés)

La Loi Madelin du 11 février 1994 favorise la protection sociale des indépendants qui ne bénéficie pas des mêmes dispositions de protection sociale que les salariés. Pour remédier à cela, elle vise à inciter les travailleurs non-salariés (TNS) à se constituer eux-mêmes leur protection sociale. De plus, elle permet aux travailleurs indépendants de déduire de leur revenu imposable les cotisations versées au titre d'un contrat d'assurance de complémentaire santé, de retraite, de prévoyance.

3. Fonctionnement des remboursements des soins

Différents éléments interviennent dans le remboursement d'un acte médical. Le graphique suivant permet de bien appréhender l'articulation d'une prise en charge.

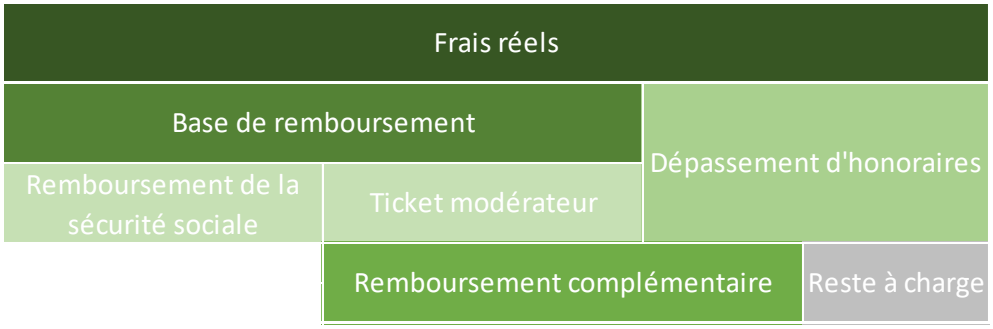


Figure 4 - Décomposition d'un remboursement

Les frais réels correspondent au prix de l'acte ou du bien médical proposé par le praticien.
La base de remboursement est le tarif de référence qui permettra de déterminer par la suite le montant qui sera remboursé pour un acte donné.
Le remboursement de la Sécurité Sociale est défini de la manière suivante :
Remboursement Sécurité Sociale = Base de remboursement × Taux de remboursement

Le ticket modérateur correspond à la part de la base de remboursement non prise en charge par la Sécurité Sociale.
Le **remboursement complémentaire** équivaut au ticket modérateur plus les dépassements d'honoraires éventuels.
Le **reste à charge**, correspond à la somme restante que l'assuré devra payer.

Exemple concret d'un remboursement de soins :
Soit un assuré qui consulte son médecin traitant, un généraliste en secteur 2[3] qui a adhéré à l'OPTAM[4]. Le tarif de convention est de 25€ auquel s'ajoutent les potentiels dépassements d'honoraires qui dans notre cas s'élèveront à 15€. L'assuré possède un contrat de complémentaire santé lui assurant un remboursement à hauteur de 100% de la base de remboursement (Sécurité Sociale incluse).
Sachant que le parcours de soins est respecté :

- Le coût total de la consultation est de 40€,
- La base de remboursement (BR) pour ce type d'acte est de 25€,
- Le remboursement de la Sécurité Sociale est de 70% BR soit $25€ \times 70\% = 17{,}5€$
- 1€ est toutefois soustrait à cette somme. Il correspond à la franchise obligatoire également appelé « participation forfaitaire ».
- La complémentaire santé prend à sa charge le ticket modérateur : 100% BR - 70% BR de la Sécurité Sociale soit 30% BR correspondant à $25€ \times 30\% = 7{,}5€$. La formule souscrite ne permet pas de prendre en charge les dépassements d'honoraires.
- Le reste à charge de l'assuré correspond donc aux dépassements d'honoraire à hauteur de 15€ et de la participation forfaitaire de 1€ soit un total de 16€.

Cas des assurés en Affection de Longue Durée (ALD) :
Les assurés touchés par des maladies graves et/ou chroniques, telles que le cancer, le VIH ou Alzheimer appelées Affections Longue Durée (ALD), bénéficient d'une prise en charge à 100% par l'Assurance Maladie. Le remboursement de la Sécurité Sociale est alors fixé à 100% de la Base de Remboursement. Toutefois cela ne suppose pas un remboursement intégral par l'Assurance Maladie : les malades atteints d'une ALD ne sont pas remboursés des dépassements d'honoraires et restent redevables de des diverses franchises médicales (comme la participation forfaitaire de 1€) et du forfait hospitalier. Ils sont en revanche exonérés du forfait de 18 euros pour les actes lourds.

4. Présentation des dernières réformes

4.1. La protection universelle maladie (PUMA)

Le 1er janvier 2016, la protection universelle maladie entre en application. Cette réforme garantit à toute personne qui travaille ou réside en France de manière stable et régulière, un droit à la prise en charge de ses frais de santé à titre personnel et de manière continue tout au long de la vie. La

[3] Les praticiens qui exercent en secteur 2 fixent eux-mêmes leurs tarifs : ils sont conventionnés à honoraires libres. L'Assurance maladie rembourse le prix de la consultation sur la base du tarif du secteur 1. On parle de praticiens pratiquant les dépassements d'honoraires.
[4] Ces dispositifs ont pour objectif d'encourager les médecins à stabiliser les dépassements d'honoraires et accroître la part des soins facturés à tarifs opposables (tarif sans dépassement d'honoraires et servant de base au remboursement de l'Assurance Maladie).

protection universelle maladie parachève ainsi la logique initiée par la couverture maladie universelle (CMU) de base de 1999, qui vise à ouvrir des droits à l'assurance maladie aux personnes résidant en France de façon stable et régulière, et qui ne relevaient d'aucune couverture maladie obligatoire.

Les objectifs de la réforme sont multiples :

- Simplifier la vie des assurés : plus besoin de justifier d'un nombre d'heures travaillées : seul l'exercice d'une activité professionnelle est pris en compte,
- Assurer la continuité des droits en cas de changement de situation personnelle / professionnelle,
- Réduire les démarches administratives,
- Garantir davantage d'autonomie et de confidentialité.

4.2. La Complémentaire santé solidaire (CSS)

La complémentaire santé solidaire (CSS) est née en novembre 2019 de la fusion de la couverture maladie universelle complémentaire (CMU-C) et de l'aide au paiement d'une complémentaire santé (ACS).

La complémentaire santé solidaire concerne les personnes couvertes par l'assurance maladie et qui disposent de faibles ressources. Le dispositif prend alors en charge la part complémentaire et les bénéficiaires sont dispensés de l'avance des frais. Selon le niveau de ressources, la Complémentaire santé solidaire peut être accordée sans participation financière soit en contrepartie d'une participation financière.

4.3. L'Aide Médicale d'Etat (AME)

Cette aide concerne les ressortissants étrangers en situation irrégulière et précaire sous condition de résidence stable (supérieure à trois mois) et de ressources (en fonction de la composition du foyer). L'aide est versée pendant un an avec une prise en charge à 100 % des soins médicaux et hospitaliers (dans la limite des tarifs de la Sécurité Sociale). Le tout avec une dispense d'avance de frais.
Certains frais de santé sont exclus du dispositif AME comme les frais de cures thermales, les soins et les médicaments liés à l'assistance médicale à la procréation ainsi que les médicaments à 15%. Pour les mineurs, les frais médicaux restent pris en charge à 100%.

4.4. La Loi ANI

L'Accord National Interprofessionnel (ANI) est un accord conclu entre les partenaires sociaux :

- Les représentants des employeurs (MEDEF, CGPME…),
- Les représentants des salariés (CGT, FO, CFDT…), auxquels sont conviées les organisations syndicales.

L'ANI concerne les entreprises dont la branche professionnelle adhère à l'une des organisations représentant les employeurs. Une fois le compromis trouvé, l'accord signé par les partenaires sociaux donne lieu à un projet de loi présenté et voté au Parlement.

Si l'accord est négocié au niveau national et concerne tous types de secteurs d'activités, on parlera d'ANI, il concerne alors l'ensemble des secteurs d'activités des entreprises du territoire national.

Deux dates sont à retenir :

- Le 11 janvier 2013 : les organisations patronales, le gouvernement et les principales confédérations syndicales ont signé l'accord national interprofessionnel sur la sécurisation de l'emploi,
- Le 14 juin 2013 : Transposition législative de l'ANI en LOI n° 2013-504. La loi sur la sécurisation de l'emploi a été promulguée.

L'accord du 11 janvier 2013 impacte de nombreux domaines en modifiant notamment :

- Le compte personnel de formation,
- Les droits rechargeables à l'assurance chômage,
- La présence des salariés dans l'organe de gouvernance,
- La généralisation de la complémentaire santé,
- Et l'encadrement des temps partiels.

C'est un nouveau modèle économique et social au service de la compétitivité des entreprises et de la sécurisation de l'emploi et des parcours professionnels.

Un cadre juridique est créé il s'appuie sur de nombreux articles dont deux concernent la santé tout particulièrement :

- L'article 1 permet à tous les salariés de bénéficier d'une couverture complémentaire santé depuis le 1er janvier 2016,
- L'article 2 qui modifie la durée de la portabilité et son financement par la mutualisation. Les signataires conviennent de généraliser le système de financement du maintien des garanties exclusivement par mutualisation. La durée maximale de la portabilité de la couverture de frais de santé et de prévoyance est portée de 9 à 12 mois.

La couverture collective santé pour tous les salariés contient 3 étapes :

- 1ère étape : ouverture des négociations de branches avant le 1er juin 2013 (uniquement recommandation d'un ou plusieurs organismes d'assurance),
- 2ème étape : à défaut d'accord de branche avant le 1er juillet 2014, obligation de négocier dans les entreprises de plus de 50 salariés (obligation annuelle de négociation prévoyance),
- 3ème étape : au 1er janvier 2016, à défaut d'accord de branche ou d'entreprise, tous les salariés doivent bénéficier d'une couverture par Décision Unilatérale de l'Employeur (DUE) dont les garanties sont fixées par la loi.

Le financement de la couverture frais de santé est partagé entre les salariés et les employeurs. L'employeur doit prendre en charge au minimum 50 % de la cotisation du socle obligatoire, le solde est versé par le salarié.

4.5. Le contrat responsable

Depuis la réforme de 2006 sur l'assurance maladie, les contrats de prévoyance complémentaire doivent, pour bénéficier d'avantages fiscaux et sociaux, proposer des prestations et des conditions de prise en charge respectant l'esprit de cette réforme. Ils doivent favoriser également le respect du parcours de soins coordonnés défini par la réglementation de la Sécurité Sociale.

Les contrats dits "responsables" doivent respecter certaines conditions. Ils doivent répondre à un cahier des charges et ne doivent notamment pas prendre en charge :

- La participation forfaitaire de 1€ par consultation ou acte médical non remboursée par la Sécurité Sociale,
- La majoration du ticket modérateur imposée au patient qui consultera un médecin sans avoir choisi de médecin traitant ou sans prescription de ce dernier,
- La majoration de la participation de l'assuré qui refusera à un professionnel de santé d'accéder ou de compléter son dossier médical personnel…

Un décret du 19 novembre 2014 a modifié les conditions à respecter par les contrats responsables en introduisant un panier de soins minimal. Les contrats responsables doivent prendre en charge :

- L'intégralité du ticket modérateur sur les consultations, actes et prestations remboursables par l'Assurance Maladie Obligatoire (à l'exception des médicaments remboursés à 30 % ou 15 %, de l'homéopathie et des cures thermales)
- Le forfait journalier hospitalier qui correspond aux frais d'hébergement d'un séjour (chambre et repas), sans limitation de durée,
- Les paniers 100 % santé en optique, prothèses, dentaires et audioprothèses.

Nouveauté 2022 : la prise en charge des consultations de psychologues adhérents au dispositif « MonPsy » :

Depuis le 5 avril 2022, le dispositif « MonPsy » permet aux personnes dès l'âge de 3 ans (enfants, adolescents et adultes) de bénéficier de séances d'accompagnement psychologique avec une prise en charge par l'Assurance maladie et les Complémentaires Santé.

Pour en bénéficier l'adhérent doit consulter un praticien référencé « Monpsy » dans l'annuaire dédié disponible sous monpsy.sante.gouv.fr/annuaire. Les psychologues partenaires sont sélectionnés par un comité d'experts sur la base de critères de formation et d'expérience, afin d'attester de leur parcours consolidé en psychologie clinique. Seuls les psychologues sélectionnés et ayant signé une convention avec l'Assurance maladie peuvent participer au dispositif « MonPsy ».

Dans le cadre du dispositif « MonPsy », l'accompagnement psychologique comprend :

- Une première séance qui est un entretien d'évaluation (facturé 40€ la séance),
- Jusqu'à 7 séances de suivi psychologique (facturé 30€ par séance) par année civile.

Concernant le remboursement, aucun dépassement d'honoraire est possible, l'Assurance Maladie Obligatoire prend en charge 60% du coût de la séance et la Complémentaire Santé complète à hauteur du ticket modérateur (soit 40%).

4.6. La réforme du 100% Santé

Pour rappel, le reste à charge est la part des dépenses de santé qui n'est couverte ni par l'Assurance Maladie Obligatoire, ni par l'Assurance Maladie Complémentaire.
L'objectif de la réforme du reste à charge zéro est d'éliminer les renoncements aux soins pour des raisons pécuniaires et donc d'améliorer l'accès à certains dispositifs de santé.
Cette réforme concerne l'optique, l'auditif et le dentaire.
Un panier de soins pour chaque secteur garantissant le remboursement intégral par la Sécurité Sociale et les complémentaires santé est mis en place.

4.6.1. Réforme du reste à charge zéro dans l'optique :

Le décret en date du 3 décembre 2018 a établi deux classes de verres et de montures :

- Les classes A[5] concernent les produits du panier « reste à charge zéro »,
- Les classes B concernent les produits avec un reste à charge.

L'opticien a l'obligation de présenter à son client un équipement de la classe A et doit exposer au minimum 35 montures de classe A pour les adultes et 20 montures pour les enfants.
Les prix libres de vente des équipements optiques sont plafonnés pour le panier A et les bases de remboursement de l'Assurance Maladie Obligatoire sont revalorisées.

	Base de remboursement (par verre)	Prix limite de vente (par verre)	Prix limite de vente équipement
Verres unifocaux	9,75 à 35,25€	32,5 à 117,5€	95 à 265€
Verres multifocaux	13,5 à 39€	45 à 130€	120 à 290€
Verres progressifs	22,5 à 51€	75 à 170€	100 à 370€
Monture	9 €	30 €	

Figure 5 - Détail de la réforme 100% santé en optique

Un renouvellement peut être effectué tous les deux ans, à l'exception des situations suivantes :

Renouvellement des équipements	
Âge de la personne	**Période de renouvellement**
16 ans et plus	2 ans
Plus de 6 ans et moins de 16 ans	1 an
Jusqu'à 6 ans	6 mois

Figure 6 – Possibilité de renouvellement des équipements optique (1/2)

[5] *Prise en charge intégrale des équipements de classe A (panier 100 % santé) : Ticket modérateurs et dépassements de tarifs à hauteur des Prix libre de vente fixés.*

Situation	Délai de renouvellement	Prescription
Dégradation des performances oculaires de l'équipement	1 an pour les 16 ans et plus Pas de délai pour les moins de 16 ans	Prescription médicale ou renouvellement avec adaptation de l'opticien (prescription de moins de 3 ans)
Conditions médicales particulières : Glaucome, hypertension intraoculaire, DMLA, cataracte évolutive, amblyopie, etc.	Pas de délai minimal	Prescription par un ophtalmologue
Association à une pathologie non oculaire : diabète, sida Concomitance avec traitement médicamenteux au long cours (corticoïdes, …)	Pas de délai minimal	Prescription par un ophtalmologue

Figure 7 – Possibilité de renouvellement des équipements optique (2/2)

La réforme du reste à charge zéro est effective en matière optique depuis le 1er janvier 2020. Seuls les équipements optiques (paires de lunettes) sont concernés, le remboursement des lentilles n'est quant à lui pas impacté par la réforme « reste à charge zéro ».

4.6.2. Réforme du reste à charge zéro dans le dentaire

Depuis le 1er avril 2019, date d'application de la nouvelle Convention nationale des chirurgiens-dentistes signée le 21 juin 2018, des « honoraires limite de facturation » pour certaines prothèses (panier de soins reste à charge zéro) sont mis en place. En contrepartie de ces plafonnements, la convention prévoit une augmentation des honoraires de certains soins dentaires.

Par exemples :

- Revalorisation de l'acte de restauration d'une incisive ou d'une canine sur une face par matériau inséré en phase plastique, sans ancrage radiculaire à hauteur de 25,06 €,
- Revalorisation de l'acte d'avulsion d'une dent temporaire sur arcade à hauteur de 25 €,
- Valorisation de l'acte d'application du vernis fluoré à hauteur de 25 €.

Depuis le 1er janvier 2020, en application de la réforme du reste à charge zéro, le chirurgien-dentiste devra remettre à son patient plusieurs devis :

- **Le panier « reste à charge zéro »**, les actes prothétiques sont plafonnés et pris en charge intégralement par l'assurance maladie et les complémentaires santé,
- **Le panier aux tarifs maîtrisés**, les prix des actes prothétiques sont plafonnés mais il n'y pas d'obligation pour les complémentaires santé de les prendre en charge intégralement,
- **Le panier aux tarifs libres**, il permet à l'assuré de choisir les matériaux et techniques de son choix sans plafonnement.

Depuis le 1er janvier 2021, la réforme du 100% santé en dentaire est entrée en pleine application. De plus, de nouveaux actes sont admis au sein des paniers de soins 100% santé tels que les dentiers amovibles à base de résine.

4.6.3. Réforme du reste à charge zéro pour l'auditif

Depuis le 1er janvier 2021, la réforme du 100 % santé dans l'auditif est entièrement effective. Les professionnels doivent proposer deux classes d'aide auditive à leur patient :

- La classe 1[6] concerne les aides auditives résultant de l'offre 100 % santé et permet à l'assuré d'être dispensé de reste à charge,
- La classe 2 concerne les aides auditives à prix libre de vente.

Afin de permettre la prise en charge intégrale des actes de classe 1, la base de remboursement des prothèses auditives a fortement augmenté depuis 2019 passant de 300€ à 400€ (pour les adultes).

Évolutions du prix de vente maximal et de la base de remboursement				
Âge du patient	**Prothèse auditive**	**1er janvier 2019**	**1er janvier 2020**	**1er janvier 2021**
20 ans et plus	Prix de vente maximal (classe I)	1 300 €	1 100 €	950 €
	Base de remboursement (Classe I et II)	300 €	350 €	400 €
Moins de 20 ans	Prix de vente maximal (classe I)	1 400 €	1 400 €	1 400 €
	Base de remboursement (Classe I et II)	1 400 €	1 400 €	1 400 €

Figure 8 - Evolution du remboursement des prothèses auditives

4.7. La résiliation infra-annuelle

4.7.1. Présentation

A l'instar de la résiliation infra-annuelle adoptée par la loi Hamon du 17 mars 2014 pour un certain type de contrats d'assurance à reconduction tacite, le législateur a amélioré les droits des consommateurs assurés avec la loi du 14 juillet 2019 relative au droit de résiliation sans frais des contrats santé. Le décret du 24 novembre 2020 est venu la compléter.
La résiliation sans frais des contrats santé est applicable depuis le 1er décembre 2020 pour les contrats existants à cette date.

4.7.2. Champ d'application

Le droit de résiliation tel que prévu par la loi s'applique à tous les contrats santé :

- Les contrats d'assurance santé prévus par le Code des assurances,
- Les mutuelles prévues par le Code de la mutualité,
- Et les complémentaires santé prévues par le Code de la Sécurité Sociale.

Il concerne tous les contrats ou règlements d'assurance couvrant des personnes physiques en dehors de leur activité professionnelle pour les risques maladie et accident, ne comportant aucune autre garantie sauf :

[6] Les produits issus de la classe 1 sont soumis à un prix de vente maximal et sont pris en charge intégralement par l'Assurance Maladie Obligatoire et les complémentaires santé.

- Le risque décès,
- L'incapacité de travail ou l'invalidité,
- L'assistance,
- La protection juridique,
- La responsabilité civile,
- La nuptialité-natalité ou,
- Les indemnités en cas d'hospitalisation.

Sont intégrés à ce dispositif, les contrats santé comportant une partie prévoyance mais pas les contrats de prévoyance pure. La loi s'appliquera à la fois aux contrats individuels et aux contrats collectifs à adhésion facultative ou obligatoire. Cependant, pour les contrats collectifs à adhésion obligatoire, seuls les souscripteurs, c'est-à-dire les employeurs ou les personnes morales, pourront faire jouer la résiliation infra-annuelle. Cette possibilité n'est donc pas ouverte aux salariés adhérents.

Cette faculté de résiliation devra être mentionnée dans chaque bulletin d'adhésion pour les assurances collectives. En tout état de cause, elle devra être rappelée à chaque avis d'échéance de cotisation ou lors de la communication annuelle prévue à la dernière phrase du premier alinéa de l'article L871-1 [7] du Code de la Sécurité Sociale.

4.7.3. Etapes de la résiliation

La faculté de résiliation est ouverte à l'expiration d'un délai d'un an à compter de la première souscription. L'employeur peut résilier lui-même le contrat ou déléguer cette tâche au nouvel assureur.

Par l'employeur : Si l'employeur souhaite résilier lui-même son contrat santé, dans ce cas, il bénéficie des règles de « notification » qui ont été assouplies pour tous les contrats d'assurance.
Pour résilier ledit contrat, la loi prévoit les dispositions suivantes :

- Une déclaration faite contre récépissé au siège social ou chez le représentant de l'assureur (union, institution de prévoyance) dans sa localité,
- Par acte extrajudiciaire,
- Par lettre recommandée ou par envoi recommandé électronique,
- Par lettre ou tout autre support durable,
- Par le même mode de communication à distance proposé lors de la conclusion de contrat.

De son côté, l'assureur doit confirmer par écrit la réception de la notification de résiliation.
Ces nouvelles modalités sont aussi bien applicables aux contrats conclus par les professionnels qu'à ceux conclus par les consommateurs.

[7] « Le bénéfice de ces mêmes dispositions est également subordonné à la condition que la mutuelle ou union relevant du code de la mutualité, l'institution de prévoyance régie par le présent code ou l'entreprise d'assurances régie par le code des assurances communique avant la souscription puis annuellement, à chacun de ses adhérents ou souscripteurs, le rapport, exprimé en pourcentage, entre le montant des prestations versées par l'organisme pour le remboursement et l'indemnisation des frais occasionnés par une maladie, une maternité ou un accident et le montant des cotisations ou primes hors taxes afférentes à ces garanties, ainsi que le montant et la composition des frais de gestion de l'organisme affectés à ces mêmes garanties, exprimé en pourcentage des cotisations ou primes hors taxes afférentes, selon des modalités précisées par arrêté du ministre chargé de la Sécurité Sociale. »

Par le nouvel assureur : Si la résiliation est à l'initiative du nouvel assureur, la première étape est la demande de l'employeur auprès du nouvel assureur. L'employeur doit, sur papier ou support durable, manifester expressément sa volonté de dénoncer l'adhésion ou de résilier son contrat en cours et de souscrire un nouveau contrat auprès du nouvel assureur.

La seconde étape est la notification de la résiliation par le nouvel assureur au précédent assureur. Cette notification doit être signifiée par lettre recommandée ou par envoi recommandé électronique et mentionner les informations suivantes :

- Référence du contrat,
- Nom et adresse de l'employeur,
- Nom du nouvel assureur choisi par l'adhérent.

Le nouvel assureur doit s'assurer de la continuité de la couverture de l'adhérent durant l'opération de résiliation.

4.7.4. Effets de la résiliation

Que la résiliation émane de l'employeur ou du nouvel assureur, elle prendra effet un mois après que le précédent assureur aura reçu la notification de la résiliation.
Le précédent assureur communique par tout support durable à l'employeur un avis de résiliation l'informant de la date de prise d'effet de la résiliation, ainsi que son droit au remboursement de son solde dans les 30 jours.
Le nouveau contrat ne peut pas débuter avant la prise d'effet de la dénonciation de l'ancienne adhésion ou la résiliation de l'ancien contrat. La difficulté que l'on peut noter, surtout pour les contrats collectifs, concerne le délai entre la demande de résiliation et la résiliation elle-même. En effet, ce délai est très court, il faut donc une bonne réactivité de toutes les parties afin que la transition soit douce pour l'employeur.

La résiliation est sans frais ni pénalité pour l'employeur, il ne pourra donc pas lui être réclamé des « frais de résiliation ». En revanche, il peut lui être demandé de régler la partie de prime correspondant à la période où le risque a continué d'être couvert.

Pour une prime réglée annuellement, le précédent assureur devra rembourser la partie de la prime au prorata du temps pendant lequel le risque n'a pas été couvert. Si le précédent assureur ne rembourse pas l'employeur dans le délai de 30 jours, le montant sera majoré des intérêts au taux légal.

Suite à l'étude des différentes réformes exposées ci-dessus, force est de constater le caractère évolutif du système d'assurance français. Prendre en compte et suivre ses évolutions règlementaires et législatives est indispensable aux organismes de complémentaire santé notamment lors d'analyses actuarielles. L'objet de notre étude n'est pas de quantifier l'impact de ces évolutions, mais de détailler la modélisation de la tarification d'un contrat de complémentaire santé. La section suivant s'attache à détailler les grands principes de la tarification en santé.

II – LES GRANDS PRINCIPES THEORIQUES EN TARIFICATION SANTÉ

Cette deuxième partie plus riche en théorie, permet de détailler l'utilisation de modèles paramétriques (modèles linéaires généralisés) et non paramétriques (via des méthodes de *Machine Learning*) dans la tarification d'un contrat santé. Avant toute chose, rappelons que les individus ne sont pas égaux devant le risque. Certains assurés ont en effet des comportements plus risqués que d'autres. Ainsi, au-delà de déterminer le juste prix, l'assureur doit réfléchir aux risques inhérents à son activité. Ainsi, après avoir présenté le principe de la mutualisation et de la segmentation nous détaillerons le risque moral et l'antisélection.

1. Principe de la mutualisation et de la segmentation

Plus un assureur mutualise, plus il se rapproche d'un système solidaire sans écart de tarification entre les individus. A contrario plus un assureur segmente son portefeuille plus il accroit sa rentabilité et l'équilibre du portefeuille (car les primes correspondront au risque réel de l'assuré). Cependant, cette segmentation entraine une perte de compétitivité sur certains segments. L'enjeu est donc de trouver le meilleur compromis entre performance et compétitivité.
Par ailleurs, chaque individu a un profil de risque bien particulier, il est donc légitime que l'assuré souhaite payer une cotisation en corrélation avec la couverture de son niveau de risque. Malheureusement, compte tenu du nombre d'adhérent et de la complexité de chaque situation personnelle cela est impossible. Ainsi, les assureurs segmentent leur portefeuille en classe homogène. En d'autres termes ils créent des groupes avec des profils de risque similaire.

<u>Prenons un exemple :</u>
Considérons une population P pouvant être divisée en deux groupes : les actifs (Groupe P_1) et les inactifs (Groupe P_2). Supposons ensuite que la population P de N personnes va voir le médecin avec une fréquence F et un coût moyen C. De la même manière, la population P_1 (respectivement P_2) de N_1 (respectivement N_2) personnes consulte leur médecin avec une fréquence F_1 (respectivement F_2) et un coût moyen C_1 (respectivement C_2) tel que : $F_2>F>F_1$ et $C_2>C>C_1$
Considérons maintenant deux sociétés A et B souhaitant assurer ce risque. La société A fait le choix de ne pas segmenter son portefeuille et donc de considérer seulement la population P. A contrario, la société B tient compte des deux populations.
La cotisation demandée par l'entreprise A sera donc : Cot A = F x C
La cotisation pour la société B sera alors :

- *Cot $B_1 = F_1 \times C_1$*
- *Cot $B_2 = F_2 \times C_2$*

Si A est seule sur le marché, elle propose un tarif moyen et son profil technique est nul.
Si B entre dans le marché, les assurés vont se répartir entre ses deux sociétés car sur un segment la société B sera plus compétitive. Ainsi, les actifs vont aller chez la société B et les inactifs dans la société A.
On obtient pour les deux sociétés les résultats techniques suivant :

- *A n'assure que les inactifs :*
  - *Recette – charge = $(N_2 \times F \times C - N_2 \times F_2 \times C_2) < 0$*
- *B assure que les actifs :*
  - *Recette – charge = $(N_1 \times F_1 \times C_1 - N_1 \times F_1 \times C_1) = 0$*

Les cotisations demandées par la société B couvrent exactement les risques assurés alors que la société A est en situation déficitaire.

La mutualisation et la segmentation constituent donc un enjeu majeur dans le calcul de la prime pure. Il s'agit de prendre en compte des caractéristiques permettant de créer un groupe homogène et de mieux définir son offre en alliant la performance et la compétitivité tarifaire.
Dans le cadre de notre étude la segmentation est réalisée par la création de gamme dans le cas présent la gamme « Retraite » sera étudiée. Les particularités de cette population sont décrites dans la section III – DESCRIPTION DU PORTEFEUILLE ÉTUDIÉ.

2. Risque moral et antisélection

2.1. Le risque moral

La théorie de l'assurance indique que la probabilité d'occurrence d'un évènement et le montant du sinistre doivent être indépendants du comportement des assurés. En santé, le risque moral intervient si les assurés peuvent influer sur la probabilité de survenance et/ou sur le coût des soins de santé.

Il existe deux risques moraux :

- Le risque moral primaire ou ex ante : la couverture du risque décourage les comportements de prévention des assurés. Il dépend des comportements des assurés vis-à-vis de tout effort susceptible d'aggraver leur état de santé ou d'augmenter la probabilité d'occurrence d'une telle aggravation,
- Le risque moral secondaire ou ex post : l'assuré possédant une bonne couverture santé a tendance à consommer plus que s'il n'était pas aussi bien couvert. Cela se traduit par des cas où le malade va demander plus de soins et ce à un prix plus élevé, d'une part parce qu'on ne pourra pas en vérifier la nécessité, d'autre part car il n'en supportera pas directement les conséquences tarifaires.

Face à cet opportunisme des assurés, l'assureur peut être tenté de récupérer un supplément de prime qui va, in fine, conduire les assurés à ne plus souhaiter s'assurer ou à surconsommer encore plus. Dès lors, comment limiter l'aléa moral ? La solution pour atténuer ce risque consiste à laisser un reste à charge pour l'assuré tels que des franchises médicales ou encore limiter l'absence d'avance de frais. Nous ne détaillerons pas plus cet effet dans la mesure où La Mutuelle Verte s'en prémuni déjà en mettant en place sur l'ensemble de sa gamme des franchises (seul le remboursement du ticket modérateur est effectué pendant un certain laps de temps si l'adhérent ne peut justifier d'une couverture de complémentaire santé depuis un an).

2.2. L'Antisélection

Le risque d'antisélection[8] est dû à une dissymétrie de l'information entre l'assureur et l'adhérent. En effet, l'adhérent connait réellement son profil de risque alors que l'assureur ne peut que l'estimer.

L'antisélection correspond au fait que l'assuré souscrit à un contrat spécifique car il sait que son risque est supérieur à l'estimation réalisée par l'assureur. Il y aura donc un déséquilibre entre le paiement de la prime d'assurance et le coût réel du contrat.

[8] *George Akerlof, The Market for "Lemons" : Quality Uncertainty and the Market Mechanism, dans Quarterly Journal of Economics, 1970, no 84, pp. 488-500.*

S'en suit un effet pervers car l'antisélection entraine la non-souscription du contrat par des individus en bonne santé avec un profil de risque plus faible qui considèrent que les primes d'assurance sont trop élevées comparativement aux prestations versées. En d'autres termes, les individus à « haut risque » chassent les individus à « faible risque ».

Pour lutter contre ce risque les assureurs peuvent mettre en place différentes solutions :

- Un contrat de complémentaire santé optionnel (en addition du contrat socle) : ainsi les adhérents choisissent la garantie leur offrant le meilleur compromis entre prix et garantie adapté à son profil (sa consommation).
- Un contrat à adhésion obligatoire (assurance santé collective) pour tous les salariés et cela quel que soit le niveau de risque. Il y aura donc une mutualisation des profils.

Modéliser le comportement des assurés en présence d'asymétrie d'information est, dès lors, le point d'appui dans la recherche de critères quantitatifs de la qualité d'un contrat. Pour l'assureur, une telle modélisation est indispensable pour la maîtrise du risque et le pilotage de son activité. Le phénomène d'antisélection est limité par la mise à disposition de différentes garanties dans la gamme « Retraite » de notre portefeuille chaque adhérent peut déterminer le meilleur compromis entre prix et remboursement de soins en fonction de sa connaissance de ses besoins de santé.

3. Tarification via le modèle Coût moyen x Fréquence

La difficulté en assurance est l'inversion du cycle de production. Traditionnellement le produit destiné à la vente est fabriqué avant d'être vendu, le vendeur connait ainsi son coût de production, les frais engendrés et donc la marge qu'il réalisera lors de la vente du produit en question. En assurance, ce cycle est inversé, l'assuré paye une prime et ne reçoit la prestation qu'à posteriori. Les montants que l'assureur devra rembourser ne sont pas connus lors du paiement de la prime, elle ne peut être évaluée qu'approximativement via des méthodes statistiques.

Tarifer un contrat santé revient donc à déterminer la prime qui sera payée par les assurés dudit contrat. Celle-ci est fonction des garanties souscrites, des caractéristiques de l'individu (âge, CSP, garantie choisie, etc.) ainsi que du régime auquel l'assuré est affilié (le Régime Général ou le Régime Local, les différences sont spécifiées dans la section *I.2.2 Le Régime Local d'Alsace-Moselle*). Il convient également de prendre en compte certaines contraintes, à la fois réglementaires (Gender Directive[9], portabilité des droits, …) ou commerciales.

La première étape de la tarification d'un contrat santé est la détermination de la prime pure. La principale méthode utilisée à cette fin est la méthode "Coût Moyen x Fréquence", également appelée "Modèle collectif". C'est une méthode très utilisée en assurance, dont les principes sont les suivants :

Notons S la somme des sinistres sur la période considérée (en général, sur une année). Dans le cas d'un contrat santé, il s'agit de la somme des remboursements effectués par l'assureur.
On a alors :

[9] Directive du le 21 décembre 2012 interdisant la mise en place de tarif différencié entre les sexes. Auparavant, les tarifs proposés aux femmes en santé étaient plus élevés que ceux des hommes à cause des postes de dépenses supplémentaires (maternité, consultations gynécologique, …).

$$S = \sum_{i=1}^{N} X_i$$

avec les notations suivantes :

- La variable aléatoire correspondant au nombre de sinistres est notée N
- La variable aléatoire correspondant au coût du i[ème] sinistre est notée (X_i) pour $i \in N$

Afin d'aller plus loin, deux hypothèses doivent être réalisées :

1. Les $(X_i)_{i\epsilon N}$ sont indépendants et identiquement distribués (i.i.d),
2. Ils sont indépendants de N.

Calculons tout d'abord $\mathbb{E}\,[S \mid N = n]$:

$$\begin{aligned}\mathbb{E}\,[S\,|N = n] &= \mathbb{E}\left[\textstyle\sum_{i=1}^{N} X_i\,\middle|N = n\right] \\ &= \mathbb{E}\,[\textstyle\sum_{i=1}^{N} X_i] \text{ car les } (X_i)_{i\epsilon N} \text{ sont indépendant de } N \\ &= n \times \mathbb{E}[X_1] \text{ car les } (X_i)_{i\epsilon N} \text{ sont i.i.d}\end{aligned}$$

$$\mathbb{E}\,[S\,] = \mathbb{E}\,[\mathbb{E}\,[S\,|N]] = \sum_{n=0}^{\infty} \mathbb{P}(N = n) \cdot \mathbb{E}[S|N = n] = \mathbb{E}[X_1] \times \sum_{n=0}^{\infty} \mathbb{P}(N = n) \cdot n$$

Ce qui conduit, finalement, à :

$$\mathbb{E}[S] = \mathbb{E}[X_1] \times \mathbb{E}[N]$$

L'espérance du montant total des sinistres est donc égale, sous les hypothèses précédemment mentionnées, au produit de l'espérance de la charge d'un sinistre et de celle du nombre total de sinistres.

Ces hypothèses permettent également d'obtenir une formule pour la variance de S :

$$\mathbb{V}[S] = \mathbb{E}[N] \times \mathbb{V}[X_1] + \mathbb{V}[N] \times \mathbb{E}^2[X_1]$$

Cette méthode permet le calcul de la prime pure de façon simple et est à ce titre très utilisée.
<u>Toutefois, elle présente certaines limites :</u>

- L'hypothèse d'indépendance et d'identique distribution des $(X_i)_{i\epsilon N}$ peut être remise en cause. Par exemple, il est possible que le coût varie au cours du temps (inflation, hausses des tarifs médicaux, désengagement de la Sécurité Sociale, ...)
- L'indépendance de N et des $(X_i)_{i\epsilon N}$ peut également être questionnée. En assurance santé, par exemple, un assuré très bien couvert (donc générant des coûts élevés) aura également tendance à aussi beaucoup consommer : coût et fréquence semblent donc à priori positivement corrélés.
- Enfin, cette méthode présente certaines limites opérationnelles. Par exemple, elle ne permet pas d'obtenir la distribution de S, nécessaire pour estimer une probabilité de ruine par exemple.

Deux méthodes sont principalement utilisées afin de prendre en compte l'effet de ces variables sur la prime pure : l'estimation empirique de coefficients correctifs (cf. Annexe 1 pour plus de détail), pour

information c'est la méthode utilisée par La Mutuelle Verte pour le calcul des primes pures) et le recours aux Modèles Linéaires Généralisés. Ces derniers sont détaillés dans la section suivante.

4. Tarification en santé : utilisation des GLM

Les Modèles Linéaires Généralisés (Generalized Linear Models, GLM) constituent le cadre de référence pour modéliser les effets des variables de segmentation sur le tarif. Ils ont été développés en 1972 par John Nelder et Robert Wedderburn. Ils sont actuellement très utilisés en assurance, notamment au niveau de la tarification des risques non-vie. Avant de présenter ce modèle une rapide description du modèle linéaire est nécessaire.

4.1. Préambule : Le modèle linéaire classique

Concrètement, ce modèle vise à déterminer les interactions linéaires entre la variable à expliquer et des variables explicatives déterministes. On dispose, pour une population de n individus, de p variables quantitatives explicatives notées X_1, ..., X_p et d'une variable quantitative à expliquer Y.

Le modèle linéaire s'écrit alors, pour le i^{ème} individu :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_i x_i + \varepsilon_i$$

Il est également possible de l'écrire sous forme matricielle :

$$\mathrm{Y} = \mathrm{X}\,\beta + \varepsilon$$

où

- $\mathrm{Y} = (y_1, y_2, ..., y_n)^t$ le vecteur aléatoire de la variable cible,
- $\mathrm{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}$ la matrice représentant les variables explicatives,
- $\beta = (\beta_0, \beta_1, ..., \beta_n)^t$ est le vecteur des paramètres du modèle ou des coefficients de régression,
- $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^t$ le vecteur des erreurs et $\mathrm{I_n}$ la matrice unitaire de rang n.

Les $(\beta_i)_{(i=1...p)}$ sont généralement estimés par la méthode des moindres carrés. En effet, sous les conditions de Gauss-Markov, l'estimateur des moindres carrés est le plus efficace des estimateurs linéaires sans biais.

Ces conditions sont les suivantes :

- $\forall i, \mathbb{E}(\varepsilon_i) = 0$,
- $\forall i, var(\varepsilon_i)) = \sigma^2$ (homoscédasticité des erreurs),
- $\forall i, j, i \neq j, cov(\varepsilon_i, \varepsilon_j) = 0$.

Ces conditions permettent d'obtenir de très bonnes propriétés sur les estimateurs (non-biaisés et efficaces). Afin de permettre l'obtention d'intervalles de confiance sur les $(\beta_i)_{(i=1...p)}$ et sur σ, une condition de normalité des erreurs est très souvent ajoutée.

Ainsi $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I_n})$ et *a fortiori* $\mathrm{Y} \sim \mathcal{N}(\mathrm{X}\beta, \sigma^2 \mathrm{I_n})$.
Cela implique notamment que l'on ait (Y_1, ..., Y_n) indépendants et identiquement distribués (i.i.d).

L'estimateur de β obtenu par la méthode des moindres carrés est alors

$$\widehat{\beta} = ({}^tXX)^{-1}\,{}^tXY$$

Sous l'hypothèse de normalité, on a alors

$$\widehat{\beta} \sim \mathcal{N}(\beta_0, \sigma^2({}^tXX)^{-1})$$

Cette hypothèse de normalité implique que l'estimateur des moindres carrés est également celui obtenu par maximum de vraisemblance.

Limites du modèle :

- La normalité des erreurs entraîne celle de la variable cible. Ce qui n'est pas toujours le cas en pratique,
- Sous les hypothèses de ce modèle, la variance de Y ne dépend pas des variables explicatives,
- Le modèle linéaire classique ne permet pas de capturer les relations non linéaires entre les variables.

Pour faire face à ces contraintes, les assureurs ont eu recours aux Modèles Linéaires Généralisés (GLM).

4.2. Théorie des Modèles Linéaires Généralisés

Les Modèles Linéaires Généralisés (GLM) sont une extension du cas particulier qu'est le modèle linéaire.
Concernant les hypothèses de ce modèle : il faut également supposer que les (Y_1, ..., Y_n) soient indépendants et identiquement distribués (comme pour le modèle linéaire) et qu'ils suivent une loi qui appartient à la famille des lois exponentielles.

Cette famille de loi englobe de très nombreuses lois usuelles. Parmi lesquelles :

- La Loi exponentielle,
- La loi normale,
- La loi binomiale,
- La loi de Poisson,
- La loi Gamma,
- …

De manière générale, la famille exponentielle regroupe les lois dont la fonction de densité peut d'écrire sous la forme :

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

Avec a, b, c des fonctions. Le paramètre θ s'appelle le paramètre naturel. Le prédicteur linéaire, noté η est défini par η = Xβ. On utilise une fonction de lien notée g, supposée monotone, afin d'exprimer une relation entre Y et le prédicteur linéaire.

$$\mathbb{E} = g^{-1}(\mathrm{X}\beta)$$

Pour chaque distribution, il existe une fonction de lien dite canonique telle que $g\ (\mathbb{E}[Y) = \theta$.
La fonction canonique est généralement retenue pour la fonction de lien g.

Famille	Support	Fonction de lien canonique
Gaussiene	$\mathbb{R}$	$X\,\beta = \mu$
Gamma	$\mathbb{R}^+$	$X\,\beta = \mu^{-1}$
Poisson	$\mathbb{N}$	$X\,\beta = \ln\mu$
Binomial	0, 1, …, N	$X\,\beta = \ln\left(\frac{\mu}{1-\mu}\right)$

Tableau 4 – Fonction de lien canonique en fonction de la distribution de loi de probabilité

Estimation des paramètres :

Les paramètres sont estimés via la méthode du maximum de vraisemblance. Dans le cas des modèles exponentiels, la log vraisemblance s'écrit :

$$ln\ L\ (\theta_1,\ldots,\theta_n,\phi,y_1,\ldots,y_n)\ =\ \sum_{i=1}^{n}\frac{y_i\theta_i\ -\ b(\theta_i)}{a(\phi)}\ +\ c(y_i,\phi)$$

En dérivant la log vraisemblance par rapport au paramètre β (pour i et j fixés) :

$$\frac{\partial \ln(\mathrm{L}_i)}{\partial \beta_j} = \frac{\partial \ln(\mathrm{L}_i)}{\partial \theta_j}\times\frac{\partial \theta_i}{\partial \mu_i}\times\frac{\partial \mu_i}{\partial \eta_i}\times\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\times\frac{y_i-\mu_i}{\mathrm{V}(Y_i)}\mathrm{X}_{\mathrm{ij}}$$

Or :

$$\frac{\partial ln(\mathrm{L}_i)}{\partial \theta_j} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i-\mu_i}{a(\phi)}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu i}{\partial \theta_i}} = \frac{1}{b''(\theta_i)}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}$$

$$\frac{\partial \eta_i}{\partial \beta_i} = x_{ij}$$

Par déduction :

$$\frac{\partial ln(\mathrm{L}_i)}{\partial \beta_j} = \frac{1}{g'(\mu_i)}\frac{y_i-\mu_i}{a(\phi)}\frac{1}{b''(\theta_i)}x_{ij} = \frac{1}{g'(\mu_i)}\times\frac{y_i-\mu_i}{V(Y_i)}x_{ij}$$

Ainsi les équations du score sont les suivantes :

Pour $j = 1, \dots, p$.

$$\sum_i \frac{\partial ln(\mathrm{L}_i)}{\partial \beta_j} = \sum_i \frac{1}{g'(\mu_i)} \times \frac{y_i - \mu_i}{V(Y_i)} x_{ij} = 0$$

En règle générale, ces équations ne peuvent pas être résolues de manière analytique et le recours à des algorithmes itératifs du type Newton-Raphson ou Fisher est nécessaire.

Sélection des variables :

Plus un modèle contient de variables explicatives, plus il est précis mais moins robuste. À l'inverse, moins un modèle a de variables explicatives, plus il est robuste mais moins précis. En effet, l'ajout d'une nouvelle variable explicative apporte des informations supplémentaires sur la variable à expliquer mais impose de fait une contrainte supplémentaire au modèle.

Il s'agit donc de déterminer un sous-ensemble de l'ensemble des variables explicatives. Un compromis est ainsi réalisé entre le souhait que le modèle sélectionné contienne peu de paramètres et le souhait que ce modèle bénéficie d'un pouvoir explicatif suffisant.

Prenons un exemple concret. Si on décrit une personne en donnant sa taille, son poids, ses caractéristiques physiques, sa coupe de cheveux et ses vêtements, on obtient une description très précise de cet individu. On comprend immédiatement de qui il s'agit. En revanche, si la personne change ses vêtements ou sa coupe de cheveux par exemple, on n'est plus sûr que ce nouvel individu soit le même que l'individu initial. Le modèle n'est donc pas robuste. Ne considérer que la taille, le poids et les différents caractères physiques comme la couleur de peau ou des yeux, aurait certainement suffi dans les deux cas à identifier l'individu.

Différentes méthodes existent afin de déterminer la combinaison de variables explicatives qui permettent d'obtenir le meilleur compromis entre précision et robustesse.

- Dans la méthode *forward*, il s'agit de rechercher la variable la plus significative au sens de la déviance[10]. Partant de ce modèle à un facteur, nous cherchons ensuite la variable qui, associée à la première, explique le mieux la sinistralité et ainsi de suite. En d'autres termes, à chaque itération la variable qui entraîne la plus grande diminution de la déviance est sélectionnée. L'introduction des variables est stoppée à partir du moment où leur effet sur le modèle n'est plus significatif, c'est-à-dire que la perte en déviance est jugée négligeable.
- Dans la méthode *backward*, il s'agit de démarrer avec le modèle complet (c'est- à-dire toutes les variables ayant un effet significatif sur le risque) puis de retirer la variable la moins significative, autrement dit celle dont l'élimination entraîne la plus faible augmentation de la déviance.
- Pour notre étude la méthode choisie pour rechercher la combinaison optimale est une méthode de type *stepwise* : il s'agit du mélange des deux techniques précédentes. Les résultats des deux méthodes sont comparés pour déterminer l'ordre d'importance des variables explicatives dans le modèle complet, en tenant compte désormais de l'influence des variables déjà présentes dans le modèle.

[10] Cet indicateur permet donc de quantifier l'écart entre la log-vraisemblance du modèle saturé et celle du modèle estimé. Il est détaillé dans la section II.6.2.

Limites :

Les GLM classiques nécessitent de disposer d'un échantillon de taille *n* relativement importante et surtout que le nombre de variables explicatives (*p*) ne soit pas trop grand. Il parait évident que si *p>n* alors il n'est plus possible de calculer l'estimateur des moindres carrés. De nombreuses techniques ont été développées pour faire face à ce problème.

Deux approches seront abordées dans les sections suivantes :

- La pénalisation de la régression.
- L'agrégation de modèle (cf. section Alternative aux Modèles Linéaires Généralisés : Machine Learning),

4.3. Les régressions pénalisées

Les estimateurs basés sur la vraisemblance conditionnelle, utilisés pour l'estimation des risques, sont instables et ont une grande variance quand le nombre d'événements n'est pas nettement plus grand que le nombre de variables explicatives, ou encore en cas de colinéarité (Greenland, 2000; Corcoran et al., 2001; Bull et al., 2007; Hansson et Khamis, 2008). Les techniques classiques de sélection de sous-ensembles telles que la sélection progressive (*stepwise*) ou la sélection pas à pas (*forward* / *backward*) sont également insatisfaisantes. Le point de vue est qu'aucune stratégie de sélection ne s'est montrée meilleure que la stratégie consistant à inclure toutes les variables explicatives dans le modèle, car ces méthodes élimineraient ou ignoreraient facilement des facteurs importants. Une approche alternative est donnée par les méthodes de pénalisation.

4.3.1. La régression Ridge

La régression qualifiée de *Ridge* consiste à minimiser la somme des carrés des coefficients, le tout pondéré par un facteur (noté λ).

On se retrouve à minimiser l'expression suivante :

Somme des résidus quadratiques (RSS) + λ * (somme des carrés des coefficients ‖.‖)

Ce paramètre λ est utilisé afin de pondérer plus ou moins l'importance de la minimisation du modèle en lui-même ou celle des coefficients.

- Lorsqu'il est égal à 0, on retombe sur le même résultat que la méthode des moindres carrés classique.
- Et lorsqu'il tend vers l'infini (+∞), les coefficients vont tendre vers 0 (sans jamais l'atteindre).
- Pour les valeurs intermédiaires, les coefficients vont être plus petits que dans le modèle non-régularisé.
- Et plus la valeur de λ augmente, au plus la complexité générale du modèle diminue.

4.3.2. La régression Lasso

La régression qualifiée de *Lasso* consiste à minimiser la somme des valeurs absolues des coefficients, le tout pondéré par un facteur (noté λ).

Cela revient à minimiser l'expression suivante :

Somme des résidus quadratiques (RSS) + λ * (somme des valeurs absolues des coefficients ‖.‖)

Les propriétés entre la régression Ridge et la régression Lasso sont assez similaires, néanmoins ci-dessous leurs différences fondamentales :

- Lorsque λ tend vers l'infini (+∞), les coefficients vont tendre vers 0 (sans jamais l'atteindre) dans le cadre de "Ridge" mais pourront atteindre 0 pour *Lasso*.
- *Ridge* est davantage adapté dans le cadre où les variables indépendantes sont fortement corrélées. *Lasso* quant à lui, permet de "sélectionner" les covariables en vue de rendre le modèle plus simple (car les coefficients peuvent devenir nuls).
- Vu que la norme L_1 est employée, cela entraîne des répercussions sur les techniques d'optimisation possibles en vue de minimiser ce système.

4.3.3. Elastic net

La régression de type *Elastic net* consiste à combiner les deux régressions précédentes (*Ridge* et Lasso) afin d'éviter la sélectivité trop forte que peut proposer Lasso tout en conservant possiblement des variables fortement corrélées.

Somme des résidus quadratiques (RSS) +
λ_1 * (somme des valeurs absolues des coefficients ‖.‖) +
λ_2 * (somme des carrés des coefficients ‖.‖)

Les régressions pénalisées permettent de contourner certaines limites des *GLM*, elles ne seront pas étudiées dans le cadre de ce mémoire. En effet, leur utilisation est plus pertinente en présence d'un grand nombre de variables explicatives (pouvant entrainer des corrélations entre variables explicatives). Or dans notre étude nous avons peu de variables explicatives. Toutefois, l'un des objectifs de notre étude est de challenger les modélisations *GLM* avec d'autres méthodes plus innovantes. La section suivante s'attache à détailler le Machine Learning.

5. Alternative aux Modèles Linéaires Généralisés : Machine Learning

Les modèles *GLM* permettent l'utilisation de tests statistiques pour qualifier la qualité d'un modèle. Il est néanmoins nécessaire de faire des hypothèses fortes aussi bien sur la loi de la variable à expliquer que sur les interactions entre les variables explicatives. De ce fait, les modélisations statistiques classiques sont restrictives et ne sont pas adaptées à l'exploration des données : les *Machine Learning* quant à elles, le sont. En effet, ces dernières permettent de capter et de retranscrire les interactions entre les données et ainsi d'affiner l'appréhension du risque. Pour ce faire nous rappelons dans un premier temps la base des méthodes ensemblistes puis nous présenterons deux approches issues d'arbre de décision.

5.1. Les arbres de décision (CART) à la base des méthodes ensemblistes :

Ce modèle ne sera pas implémenté lors des applications numériques, mais sera tout de même introduit dans la mesure où il est à la base des deux modèles présentés par la suite.

Cette méthode présente de nombreux atouts : elle est performante, non linéaire, non paramétrique et permet de visualiser et de comprendre les résultats, présentant ainsi un avantage considérable sur

d'autres modèles plus complexes dont l'algorithme fonctionne comme une boite noire. Son utilisation peut se faire soit en cas de classification, soit en cas de régression.

Les arbres de décision peuvent cependant avoir une forte dépendance aux données d'apprentissage et donc une variance élevée.

Concrètement, les arbres de décisions permettent de partitionner les variables explicatives en groupes homogènes en fonction de la variable à prédire. Ces groupes sont formés en prenant en compte une hiérarchie basée sur la capacité prédictive des variables explicatives. L'algorithme d'un arbre de décision fonctionne selon le principe suivant :

1. Les individus sont divisés en k classes pour expliquer la variable de sortie. La première division est obtenue en choisissant la variable explicative qui fournit la meilleure explication de la variable de sortie. L'échantillon initial est alors divisé en k sous-échantillons de populations. Ces sous-populations définissent des nœuds de l'arbre. À chaque nœud est associée une mesure de proportion qui permet d'expliquer l'appartenance à une classe ou la signification d'une variable de sortie.
2. L'opération est renouvelée, ainsi chaque sous-population est à nouveau divisée selon la variable la plus pertinente.
3. A un moment plus aucune séparation ne sera possible. Au terme du processus les nœuds terminaux sont déterminés, appelés « feuilles » de l'arbre.

L'algorithme *CART* procède en plusieurs étapes. Tout d'abord, *CART* construit l'arbre de taille maximale: chaque feuille de cet arbre ne contient alors qu'une unique observation. Pour ce faire, il divise à chaque étape les données en deux sous-échantillons. La règle de division est choisie selon un critère d'homogénéité : à chaque étape on choisit la règle de division qui permet d'obtenir les groupes de données les plus homogènes. On cherche en particulier à ce que dans chaque sous-échantillon les données soient le moins dispersées possibles. Une fois cette étape terminée, l'arbre final ne contient qu'une unique donnée par feuille et le biais est alors nul.

CART propose ensuite un élagage de l'arbre saturé. Il vise à supprimer des séparations afin de regrouper des données, en partant des feuilles vers les racines. Une nouvelle mesure d'erreur est mise en place : il s'agit de « l'erreur de complexité ».

La notion de fonction objectif :

Les algorithmes supervisés visent à obtenir la meilleure performance possible lors de leur construction selon une « fonction objectif ». Par exemple, tel que présenté précédemment, l'arbre de décision minimise la fonction objectif qui est le critère de *Gini*[11]. Cette mesure ne tient pas compte de la complexité du modèle et amène à une variance potentiellement trop élevée.

Pour en tenir compte, les fonctions objectif ont deux composantes : une fonction de coût α et un

[11] L'indice (ou coefficient) de Gini est un indicateur synthétique permettant de rendre compte du niveau d'inégalité pour une variable et sur une population donnée. Il varie entre 0 (égalité parfaite) et 1 (inégalité extrême). Entre 0 et 1, l'inégalité est d'autant plus forte que l'indice de Gini est élevé.
Il est égal à 0 dans une situation d'égalité parfaite où la variable prend une valeur identique sur l'ensemble de la population. À l'autre extrême, il est égal à 1 dans la situation la plus inégalitaire possible, où la variable vaut 0 sur toute la population à l'exception d'un seul individu.

paramètre de régularisation β. Ainsi, la fonction objectif Ω est définie de la manière suivante :

$$\Omega^{(t)}(\Theta) = \alpha^{(t)}(\Theta) + \beta^{(t)}(\Theta)$$ avec : Θ le jeu de paramètres retenu.

L'objectif est de déterminer les paramètres Θ du modèle qui minimisent la fonction objectif. En complexifiant le modèle on diminue souvent la fonction de coût au détriment du paramètre de régularisation. Pour l'erreur de complexité, le paramètre de régularisation est le suivant :

α × nombre de feuilles

Ainsi, si α est nul, l'arbre maximal minimise l'erreur de complexité. À l'inverse, un α élevé va favoriser des arbres « simples » contenant peu de subdivisions et donc de feuilles. Dans une certaine mesure le paramètre α va ainsi pénaliser le biais au profit de la variance.

L'algorithme *CART* détermine alors avec cette mesure d'erreur de complexité une suite de sous-arbres optimaux minimisant l'erreur de complexité pour α qui varie en partant de 0. Dans cette suite d'arbres optimaux, *CART* détermine l'arbre optimal en le testant sur une base de validation. L'arbre qui minimise l'erreur sur la base de validation est alors retenu.

Nous allons maintenant présenter deux méthodes ensemblistes basées sur des agrégations d'arbres de décision. Le premier est le *Random Forest* (ou forêts aléatoires) il repose sur des stratégies aléatoires. Tandis que le second, *XGBoost*, repose sur une stratégie adaptative. Ces méthodes permettent de combler l'un des principaux défauts des arbres de décision : la dépendance de l'arbre de décision aux données d'apprentissage. L'approche est cependant différente, dans le cadre de forêts aléatoires on utilise des arbres développés de faible biais et nous cherchons à réduire la variance, alors que pour *XGBoost*, nous agrégeons des arbres plus simples de faible variance et nous cherchons à diminuer le biais du modèle.

5.2. Les *Random Forest*

Les *Random Forest* consistent à utiliser plusieurs arbres de décision pour en faire des « forêts ». L'algorithme consiste à construire une famille d'arbres de décision sur des échantillons *boostrap*, suivi de l'agrégation des prédictions des modèles.

Le principe de l'algorithme est de chercher, pour chaque scission, non plus la meilleure scission parmi toutes variables explicatives (n), mais la meilleure scission pour p variables explicatives tirée aléatoirement parmi n. Cette double randomisation a été introduite par L. BREIMAN en 2001.

Les paramètres à optimiser :
Plusieurs paramètres sont à optimiser afin d'obtenir les meilleurs résultats :

- Le nombre de variables sélectionnées par scissions.
- Le nombre de feuilles de chaque arbre.
- Le nombre d'arbres dans la forêt.

5.3. Le Gradient Boosting Machine

Le *Boosting* comme le *Random Forest* se base sur une agrégation d'arbres. Contrairement au *Random Forest*, le *Boosting* génère des arbres en séries, c'est à dire que chaque arbre généré (excepté le premier) a accès à son prédécesseur et notamment à l'erreur commise par son prédécesseur. Le nouvel arbre aura alors pour objectif de combler les lacunes de son prédécesseur en donnant plus de poids aux données mal prédites. Le *Boosting* concentre ainsi ses efforts sur les observations les plus difficiles à ajuster tandis que l'agrégation de l'ensemble des modèles permet d'éviter le sur-apprentissage. Il existe différents algorithmes de *Boosting*, qui diffèrent selon :

- Leur manière de pondérer (pour renforcer l'apprentissage des données mal ajustées),
- Leur objectif (notamment classification/régression),
- La fonction de perte (pour mesurer l'ajustement et la façon d'agréger les modèles successifs).

5.3.1. Le Gradient Boosting :

Afin de définir la méthode une agrégation d'arbres est présentée.

Soit B_k l'arbre construit à l'étape k et $f_k = \sum_{j=1}^{k} B_j$ le modèle à l'étape k.

Les couples entrée/sortie (X, Y) composent la base d'apprentissage. Par simplification, est noté X_i les données de la ligne i de la matrice X. A l'étape k, le modèle donne l'estimation $f_k(X)$ de Y. Prenons comme fonction de coût la fonction des moindres carrés définie de la manière suivante :

$$J = \sum_{i=1}^{m} j\,(Y_i, f(X_i))$$

Avec : $j\,(a, b) = \frac{(b-a)^2}{2}$

Le gradient de J par rapport à $f(X_i)$ est :

$$\frac{\partial \mathrm{J}}{\partial f(X_i)} = \frac{\partial \sum_{l=1}^{m} j\,(Y_l, f(X_l))}{\partial f(X_i)} = f(X_i) - Y_i$$

Le modèle Gradient *Boosting* est organisé de la manière suivante :

- Etape 1 : un premier arbre de décision B_1 est construit, il vise à prédire Y à partir de X.
  Ainsi : $B_1\,(X_i) = Y_i + B_1\,(X_i) - Y_i$
  On appelle résidus la valeur observée des insuffisances du modèle. Pour B_1, les résidus s'écrivent :
  $$res_1 = -\frac{\partial \mathrm{J}}{\partial f(X_i)} = Y_i - f_1(X_i)$$

- Etape 2 : Pour combler les manques du premier arbre, on construit B_2 qui vise à prédire les résidus de l'étape précédente. Une fois le deuxième arbre construit, notre modèle est donc égal à $f_2(X) = B_1(X) + B_2(X)$. Ainsi $B_2\,(X) = (Y_i - f_1(X_i)) + (f_2(X_i) - Y_i)$. Et les résidus de l'étape 2 s'écrivent :

$$res2 = -\frac{\partial \mathrm{J}}{\partial f(X_i)} = Y_i - f_2(X_i)$$

- Etape 3 : Pour combler les manques du deuxième arbre, on construit B_3 qui vise à prédire res_2.
- … Cette procédure est réitérée suivant la descente de gradient. À l'étape k, on construit B_k qui cherche à prédire $res_{k-1} = -\frac{\partial \mathrm{J}}{\partial f(X_i)} = Y_i - f_{k-1}(X_i)$.

5.3.2. Une variante : XGBoost

XGBoost est une version « extrême » du *Gradient Boosting*. Il s'agit d'un algorithme particulièrement populaire à l'heure actuelle. Il obtient d'excellentes performances dans un nombre très varié de situations, en classification comme en régression. *XGBoost* permet de généraliser le *Gradient Boosting* à d'autres fonctions que des arbres de régression. Toutefois nous gardons cette dernière méthode pour présenter l'algorithme.

Dans ce cadre-là et de la même manière que dans le *Gradient Boosting*, *XGBoost* construit des arbres en séries en vue de minimiser le biais, tout en contrôlant la variance.
Définissons :

- la fonction de perte l (*MSE*[12] par exemple),
- les couples entrée/sortie (X, Y),
- B_k l'arbre construit à l'étape k.

Ainsi à l'étape t, la prédiction du modèle est : $\widehat{Y_i}^{(t)} = \widehat{Y_i}^{(t-1)} + \eta B_t(X_i)$
η étant un paramètre permettant de diminuer le phénomène de surapprentissage lorsque le nombre d'arbres est élevé.

Par simplification, est présenté l'algorithme *XGBoost* dans le cas particulier où η = 1.
À chaque étape, l'arbre à ajouter est l'arbre qui va minimiser la fonction objectif. Pour rappel elle se décompose en une fonction de coût plus un paramètre de régularisation. Ainsi à l'étape t :

$$\Omega^{(\mathrm{t})}(\Theta) = \alpha^{(\mathrm{t})}(\Theta) + \beta^{(\mathrm{t})}(\Theta) = \sum_{j=1}^{n} l(\widehat{Y_i}^{(t)}, Y_i) + \sum_{j=1}^{K} \beta(\mathrm{B}_j)$$

Pour *XGBoost*, puisque nous prenons la *MSE*[11] comme fonction de coût l, on utilise :

$$\alpha^{(\mathrm{t})}(\Theta) = \sum_{j=1}^{n} l(\widehat{Y_i}^{(t)}, Y_i) = \sum_{j=1}^{n} (\widehat{Y_i}^{(t)} - Y_i)^2$$

La fonction de régularisation utilisée est la suivante :

$$\beta(\mathrm{B}_j) = \gamma\,\mathrm{T} + \frac{1}{2}\left[\alpha \sum_{j=1}^{T} |c_j| + \lambda \sum_{j=1}^{T} c_j^2\right]$$

avec :

- γ, α, λ des paramètres de pénalisation de la complexité du modèle (à optimiser lors de la phase de calibration par validation croisée[13]),

[12] *Mean Squared Error* (*MSE*) correspond à l'erreur quadratique moyenne d'un estimateur
[13] Cette méthode consiste à séparer aléatoirement la base d'étude en k groupes de tailles égales. Le premier fold est traité en tant que base de validation et les k-1 folds restants sont utilisés en tant que base d'entraînement. Procédure détaillée en annexe 5.

- T le nombre de feuilles de l'arbre B_t ,
- c_j la valeur associée à chaque feuille de l'arbre.

Ainsi, à chaque étape, il s'agit d'ajouter un arbre afin de minimiser la fonction objectif. A des fins de simplifications, considérons le paramètre de pénalisation α comme nul. Ainsi, à l'étape t, la fonction objectif est la suivante :

$$\Omega^{(t)} = \sum_{j=1}^{n} (\hat{Y}_i^{(t-1)} + B_t(X_i) - Y_i)^2 + \beta(B_j)$$

En pratique, le développement de Taylor à l'ordre 2 est utilisé afin d'approcher la fonction objectif :

$$\Omega^{(t)} \approx \sum_{j=1}^{n} [g_i \times B_t(X_i) + \frac{1}{2} \times h_i \times {B_t}^2(X_i)] + \beta(B_j)$$

avec :

- $g_i := \frac{\partial l(\hat{Y}_i^{(t-1)}, Y_i)}{\partial {Y_i}^{(t-1)}} = 2 \times (\hat{Y}_i^{(t-1)} - Y_i)$
- $h_i := \frac{\partial^2 l(\hat{Y}_i^{(t-1)}, Y_i)}{\partial^2 {Y_i}^{(t-1)}} = 2$

D'où :

$$\Omega^{(t)} \approx \sum_{j=1}^{n} \left[2 \times \left(Y_i - \hat{Y}_i^{(t-1)}\right) \times B_t(X_i) + {B_t}^2(X_i)\right] + \beta(B_j)$$

On définit $I_j = \{i \; tel \; que \; q(X_i) = j \; ie \; tel \; que \; X_i \in \text{à } la \; feuille \; j\}$: l'ensemble des éléments appartenant à la feuille j. Sur cet ensemble, la fonction objectif devient :

$$\Omega^{(t)} \approx \sum_{j=1}^{T} \left[G_j \times w_j + \frac{1}{2} \times (H_j + \lambda) \times w_j^2\right] + \gamma \times T$$

avec :

- w_j la valeur associée à chaque feuille,
- $G_j = \sum_{i \in I_j} g_i$,
- $H_j = \sum_{i \in I_j} h_i$.

Les w_j sont indépendants les uns par rapport aux autres et $G_j \times w_j + \frac{1}{2} \times (H_j + \lambda) \times w_j^2$ est une forme quadratique. Pour une structure d'arbre q fixe, on a alors la solution suivante :

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

$$\Omega^* = -\frac{1}{2} \times \sum_{j=1}^{t} \frac{G_j^2}{H_j + \lambda} + \gamma \times T.$$

La dernière équation permet de mesurer à quel point la structure q(x) est performante. Cependant, il est impossible de comparer toutes les structures d'arbre possible (il y en a une infinité). En pratique, nous débutons avec un arbre de profondeur nulle et ajoutons des nœuds de manière récursive. Lorsque l'on ajoute un nœud à une feuille, nous découpons alors la feuille en deux feuilles. La baisse de la fonction objectif est alors la suivante :

$$\frac{1}{2} \times \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

avec :

- $\frac{G_L^2}{H_L + \lambda}$ (resp. $\frac{G_R^2}{H_R + \lambda}$) le score sur la partie gauche (resp. droite) de la séparation,
- $\frac{(G_L + G_R)^2}{H_L + H_R + \lambda}$ le score sur la feuille originale de l'arbre, avant séparation,
- γ le surplus de complexité du modèle induit par la séparation.

Nous pouvons donc mesurer cette baisse pour chaque variable. Pour chaque variable, un parcours linéaire permet de déterminer la meilleure séparation. On réitère cette procédure sur tous les nœuds tant qu'il est possible d'améliorer la fonction objectif.

Les deux méthodes ensemblistes que nous allons utiliser dans notre étude sont le *Random Forest* et le *XGBoost*, qui permettent de visualiser les variables les plus discriminantes. Cette information est précieuse car ces modèles sont capables de détecter des interactions entre variables qu'un test univarié d'importance n'aurait pas nécessairement révélé. Ainsi une comparaison des variables les plus discriminantes en fonction de la méthode choisie sera réalisée.

6. Indicateur de performance

L'un des enjeux de cette étude est de pouvoir comparer les différentes modélisations. Ainsi, il est nécessaire de déterminer des indicateurs nous permettant d'évaluer et de comparer les performances de différentes modélisations. Notre choix se porte sur deux indicateurs le *RMSE* et la déviance.

6.1. Le *MSE* et le *RMSE*

Le *Mean Squared Error* (*MSE*), correspond à l'erreur quadratique moyenne d'un estimateur. Dans notre cas, si on note $(Y_i)_{i=1,\dots,n}$ les fréquences de sinistres observées dans la base et $(\hat{Y}_i)_{i=1,\dots,n}$ les fréquences de sinistres prédites par un modèle, alors l'erreur quadratique du modèle est calculé comme suit :

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

Tel qu'il est défini, l'interprétation du *MSE* est aisée : plus le *MSE* est élevé plus la fréquence de sinistres prédite est éloignée de la fréquence de sinistres observée, et donc moins le modèle est ajusté. Un *MSE* nul signifie que l'estimation est parfaite car toutes les valeurs prédites sont égales aux valeurs

observées. Pour comparer deux modèles, on compare leur *MSE* pour retenir celui qui a la valeur la plus petite.

Le *RMSE* (*Root Mean Squared Error*) représente la racine carrée du *MSE*. Le *RMSE* mesure la différence entre les valeurs prédites par le modèle et les valeurs observées (réelles) comparable à « l'écart type des erreurs ».

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right)^2}$$

6.2. La déviance

La déviance Δ d'un modèle statistique se définit comme suit :

$$\Delta = 2\log\left(\tilde{\mathrm{I}} - \hat{\mathrm{I}}\right)$$

Avec :

- $\tilde{\mathrm{I}}$ la log-vraisemblance du modèle saturé,
- $\hat{\mathrm{I}}$ la log-vraisemblance du modèle estimé.

Cet indicateur permet donc de quantifier l'écart entre la log-vraisemblance du modèle saturé et celle du modèle estimé. Pour rappel, le modèle saturé correspond à un modèle possédant autant de paramètres que d'observations et donc qui prédit exactement les valeurs observées. Ainsi, la déviance permet de déterminer la déviation entre notre modèle et le modèle parfait, c'est pourquoi la déviance est un indicateur à minimiser.

Cependant, l'indicateur utilisé ne sera pas tout à fait la déviance des modèles mais plutôt la moyenne des résidus de déviance des modèles. Plus petite sera la moyenne des déviances résiduelles, plus petite sera la déviance. C'est pourquoi nous cherchons aussi à minimiser la moyenne des déviances résiduelles.

6.3. Critère *AIC* et *BIC*

Le critère *AIC* (*Akaike Information Criterion*) :

C'est une mesure statistique de la qualité d'un modèle. Nous pouvons l'utiliser comme un indicateur pour comparer deux modèles estimés. Le meilleur modèle sera celui minimisant le critère AIC.

$$AIC = -2I + 2q$$

Avec I la log-vraisemblance du modèle et q le nombre de paramètres.
Plus le modèle possède un nombre élevé de paramètres, plus le critère AIC est pénalisé.

Le critère *BIC* (*Bayesian Information Criterion*) :

C'est une variante du critère AIC. La différence se traduit par la prise en compte du nombre d'observations n :

$$BIC = -2I + log(n)$$

La taille de l'échantillon est donc un facteur supplémentaire de la détérioration du critère BIC. Là encore il s'agit de minimiser ce critère pour avoir le meilleur modèle.

Ces indicateurs sont sensibles à la taille de l'échantillon. Il convient donc de les analyser en regard de la taille de l'échantillon ou bien de les normaliser par cette taille. Le critère AIC est une pénalisation adaptée de la déviance mais il a tendance à sélectionner un nombre trop important de variables (et ce défaut ne s'atténue pas avec un nombre plus élevé d'observations). Le critère BIC traite le problème de sélection des variables sous un autre angle. Il se base sur la théorie bayésienne et optimise la distribution à posteriori. Cet indicateur corrige le principal défaut de *l'AIC*, à savoir la sélection d'un nombre trop important de variables. Le *BIC* est asymptotiquement convergent, mais la sélection des variables selon cet indicateur est trop restrictive.

Cette section a permis de faire un rappel théorique des différentes manières de tarifier un contrat de complémentaire santé et de comparer nos modèles. Ces modèles seront implémentés dans la suite de notre étude. Mais avant cela, notre portefeuille doit être analysé et notre base de données assainie.

III - DESCRIPTION DU PORTEFEUILLE ÉTUDIÉ

Tarifer un contrat impose une exploration des données au préalable afin de mieux connaître son portefeuille et notamment ses risques et ses spécificités. Cette première étape est primordiale afin d'envisager un retraitement de certaines variables. Cette section met en lumière le cheminement réalisé dans le retraitement des données ainsi qu'une description des principales caractéristiques du portefeuille étudié.

1. Les hypothèses du profil de risque du portefeuille de l'étude

Pour que le principe d'assurance fonctionne il faut un risque réalisable ainsi qu'un aléa. L'assureur a donc besoin de connaitre les risques auxquels ses assurés sont soumis mais également le type d'aléa. A cette fin, une bonne connaissance du profil assuré est obligatoire.

Concrètement, lors du processus de souscription, les assureurs accompagnent leurs clients en réalisant une phase de « découverte ». Cette étape est cruciale car elle permet de récupérer des informations sur le prospect et ainsi mieux cerner ses attentes et ses besoins afin de l'orienter sur une garantie appropriée mais surtout, elle permet de déterminer son profil de risque. L'âge, la zone géographique, l'activité professionnelle font notamment partie des informations indispensables à la mise en place de la police.

Outre les données propres à chaque individus (l'âge, la zone géographique, l'activité professionnelle) les adhérents souscrivant à la gamme « Retraite » possèdent des caractéristiques similaires.

- Un âge avancé comparativement à d'autres populations. En effet, les produits de la gamme sont disponibles à partir de 60 ans,
- Une consommation médicale plus importante que d'autres populations. En effet, du fait de leur âge, les assurés auront tendance à avoir recours plus régulier à leur contrat de complémentaire santé. Ce qui explique également le fait que les cotisations de cette population sont plus élevées,
- Un profil plus « fidèle » (moins de radiation).

Afin de débuter l'étude, voici ci-dessous le détail des informations à notre disposition.

2. Les données initiales

Les données de notre étude concernent les produits de la gamme « Retraite », soit quatre garanties. Une extraction des données de notre portefeuille a été réalisée pour les survenances 2019, 2020 et 2021. En pratique, deux bases sont extraites : la première concerne les informations des adhérents et de leur contrat, la seconde détaille quant à elle les prestations versées sur les périodes analysées.

Plus précisément, la base contrat comprend les champs suivants :

- Le numéro de l'adhérent principal,
- Le rang du bénéficiaire,
- Le type de bénéficiaire (adhérent principal, conjoint, enfant),
- L'âge de l'adhérent principal à la date de situation (date d'extraction)

- Le type d'activité de l'adhérent principal,
- La garantie choisie,
- Le département de résidence de l'adhérent,
- Le sexe du bénéficiaire.

La base prestation a la forme suivante :

- Le numéro de l'adhérent principal,
- Le rang du bénéficiaire,
- Le type de bénéficiaire (adhérent principal, conjoint, enfant),
- La garantie choisie,
- Les remboursements de prestation avec :
    - La date de soins,
    - La date de règlement de l'acte,
    - Le montant des actes réalisés (frais réel par acte),
    - La base de remboursement (BR) des actes réalisés (BR par acte),
    - Les remboursements de la complémentaire santé sur les actes réalisés (montant par acte),
    - Les remboursements reçus par l'adhérent vis-à-vis d'une autre complémentaire santé dans le cas où le contrat de l'assuré est en mutuelle de second rang,
    - Le regroupement statistique afin de catégoriser les actes (optique, dentaire, …).

Une fois les bases extraites, une analyse de celles-ci doit être effectuée. En effet, certaines variables doivent être ajoutées, d'autres doivent être épurées et/ou corrigées. La section suivante présente ces actions.

3. Retraitement des bases de données

Dans notre cas, ont été écartées à la suite de cette première analyse les données négatives notamment dues aux annulations de remboursement. Ces annulations peuvent être liées :

- À un défaut de paiement de la prime d'assurance, conduisant à l'annulation des remboursements de l'adhérent,
- Ou encore à une erreur opérationnelle telle que les erreurs de saisie,
- …

Une correction et suppression de certaines lignes remontées ont été réalisées. En effet l'extraction a mis en évidence des lignes de prestation qui n'auraient pas dû être présentes (le produit de l'adhérent ne correspondant pas à l'étude). Ce biais provient du fait que sur la période d'extraction, un adhérent peut avoir souscrit successivement différents produits (passage à la retraite entrainant un changement de produit par exemple).

Une fois la base assainie, la phase de préparation des données pour notre étude peut commencer. A ce titre, ont été réalisées les actions suivantes :

- Calcul de l'exposition pour tous les bénéficiaires présents dans la base de données,
- Recalcul de tous les sinistres en « as if » pour les années 2019 et 2020 en prenant en compte l'évolution des bases de remboursement entre ces périodes et l'inflation des dépenses de santé. Habituellement l'indicateur servant de base à cette mesure est le Plafond Mensuel de

la Sécurité Sociale (PMSS) or avec la crise du Covid19 cet indicateur est resté fixe à 3428€ de 2020 à 2022. L'évolution moyenne du PMSS sur 10 ans a donc été utilisée,

- Regroupement par zone tarifaire en utilisant le zonier de La Mutuelle Verte. Un zonier sur notre base de données a été réalisé malheureusement le nombre d'adhérent dans certains départements ne permet pas une analyse fiable. Le zonier de La Mutuelle Verte a donc été retenu.
- Création de classes d'âge.

Une fois les retraitements effectués, une analyse plus fine du portefeuille est réalisée à travers l'étude des variables présentes dans notre base de données. Cette étape est importante car elle permet de comprendre la structure du portefeuille.

4. Statistiques descriptives du portefeuille

Cette section met en avant les spécificités du portefeuille étudié à travers l'analyse des variables de la base de données.

4.1. Description de la population par survenance

Le tableau suivant présente le nombre de chef de famille ainsi que les bénéficiaires depuis 2019 adhérent à l'une des gammes « Retraite » :

	2019	2020	2021
Chef de famille	2021	2349	2876
Conjoint	469	540	640
Enfant	28	29	34
Effectif Bénéficiaire	2515	2915	3548
Charge familiale	1,24	1,24	1,23

Tableau 5 - Analyse des bénéficiaires par survenance

Le chef de famille est l'adhérent principal du contrat. Il a un numéro d'adhérent propre définit par l'assureur. Il peut ensuite ajouter sur son contrat son conjoint et(ou) son(ses) enfant(s) à charge. L'ensemble de la famille constitue l'effectif bénéficiaire.

Nous constatons une croissance de l'effectif bénéficiaire depuis 2019 et ce malgré la réforme de la RIA. Le profil « fidèle » de nos adhérents est bien vérifié. Par ailleurs, du fait d'une population retraitée ne possédant pas ou peu d'enfant sur leur contrat de complémentaire santé, la charge familiale est faible. Par ailleurs, elle est également stable sur l'ensemble des survenances analysées. A noter que compte tenu de l'âge de l'adhérent principal, les enfants à charge correspondent en majorité à des « enfants majeurs handicapés » pouvant être couverts.

4.2. Répartition de la population en fonction des classes d'âge

Afin d'appréhender au mieux la consommation du portefeuille, l'âge de l'effectif est analysé à l'aide d'une pyramide des âges. En effet, les besoins en frais médicaux augmentent avec l'évolution de l'âge. Par exemple, les jeunes actifs ont des dépenses moins importantes que notre portefeuille composé de seniors.

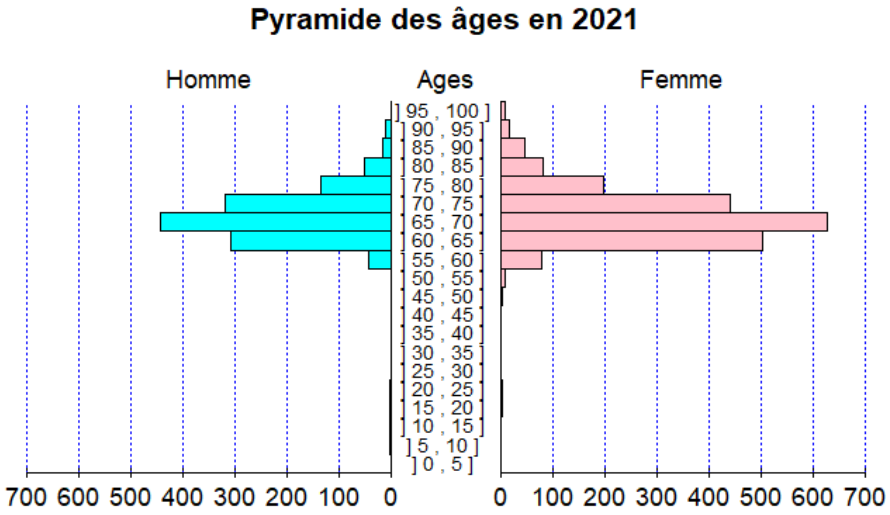


Figure 9 - Pyramide des âges en 2021

Sans surprise, notre portefeuille est composé d'une majorité de personnes retraitées. Les conjoints et les enfants sont très peu représentatifs sur ce portefeuille ; 82% des bénéficiaires se situant entre 60 ans et 75 ans.

4.3. Analyse des dépenses de santé en fonction de l'âge

Comme évoqué précédemment, les dépenses de santé augmentent avec l'âge. Le graphique ci-dessous présente/recense les dépenses de santé des adhérents de la gamme « Retraite » en 2021 par classe d'âge :

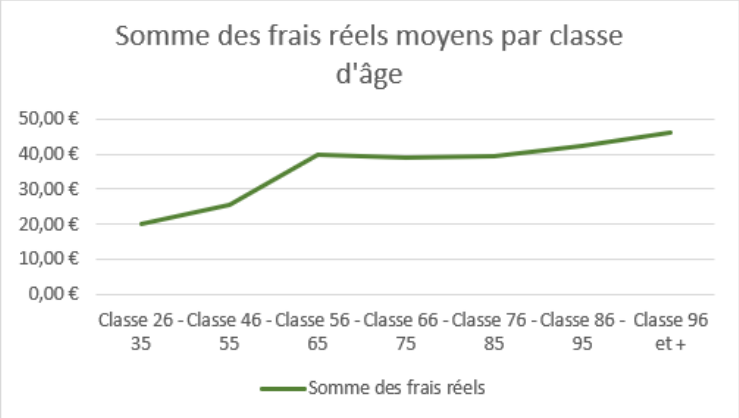


Figure 10 - Somme des frais réels par classe d'âge pour la survenance 2021

Le phénomène d'augmentation des dépenses de santé est notable. Le plateau constaté entre les classes d'âge centrales s'explique par la faible représentativité de ce groupe, biaisant de ce fait l'analyse.

4.4. Consommation moyenne par garantie

Afin de rentrer plus en détail dans l'analyse de la consommation, analysons la répartition des montants de frais réels moyen et de consommation moyenne par garantie et par année de survenance. Le détail est présenté dans le tableau ci-dessous :

	2019		2020		2021		Sur 3 ans	
	Moyenne MT_FR[14]	Moyenne MT_AMC[15]	Moyenne MT_FR[8]	Moyenne MT_AMC[9]	Moyenne MT_FR[8]	Moyenne MT_AMC[9]	Moyenne MT_FR[8]	Moyenne MT_AMC[9]
Retraite 1	1 284 €	389 €	1 425 €	475 €	1 507 €	589 €	1 425 €	503 €
Retraite 2	1 622 €	525 €	1 588 €	552 €	1 644 €	629 €	1 620 €	575 €
Retraite 3	2 437 €	836 €	2 319 €	888 €	2 506 €	943 €	2 426 €	895 €
Retraite 4	3 224 €	1 289 €	3 175 €	1 243 €	3 198 €	1 363 €	3 199 €	1 301 €

Tableau 6 - Frais réels & Remboursement AMC par garantie et par survenance

Quatre formules sont proposées : une entrée de gamme, deux intermédiaires et une haut de gamme. Naturellement, dès que la couverture santé augmente (produit plus haut de gamme) nous constatons également une hausse des remboursements de la complémentaire santé. Il est intéressant de noter qu'il en est de même pour les frais réels, soit la dépense totale de l'adhérent. En d'autres termes, la dépense de santé est corrélée positivement au niveau de garantie de l'adhérent. En réalité, cet effet s'explique par la possibilité d'obtenir des remboursements santé supplémentaires en souscrivant à un produit plus « haut de gamme » les adhérents peuvent bénéficier de soins plus onéreux avec potentiellement peu ou pas de reste à charge. De ce fait ils consomment plus. Ce point a été détaillé dans la section II.2.1. Le risque moral.

En s'attachant à analyser la consommation en santé par survenance (le tableau a été réalisé avant le retraitement en « as if » des remboursements santé), nous constatons une hausse significative des remboursements santé. La hausse la plus significative est observée sur la garantie entrée de gamme. Cette hausse s'explique par l'entrée en vigueur de la réforme du 100% santé de 2020, entraînant un remboursement des équipements optiques et prothèses dentaires par la complémentaire santé, sans reste à charge pour l'adhérent. Sur ces postes la garantie Retraite 1 est une garantie dite « d'entrée de gamme », à ce titre elle ne rembourse que peu les adhérents. L'impact de la réforme du 100% santé est donc particulièrement notable et conséquent.

Il est intéressant de noter que l'impact est moindre sur les garanties haut de gamme qui remboursent déjà considérablement l'adhérent.

Par ailleurs, il faut également noter la présence d'une dérive médicale entre ces différentes survenances. Cette dérive est observée à travers l'évolution des frais réels. Elle prend en compte différents effets :

[14] *Moyenne MT_FR = Moyenne des frais réels par bénéficiaire en 2021.*
[15] *Moyenne MT_AMC = Moyenne des remboursements de la complémentaire santé par bénéficiaire en 2021.*

- L'inflation dans les dépenses de santé,
- Le désengagement de la Sécurité Sociale engendrant une prise en charge supérieure par les complémentaires santé (déremboursement des actes d'homéopathie par exemple).

Ces analyses expliquent le recours aux sinistres « as if » dans notre étude.

4.5. Répartition des formules

Pour rappel, quatre formules sont proposées avec un produit « entrée de gamme », deux garanties intermédiaires et la dernière considérée comme une garantie « haut de gamme ». Comme l'indique le graphique ci-dessous, les assurés sont majoritairement présents sur les gammes intermédiaires Retraite 2 et Retraite 3.

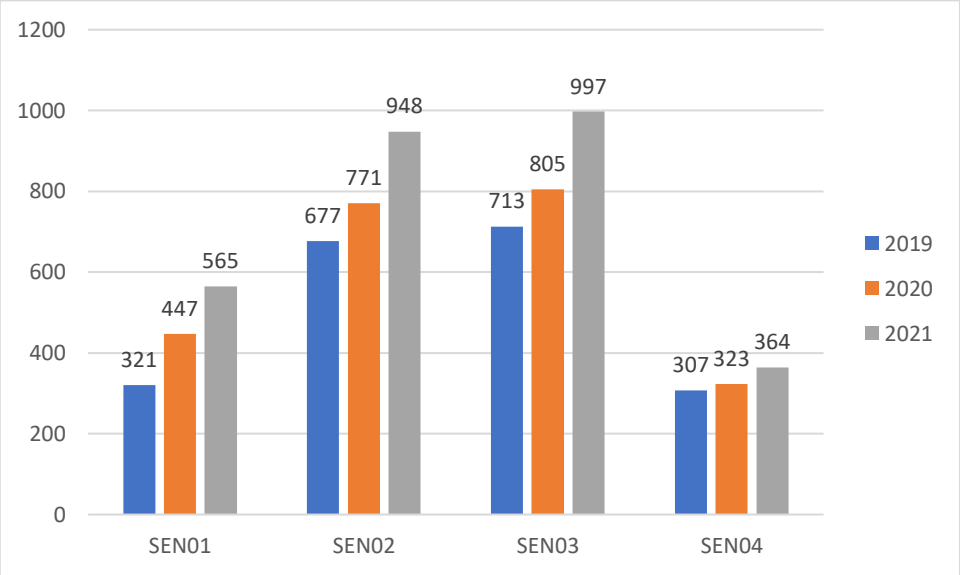


Figure 11 - Répartition des codes produit par survenance

Cette analyse peut être complétée par une analyse de la répartition des produits par classe d'âge. Procédons à cette analyse sur la survenance 2021 :
La faible présence d'enfant nous conduit à retravailler les classes d'âge afin de regrouper cette population. La répartition des différentes formules de la gamme « Retraite » par classe d'âge est présentée ci-dessous :

	SEN01	SEN02	SEN03	SEN04
Classe 0 - 25	2%	1%	1%	0%
Classe 26 - 35	0%	0%	0%	0%
Classe 46 - 55	1%	1%	0%	0%
Classe 56 - 65	28%	29%	29%	29%
Classe 66 - 75	45%	52%	56%	62%
Classe 76 - 85	18%	13%	11%	8%
Classe 86 - 95	6%	4%	2%	1%
Classe 96 et +	0%	0%	0%	0%

Tableau 7 - Répartition des formules par classe d'âge

Il est intéressant de noter que :

- La classe d'âge 56 - 65 ans est répartie de manière uniforme sur les différents produits de la gamme « Retraite »,
- La répartition sur la classe d'âge 66 – 75 ans est différente avec en majorité une présence sur la garantie haut de gamme puis les gammes intermédiaires,
- La classe 76 – 85 ans est majoritairement répartie sur le produit Retraite 1.

4.6. Répartition des adhérents par département

La répartition des adhérents par département est présentée ci-dessous :

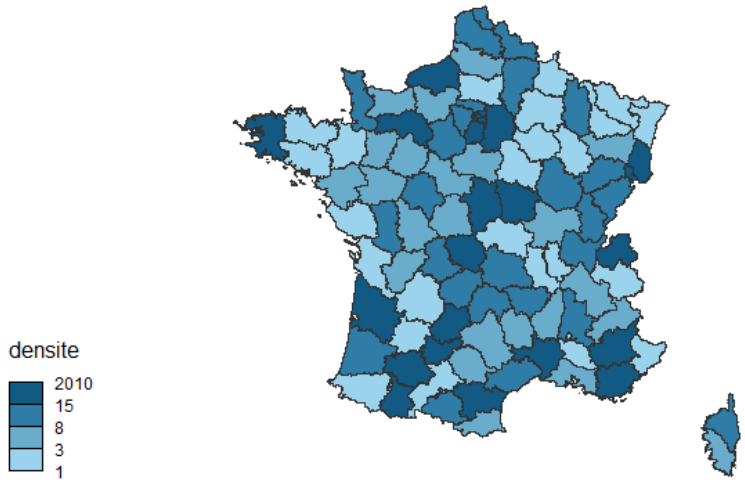


Figure 12 - Densité des adhérents de la gamme « Retraite »par département en 2021

Compte tenu de la faible population d'adhérent sur un grand nombre de département l'analyse graphique n'est pas aisée. Les adhérents sont majoritairement présents sur les départements suivants (dans l'ordre décroissant) :

- Le Var : correspondant au siège de La Mutuelle Verte,
- Le Pas-de-Calais : une agence de La Mutuelle Verte est implantée à Arras,
- Les Alpes-Maritimes,
- Les Bouches-du-Rhône,
- La région parisienne.

Ces 5 départements représentent 75% de la population des seniors de La Mutuelle Verte.

4.7. Analyse de la consommation par famille d'actes

La consommation en santé peut être analysée par grande famille d'actes :

- Le poste « soins courants » :
    - Soins courants honoraires : cette famille englobe les consultations, visites des médecins généralistes / spécialistes mais également les actes techniques, d'imagerie médicale et les frais de transport.

    - Soins courants autres : ce groupe rassemble les infirmiers, les kinésithérapeutes, les orthophonistes et les orthoptistes.
- Le poste « pharmacie » : composé des différents actes réalisés en officine.
- Le poste hospitalisation : répertoriant tous les actes d'hospitalisation avec notamment les actes de chirurgie, obstétriques, de psychiatrie, … Les actes connexes sont également pris en compte comme la chambre particulière, les frais journaliers (directement à la charge des complémentaires santé)
- Le poste « optique » : regroupant notamment l'achat d'équipement optique (lunette), les lentilles et la chirurgie réfractive.
- Le poste « dentaire » : composé des soins conservateurs (détartrage, extraction d'une dent, soin d'une carie, …), des prothèses dentaires, de l'orthodontie, de la parodontologie et de l'implantologie dentaire.
- Le poste « prothèse auditive » : répertoriant les actes liés à l'appareillement auditif et des actes connexes (piles par exemple).
- Le poste « Médecine douce » : englobant les actes de médecine non conventionnelle (ostéopathie, étiopathie, chiropractie, …).
- Les autres soins : cette famille regroupe les autres actes de santé avec notamment le grand et petit appareillage (fauteuils roulant, attelles, …), les actes de cures thermales, les actes hors parcours de soins[16].

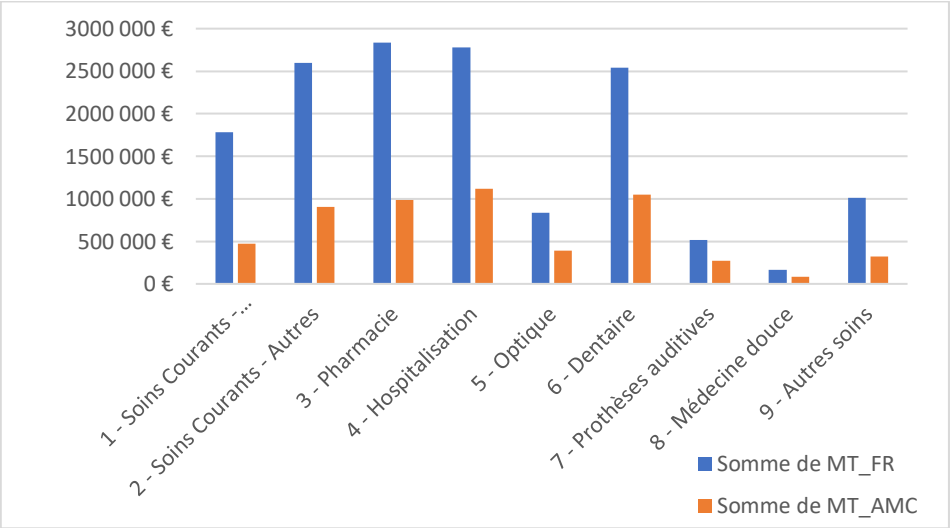


Figure 13 - Analyses des frais réels et des remboursements de la complémentaire santé par famille d'actes en 2021

<u>L'analyse du graphique ci-dessus met en lumière différents points :</u>

Premièrement notons le poids que représentent les remboursements de complémentaire santé vis-à-vis de la dépense totale de l'adhérent. Bien évidemment, il faut prendre en compte le fait que les remboursements du Régime Obligatoire ne sont pas mentionnés dans ce graphique. En effet, prenons le cas des soins courants dans le Régime Général : ils sont remboursés à hauteur de 70% (60% selon les actes) de la base de remboursement par le Régime Obligatoire. Cette forte participation de la part du Régime Obligatoire explique la plus faible part de prise en charge de la complémentaire santé sur

[16] *Lorsqu'un adhérent consulte un spécialiste sans avoir au préalable été orienté par un médecin généraliste. La base de remboursement est modifiée. De ce fait, ce type d'actes est séparé du poste soins courants.*

ce type de remboursement. Dans cet exemple, la complémentaire santé intervient pour la prise en charge du ticket modérateur et des dépassements d'honoraire.
A contrario, la prise en charge sur le poste optique est beaucoup plus importante pour la complémentaire santé. En effet, dans le cas de l'achat d'un équipement optique hors 100% santé, le Régime Obligatoire intervient à hauteur de 60% de la base de remboursement, soit au final une prise en charge à hauteur de 9 centimes pour une monture et deux verres. Le prix d'un équipement optique est bien évidemment très supérieur, de ce fait seul le contrat de complémentaire santé est mis à contribution dans le remboursement de ce type d'acte.

Dans un second temps, il est intéressant de comparer les différentes familles de soins. Habituellement les postes les plus importants d'un contrat de complémentaire santé sont : l'optique et le dentaire. Dans notre cas, les remboursements de la gamme « Retraite » sont principalement les suivants (classés par ordre d'importance) :

- L'hospitalisation,
- Le dentaire,
- La pharmacie,
- Les autres soins courants.

Cette répartition de la consommation est propre à cette population. En effet, avec une avancée de l'âge les hospitalisations sont plus courantes. De plus, ces hospitalisations entrainent des actes connexes :

- Elles peuvent entrainer de la rééducation ou des séances de kinésithérapie et donc de la consommation dans le poste autres soins courants,
- Le prise de médicament et donc le volume du poste pharmacie est impacté.

Ces différences démontrent bien pourquoi cette population a été segmentée vis-à-vis du reste des adhérents de la compagnie d'assurance. Après avoir constaté des spécificités de la gamme « Retraite », l'assureur doit tarifer les contrats. La section suivante est une application des modèles présentés dans la section II, permettant de déterminer la prime pure du portefeuille étudié.

IV – APPLICATION À NOTRE JEU DE DONNÉES

Cette section présente une application de la théorie des modèles linéaires généralisés (*GLM*) ainsi qu'une utilisation des algorithmes de *Random Forest* et de *XGBoost*.
Afin de comparer les capacités prédictives des modèles, la base de données est scindée en deux : 80% des observations servent à calibrer les modèles, les 20% restant à les évaluer. Les termes « apprentissage » et « test » sont respectivement employés pour les désigner. Les analyses présentées dans la première partie de cette section concernent la base d'apprentissage qui permet de calibrer les modèles. Dans un second temps les modèles sont implémentés sur la base test afin de de comparer leur pouvoir prédictif.

1. Méthodes *GLM*

Les modèles linéaires généralisés contiennent trois composantes : la composante aléatoire, la composante déterministe ainsi que la relation fonctionnelle. Afin d'effectuer une modélisation il faut au préalable choisir la loi de la composante aléatoire à modéliser, déterminer une base sur laquelle entrainer le modèle et enfin choisir la fonction de lien. Une analyse des corrélations entre les variables ainsi qu'une sélection des variables implémentées dans les modèles sont ensuite réalisées. Enfin une analyse fine des résultats permettra de valider puis de conclure sur cette première modélisation de la prime pure.

1.1. Choix de la composante aléatoire et de la fonction de lien sur le modèle fréquence

L'objectif ici est de modéliser le nombre de sinistres, ce qui implique que la loi de la composante aléatoire doit être « discrète ». Il est coutume en assurance non-vie de choisir la loi de *Poisson* ou la loi *Binomiale Négative* pour l'étude de la fréquence. La loi de *Poisson* est privilégiée lorsque la distribution est équidispersée, ce qui signifie que sur un échantillon donné l'espérance et la variance de la composante sont identiques. En revanche, lorsque la variance est plus importante que l'espérance on parle de surdispersion et la loi *Binomiale Négative* est alors privilégiée (ou la loi quasi-Poisson).

Espérance	Variance
56.36	4004.24

Tableau 8 - Analyse de l'espérance et de la fréquence de la variable nombre de sinistre

Le phénomène de surdispersion présent dans le tableau ci-dessus entraîne donc l'utilisation de la loi *Binomiale Négative*. Une analyse graphique nous permet de confirmer notre première intuition.

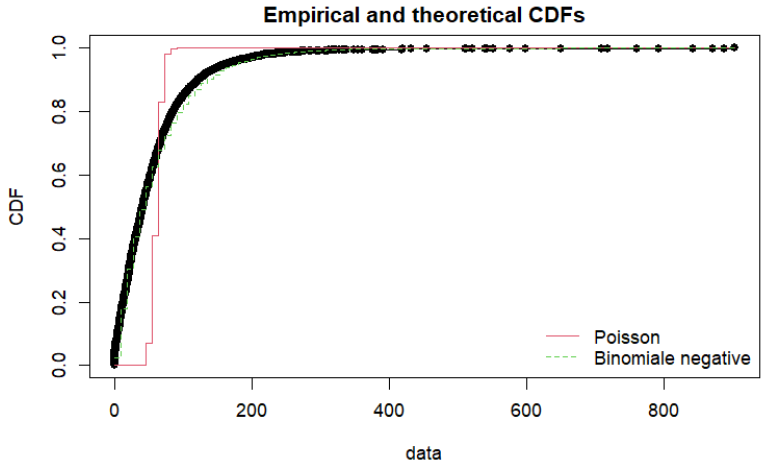


Figure 14 - Comparaison de la densité empirique avec la loi de Poisson et la loi Binomiale Négative

La loi *Binomiale Négative* sera donc utilisée comme composante aléatoire pour la modélisation de la fréquence. Par ailleurs, la fonction de lien utilisée est le « log » afin d'obtenir un modèle multiplicatif.

1.2. Choix de la composante aléatoire et de la fonction de lien sur le modèle coût moyen

Concernant le modèle coûts de sinistre il est coutume en assurance non-vie de choisir un GLM *Gamma* ou *Log-normal*.

Pour chaque loi une analyse de quatre graphiques est réalisée :

- Analyse de la densité empirique et théorique,
- Analyse du *QQ plot* : il consiste à tracer les quantiles de données empiriques par rapport aux quantiles théoriques. Ce graphique permet d'évaluer dans quelle mesure un modèle théorique particulier décrit une distribution de données,
- Analyse de la fonction de répartition : ce graphique fournit un tracé de la distribution empirique et de la distribution de la loi choisie,
- Analyse du *PP plot* : permet de comparer nos données empiriques à un modèle théorique. Il trace la proportion théorique inférieure ou égale à chaque valeur observée par rapport à la proportion réelle.

Loi *Log-normale* :

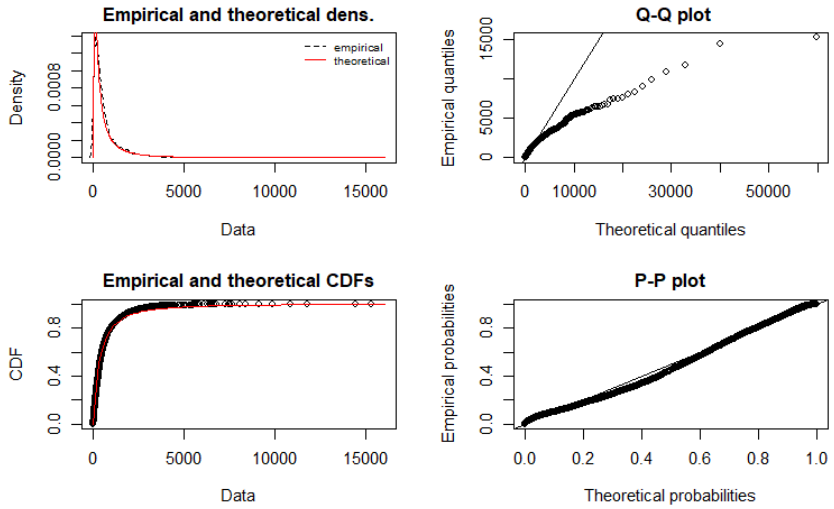


Figure 15 - Analyse de l'adéquation de nos données empiriques à la loi Log-normale

L'analyse graphique met en lumière le fait que la loi *Log-normale* permet de bien modéliser les données qui possèdent un coût moyen peu élevé. Nous constatons sur le Q-Q plot que cette loi surestime très fortement les sinistres ayant un coût élevé.

Loi *Gamma* :

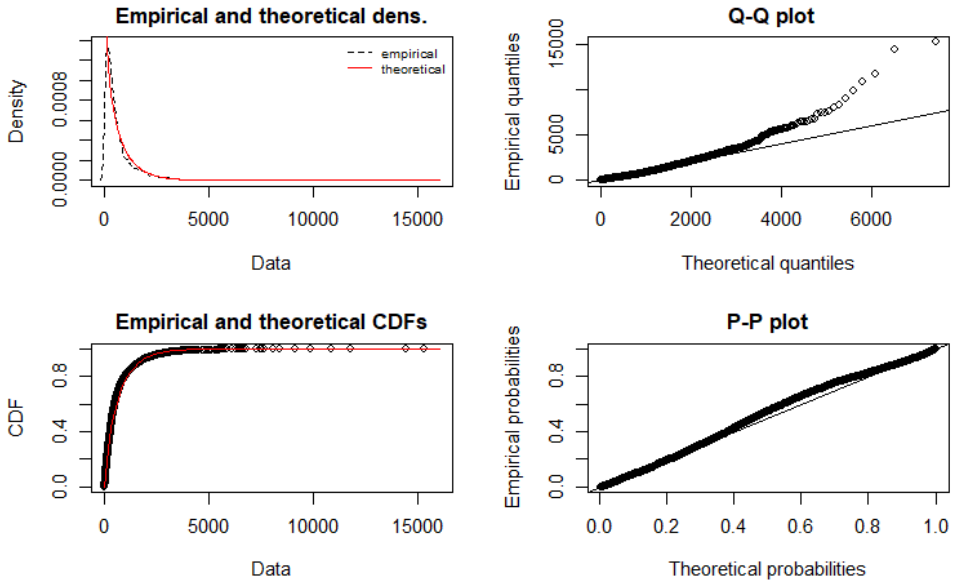


Figure 16 - Analyse de l'adéquation de nos données empiriques à la loi Gamma

L'analyse graphique de l'adéquation de la loi *Gamma* indique que cette loi permet également de bien modéliser les données qui ont un coût moyen peu élevé. A contrario, nous constatons sur le *Q-Q Plot* que cette loi sous-estime très fortement les sinistres ayant un coût élevé.

L'analyse graphique ne permet pas de choisir objectivement la loi qui modélise le mieux nos données empiriques. L'utilisation du test *Goodness of statistics* semble donc nécessaire afin de choisir objectivement la loi qui modélisera le coût moyen des sinistres.

Goodness-of-fit statistics	Loi *log-normale*	Loi *Gamma*
Kolmogorov-Smirnov statistic	0.06258518	0.05735264
Akaike's Information Criterion	85530.55	85338.63
Bayesian Information Criterion	85543.85	85351.92

Tableau 9 - Statistique d'adéquation des lois de probabilité

Malheureusement, le test de d'adéquation des lois n'est pas concluant ; les p-values ne nous permettent pas d'accepter l'hypothèse d'une loi *Log-normale* ou gamme. Après recherche, la loi de *Pareto* semble plus indiquée :

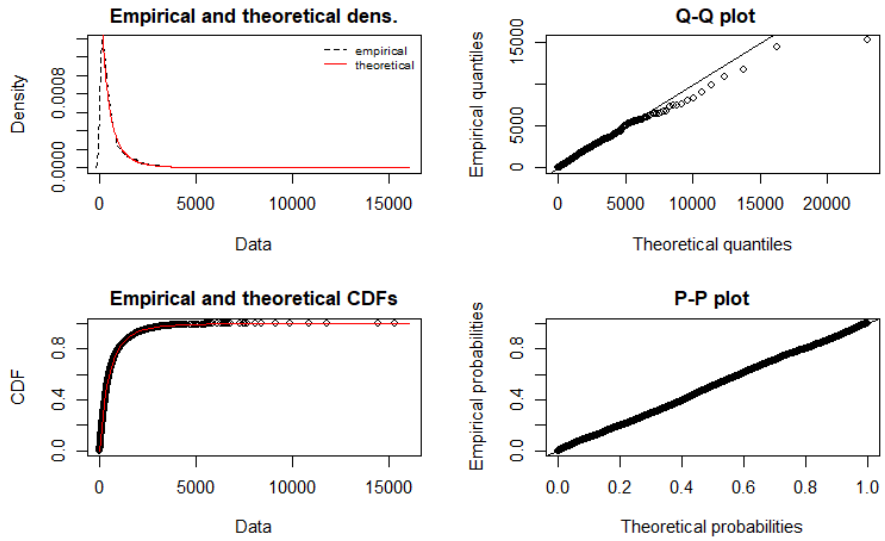


Figure 17 - Analyse de l'adéquation de nos données empiriques à la loi de Pareto

L'explication provient du fait que le portefeuille analysé correspond à une population de senior. Et comme vu précédemment les typologies et fréquences de consommations sont différentes. Ainsi le poste hospitalisation est plus élevé en termes de coût des actes (le coût des hospitalisations est plus lourd que le coût moyen des autres familles d'actes). Comparativement aux autres postes de soins, ces sinistres peuvent être considérés comme des sinistres « graves » au sens de l'assurance non-vie.

Par ailleurs, deux solutions sont envisageables pour contourner le problème :

- Réaliser une analyse des postes de santé séparément (analyse modulaire),
- Exclure les lignés liées aux remboursements du poste « Hospitalisation » qui seraient traitées à part.

Afin de valider cette dernière hypothèse, les remboursements liés à la variable « Hospitalisation » ont été exclus et l'adéquation des lois *Log-normale* et *Gamma* vérifiée.

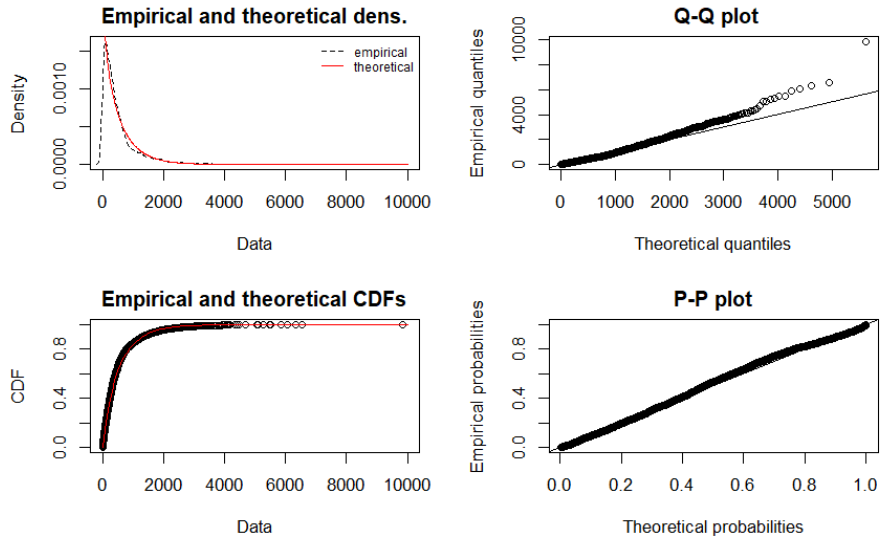


Figure 18 - Analyse de l'adéquation de nos données empiriques à la loi de Gamma avec exclusion des actes « Hospitalisation »

Goodness-of-fit statistics	Loi *Log-normale*	Loi *Gamma*
Kolmogorov-Smirnov statistic	0.06658775	0.04544076
Akaike's Information Criterion	82801.86	82418.82
Bayesian Information Criterion	82815.14	82432.10

Tableau 10 - Statistique d'adéquation des lois de probabilité avec exclusion des actes « Hospitalisation »

Nous constatons cette fois que l'analyse graphique du *Q-Q plot* est plus en phase avec nos données (bien que le traitement de valeurs extrêmes ne soit toujours pas satisfaisant). De plus, la loi *Gamma* valide l'hypothèse nulle du test de *Kolmogorov-Smirnov*.

Dans le contexte de cette étude et afin de ne pas tronquer le modèle, nous choisissons de garder les actes d'Hospitalisation du modèle initial, cela malgré une adéquation imparfaite de la loi *Gamma* et de la loi *Log-normale*. Nous n'utiliserons pas la loi de *Pareto* car celle-ci ne figure pas dans la famille exponentielle et à ce titre nous ne pouvons pas réaliser de régression linéaire généralisée.

A des fins de simplification nous appliquerons la loi *Gamma* dans notre étude car elle fait partie de la famille exponentielle et c'est la loi qui minimise le critère *AIC* et *BIC*. Naturellement, dans un cadre classique de tarification d'un contrat santé, il faudrait réaliser différents modèles :

- Soit un modèle pour les sinistres attritionnels et un autre pour les graves
- Soit un modèle par typologie d'acte : un modèle sur les soins courants, un modèle pour l'optique, …

Par ailleurs, pour le modèle coût moyen de notre étude, la fonction de lien utilisée est le « *log* ».
Avant de poursuivre dans la modélisation il est nécessaire de sélectionner les variables explicatives et de vérifier la corrélation entre ces variables. La section suivante présente cette analyse.

1.3. Étude de corrélation sur les deux modèles

Lorsque deux variables sont fortement corrélées (en pratique le seuil est fixé à 0,70), alors l'information recueillie est redondante car apportée par les deux variables. Intégrer dans le modèle des informations identiques présente quelques risques pour l'estimation des coefficients de régression. En effet différentes problématiques peuvent survenir :

- Le signe des coefficients de régression peut ne pas être en accord avec les connaissances du domaine. Par exemple dans un modèle de crédibilité si le coefficient associé à la modalité d'un bon conducteur (bonus 50) vaut 1 alors que celui de la modalité d'un conducteur sinistré (bonus 1,5) vaut 0,8 cela signifie qu'un bon conducteur payerait 20% plus cher que le conducteur sinistré ce qui n'est pas logique,
- Les variances peuvent être supérieures à tel point que les variables ne seraient potentiellement plus significatives,
- Instabilité des résultats : l'ajout ou la suppression d'une variable peut entrainer une modification significative des coefficients de régression.

Toutes ces raisons conduisent à réaliser une analyse fine des corrélations potentielles entre les variables. Une matrice de corrélation a été réalisée sur chaque modèle afin de déterminer la présence ou non de corrélation.

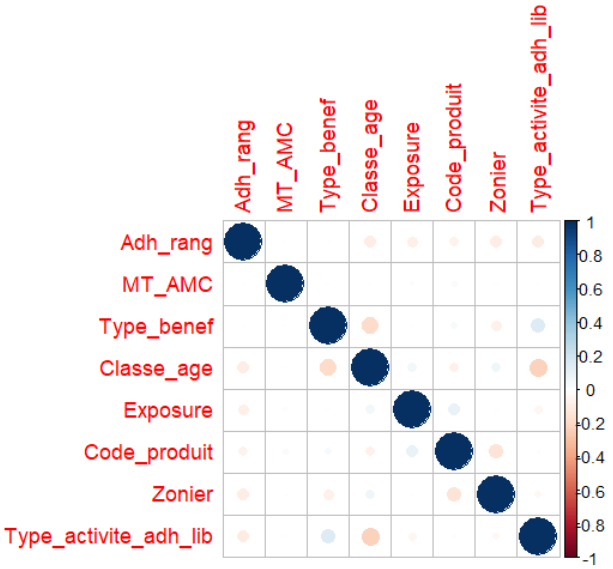


Figure 19 – Matrice de corrélation modèle coût de sinistre

Aucune corrélation notable n'est spécifiée à travers l'étude de la matrice de corrélation, les modèles peuvent ainsi être mis en place.

1.4. Les résultats des modélisations

1.4.1.1. Modèle Coût Moyen

Un premier GLM a été implémenté sur la base d'apprentissage avec l'ensemble des variables explicatives. Les résultats sont présentés ci-dessous :

```
> summary(glm_mnt_ini)

Call:
glm(formula = MT_CM ~ Classe_age + Type_activite_adh_lib + Type_benef +
    Code_produit + Zonier + offset(log(Exposure)), family = Gamma(link = "log"),
    data = complet_CM_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8519  -1.0967  -0.5074   0.1333   7.2690

Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                            6.51802    0.99673   6.539 6.63e-11 ***
Classe_ageClasse 26 - 35                               0.87387    1.08571   0.805 0.420916
Classe_ageClasse 46 - 55                              -1.22701    0.89749  -1.367 0.171620
Classe_ageClasse 56 - 65                              -0.84778    0.94232  -0.900 0.368326
Classe_ageClasse 66 - 75                              -0.73568    0.94249  -0.781 0.435083
Classe_ageClasse 76 - 85                              -0.43428    0.94364  -0.460 0.645374
Classe_ageClasse 86 - 95                               0.05513    0.94729   0.058 0.953593
Classe_ageClasse 95 et +                               0.02960    1.01325   0.029 0.976696
Classe_ageClasse 96 et +                              -0.42654    1.04364  -0.409 0.682772
Type_activite_adh_libARTISAN [AR]                      0.73243    0.48227   1.519 0.128876
Type_activite_adh_libAUTO ENTREPRENEUR [AE]           -0.17453    0.44998  -0.388 0.698131
Type_activite_adh_libCOMMERCANT [CO]                  -0.62970    0.68734  -0.916 0.359626
Type_activite_adh_libDEMANDEUR D'EMPLOI [DE]           0.18893    0.33885   0.558 0.577166
Type_activite_adh_libETUDIANT [ET]                     0.20330    0.54102   0.376 0.707095
Type_activite_adh_libPROFESSION LIBERALE [PL]          0.34532    0.42531   0.812 0.416856
Type_activite_adh_libPROFESSION SANTE SALARIE [SS]     0.40843    0.36773   1.111 0.266748
Type_activite_adh_libPROFESSIONNEL DE SANTE LIBERAL [PS]  0.51833    0.44143   1.174 0.240361
Type_activite_adh_libRETRAITE [RE]                     0.36424    0.32003   1.138 0.255098
Type_activite_adh_libSALARIE [SA]                      0.41223    0.32688   1.261 0.207318
Type_activite_adh_libSANS ACTIVITE []                  0.50052    0.32397   1.545 0.122401
Type_benefConj                                        -0.11187    0.04862  -2.301 0.021419 *
Type_benefEnf                                         -0.84370    0.90656  -0.931 0.352061
Code_produitSEN02                                     -0.02922    0.05443  -0.537 0.591373
Code_produitSEN03                                      0.32176    0.05373   5.989 2.22e-09 ***
Code_produitSEN04                                      0.75441    0.06702  11.256  < 2e-16 ***
ZonierZone 2                                           0.07757    0.07760   1.000 0.317502
ZonierZone 3                                           0.05914    0.08435   0.701 0.483257
ZonierZone 4                                           0.22255    0.06759   3.293 0.000997 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 2.215929)

    Null deviance: 9849.4  on 6752  degrees of freedom
Residual deviance: 9080.0  on 6725  degrees of freedom
AIC: 100284

Number of Fisher Scoring iterations: 13
```

Seules cinq variables sont significatives :

- `Intercept` : la constante du modèle,
- `Type_benefConj` : les conjoints présents sur le contrat de l'adhérent,
- `Code_produitSEN03` : La garantie Retraite de niveau 3,
- `Code_produitSEN04` : La garantie Retraite de niveau 4,
- `ZonierZone 4` : Les départements de la zone 4. Il s'agit des départements administratifs où La Mutuelle Verte a constaté le niveau de consommation par bénéficiaire le plus élevé.

Malgré la création de classe d'âge nous constatons qu'aucune variable `Classe_ageClasse` n'est significative. De la même manière aucune variable concernant le type d'activité de l'adhérent n'est significative.

Une analyse globale des variables explicatives est réalisée via la procédure *Anova* :

```
> anova(glm_mnt_ini,test="Chisq")
Analysis of Deviance Table

Model: Gamma, link: log

Response: MT_CM

Terms added sequentially (first to last)


                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                  6752     9849.4
Classe_age            8   223.43      6744     9626.0 < 2.2e-16 ***
Type_activite_adh_lib 11    45.27      6733     9580.7 0.0397661 *
Type_benef            2      9.51      6731     9571.2 0.1169660
Code_produit          3   448.92      6728     9122.3 < 2.2e-16 ***
Zonier                3    42.30      6725     9080.0 0.0002622 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

La variable `Type_bénef` représentant les types de bénéficiaire (adhérent, conjoint, enfant) n'apparait pas significative (rejet de l'hypothèse nulle). En revanche la classe d'âge est bien significative ainsi de nouveaux groupes doivent être crées afin de rendre significative cette variable. La variable `Type_activite` est significative de la même manière de nouveaux groupes doivent être réalisés.

Concomitamment, la procédure *stepwise* est utilisée. Elle permet de rechercher la combinaison optimale et donc de sélectionner les variables explicatives.

Les résultats de la procédure sont détaillés ci-dessous :

```
> summary(glm_mnt_Stepwise)

Call:
glm(formula = MT_CM ~ Classe_age + Code_produit + Zonier + offset(log(Exposure)),
    family = Gamma(link = "log"), data = complet_CM_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8794  -1.0967  -0.5141   0.1293   7.3017

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.07149    0.23289  26.070  < 2e-16 ***
Classe_ageClasse 26 - 35  0.89558    1.08971   0.822 0.411193
Classe_ageClasse 46 - 55 -0.48359    0.35060  -1.379 0.167849
Classe_ageClasse 56 - 65 -0.06577    0.22557  -0.292 0.770612
Classe_ageClasse 66 - 75  0.05000    0.22412   0.223 0.823471
Classe_ageClasse 76 - 85  0.35423    0.22813   1.553 0.120536
Classe_ageClasse 86 - 95  0.84478    0.24284   3.479 0.000507 ***
Classe_ageClasse 95 et +  0.84881    0.43782   1.939 0.052577 .
Classe_ageClasse 96 et +  0.41300    0.50639   0.816 0.414771
Code_produitSEN02        -0.02743    0.05502  -0.499 0.618114
Code_produitSEN03         0.32367    0.05417   5.975 2.41e-09 ***
Code_produitSEN04         0.75537    0.06774  11.152  < 2e-16 ***
ZonierZone 2              0.08476    0.07822   1.084 0.278592
ZonierZone 3              0.07628    0.08522   0.895 0.370744
ZonierZone 4              0.24399    0.06816   3.580 0.000346 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 2.273548)

    Null deviance: 9849.4  on 6752  degrees of freedom
Residual deviance: 9122.6  on 6738  degrees of freedom
AIC: 100295

Number of Fisher Scoring iterations: 12
```

La sélection de variable a permis de supprimer les variables non pertinentes : `Type_benef` ainsi que `Type_activite_adh_lib`. Nous constatons toutefois que les coefficients ne sont pas tous significatifs.

Afin d'y remédier de nouveaux groupes ont été créés.

Retraitements effectués :

- Etant sur une gamme « Retraite » seules les classes d'âge des adhérents principaux ont été sélectionnées, toutes les autres classes d'âge ont été regroupées,
- Les enfants et les conjoints ont été regroupés dans un groupe « Ayant droit »,
- Les zones géographiques 1,2 et 3 ont été regroupées car trop peu représentatives,
- Les produits (garanties) Retraite 1 et Retraite 2 ont été regroupés.

```
> glm_mnt_ini <- glm(formula = MT_CM ~  Classe_age_rec +Type_benef_rec + Code_produit_rec + Zonier_rec +
offset (log(Exposure)), data = complet_CM_train, family =Gamma(link="log"))
Call:
glm(formula = MT_CM ~ Classe_age_rec + Type_benef_rec + Code_produit_rec +
    Zonier_rec + offset(log(Exposure)), family = Gamma(link = "log"),
    data = complet_CM_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8632  -1.0935  -0.5106   0.1351   7.5431

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.06810    0.05045 120.287  < 2e-16 ***
Classe_age_recClasse 66 - 75  0.11430    0.04393   2.602  0.00929 **
Classe_age_recClasse 76 - 85  0.41747    0.06242   6.688 2.44e-11 ***
Classe_age_recClasse 86 - 95  0.91438    0.10545   8.671  < 2e-16 ***
Classe_age_recClasse 95 et +  0.74713    0.29602   2.524  0.01163 *
Type_benef_recAyant droit    -0.08427    0.04753  -1.773  0.07625 .
Code_produit_recSEN03         0.34858    0.04083   8.537  < 2e-16 ***
Code_produit_recSEN04         0.77292    0.05754  13.433  < 2e-16 ***
Zonier_recZone 4              0.17556    0.03975   4.417 1.02e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 2.324486)

    Null deviance: 9849.4  on 6752  degrees of freedom
Residual deviance: 9127.4  on 6744  degrees of freedom
AIC: 100288

Number of Fisher Scoring iterations: 7
```

A cette étape, tous les coefficients du modèle sont significatifs.
Nous pouvons déterminer un intervalle de confiance de nos coefficients.

```
> confint(glm_mnt_Stepwise)
Attente de la réalisation du profilage...
                               2.5 %     97.5 %
(Intercept)                5.63651585 6.56375300
Classe_ageClasse 26 - 35 -0.74514927 4.05763051
Classe_ageClasse 46 - 55 -1.16754026 0.23112495
Classe_ageClasse 56 - 65 -0.54649919 0.35521884
Classe_ageClasse 66 - 75 -0.42861859 0.46820729
Classe_ageClasse 76 - 85 -0.13088047 0.78103865
Classe_ageClasse 86 - 95  0.33574950 1.30383967
Classe_ageClasse 95 et +  0.02865831 1.78809323
Classe_ageClasse 96 et + -0.50202851 1.53330561
Code_produitSEN02        -0.13667699 0.08067202
Code_produitSEN03         0.21625730 0.42977931
Code_produitSEN04         0.62244476 0.88884922
ZonierZone 2             -0.07079451 0.23777667
ZonierZone 3             -0.09178016 0.24329755
ZonierZone 4              0.10748803 0.37605699
```

Toutefois, il est primordial de vérifier la pertinence du modèle. Un premier test rapide repose sur la vérification de la déviance et les résidus de déviance. Le rapport de ces deux valeurs doit être proche de 1.

```
> deviance(glm_mnt_Stepwise) / df.residual(glm_mnt_Stepwise)
[1] 1.150382
```

Le modèle semble correct. Pour aller plus loin, une méthode classique consiste à analyser les résidus issus de ce modèle.

Dans un modèle linéaire classique, la variable expliquée est décomposée en une partie explicative et une partie résiduelle supposée vérifier des propriétés spécifiques, notamment la normalité. Dans cette étude nous travaillons avec des modèles linéaires généralisés et dans notre cas il n'y a pas ce type de décomposition. Il est ainsi plus difficile d'apprécier la validité des hypothèses formulées aussi bien sur le modèle lui-même que sur la loi des observations.

En s'inspirant de la théorie du modèle linéaire classique, deux types de résidus peuvent être analysés :

- Les Résidus de la Déviance : $\widehat{D} = 2\sum_{i=1}^{n} Y_i \ln\left(\frac{Y_i}{\hat{p}(xi)}\right) + (1 - Y_i) \ln\left(\frac{1 - Yi}{1 - \hat{p}(xi)}\right))$
  avec $x_i = (1, x_{1,i}, \ldots, x_{p,i})$.

- Les Résidus de *Pearson* : *Pour tout i ∈{1, . . . , n}, on appelle* $i^{ème}$ *résidus de Pearson la réalisation de* $\hat{e}_i = \frac{Y_i - \hat{\lambda}(x_i)}{\sqrt{\hat{\lambda}(x_i)}}$ *avec Y la variable à expliquer et x les variables explicatives*

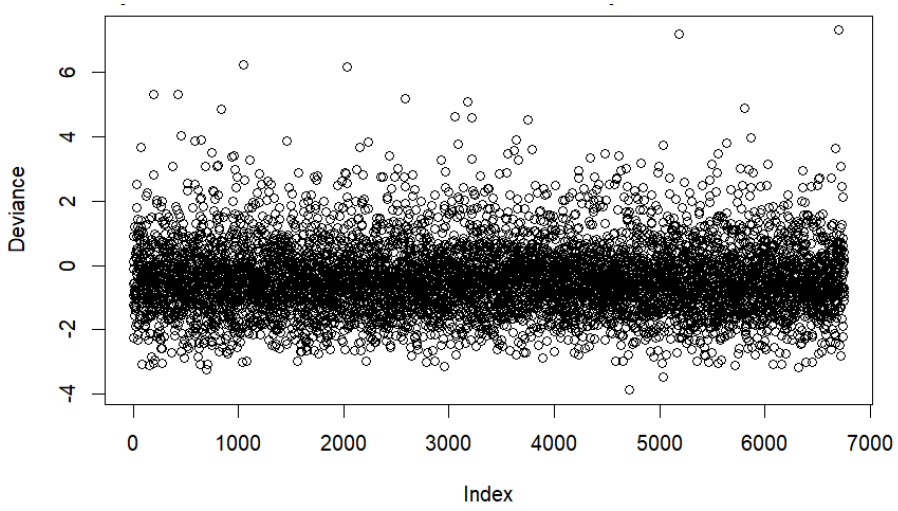


Figure 20 - Modèle coût de sinistre - Représentation des résidus de la déviance

Concernant les résidus de déviance : l'analyse graphique permet d'estimer la validité du modèle comme « correcte » (mais pas certaine). En effet, les résidus observés se situent autour de l'axe des abscisses, avec une variance constante selon l'observation i, autrement dit si le nuage de points est de forme cylindrique autour de l'axe des abscisses. Le critère est assez subjectif mais facilement lisible sur le graphique.

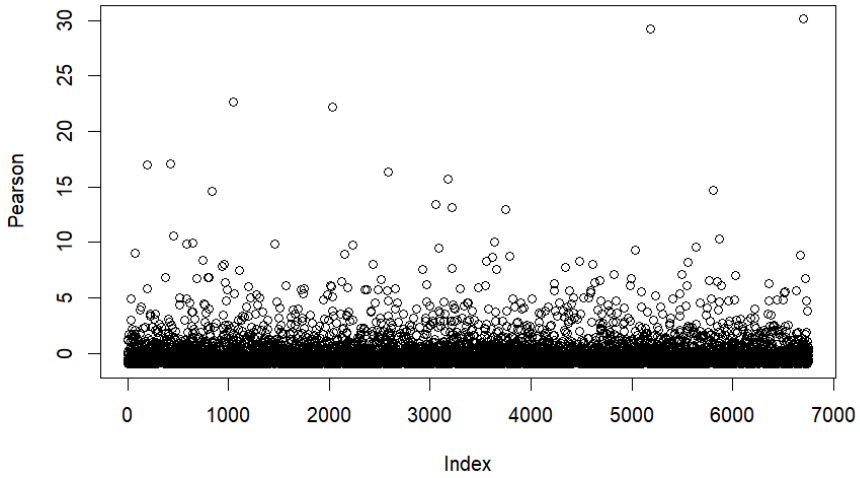


Figure 21 - Modèle coût de sinistre - Représentation des résidus de Pearson

Le graphique des résidus de *Pearson* ne nous permet pas de conclure sur la validité du modèle mais indique qu'une grande majorité des résidus sont nulles.

Notons que certains résidus sont significativement non nul et indiquent que la modélisation effectuée n'est pas parfaire. Naturellement ce résultat nous semble logique dans la mesure où nous avions constaté, lors de l'étude de l'adéquation de nos données empiriques à la loi *Gamma,* une incohérence notamment lorsque le coût moyen des sinistres est élevé.

Pour vérifier le pouvoir prédictif du modèle celui-ci est implémenté sur la base de test. Le *RMSE* est ainsi calculé :

Modèle Coût de sinistre	RMSE
Méthode GLM	24,14

Tableau 11 - Modèle Coût moyen : Valeur du RMSE sur la base de test

Cette information nous servira à comparer les modèles par la suite.

1.4.1.2. Modèle Fréquence

Pour le modèle fréquence, la survenance 2020 a été supprimée. En effet, l'année 2020 est atypique à cause de la pandémie du *Covid-19*. La France a connu un premier confinement strict en mars-avril 2020 de ce fait les déplacements des Français étaient limités et une partie des professionnels de santé n'ont pas exercé sur cette période biaisant ainsi la fréquence de la survenance 2020.

Pour le reste, la même méthodologie que précédemment est appliquée. Le premier modèle n'a pas été concluant et a également demandé un retraitement des variables explicatives. La démarche est détaillée en annexe 2, les résultats finaux sont présentés ci-dessous :

```
> glm_freq_Stepwise_bi <-step(glm_freq_bi,direction = c("both", "backward", "forward"),trace=0)
> summary(glm_freq_Stepwise_bi)

Call:
glm.nb(formula = nb_acte ~ Classe_age_rec + Code_produit + Zonier_rec +
    offset(log(Exposure)), data = complet_Freq_train, init.theta = 1.139640485,
    link = "log")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2582  -0.9271  -0.2811   0.3087   5.7893

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  2.80839    0.14861  18.898  < 2e-16 ***
Classe_age_recClasse 26 - 55  0.52270    0.24657   2.120   0.0340 *
Classe_age_recClasse 56 - 65  0.78316    0.14359   5.454 4.93e-08 ***
Classe_age_recClasse 66 - 75  0.97140    0.14235   6.824 8.84e-12 ***
Classe_age_recClasse 76 - 85  1.31827    0.14591   9.035  < 2e-16 ***
Classe_age_recClasse 86 - 95  1.67396    0.15964  10.486  < 2e-16 ***
Classe_age_recClasse 95 et +  1.54098    0.29767   5.177 2.26e-07 ***
Classe_age_recClasse 96 et +  0.98413    0.41158   2.391   0.0168 *
Code_produitSEN02             0.06859    0.04095   1.675   0.0939 .
Code_produitSEN03             0.21252    0.04054   5.243 1.58e-07 ***
Code_produitSEN04             0.31387    0.04985   6.296 3.06e-10 ***
Zonier_recZone 2 - 3          0.09944    0.05482   1.814   0.0697 .
Zonier_recZone 4              0.22774    0.05139   4.431 9.36e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.1396) family taken to be 1)

    Null deviance: 5977.6  on 4816  degrees of freedom
Residual deviance: 5628.2  on 4804  degrees of freedom
AIC: 47507

Number of Fisher Scoring iterations: 1


              Theta:  1.1396
          Std. Err.:  0.0229

 2 x log-likelihood:  -47479.2780
```

Pour rappel la loi *Binomiale Négative* a été utilisée car nous sommes dans un cas de surdispersion.

Tous les coefficients sont significatifs, nous pouvons sélectionner le modèle et réaliser un intervalle de confiance de nos paramètres :

```
> confint.default(glm_freq_Stepwise_bi)
                                2.5 %    97.5 %
(Intercept)                  2.51712921 3.0996571
Classe_age_recClasse 26 - 55  0.03943023 1.0059602
Classe_age_recClasse 56 - 65  0.50172394 1.0646049
Classe_age_recClasse 66 - 75  0.69240758 1.2503953
Classe_age_recClasse 76 - 85  1.03229855 1.6042363
Classe_age_recClasse 86 - 95  1.36107526 1.9868499
Classe_age_recClasse 95 et +  0.95755398 2.1243977
Classe_age_recClasse 96 et +  0.17745837 1.7908109
Code_produitSEN02            -0.01166294 0.1488529
Code_produitSEN03             0.13307605 0.2919736
Code_produitSEN04             0.21615550 0.4115814
Zonier_recZone 2 - 3         -0.00800672 0.2068884
Zonier_recZone 4              0.12701085 0.3284596
```

De nouveau, il est important de vérifier la pertinence du modèle.
Un premier test rapide est de vérifier la déviance et les résidus de déviance. Le rapport des deux valeurs doit être proche de 1.

```
> deviance(glm_freq_Stepwise_bi) / df.residual(glm_freq_Stepwise_bi)
[1] 1.171555
```

Le modèle semble correct car très proche de 1. L’analyse des résidus est ensuite réalisée :

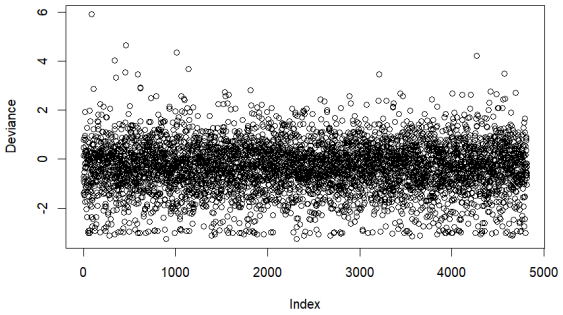


Figure 22 - Modèle Fréquence - Représentation des résidus de la déviance

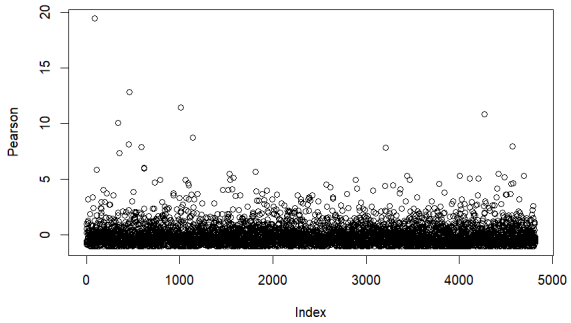


Figure 23 - Modèle Fréquence - Représentation des résidus de Pearson

Le nuage de points des résidus de Déviance est bien de forme cylindrique autour de l'axe des abscisses accréditant dans une certaine mesure la justesse du modèle. Les résidus de *Pearson* indiquent également qu'une grande partie des résidus sont nulles ou proches de 0 signe d'une modélisation correcte de nos données empiriques.

Pour vérifier le pouvoir prédictif du modèle celui-ci est implémenté sur la base de test. Le RMSE est ainsi calculé :

Modèle Fréquence	RMSE
Méthode GLM	57,82

Tableau 12 - Modèle Fréquence : Valeur du RMSE sur la base de test

Compte tenu de l'absence de validité des modèles par un test statistique il parait pertinent de challenger ces résultats à travers d'autres méthodes de modélisation.

2. Modélisation via les *Random Forest*

Cette section présente la modélisation de la prime pure par la méthode des *Random Forest*. Le principe est identique ; nous avons deux modèles : un modèle coût moyen et un modèle fréquence. De la même manière que précédemment le modèle est implémenté sur la base d'apprentissage puis amélioré (ici via le tuning des hyperparamètres) et dans un dernier temps nous analysons son pouvoir prédictif sur la base de test.

2.1. Modèle Coût Moyen

La base de données a été modifiée de manière à ne présenter que des variables numériques. Le modèle a ensuite été implémenté sur R :

```
> rf <- randomForest(data=db, MT_CM ~ ., ntree = 50, mtry = sqrt(ncol(db)-1),
+                    nodesize = 1,maxnodes=NULL)
```

Les quatre hyperparamètres ont été subjectivement sélectionnés pour cette première analyse :

- `ntree` correspondant au nombre d'arbres. Ce paramètre a été initialisé à 50.
- `mtry` détermine le nombre de variables à prendre dans chaque nœud. Il permet de réduire l'effet de la corrélation entre les arbres d'une forêt.
- `nodesize` indique le nombre d'observations minimal devant être présent par feuille. S'il n'y a pas de critère d'arrêt, la profondeur des arbres peut mener dans un cas extrême à un modèle saturé, c'est pourquoi on cherche à maîtriser l'effet de surapprentissage. Pour cette première forêt le paramètre a été initialisé à 1.
- `maxnodes` : permet de stopper la construction de l'arbre. Dans ce cas, la construction dudit arbre se terminera lorsque le nombre de nœuds terminaux atteint maxnodes. Valeur initialisée à NULL.

Afin d'optimiser le modèle les hyperparamètres vont être calibrés. Concrètement, le réglage des hyperparamètres repose davantage sur des résultats expérimentaux que sur la théorie. Par conséquent, la meilleure méthode pour déterminer les paramètres optimaux consiste à essayer de nombreuses combinaisons différentes pour évaluer les performances de chaque modèle.

Cependant, évaluer et optimiser notre modèle uniquement sur notre jeu de données peut conduire à l'un des problèmes les plus fondamentaux de l'apprentissage automatique : le surapprentissage.
En effet, si nous optimisons le modèle sur notre base de données, notre modèle obtiendra de très bons résultats, sans être automatiquement généralisable pour autant. Afin de pallier ce problème une validation croisée[17] doit être réalisée. Cependant, sur ce modèle précis ce point n'est pas obligatoire. En effet, le défaut majeur de l'arbre de décision est que sa performance est fortement dépendante de l'échantillon de données d'initial. Or en utilisant une multitude d'arbres (d'où le terme de « Forêt » aléatoire) nous parons ce problème.

Afin de calibrer nos hyperparamètres, chaque combinaison de paramètres à essayer peut être répertoriée au moyen d'une *Grid Search* ; une méthode qui évalue toutes les combinaisons que nous définissons. L'analyse de *RMSE* nous servira à déterminer le meilleur paramètre.

[17] Se référer à l'annexe 5 pour le détail de cette procédure.

Analyse du meilleur `mtry` :

```
> set.seed(1234)
> tuneGrid <- expand.grid(.mtry = c(1:5 ))
> rf_mtry <- train(CM~.,
+                  data = db,
+                  method = "rf",
+                  metric = "RMSE",
+                  tuneGrid = tuneGrid,
+                  importance = TRUE,
+                  nodesize = 1,
+                  ntree =100)
> print(rf_mtry)
Random Forest

6753 samples
   6 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 6753, 6753, 6753, 6753, 6753, 6753, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared     MAE
  1     21.91194  0.016741600  8.746323
  2     22.74953  0.015056160  8.980249
  3     23.73066  0.012495476  9.321863
  4     24.64775  0.010885238  9.633539
  5     25.75464  0.009967031  9.852675

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 1.
```

La valeur de `mtry` minimisant le *RMSE* est 1, soit une seule variable à prendre dans chaque nœud.

```
> store_nodesize <- list()
> tuneGrid <- expand.grid(.mtry = best_mtry)
> for (nodesize in c(1:30)) {
+   set.seed(1234)
+   rf_nodesize <- train(CM~.,
+                        data = db,
+                        method = "rf",
+                        metric = "RMSE",
+                        tuneGrid = tuneGrid,
+                        importance = TRUE,
+                        maxnodes = NULL,
+                        nodesize = nodesize,
+                        ntree = 100)
+   current_iteration <- toString(nodesize)
+   store_nodesize[[current_iteration]] <- rf_nodesize
+ }
> results_mtry <- resamples(store_nodesize)
> summary(results_mtry)

Call:
summary.resamples(object = results_mtry)

Models: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26,
27, 28, 29, 30
Number of resamples: 25
RMSE
       Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
1  14.92092 20.18059 22.27654 22.06836 24.34769 30.23406    0
2  14.88140 20.11040 22.53962 22.07072 24.36816 30.06490    0
3  14.78433 19.93666 22.28287 22.05295 24.27505 30.08027    0
4  14.89391 19.97727 22.50001 22.05991 24.24659 30.12816    0
5  14.79661 19.78334 22.29874 22.04604 24.37126 30.07092    0
6  14.80586 20.38956 22.18449 22.00557 24.31202 30.02887    0
7  14.56193 19.95782 22.36394 22.05087 24.33770 30.01992    0
8  14.94027 20.29554 21.96926 22.03484 24.34994 30.08906    0
9  14.76283 19.87557 21.91641 22.00908 24.23978 30.06177    0
10 14.67345 20.03708 22.12035 22.01903 24.25861 29.88810    0
11 14.82600 19.91798 21.87217 22.01417 24.36502 29.99782    0
12 14.69861 19.76836 21.88246 21.95756 24.30823 30.03033    0
13 14.67616 19.69773 22.20964 21.98351 24.36828 30.02196    0
14 14.86509 19.92512 22.14010 21.97749 24.32317 29.92996    0
15 14.72739 19.90137 22.35215 22.01576 24.20871 30.00145    0
16 14.76639 19.80384 21.99983 21.96083 24.33885 29.86784    0
17 14.61375 19.84359 22.62761 21.99771 24.27866 29.98324    0
18 14.73785 19.93430 22.05127 22.04889 24.32218 30.09766    0
19 14.70690 19.91220 22.09177 21.99066 24.17216 30.02014    0
20 14.63524 20.43575 21.98658 21.98875 24.27315 30.01112    0
21 14.81014 19.88503 22.31638 21.99306 24.24798 29.97346    0
22 14.82500 19.96583 22.26794 22.00189 24.24630 30.06964    0
23 14.65821 19.79305 21.93876 22.00761 24.61400 29.98367    0
24 14.61318 20.06452 22.08362 21.99970 24.27731 29.99357    0
25 14.66773 19.83361 21.90354 21.97919 24.31723 30.08849    0
26 14.73800 20.01199 22.08454 21.99383 24.31299 30.03981    0
27 14.71693 19.74188 22.43976 21.96111 24.30483 30.05179    0
28 14.88577 19.96928 22.08157 22.07809 24.18582 29.99927    0
29 14.68394 19.85678 21.88076 21.96110 24.26302 29.97319    0
30 14.59812 20.05470 21.92484 21.95090 24.22932 30.02631    0
```

De la même manière, la valeur `nodesize` qui minimise le *RMSE* maximum (valeur surlignée en jaune dans la sortie R ci-dessus) est 16.

```
> ## Search the best ntrees
> store_maxtrees <- list()
> for (ntree in c(5, 10, 50, 100, 250, 500)) {
+   set.seed(5678)
+   rf_maxtrees <- train(CM ~.,
+                        data = db,
+                        method = "rf",
+                        metric = "RMSE",
+                        tuneGrid = tuneGrid,
+                        importance = TRUE,
+                        nodesize = 1,
+                        maxnodes = 16,
+                        ntree = ntree)
+   key <- toString(ntree)
+   store_maxtrees[[key]] <- rf_maxtrees
+ }
> results_tree <- resamples(store_maxtrees)
> summary(results_tree)

Call:
summary.resamples(object = results_tree)

Models: 5, 10, 50, 100, 250, 500
Number of resamples: 25

RMSE
        Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
5   15.68978 21.74285 25.26208 24.74046 28.09255 31.19346    0
10  16.02709 21.79386 24.91802 24.68960 27.93362 31.10833    0
50  15.84576 20.81406 24.72719 24.51498 27.85544 31.12500    0
100 15.73156 20.75871 24.76173 24.49744 27.83922 31.07181    0
250 15.68239 20.73389 24.73921 24.49085 27.85361 31.07551    0
500 15.67518 20.71136 24.70423 24.48361 27.85131 31.07691    0
```

Avec la même logique, le nombre d'arbres `ntrees` optimal est de 100.
Par ailleurs, le paramètre `maxnode` ne sera pas utilisé dans la mesure où il n'a pas permis de diminuer le *RMSE*.

A cette étape, tous les paramètres ont été optimisés. Ainsi la *Random Forest* finale peut être effectuée :

```
> fit_rf <- train(CM ~.,
+                 db,
+                 method = "rf",
+                 metric = "RMSE",
+                 tuneGrid = tuneGrid,
+                 importance = TRUE,
+                 nodesize = 16,
+                 ntree = 100,
+                 maxnodes = NULL)
```

Le modèle de *Random Forest* renvoie un objet « importance » : il s'agit de la diminution moyenne de l'impureté apportée par chaque variable. Elle est calculée par l'indice de *Gini*[18] : la diminution pour chaque nœud est cumulée, puis une moyenne sur l'ensemble des arbres est effectuée.

[18] L'indice (ou coefficient) de Gini est un indicateur synthétique permettant de rendre compte du niveau d'inégalité pour une variable et sur une population donnée. Il varie entre 0 (égalité parfaite) et 1 (inégalité extrême). Entre 0 et 1, l'inégalité est d'autant plus forte que l'indice de Gini est élevé.
Il est égal à 0 dans une situation d'égalité parfaite où la variable prend une valeur identique sur l'ensemble de la population. À l'autre extrême, il est égal à 1 dans la situation la plus inégalitaire possible, où la variable vaut 0 sur toute la population à l'exception d'un seul individu.

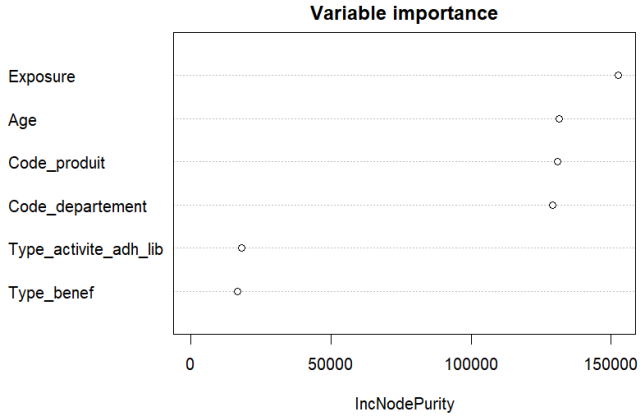


Figure 24 - Random Forest finale - Importance des variables explicatives

Quatre variables ont un poids significatif dans ce modèle. L'apport d'une variable supplémentaire ne diminue plus significativement l'impureté du modèle. Ainsi les variables explicatives les plus significatives pour le modèle coût de sinistre sont :

- L'exposition calculée,
- L'âge du bénéficiaire,
- Le code produit (la garantie du bénéficiaire),
- Le département de l'adhérent.

Les résultats du modèle sur la base de test sont présentés ci-dessous :

Modèle Coût de sinistre Méthode *Random Forest*	***RMSE***
Sans optimisation	24,88
Avec optimisation	23,62
Pourcentage d'écart	-5%

Tableau 13 - Modèle Coût de sinistre : Récapitulatif des modèles Random Forest réalisés

En comparant les prévisions vis-à-vis de la base test nous constatons que l'optimisation réalisée du modèle a permis d'améliorer le RMSE de 5%. Pour information, le temps de calcul est d'environ 3h.

2.2. Modèle Fréquence

La même méthode est utilisée pour le modèle Fréquence. La base de données initiale a été modifiée de manière à ne présenter que des variables numériques. Le modèle a ensuite été implémenté sur R :

```
rf.freq <- randomForest(data=db.freq, nb_acte ~ ., ntree = 10, mtry = sqrt(ncol(db)-1),
           nodesize = 1,maxnodes=NULL)+ nodesize = 1,maxnodes=NULL)
```

Les quatre hyperparamètres ont ensuite été calibrés (le détail est présenté en annexe 3) :

- `ntree` = 250,
- `mtry` = 1,

- `nodesize` = 18,
- `maxnodes` = NULL.

La représentation des différentes segmentations sous forme d'arbre de décision n'est pas réalisable. Toutefois, l'impact des variables explicatives sur la variable réponse (ici coût moyen) peut être envisagé par la mesure de leur importance.

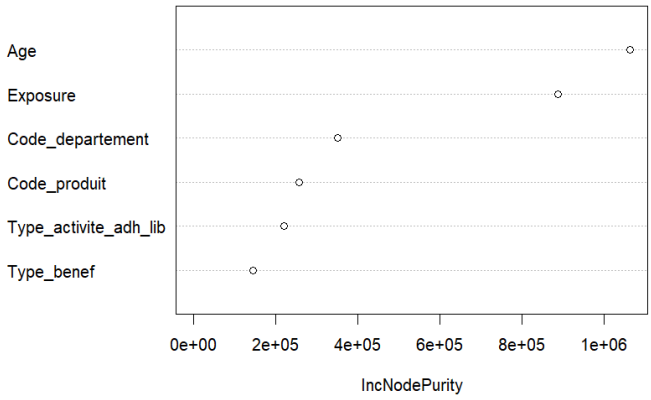


Figure 25 - Modèle fréquence : Importance des variables

Les variables explicatives les plus significatives pour le modèle fréquence sont :

- L'exposition,
- L'âge du bénéficiaire.

Le modèle optimisé est appliqué à la base test afin d'effectuer la validation croisée et d'évaluer le pouvoir prédictif du modèle. Une comparaison avec le modèle initial est également réalisée permettant de quantifier l'amélioration apportée par l'optimisation des hyperparamètres :

Modèle Fréquence Méthode *Random Forest*	***RMSE***
Sans optimisation	57,87
Avec optimisation	57,43
Pourcentage d'écart	-0.76%

Tableau 14 – Modèle Fréquence : Récapitulatif des modèles Random Forest réalisés

L'optimisation permet d'améliorer le *RMSE* vis-à-vis de la base test. Toutefois l'impact est assez faible. Compte tenu du temps de calcul (environ 3h) le gain est marginal.

3. XGboost

L'algorithme *XGBoost* est un algorithme ensembliste agrégeant des arbres de décision. A chaque itération, l'arbre construit apprend de l'erreur de son prédécesseur et la corrige dans le sens du gradient. Deux modèles sont analysés (sur la base d'apprentissage) : le coût moyen et la fréquence. Ils

seront ensuite implémentés sur la base de test et leurs *RMSE* seront comparés aux autres modélisations réalisées.

3.1. Modèle Coût Moyen

La base de données a été modifiée de manière à ne présenter que des variables numériques. Le modèle a ensuite été implémenté sur R :

```
bst_CM_ini= xgboost(data=xgbMatrix_CM,method = "xgbTree", eta = 0, gamma =0,max_depth = 10,
nrounds=100, colsample_bytree=1, verbose = 0, eval_metric = "error")
```

Le modèle *XGBoost* s'optimise à travers ses différents hyperparamètres.

- `nround` est le nombre d'arbres à implémenter,
- `max_depth` correspondant à la profondeur d'arbre maximale,
- `colsample_bytree` est le pourcentage des variables utilisées pour construire un modèle,
- `eta` est le taux d'apprentissage,
- `gamma` correspond à la régularité du modèle ; plus le paramètre est grand plus le modèle sera lisse,
- `min_child_weight` est le nombre d'observations minimum dans chaque nœud pour poursuivre le développement de l'arbre,
- `subsample` est le pourcentage des observations utilisées pour construire un arbre.

Les paramètres sont indépendants les uns envers les autres. Par conséquent, l'optimisation du modèle *XGBoost* a été effectuée en testant toutes les combinaisons des paramètres. La combinaison minimisant le *RMSE* sur l'échantillon d'apprentissage est sélectionnée. Et afin de déterminer le meilleur paramètre en évitant le surapprentissage l'approche de la validation croisée été réalisée sur l'échantillon d'apprentissage (le principe de la validation croisée est présentée en annexe 5).

```
xgb_trcontrol = trainControl(method = "cv", number = 5, allowParallel = TRUE,
    verboseIter = FALSE, returnData = FALSE)

xgbGrid <- expand.grid(nrounds = c(10,50,100,200,500),
                       max_depth = c(1:10),
                       colsample_bytree = 0.5,
                       eta = c(0.1,0.2,0.3,0.4,0.5),
                       gamma=c(0:10),
                       min_child_weight = 1,
                       subsample = 1
                      )

set.seed(0)
xgb_model = train(dataXGB_CM, complet_CM_trainXG$MT_CM, trControl = xgb_trcontrol, tuneGrid = xgbGrid,
    method = "xgbTree", drop = FALSE)
```

Finalement, la meilleure combinaison est la suivante :

Hyperparamètre	Valeur optimale
nround	50
max_depth	1
colsample_bytree	0,74
eta	0,1
gamma	7
min_child_weight	1
subsample	1

Tableau 15 - Modèle Coût de sinistre : Valeurs optimisées des hyperparamètres

Comme pour les *Random Forest* le modèle *XgBoost* permet d'appréhender l'importance des variables.

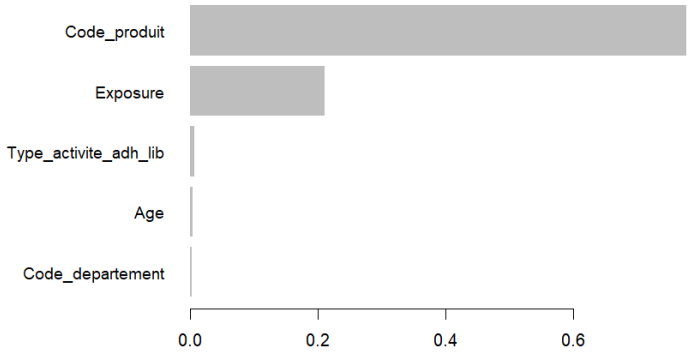


Figure 26 - Xgboost : Importance des variables

Ainsi, le code produit et l'exposition sont les variables les plus significatives. Les autres variables sont nettement moins importantes.

Afin de vérifier la qualité de prédiction du modèle, celui-ci est appliqué à notre échantillon test. Une comparaison avec le modèle initial est également réalisée permettant de quantifier l'amélioration apportée par l'optimisation des hyperparamètres :

Modèle Coût du sinistre Méthode *XgBoost*	*RMSE*
Sans optimisation	26,16
Avec optimisation	23,52
Pourcentage d'écart	-10%

Tableau 16 - Récapitulatif des modèles XgBoost réalisés sur le modèle coût de sinistre

L'optimisation du modèle a permis de gagner 10% de précision. Il est donc primordial d'optimiser les paramètres afin d'avoir le modèle le plus précis possible. Une attention particulière doit être portée au surapprentissage, c'est la raison pour laquelle les paramètres sont choisis en validation croisée (*V5-Fold* dans notre cas). Notons par ailleurs que cette opération entraine un temps de calcul conséquent (8h) et donc non négligeable pour l'entreprise.

3.2. Modèle Fréquence

La même méthode a été utilisée sur le modèle Fréquence. Ainsi, la meilleure combinaison des hyperparamètres est la suivante :

Hyperparamètre	Valeur optimale
nround	100
max_depth	10
colsample_bytree	0,74
eta	0,1
gamma	0
min_child_weight	1
subsample	1

Tableau 17 - Modèle Fréquence : Valeurs optimisées des hyperparamètres

Le modèle optimisé nous permet d'appréhender l'importance des variables :

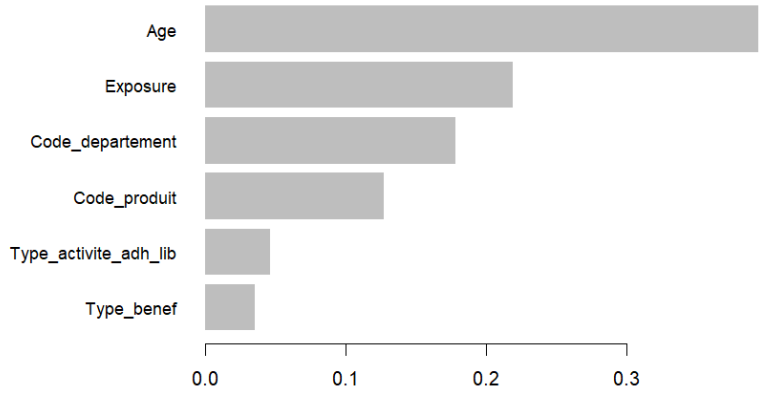


Figure 27 - Xgboost : Importance des variables

Ainsi, la variable « âge » est la variable la plus importante du modèle Fréquence. Les variables exposition, département et produit permettent d'affiner le modèle.

Afin de vérifier la qualité de prédiction du modèle, celui-ci est appliqué à notre échantillon test. Une comparaison avec le modèle initial est également réalisée permettant de quantifier l'amélioration apportée par l'optimisation des hyperparamètres :

Modèle Fréquence Méthode *XgBoost*	*RMSE*
Sans optimisation	82,86
Avec optimisation	57,77
Pourcentage d'écart	-30%

Tableau 18 - Récapitulatif des modèles XgBoost réalisés sur le modèle Fréquence

Le gain en termes de précision est significatif (-30%) confirmant le besoin d'optimiser les modèles initiaux.

4. Comparaison des modèles

La comparaison finale des modèles se fait sur l'échantillon de test, jusqu'à présent inutilisé. Cette étape a pour but de déterminer quel modèle est le plus efficient pour répondre aux objectifs de l'étude.

Un premier point de comparaison est le pouvoir prédictif associé à chaque modèle. Les modèles optimisés sont utilisés pour prédire les coûts moyens des sinistres et la fréquence de l'échantillon de test. Les résultats obtenus sont présentés dans le tableau ci-dessous :

	***RMSE* - Modèle Coût de sinistre**	***RMSE* - Méthode Fréquence**	**Temps de calcul**
GLM	24,14	57,82	Immédiat
Random Forest	23,62	57,91	5h
XgBoost	23,52	57,77	16h

Tableau 19 - Récapitulatif des différentes modélisations réalisées

Dans un premier temps, nous constatons que le pouvoir prédictif des modèles est assez similaire :

- Sur le modèle de Coût : le modèle *GLM* de Coût est légèrement en deçà dû à l'utilisation de la loi *Gamma* qui n'est pas optimale. Toutefois le recours à l'usage d'autres méthodes ne permet pas d'améliorer de manière significative (moins de 0,5%) le modèle *GLM* et compte tenu du nombre d'adhérents sur la gamme « Retraite » le gain est marginal.
- Sur le modèle Fréquence : les résultats sont très proches (moins de 0,2%)

Le temps de calcul nécessaire à l'établissement des différentes modélisations est également un facteur à prendre en compte.

- Pour le modèle *GLM* : les résultats sont instantanés mais nécessitent un retraitement manuel avec un regroupement de variables afin de toutes les rendre significatives.
- Pour les algorithmes de *Random Forest* et de *XgBoost* : ces méthodes nécessitent un tuning des hyperparamètres et une *cross validation* pour éviter le phénomène de surapprentissage. Cette méthodologie s'avère longue et donc coûteuse pour l'entreprise.

Comparaison de l'importance des variables entre les deux algorithmes de *Machine Learning* :

Il est intéressant de noter que l'importance des variables n'est pas forcément similaire entre la méthode *Random Forest et le XgBoost*.

- Pour le modèle coûts : Quatre variables sont significatives dans le modèle *Random Forest* : L'exposition calculée, l'âge du bénéficiaire, le code produit (la garantie du bénéficiaire), le département de l'adhérent. Tandis que dans le modèle *XgBoost* seules les variables code produit et exposition sont significatives. En modifiant `max_depth` (la profondeur d'arbre maximale) du modèle *XgBoost* nous obtenons les quatre mêmes variables significatives que dans le modèle *Random Forest*. Le calibrage des modèles est donc très sensible aux modifications des paramètres de tuning.

- Pour le modèle fréquence : L'ordre des variables (en poids d'importance) est identique, cependant le poids attribué à certaines variables varie (par exemple le poids de la variable exposition est plus important dans le modèle *Random Forest)*.

Globalement les deux modèles sont assez différents dans la mesure où les variables explicatives n'ont pas le même poids d'importance. Toutefois les résultats finaux sont assez similaires notons seulement un temps de calcul beaucoup plus important pour Le XgBoost qui nécessite une validation croisée pour éviter le phénomène de surapprentissage.

Avantages et limites des méthodes :

- **Pour le modèle *GLM* :** Présentant des résultats aisément interprétables, les modèles linéaires généralisés font ainsi preuve d'une très bonne applicabilité opérationnelle. De plus, ces modèles paramétriques sont construits dans un cadre théorique leur permettant d'effectuer des tests statistiques capables d'évaluer leur qualité, jouissant alors d'une robustesse assez satisfaisante. Cependant, deux points essentiels ne sont pas couverts par les *GLM* : la prise en compte des effets non linéaires de certaines variables nominales et la détection (et la modélisation) des interactions entre certaines variables quantitatives ou qualitatives. En effet, par construction, les impacts des variables explicatives sur la variable cible ne peuvent être modélisés que linéairement dans le cas d'un *GLM*.

- **Pour le modèle *Random Forest* :** il permet de classer les variables explicatives par ordre d'importance dans la prévision. Cette méthode est particulièrement avantageuse pour la gestion de données volumineuses. De plus, contrairement à d'autres algorithmes, il n'y a que très peu d'hyperparamètres à gérer. En effet, il faut simplement se contenter du nombre d'arbres décisionnels qui la compose ainsi que de la profondeur maximale de chaque arbre. Il est ainsi possible de réaliser un calibrage plus aisé des paramètres. Le dernier avantage n'est pas des moindres : avec le *Random Forest*, il n'y a pas de surapprentissage. L'inconvénient réside dans le fait que les résultats sont moins facilement interprétables que les modèles *GLM* et que le *tuning* des hyperparamètres nécessite un temps de calcul important.

- **Pour le modèle *XgBoost* :** à l'heure actuelle lorsqu'il s'agit de données de petite à moyenne taille, l'algorithme *XGBoost* est considéré comme le meilleur de sa catégorie. Grâce à son principe d'auto-amélioration séquentielle, il peut être utilisé pour résoudre des problèmes de régression, de classification et même de classement. Toutefois, ce modèle sacrifie l'intelligibilité en se comportant comme une boite noire. Notons néanmoins qu'il permet de classer les variables explicatives par ordre d'importance dans la prévision (information cruciale dans notre étude). Il est également important de préciser que ce modèle nécessite comme beaucoup d'algorithmes basés sur des arbres de décision, une attention particulière face au phénomène de surapprentissage. De ce fait une validation croisée doit être réalisée, augmentant considérablement le temps de calcul pour l'optimisation des paramètres (8h par modèle).

Améliorations possibles des modèles dans le cas de notre étude :

Une amélioration potentielle de nos prédictions réside dans le fait d'intégrer des données externes à notre base de données, car elles permettraient de capter des informations supplémentaires non communiquées à la souscription du contrat.

De plus, nous avons fait le choix de conserver la loi *Gamma* dans notre étude malgré le fait qu'elle ne soit pas adaptée au sinistre grave présent dans notre portefeuille. En assurance santé, il est d'usage de calculer pour chaque catégorie d'acte médical la prime pure associée. Ainsi la prime pure totale du contrat s'obtient en sommant les primes pures de chacun des actes :

$$Prime\ pure\ totale\ = \sum_{i\ \epsilon\ Famille\ d'Actes} Prime\ pure\ de\ la\ Famille\ d'Acte\ i$$
$$= \sum_{i\ \epsilon\ Famille\ d'Actes} Fréquence(i)\ \times Coût\ moyen(i)$$

La principale source d'amélioration de notre modélisation est donc de réaliser un modèle pour chaque famille de soins. Cela permettra de limiter les erreurs liées à la différence de coût entre les sinistres attritionnels et graves.

CONCLUSION

L'objet de cette étude est de mettre en avant la modélisation de la tarification d'un contrat santé, en détaillant le fonctionnement du secteur de l'assurance santé en France. Dès lors, nous avons pu constater son caractère évolutif, notamment par la mise en place de nombreuses réformes ces dernières années, engendrant inéluctablement un impact significatif sur les complémentaires santé. C'est pour cette raison que ces dernières doivent impérativement prendre en compte et suivre rigoureusement ces évolutions règlementaires, lors de leurs analyses actuarielles et notamment la tarification d'un contrat santé.

Après avoir présenté les différentes étapes de la tarification d'un contrat santé ainsi que les risques inhérents à l'activité de l'assureur ont donc été présentés. Nous avons pris soin de détailler le besoin de segmentation, tout en portant une attention particulière à l'aléa moral et à l'antisélection.

Différentes modélisations possibles de la prime pure ont ensuite été décrites au moyen de méthodes paramétriques et non paramétriques. Suite à l'étude du portefeuille nous avons appliqué les modélisations suivantes : un modèle classique *GLM* puis deux algorithmes non paramétriques le *Random Forest* et le *XgBoost.*

Le modèle *GLM* est le plus utilisé en assurance de par sa relative simplicité de mise en œuvre et d'interprétabilité des résultats. En analysant les données empiriques, nous nous sommes aperçus que l'approche d'une tarification globale n'est pas pertinente. En effet, l'adéquation de la loi *Gamma* n'est pas optimale, tenant compte du fait qu'elle sous-estime les sinistres graves. Il est ainsi nécessaire de segmenter les modèles par famille d'acte afin d'être plus pertinent dans l'approche tarifaire. Deux autres modélisations ont été réalisées ; ces algorithmes non paramétriques permettent de ne pas dépendre de loi de probabilité théorique comme la loi *Gamma*. A des fins de comparaison entre les différentes modélisations, nous avons conservé une tarification globale et non par famille d'acte.

Les résultats obtenus démontrent que les algorithmes de *Machine Learning* peuvent améliorer les performances d'un *GLM* sans pour autant être incontestablement les meilleurs. De plus, compte tenu du temps de calcul et d'optimisation important que nécessitent les algorithmes de *Machine Learning*, nous estimons qu'à ce niveau de segmentation leur utilisation n'est pas justifiée. Néanmoins, dans le cas où nous disposerions d'une base de données plus conséquente, il est probable que les résultats obtenus à partir des algorithmes de *Machine Learning* seraient bien meilleurs que les résultats obtenus par *GLM*. La précision de ces résultats justifierait alors un temps de calcul plus conséquent.

Par ailleurs, la réalisation de cette étude met en lumière l'insuffisance du nombre d'informations récoltées sur les assurés dans la base de données. Il est donc nécessaire pour les compagnies d'assurance d'obtenir des informations pertinentes sur les assurés, pouvant être utiles à la tarification d'un contrat. Il serait également profitable pour les assureurs, de recourir à l'ajout de variables exogènes afin de compléter leur modèle : le nombre de lits médicalisés, de praticiens, des informations climatiques, … Toutes ces données enrichiraient alors considérablement les modèles de Machine Learning, augmentant possiblement leur usage par les assureurs.

BIBILIOGRAPHIE

- CHARPENTIER A., DUTANG C. [2012] « L'actuariat avec R », https://cran.r-project.org/doc/contrib/Charpentier_Dutang_actuariat_avec_R.pdf, page 39 à 52.
- CHÉRY C. [2012] « Construction d'un outil de tarification de contrats complémentaire santé », http://www.ressources-actuarielles.net/.
- CHESNEAU C. [2022] « Modèles de régression », https://chesneau.users.lmno.cnrs.fr/Reg-M2.pdf, page 97 à 107.
- CTIP [2022] « Cahier statistique 2021 des institutions de prévoyance, édition 2022 », https://ctip.asso.fr/wp-content/uploads/2022/10/20220510CTIP-Cahier-statistiques-2021-PAP.pdf.
- DELLA-VEDOVA C. [2019] « nettoyer-et-valider-les-donnees-avec-r », https://delladata.fr/nettoyer-et-valider-les-donnees-avec-r/.
- EL JERDY M. [2008] « Tarification des groupes en assurance santé », http://www.ressources-actuarielles.net/.
- EL KHALOUI A. [2018] « [Tuto] Boost ton ML : XGBoost facile & efficace avec R ! », https://datafuture.fr/post/faire-tourner-xgboost-sous-r/.
- FAVRE-BEGUET M. [2021] : « Cours Protection Sociale », Cours ISFA.
- KARAMOKO FOFANA C.H. [2015] « Approche tarifaire des contrats collectifs Frais de Santé à l'aide des méthodes d'apprentissage », http://www.ressources-actuarielles.net/.
- LAGADEC F. [2009] « Tarification d'un contrat de complémentaire santé par un Modèle Linéaire Généralisé », http://www.ressources-actuarielles.net/.
- LA SÉCURITÉ SOCIALE [2022] « Notre environnement : la Sécurité sociale » https://assurance-maladie.ameli.fr/qui-sommes-nous/organisation/securite-sociale/securite-sociale.
- LA SÉCURITÉ SOCIALE [2021] « Les chiffres clés de la sécurité sociale 2020 », https://www.securite-sociale.fr/files/live/sites/SSFR/files/medias/DSS/2021/CHIFFRES%20CLES%202020%20ED2021.pdf.
- MASSÉ A. « Aide à l'utilisation du logiciel R », https://sites.google.com/site/rgraphiques/home/filtrer-et-trier-des-donn%C3%A9es-avec-r#h.p_ZucR_n2v70Yc, site internet.
- MILHAUD X. [2020] : « Pratiques avancée de tarification et de provisionnement en assurance non-vie », cours ISFA.
- MORIN J.B. [2012] « La tarification en santé », http://www.ressources-actuarielles.net/.
- MUTUALITÉ FRANCAISE [2021] « RAPPORT D'ACTIVITÉ 2021 Assemblée générale 2022 », https://docs.mutualite.fr/rapport-activite-2021/2/.
- NDIAYE A. [2020] « Estimation des Prestations, PSAP et Intervalles de confiance en assurance santé : méthodes d'agrégation et réseaux de neurones », http://www.ressources-actuarielles.net/.
- PLANCHET F., MISERAY A. [2017] « Tarification IARD Introduction aux techniques avancées », cours ISFA
- PIETTE P. [2022] « Statistical Learning for Actuaries », cours ISFA.
- ROUVIERE L. [2015] : « Sélection-"validation" de modèles », cours univ-rennes2.
- SAN MARTIN G. [2016] « GLM : Generalized Linear Models », https://www.cellulestat.cra.wallonie.be/wpcontent/uploads/2016/12/Formation_Stats_3_1_GLM.pdf Cours Centre Wallon de Recherche Agronomique.
- TOUTAIN F.H. [2018] « Création d'un outil de tarification santé », http://www.ressources-actuarielles.net/.

LISTE DES FIGURES

LISTE DES TABLEAUX

ANNEXES

Annexe 1 : Estimation des coefficients empiriques

Le facteur de l'âge influe fortement sur la consommation médicale d'un assuré. C'est d'ailleurs la raison pour laquelle les assureurs segmentent cette population et proposent des gammes axées vers ces profils de risque.

Par exemple, tous postes confondus, un assuré de 80 ans consommera en moyenne plus qu'un assuré de 25 ans car de manière générale, la consommation médicale s'accroît avec l'âge. Il existe néanmoins quelques exceptions à cette règle :

- Les nouveau-nés consomment plus que les jeunes enfants,
- Un pic de consommation est atteint autour de 15 ans : il est lié aux dépenses d'orthodontie importantes à cet âge.

Partant de ce constat, il est d'usage de calculer la consommation moyenne (somme des fréquences empiriques × coût moyens empiriques sur l'ensemble des actes) en fonction de l'âge.

La prime pure est alors dépendante de l'âge. A ce stade, certains retraitements sont nécessaires :

- Un regroupement en classes d'âge pour réunir les individus et éviter une volatilité trop importante (principalement pour les âges élevés ou très jeunes ; globalement là où les effectifs sont très réduits),
- Un lissage de la courbe afin d'éliminer les sauts entre les âges, liés à des consommations atypiques. Des méthodes couramment utilisées à cet effet existent, telles que l'utilisation de *splines* ou la méthode de *Whittaker-Henderson.*

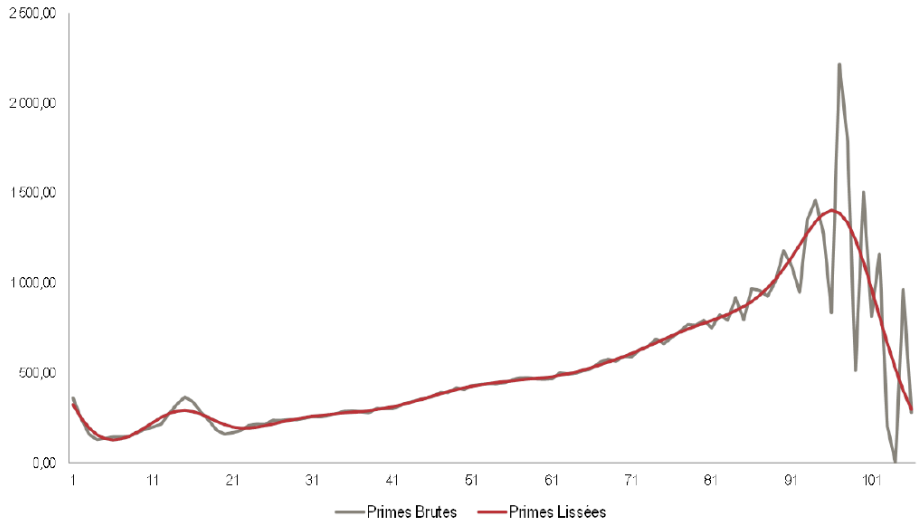


Figure 28 - Illustration de la méthode de Whittaker-Henderson sur la courbe de la prime pure selon l'âge

L'effet du lissage sur les primes (en rouge sur le graphique) permet d'éviter les fluctuations. La forte volatilité des âges avancés s'explique par le manque d'adhérents sur ces tranches d'âge entraînant une plus grande volatilité des primes. La consommation décroît aux âges élevés : cela se justifie en partie par la forte proportion d'assurés âgés en Affection de Longue Durée (ALD), mieux pris en charge par

l'Assurance Maladie (limitant mécaniquement l'utilisation du contrat de complémentaire santé). Par ailleurs les personnes ayant un âge avancé effectuent moins d'actes onéreux (faible consommation en dentaire car toutes les dents ont potentiellement étaient remplacées par un dentier par exemple, difficulté à se déplacer pour faire modifier sa correction optique, …) ce qui explique également la diminution de la consommation de santé.

Calcul de coefficients correcteurs

Une fois obtenue cette prime pure par âge, il est possible d'y ajouter l'impact d'autres variables par l'application de coefficients correcteurs. Ces coefficients s'obtiennent en calculant le rapport entre la consommation moyenne d'une catégorie et celle de l'ensemble de la population.

Généralement, les variables fortement discriminantes et donc retenues pour le calcul des correctifs sont la CSP et la zone géographique. Dans notre cas les adhérents sont des séniors donc la CSP n'est pas utilisée, seule la zone géographique impacte la prime pure.

Annexe 2 : Modélisation GLM : Modèle Fréquence

Le modèle initial présente peu de variables significatives :

```
> glm_freq_bi <- glm.nb(formula = nb_acte ~  Classe_age + Type_benef + Code_produit + Zonier +
Type_activite_adh_lib + offset (log(Exposure)), data = complet_Freq_train, link="log")
> summary(glm_freq_bi)

Call:
glm.nb(formula = nb_acte ~ Classe_age + Type_benef + Code_produit +
    Zonier + Type_activite_adh_lib + offset(log(Exposure)), data = complet_Freq_train,
    link = "log", init.theta = 1.14491146)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2665  -0.9319  -0.2788   0.3080   6.0411

Coefficients:
                                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                             2.631e+00  1.018e+00   2.585  0.00975 **
Classe_ageClasse 26 - 35                                1.063e+00  5.827e-01   1.825  0.06798 .
Classe_ageClasse 36 - 45                               -1.784e+01  1.276e+03  -0.014  0.98884
Classe_ageClasse 46 - 55                                5.299e-01  9.627e-01   0.550  0.58203
Classe_ageClasse 56 - 65                                7.895e-01  9.905e-01   0.797  0.42540
Classe_ageClasse 66 - 75                                9.595e-01  9.906e-01   0.969  0.33275
Classe_ageClasse 76 - 85                                1.305e+00  9.913e-01   1.317  0.18797
Classe_ageClasse 86 - 95                                1.663e+00  9.934e-01   1.674  0.09418 .
Classe_ageClasse 95 et +                                1.508e+00  1.024e+00   1.472  0.14094
Classe_ageClasse 96 et +                                9.720e-01  1.063e+00   0.914  0.36067
Type_benefConj                                         -5.606e-02  3.671e-02  -1.527  0.12679
Type_benefEnf                                          -8.918e-02  9.774e-01  -0.091  0.92730
Code_produitSEN02                                       6.892e-02  4.095e-02   1.683  0.09236 .
Code_produitSEN03                                       2.118e-01  4.065e-02   5.210 1.89e-07 ***
Code_produitSEN04                                       3.115e-01  4.997e-02   6.234 4.55e-10 ***
ZonierZone 2                                            1.248e-01  5.863e-02   2.129  0.03323 *
ZonierZone 3                                            4.656e-02  6.455e-02   0.721  0.47074
ZonierZone 4                                            2.194e-01  5.150e-02   4.261 2.04e-05 ***
Type_activite_adh_libARTISAN [AR]                      -3.478e-02  3.514e-01  -0.099  0.92114
Type_activite_adh_libAUTO ENTREPRENEUR [AE]             8.062e-02  3.365e-01   0.240  0.81064
Type_activite_adh_libCOMMERCANT [CO]                   -3.751e-03  4.548e-01  -0.008  0.99342
Type_activite_adh_libDEMANDEUR D'EMPLOI [DE]            5.333e-02  2.460e-01   0.217  0.82838
Type_activite_adh_libETUDIANT [ET]                      4.148e-01  3.679e-01   1.128  0.25950
Type_activite_adh_libPROFESSION LIBERALE [PL]           6.077e-01  3.168e-01   1.918  0.05506 .
Type_activite_adh_libPROFESSION SANTE SALARIE [SS]      2.452e-01  2.719e-01   0.902  0.36725
Type_activite_adh_libPROFESSIONNEL DE SANTE LIBERAL [PS] -1.065e-01  3.268e-01  -0.326  0.74452
Type_activite_adh_libRETRAITE [RE]                      2.034e-01  2.322e-01   0.876  0.38091
Type_activite_adh_libSALARIE [SA]                       2.017e-01  2.375e-01   0.849  0.39565
Type_activite_adh_libSANS ACTIVITE []                   2.509e-01  2.356e-01   1.065  0.28702
Type_activite_adh_libSANS ACTIVITE [NA]                 1.818e-02  2.630e-01   0.069  0.94491
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.1449) family taken to be 1)
    Null deviance: 6002.1  on 4816  degrees of freedom
Residual deviance: 5625.4  on 4787  degrees of freedom
AIC: 47516
Number of Fisher Scoring iterations: 1


              Theta:  1.1449
          Std. Err.:  0.0230

 2 x log-likelihood:  -47453.6370
```

A cette étape, nous réalisons une *Anova* pour vérifier la significativité des variables explicatives :

```
> anova(glm_freq_bi,test="Chisq")
Avis : tests réalisés sans ré-estimer 'theta'Analysis of Deviance Table

Model: Negative Binomial(1.1449), link: log

Response: nb_acte

Terms added sequentially (first to last)


                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                  4816     6002.1
Classe_age            9  285.312      4807     5716.8 < 2.2e-16 ***
Type_benef            2    1.491      4805     5715.3    0.4746
Code_produit          3   45.059      4802     5670.3 8.991e-10 ***
Zonier                3   30.473      4799     5639.8 1.097e-06 ***
Type_activite_adh_lib 12   14.366      4787     5625.4    0.2780
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Il apparait que les variables `Type_benef` et `Type_activite_adh_lib` ne sont pas significatives, nous décidons donc de les supprimer de notre analyse. Afin de confirmer le point, nous réalisons une régression de type *Stepwise*. Les deux variables sont bien supprimées, nous obtenons alors un modèle avec beaucoup plus de variables significatives.

```
> glm_freq_Stepwise_bi <-step(glm_freq_bi,direction = c("both", "backward", "forward"),trace=0)
> summary(glm_freq_Stepwise_bi)

Call:
glm.nb(formula = nb_acte ~ Classe_age + Code_produit + Zonier +
    offset(log(Exposure)), data = complet_Freq_train, init.theta = 1.141350428,
    link = "log")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2605  -0.9265  -0.2757   0.3084   5.8939

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               2.83199    0.14840  19.084  < 2e-16 ***
Classe_ageClasse 26 - 35  0.90035    0.57892   1.555   0.1199
Classe_ageClasse 36 - 45 -17.37991  988.19515  -0.018   0.9860
Classe_ageClasse 46 - 55  0.47461    0.26289   1.805   0.0710 .
Classe_ageClasse 56 - 65  0.76074    0.14338   5.306 1.12e-07 ***
Classe_ageClasse 66 - 75  0.94729    0.14214   6.665 2.65e-11 ***
Classe_ageClasse 76 - 85  1.29698    0.14569   8.902  < 2e-16 ***
Classe_ageClasse 86 - 95  1.65255    0.15943  10.366  < 2e-16 ***
Classe_ageClasse 95 et +  1.51444    0.29741   5.092 3.54e-07 ***
Classe_ageClasse 96 et +  0.96636    0.41145   2.349   0.0188 *
Code_produitSEN02         0.06738    0.04093   1.646   0.0997 .
Code_produitSEN03         0.21108    0.04052   5.210 1.89e-07 ***
Code_produitSEN04         0.31663    0.04992   6.343 2.26e-10 ***
ZonierZone 2              0.12875    0.05849   2.201   0.0277 *
ZonierZone 3              0.04785    0.06452   0.742   0.4584
ZonierZone 4              0.22778    0.05138   4.433 9.30e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.1414) family taken to be 1)

    Null deviance: 5985.6  on 4816  degrees of freedom
Residual deviance: 5626.0  on 4801  degrees of freedom
AIC: 47504

Number of Fisher Scoring iterations: 1


              Theta:  1.1414
          Std. Err.:  0.0229

 2 x log-likelihood:  -47469.6900
```

Certaines variables ont toujours une p-value supérieure à 0,05. Nous choisissons donc de les regrouper avec d’autres classes.

```
> glm_freq_Stepwise_bi <-step(glm_freq_bi,direction = c("both", "backward", "forward"),trace=0)
> summary(glm_freq_Stepwise_bi)

Call:
glm.nb(formula = nb_acte ~ Classe_age_rec + Code_produit + Zonier_rec +
    offset(log(Exposure)), data = complet_Freq_train, init.theta = 1.139640485,
    link = "log")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2582  -0.9271  -0.2811   0.3087   5.7893

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  2.80839    0.14861  18.898  < 2e-16 ***
Classe_age_recClasse 26 - 55  0.52270    0.24657   2.120   0.0340 *
Classe_age_recClasse 56 - 65  0.78316    0.14359   5.454 4.93e-08 ***
Classe_age_recClasse 66 - 75  0.97140    0.14235   6.824 8.84e-12 ***
Classe_age_recClasse 76 - 85  1.31827    0.14591   9.035  < 2e-16 ***
Classe_age_recClasse 86 - 95  1.67396    0.15964  10.486  < 2e-16 ***
Classe_age_recClasse 95 et +  1.54098    0.29767   5.177 2.26e-07 ***
Classe_age_recClasse 96 et +  0.98413    0.41158   2.391   0.0168 *
Code_produitSEN02             0.06859    0.04095   1.675   0.0939 .
Code_produitSEN03             0.21252    0.04054   5.243 1.58e-07 ***
Code_produitSEN04             0.31387    0.04985   6.296 3.06e-10 ***
Zonier_recZone 2 - 3          0.09944    0.05482   1.814   0.0697 .
Zonier_recZone 4              0.22774    0.05139   4.431 9.36e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.1396) family taken to be 1)

    Null deviance: 5977.6  on 4816  degrees of freedom
Residual deviance: 5628.2  on 4804  degrees of freedom
AIC: 47507

Number of Fisher Scoring iterations: 1


              Theta:  1.1396
          Std. Err.:  0.0229

 2 x log-likelihood:  -47479.2780
```

A cette étape, notre modèle est bien paramétré, il faut ensuite valider les résidus du modèle. (cf. Section IV.1.4.1.2).

Annexe 3 : Random Forest : Modèle Fréquence

Choix du paramètre `mtry` :

```
> set.seed(1234)
> tuneGrid <- expand.grid(.mtry = c(1:6 ))
> rf_mtry <- train(nb_acte~.,
+                  data = db.freq,
+                  method = "rf",
+                  metric = "RMSE",
+                  tuneGrid = tuneGrid,
+                  importance = TRUE,
+                  nodesize = 18,
+                  ntree =100)
> print(rf_mtry)
Random Forest

4817 samples
   6 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 4817, 4817, 4817, 4817, 4817, 4817, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared    MAE
  1     58.78943  0.10650742  37.37662
  2     59.18974  0.09674767  37.24674
  3     60.03579  0.08705636  37.73281
  4     60.67875  0.08112318  38.17163
  5     61.05338  0.07773063  38.38781
  6     61.37937  0.07564281  38.54564

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 1.
```

Choix du paramètre `mtry` :

```
> store_maxtrees <- list()
> for (ntree in c(5,10,50,100,250,400,500)) {
+   set.seed(5678)
+   rf_maxtrees <- train(nb_acte~.,
+                        data = db.freq,
+                        method = "rf",
+                        metric = "RMSE",
+                        tuneGrid = tuneGrid,
+                        importance = TRUE,
+                        nodesize = 18,
+                        maxnodes = NULL,
+                        ntree = ntree)
+   key <- toString(ntree)
+   store_maxtrees[[key]] <- rf_maxtrees
+ }
> results_tree <- resamples(store_maxtrees)
> summary(results_tree)

Call:
summary.resamples(object = results_tree)

Models: 5, 10, 50, 100, 250, 400, 500
Number of resamples: 25

RMSE
        Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
5   50.23545 58.47925 61.05157 60.62143 63.22901 70.06463    0
10  49.92906 58.29305 60.51510 60.23202 62.85524 69.16020    0
50  49.85417 57.89396 60.29497 59.96175 62.62078 68.66647    0
100 49.81297 57.89123 60.32723 59.92207 62.49455 68.51084    0
250 49.76362 57.89031 60.32557 59.89319 62.41309 68.48728    0
400 49.80141 57.90288 60.33524 59.89774 62.41088 68.52639    0
500 49.80839 57.87675 60.32920 59.89446 62.38364 68.52994    0
```

Choix du paramètre `nodesize` :

```
> store_nodesize <- list()
> tuneGrid.freq <- expand.grid(.mtry = best_mtry)
> for (nodesize in c(1:30)) {
+   set.seed(5678)
+   rf_nodesize <- train(nb_acte~.,
+                        data = db.freq,
+                        method = "rf",
+                        metric = "RMSE",
+                        tuneGrid = tuneGrid,
+                        importance = TRUE,
+                        nodesize = nodesize,
+                        maxnodes = 25,
+                        ntree = 20)
+   current_iteration <- toString(nodesize)
+   store_nodesize[[current_iteration]] <- rf_nodesize
+ }
> results_mtry <- resamples(store_nodesize)
> summary(results_mtry)

Call:
summary.resamples(object = results_mtry)

Models: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26,
27, 28, 29, 30
Number of resamples: 25

RMSE
       Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
1  50.57259 57.55095 60.23035 59.88392 62.21812 68.43366    0
2  50.33529 57.79659 60.37648 59.91796 62.25141 68.55740    0
3  50.05922 57.86474 60.35289 59.89313 62.15600 68.45158    0
4  50.28881 57.78650 60.34724 59.88701 62.30562 68.03685    0
5  49.96785 57.72948 59.85814 59.85355 62.33471 68.14604    0
6  50.16823 57.82737 60.23780 59.89157 62.23612 68.07271    0
7  49.84192 57.68167 60.19187 59.86582 62.27448 68.25523    0
8  49.88966 57.75033 60.27905 59.84413 62.34555 68.37503    0
9  49.91257 57.83226 60.12824 59.84747 62.07884 68.49921    0
10 50.07154 57.84581 60.39501 59.85988 62.21585 68.53342    0
11 50.63115 57.85174 60.19132 59.86258 62.29656 68.39426    0
12 50.16977 57.76250 60.15625 59.84863 62.34627 68.20985    0
13 49.86315 57.80858 60.27781 59.87696 62.13319 68.52467    0
14 49.98256 57.93947 60.25503 59.89687 62.16717 68.43600    0
15 50.06969 57.45441 60.24388 59.83249 62.16128 68.32291    0
16 50.14824 57.83517 60.12759 59.88048 62.20721 68.53478    0
17 49.89557 57.90645 60.34396 59.86688 62.23144 68.48675    0
18 50.02490 57.67589 60.42387 59.87747 62.17555 67.96258    0
19 50.04375 57.61217 60.15347 59.86481 62.28266 68.23375    0
20 49.96198 57.89901 60.34131 59.86204 62.34859 68.15228    0
21 49.79061 57.52416 60.35317 59.84846 62.36389 68.15228    0
22 49.93280 57.82988 60.31447 59.86204 62.23595 68.05986    0
23 49.83620 57.80730 60.49083 59.86516 62.21166 68.36625    0
24 50.03700 57.58176 60.30964 59.84573 62.27879 68.10030    0
25 50.00128 57.87385 60.25781 59.84908 62.27391 68.50540    0
26 50.19909 57.86897 60.23370 59.88646 62.23162 68.43366    0
27 49.98900 57.85434 60.21532 59.87198 62.41561 68.43366    0
28 49.95406 57.81659 60.21532 59.81369 62.28356 68.43366    0
29 49.96462 57.60049 60.16012 59.84626 62.00158 68.66475    0
30 50.31255 57.61454 60.34771 59.86600 62.12039 68.07381    0
```

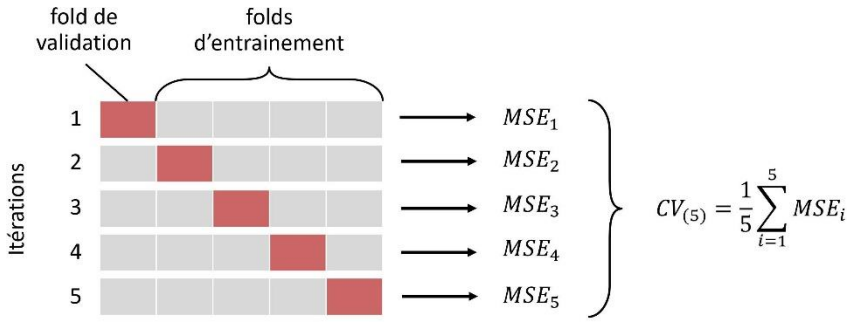
Annexe 4 : Xgboost : Modèle Fréquence

```
> #Xgboost Modèle Fréquence
> Bst_ini = xgboost(data=xgbMatrix, method = "xgbTree", eta = 0.35,  gamma=0, max_depth = 2,
+ nrounds=100, verbose = 0, eval_metric = "error")


> #Optimisation avec validation croisée V5-fold

> xgb_trcontrol = trainControl(method = "cv", number = 5, allowParallel = TRUE,
+ verboseIter = FALSE, returnData = FALSE)


> xgbGrid <- expand.grid(nrounds = c(5,10,50,100,150,200,300,400,500),
+                        max_depth = c(1:10),
+                        colsample_bytree = seq(0.5, 0.9, length.out = 6),
+                        eta = c(0.1,0.2,0.3,0.4,0.5),
+                        gamma=c(0:10),
+                        min_child_weight = 1,
+                        subsample = 1
+                       )

> set.seed(0)
> xgb_model = train(xgbMatrix, complet_Freq_trainXG$nb_acte, trControl = xgb_trcontrol, tuneGrid =
+ xgbGrid, method = "xgbTree", drop = FALSE)

> xgb_model$bestTune
nrounds       max_depth      eta         gamma      colsample_bytree       min_child_weight       subsample
  100            10          0.1          0             0.74                    1                    1

> bst = xgboost(data=xgbMatrix,method = "xgbTree",nrounds=100, verbose = 0,colsample_bytree=0.74,
+ eval_metric = "error")

> predict_Boosting_ini <- predict(model, newdata = dataXGB_test, type="response")
> RMSE.boost.freq = rmse(complet_Freq_test$nb_acte, predict_Boosting)

> predict_Boosting <- predict(bst, newdata = dataXGB_test, type="response")
> RMSE.boost.freq = rmse(complet_Freq_test$nb_acte, predict_Boosting)
```

Annexe 5 : Validation croisée, l'approche *V-fold*

Cette méthode consiste à séparer aléatoirement la base d'étude en k groupes de tailles égales. Le premier fold est traité en tant que base de validation et les k-1 folds restants sont utilisés en tant que base d'entraînement. Il est donc possible de calculer les indicateurs avec la base de validation (le fold restant).

Voici un exemple avec l'indicateur *MSE* :
On peut schématiser le procédé de la manière suivante si l'on considère une validation croisée V5-fold :



La procédure est répétée k fois en choisissant à chaque fois un fold différent en tant que base de validation. Cette opération résulte au calcul de k différents indicateurs, MSE_1, …, MSE_k. On peut ainsi obtenir l'estimation du *MSE* en moyennant les k *MSE*.