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

le 15 Décembre 2020

Par : Djole Henoc AKAFFOU
Titre : Méthode alternative de tarification santé: GLM/XGBoost.

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

Membres présents du jury de l'Institut des Actuaires :
Anaëlle LE BERRE
Signature :

Alain MOEGLIN
Signature :

Entreprise :
PricewaterhouseCoopers
Signature :

Membre présent du jury de l'EURIA :
Franck VERMET
Signature :

Directeurs de mémoire en entreprise :
Kevin KOUO

Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
F +33 (0)2 98 01 66 57
euria@univ-brest.fr

Résumé

En France, le marché de l'assurance santé est un marché très concurrentiel avec des réformes qui protègent de plus en plus les assurés. Cela contraint les assureurs à optimiser leurs modèles de tarification afin d'être plus compétitifs.

En règle générale, les tarifs d'assurance non vie sont modélisés par des méthodes de régression paramétrique comme le GLM. Ces méthodes ont l'avantage d'être opérationnellement faciles à mettre en œuvre et directement interprétables. Cependant, ces modèles restent tout de même très contraignants car ils supposent, entre autres, l'existence d'une relation linéaire entre les variables explicatives et une transformée (généralement logarithmique) de l'espérance de la variable d'intérêt.

D'autre part, les assureurs disposent des méthodes de Machine Learning qui sont non-paramétriques. Cela permet à ces modèles d'être plus flexibles et d'avoir un pouvoir de prédiction généralement plus important que celui des GLM. Cependant, ces modèles souffrent de leur complexité parce qu'ils sont difficiles à mettre en œuvre et difficilement interprétables.

L'objectif de ce mémoire est de combiner ces différentes méthodes afin de proposer une alternative pour construire un modèle avec un meilleur pouvoir de segmentation et de prédiction tout en y apportant des éléments d'interprétation et qui soient opérationnellement peu coûteux.

Pour ce faire, la démarche utilisée consiste dans un premier temps à incorporer aux algorithmes de Machine Learning en particulier le Boosting, des hypothèses de loi pour en déduire des fonctions de perte similaires aux fonctions de déviance des modèles GLM. Le but est d'adapter ces méthodes à des problématiques actuarielles et ainsi capter des interactions complexes qui ne peuvent l'être directement dans la modélisation GLM classique. Les modèles ainsi construits seront par la suite évalués à l'aide d'une base test et d'un processus d'évaluation afin d'évaluer la qualité de segmentation et de prédiction de ceux-ci. Pour finir, l'objectif sera d'utiliser différents outils d'interprétabilité (PDP, ICE, Feature Interaction, Feature Importance) pour rendre ces modèles transparents afin d'évaluer l'apport de cette approche.

Mots clefs: GLM / Machine Learning / Gradient Boosting / Tweedie / Tarification santé / Analyse prédictive / GBM / XGBoost

Abstract

In France, the health insurance market is a highly competitive market with reforms that increasingly protect policyholders. Therefore insurers have to optimize their pricing models in order to be more competitive.

Generally, non-life insurance rates are modeled by parametric regression methods such as GLM. These methods have the advantage of being operationally easy to implement and directly interpretable. However, these models are still very constraining because they assume, among other things, the existence of a linear relationship between the explanatory variables and a transform (generally logarithmic) of the expectation of the variable of interest.

On the other hand, insurers have Machine Learning methods which are non-parametric. This allows these models to be more flexible and to have a predictive power generally greater than that of GLM. However, these models suffer from their complexity because they are difficult to implement and difficult to interpret.

The objective of this thesis is to combine these different methods in order to propose an alternative to build a model with a better segmentation and prediction power, while providing elements of interpretation and which is operationally inexpensive.

To do this, the approach used initially consists of incorporating into Machine Learning algorithms, in particular Boosting, law hypotheses to deduce loss functions similar to the deviance functions of the GLM models. The aim is to adapt these methods to actuarial problems and thus capture complex interactions that cannot be directly captured in conventional GLM modeling. The models thus constructed will then be evaluated using a test base and an evaluation process to assess the quality of segmentation and prediction of these models. Finally, the objective will be to use different interpretability tools (PDP, ICE, LIME, Feature Interaction, Feature Importance) to make these models transparent in order to evaluate the contribution of this approach.

Keywords: GLM / Machine Learning / Gradient Boosting / Tweedie / Health Insurance/ Health pricing/ Predictive analysis / GBM / XGBoost

Synthèse

Introduction
En France, le marché de l'assurance est un marché très concurrentiel. En particulier, sur le marché de l'assurance santé, on observe une homogénéisation des garanties de plus en plus forte essentiellement due à des reformes réglementaires régulières (ANI, contrats responsables, reforme 100 % santé, etc.).

Dans ce contexte, les organismes d'assurance complémentaire se voient obligés de s'adapter constamment en proposant des produits de plus en plus attractifs afin de rester compétitifs. Pour cela, il leur faut construire des modèles de tarification capables de proposer des tarifs en adéquation avec les risques individuels des assurés. En d'autres termes, ces compagnies ont besoin d'un modèle avec une bonne capacité de segmentation des risques, un bon pouvoir de prédiction tout en étant interprétable.

L'approche proposée dans le cadre de ce mémoire consiste dans un premier temps à incorporer aux algorithmes de Machine Learning en particulier le Boosting, des hypothèses de loi pour en déduire des fonctions de perte similaires à celles des modèles GLM. Le but est d'adapter ces méthodes à des problématiques actuarielles et ainsi capter des interactions complexes qui ne peuvent l'être directement dans la modélisation GLM classique. Les modèles ainsi construits seront par la suite évalués à l'aide d'une base test et d'un processus d'évaluation afin d'évaluer la qualité de segmentation et de prédiction de ceux-ci. Pour finir, l'objectif sera d'utiliser différents outils d'interprétabilité afin de rendre ces modèles interprétables.

Le choix de la modélisation

Dans le cadre de ce mémoire, le choix de la modélisation s'est porté sur la modélisation de la dépense réelle plutôt que sur celle de la part complémentaire. Ce choix se motive par la flexibilité que donne la modélisation de la dépense réelle. En effet, modéliser la dépense réelle permet par la suite d'appliquer toute sorte de garantie ce qui n'est pas le cas de la modélisation de la part complémentaire qui comprend déjà une garantie spécifique.

On montre entre autres dans ce mémoire qu'il existe une équivalence entre la moyenne des remboursements complémentaires et la modélisation des dépenses réelles qui prend en compte la fréquence des sinistres pour un ensemble d'assurés avec des garanties homogènes.

Les modèles utilisés
En notant Y la variable à expliquer (la dépense réelle dans notre cas) et X la matrice des variables explicatives (les caractéristiques des assurés dans notre cas) les modèles utilisés dans ce mémoire sont de deux types et sont caractérisés de la manière suivante :

- **Le modèle linéaire généralisé** : ce modèle est une généralisation du modèle linéaire qui

fait l'hypothèse de loi, selon laquelle Y suit une loi de la famille exponentielle (Gaussienne, binomiale, Poisson, Gamma...). En outre ce modèle fait l'hypothèse de loi selon laquelle, Il existe une fonction de lien g telle que : $g(E[Y|X]) = F(X) = X^T\beta$, avec β à estimer.

- **Le modèle du Gradient Boosting** : ce modèle fait l'hypothèse selon laquelle il existe une fonction F telle que $Y = F(X)$. Cette fonction inconnue F a pour estimateur $\hat{F} = \underbrace{arg\,min}_{F(\cdot)\in\mathcal{F}} E_{y,x}[\mathscr{L}(y, F(x))]$, où L est appelée fonction de perte. Cette fonction de perte peut être définie de sorte à obtenir l'hypothèse de loi équivalente sous le modèle GLM.

L'évaluation des modèles

Pour évaluer les modèles de Machine Learning, on dispose de différentes métriques d'évaluation. Les plus connues pour les problèmes de régression sont le RMSE, le MSE, le MAE, l'indice de Gini etc. Cependant, ces métriques donnent une vision globale de la qualité de la prédiction du modèle sur l'ensemble des données. De plus, certaines métriques sont moins adaptées pour des données assurantielles.

Dans le cadre de ce mémoire, une approche différente est proposée afin d'évaluer efficacement la qualité du modèle. Celle-ci est composée de trois niveaux d'évaluation qui sont :

- **Premier niveau :** ce niveau permet d'évaluer l'éventuel biais qui pourrait exister dans l'apprentissage des données en évaluant sur les données d'apprentissage, le rapport $\frac{\sum observations}{\sum predictions}$ qu'on défini comme le ration O/P (Observations sur prédictions). Cela permet de s'assurer que la mutualisation est bien effective sur les données d'apprentissage, auquel cas, le rapport doit respecter la condition suivante : $|O/P - 1| < seuil$. En cas de non-adéquation aux données, le modèle est considéré comme biaisé et sa prédiction n'est pas acceptée. Ce seuil de tolérance choisi dans notre cas est de 1 point par rapport à 100%.
- **Deuxième niveau** : ce niveau permet d'évaluer la qualité de la prédiction sur la base de test, segment par segment (ensemble des individus ayant les mêmes caractéristiques) en mesurant la valeur $|O/P - 1|$ sur chaque segment. Cette valeur permet de mesurer l'écart entre la prédiction et l'observation sur chaque segment de la base de test.

  On note que pour des observations nulles, la valeur de $|O/P - 1|$ est 1. Ce cas de figure n'est pas avantageux pour un modèle au détriment des autres donc n'entrave pas la comparaison entre les modèles.
- **Troisième niveau :** le troisième niveau est un agrégat du niveau précédent qui permet d'évaluer la qualité globale du modèle sur les données de test. Cette évaluation globale est obtenue par une moyenne sur l'ensemble des segments analysés sur le deuxième niveau.

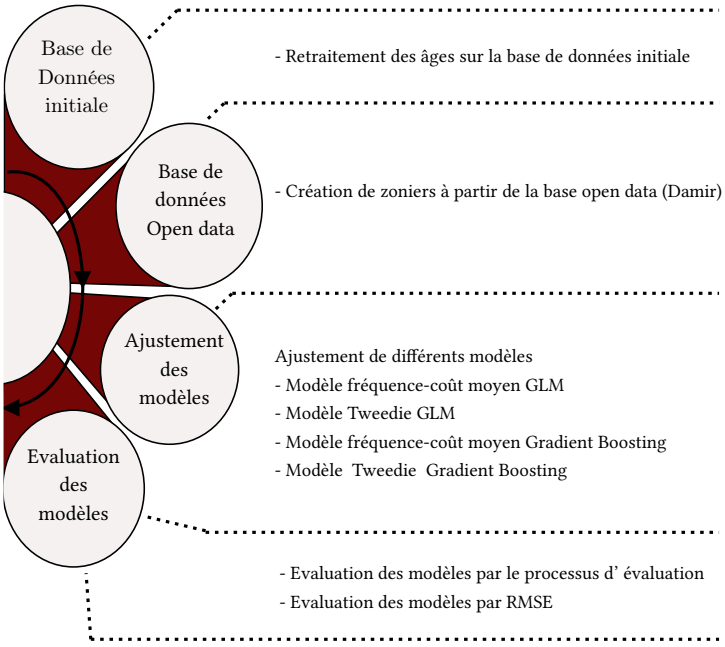
La base de données

La base de données utilisée dans le cadre de ce mémoire est une base de données issue de la fusion de deux bases de données anonymisées d'un portefeuille d'adhésions individuelles provenant d'un organisme complémentaire d'assurance maladie.

Dans le cadre de ce mémoire, les postes de risque modélisés sont :

- **Généraliste** : Ce poste correspond aux dépenses engendrées pour la visite chez un médecin généraliste. Sa composition est très structurée parce que les dépenses sur ce poste sont pour la plupart, conventionnées.
- **Optique** : Ce poste correspond aux dépenses engendrées pour l'Optique (verre Optique, monture, lentille, etc. Il fait partie des postes de dépenses concernés par la nouvelle reforme du 100% santé.)

Après l'analyse préliminaire de la base de données effectuée dans ce mémoire, la modélisation retenue se résume à travers les différentes étapes suivantes :



Les résultats
Cette partie présente les résultats obtenus après la construction de plusieurs modèles pour chacun des différents postes analysés. La démarche utilisée vise à comparer plusieurs modèles entre eux afin d'en tirer le maximum d'informations permettant une segmentation optimale. Pour cela, cinq modèles sont construits sur chaque poste : le modèle GLM Tweedie, le modèle GLM Fréquence Coût-Moyen, le modèle XGBoost (Gradient Boosting) Fréquence Coût-Moyen, le modèle XGBoost Tweedie, et le modèle XGBoost MSE (estimation du coût total avec la fonction de perte « square error »).

Pour le poste **Généraliste**, après l'application du processus d'évaluation, on obtient les résultats suivants :

- Le premier niveau d'évaluation a permis de retenir quatre modèles qui sont : le modèle GLM Tweedie, le modèle XGBoost Fréquence Coût-Moyen et le modèle XGBoost MSE.

- Le deuxième niveau d'évaluation a permis d'observer qu'aucun modèle n'a la supériorité absolue en termes de performance sur l'ensemble de la base de test. Cependant, chaque modèle est meilleur sur une partie des profils de la base de données avec la répartition suivante : le XGBoost MSE (39,4%), le XGBoost Fréquence Coût-Moyen (25,3%), le modèle GLM Fréquence Coût-Moyen (18,56%), le modèle GLM Tweedie (16,53%).

  Ces résultats permettent de construire le modèle optimal à partir des quatre modèles qui représente l'indicatrice par segment préférentiel des prédictions de chacun des quatre modèles.

- Le troisième niveau d'évaluation a permis d'observer que le modèle GLM fréquence Coût-Moyen a une meilleure performance moyenne en termes de ratio O/P sur l'ensemble de la base de test (56,7%). Ce modèle a une performance moyenne meilleure que celle de son équivalent XGBoost (56,9%) qui est suivi du modèle XGBoost MSE (57,51%). En dernière position, se trouve le modèle GLM Tweedie (59,5%).

Pour évaluer la stabilité des résultats obtenus sur le troisième niveau du processus d'évaluation, la méthode du Bootstrap est appliquée afin d'obtenir des sous-échantillons aléatoires de notre base de test. Par la suite, les différents modèles sont évalués sur ces sous-échantillons afin d'obtenir de nouvelles évaluations en termes de moyenne d'écarts moyens |O/P-1| et de RMSE.

Les résultats obtenus mettent en évidence la quasi-stabilité des rangs entre les différents modèles ce qui renforce les résultats moyens obtenus précédemment.

Par ailleurs, les résultats obtenus permettent d'observer que pour toutes les itérations, le modèle optimal construit reste le meilleur en termes de moyenne d'écarts moyens |O/P-1| et de RMSE.

Pour le poste **Optique**, après l'application du processus d'évaluation, on obtient les résultats suivants :

- Le premier niveau d'évaluation a permis de retenir tous les cinq modèles ajustés.

- Le deuxième niveau d'évaluation a permis d'observer que le modèle GLM Fréquence Coût-Moyen, est de loin, le modèle avec la meilleure performance sur la plus grande part de sinistres observés (41,54%). A ce niveau, ce modèle semble être le plus performant. En deuxième position, c'est le modèle XGBoost Tweedie avec 19,86% des sinistres observés, suivi du modèle XGBoost Fréquence Coût-Moyen avec 16,01% des sinistres observés. En quatrième position, on observe le modèle GLM Tweedie (14,5%) suivi en dernière position du modèle XGBoost MSE.

  Ces observations permettent de construire un modèle plus optimal qui correspond au meilleur des cinq modèles sur chaque segment afin d'aboutir à une meilleure qualité de prédiction.

- Le troisième niveau d'évaluation a permis d'observer que le modèle XGBoost Fréquence Coût-Moyen a une meilleure performance moyenne sur l'ensemble de la base de test (84,31%). Cette performance est très proche de celle du modèle GLM Fréquence Coût-

Moyen (84,73%) qui vient en deuxième position. En troisième position, nous avons le modèle GLM Tweedie (86,32%) suivi du modèle XGBoost Tweedie (88,19%). En dernière position, vient le modèle XGBoost MSE.

On observe par ailleurs un gain de performance significatif (- 10% d'erreur) avec le modèle optimal par rapport à la meilleure performance des cinq modèles ajustés.

Par la méthode Bootstrap, les différents modèles sont évalués sur des sous-échantillons aléatoires afin d'obtenir de nouvelles évaluations en termes de moyenne des écarts moyens |O/P-1| et de RMSE.

Les résultats obtenus mettent en évidence la quasi-stabilité des rangs entre les différents modèles ce qui renforce les résultats moyens obtenus précédemment.

Par ailleurs, les résultats obtenus à l'aide de la méthode Bootstrap permettent d'observer que pour toutes les itérations, le modèle optimal construit reste le meilleur en termes de moyenne d'écarts moyens |O/P-1| et de RMSE.

L'interpretabilité des modèles

Le régime RGPD de "responsabilité algorithmique" et le "droit à l'explication" qui en résulte soulignent l'importance d'avoir des modèles de tarification transparents. Cependant, les techniques de Machine Learning sont souvent considérées comme des boîtes noires (black box) contrairement aux modèles statistiques tels que les GLMs.

Plusieurs outils de transparence de modèles existent. Cependant, dans le cadre de ce mémoire, les méthodes utilisées sont les suivantes :

- le graphique de dépendance partielle (Partial Dépendance Plots - PDP),
- le graphique de l'espérance conditionnelle individuelle (Individual Conditional Expectation (ICE) plots),
- la mesure de la force de l'interaction des variables (Variable interaction strength (IS)),
- l'importance des variables par permutation (Permutation feature importance - VIP).

Pour les deux postes analysés, les résultats obtenus lors de la mise en œuvre de ces différents outils d'interpretabilité sont très cohérents avec l'analyse préliminaire des données. En effet, on observe une structure de la dépense moyenne très proche de ce que l'on a pu observer empiriquement pour toutes les variables explicatives en fonction de leurs différences modalités.

De plus, on observe que tous les modèles ont des segmentations très similaires, avec des intensités de différenciation des prédictions qui différent en fonction des modalités des différentes variables.

Conclusion

Au vu des résultats obtenus, l'approche alternative proposée pour la construction d'un modèle optimal permettrait d'améliorer la qualité de prédiction d'un outil de tarification santé tout en permettant son interprétablité.

Cependant, il est bien de noter que la base de données utilisée dans le cadre de ce mémoire reste assez restreinte. Par conséquent, pour mesurer efficacement la stabilité des résultats obtenus ainsi que l'apport marginal des méthodes de Gradient Boosting, l'approche utilisée dans ce mémoire doit être appliquée à une base de données plus grande, avec plus de variables explicatives. Ce qui sera en tout état de cause de plus en plus le cas au vu des possibilités qu'offre le Big Data quant à l'utilisation de sources de données externes.

Executive summary

Introduction

In France, the insurance market is very competitive. In particular, in the health insurance market, there is an increasingly strong homogenisation of guarantees mainly due to regular regulatory reforms (ANI, responsible contracts, 100% health reform, etc.).

In this context, insurance organisations are being forced to constantly adapt by offering attractive products in order to remain competitive. To do this, they must build pricing models capable of offering rates that are in line with the individual risks of the insured. In other words, these companies need a model with a good risk segmentation capacity, a good predictive power, while being interpretable.

The approach proposed in this paper consists first of all in incorporating into Machine Learning algorithms, in particular Boosting, hypotheses of law to deduce loss functions similar to those of GLM models. The aim is to adapt these methods to actuarial issues and thus capture complex interactions that cannot be directly captured in conventional GLM modelling. The models thus constructed will then be evaluated using a test base and an evaluation process to assess the quality of their segmentation and prediction. Finally, the objective will be to use different interpretability tools in order to make these models interpretable.

The choice of modelling

In the context of this paper, the choice of modelling was made to model the real expenditure rather than the complementary part. This choice is motivated by the flexibility provided by the modelling of real expenditure. In fact, modelling the real expenditure makes it possible to subsequently apply any kind of guarantee, which is not the case with the modelling of the complementary part, which already includes a specific guarantee.

Among other things, this paper shows that there is an equivalence between the average of complementary reimbursements and the modelling of real expenditure, which takes into account the frequency of claims for a group of insured persons with homogeneous guarantees.

The models used

By denoting Y the variable to be explained (the real expenditure in our case) and X the matrix of explanatory variables (the characteristics of the insured in our case), the models used in this paper are of two types and are characterized as follows :

- **The generalised linear model** : this model is a generalisation of the linear model which makes the law hypothesis, which states that Y follows an exponential family law

(Gaussian, binomial, Poisson, Gamma...). Furthermore, this model assumes that there is a link function g such that : $g(E[Y|X]) = F(X) = X^T\beta$, with β to be estimated.

- **The Gradient Boosting Model** : this model assumes that there is a F function such that $Y = F(X)$. This unknown function F has the estimator $\hat{F} = \underbrace{arg\,min}_{F(\cdot)\in\mathcal{F}} E_{y,x}[\mathscr{L}(y, F(x))]$, where L is called the loss function. This loss function can be defined so as to obtain the equivalent law assumption under the GLM model.

Model evaluation

Different assessment metrics are available for evaluating Machine Learning models. The best known for regression problems are the RMSE, MSE, MAE, Gini index etc. However, these metrics give an overall view of the quality of the model prediction on all the data. In addition, some metrics are less suitable for insurance data.

In this paper, a different approach is proposed in order to efficiently assess the quality of the model. This approach is composed of three levels of evaluation which are :

- **First level :** This level allows to evaluate the possible bias that could exist in the learning data by evaluating on the learning data, the $\frac{\sum observations}{\sum predictions}$ ratio that we define as the O/P ration (Observations over predictions). This makes it possible to ensure that the mutualisation is indeed effective on the learning data, in which case, the ratio must respect the following condition : $|O/P - 1| < threshold$. In case of non-adaptation to the data, the model is considered biased and its prediction is not accepted. This tolerance threshold chosen in our case is 1 point compared to 100%.
- **Second level** : this level allows the quality of prediction to be assessed on a test basis, segment by segment (all individuals with the same characteristics) by measuring the value $|O/P - 1|$ on each segment. This value allows to measure the gap between prediction and observation on each segment of the test base.

  Note that for null observations, the value of $|O/P - 1|$ is 1. This case is not advantageous for one model to the detriment of the others and therefore does not hinder the comparison between models.
- **Third level :** the third level is an aggregate of the previous level which allows to evaluate the overall quality of the model on the test data. This global evaluation is obtained by averaging all the segments analysed on the second level.

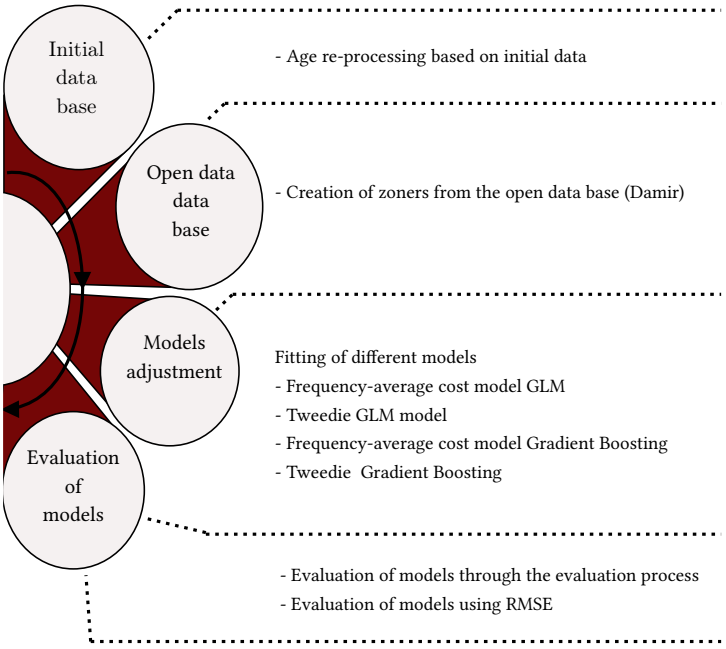
The database

The database used in this brief is a database resulting from the merger of two anonymised databases of a portfolio of individual memberships from a supplementary health insurance organisation.

The risk items modelled in this paper are :

- **General practitioner** : This item corresponds to the expenses generated for the visit to a general practitioner. Its composition is very structured because most of the expenses on this item are conventional.
- **Optics** : This item corresponds to expenditure generated for optics (optical lenses, frames, lenses, etc.). It is one of the expenditure items concerned by the new 100

After the preliminary analysis of the database carried out in this dissertation, the chosen modeling is summarized through the following different steps :



Results

This section presents the results obtained after the construction of several models for each of the different positions analyzed. The approach used aims to compare several models with each other in order to extract the maximum amount of information for optimal segmentation. For this purpose, five models are built for each substation : the GLM Tweedie model, the GLM Frequency-Cost-Medium Frequency model, the XGBoost (Gradient Boosting) Frequency-Cost-Medium model, the XGBoost Tweedie model, and the XGBoost MSE model (estimation of the total cost with the "square error" loss function).

For the **General practitioner** position, after applying the evaluation process, the following results are obtained :

- The first level of evaluation allowed the selection of four models which are : the GLM Tweedie model, the XGBoost Frequency-Cost-Average model and the XGBoost MSE model.

- The second level of evaluation allowed us to observe that no model has absolute superiority in terms of performance over the entire test base. However, each model is better on a part of the profiles of the database with the following distribution : XGBoost MSE (39.4%), Gradient Boosting Frequency Cost-Mean (25.3%), GLM Frequency Cost-Mean (18.56%), GLM Tweedie (16.53%).

  These results make it possible to build the optimal model from the four models, which represents the indicator function per preferential segment of the predictions of each of the four models.

- The third level of evaluation allowed us to observe that the GLM frequency-average cost model has a better average performance over the whole test base (56.7%). This model has a better average performance than its equivalent XGBoost model (56.9%), followed by the XGBoost MSE model (57.51%). In last position is the GLM Tweedie model (59,5%).

To assess the stability of the results obtained on the third level of the evaluation process, the Bootstrap method is applied in order to obtain random sub-samples of our test base. Subsequently, the different models are evaluated on these sub-samples in order to obtain new evaluations in terms of mean deviation |O/P-1| and RMSE.

The results obtained show that the ranks between the different models are almost stable. models, which reinforces the average results obtained previously.

Moreover, the results obtained allow us to observe that for all iterations, the Optimal constructed model remains the best in terms of mean deviation |O/P-1| and RMSE.

For the **Optical** position, after the application of the evaluation process, the following results are obtained :

- The first level of evaluation allowed to retain all five adjusted models.

- The second level of evaluation showed that the GLM Frequency-Cost-Average model is by far the model with the best performance on the largest share of observed claims (41.54%). At this level, this model seems to be the most efficient. In second position, it is the XGBoost Tweedie model with 19.86% of the observed claims, followed by the XGBoost Cost Average Frequency model with 16.01% of the observed claims. In fourth position is the GLM Tweedie model (14.5%) followed in last position by the XGBoost MSE model.

  These observations make it possible to build a more optimal model which corresponds to the best of five models in each segment to achieve better prediction quality.

- The third level of evaluation showed that the XGBoost Frequency-Cost-Average model has a better average performance over the entire test base (84.31%). This performance is very close to that of the GLM-Mean Cost Frequency model (84.73%) which comes in second place. In third position comes the GLM Tweedie model (86.32%) followed by the XGBoost Tweedie model (88.19%). In last position comes the XGBoost MSE model.

  A significant gain in performance (-10% error rate) is also observed with the optimal model in relation to the best performance.

Using the Bootstrap method, the different models are evaluated on random sub-samples in order to obtain new evaluations in terms of mean deviation |O/P-1| and RMSE.

The results obtained show that the ranks between the different models are almost stable. models, which reinforces the average results obtained previously.

Moreover, the results obtained allow us to observe that for all iterations, the Optimal constructed model remains the best in terms of mean deviation |O/P-1| and RMSE.

The interpretability of models

The RGPD regime of "algorithmic liability" and the resulting "right to explanation" underlines the importance of having transparent pricing models. However, Machine Learning techniques are often seen as "black boxes" as opposed to statistical models such as GLM.

Several tools for transparent models exist. However, for the purposes of this paper, the methods used are as follows :

- the Partial Dependency Plots (PDP) graph,
- Individual Conditional
- Expectation (ICE) plots,
- the measure of the strength of the interaction of variables (Variable interaction strength (IS)),
- the importance of variables by permutation (Permutation feature importance - VIP).

For the two positions analysed, the results obtained during the implementation of these different interpretability tools are very consistent with the data analysis. In fact, the structure of the average expenditure is very close to what has been empirically observed for all the explanatory variables according to their different modalities.

Moreover, we observe that all the models have very similar segmentations, with different intensities of differentiation of the predictions with respect to the modalities of the different variables.

Conclusion

In view of the results obtained, the alternative approach proposed for the construction of an optimal model would make it possible to improve the quality of prediction of a pricing tool while allowing its interpretation.

However, it should be noted that the database used for this paper is still rather limited. Therefore, in order to effectively measure the stability of the results obtained as well as the marginal contribution of Gradient Boosting methods, the approach used in this paper must be applied to a larger database with more explanatory variables. This will in any case be increasingly the case in view of the possibilities offered by Big Data for the use of external data sources.

Remerciements

Je tiens à remercier tout d'abord, ceux qui m'ont transmis leur connaissance au cours de ce stage, et même à ceux qui ont eu la gentillesse de faire de cette expérience un moment très profitable.

Mes remerciements s'adressent tout particulièrement à mon tuteur Kevin KOUO, Actuaire, consultant senior de l'équipe RVMS de PwC France, pour m'avoir accordé sa confiance, pour toute l'aide qu'il m'a apportée et pour la patience dont il a fait preuve afin de me permettre de mieux cerner le travail à réaliser tout au long de ce stage.

J'adresse tout autant mes plus vives gratitudes à David Cadoux Partner ainsi qu'à Guillaume Beneteau, associés RVMS, pour leur accueil chaleureux et leur bienveillance, qui ont été indispensables à mon intégration.

De même, j'adresse mes remerciements à l'ensemble des membres de l'equipe RVMS pour leur sympathie, leur aide spontanée, ce qui m'a permis de passer une belle expérience.

Je désire aussi adresser mes remerciements à ma femme, Mélissa FOURNET, pour son soutien et ses encouragements qui ont été indispensables à la réussite de mon projet professionnel.

Par ailleurs, j'adresse mes remerciement à Eric SMUTEK, Actuaire, Consultant chez FIXAGE, pour ses précieux conseils et ses encouragements qui ont été d'un apport considérable à la construction de mon projet professionnel.

Je souhaite aussi remercier l'ensemble de mes professeurs de l'EURIA qui m'ont apporté une formation indispensable à ma réussite professionnelle.

Enfin, Je remercie toutes les personnes qui ont contribué de près ou de loin à l'enrichissement de ce mémoire.

Table des matières

Introduction

L'une des caractéristiques principales du business de l'assurance est le cycle inversé de production. L'assureur reçoit des cotisations pour des prestations dont les versements sont conditionnés à la réalisation des risques assurés. Les assureurs sont donc contraints de fixer le montant de ces cotisations lors de la création de leurs offres d'assurance et avant même la connaissance des montants réels des prestations futures. C'est l'une des problématiques majeures du business de l'assurance et plus particulièrement de l'assurance santé. Par conséquent, la tarification occupe une place centrale dans le métier de l'assurance en général, et de l'assurance santé en particulier.

Par ailleurs, l'assurance santé en France est d'autant plus spécifique parce qu'elle présente un caractère social. En effet, cela se matérialise au travers de la couverture universelle qui permet à tous les individus d'avoir accès aux services de santé dont ils ont besoin sans que cela n'entraîne pour eux, des difficultés financières. Ce caractère social du système de santé Français s'est très bien illustré en épargnant des difficultés financières aux ménages Français contraints de se faire hospitaliser pendant plusieurs semaines lors de la pandémie mondiale de la Covid19.

De plus, le marché de l'assurance santé en France est un marché marqué par des évolutions réglementaires qui créent de nouvelles opportunités de plus en plus nombreuses pour les assureurs.

Ce contexte économique et social, couplé à une situation de forte concurrence sur le marché de l'assurance santé, met en exergue les contraintes qui s'imposent aux assureurs dans la conception de leurs produits. Cela oblige donc les actuaires, dont le rôle est de concevoir ces produits, à adapter leurs méthodes de tarification pour être plus compétitifs.

En règle générale, la tarification en santé s'appuie sur des méthodes de régression qui tendent à vouloir prédire les tarifs des assurés sur la base d'un lien paramétrique entre l'estimation de ces tarifs et certaines caractéristiques des assurés. Ces méthodes présentent des limites qui peuvent biaiser l'estimation du coût de la couverture d'assurance que cherche à estimer l'assureur.

Avec l'avènement du Big Data et le développement de certains algorithmes de prédiction de plus en plus performants, certains assureurs s'orientent vers des méthodes dites non-paramétriques. Ces méthodes comportent parfois des limites lors de leur mise en œuvre opérationnelle parce qu'elles sont coûteuses en temps d'exécution et sont difficiles à interpréter.

Ce mémoire a pour but de proposer une méthode de tarification alternative avec une application à un portefeuille d'assurance santé. La démarche utilisée dans ce mémoire consiste à combiner des modèles de régression paramétriques et non paramétriques afin de pouvoir construire un modèle hybride permettant d'avoir une maîtrise sur notre objectif d'optimalité.

Pour ce faire, ce mémoire s'articule autour de quatre parties. Dans la première partie, nous

présenterons le contexte général de l'assurance santé en France. Dans la seconde partie, le cadre théorique de la tarification santé sera posé. Dans la troisième partie, il sera question de la mise en application des méthodes utilisées et une étude approfondie des résultats obtenus. Pour finir, dans la quatrième partie, il s'agira d'analyser la transparence du modèle retenu.

1 Quelques caractéristiques de l'assurance maladie en France

"O santé ! Santé ! Bénédiction des riches ! Richesse des pauvres ! Qui peut t'acquérir à un prix trop élevé, puisqu'il n'y a pas de joie dans ce monde sans toi ?"

Ben Jonson, 1572 -1637

En France, la protection sociale est l'ensemble des mécanismes qui, dans un cadre de solidarité nationale, permettent aux individus vivant en France, de se couvrir contre les conséquences financières des événements liés aux risques sociaux (maladie, accident de travail, Maternité, Vieillesse, etc.). Elle est financée d'une part par les cotisations sociales et d'autre part par la collecte d'impôts, de taxes et de contributions publiques.

La protection sociale est organisée selon quatre niveaux qui sont :

- **la Sécurité sociale**, en charge de la couverture de base pour les risques maladie, retraite, famille, accidents du travail, maladies professionnelles et vieillesse.
- **les régimes complémentaire**s, qui peuvent venir compléter les risques couverts par la Sécurité sociale ;
- L'**UNEDIC** (union nationale interprofessionnelle pour l'emploi dans l'industrie et le commerce) pour l'assurance chômage ;
- **Les aides de l'État et des département**s pour les plus démunis.

Dans le cadre du risque maladie, la prise en charge des dépenses de santé se décompose en deux parties :

- Une part assurée par la branche assurance maladie de la Sécurité sociale ;
- Une part complémentaire appelée « ticket modérateur », à la charge de l'assuré et pris en charge partiellement ou totalement, le cas échéant, par une complémentaire santé.

Il faut donc savoir distinguer le régime obligatoire (Sécurité Sociale) de la complémentaire santé.

1.1 Le régime obligatoire de l'assurance santé

Le premier niveau de protection sociale en santé est assuré par la Sécurité sociale. Comme son nom l'indique, le régime obligatoire de santé est imposé par la loi à toutes les personnes vivant et travaillant en France de manière régulière. C'est la couverture minimale légale de santé. Tous les résidents Français y sont affiliés et cotisent selon leurs revenus. En contrepartie, ils bénéficient avec leurs ayant droits d'un remboursement partiel de leurs frais de santé.

Cependant, si tous les résidents Français bénéficient d'une couverture santé de base, celle-ci diffère selon les situations personnelles, et le cas échéant, le secteur d'activité.

1.1.1 Les différents régimes de la Sécurité sociale

Il existe trois types de régime de sécurité sociale classés selon les catégories socioprofessionnelles. On distingue :

- **Le régime général**, géré par la CNAM (Caisse nationale de l'assurance maladie des travailleurs salariés). Il s'adresse aux salariés du secteur privé, aux travailleurs indépendants ainsi qu'à leurs ayants droit. Il couvre huit personnes sur neuf en France ;
- **Le régime agricole**, géré par la caisse centrale de la Mutualité sociale agricole (MSA). Il accompagne les exploitants, les salariés agricoles et les entreprises agricoles et couvre une personne sur vingt en France.
- **Les régimes spéciaux**, qui sont au nombre de 27, et qui regroupent les régimes de la SNCF, des fonctionnaires, des mines, des cultes, des députés, etc.

L'existence jusqu'à ce jour de cette multitude de régimes s'explique par le fait qu'un certain nombre de professions et de corps sociaux, se sont opposés à l'intégration d'une Sécurité sociale unique instaurée en 1945.

1.1.2 Le mécanisme de remboursement des frais de santé

Pour mieux comprendre le mécanisme de remboursement des frais de santé, il est primordial de définir des notions essentielles qui sont :

- **Le tarif de convention (TC), la base de remboursement (BR) et le tarif d'autorité (TA) :**

  Le tarif de convention est le tarif issu des différentes conventions qui existent entre la sécurité sociale, le gouvernement et les principaux syndicats des professions médicales. C'est ce tarif qui régit le système de remboursement utilisé par la Sécurité sociale pour les remboursements des soins et les honoraires des professionnels de soins, soumis aux conventions nationales. On parle de base de remboursement.

  Il existe deux catégories de professionnels de santé conventionnés. On parle de professionnels de santé appartenant au **secteur 1** pour ceux qui appliquent le tarif de convention

et de professionnels de santé appartenant au **secteur 2** pour les professionnels de santé conventionnés à honoraires libres.

Par ailleurs, les professionnels de santé non conventionnés sont dits de **secteur 3**. Ils pratiquent des tarifs totalement libres, avec dépassement d'honoraires. Le remboursement de la Sécurité sociale de leurs consultations est très inférieur à la base de remboursement pour les médecins conventionnés. On parle alors de **tarif d'autorité**.

- **Le ticket modérateur (TM) :** Il correspond à la portion de la base de remboursement non prise en charge par l'Assurance Maladie.
- **Les participations forfaitaires et franchises médicales :**
  - Depuis 2004, pour chaque consultation ou acte réalisé par un professionnel de santé ainsi que sur les examens radiologiques et analyses de biologie médicale, une participation forfaitaire de 1 € est à la charge des assurés de plus de 18 ans. Cette participation est plafonnée à 50 € par an et par personne.
  - Depuis 2008, une somme est déduite des remboursements de la sécurité sociale sur les médicaments, les actes paramédicaux et les transports sanitaires. Ce montant est de 0,5€ pour les médicaments et les actes paramédicaux et de 2€ pour les transports sanitaires.
  - Le forfait hospitalier, crée en 1983, est la somme déduite des remboursements par la Sécurité sociale pour les frais journaliers d'hébergement et d'entretien entraînés par hospitalisation de l'assuré. Il est de 20€ pour les séjours en hôpital et de 15€ pour le service psychiatrique.
    Par ailleurs, on note que la Sécurité sociale prend en charge 80% des frais d'hospitalisation.
- **Le parcours de soins coordonnés :** Crée en 2004, le parcours de soins coordonnés oblige chaque assuré à désigner son médecin traitant sous peine de voir ses remboursements de consultations minorés. Le médecin traitant a pour rôle de centraliser toutes les informations de soins et de mettre à jour le dossier médical de l'assuré. Tout médecin peut remplir ce rôle, à condition d'être un médecin conventionné (secteur 1 ou 2). Lorsque la nécessité se présente, c'est le médecin traitant qui envoie l'assuré chez un spécialiste.

Ci-dessous, un résumé graphique de la prise en charge des frais de santé par la sécurité sociale.

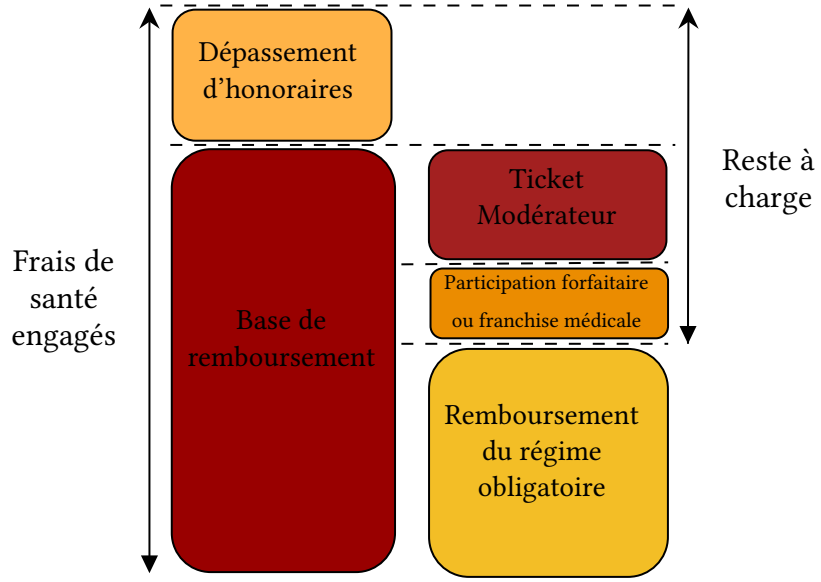


FIGURE 1.1 – Prise en charge des frais de santé par la sécurité sociale

1.2 Le régime complémentaire de l'assurance santé

Comme décrit précédemment, la Sécurité sociale ne prend pas en charge la totalité des dépenses de santé. Il reste donc un reste à la charge de l'assuré qui peut être couvert par l'adhésion à une complémentaire santé. Cette complémentaire santé complétera partiellement ou totalement le remboursement des dépenses de santé en fonction du contrat choisi.

Le tableau ci-dessous montre des exemples de prise en charge par une complémentaire santé, d'une consultation chez un généraliste conventionné Secteur 2. Le tarif de convention est alors de 23 € et la franchise médicale appliquée par la sécurité sociale est de 1 € par consultation.

Taux de remboursement de la sécurité sociale	Remboursement Sécurité sociale après déduction de la franchise		Taux de remboursement de la complémentaire santé	Remboursement de la complémentaire santé après déduction du remboursement de la sécurité sociale avant la franchise		Remboursement total maximum
70%	15,10 €	+	100 %	18.40	=	33.50
70%	15,10 €	+	150 %	29.90	=	45.00
70%	15,10 €	+	200 %	41.40	=	56.50
70%	15,10 €	+	300 %	52.90	=	68.00

TABLE 1.1 – Exemple de prise en charge par une complémentaire santé

Plusieurs organismes d'assurance peuvent permettre la souscription à une complémentaire santé. Il s'agit des organismes suivants :

- **Les mutuelles santé** qui sont des sociétés de personnes à but non lucratif organisant la solidarité entre leurs membres et régies par le code de la mutualité ;
- **Les compagnies d'assurance santé** qui sont des sociétés de droit privé à but lucratif qui fournissent des services d'assurance. Elles sont régies par le code des assurances ;
- **Les institutions de prévoyance** qui sont des sociétés de personnes, qui gèrent des contrats collectifs d'assurance de personnes couvrant les risques de maladie, incapacité de travail, invalidité, dépendance et décès. Elles sont à but lucratif et régies par le code de la sécurité sociale. ;

Ces trois acteurs sont associés à la gestion de l'assurance maladie par le biais de l'Union nationale des organismes d'assurance maladie complémentaire (UNOCAM).

Il existe deux types de contrats de complémentaire santé : les contrats individuels et les contrats collectifs.

1.2.1 Les différents types de contrats de complémentaire santé

1.2.1.1 Les contrats collectifs

Un contrat collectif de complémentaire santé est un contrat souscrit par une entreprise ou une association au profit de ses employés et de leurs ayants droit.

L'Accord National Interprofessionnel de janvier 2013 (ANI), entré en vigueur le 1er janvier 2016, oblige tout employeur du secteur privé (entreprise et association) à proposer à l'ensemble de ses salariés, une couverture complémentaire santé collective. Toutefois, le salarié a le droit de refuser l'adhésion s'il possède déjà une complémentaire santé.

Le contrat collectif de complémentaire santé doit respecter un ensemble de garanties minimum, à savoir :

- Le remboursement à 100% du ticket modérateur.
- Le remboursement à 100% du forfait hospitalier journalier.
- Le remboursement des frais dentaires à hauteur de 125% du tarif conventionnel.
- Le remboursement des frais forfaitaires d'Optiques par période de deux ans pour les adultes et un an pour les enfants. Le minimum de prise en charge est de 100 € pour les verres simples et la monture et de 150 € pour les verres complexes et la monture.

1.2.1.2 Les contrats individuels

Un contrat individuel de complémentaire santé est un contrat grand-public qui peut être souscrit directement par l'assuré à titre individuel ou au profit d'un membre de sa famille. Généralement moins avantageux que les contrats collectifs, ces contrats s'adressent principalement aux personnes n'ayant pas accès aux contrats collectifs (étudiants, indépendants, chômeurs, etc.).

Ils peuvent aussi faire office de surcomplémentaire pour les salariés qui souhaitent compléter leur assurance collective obligatoire.

1.3 Quelques reformes importantes de l'assurance santé en France

1.3.1 Solvabilité 2

Entrée en vigueur le 1er janvier 2016, la réforme réglementaire européenne Solvabilité 2 a pour objectif de renforcer les critères de solvabilité des compagnies d'assurance en matière de volume de liquidité, de surveillance, de gestion des risques, de reporting et de transparence.

Dans le contexte du marché de l'assurance santé particulièrement très concurrentiel avec des niveaux de rentabilité faibles, cette norme prudentielle fait croître le besoin en financement des organismes de complémentaire santé.

1.3.2 L'Accord National Interprofessionnel (ANI)

Entrée en vigueur le 1er janvier 2016, la loi sur l'Accord national interprofessionnel prévoit la généralisation de la couverture complémentaire santé à tous les salariés. De ce fait, toutes les entreprises sont dans l'obligation de proposer à l'ensemble de leurs salariés une complémentaire santé et de financer celle-ci à hauteur de 50 % minimum.

Selon une étude menée par la Fédération Française des Assurances (FFA), cette loi a conduit à un transfert de 4 millions de contrats individuels vers des contrats collectifs. Cette réforme impacte ainsi le marché de l'assurance complémentaire en la rendant plus concurrentielle avec une part de marché des compagnies d'assurance qui passe de 27,8 % à 30 %.

1.3.3 Le contrat responsable

Le contrat responsable est une forme spécifique de contrat de complémentaire santé dont le but est d'inciter les assurés à suivre le parcours de soin conseillé par l'Assurance Maladie.

Un contrat de complémentaire santé est dit responsable lorsqu'il comporte les garanties suivantes :

- la prise en charge totale ou partielle des consultations et prescriptions du médecin traitant afin d'encourager les patients à respecter le parcours de soins coordonnés ;
- L'absence de prise en charge de la participation forfaitaire de 1 € ;
- Pour les patients qui consultent un spécialiste sans passer par leur médecin traitant, l'exclusion totale ou partielle de la prise en charge des dépassements d'honoraires sur le tarif des actes et consultations ;
- La prise en charge totale de certaines prestations liées à la prévention.

Le contrat responsable offre également des avantages fiscaux pour les entreprises et professionnels indépendants.

Ce dispositif, imposant des planchers et des plafonds de remboursement conduit à une standardisation des contrats

1.3.4 La réforme 100 % santé : le reste à charge de zéro (RAC 0)

Entrée en vigueur en Janvier 2019 avec une mise en application progressive jusqu'en 2021, la réforme 100 % santé a pour but de faciliter à tous les Français, l'accès aux soins. Sa mise en œuvre passe par la suppression du reste à charge sur certains équipements qui entrent dans le cadre de paniers définis par les pouvoirs publics notamment l'Optique, l'audio et les prothèses dentaires.

Cette reforme vient renforcer la structure des contrats de complémentaire santé déjà impactée par certaines reformes comme l'ANI et le contrat responsable.

1.4 Les dépenses de santé en France

En France, la dépense courante de santé (**DCS**) est la somme de toutes les dépenses courantes engagées par les financeurs publics et privés pour la fonction santé (la Sécurité sociale et les complémentaires santé). Elle est principalement composée (à environ 75%) de la consommation de soins et biens médicaux (**CSBM**).

La CSBM est structurée en 2018 de manière suivante :

- Les soins hospitaliers en structure publique et privée, hors soins de longue durée pour 46,4 % ;
- Les soins ambulatoires (médecins, dentistes, auxiliaires médicaux, laboratoires d'analyse, thermalisme) pour 27 % ;
- Les transports sanitaires pour 2.5 % ;
- Les médicaments en ambulatoire pour 16.1 % ;
- Les autres biens médicaux en ambulatoire (Optique, prothèses, petit matériel et pansements), pour 7,9 %.

Bien qu'ayant une structure globalement stable, la CSBM s'est progressivement déformée entre 2009 et 2018. La part de médicaments a reculé en moyenne de 0,4 points par an au profit d'une hausse moyenne annuelle des soins de ville de 0,2 points et des autres biens médicaux de 0,1 points.

Les résultats des comptes de la sécurité sociale pour les dépenses de santé en 2018 montrent une baisse de la croissance de la consommation de Soins et de Biens Médicaux (CSBM). Cette consommation est estimée à 203.5 milliards d'euros soit + 1,5 %, contre + 1,7 % en 2017. Elle représente 8,6 % du PIB, soit 3 037 euros par habitant. Ci-dessous, le graphique montrant l'évolution de la CSBM entre 2009 et 2018. [1]

1. *Insee, Tableaux de l'économie française Édition 2020, Dépenses de santé* 2018.

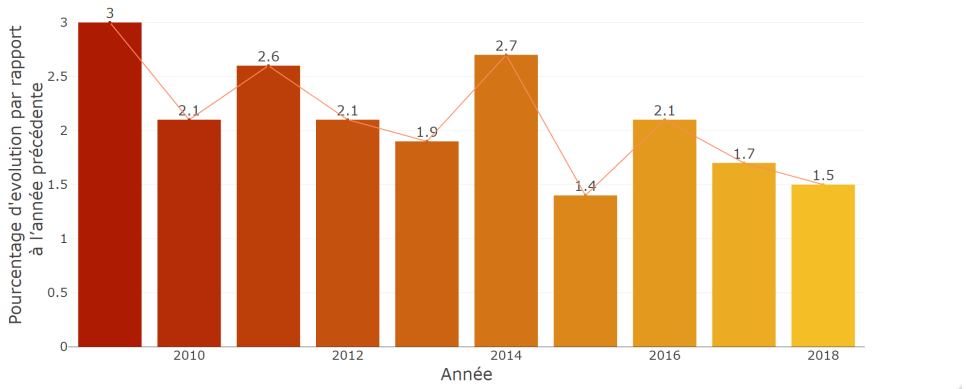


FIGURE 1.2 – Évolution de la dépense en santé en France de 2009 à 2018

Outre la CSBM, la DCS est composée pour environ 25% des éléments suivant :

- Les soins aux personnes âgées dispensés dans les EHPAD (Établissements d'Hébergement pour Personnes Agrées Dépendantes) ainsi que les soins des infirmiers à domicile
- les indemnités journalières d'arrêt de travail (maladie, maternité et accidents du travail) ;
- les coûts de gestion du système ;
- les dépenses de prévention institutionnelle (campagnes de prévention non comptabilisées dans la CSBM).

En 2017, de même que l'Allemagne, la France a consacré plus de 11,3 % de son PIB à la santé selon une étude [2] de l'OCDE et de la Commission européenne publiée en novembre 2019. Cela représente la plus grande part parmi les pays de l'UE qui consacrent en moyenne 9,8% de leur richesse nationale en matière de santé.

En outre, comme le montre le graphique ci-dessous, la France se place en troisième rang des pays de l'OCDE.

2. *State of Health in the EU, France, Profils de santé par pays* 2019.

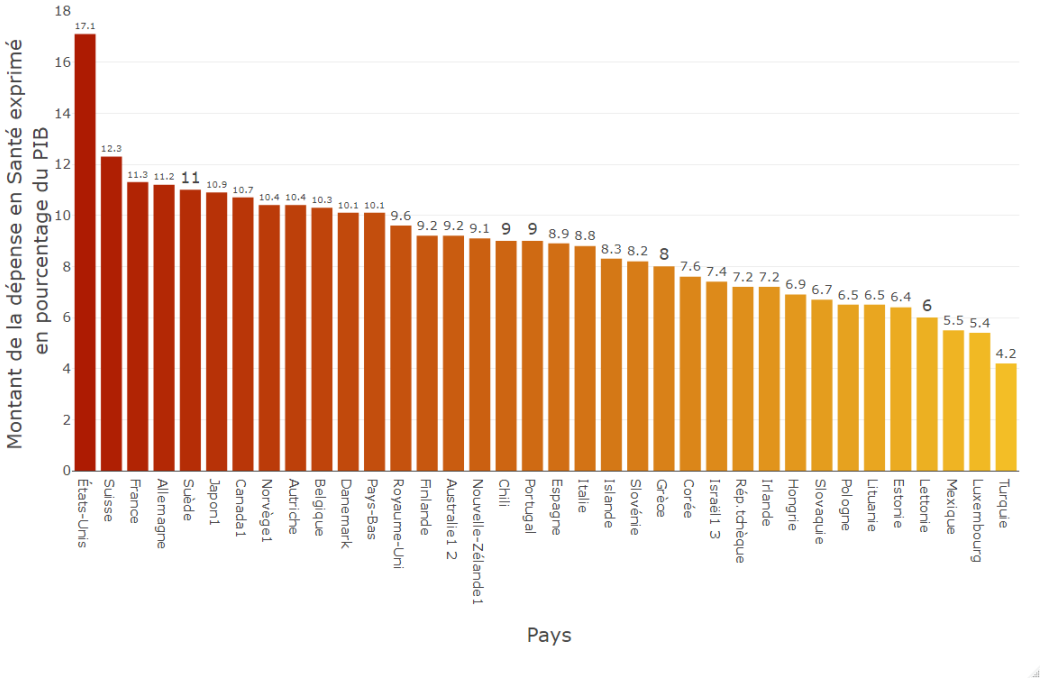


FIGURE 1.3 – Comparaison de la dépense courante en santé dans les pays de l'OCDE en 2017

1.5 Conclusion

La présentation succincte du fonctionnement du système d'assurance santé en France faite dans ce chapitre met en exergue un système d'une rare complexité. Cette complexité est le fait de plusieurs facteurs dont l'existence d'une multitude de régimes gérés par plusieurs organismes. Par ailleurs, le caractère très concurrentiel de ce système, couplé à des réformes très récurrentes et de plus en plus en faveur des assurés, accentuent cette complexité.

Dans ce contexte, les organismes d'assurance complémentaire se voient obligés de s'adapter constamment en proposant des produits de plus en plus attractifs. Pour cela, il leur faut construire des modèles de tarification capables de proposer des tarifs en adéquations avec les risques individuels des assurés.

La suite de ce mémoire est consacrée à la construction d'un tel modèle.

2 Contexte théorique de l'étude

"Une théorie mathématique ne doit être regardée comme parfaite que si elle a été rendue tellement claire qu'on peut la faire comprendre au premier individu rencontré dans la rue."
David Hilbert, 1862-1943

2.1 La tarification santé

Le rôle d'une compagnie d'assurance et plus particulièrement d'une complémentaire santé est d'assurer un certain risque. En contrepartie, elle reçoit des primes payées par ses assurés. Elle est donc emmenée à indemniser, au fil de l'eau, les assurés pour les sinistres qu'ils déclarent. Pour ce faire, les compagnies d'assurances souhaitent estimer le coût total engendré par les sinistres futurs de chaque assuré. Ce chapitre introduit les notations et les modèles mathématiques utilisés dans ce mémoire.

2.1.1 L'antiselection et la nécessité de segmentation

Par définition, l'antisélection ou sélection adverse est une situation dans laquelle une partie possède plus d'informations qu'une autre au moment de la signature d'un contrat (ex-ante). Dans le cadre de l'assurance, les assureurs n'observent que partiellement le niveau de risque des assurés. Ainsi, lorsque l'assuré est à risque élevé, il n'a pas intérêt à déclarer son niveau de risque.

Dans le cas où un assureur fixe une prime de risque moyenne applicable à l'ensemble de ses assurés, cela l'expose à une surestimation de la prime de risque pour les assurés à faible risque et à une sous-estimation de la prime de risque pour les assurés à risque élevé. Cette situation pourrait conduire les assurés qui ont un niveau de risque faible, à renoncer à s'assurer tandis que l'effet inverse pourrait s'observer chez les assurés à risque élevé.

D'après Rothschild et Stiglitz [1], l'assureur peut endiguer l'effet de l'asymétrie d'information en proposant des contrats différenciés vers lesquels les individus vont se diriger en fonction de leurs niveaux de risques. Les individus les moins à risque se dirigeront vers les contrats qui

1. Rothschild et Stiglitz 1976.

proposent des niveaux de couverture bas tandis que les individus les plus à risque se dirigeront vers les contrats qui proposent des niveaux de couvertures élevés. Cette situation pourrait conduire à ce que l'assureur ne récupère que les risques les plus élevés.

Il est donc indispensable pour un assureur de faire une bonne segmentation des risques qu'il assure afin de proposer des contrats adéquats, rester compétitif et d'assurer sa rentabilité.

2.1.2 Modélisation des dépenses de santé

Pour un portefeuille d'assurance, le coût total engendré par l'ensemble des sinistres appelé aussi charge totale de sinistre s'écrit comme la somme de toutes les charges engendrées par l'ensemble des sinistres de chaque contrat pris individuellement. Il existe deux approches pour calculer cette somme : le modèle **collectif** et le modèle **individuel**.

2.1.2.1 Choix de modélisation : la dépense réelle.

Dans le cadre de ce mémoire, le choix de la modélisation s'est porté sur la modélisation de la dépense réelle plutôt que sur celle de la part complémentaire.

Ce choix se motive par la flexibilité que donne la modélisation de la dépense réelle. En effet, modéliser la dépense réelle permet par la suite d'appliquer toute sorte de garantie ce qui n'est pas le cas de la modélisation de la part complémentaire qui comprend déjà une garantie spécifique.

En prenant comme exemple le cas de la garantie « visite chez un généraliste conventionné Secteur 1 » défini ci-dessous, on obtient les résultats suivants :

Garantie	60% BR-SS
Base de remboursement (BR)	23 €
Part complémentaire maximale (PCM)	13,8 €
Part Sécurité sociale (PSS)	70% BR
Part Sécurité sociale maximale (PSSM)	16,1 €
Participation forfaitaire obligatoire (PFO)	1 €

	Sinistre 1	Sinistre 2	Sinistre 3	Moyenne des sinistres	Sinistre Moyen
Dépense Réelle	100 €	50 €	0 €	50 €	50 €
Base de Remboursement	23 €	23 €	0 €	15,33 €	15,33 €
Fréquence	1	1	0	2/3	2/3
Participation forfaitaire obligatoire	1 €	1 €	0 €	0,67 €	0,67 €
Part Sécurité sociale	15,1 €	15,1€	0 €	10,1 €	10,1 €
Reste à charge avant remboursement complémentaire	84,9 €	84,9 €	0 €	39.9 €	39.9 €
remboursement complémentaire	13,8 €	13,8 €	0 €	9,2 €	9,2 €
Reste à charge après remboursement complémentaire	71,1 €	21,1 €	0 €	30,7 €	30,7 €

TABLE 2.1 – Exemple de modélisation de la dépense moyenne

Ici, les différents calculs effectués pour obtenir la moyenne des sinistres se font en prenant en compte la moyenne des données des trois sinistres à chaque étape du calcul tandis que les calculs effectués pour obtenir le sinistre moyen se font en multipliant les paramètres par la fréquence des sinistres.

L'objectif de cette approche est de contourner la problématique qui consiste à modéliser directement le remboursement complémentaire en y intégrant de facto le niveau de la garantie complémentaire. Une alternative consiste donc à modéliser la dépense réelle ainsi que la fréquence pour en déduire le remboursement complémentaire équivalent associé à cette garantie.

En définitive, par ces différents calculs, on montre qu'il existe une équivalence entre la moyenne des remboursements complémentaires et la modélisation des dépenses réelles qui prend en compte la fréquence des sinistres pour un ensemble d'assurés avec des garanties homogènes.

2.1.2.2 Le modèle individuel

Le modèle individuel permet d'avoir une vision police par police du portefeuille étudié. En notant n le nombre total de police dans le portefeuille et S_{ind} la charge totale de sinistre correspondante sur une année donnée, on a :

$$S_{ind} = X_1 + ... + X_n = \sum_{k=1}^{n} X_k \tag{2.1}$$

où X_k pour $k \in \{1, ..., n\}$ représente la somme totale des sinistres engendrés par le $k - ieme$ police du portefeuille.

Par ailleurs, on fait aussi l'hypothèse que les $X_1, ..., X_n$ sont des **variables aléatoires indépendantes** et **identiquement distribuées**. Il faut aussi noter que pour $k \in \{1, ..., n\}$, $X_k \in [0, +\infty[$. Dans ces conditions, S_{ind} est une variable aléatoire. Son espérance, lorsqu'elle existe est définie comme suit :

$$E[S_{ind}] = E[\sum_{k=1}^{n} X_k] = \sum_{k=1}^{n} E[X_k] = nE[X_1] \tag{2.2}$$

Cette espérance correspond à la **prime pure** à estimer.

2.1.2.3 Le modèle collectif

Contrairement au modèle individuel, le modèle collectif de risque qui est une extension du modèle individuel de risque, a une vision sinistre par sinistre, indépendamment des polices correspondantes. Ainsi, en notant N le nombre de sinistres sur tout le portefeuille pour une

année donnée et S_{coll} la charge totale de sinistre correspondante, on a :

$$S_{coll} = X_1 + ... + X_n = \sum_{k=1}^{N} X_k \tag{2.3}$$

où X_k pour $k \in \{1, ..., N\}$ représente le montant engendré par la k-ieme sinistre du portefeuille. En outre, ce modèle fait les hypothèses suivantes :

- N est une variable aléatoire à valeurs dans $\mathbb{N}$
- $X_1, ..., X_N$ sont des **variables aléatoires indépendantes** et **identiquement distribuées** à valeur dans $[0, +\infty[$.

Dans ces conditions, **la prime pure** correspond à :

$$E[S_{coll}] = E[\sum_{k=1}^{N} X_k] \tag{2.4}$$

Lorsqu'on fait l'hypothèse supplémentaire de l'indépendance entre la variable aléatoire N et les variables aléatoires X_k pour $k \in \{1, ..., N\}$, on obtient :

$$E[S_{coll}] = E[\sum_{k=1}^{N} X_k] = \underbrace{E[N]}_{\text{fréquence}} \times \underbrace{E[X_1]}_{\text{Coût moyen}} \tag{2.5}$$

Cette hypothèse supplémentaire permet donc d'écrire la prime pure comme le produit de la fréquence d'occurrence du sinistre sur l'année par son coût moyen.

2.1.2.4 Le modèle fréquence - coût moyen versus le modèle coût total

Dans le modèle collectif, il existe deux approches pour modéliser la prime pure :

- **L'approche fréquence - coût moyen** : Cette approche consiste à créer deux modèles distincts. Un modèle pour la fréquence des sinistres qui correspond au nombre de sinistres par exposition et un autre pour la sévérité des sinistres qui correspond au coût des sinistres. Par la suite, les prédictions des deux modèles sont multipliées afin d'obtenir la prime pure.
- **L'approche coût total** : Cette approche consiste à modéliser directement la prime pure, c'est-à-dire le coût total par exposition.

L'approche fréquence – coût moyen à l'avantage de permettre d'avoir une vision individuelle du risque sur la fréquence et sur la sévérité ce qui permet de maîtriser et d'ajuster la prime pure selon les effets portés sur la fréquence ou sur la sévérité des coûts.

À contrario, l'approche coût total permet d'avoir une vision plus globale du risque et ne permet pas de maîtriser les effets séparément sur la fréquence et la sévérité des coûts. Ainsi, lorsqu'une variable a un effet positif fort sur la fréquence et un effet négatif tout aussi fort sur

la sévérité, cette variable aura un effet nul dans la modélisation coût total ce qui constitue une perte d'informations.

En définitive, Le choix de l'approche à utiliser est guidé par la nature des données et du risque à modéliser.

2.2 Les Modèles linéaires généralisés

2.2.1 Théorie des modèles linéaires généralisés

2.2.1.1 Introduction

Comme son nom l'indique, le modèle linéaire généralisé, formulé par ***J. A. Nelder et R. W. M. Wedderburn***[2], est une généralisation du modèle linéaire qui permet de s'affranchir de plusieurs hypothèses fondamentales du modèle linéaire qui sont entre autres :

- **Les observations de la variable à expliquer sont distribuées suivant une loi normale :**
  Cette hypothèse s'avère très vite limitée lorsqu'on souhaite modéliser des observations issues d'une loi discrète ou strictement positive comme cela est souvent le cas pour les données d'assurance.
- **Il existe une relation linéaire entre l'espérance de la variable à expliquer et les variables explicatives :**
  Cette deuxième hypothèse est elle aussi très forte et difficilement vérifiable. En effet, elle implique que l'espérance de la variable à expliquer ne peut être bornée. Ainsi, si celle-ci est binaire (survenance ou pas de sinistre) ou strictement positive (sévérité d'un sinistre), la prédiction par un modèle linéaire peut produire des valeurs en dehors de l'ensemble des valeurs acceptables.
- **La variance des variables aléatoires représentant les observations est constante :**
  Cette hypothèse n'est pas vérifiée pour les lois dont la variance est dépendante de l'espérance. C'est notamment le cas pour les lois de poisson utilisées pour modéliser les lois de comptage et les lois de gamma qui sont utilisées pour modéliser des phénomènes strictement positifs.

Plus formellement, soit $(y_1, y_2, ..., y_n)$ le vecteur d'observation, issue de la réalisation de la variable aléatoire $(Y_1, Y_2, ..., Y_n)$. et $(x_{i1}, x_{i2}, ..., x_{im})$ le i-ème vecteur ligne des variables explicatives associées à l'observation i. Alors, le modèle linéaire spécifie que :

$$Y_i = x_i\beta + \epsilon \tag{2.6}$$

Avec β le vecteur des paramètres et $Y_i \sim \mathcal{N}(x_i\beta, \sigma^2)$ pour l'observation i. Cela implique que $\mu_i = \mathbb{E}[Y_i] = x_i\beta$. Le modèle linéaire généralisé va plus loin en considérant que μ_i est lié à $x_i\beta$ par une fonction non nécessairement linéaire appelée fonction de lien telle que :

$$g(\mu_i) = x_i\beta \tag{2.7}$$

2. Nelder et Wedderburn 1972.

Ainsi, étant donné la variable réponse Y, construire un modèle GLM consiste à effectuer ces différentes étapes :

- Choisir la distribution de $f(y)$ (resp. P(Y=y)) au sein de la famille exponentielle qui modélise la variable aléatoire Y
- Choisir une fonction de lien $g(\mu_i)$
- Choisir les variables explicatives x correspondantes à la variable y qui expliquent $g(\mu)$
- Ajuster le modèle GLM en estimant le vecteur de paramètre β par maximum de vraisemblance
- Sachant $\hat{\beta}$ l'estimateur de β, générer des prédictions de μ en utilisant la relation $g(\hat{Y}_i) = x_i\hat{\beta}$

Cette généralisation permet de considérer des variables aléatoires issues d'un panel de lois appelé la famille exponentielle. Le tableau ci-dessous résume la différence entre le modèle linéaire et le modèle linéaire généralisé.

	Modèle linéaire Gaussien Modèle	linéaire Généralisé
Réponse	continue	continue ou discrète
Loi	gaussienne	famille exponentielle
Espérance	E(Y_i) = $x_i\beta$	E(Y_i) = $g^{-1}(x_i\beta)$
Variance	$\mathbb{V}ar(Y_i) = \sigma^2$	$\mathbb{V}ar(Y_i) = \mathbb{V}ar(x_i\beta)$

TABLE 2.2 – Comparaison modèle linéaire Gaussien - modèle linéaire généralisé

2.2.1.2 Famille exponentielle

On dit que la variable aléatoire Y, de densité f_Y (resp. de loi P_Y) appartient à la famille exponentielle si f_Y (resp. P_Y) s'écrit sous la forme

$$f_Y(y_i|\theta_i,\phi)\ (resp.\ P_Y(Y=y_i|\theta_i,\phi)) = exp\Big(\frac{y_i\theta_i - b(\theta_i)}{\gamma_i(\phi)} + c(y_i,\phi)\Big) \tag{2.8}$$

Où γ_i, b et c sont des fonctions déterministes connues. Le paramètre θ_i appelé paramètre naturel est le paramètre d'intérêt tandis que le paramètre ϕ est appelé paramètre de nuisance ou de dispersion et est supposé connu. Ainsi, pour une variable aléatoire Y appartenant à la famille exponentielle, on a les résultats suivants :

$$\begin{aligned} \mathbb{E}(Y) &= b'(\theta) \\ \mathbb{V}ar(Y) &= \gamma(\phi)b''(\theta) \end{aligned} \tag{2.9}$$

avec b' et b'' les dérivées premières et secondes de la fonction b .

Cette formulation inclut la plupart des lois usuelles, en l'occurrence, la loi Gamma, Poisson, Binomiale, Gaussienne, inverse Gaussienne etc... Ci-dessous, un tableau présentant ces exemples de lois.

Distribution de Y_i	ϕ	$\gamma_i(\phi)$	$b(\theta_i)$	$c(y_i, \phi)$
Normale $(\mu_i; \sigma^2)$	σ^2	ϕ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\{\frac{y_i^2}{\sigma^2} + log(2\pi\sigma^2)\}$
Poisson(μ_i)	1	ϕ	$\exp(\theta_i)$	- $\log y_i!$
Binomiale $\frac{1}{mi}(m_i; \mu_i)$	$\frac{1}{\mu_i}$	ϕ	$log(1 + exp(\theta_i))$	$\log\binom{m_i}{m_i y_i}$
Gamma $(\mu_i; \alpha)$	$\frac{-1}{\mu_i}$	$-log(-\theta)$	$\alpha log(\alpha y_i) - log(y_i) - log(\Gamma(\alpha))$	
Inverse Gaussienne$(\mu_i; \sigma^2)$	σ^2	ϕ	$-(-2\theta)^{\frac{1}{2}}$	$-\frac{1}{2}\{log(2\pi\phi y^3) + \frac{1}{\phi y_i}\}$

TABLE 2.3 – Exemple de lois de la Famille exponentielle

2.2.1.3 La fonction de lien

La fonction de lien g est une fonction bijective, définie de l'ensemble de définition de μ_i vers $\mathbb{R}$ telle que $g(\mu_i) = x_i\beta$.

Cette fonction permet d'établir un lien non nécessairement linéaire entre l'espérance de la variable d'intérêt et une combinaison linéaire des variables explicatives. D'après la relation 2.9, cela revient à choisir $g(\mu_i) = (b')^{-1}(\mu_i)$. Dans le cas de la régression linéaire, cette fonction est l'identité.

De façon générale, le choix de la fonction de lien est une liberté supplémentaire qu'offre la démarche de modélisation GLM. Toutefois, la fonction de lien canonique s'avère être un choix judicieux parce qu'elle permet d'assurer la convergence de l'algorithme d'estimation classiquement utilisé pour la maximisation de la vraisemblance (Algorithme de Newton-Raphson).

Le tableau ci-dessous présente des exemples de fonction de lien

Distribution de Y_i	Fonction de lien canonique $g(\mu_i)$
Normale $(\mu_i; \sigma^2)$	μ_i
Poisson(μ_i)	$log(\mu_i)$
Binomiale $\frac{1}{mi}(m_i; \mu_i)$	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$
Gamma $(\mu_i; \alpha)$	$\frac{-1}{\mu_i}$
Inverse Gaussienne $(\mu_i; \sigma^2)$	$\frac{-1}{2\mu_i^2}$

TABLE 2.4 – Exemple de fonctions de lien canonique

2.2.1.4 Estimation des paramètres par maximum de vraisemblance

En considérant $y = (y_1, ..., y_n)$ comme étant une réalisation de l'échantillon de n variables aléatoires indépendantes, $Y_1, ..., Y_n$ dont les fonctions de densité f_{Y_i} sont issues de la famille

exponentielle et pour chaque i, y_i la réponse en $x_i = (x_1, ..., x_n)$, la vraisemblance en y s'écrit

$$\mathcal{L}(y; \beta, \phi) = \prod_{i=1}^{n} f(y_i; \omega_i, \phi) = \prod_{i=1}^{n} f(y_i; x_i\beta, \phi) \tag{2.10}$$

Avec $\omega_i = g(\mathbb{E}(Y_i)) = x_i\beta$ et g la fonction de lien canonique. Dans ces conditions, la log-vraisemblance s'écrit :

$$l(y; \beta, \phi) = \sum_{i=1}^{n} \Big[\frac{y_i x_i \beta - b(x_i\beta)}{\gamma_i(\phi)} + c(y_i, \phi)\Big] \tag{2.11}$$

La valeur maximale de $l(y; \beta, \phi)$ est obtenue en résolvant l'équation aux dérivées partielles suivante :

$$\begin{cases} \frac{\partial l(y;\beta,\phi)}{\partial \beta_j} = 0 \; pour \; j = 1, ..., p \\ \frac{\partial l(y;\beta,\phi)}{\partial \phi} \end{cases}$$

La résolution d'un tel système ne possède pas de solution explicite. Toutefois, il existe plusieurs algorithmes d'optimisation itératifs pour obtenir les estimations du maximum de vraisemblance. Les deux algorithmes les plus utilisés sont l'algorithme de Newton-Raphson et l'algorithme du Fisher-scoring

2.2.1.5 La qualité d'ajustement

L'une des questions fondamentales qui ressort de l'ajustement d'un modèle à des données est la qualité de l'ajustement de celui-ci. Dans le cadre de la modélisation GLM, la qualité de l'ajustement est essentiellement mesurée par deux différentes statistiques : La déviance et le χ^2 de Pearson généralisé.

2.2.1.5.1 La déviance

L'idée derrière la notion de déviance est de mesurer l'écart entre un modèle GLM ajusté et le modèle parfaitement ajusté aux données appelé modèle saturé. Le modèle saturé se définit comme étant le modèle possédant autant de paramètres que d'observations.

Formellement, la déviance se définit comme étant la différence entre la log-vraisemblance du modèle ajusté et celle du modèle saturé. En notant M le modèle ajusté et S le modèle saturé, la déviance du modèle M se définit alors par :

$$D(M) = -2\phi\Big(\mathcal{L}(y; \hat{\beta}_M) - \mathcal{L}(y; \hat{\beta}_s)\Big) = -2\phi\Big(log(\frac{\mathcal{L}_M}{\mathcal{L}_S})\Big) \tag{2.12}$$

où $\hat{\beta}_M$ est l'estimateur du maximum de vraisemblance de β dans le modèle M et $\hat{\beta}_s$ désigne l'estimateur du maximum de vraisemblance de β dans le modèle saturé.

$\mathcal{L}(y;\hat{\beta}_m)$ étant toujours inférieur à $\mathcal{L}(y;\hat{\beta}_s)$, on a $D(M) \geq 0$ et $D(M) = 0$ lorsque le modèle s'ajuste parfaitement aux données.

On note aussi qu'asymptotiquement, D suit une loi du χ^2 à $n-p$ degrés de liberté, où p est le nombre de variables explicatives. Cette propriété permet de construire un test de significativité de l'écart mesuré.

On peut aussi définir la ***scaled deviance*** D^* par :

$$D^* = \frac{D}{\phi} = -2\Big(log(\frac{\mathcal{L}_M}{\mathcal{L}_S})\Big) \tag{2.13}$$

le tableau ci-dessous présente des exemples de scaled deviances

Distribution de Y_i	Déviance D*
Normale $(\mu_i;\sigma^2)$	$\sum_{i=1}^n \omega_i(y-\mu_i)^2$
Poisson(μ_i)	$2\sum_{i=1}^n \omega_i\Big\{y_i ln\frac{y_i}{\mu_i} - (y_i-\mu_i)\Big\}$
Gamma $(\mu_i;\alpha)$	$2\sum_{i=1}^n \omega_i\Big\{ - ln\frac{y_i}{\mu_i} + ln\frac{y_i-\mu_i}{\mu_i}\Big\}$
Inverse Gaussienne $(\mu_i;\sigma^2)$	$2\sum_{i=1}^n \omega_i\Big\{\frac{(y_i-\mu_i)^2}{y_i\mu_i^2}\Big\}$
Binomial $\frac{1}{mi}(m_i;\mu_i)$	$2\sum_{i=1}^n \omega_i m_i\Big\{y_i ln\frac{y_i}{\mu_i} + (1-y_i)ln\frac{1-y_i}{1-\mu_i}\Big\}$

TABLE 2.5 – Exemples de scaled Deviance

2.2.1.5.2 Le χ^2 de Pearson

Le χ^2 de Pearson est également utilisé pour comparer l'écart entre le modèle ajusté et les données observées. Le χ^2 de Pearson la statistique définie par :

$$\chi^2 = \sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{\mathbb{V}ar(\hat{\mu}_i)} \tag{2.14}$$

Avec $\hat{\mu}_i = g^{-1}(x_i\hat{\beta})$. L'écart observé n'est pas significatif au niveau α si la valeur observée de χ^2 est supérieure au quantile $\chi^2_{n-p,1-\alpha}$

2.2.2 Modélisation des fréquences

Soit un assuré i avec un nombre de sinistres noté K_i. Alors K_i est une variable aléatoire discrète à valeur dans $\{0,1,2,....\}$. La probabilité que l'assuré i déclare k sinistres est alors noté $P(K_i = k)$.

2.2.2.1 Le modèle de poisson

Le modèle de régression de Poisson est un modèle issu de la distribution de Poisson qui permet de modéliser le nombre d'occurrences d'un évènement rare dans un intervalle de temps donné.

La variable aléatoire K_i suit une loi de poisson de paramètre λ_i si pour $k \in \mathbb{N}$, la distribution de K_i s'écrit :

$$P(K_i = k|\lambda_i) = \frac{\lambda_i^k e^{-\lambda_i}}{k!} \tag{2.15}$$

De plus, l'espérance et la variance de K_i sont telles que $\mathbb{E}(K_i) = \mathbb{V}ar(K_i) = \lambda_i$.

Dans ces conditions, la vraisemblance du modèle de poisson s'écrit :

$$\mathcal{L}(\beta|K, X) = \prod_{i=1}^{n} P(K_i = k_i|\lambda_i) \prod_{i=1}^{n} \frac{\lambda_i^{k_i} e^{-\lambda_i}}{k_i!} \tag{2.16}$$

Où $\lambda_i = \mathbb{E}(K_i|x_i) = e^{x_i\beta}$.

Le modèle de régression de Poisson fournit un cadre de base pour l'analyse des données de comptage. Toutefois, l'une de ces hypothèses, en l'occurrence l'égalité entre la variance et l'espérance, reste très forte. En effet, on observe en pratique que les données de comptage présentent souvent une variance plus importante que celle prédite par la moyenne. On parle alors de surdispersion. Pour prendre en compte ce phénomène de surdispersion, d'autres lois plus adaptées sont utilisées.

2.2.2.2 La loi Négative Binomiale

En assurance et plus particulièrement en santé, les données mettent très souvent en évidence un phénomène de surdispersion qui peut être généré par des facteurs de risques non observés (qualité de vie, consommation de drogue, d'alcool, de tabac ...). Ce phénomène peut être modélisé en introduisant un facteur d'hétérogénéité aléatoire.

Formellement, sachant $\Theta_i = \theta$, la variable aléatoire K_i modélisant le nombre de sinistres de l'assuré i suit une loi de poisson de paramètre $\mu_i\theta$. Ici, on suppose que $\Theta_1, ..., \Theta_n$ sont des variables aléatoires indépendantes telles que $\forall\ i\ \in\ \{1, ..., n\}$, Θ_i suit une loi Gamma de paramètres identiques α_i (tels que $\mathbb{E}(\Theta_i) = 1$) avec α_i qui dépend des valeurs des variables explicatives x_i.

Dans ces conditions, on obtient une loi Négative Binomiale définie par :

$$P(K_i = k|\mu_i, \alpha_i) = \frac{\Gamma(k + \alpha_i^{-1})}{\Gamma(\alpha_i^{-1})\ k!} \left[\frac{\alpha_i^{-1}}{\alpha_i^{-1} + \mu_i}\right]^{\alpha_i^{-1}} \left[\frac{\mu_i}{\alpha_i^{-1} + \mu_i}\right]^{k} \forall\ k\ \in\ \mathbb{N} \tag{2.17}$$

où $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt\ \forall\ k > 0$.

La vraisemblance du modèle de régression Binomiale négative est dans ces conditions définie par :

$$\mathcal{L}(\beta|X) = \prod_{i=1}^{n} P(K_i = k_i|x_i, \alpha) = \prod_{i=1}^{n} \frac{\Gamma(k_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\, k!} \left[\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right]^{\alpha^{-1}} \left[\frac{\mu}{\alpha^{-1} + \mu}\right]^{k} \tag{2.18}$$

Où $\mu_i = \mathbb{E}(K_i|x_i) = e^{x_i\beta}$, $\alpha_i > 0$. Pour $\alpha_i = 0$ on retrouve le modèle de poisson.

On note par ailleurs que $\mathbb{V}ar(K_i|x_i) = \mu_i(1 + \alpha_i\mu_i)$. Cette paramétrisation correspond au modèle Binomiale négative de type 2.

Plus généralement, Cameron et Trivedi [3] considèrent une classe plus générale de distribution binomiale négative (NBp) ayant la même moyenne μ_i, mais une variance de la forme $\mu_i(1 + \alpha_i\mu_i^{p-1})$ avec p le troisième paramètre à estimer.

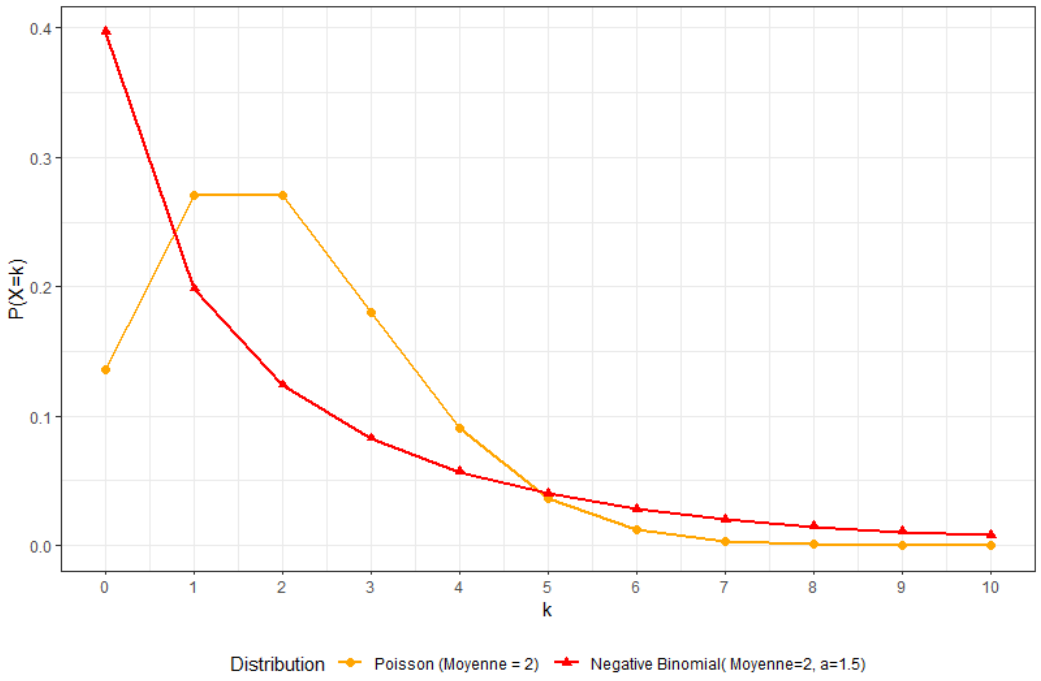


FIGURE 2.1 – Comparaison de la loi de Poisson versus Negative Binomial pour une moyenne = 2

2.2.2.3 Les lois Zero inflated

Un autre problème que l'on peut rencontrer lors de la modélisation des données de comptage en assurance santé est une surreprésentation de zéros (la variable à expliquer possède une proportion de zéros beaucoup plus importante que ce que prévoit l'hypothèse de loi).

3. CAMERON et TRIVEDI 1986.

Une solution à ce problème est d'utiliser les modèles zéro-inflated qui permettent de tenir compte de deux sources distinctes zéros. Une source est supposée être générée par des individus qui n'entrent pas dans le processus de comptage et l'autre source est générée par ceux qui entrent dans le processus de comptage, mais ne déclarent aucun sinistre.

En prenant l'exemple de données relatives aux visites chez un médecin spécialiste avec pour objectif de modéliser le nombre de visites effectuées par les patients sur une période donnée, supposons que ces données contiennent également des informations sur des personnes qui n'ont pas accès à des médecins pendant la période spécifiée. Ces patients reçoivent alors un nombre de sinistre égal à zéro.

Lorsqu'il y a beaucoup plus de résultats de zéros observés (les individus n'ayant pas visité un médecin) que ceux que prévoit la distribution de Poisson ou binomiale négative sur laquelle la modélisation est basée, le modèle Zéro-inflated propose de diviser le processus de comptage en deux parties. Une première partie du modèle se concentre sur la question de savoir si l'individu est entré dans le processus de comptage et une autre partie modélise (lorsque l'individu est entré dans le processus de comptage) le nombre de visites effectuées chez un médecin spécialiste.

Ce modèle initialement introduit par **Diane Lambert** [4] dans le cadre du contrôle des processus en fabrication a depuis été utilisé dans de nombreuses applications notamment en assurance. Une variable aléatoire K_i suit une loi Zéro-inflated si K_i peut être écrite comme un mélange d'une probabilité π_i de Dirac en 0 et une distribution de comptage p_i (Poisson ou négative binomial)

$$\mathbb{P}(K_i = k) = \begin{cases} \pi_i + [1 - \pi_i] \cdot p_i(0) \; pour \; k = 0 \\ [1 - \pi_i] \cdot p_i(k) \; pour \; k = 1, 2, \dots \end{cases} \tag{2.19}$$

Ainsi, pour une loi de comptage p_i d'espérance μ_i on a :

$\mathbb{E}(K_i) = [1 - \pi_i]\mu_i$ et $\mathbb{V}ar(K_i) = \pi_i\mu_i + \mu_i^2[1 - \pi_i] > \mathbb{E}(K_i)$

2.2.2.4 Les modèles Hurdle

De même que les modèles Zéro-inflated, les modèles Hurdle sont des modèles de comptage utilisés pour prendre en compte la surreprésentation des zéros observés dans les données de comptage. Initialement développés par **John Mullahy** [5], les modèles Hurdle sont basés sur un processus dichotomique, où les assurés qui déclarent au moins un sinistre sont considérés comme complètement différents de ceux qui ne déclarent rien. Pour des fonctions de masse $p_i^{(1)}$

4. Lambert 1992.
5. Mullahy 1986.

et $p_i^{(2)}$, la loi de la variable aléatoire K_i qui suit une loi Hurdle est définie par :

$$\mathbb{P}(K_i = k) = \begin{cases} p_i^{(1)}(0) \; pour \; k = 0 \\ \frac{1-p_i^{(1)}(0)}{1-p_i^{(2)}(0)} p_i^{(2)}(k) \; pour \; k = 1, 2, ... \end{cases} \tag{2.20}$$

Ainsi, pour une loi de comptage $p_i^{(1)}$ et $p_i^{(2)}$ d'espérances $\mu_i^{(1)}$ et $\mu_i^{(2)}$ respectivement, on a :

$$\mathbb{E}(K_i) = \frac{1 - p_i^{(1)}(0)}{1 - p_i^{(2)}(0)} \mu_i^{(2)}$$
$$\mathbb{V}ar(K_i) = \mathbb{P}(K_i = 0) \cdot \mathbb{V}ar(K_i | K_i > 0) + \mathbb{P}(K_i = 0) \cdot \mathbb{E}(K_i | K_i > 0)$$

2.2.3 Modélisation du coût moyen

La modélisation de la sévérité des sinistres peut être effectuée en utilisant plusieurs modèles possibles qui diffèrent dans leurs hypothèses et sur la répartition des sinistres. Il convient de noter qu'une distribution de sévérité de sinistre est généralement composée de valeurs continues et positives. De plus, la distribution de la sévérité des sinistres est généralement concentrée à gauche (sur les petits sinistres) et à queue épaisse à droite (sur les grands sinistres). En effet, les sinistres importants ne sont généralement pas fréquents, tandis qu'un plus grand nombre de sinistres relativement petits sont courants en assurance en général et plus particulièrement en assurance santé. Des exemples de telles distributions sont les distributions Log-normal et Gamma (**Ohlsson, Johansson et John**) [6].

Cependant, il existe des risques pour lesquels la distribution des sinistres n'est pas continue. C'est notamment le cas en assurance santé lorsque les coûts engendrés par la visite chez un médecin généraliste sont modélisés.

En effet, comme décrit en 1.1.2, les tarifs en vigueur chez les médecins généralistes sont conventionnés. Dans ce cas de figure, si Y est la variable d'intérêt et $y_1, y_2, ..., y_n$ sont les valeurs prise par Y, alors le coût moyen est calculé de la manière suivante : $\mathbb{E}[Y] = \sum_{i=1}^{n} y_i \mathbb{P}(Y = y_i)$. Les différentes probabilités $\mathbb{P}(Y = y_i)$ sont estimés par des modèles binomiaux ou multinomiaux.

Dans cette partie, X_i pour $i \in \{1, ..., N\}$ représente le montant engendré par le k-ieme sinistre du portefeuille.

2.2.3.1 Loi Gamma

La variable aléatoire continue Y_i a une distribution Gamma avec le paramètre de forme α_i et le paramètre d'échelle θ_i si sa fonction de densité de probabilité est donnée par :

$$f_{Y_i}(y) = \frac{(y/\theta)^{\alpha}}{y\Gamma(\alpha)} exp(-y/\theta) \; pour \; y > 0. \tag{2.21}$$

6. Ohlsson et Johansson 2010.

où $\alpha > 0, \theta > 0$.

La moyenne et la variance de cette distribution sont données par

$$\mathbb{E}(Y) = \alpha\theta$$
$$\mathbb{V}ar(Y) = \alpha\theta^2$$

En modifiant α_i et θ_i, la distribution Gamma permet de représenter de nombreuses formes (2.2). La forme polyvalente de la distribution Gamma en fait un bon candidat pour représenter la sévérité des sinistres en assurance.

La fonction de vraisemblance de N observations $(Y_1, ..., Y_N)$ s'écrit

$$l(\alpha, \theta) = \prod_{i=1}^{N} log(f(Y_i; \alpha, \theta)) = (k-1)\sum_{i=1}^{N} log(Y_i) - \sum_{i=1}^{N} \frac{Y_i}{\theta} - Nklog(\theta) - Nlog(\Gamma(k)) \quad (2.22)$$

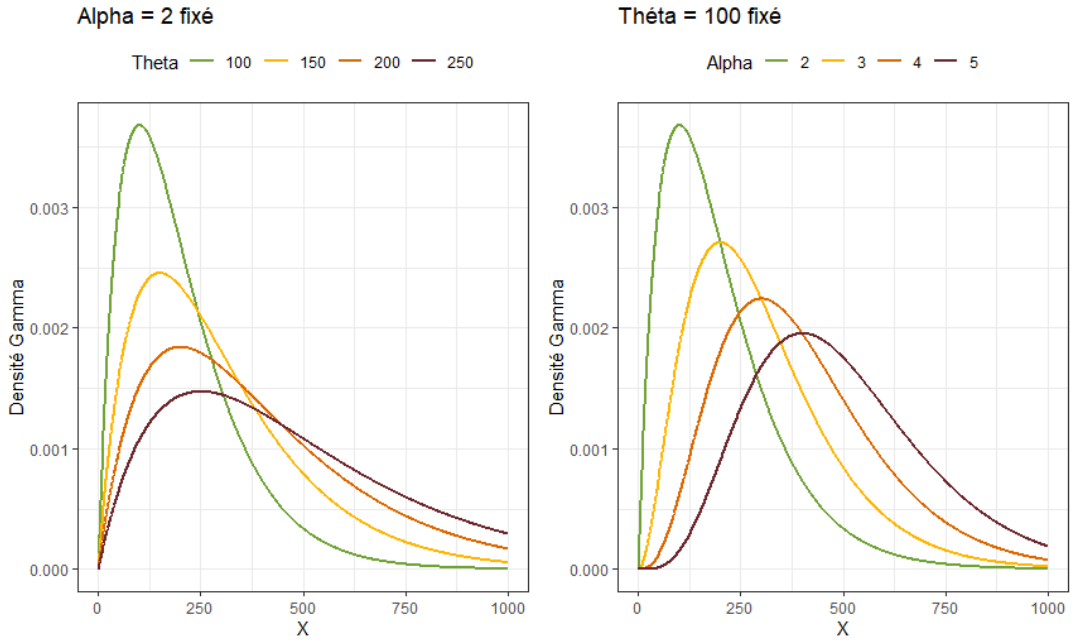


FIGURE 2.2 – Densité de la loi Gamma en fonction des paramètres θ et α

2.2.3.2 Réponses binaires et multinomiales

Dans un certain nombre de cas, la distribution des sinistres présente une structure discrète. Dans ce cas de figure, modéliser le coût moyen nécessite la modélisation de la probabilité d'occurrence de chaque modalité de la distribution.

Lorsque la distribution est composée de deux modalités, le modèle Binomial est utilisé pour modéliser les probabilités d'occurrence. Lorsque le nombre de modalités $m > 2$, le modèle mutinomiale est utilisé

2.2.3.2.1 Modèle binomial :

La variable aléatoire discrète Y a une distribution Bernoulli de paramètre $p_\beta(X)$ lorsque Y prend la valeur 1 avec probabilité $p_\beta(X)$ et 0 avec la probabilité $1 - p_\beta(X)$. En d'autres termes :

$$\mathbb{P}(Y = y_i|x_i;\beta) = p_\beta(x_i)^{y_i}\big(1 - p_\beta(x_i)\big)^{y_i} \tag{2.23}$$

Dans ces conditions, la vraisemblance de ce modèle s'écrit

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} log(\mathbb{P}(Y = y_i|x_i;\beta)) = \sum_{i=1}^{n} log(p_\beta(x_i)^{y_i}\big(1 - p_\beta(x_i)\big)^{y_i} \tag{2.24}$$

Dans le cas de la régression Logistique, $\mathbb{P}(Y = y_i|x_i;\beta)$ s'ecrit :

$$\mathbb{P}(Y = y_i|x_i;\beta) = \frac{1}{1 + exp(-\beta^T x_i)} \tag{2.25}$$

2.2.3.2.2 Modèle multinomial :

Considérons maintenant le cas où la variable aléatoire discrète Y possède $M > 2$ modalités.

En notant $(Y = m)$ la variable aléatoire qui vaut 1 lorsque $(Y = m)$ est vrai et 0 sinon, la loi jointe de $((Y = 1), ..., (Y = M))$ est une loi multinomiale de paramètre $p = (p_1, p_2, ..., p_M)$ où pour $\forall\ m \in \{1, ..., M\}$, $p_m = P[Y = m]$. On note par ailleurs que $\sum_{m=1}^{M} p_k = 1$.

Ainsi, pour estimer $p_{m|x} = P[Y = m|x]$, on choisit un évènement de référence (dans notre cas de figure, M) et on considère les $M - 1$ modèles de régression logistiques dans lesquels on effectue des régressions par rapport à la référence choisie. On obtient :

$$p_{m|x} = P[Y = m|x] = \frac{exp(\beta_m^T x)}{1 + \sum_{i=1}^{M-1} exp(\beta_i^T x)}\ \forall\ m = 1, ..., M-1,\ et, \tag{2.26}$$

$$p_{M|x} = P[Y = M|x] = 1 - \sum_{m=1}^{M-1} p_{m|x}\ pour\ M \tag{2.27}$$

2.2.3.3 Loi log-normal

La variable aléatoire continue X_i a une distribution log-normale si sa fonction de densité de probabilité est donnée par :

$$f_{X_i}(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} exp(\frac{(ln(x) - \mu)}{2\sigma^2})\ pour\ x > 0. \tag{2.28}$$

où les paramètres $\mu > 0$, $\sigma > 0$ sont l'espérance et l'écart type du lagorithme de X_i. La moyenne et la variance de cette distribution sont données par

$$\mathbb{E}(X_i) = exp(\mu + \frac{1}{2}\sigma^2)$$
$$\mathbb{V}ar(X_i) = exp(2\mu + 2\sigma^2) - exp(2\mu + \sigma^2)$$

Par ailleurs, on note aussi que si $X_i \sim lognormale(\mu, \sigma^2)$ alors $X_i^* = log(X_i) \sim \mathcal{N}(\mu, \sigma^2)$.

De même que la loi Gamma, la loi log-normale à une variance proportionnelle au carré de son espérance. On observe par ailleurs que pour des petites valeurs de σ^2 la loi log-normal se rapproche de la loi Gamma.

2.2.4 Modélisation du coût total avec la loi de Tweedie

En assurance, le coût total engendré par une police d'assurance sur un risque couvert, a traditionnellement une distribution continue, avec des valeurs positives ou nulles. Une approche standard pour modéliser ces types de données est l'utilisation d'un modèle de Tweedie.

En considérant la somme $S = X_1 + ... + X_N = \sum_{k=1}^{N} X_k$ définie en 2.3, représentant le coût total de sinistres et en faisant les hypothèses suivantes :

- N, représentant le nombre de sinistres, suit une loi de poisson de paramètre λ
- les observations $X_1, ..., X_N$ sont indépendantes et issues d'une variable aléatoire de loi Gamma(α, γ) et indépendante de N

Alors, S suit une loi de Tweedie introduit par Jorgensen [7].

La distribution de Tweedie est une distribution de la famille exponentielle telle que :

$$\mathbb{V}ar(S) = \phi[\mathbb{E}(S)]^p,$$

où ϕ est le paramètre de dispersion et p le paramètre de puissance.

Lorsque $p = 0$, la distribution de Tweedie a une variance constante et se réduit à la loi normale. Lorsque $p = 1$, la distribution obtenue est une distribution de poisson et $p = 2$ correspond à la loi Gamma. Enfin, lorsque $p \in]1, 2[$, S est un mélange de distribution de poisson et de Gamma idéal pour modéliser le coût total par assuré.

2.3 Grandiant boosting

Plusieurs modèles d'apprentissage machine (Machine Learning) sont fondés sur un seul modèle prédictif (la régression linéaire, la régression linéaire généralisée, les machines à vecteurs de support, etc.). D'autres approches, dites *ensemblistes*, telles que les algorithmes de Bagging

7. Jørgensen 1987.

(Bootstrap aggregation) et de forêts aléatoires (Random Forest), sont fondés sur plusieurs modèles prédictifs. Ces approches reposent sur l'idée de construire un ensemble de modèles où chaque modèle prédit un résultat individuel, puis ces résultats sont agrégés par une moyenne pondérée.

Basés sur une stratégie différente, les modèles de Gradient Boosting sont des modèles dits *ensemblistes*, qui reposent sur une approche séquentielle dite *adaptative* qui consiste à combiner plusieurs modèles avec des pouvoirs de prédiction faibles (weak learners) pour obtenir un modèle avec un pouvoir de prédiction puissant.

Le graphique ci-dessous résume les différences de fonctionnement de ces différents types de modèle.

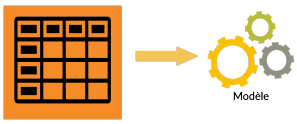
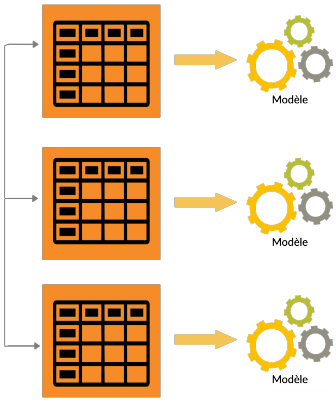
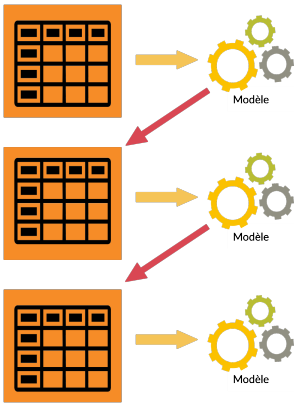
Modèle à un prédicteur	Bagging	Boosting
		
Construction unique	Construction parallèle	Construction Séquentielle

TABLE 2.6 – Tableau comparatif entre les modèles à un prédicateur, les modèles Bagging et les modèles Boosting

2.3.1 L'idée de base : AdaBoost

Ada-boost ou Adaptive Boosting est la méthode de classification élémentaire à deux classes à l'origine des algorithmes de boosting proposé par Yoav Freund et Robert Schapire en 1996 [8]. Elle permet de combiner plusieurs classificateurs pour obtenir un meilleur pouvoir prédictif.

Freund et Schapire ont illustré leur algorithme à travers un exemple d'une course de chevaux. Dans cette course, un joueur souhaite parier sur le cheval qui a le plus de chance de gagner. Afin d'augmenter la probabilité de gain de son pari, le joueur souhaite recueillir les avis d'experts avant de parier. Un tel processus de collecte d'informations auprès de différents experts est similaire à la collecte d'informations effectuée dans l'algorithme AdaBoost par le biais d'un ensemble de classificateurs faibles .

8. FREUND, SCHAPIRE et al. 1996.

l'algorithme AdaBoost ajuste un classificateur faible aux versions pondérées des données de manière itérative. À chaque itération, les données sont pondérées à nouveau de sorte à donner un poids plus important aux points mal classés. Le modèle qui en résulte s'écrit :

$$\hat{f}_M(x) = \sum_{i=1}^{M} \theta_i c_i(x) \tag{2.29}$$

où M est le nombre de classificateurs final, $c_i(x) \in \{-1, 1\}$ le i-ème classificateur et θ_i la i-ème pondération.

Opérationnellement, l'algorithme se présente comme suit :

Algorithme 1 : AdaBoost ou (adaptive boosting)

Soit x_0 à prévoir et ;
$z = \{(x_1, y_1), ..., (x_n, y_n)\}$ un échantillon;
INITIALISER les poids $w = \{w_i = 1/n; i = 1, ..., n\}$;
for $i = 1$ *à* M **do**
  Estimer c_i sur l'échantillon pondéré par w ;
  Calculer le taux d'erreur apparent :

$$\hat{\epsilon}_p = \frac{\sum_{k=1}^{n} w_k \mathbf{1}\{c_i(x_k) \neq y_k\}}{\sum_{k=1}^{n} w_k}$$

  ;
  Calculer les logits $\theta_i = log((1 - \hat{\epsilon}_p)/\hat{\epsilon}_p)$;
  Calculer les nouvelles pondérations :

$$w_k \leftarrow w_k \cdot exp[\theta_i \mathbf{1}\{c_i(x_k) \neq y_k\}; k = 1, ..., n\}]$$

end
Résultat du vote : $\hat{f}_M(x_0) = signe[\sum_{i=1}^{I} \hat{\theta}_i \hat{c}_i(x_0)]$

Dans cet algorithme, le paramètre M est le paramètre à optimiser par validation croisé pour éviter le surajustement. Le graphique ci-dessous présente une illustration de l'algorithme AdaBoost.

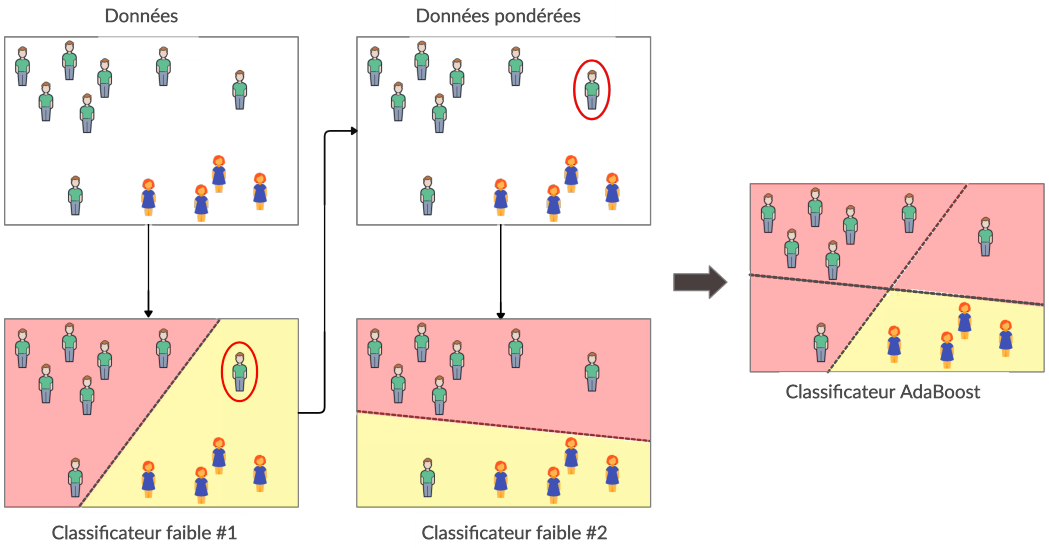


FIGURE 2.3 – Illustration de l'algorithme AdaBoost

2.3.2 Gradient Boosting Machine (GBM)

2.3.2.1 Principe général

Dans la continuité des travaux de Freund et Schapire [9], Breiman a mis en évidence le fait que l'AdaBoost est en fait un algorithme de type descente de gradient dans un espace de fonctions, identifiant ainsi le boosting à la frontière de l'optimisation numérique et l'estimation statistique. Plus tard, dans la même continuité, des algorithmes de classifications et de régression, ont été proposés par Friedman [10] sous l'acronyme MART (multiple additive regression trees) puis sous celui de GBM (Gradient Boosting models) généralisant ainsi l'AdaBoost.

Concrètement, considérons le vecteur de variables explicatives $x = (x_1, x_2, ..., x_p)^T$ et y la variable à expliquer. Le but de l'algorithme de Gradient Boosting est d'estimer la fonction optimale $\hat{F}$ qui permet de faire un lien entre y et x en minimisant l'espérance d'une certaine fonction de perte $\mathscr{L}$ qui appartient à une classe de fonction $\mathcal{F}$. $\hat{F}$ s'écrit :

$$\hat{F} = \underbrace{arg\,min}_{F(\cdot)\in\mathcal{F}} E_{y,x}[\mathscr{L}(y, F(x))] \tag{2.30}$$

où $\mathscr{L}$ est supposée être différentiable par rapport à F. Ainsi, sachant les observations $\{y_i, x_i\}_{i=1}^n$ où $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$, $\hat{F}$ peut s'écrire en minimisant la fonction de risque

9. FREUND, SCHAPIRE et al. 1996.
10. FRIEDMAN 2002.

empirique de la manière suivante :

$$\hat{F} = \underbrace{arg\,min}_{F(\cdot)\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}(y, F(x)) \tag{2.31}$$

Dans cet algorithme, on fait l'hypothèse que toute fonction $F \in \mathcal{F}$ est un ensemble de M ajustements de base tels que :

$$F(x) = F_0 + \sum_{m=1}^{M} \gamma_m c_m(x) \tag{2.32}$$

où $\forall\ m \in \{1, ..., M\}$, c_m est un ajustement de base et γ_m

Cette approche est généralement utilisée avec des arbres de décision de taille fixe en tant que classificateurs faibles (ou ajustement de base dans le cadre d'une régression). Dans ce contexte, on parle alors de **gradient tree boosting**.

Ci-dessous, la description de l'algorithme de Gradient Boosting pour la régression.

Algorithme 2 : Gradient Tree Boosting pour la régression

Soit x_0 à prévoir ;
INITIALISER s $\hat{F}_0 = arg\,min_\gamma \sum_{i=1}^{n} \mathscr{L}(y_i, \gamma)$;
for $m = 1$ *à* M **do**
  Calculer $r_{mi} = -\Big[-\frac{\partial \mathscr{L}(y_i, F(x_i))}{\partial F(x_i)} \Big]_{F=F_{m-1}}$; i=1,...,m ;
  Ajuster un arbre de régression c_m aux couples $(x_i, r_{mi})_{i=1,...,n}$;
  Calculer γ_m en résolvant : $arg\,min_\gamma \sum_{i=1}^{n} \mathscr{L}(y_i, F_{m-1}(x_i) + \gamma c_m(x_i))$;
  Mise à jour : $\hat{F}_m(x) = \hat{F}_{m-1}(x) + \gamma_m c_m(x)$
end
Résultat : $\hat{F}_M(x_0)$

2.3.2.2 Régularisation : éviter le surajustement

Un ajustement trop proche des données d'apprentissage peut entraîner une détérioration de la capacité de généralisation du modèle. Plusieurs techniques dites de régularisation réduisent cet effet de surajustement en encadrant la procédure d'ajustement. Parmis ces techniques, on peut citer :

- **La méthode du rétrécissement (Shrinkage)** est une technique couramment utilisée dans les algorithmes de Gradient Boosting qui consiste à modifier la règle de mise à jour à chaque itération de la manière suivante :

$$\hat{F}_m(x) = \hat{F}_{m-1}(x) + \delta\gamma_m c_m(x),\ 0 < \delta \leq 1,$$

Où δ est appelé "taux d'apprentissage" et permet de controler la vitesse de convergence. Si sa valeur est petite cela conduit à accroître le nombre d'ajustements de base nécessaire et entraîne généralement une amélioration de la qualité de prédiction. Le boosting pouvant converger exactement, δ grand peut conduire à une situation de surajustement

- **Stochastic Gradient Boosting** : peu après l'introduction au Gradient Boosting, Friedman a proposé une modification mineure de l'algorithme, inspirée de la méthode d'agrégation bootstrap (bagging). Il a notamment proposé qu'à chaque itération de l'algorithme, un ajustement de base soit ajusté sur un sous-échantillon des données d'apprentissage tiré au hasard sans remise afin de construire une séquence de prédicteurs plus indépendants. Cette approche conduit à une amélioration substantielle de la précision de l'algorithme de Gradient boosting. Le taux de sous-échantillonnage est un autre paramètre à optimiser.

Par ailleurs, un paramètre de régularisation naturel est le nombre d'itérations M de Gradient Boosting (le nombre d'ajustements de base dans le modèle). Augmenter M réduit l'erreur de prédiction sur les données d'apprentissage, mais peut entraîner un surajustement. Une valeur optimale de M est souvent sélectionnée par validation croisée.

Dans le cas où l'ajustement de base est un arbre, la profondeur maximale de l'arbre est un autre paramètre à optimiser pour contrôler la complexité de l'algorithme. Plus la profondeur maximale est grande, plus l'algorithme devient complexe, mais est capable de capter les interactions complexes entres les variables.

2.3.2.3 eXtreme Gradient Boosting : XGBoost

Initialement développé par Tianqi Chen et C. Guestrin en 2016, l'eXtreme Gradient Boosting (XGBoost) fait maintenant partie d'une collection plus large de bibliothèques open-source développées par la Distributed Machine Learning Community (DMLC). XGBoost est une implémentation évolutive et très puissante du Gradient Boosting Machine.

L'implémentation du XGBoost offre plusieurs fonctionnalités avancées pour l'optimisation des modèles, l'optimisation de l'utilisation des ressources de calcul et l'optimisation des algorithmes. Il est capable d'exécuter les trois principales formes de Gradient Boosting (Gradient Boosting, Stochastic Gradient Boosting et le Regularized Gradient Boosting) et il est suffisamment robuste pour supporter un réglage fin et l'ajout de paramètres de régularisation. En résumé, les principaux atouts et fonctionnalités de cette implémentation du Gradient Boosting sont les suivants :

- **Régularisation :** Peut effectuer la régularisation L1 ou L2 pour contrôler le surajustement.
- **Traitement de données creuses :** Intègre un algorithme de recherche de fractionnement de données en fonction du type de données, pour traiter différents types de modèles de données creuses.
- **Algorithme schéma de quantile pondéré (weighted quantile sketch algorithm) :** Utilise un algorithme de schéma de quantile pondéré distribué pour traiter efficacement les données pondérées.

- S**tructure en blocs pour la parallélisation de l'apprentissage :** Rend possible l'utilisation de plusieurs coeurs sur le CPU grâce à une structure en blocs dans la conception du système. La structure en blocs permet de réutiliser les données déjà traitées.
- **Sensibilisation au cache :** Attribue des tampons internes dans chaque cœur du CPU, dans lesquels les statistiques de gradient peuvent être stockées.
- **Calcul hors des cœurs du CPU :** Optimise l'espace disque disponible et maximise son utilisation lors de la manipulation de grande base de données qui ne peuvent tenir dans la mémoire RAM.

2.3.3 Boosted GLM

L'algorithme de Gradient Boosting introduit en 2.3.2 nécessite la spécification d'une fonction de perte qui doit être minimisée pendant la phase de d'apprentissage du modèle. Cette section fait une introduction générale d'une telle fonction pour des tâches assimilables au modèle GLM.

2.3.3.1 Fonction de perte (Loss function)

La fonction de perte est une fonction qui permet d'évaluer la qualité de la modélisation des données d'apprentissage par un algorithme de Machine Learning. Lorsque les prédictions s'écartent beaucoup des observations réelles, la fonction de perte prend de grandes valeurs. A contrario, lorsque les prédictions sont assez bonnes, la fonction de perte prend des valeurs plus petites. Ainsi, Progressivement, à l'aide d'une telle fonction, l'algorithme de Gradient Boosting améliore sa qualité de prédiction pour atteindre son optimum.

Dans le cadre des problèmes de régressions, la fonction de perte standard est la fonction des erreurs au carré (sum squared error loss) définie par :

$$\mathscr{L}(y, F(x)) := \sum_{i=1}^{n} (y_i - F(x_i))^2 \tag{2.33}$$

Cette fonction de perte n'est pas nécessairement un choix pertinent pour la modélisation des données de comptage, des données positives et asymétriques comme on en rencontre lors de la modélisation des données assurentielles. En effet, comme le montre l'article ***Boosting insights in insurance tariff plans with tree-based Machine Learning methods*** [11], en considérant la notion de déviance définie en (2.12), la déviance d'une distribution gaussienne de variance

11. Henckaerts et al. 2019.

constante s'exprime de la manière suivante :

$$\begin{aligned}\mathcal{D}[y, F(x)] &= -2ln\Big[\frac{\mathcal{L}(F(x_i))}{\mathcal{L}(y_i)}\Big] \\ &= 2ln\prod_{i=1}^{n} exp\Big\{-\frac{1}{2\sigma^2}(y_i - y_i)^2\Big\} - 2ln\prod_{i=1}^{n} exp\Big\{-\frac{1}{2\sigma^2}(y_i - F(x_i))^2\Big\} \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - F(x_i))^2\end{aligned}$$

Ce qui est équivalent à une somme des erreurs au carré définie précédemment.

Ce résultat implique qu'une fonction de perte basée sur les erreurs au carré est appropriée lorsque les données à prédire sont suffisamment Gaussien. Plus généralement, cette fonction de perte est appropriée pour des données à prédire, centrée autour de la moyenne, avec une variance constante. Cependant, les données généralement rencontrées en assurance (Gamma, Log-normal, Poisson, etc.) n'ont pas de telles caractéristiques. Une solution alternative consiste à utiliser des fonctions de perte plus adaptées et inspirées par les modèles GLM.

En effet, en considérant par exemple que les coûts moyens sont distribués selon une loi Gamma $(F(x_i), \alpha)$, la fonction de perte inspirée du modèle GLM serait :

$$\begin{aligned}\mathcal{D}[y, F(x)] &= -2ln\Big[\frac{\mathcal{L}(F(x_i))}{\mathcal{L}(y_i)}\Big] \\ &= 2ln\prod_{i=1}^{n}\frac{1}{y_i\Gamma(\alpha)}\Big(\frac{\alpha y_i}{y_i}\Big)^{\alpha}\exp\Big(-\frac{\alpha y_i}{y_i}\Big) - 2ln\prod_{i=1}^{n}\frac{1}{y_i\Gamma(\alpha)}\Big(\frac{\alpha y_i}{F(x_i)}\Big)^{\alpha}\exp\Big(-\frac{\alpha y_i}{F(x_i)}\Big) \\ &= 2\alpha\sum_{i=1}^{n}\Big\{\frac{y_i - F(x_i)}{F(x_i)} - \ln\Big(\frac{y_i}{F(x_i)}\Big)\Big\} = 2\alpha\mathscr{L}_{Gamma}(y, F(x))\end{aligned}$$

Le paramètre d'échelle α étant constant, il peut être ignoré. Cette fonction de perte ainsi définie est dérivable par rapport à $F(x_i)$ et minimisée lorsque le maximum de vraisemblance est maximisé.

En définitive, on pourrait construire des fonctions de perte de la même façon, pour toutes les distributions appartenant à la famille exponentielle. Ces fonctions de perte ont l'avantage de tenir compte de la courbure de la distribution supposée être sous-jacente.

2.3.3.2 L'approche boosted GLM

Comme défini en (2.7), le modèle linéaire généralisé fait l'hypothèse que :

$$g(\mu) = x\beta = F(x)$$

Où F est une fonction linéaire de x_i. L'approche Boosted GLM consiste à s'affranchir de l'hypothèse de linéarité tout en conservant l'hypothèse de loi à travers des algorithmes de Gradient Boosting. l'hypothèse de loi est ainsi conservée en utilisant des fonctions de perte

inspirées du GLM définies précédemment. Ainsi, on obtient une la formulation du boosted GLM de la façon suivante :

$$g(\mu) = F(x),$$

Où F est une fonction non nécessairement linéaire de x. Cette flexibilité de F permet de capter toute forme de relation qui lie F à x ainsi que toutes les interactions possibles entre les variables explicatives.

2.4 Mesure de la qualité de la prédiction

Si l'analyse des données ainsi que la phase d'apprentissage du modèle sont des étapes importantes dans la construction d'un modèle de Machine Learning, c'est aussi le cas de la mesure de la qualité de la prédiction. La mesure de la qualité de la prédiction permet de mesurer la capacité d'un modèle à se généraliser sur de nouvelles données. Cette mesure s'obtient à l'aide d'une métrique d'évaluation appelée aussi métrique de validation.

Plusieurs métriques d'évaluation existent. Les plus connues pour les problèmes de régression sont le RMSE, MSE, MAE, l'indice de Gini etc. Cependant, ces métriques donnent une vision globale de la qualité de la prédiction du modèle sur l'ensemble des données. De plus, certaines métriques sont moins adaptées pour des données assurantielles. En l'occurrence, le RMSE, qui est l'une des métriques les plus utilisées pour l'évaluation des modèles de Machine Learning, est une métrique très sensible aux valeurs extrêmes et mieux adaptée aux distributions symétriques.

Dans le cadre de ce mémoire, nous avons élaboré un processus d'évaluation mieux adapté à nos objectifs du métier afin d'évaluer efficacement nos différents modèles.

2.4.1 Quelques définitions pour mieux comprendre le processus d'évaluation

2.4.1.1 Les segments élémentaires

Dans le cadre de ce mémoire, on définit comme segment élémentaire, toute combinaison d'instance des variables explicatives. Plus concrètement, considérons que l'on a deux variables explicatives (**Sexe** et **Type de bénéficiaires**). Un segment élémentaire serait par exemple, l'ensemble des individus de **Sexe** Masculin de **Type de bénéficiaire** Conjoint.

2.4.1.2 Les segments de modèle

Dans le cadre de ce mémoire, on définit le segment d'un modèle comme étant, l'ensemble des segments élémentaires sur lesquels ce modèle a la meilleure performance. Plus concrètement, considérons que l'on a deux variables explicatives (**Sexe** et **Type de bénéficiaires**). De plus, considérons deux modèles (modèle 1 et modèle 2) avec les performances suivantes :

N° des segments élémentaires	Sexe	Type de bénéficiaire	Meilleure prédiction
1	M	Adhérent	Modèle 1
2	M	Conjoint	Modèle 2
3	M	Enfant	Modèle 1
4	F	Adhérent	Modèle 1
5	F	Conjoint	Modèle 1
6	F	Enfant	Modèle 2

TABLE 2.7 – Exemples de segments de modèle

Ainsi, on en déduit les segments de modèle suivants :

- **Segment du modèle 1** : ce segment correspond à l'union des segments élémentaires pour lesquels le modèle 1 a la meilleure prédiction, soit l'union des segments 1, 3, 4 et 5.
- **Segment du modèle 2** : ce segment correspond à l'union des segments pour lesquels le modèle 2 a la meilleure prédiction, soit l'union des segments 2 et 6.

2.4.2 Définition du processus d'évaluation

- **Premier niveau :** Ce niveau permet d'évaluer l'éventuel biais qui pourrait exister dans l'apprentissage des données en évaluant sur les données d'apprentissage, le rapport $\frac{\sum observations}{\sum predictions}$ qu'on défini comme le ration O/P (Observations sur prédictions). Cela permet de s'assurer que la mutualisation est bien effective sur les données d'apprentissage, auquel cas, le rapport doit respecter la condition suivante : $|O/P - 1| < seuil$. En cas de non-adéquation aux données, le modèle est considéré comme biaisé et sa prédiction n'est pas acceptée.

  On note par ailleurs que le seuil de tolérance choisi dans notre cas est de 1 point par rapport à 100%.

- **Deuxième niveau** : Ce niveau permet d'évaluer la qualité de la prédiction sur la base de test, sur tous les segments élémentaires en mesurant la valeur $|O/P - 1|$ sur chaque segment élémentaire. Cette valeur permet de mesurer l'écart entre la prédiction et l'observation sur chaque segment élémentaire de la base de test.

  On note que pour des observations nulles, la valeur de $|O/P - 1|$ est 1. Ce résultat constitue un biais considéré comme acceptable. En effet, les observations nulles correspondent à l'absence de sinistre pour un segment donné. Dans ce cas, cela voudrait dire qu'on a une probabilité quasi nulle d'avoir un sinistre pour les individus ayant ces caractéristiques, ou que ces individus sont mal représentés, ou encore qu'il existe un biais de sélection dans le choix de notre base de test. Dans tous les cas, cette situation n'est pas avantageuse pour un modèle au détriment des autres pour cette évaluation.

- **Troisième niveau :** Le troisième niveau est un agrégat du niveau précédent qui permet d'évaluer la qualité globale du modèle sur les données de test. Cette évaluation globale est

obtenue par une moyenne sur l'ensemble des segments analysés sur le deuxième niveau.

Dans la suite, ce processus ainsi défini servira de base d'évaluation des différents modèles construits.

3 Mise en application

"Les statistiques sont vraies quant à la maladie et fausses quant au malade ; elles sont vraies quant aux populations et fausses quant à l'individu."
Léon Schwartzenberg 1923 - 2003

3.1 Présentation de la base de données

3.1.1 Base de données Initiale

La base de données utilisée dans le cadre de ce mémoire est une base de données issue de la fusion de deux bases de données anonymisées d'un portefeuille d'adhésions individuelles provenant d'un organisme complémentaire d'assurance maladie. Les deux bases de données fusionnées sont composées d'une part d'informations strictement statistiques sur les adhésions et d'autre part d'informations concernant les sinistres du portefeuille pour la période allant du 01/01/2018 au 31/12/2018.

Dans le cadre de ce mémoire, les postes de risque modélisés sont :

- Généraliste : Ce poste correspond aux dépenses engendrées pour la visite chez un médecin généraliste. Sa composition est très structurée parce que les dépenses sur ce poste sont pour la plupart, conventionnées.
- Optique : Ce poste correspond aux dépenses engendrées pour l'Optique (verre Optique, monture, lentille, etc.). Il fait partie des postes de dépenses concernés par la nouvelle reforme du 100% santé.

Pour ces deux postes, la base de données finale est composée des éléments suivants :

Risque	Effectif	Nombre de sinistre
Généraliste	8105	18580
Optique	8105	2317

La première étape de l'analyse de la base de données est la sélection des informations

pertinentes pour aboutir à une base de données saine pour la modélisation. Pour cela, les variables de la base de données retenues sont les suivantes :

Catégorie	Variables	Commentaires
Base des adhésions		
CONTRATS	Clé d'adhésion	Numéro unique de l'adhésion permettant l'identification du contrat
	Date d'adhésion	Date de début du contrat au format (Jour/Mois/Année)
	Date de radiation	Date de fin de l'adhésion le cas échéant au format (Jour/Mois/Année)
BÉNÉFICIAIRES	Type de bénéficiaire	Variable à trois valeurs : Adhérent ; Conjoint ; Enfant.
	Sexe	Variable à deux valeurs : F := Féminin ; M := Masculin
	Date de naissance	Date de naissance du bénéficiaire au format (Jour/Mois/Année)
	Code postal	Code postal du bénéficiaire
	Âge	Age au 31/12/2018 du bénéficiaire
Base des prestations		
PRESTATIONS	Type d'acte	Description du tyoe d'acte
	Code du type d'acte	Code unique du type d'acte
	Poste du risque	Poste de dépense en santé représentant un ensemble de type d'acte
	Dépense engagée	Coût réel de la dépense engendrée par le sinistre
	Remboursement sécurité social	Montant remboursé par la sécurité sociale
	Remboursement organisme complémentaire	Montant remboursé par l'organisme assureur
	Reste à charge	Montant restant à la charge de l'assuré

TABLE 3.1 – Détail de la base de données

Ces informations ainsi sélectionnées permettent par la suite d'effectuer des retraitements et la création de variables.

Dans la suite de ce mémoire, les variables d'intérêts seront les dépenses engagées par poste de risque.

3.1.2 Traitement des données

3.1.2.1 Retraitement et création de variables

Les informations précédemment sélectionnées ont permis la création de différentes variables qui sont :

- **L'exposition** : cette variable permet de mesurer l'exposition au portefeuille sur l'année 2018 de chaque assuré. Elle vaut 1 lorsque l'assuré est présent toute l'année sur le portefeuille et (nombre de jours de présence)/(nombre de jours dans l'année) si l'assuré est entré ou sorti du portefeuille au cours de l'année.

  La création de cette variable est importante et indispensable pour une meilleure estimation de la fréquence d'occurrence des sinistres. Afin de créer cette variable, les variables, Date d'adhésion, Date de radiation sont utilisées.

- **Départements et Régions** : la variable Code postal est retraitée afin de créer les variables Département et Régions pour avoir une vision plus agrégée de la répartition géographique des assurées.

3.1.2.2 Données manquantes et aberrantes

L'analyse exploratoire des données a permis de s'assurer de l'absence de valeurs manquantes dans la base de données. En outre, l'analyse des différentes variables explicatives ainsi que celle des différentes variables d'intérêt ont permis de s'assurer aussi de l'absence de données aberrantes.

Ces observations permettent de poursuivre plus sereinement, l'analyse de la base de données.

3.1.3 Analyse descriptive de la base de données

3.1.3.1 Répartition des assurés par sexe

Le graphique ci-dessous présente le nombre d'adhérents (en termes d'exposition) en fonction de la variable Sexe.

L'analyse de ce graphique permet d'observer que la base de données étudiée présente une répartition quasiment équidispersée entre les hommes et les femmes. Cette observation permet de mesurer correctement l'effet du sexe sur les différentes variables d'intérêt.

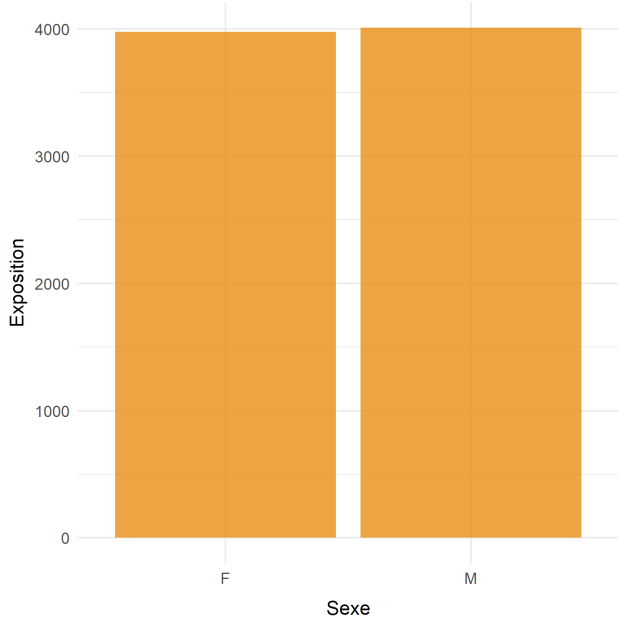


FIGURE 3.1 – Exposition dans le portefeuille en fonction de la variable sexe

Par ailleurs, on note que d'après la directive européenne « ***gender directive*** » il n'est pas possible de tarifer selon le genre. Cependant, cette variable peut servir de mesure de risque notamment pour la tarification collective.

3.1.3.2 Répartition des assurés par type de bénéficiaire

Le graphique ci-dessous présente le nombre d'adhérents (en termes d'exposition) en fonction du type d'adhérents.

Ce graphique met en exergue une faible représentation des conjoints ainsi que des enfants dans la base par rapport à celle des adhérents.

Ce portefeuille étant constitué d'adhésions individuelles, cette représentation parait cohérente. En effet, l'analyse de la base de données montre que les **adhérents** sont des personnes avec un âge moyen de 65 ans ce qui laisse aussi supposer qu'ils sont, pour une grande partie, à la retraite. Cette forte concentration des séniors dans la population d'adhérents diminue la probabilité de voir apparaître des bénéficiaires enfants dans le portefeuille.

Par ailleurs, on note aussi que , l'analyse de la base de données montre que les proportions de bénéficiaires de type **conjoint** et de bénéficiaires de type **enfant** sont respectivement de 24% et 13%. Ces éléments laissent supposer que la plupart des conjoints des adhérents du portefeuille préfèrent souscrire à titre individuel par le biais d'une adhésion collective lorsqu'ils sont encore en activité ce qui est en général, est plus avantageux.

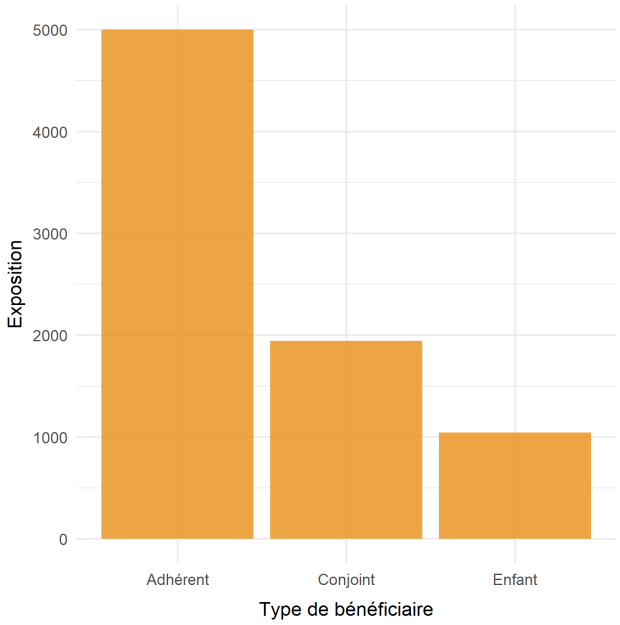


FIGURE 3.2 – Exposition dans le portefeuille en fonction de la variable type de bénéficiaire

3.1.3.3 Pyramide des âges

Le graphe ci-dessous représentant le nombre de bénéficiaires par âge, met en évidence un portefeuille majoritairement composé de séniors (âge moyen à 56 ans). Cette répartition par âge s'explique par la nature du portefeuille étudié. Ce portefeuille est composé en majorité de souscription à adhésion individuelle comme indiqué précédemment.

La sous-représentation des âges inférieurs à 55 ans ainsi que d'autres raisons opérationnelles nous conduisent à considérer des classes d'âge construites par connaissance métier.

Par ailleurs, la pyramide des âges ci-dessous permet d'observer qu'on a quasiment autant d'hommes que de femmes pour chaque classe d'âge.

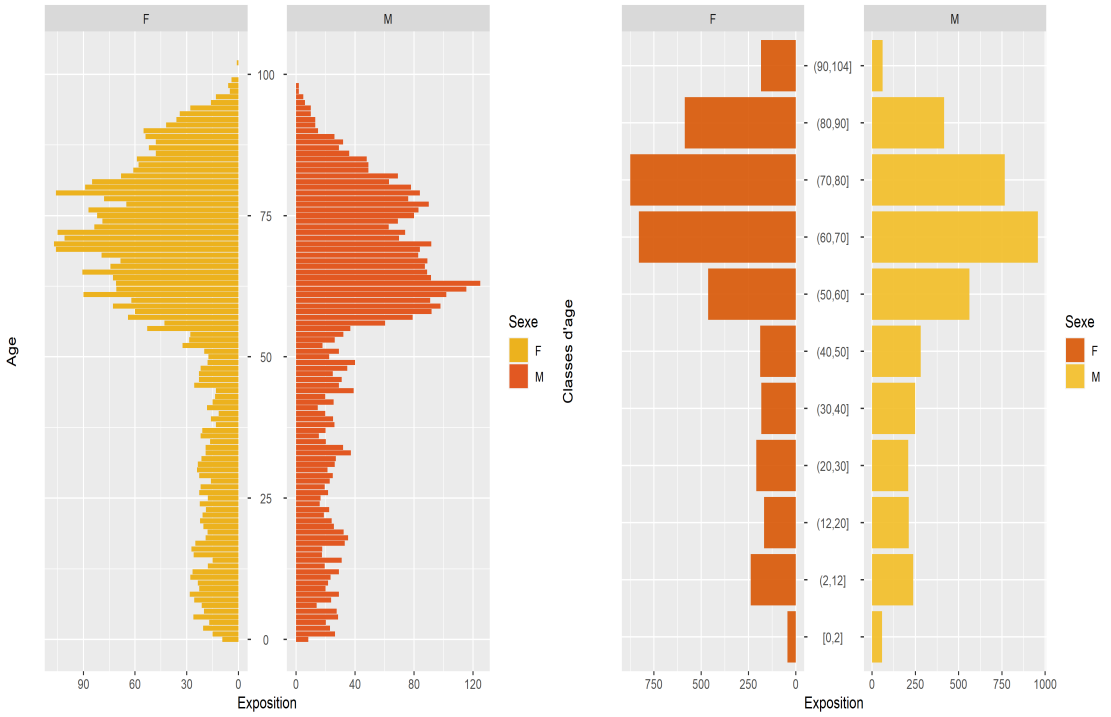


FIGURE 3.3 – Pyramide des âges du portefeuille étudié

3.1.3.4 Répartition géographique des assurés

La figure 3.4 ci-dessous représente la répartition des assurés par département en fonction des expositions en 2018. La figure 3.5 quant à elle représente la répartition des assurés par région en fonction des expositions en 2018.

Ces deux figures mettent en exergue une forte concentration de notre portefeuille dans l'ouest de la France et plus particulièrement dans la région ***Pays de la loire***. Cette forte concentration introduit un biais qui ne permet pas d'avoir une vision globale du risque sur toute l'étendue du territoire Français métropolitain.

En effet, l'on pourrait observer par exemple, une faible sinistralité voire même une absence de sinistralité sur certaines régions sous représentées. Ces effets qui peuvent être dus à la sous représentativité de ces régions peuvent constituer un biais important.

Pour atténuer l'effet de ce biais, la base open-damir est utilisée afin d'avoir une répartition plus représentative du risque sur le territoire français métropolitain sur la base des coûts moyens par habitants.

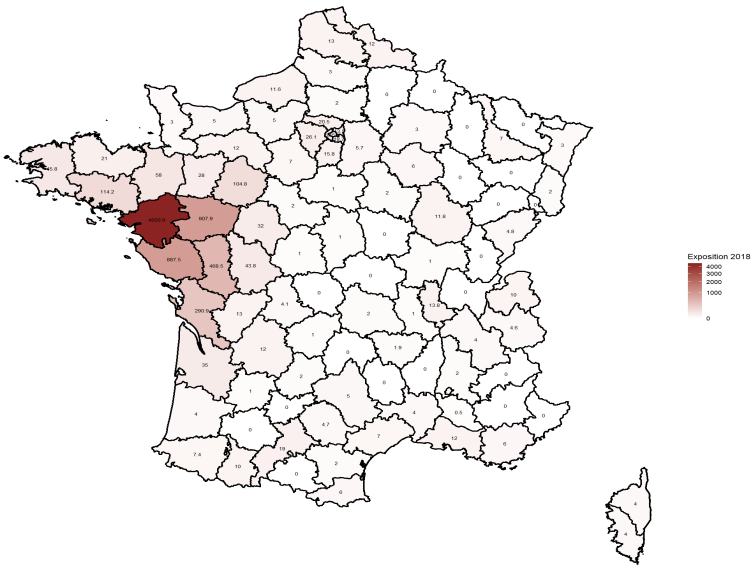


FIGURE 3.4 – Démographie du portefeuille par département

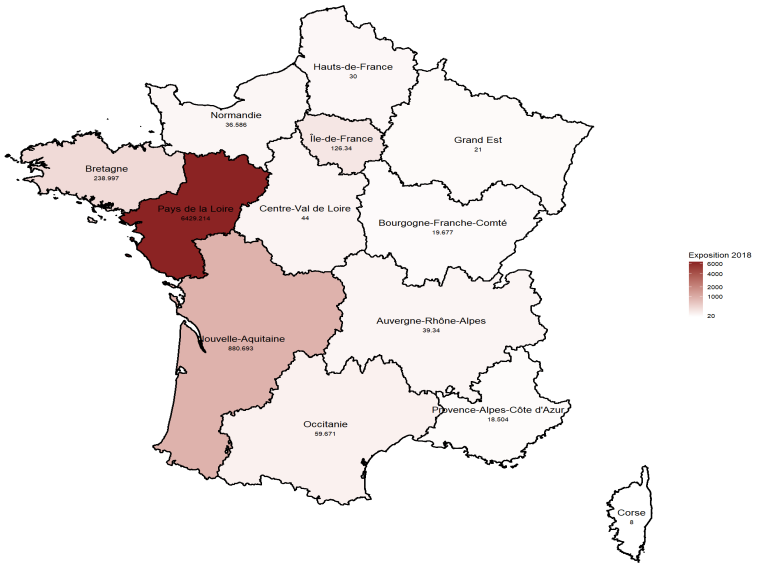


FIGURE 3.5 – Démographie du portefeuille par région

3.1.4 Apport de la base Open Damir

3.1.4.1 Présentation de la base Damir

D'après la plateforme ouverte des données publiques Française data.gouv.fr, La base de données Open DAMIR est une base de données produite par la Caisse nationale de l'assurance maladie (CNAM). Cette base de données, qui est une extraction du Système National Inter-Régimes d'Assurance Maladie (Sniiram), est une base complète sur les dépenses d'assurance maladie tous régimes confondus.

Cette base de données concerne l'ensemble des prestations prises en charge par l'Assurance Maladie obligatoire y compris les prestations hospitalières facturées directement à l'Assurance Maladie pour l'ensemble des régimes. Les dépenses sont détaillées selon six axes d'analyse (période, prestation, organisme de prise en charge, bénéficiaire des soins, professionnel de santé exécutant, professionnel de santé prescripteur) et sept indicateurs de montant (total de la dépense, base de remboursement, remboursé, dépassement) et de volume (quantité, dénombrement, coefficient).

Afin de préserver l'anonymat des professionnels de santé et des bénéficiaires des soins, les axes géographiques sont limités :

- à 9 zones géographiques (zones d'études et d'aménagement des territoires) qui sont des regroupements de régions administratives de 2009 à 2014,
- à 13 zones géographiques (proches des grandes régions administratives nouvellement créées) à partir de 2015.

3.1.4.2 Regroupement des régions par le biais de la base open-Damir

Comme évoqué en 3.1.3.4, la base de données initiale ne permet pas d'avoir une vision claire des différents risques sur tout le territoire métropolitain de la France. Pour atténuer cet effet, l'approche utilisée dans ce mémoire consiste à créer des zoniers pour les différents risques sur la base de la dépense moyenne par habitant.

Pour ce faire, les opérations suivantes sont effectuées :

- La base de données Open DAMIR est triée en faisant correspondre les codes d'actes disponibles pour chaque poste de la base initiale aux codes d'actes disponibles dans la base Open Damir.
- Pour chaque région, la somme de la dépense par poste est divisée par la population totale de la région en 2018.
- Les différentes régions sont regroupées en subdivisant, pour chaque poste, l'intervalle des valeurs de la dépense par habitant des différentes régions en 5 (cinq) intervalles de même longueur.

Ces intervalles ainsi obtenus permettent d'avoir des groupes de régions homogènes. Ci-dessous le tableau des résultats obtenus pour chaque région.

Régions	Code région	Dépense par habitant généraliste	intervalle Dépense par habitant généraliste	Groupes de régions généraliste	Dépense par habitant Optique	intervalle Dépense par habitant Optique	Groupes de régions Optique
Auvergne-Rhône-Alpes	84	134.70	[-Inf,136.2)	84-53-24-52	97.01	[97.01, Inf]	84-24
Bourgogne-Franche-Comté	27	138.70	[136.2,151.6)	27-28	94.60	[94.60,97.01)	27-44-52
Bretagne	53	136.22	[-Inf,136.2)	84-53-24-52	92.27	[89.02,93.62)	53-32-75
Centre-Val de Loire	24	131.51	[-Inf,136.2)	84-53-24-52	97.45	[97.01, Inf]	84-24
Corse	93	170.10	[157.2,171.2)	93-44-93	85.69	[-Inf,89.02)	93-76-94
Grand Est	44	171.21	[157.2,171.2)	93-44-93	95.81	[94.60,97.01)	27-44-52
Hauts-de-France	32	177.61	[171.2, Inf]	32	89.02	[89.02,93.62)	53-32-75
Normandie	28	138.96	[136.2,151.6)	27-28	93.62	[93.62,94.60)	28-11
Nouvelle-Aquitaine	75	154.31	[151.6,157.2)	75-76-11	90.48	[89.02,93.62)	53-32-75
Occitanie	76	157.20	[151.6,157.2)	75-76-11	86.44	[-Inf,89.02)	93-76-94
Pays de la Loire	52	128.22	[-Inf,136.2)	84-53-24-52	96.34	[94.60,97.01)	27-44-52
Provence-Alpes-Côte d'Azur	93	170.10	[157.2,171.2)	93-44-93	85.69	[-Inf,89.02)	93-76-94
Île-de-France	11	151.63	[151.6,157.2)	75-76-11	93.77	[93.62,94.60)	28-11

TABLE 3.2 – Regroupement des région par rapport à la dépense moyenne par habitant pour les postes Généraliste et Optique

Les résultats obtenus permettent de construire les zoniers ci-dessous, pour les postes Généraliste et Optique.

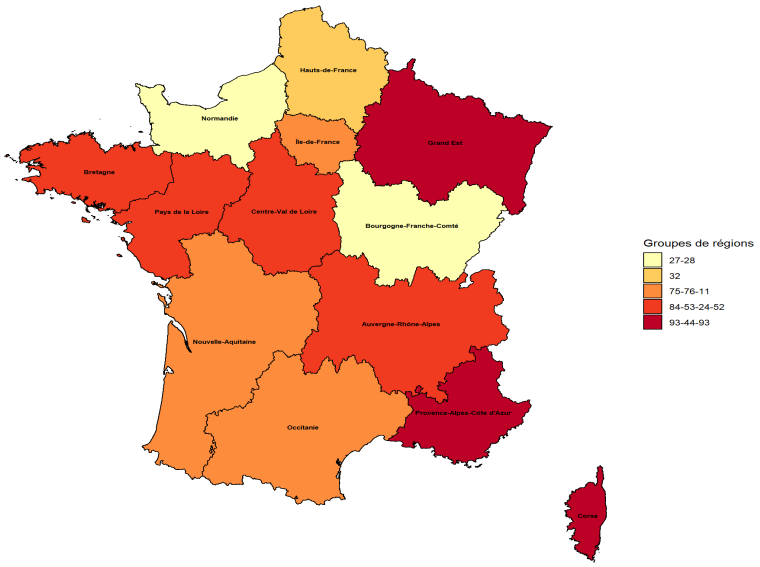


FIGURE 3.6 – Regroupement des régions pour le poste Généraliste, obtenu avec la base Open-damir sur l'année 2018

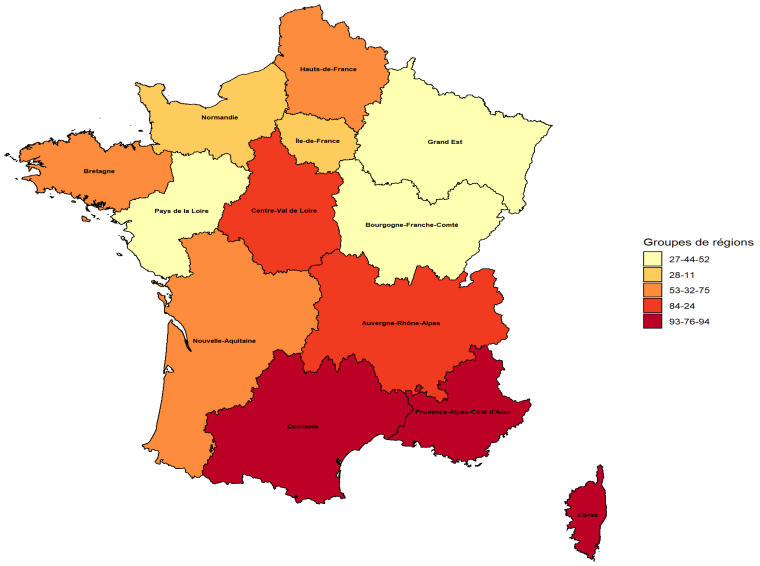
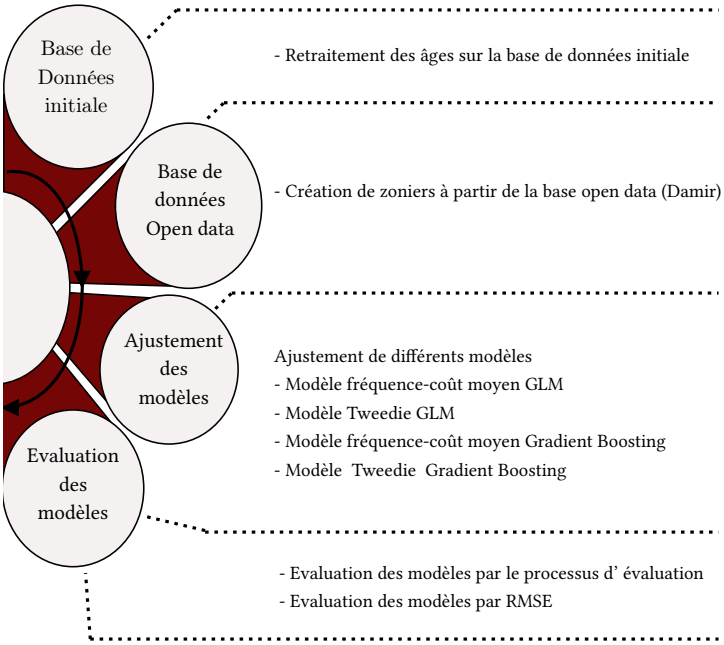


FIGURE 3.7 – Regroupement des régions pour le poste Optique, obtenu avec la base Open-damir sur l'année 2018

3.1.5 Modélisation retenue

L'analyse préliminaire de la base de données étant effectuée, la modélisation retenue se résume à travers les différentes étapes suivantes :



3.2 Description des différents postes

3.2.1 Le poste Généraliste

3.2.1.1 La dépense réelle pour le poste Généraliste

Le graphe ci-dessous représente la distribution de la dépense réelle pour le poste Généraliste.

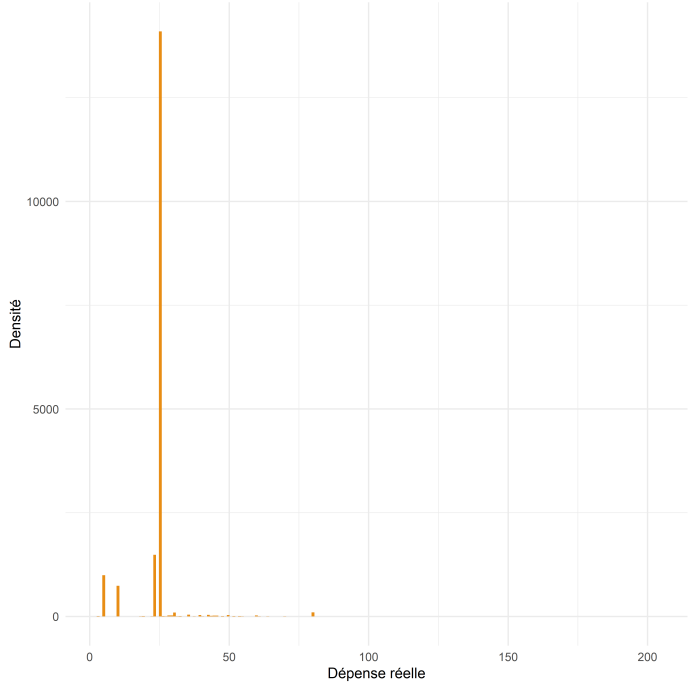


FIGURE 3.8 – Distribution de la dépense réelle pour le poste Généraliste

L'analyse de ce graphique permet d'observer une structure quasi-discrète de la dépense réelle pour le poste Généraliste. En effet, on observe une concentration des valeurs de la dépense réelle sur les observations 25 €, 23 €, 10 €, 5 €. De plus, l'analyse du tableau 3.3 ci-dessous permet d'observer que cette concentration représente 96 % des observations.

Dépenses réelles	25€	23€	5€	10€	Autres
Pourcentages	78,75%	8,29%	5,54%	4,13%	2,76%
Pourcentage cumulé	78,75%	87,04%	92,58%	96,7%	100%
Sinistralité totale	352 200€	34 109€	7 380€	4 955€	26 789.3€
Pourcentage de la sinistralité totale	82,79 %	8,017%	1,73%	1,165%	6,30%
Pourcentage Cumulé de la sinistralité totale	82,79%	90,80%	92,54%	93,70%	100%

TABLE 3.3 – Répartition de la dépense réelle pour le poste Généraliste

Au vu de ces observations, le choix de la modélisation de la dépense réelle pour le poste Généraliste se porte sur une modélisation discrète par le biais d'une régression multinomiale.

En considérant Ω comme étant l'ensemble des valeurs prises par la dépense réelle, les observations autres ($\Omega \setminus \{$ 25€, 23€, 10€, 5€ $\}$) de la dépense réelle sont considérées comme étant une seule modalité représentée par leur moyenne (45,4€).

Cette approximation parait raisonnable, car l'ensemble de ces valeurs représente $2,76\%$ des observations et $6,30\%$ de la sinistralité totale.

Dans ces conditions, en notant Y la variable aléatoire représentant la dépense réelle pour le poste Généraliste et X la matrice colonne des variables explicatives, alors :

$$\begin{aligned}\mathbb{E}[Y/X] = 25 \cdot \mathbb{P}((Y = 25|X) + 23 \cdot \mathbb{P}((Y = 25|X) + 10 \cdot \mathbb{P}((Y = 25|X) + \\ 5 \cdot \mathbb{P}((Y = 25|X) + 45,4 \cdot \mathbb{P}((Y \in \Omega \backslash \{25, 23, 10, 5\}|X)\end{aligned} \tag{3.1}$$

Ici, les différentes probabilités sont obtenues par le biais d'une régression multinomiale.

3.2.1.2 La dépense réelle en fonction du sexe

L'analyse de la dépense moyenne sur le poste Généraliste permet d'observer un coût moyen empirique plutôt stable par rapport au sexe et une fréquence des dépenses plus importante chez les femmes que hommes. Le sexe a donc un effet plus important sur la fréquence des sinistres que sur la dépense moyenne.

Par ailleurs, on observe un coût total empirique plus important chez les femmes et essentiellement porté par la fréquence empirique.

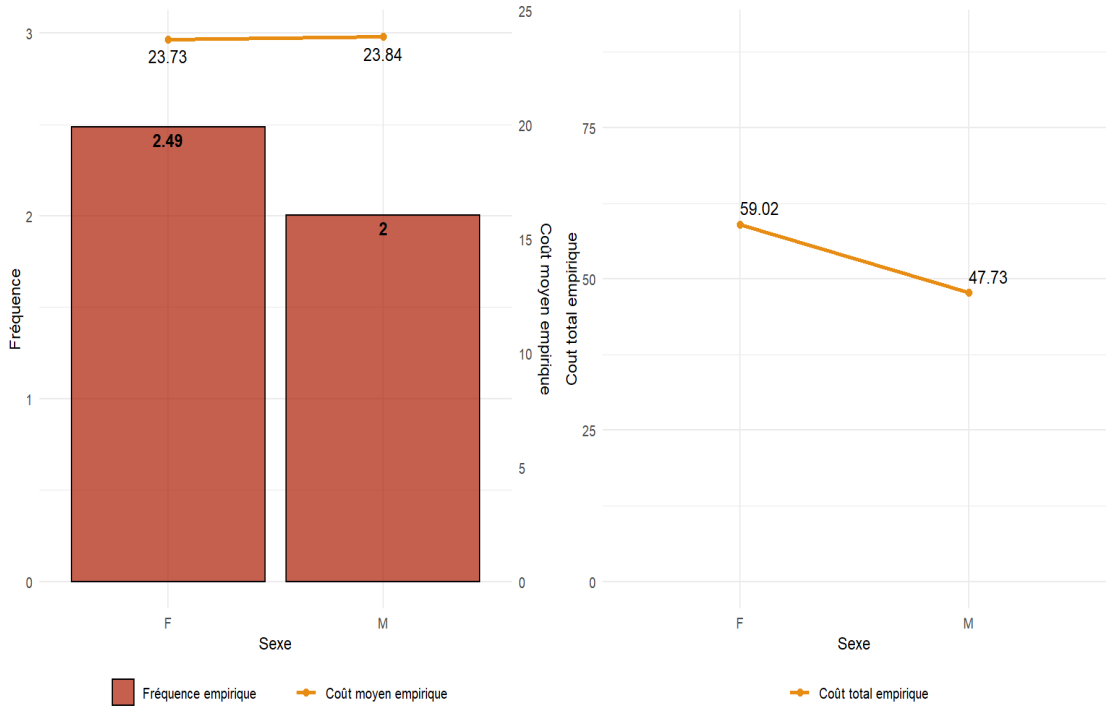


FIGURE 3.9 – Analyse de la dépense réelle moyenne pour le poste Généraliste en fonction de la variable Sexe

3.2.1.3 La dépense réelle en fonction du type de bénéficiaire

L'analyse graphique de l'effet du type de bénéficiaire sur la dépense moyenne montre une dépense moyenne légèrement plus faible chez les adhérents que chez les conjoints et une dépense moyenne nettement plus faible chez les enfants que chez les adhérents.

Par ailleurs, on observe une fréquence de sinistres légèrement plus faible chez les adhérents que chez les conjoints qui ont eux mêmes une fréquence de sinistre nettement plus faible que celle des enfants.

Le caractère discriminant de la variable type de bénéficiaire est très bien marqué sur la dépense moyenne ainsi que sur la fréquence des sinistres.

On observe un prix empirique qui est porté aussi bien par la dépense moyenne que par la fréquence des sinistres avec une plus forte valeur chez les conjoints et une plus faible valeur chez les adhérents.

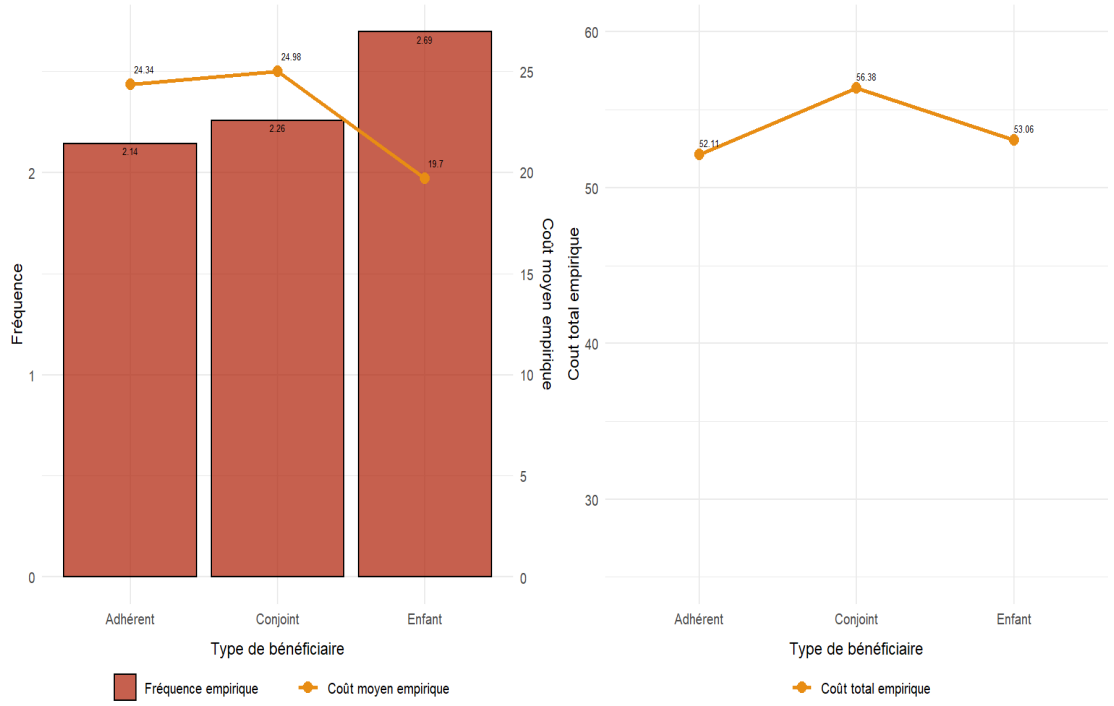


FIGURE 3.10 – Analyse de la dépense réelle moyenne pour le poste Généraliste en fonction de la variable Type de bénéficiaire

3.2.1.4 La dépense réelle en fonction de l'âge

L'analyse des graphiques ci-dessous permet d'observer une dépense moyenne très faible et une fréquence des sinistres très forte chez les enfants de la tranche d'âge [0,2]. Ces observations sont dues à une présence importante d'acte de majoration pour les enfants de 0 à 6 ans (MEG).

L'effet de l'âge sur la dépense moyenne et la fréquence des sinistres n'étant pas linéaire, la discrétisation de l'âge permet d'avoir une meilleure représentation de ces effets. Les effets combinés sur la fréquence des sinistres et sur la dépense moyenne permettent d'avoir un prix empirique avec une tendance décroissante jusqu'à 50 ans puis croissante jusqu'à 80 ans et enfin décroissante jusqu'à 104 ans.

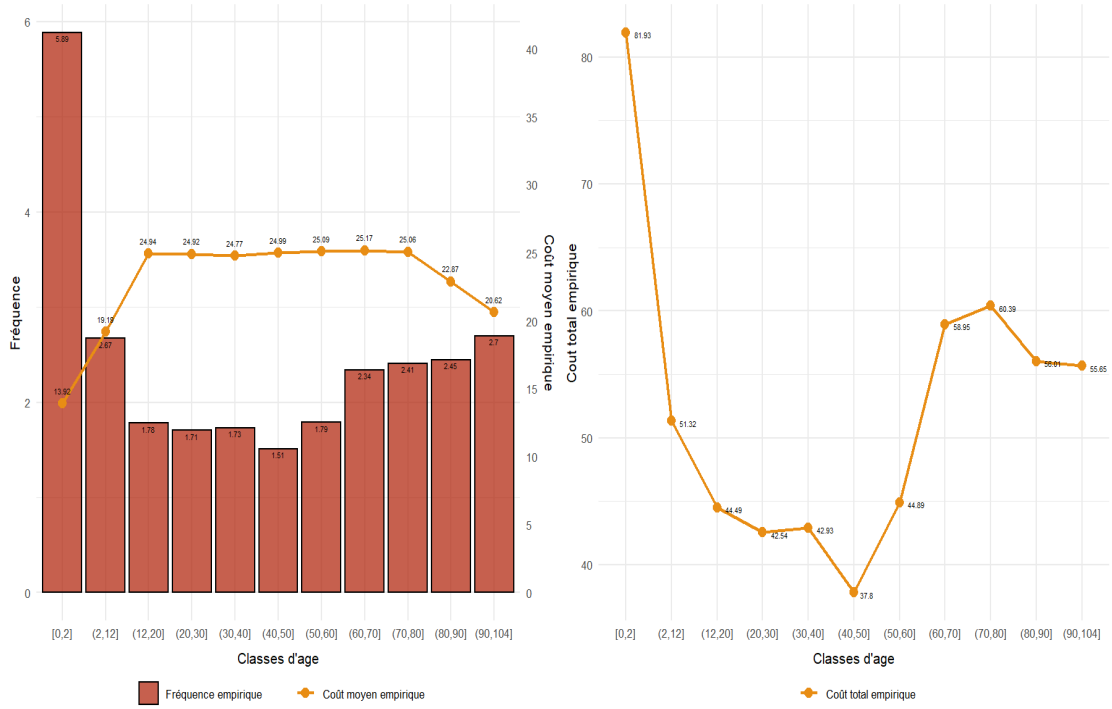


FIGURE 3.11 – Analyse de la dépense réelle moyenne pour le poste Généraliste en fonction de la variable Classe d'âges

3.2.1.5 La dépense réelle en fonction de la région

Contrairement à la représentation par région, la représentation par groupe de régions permet de corriger la sensibilité aux valeurs extrêmes causée par la faible représentation de certaines régions dans le portefeuille.

Les graphiques ci-dessous en 3.12 permettent d'observer que le tarif moyen oscille entre 23 et 25 euros pour ces groupes ce qui se rapproche du tarif de convention secteur 1. On observe en revanche des effets distincts en fonction des groupes de régions sur la fréquence. L'effet des groupes de région sur le coût total est donc principalement guidé par la fréquence des consultations.

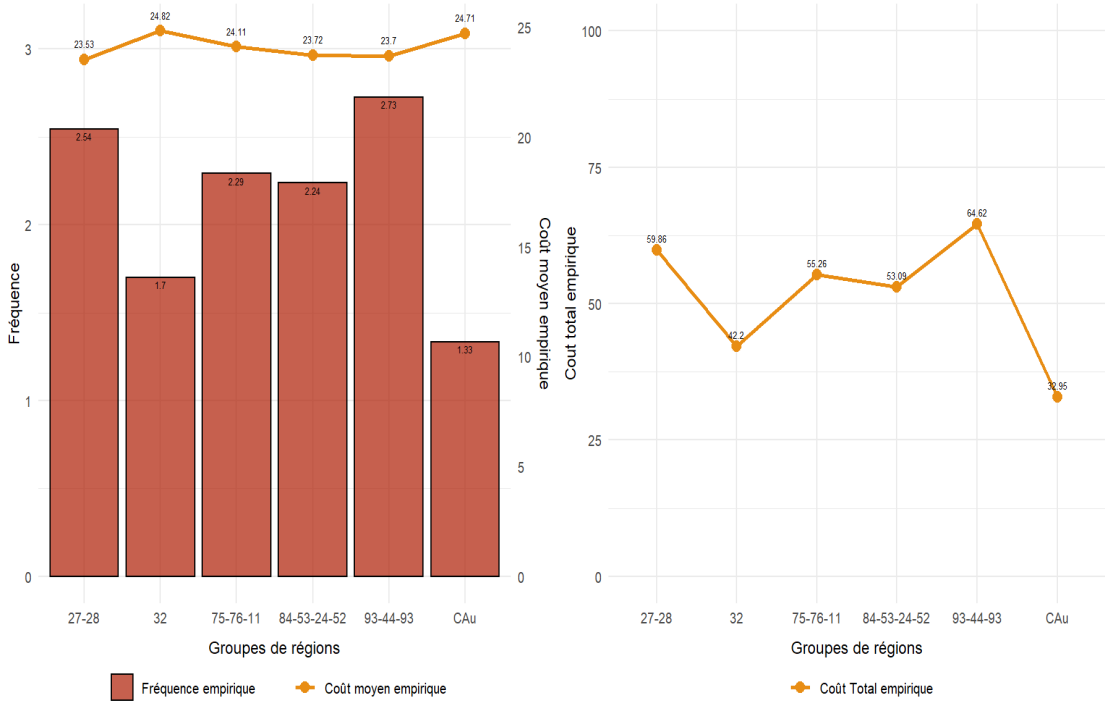


FIGURE 3.12 – Analyse de la dépense réelle moyenne pour le poste Généraliste en fonction de la variable de région

3.2.1.6 Corrélation entre les variables explicatives

Dans cette partie, la méthode utilisée pour mesurer la corrélation entre les différentes variables explicatives est la méthode « V de Cramer ». Le choix de cette méthode s'explique par le fait que toutes les variables explicatives utilisée dans ce mémoire sont des variables catégorielles.

La méthode « V de Cramer » est une mesure d'association entre deux variables catégorielles introduite par Harald Cramér [1], basée sur la statistique χ^2 de Pearson et qui prend ses valeurs dans $[0, 1]$.

Contrairement au χ^2 de Pearson, cette statistique reste stable si la taille de l'échantillon augmente dans les mêmes proportions pour chaque modalité.

En considérant un échantillon simultané de taille n x 2 de deux variables aléatoires X et Y avec r et k le nombre de modalités respectives des variables, La statistique « V de cramer » est calculée de la manière suivante :

$$V = \sqrt{\frac{\chi^2/n}{min(k-1, r-1)}} \tag{3.2}$$

1. CRAMÉR 1946.

Où χ^2 est la statistique définie par $\chi^2 = \sum_{i,j} \frac{(n_{i,j} - \frac{n_i, n_j}{n})^2}{\frac{n_i, n_j}{n}}$, et $n_{i,j}$ le nombre d'observations du couple (X_i, Y_i)

En appliquant cette méthode à la base de données du poste Généralise, on obtient le graphe ci-dessous :

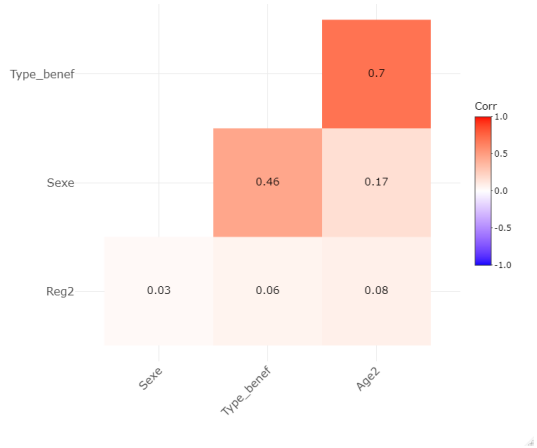


FIGURE 3.13 – Corrélations entre les variables explicatives pour le poste Généraliste

L'analyse de cette matrice de corrélation permet d'observer d'une part, une corrélation élevée entre les variables ***Type de bénéficiaire*** (Type_benef) et ***Classe d'âge*** (Age2) et entre les variables ***Sexe*** et ***Type de bénéficiaire***. D'autre part, les autres corrélations observées restent très faibles.

Ces résultats paraissent cohérents avec la structure des données. En effet, en regardant de plus près les données, on observe que les enfants des adhérents ont des âges en dessous de 27 ans. Les adhérents ont des âges au-dessus de 19 ans et les conjoints ont des âge au-dessus de 19 ans.

3.2.2 Le poste Optique

3.2.2.1 La dépense réelle pour le poste Optique

Le graphique ci-dessous présente la distribution de la dépense réelle pour le poste Optique. Son analyse permet d'observer une structure continue de la dépense réelle . Cette distribution est de plus, à valeurs positives et présente une forme asymétrique.

L'ensemble de ces informations permettent de faire l'hypothèse de la loi Gamma, pour la distribution conditionnelle de la dépense réelle, ce qui semble une hypothèse raisonnable.

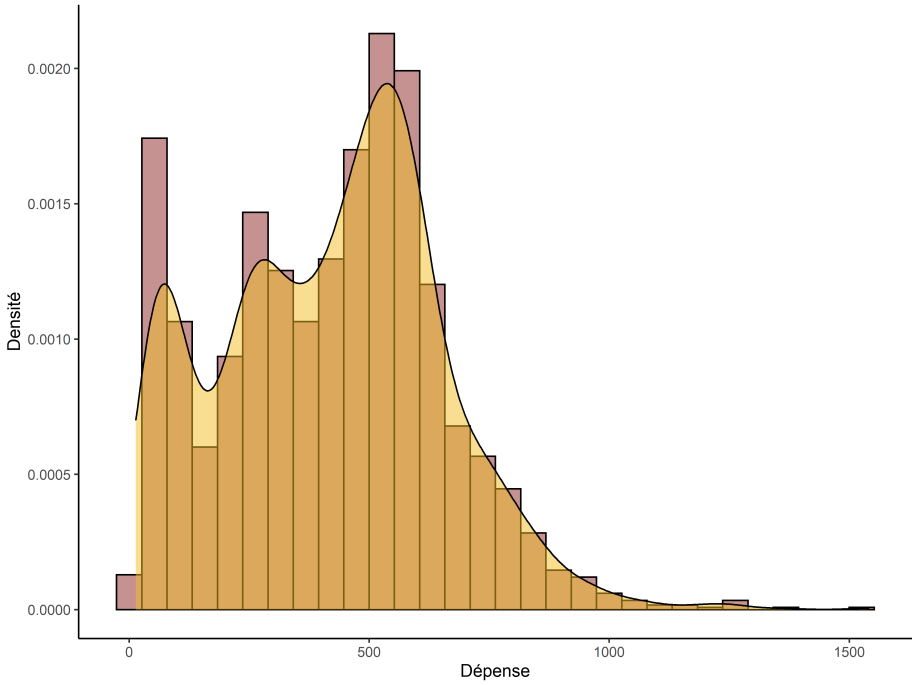


FIGURE 3.14 – Distribution de la dépense réelle pour le poste Optique

3.2.2.2 Dépense réelle en fonction de la variable sexe

L'analyse graphique de l'effet de la variable sexe sur la dépense moyenne pour le poste Optique permet d'observer une dépense moyenne plus faible chez les femmes que chez les hommes tandis que la fréquence des sinistres est plus importante chez les femmes que chez les hommes.

Ces effets combinés permettent d'aboutir à un prix empirique dicté par l'effet de la fréquence et nettement plus élevé chez les femmes que chez les hommes.

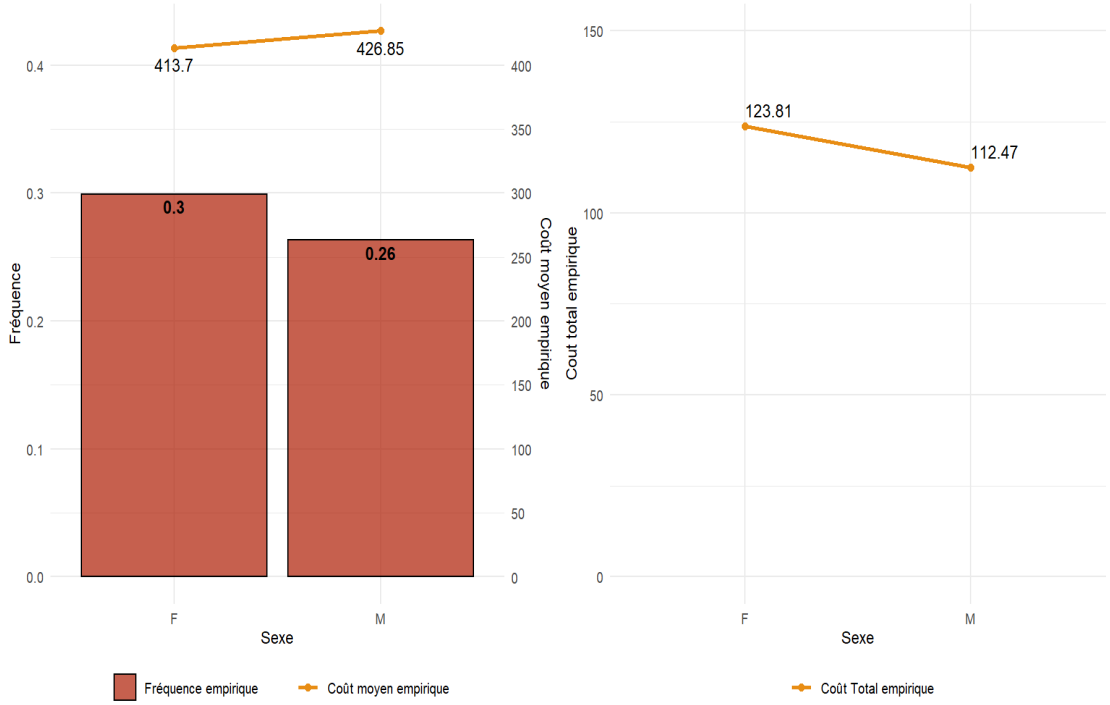


FIGURE 3.15 – Analyse de la dépense réelle moyenne pour le poste Optique en fonction de la variable Sexe

3.2.2.3 Dépense réelle en fonction du type de bénéficiaire

L'analyse graphique de l'effet du type de bénéficiaire sur la dépense moyenne montre une dépense moyenne nettement plus faible chez enfants que chez les adhérents qui ont eux même une dépense moyenne plus faible que celle des conjoints. Le caractère discriminant de la variable type d'adhérent est très marqué sur la dépense moyenne.

Par ailleurs on observe une fréquence de sinistres plus faible chez les enfants que chez les conjoints qui ont eux même une fréquence des sinistres plus faible de celle des adhérents. De même que pour la dépense moyenne ; Le caractère discriminant de la variable type d'adhérent est très bien marqué sur la la fréquence des sinistres.

On observe un coût total moyen empirique essentiellement porté par la dépense moyenne avec une plus forte moyenne chez les conjoints et une très faible moyenne chez les enfants.

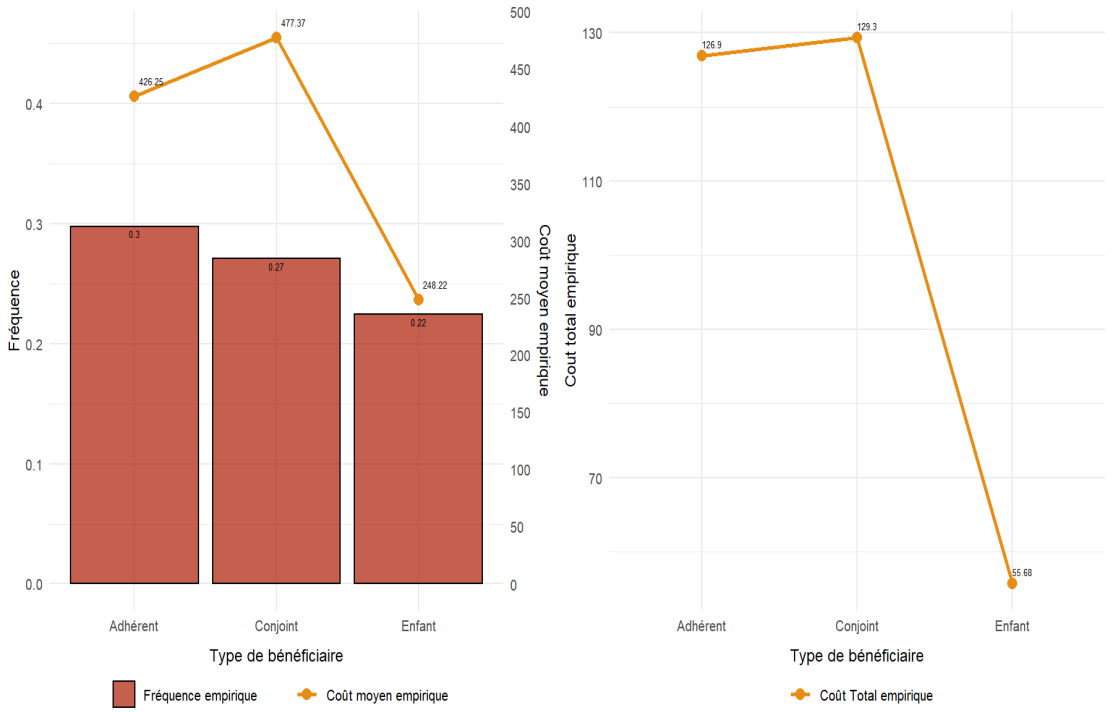


FIGURE 3.16 – Analyse de la dépense réelle moyenne pour le poste Optique en fonction de la variable Type de bénéficiaire

3.2.2.4 Dépense réelle en fonction de l'âge

Les graphiques, ci-dessous permettent d'observer l'effet de l'âge sur la dépense moyenne ainsi que sur la fréquence des sinistres. Ces effets sont chacune non-linéaire. La discrétion de l'âge parait donc pertinente pour prendre en compte le caractère discriminant de l'âge sur la dépense moyenne ainsi que sur la fréquence.

Par ailleurs, on observe une très faible fréquence de sinistre pour les individus dans la tranche d'âges [0,2]. Cette observation est tout à fait cohérente, car ces individus sont très rarement sujets au port de lunettes et de lentilles correctrices.

Au final, on observe un effet distinct des tranches d'âge sur le prix empirique essentiellement porté par la fréquence des sinistres.

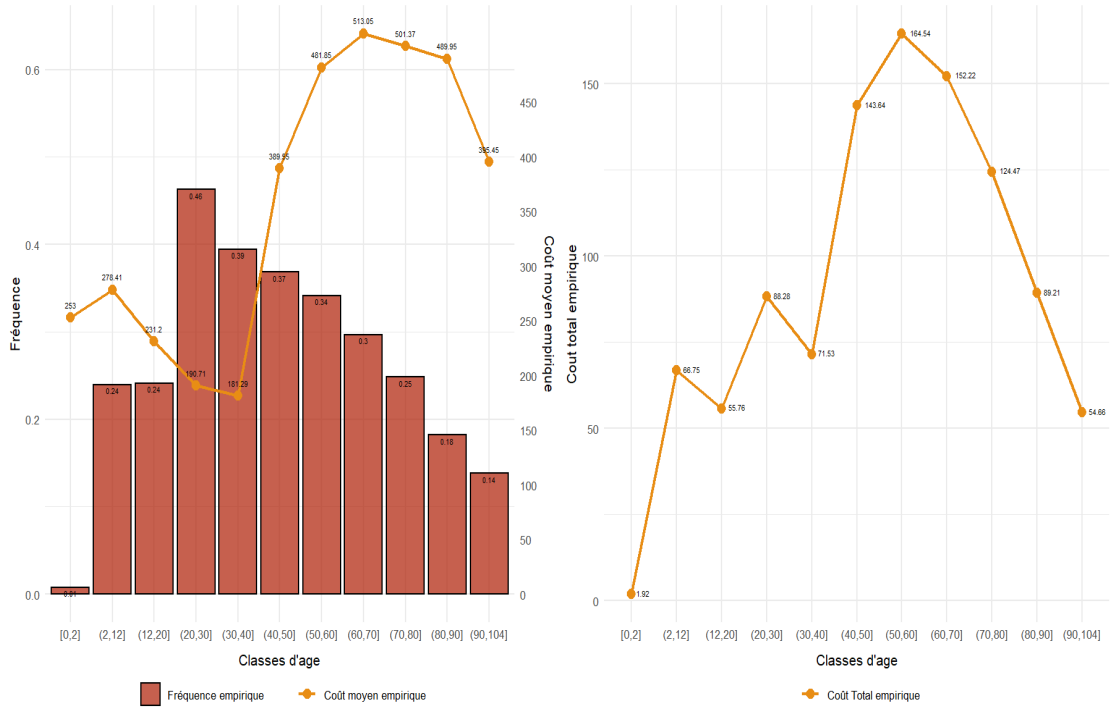


FIGURE 3.17 – Analyse de la dépense réelle moyenne pour le poste Optique en fonction de la variable Age

3.2.2.5 Dépense réelle en fonction de la région

De même que pour le poste Généraliste, la représentation par groupe de région permet de corriger la sensibilité aux valeurs extrêmes causée par la faible représentation de certaines régions dans le portefeuille.

On observe des effets distincts en fonction des groupes de région sur la fréquence et sur la dépense moyenne avec un caractère plus marqué sur la fréquence. L'effet des groupes de région sur le prix empirique est principalement guidé par l'effet sur la fréquence.

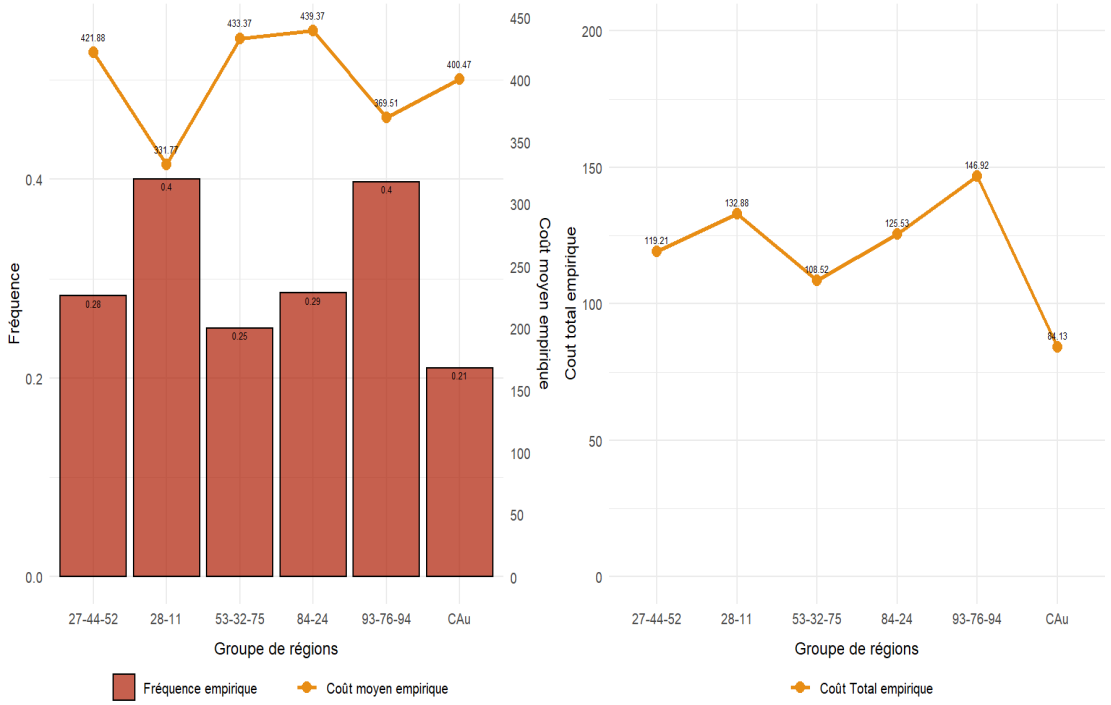


FIGURE 3.18 – Analyse de la dépense réelle moyenne pour le poste Optique en fonction de la variable Groupes de régions

3.2.2.6 Corrélation entre les variables explicatives

De même que dans la partie 3.2.1.6, les résultats de cette partie sont obtenus par la méthode « V de Cramer ».

Le graphe ci-dessous présente les corrélations entre les différentes variables explicatives pour le poste Optique.

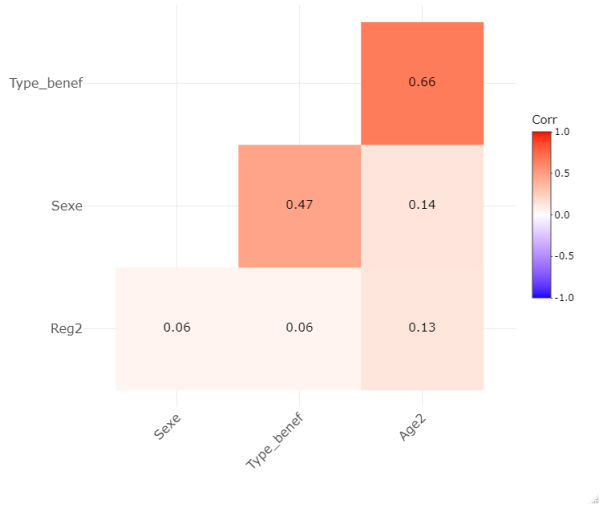


FIGURE 3.19 – Corrélations entre les variables explicatives pour le poste Optique

L'analyse de cette matrice de corrélation permet d'observer dans un premier temps, une corrélation moyenne entre les variables ***Type de bénéficiaire*** (Type_benef) et ***Classe d'âge*** (Age2) et entre les variables ***Sexe*** et ***Type de bénéficiaire***. Dans un deuxième temps, les autres corrélations observées restent très faibles.

3.3 Approches de modélisation

3.3.1 Constructions des modèles

Dans ce mémoire, les modèles sont construits selon deux axes :

- **Le type de modèle (GLM vs XGBoost)** :
  - **Les modèles issus du modèle linéaire généralisé (GLM)** qui permettent d'établir une relation linéaire entre une transformée de l'espérance de la variable d'intérêt et les variables explicatives.
  - **Les modèles issus du eXtrem Gradient Boosting (XGBoost)** qui permettent d'établir une relation fonctionnelle (non nécessairement linéaire) entre l'espérance de la variable d'intérêt et les variables explicatives. Pour rappel, cette méthode utilise l'algorithme de descente de gradient pour minimiser une fonction de perte. Par ailleurs, le choix de cette fonction dépend de la nature de données à modéliser.
- **L'approche de détermination de la prime pure (Coût total vs Fréquence coût moyen)** :
  - **Coût total** : pour chaque type de modèle, on utilise cette approche pour obtenir une prédiction directe de la prime pure.

– **Fréquence Coût-moyen** : pour chacun des modèles, on utilise cette approche pour obtenir d'une part la fréquence et d'autre part le coût moyen. La prime pure est par la suite obtenue en faisant le produit de ces deux quantités.

En résumé, selon ces deux axes, les modèles suivants sont construits pour chaque poste :

	Coût total	Fréquence - Coût moyen
GLM	GLM Tweedie	GLM FCM
XGBoost	• XGBoost Tweedie *(Fonction perte spécifique aux données assurantielles)* • XGBoost MSE *(Fonction perte par défaut pour les problèmes de régression (MSE))*	XGBoost FCM

Table 3.4 – Résumé des modèles construits

Ces différents modèles sont construits sous **R** et évalués dans la suite de ce mémoire.

3.3.2 Description des modèles à construire

3.3.3 Le modèle GLM FCM

Le modèle ***GLM FCM*** est l'un des modèles les plus couramment utilisés pour modéliser la prime pure. Ce modèle a l'avantage d'être facile à interpréter et à mettre en œuvre. Il permet d'avoir une vision détaillée aussi bien sur la fréquence des sinistres que sur leur coût moyen. Cependant, les hypothèses fortes de ce modèle pourraient limiter la qualité de prédiction de celui-ci.

Ce modèle est obtenu en construisant deux modèles GLM :

- Un modèle GLM Poisson pour modéliser la fréquence des sinistres,
- Un modèle GLM pour modéliser la sévérité des sinistres :
  - Un modèle GLM multinomial afin de pouvoir calculer la moyenne définie en 3.1 pour le poste Généraliste.
  - Un modèle GLM Gamma pour le poste Optique.

Pour rappel, la construction de ces modèles GLM s'obtient en minimisant les fonctions de déviance définies en 2.5 à travers l'algorithme de Newthon Raphson.

La prédiction finale de ce modèle qui correspond à la prime pure s'obtient par le produit des deux prédictions.

3.3.4 Le modèle XGB FCM

Ce modèle qui est inspiré du modèle ***GLM FCM*** a l'avantage de s'affranchir de linéarité de ce dernier. Cette flexibilité constitue un avantage qui permet de pouvoir gagner en pouvoir de

prédiction par rapport au ***GLM FCM*** lorsque les hypothèses de loi sont suffisamment respectées. Cependant, ce modèle perd en explicatibilité par rapport au modèle ***GLM FCM***. La construction de ce modèle passe par la construction de deux modèles XGBoost :

- Un modèle XGBoost **Poisson** pour modéliser la fréquence des sinistres :
  La construction de ce modèle consiste à utiliser une fonction de perte semblable à la déviance du modèle GLM Poisson de sorte que cela soit équivalent à faire l'hypothèse de loi sous-jacente. Cette fonction de perte est définie comme suit :

$$\mathscr{L}(y_i, \hat{y}_i) := \sum_{i=1}^{k} \{\hat{y}_i + y_i ln(\hat{y}_i)\}$$

  .

- Un modèle XGBoost est construit pour modéliser la sévérité des sinistres :

  - Un modèle **XGBoost Multinomial** pour la sévérité sur le poste Généraliste :
    La construction de ce modèle consiste à utiliser une fonction de perte inspirée de la déviance du modèle GLM Multinomial logistique afin de construire la moyenne définie en 3.1. Cette fonction de perte, aussi appelée **Softmax** est définie pour k modalités comme suit :

$$\mathscr{L}(y_i, \hat{y}_i) := -\sum_{i=1}^{m}\sum_{j=1}^{k} 1_{\{y_i=j\}} log(\hat{P}(y_i = j|x_i)) = -\sum_{i=1}^{m}\sum_{j=1}^{k} 1_{\{y_i=j\}} log(\hat{y}_i^j)$$

    où m est la taille de l'échantillon et k le nombre de classes modélisées.

  - Un modèle **XGBoost Gamma** pour la sévérité sur le poste optique : La construction de ce modèle consiste à utiliser une fonction de perte inspirée de la déviance du modèle GLM Gamma de sorte que cela soit équivalent à faire l'hypothèse de loi sous-jacente. Cette fonction de perte est définie comme suit :

$$\mathscr{L}(y_i, \hat{y}_i) := \sum_{i=1}^{k} \{x_i \frac{y_i - \hat{y}_i}{\hat{y}_i} + ln(\frac{y_i}{\hat{y}_i})\}$$

3.3.5 Le modèle GLM Tweedie

Pour rappel, le modèle ***GLM Tweedie*** est un modèle permettant de modéliser directement la prime pure. Cette modélisation se fait en deux étapes :

- La première étape qui consiste à estimer par maximum de vraisemblance, le paramètre **p**, qui représente la puissance de la variance et qui constitue une hypothèse de la loi de Tweedie.
- La deuxième étape consiste à ajuster le modèle GLM en faisant l'hypothèse que la variable cible suit une loi de Tweedie d'ordre **p**. On note aussi que cet ajustement GLM se fait en minimisant la déviance associée.

3.3.6 Le modèle XGB Tweedie

Le modèle ***XGB Tweedie*** est un modèle inspiré du modèle ***GLM Tweedie*** et permet de modéliser directement la prime pure en deux étapes :

- La première étape qui consiste à estimer par maximum de vraisemblance, le paramètre **p**, qui représente la puissance de la variance et qui constitue une hypothèse de la loi de Tweedie.
- La deuxième étape consiste à ajuster le modèle XGBoost en faisant l'hypothèse que la variable cible suit une loi de Tweedie d'ordre **p**. pour ce faire, l'on utilise une fonction de perte inspirée de la déviance du modèle ***GLM Tweedie*** et définie comme suit :

$$\mathscr{L}(y_i, \hat{y}_i) = \sum_{i=1}^{k} \{ y_i \frac{\hat{y}_i^{1-p}}{1-p} + \frac{\hat{y}_i^{2-p}}{2-p} \}$$

3.3.7 Un modèle XGB MSE

Ce modèle est construit comme méthode de référence pour mesurer l'apport marginal des fonctions de perte définies pour des problématiques assurantielles. Cette modélisation est couramment utilisée dans la littérature comme approche XGBoost alternative à la modélisation paramétrique. Il utilise l'approche coût totale pour estimer directement la prime pure par le biais de la fonction de perte définie par :

$$\mathscr{L}(y_i, \hat{y}_i) = \sum_{i=1}^{k} \{ y_i - \hat{y}_i \}^2$$

Nous montrerons par la suite que cette méthode n'est pas toujours adaptée à la tarification en assurance.

3.4 Résultats

3.4.1 Évaluation des modèles du poste Généraliste

3.4.1.1 Résultats du processus d'évaluation des modèles

3.4.1.1.1 Premier niveau d'évaluation

Comme défini en 2.4, le premier niveau d'évaluation des modèles permet d'évaluer les ratios O/P définis précédemment. Le tableau ci-dessous présente les résultats de l'évaluation des ratios O/P sur la base d'apprentissage pour le poste Généralise.

Modèles	XGBoost Tweedie	GLM Tweedie	XGBoost Frequence-CM	GLM Frequence-CM	XGBoost MSE
Observation	322 728,60 €	322 728,60 €	322 728,60 €	322 728,60 €	322 728,60 €
Prédiction	314 320,53 €	322 502,85 €	322 545,39 €	322 657,62 €	325 582,26 €
Ratio O/P	102,68%	100,07%	100,06%	100,02%	99,12%
Écarts O-P	-8 408,07 €	-225,75 €	-183,21 €	-70,98 €	2 853,66 €

TABLE 3.5 – Résultats du premier niveau d'évaluation pour le poste Généraliste

L'analyse des résultats du premier niveau d'évaluation permet d'observer des ratios O/P dans l'intervalle d'acceptation pour les modèles GLM Tweedie, XGBoost Fréquence Coût-Moyen, GLM Fréquence-CM et XGBoost MSE. Cependant, pour le modèle XGBoost Tweedie, le ratio O/P n'est pas dans l'intervalle d'acceptation défini.

Cette observation parait tout à fait cohérente. En effet, comme observé lors de la phase d'analyse des données, la distribution des coûts moyens est une distribution discrète et non Gamma. Ainsi, l'hypothèse d'une distribution de Tweedie n'est pas suffisamment respectée. Ces conditions créent donc un biais dans l'adéquation, car les modèles XGBoost utilisent différentes méthodes de rééchantillonnage afin d'éviter l'overfitting et d'être plus fidèle possible, aux hypothèses du modèle. Ce modèle sera donc écarté dans la suite de l'analyse.

Le modèle GLM Tweedie quant à lui s'ajuste bien à la base de données d'apprentissage malgré le biais dans l'hypothèse de loi des côuts moyens. L'analyse du deuxième niveau d'évaluation permettra d'évaluer la qualité de ce modèle à se généraliser.

3.4.1.1.2 Deuxième niveau d'évaluation

Comme défini en 2.4, ce deuxième niveau d'évaluation permet d'avoir une vision par segment des quatre modèles retenus.

Dans cette partie, on défini quatre segments. Ces segments (segment 1, segment 2, segment 3 et segment 4) correspondent à l'ensemble des profils pour lesquelles les modèles GLM Tweedie, XGBoost Fréquence Coût-Moyen, GLM Fréquence-CM et XGBoost MSE ont respectivement, la meilleure évaluation. Cette évaluation correspond à la moyenne des valeurs $|O/P - 1|$ qui sont calculés sur les différents groupes d'individus avec des profils identiques.

Les résultats de cette évaluation sont présentés dans le tableau ci-dessous.

Modèles	Sinistres observés	GLM Tweedie	XGBoost Frequence-CM	GLM Frequence-CM	XGBoost MSE	Modèle optimal
Segment 1	17 164	50,1%	58,4%	53,1%	61,1%	50,1%
Segment 2	26 295	59,3%	46,1%	56,9%	55,6%	46,1%
Segment 3	19 273	98,1%	122,1%	89,5%	110,1%	89,5%
Segment 4	41 130	50,9%	40,4%	44,8%	32,3%	32,3%

TABLE 3.6 – Résultats du deuxième niveau d'évaluation pour le poste Généraliste

L'analyse des résultats permet d'observer qu'il n'existe pas de supériorité absolue d'un modèle sur les autres sur l'ensemble de la base de test. Cependant, le segment 4, qui correspond à l'ensemble des profils où le modèle XGBoost MSE a la meilleure prédiction du montant de sinistre au vu du montant de sinistres empiriques qui correspond à (41 130/103 862 =39,6%) du montant total de sinistres sur le poste Généraliste.

À cette étape, ce modèle semble être le plus performant de quatre modèles. Cela pourrait s'expliquer par l'adéquation insuffisante des autres modèles aux hypothèses de loi formulées.

Par ailleurs, on note que le deuxième meilleur modèle des quatre est également un modèle non-paramétrique au vu du montant de sinistres mieux estimés (26 295/103 862=25,3%) suivi du modèle GLM fréquence Coût-Moyen (18,56%), puis du modèle GLM Tweedie (16,53%).

Ces résultats permettent de construire le modèle optimal à partir des quatre modèles qui représente l'indicatrice par segment préférentiel des prédictions de chacun des quatre modèles.

3.4.1.1.3 Troisième niveau d'évaluation

Ce troisième niveau d'évaluation qui est un agrégat du précédent, permet de mesurer la qualité de la prédiction sur l'ensemble de la base de test pour tous les modèles. Les résultats obtenus sont présentés dans le tableau ci-dessous.

Choix de modèle	GLM Tweedie	XGBoost Frequence-CM	GLM Frequence-CM	XGBoost MSE	Modèle optimal
moyenne de l'erreur	59,5%	56,9%	56,7%	57,5%	48,8%

TABLE 3.7 – Résultats du troisième niveau d'évaluation pour le poste Généraliste

De ces résultats obtenus, on en déduit que le modèle GLM fréquence Coût-Moyen a une meilleure performance moyenne sur l'ensemble de la base de test (56,7%). Ce modèle a une meilleure performance moyenne que celle de son équivalent XGBoost (56,9%) qui est suivi du modèle suivi du modèle XGBoost MSE qui a une performance moyenne (57,51%).

Cette observation est en partie due au fait que toutes les variables explicatives sont des variables discrètes plus favorables aux modèles GLM. En effet, en considérant la forme continue de la variable Age dans les modèles GLM Fréquence Coût-Moyen et XGBoost Fréquence Coût-Moyen, on observe une dégradation de la performance globale des modèles plus importante sur le modèle GLM que sur le modèle Gradient Boosting comme le montre le tableau ci-dessous.

Choix de modèle	XGBoost Frequence-CM	XGBoost Frequence-CM (Age continue)	GLM Frequence-CM	GLM Frequence-CM (Age continue)
Moyenne de l'erreur	56,9%	57,57%	56,7%	62,25%
Dégradation de l'erreur		+0,63%		+5,53%

TABLE 3.8 – Comparaison de la dégradation de la moyenne de l'erreur sur les modèles GLM et XGBoost Fréquence-Coût moyen, après l'intégration de la variable continue Âge

On observe sur cet exemple que le modèle GLM est beaucoup plus sensible à la structure des données et cela est dû à l'hypothèse de linéarité qui reste très forte.

Aussi, on observe que le modèle GLM Tweedie a la plus mauvaise performance. Cette observation est tout à fait cohérente avec l'observation faite sur la structure de la dépense réelle qui n'est pas continue.

Par ailleurs, on observe un gain significatif de performance (14%) avec le modèle optimal par rapport au modèle XGBoost MSE. Ce résultat montre que le modèle optimal construit, qui pour rappel est un mélange des différents modèles construits, permet d'avoir un gain de performance significatif par rapport aux autres modèles.

3.4.1.2 Résultats de l'évaluation RMSE sur la base de test

Le tableau ci-dessous présente les résultats des évaluations des différents modèles avec la métrique RMSE, sur la base de test pour le risque Généraliste.

Choix de modèle	GLM Tweedie	XGBoost Frequence-CM	GLM Frequence-CM	XGBoost MSE	Modèle optimal
RMSE	68,68	68,29	68,60	68,48	67,90

TABLE 3.9 – Résultats RMSE sur la base de test pour le poste Généraliste

Selon les résultats obtenus par le biais de cette métrique, le modèle avec la meilleure performance est le XGBoost Fréquence Coût-Moyen (68,29).

En deuxième position, c'est le deuxième modèle Gradient Boosting, XGBoost MSE qui obtient la meilleure performance (68,48). Le modèle GLM XGBoost est en troisième position (68,60) suivi du modèle GLM Tweedie (68,68).

En outre, on observe un gain de performance avec le modèle optimal précédemment construit qui a un RMSE $0,5\%$ plus bas que le RMSE du modèle XGBoost Fréquence Coût-Moyen.

On note par ailleurs que cette métrique, malgré le fait qu'elle soit largement utilisée en Machine Learning de manière générale, elle n'est pas toujours adaptée. En effet, cette métrique est sensible aux grandes valeurs. Elle n'est donc pas très adaptée lorsque la distribution des données est une distribution asymétrique avec une queue comme les distributions de type Gamme ou Tweedie. Elle sert toutefois de benchmark à notre processus d'évaluation.

3.4.1.3 Résultats Bootstrap sur la base de test

Le but de cette partie est de tester la stabilité et la pertinence des résultats obtenus. Pour cela, la méthode Bootstrap est utilisée afin d'obtenir des nouvelles évaluations sur de nouvelles bases de test que l'on obtient en effectuant plusieurs tirages aléatoires avec remise, à partir de la base de test pour le poste Généraliste.

Ainsi, sur chaque nouvelle base de données obtenue, le RMSE ainsi que le ratio O/P sont calculés pour chaque modèle.

3.4.1.3.1 Écart moyen |O/P-1|

L'analyse des résultats obtenus permet d'observer que les écarts moyens |O/P-1| des différentes méthodes ont la même tendance en fonction des différentes itérations. De plus, on observe aussi que les rangs entre les modèles évoluent beaucoup. Toutefois, les meilleures performances parmi les quatre modèles implémentés sont majoritairement obtenus avec le modèle XGBoost Fréquence Coût-Moyen, suivi du modèle GLM Fréquence Coût-Moyen en deuxième position, suivi du modèle XGBoost MSE en troisième position, et enfin, le modèle GLM Tweedie.

Ces observations mettent en évidence la quasi-stabilité des rangs entre les différents modèles ce qui renforce les résultats moyens obtenus précédemment.

Par ailleurs, les résultats obtenus permettent d'observer que pour toutes les itérations, le modèle optimal construit reste le meilleur en termes de moyenne d'écarts moyens |O/P-1|.

Le graphique ci-dessous présente les résultats obtenus.

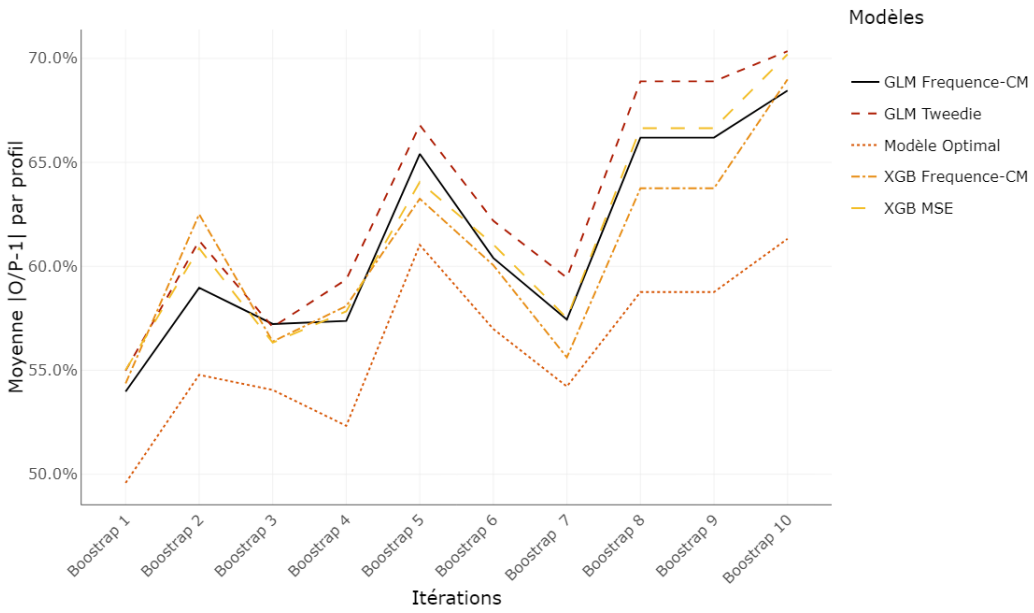


FIGURE 3.20 – Bootstrap de la base de test du poste Généraliste pour le calcul des écarts |O/P-1|

3.4.1.3.2 RMSE

Le graphique ci-dessous présente les résultats obtenus par évaluation avec la métrique RMSE des différents modèles, sur des bases de données Bootstrap de la base de test.

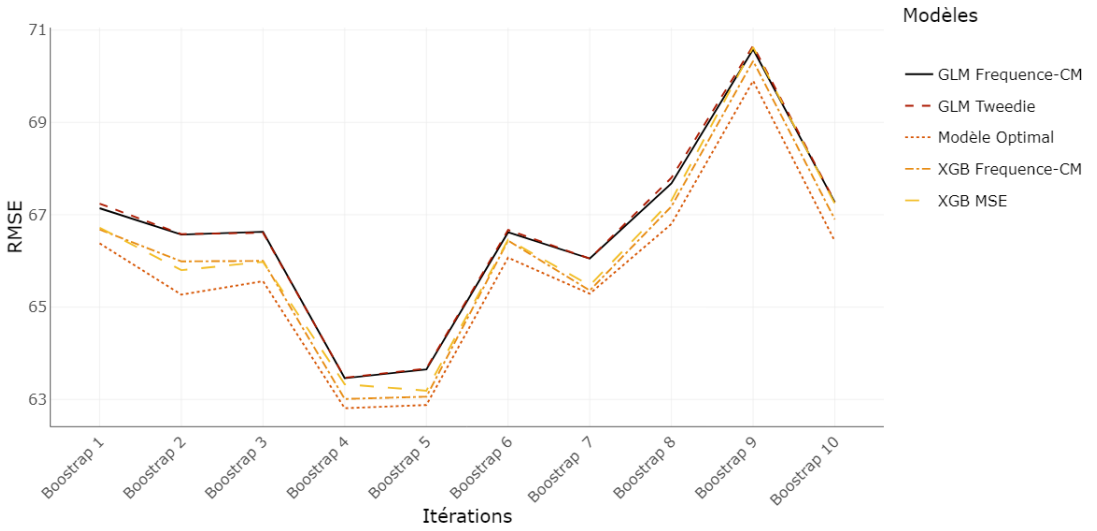


FIGURE 3.21 – Bootstrap de la base de test du poste Généraliste pour le calcul du RMSE

L'analyse des résultats obtenus permet d'observer que les évaluations RMSE des différents modèles ont une même tendance par rapport aux différentes itérations Bootstrap. Par ailleurs, on note aussi que les rangs des différents modèles sont quasiment toujours conservés avec en première position, le modèle XGBoost Fréquence Coût-Moyen, suivi du modèle XGBoost MSE en deuxième position, suivi du modèle GLM Fréquence Coût-Moyen en troisième position, et enfin, le modèle GLM Tweedie. Ainsi, on déduit une stabilité dans la comparaison des modèles en termes de RMSE.

La bonne performance des modèles XGBoost en termes de RMSE s'explique en partie par le fait que ces modèles ont la capacité à s'adapter en présence de données volatiles ce qui n'est pas le cas des modèles GLM.

Par ailleurs, on note que pour toutes les itérations, le modèle optimal construit a une meilleure performance en termes de RMSE par rapport à tous les autres modèles.

Ces résultats viennent consolider les résultats précédemment obtenus avec le Bootstrap des écarts moyens |O/P-1|.

3.4.2 Évaluation des modèles du poste Optique

3.4.2.1 Résultats du processus d'évaluation des modèles

3.4.2.1.1 Premier niveau d'évaluation

De même que pour le poste Généraliste, ce premier niveau d'évaluation permet d'évaluer les ratios O/P pour tous les modèles du poste Optique. Le tableau ci-dessous présente les résultats de l'évaluation des ratios O/P sur la base d'apprentissage de ce poste.

Modèles	XGBoost Tweedie	GLM Tweedie	XGBoost Frequence-CM	GLM Frequence-CM	XGBoost MSE
Observation	691 284,28 €	693 140,69 €	687 219,54 €	693 081,06 €	690 303,65 €
Prédiction	692 948,20 €	692 948,20 €	692 948,20 €	692 948,20 €	692 948,20 €
Ratio O/P	100,24%	99,97%	100,8%	99,98%	100,383%
Écarts O-P	-1 663,92 €	192,49 €	-5 728,66 €	132,86 €	-2 644,55

TABLE 3.10 – Résultats du premier niveau d'évaluation pour le poste Optique

Au vu de ces résultats, tous les modèles sont acceptés pour la suite du processus d'évaluation.

3.4.2.1.2 Deuxième niveau d'évaluation

De même que pour le poste Généraliste, dans cette partie, on cherche à avoir une vision par segment des cinq modèles retenus.

Pour cela, on définit cinq segments. Ces segments (segment 1, segment 2, segment 3, segment 4 et segment 5) qui correspondent à l'ensemble des profils pour lesquels les modèles XGBoost Tweedie, GLM Tweedie, XGBoost Fréquence Coût-Moyen, GLM Fréquence-CM et XGBoost MSE ont respectivement la meilleure évaluation ($|O/P - 1|$).

Modèles	Sinistres observés	XGBoost Tweedie	GLM Tweedie	XGBoost Frequence-CM	GLM Frequence-CM	XGBoost MSE	Modèle optimal
Segment 1	47 025,07 €	78,64%	92,19%	93,47%	95,75%	82,17%	78,64%
Segment 2	34 327,63 €	61,71%	57,26%	59,89%	58,42%	65,66%	57,26%
Segment 3	37 908,75 €	112,89%	107,83%	91,21%	101,93%	117,42%	91,21%
Segment 4	98 375,78 €	97,55%	81,69%	81,33%	76,44%	101,41%	76,44%
Segment 5	19 176,83 €	72,94%	82,00%	86,60%	83,58%	67,88%	67,88%

TABLE 3.11 – Résultats du deuxième niveau d'évaluation pour le poste Optique

L'analyse des résultats permet d'observer qu'il n'existe pas de supériorité absolue d'un modèle sur les autres sur l'ensemble de la base de test. Cependant, le segment 4, qui correspond au modèle GLM Fréquence Coût-Moyen, est de loin, le segment avec la plus grande part de sinistres observés (41,54%). A ce niveau, ce modèle semble être le plus performant.

En deuxième position, c'est le modèle XGBoost Tweedie avec 19,86% des sinistres observés, suivi du modèle XGBoost Fréquence Coût-Moyen avec 16,01% des sinistres observés. En quatrième position, on observe le modèle GLM Tweedie (14,5%) suivi en dernière position du modèle XGBoost MSE.

Ces observations permettent de construire un modèle plus optimal qui correspond au meilleur des cinq modèles sur chaque segment afin d'aboutir à une meilleure qualité de prédiction.

3.4.2.1.3 Troisième niveau d'évaluation

Le tableau ci-dessous présente les moyennes des écarts moyens |O/P-1| calculés précédemment sur toute la base de test.

Modèles	XGBoost Tweedie	GLM Tweedie	XGBoost Frequence-CM	GLM Frequence-CM	XGBoost MSE	Modèle optimal
Moyenne	88,19%	86,32%	84,31%	84,73%	90,59%	75,94%

Table 3.12 – Résultats du troisième niveau d'évaluation pour le poste Optique

Les résultats obtenus au troisième niveau de cette évaluation permettent d'observer que le modèle XGBoost Fréquence Coût-Moyen a une meilleure performance moyenne sur l'ensemble de la base de test (84,31%).

Cette performance est très proche de celle du modèle GLM-Fréquence Coût-Moyen (84,73%) qui vient en deuxième position.

En troisième position, vient le modèle GLM Tweedie (86,32%) suivi du modèle XGBoost Tweedie (88,19%). En dernière position, vient le modèle XGBoost MSE.

On observe par ailleurs un gain de performance significatif (- 10% de d'erreur) avec le modèle optimal par rapport à la meilleure performance.

3.4.2.2 Résultats de l'évaluation RMSE sur la base de test

Le tableau ci-dessous présente les résultats des évaluations des différents modèles avec la métrique RMSE, sur la base de test du poste Optique.

Modèles	XGBoost Tweedie	GLM Tweedie	XGBoost Frequence-CM	GLM Frequence-CM	XGBoost MSE	Modèle optimal
Moyenne	228,99	228,26	227,99	228,10	229,62	226,89

Table 3.13 – Résultats RMSE sur la base de test pour le poste Optique

Les résultats obtenus par le biais de cette métrique montrent que le modèle avec la meilleure performance est le XGBoost Fréquence Coût-Moyen avec un RMSE de 227,99 suivi du modèle GLM Fréquence Coût-Moyen en deuxième position avec un RMSE de 229,62.

En troisième position, vient le modèle GLM Tweedie (228,26) suivi du modèle XGBoost Tweedie (228,99). En dernière position, vient le modèle XGBoost MSE (229,62).

On remarque par ailleurs que pour le processus d'évaluation ainsi que pour la métrique RMSE, le modèle XGBoost MSE a obtenu des performances en dessous des autres modèles. Cela est dû au fait que la fonction de perte MSE, comme démontrée en 2.3.3.1, ne s'adapte pas très

bien aux distributions asymétriques comme c'est le cas pour la distribution de la dépense réelle pour le poste Optique.

Par ailleurs, on observe une amélioration de la performance avec le modèle optimal précédemment construit qui permet d'obtenir un gain de performance (baisse de 1,1 point de RMSE) par rapport au modèle XGBoost Fréquence Coût-Moyen.

3.4.2.3 Résultats Bootstrap sur la base de test

3.4.2.3.1 Écarts moyens |O/P-1|

Les résultats obtenus par Bootstrap sur la base de test du poste Optique pour la moyenne des écarts |O/P-1| sont représentés sur le graphe ci-dessous.

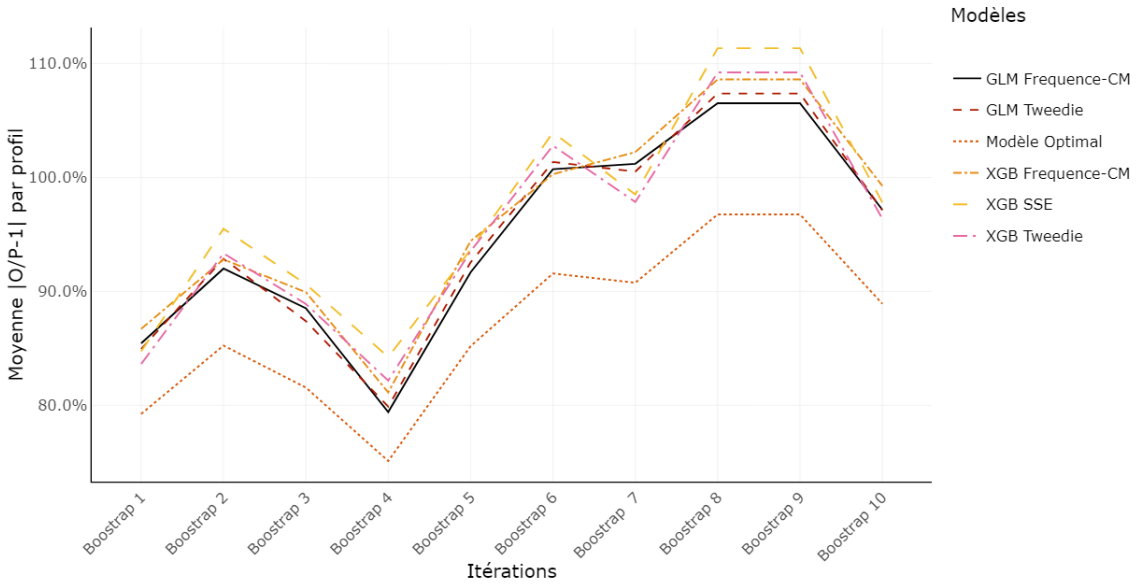


FIGURE 3.22 – Bootstrap de la base de test du poste Optique pour le calcul des écarts moyens |O/P-1|

L'analyse du graphique ci-dessus permet d'observer que les moyennes de ces écarts pour les différents modèles ont une même tendance en fonction des itérations avec un rang qui est quasiment conservé pour toutes les itérations. De plus, la meilleure performance est majoritairement obtenue avec le modèle GLM Fréquence Coût-Moyen, suivi du modèle GLM Tweedie, suivi du XGBoost Fréquence Coût-Moyen, suivi du modèle XGBoost Tweedie et enfin, en dernière position, le modèle XGBoost MSE.

Cependant, il est bien de noter que les performances des quatre premièrs modèles cités sont très proches. Aussi, le modèle XGBoost MSE est une fois de plus en dernière position, ce qui montre une fois de plus le caractère inadapté aux données du poste Optique.

Par ailleurs, le modèle optimal construit possède la meilleure performance en termes d'écart |O/P-1| pour toutes les itérations.

3.4.2.3.2 RMSE

Les résultats obtenus par Bootstrap sur la base de test du poste Optique pour la métrique RMSE sont représentés sur le graphe ci-dessous.

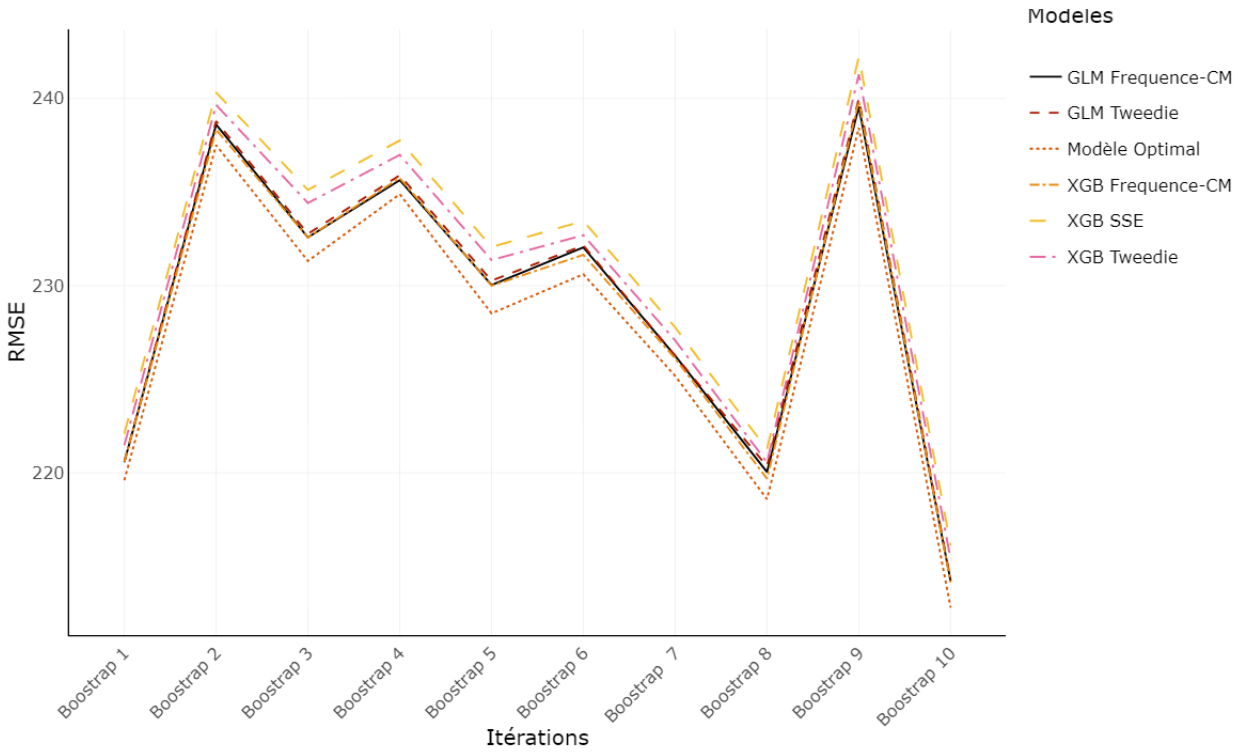


FIGURE 3.23 – Bootstrap de la base de test du poste Optique pour le calcul du RMSE

L'analyse de ces résultats permet de mettre en évidence une évolution homogène des performances de tous les modèles avec un rang entre les modèles qui est quasiment toujours conservé. On note par ailleurs que les performances des différents modèles en termes de RMSE sont très proches sur la plupart des itérations.

On observe en outre que le modèle optimal construit reste le meilleur pour toutes les itérations.

3.5 Conclusion de la mise en application

L'analyse des résultats obtenus dans cette partie du mémoire a permis d'observer ces résultats importants :

- Le processus d'évaluation défini dans ce mémoire a permis d'avoir une vision aussi bien détaillée que agrégée des performances des modèles.
- L'analyse détaillée des performances des modèles par le biais du processus d'évaluation a pu mettre en évidence qu'il n'existe pas de supériorité absolue d'un des modèles sur les autres sur l'ensemble des profils de la base de test. Cette analyse a aussi permis la construction d'un modèle optimal qui permet d'avoir un gain marginal de performance par rapport à tous les autres modèles.
- L'utilisation des modèles de Gradient Boosting a permis d'avoir un gain marginal par rapport aux GLMs en termes de performance moyenne des modèles. C'est notamment le cas sur le poste Généraliste où ces modèles ont presque tout le temps été meilleurs que les modèles GLM en termes de performance. En outre, l'utilisation des fonctions de perte spécifique aux problèmes assurantielles a permis d'avoir un gain significatif en termes de performance notamment sur le poste Optique où la fonction de perte classique s'est avérée peu efficace.
- Aussi, on note la stabilité et la capacité des modèles de Gradient Boosting à pourvoir s'adapter aux données. c'est notamment le cas lors de l'utilisation de la variable continue ***Age***, au détriment de la variable catégorielle ***Classe d'âges*** sur le poste Généraliste. Cela a fortement dégradé le modèle GLM mais cela fut moins le cas pour le modèle Gradient Boosting équialent.

La suite de ce mémoire sera consacrée à l'interpretabilité des différents modèles construits.

4 Transparence des modèles

"En mathématiques, "évident" est le mot le plus dangereux."
Eric Temple Bell 1883 - 1960

4.1 Introduction

Le régime RGPD de "responsabilité algorithmique" et le "droit à l'explication" qui en résulte soulignent l'importance d'avoir des modèles de tarification transparents. Cependant, les techniques de Machine Learning sont souvent considérées comme des boîtes noires (***black box***) contrairement aux modèles statistiques tels que les GLMs.

En effet, pour les modèles de type GLM, les estimations des paramètres et leurs erreurs types donnent des informations sur l'effet, l'incertitude et la significativité de toutes les variables explicatives.

Les modèles de Machine Learning quant à eux s'accompagnent d'une difficulté à interpréter d'une part le processus de génération du modèle et d'autre part la prédiction de la variable d'intérêt, la prime dans notre cas de figure.

Il existe plusieurs méthodes pour permettre l'interpretabilité des modèles de prédiction dits ***black box***, dont le Gradient Boosting est l'un des plus populaires. Ces méthodes d'interprétabilité sont catégorisées selon différents dualismes :

- **Intrinsèques vs post-hoc :** les méthodes intrinsèques pour l'interprétabilité sont des méthodes qui produisent des modèles interprétables tandis que les méthodes post-hoc produisent des méthodes black box.
- **Locales vs globales :** les modèles d'interprétabilité locales sont des méthodes permettant d'expliquer une prédiction selon un input donné. A contrario, lorsque la méthode d'interprétation concerne le fonctionnement global de l'algorithme sur l'ensemble des inputs de la base de données, on dira que la méthode est globale.
- **Agnostique vs spécifique :** les modèles d'interprétabilité agnostiques sont des modèles pouvant être utilisés pour n'importe quelles classes de méthodes d'apprentissage. A contrario, les modèles d'interprétabilité spécifiques ne peuvent être utilisés que pour interpréter une famille spécifique d'algorithmes, tels que le Gradient Boosting.

Enfin, nous nous intéressons aux méthodes agnostiques du fait de leur capacité à s'adapter à tous les modèles.

Plusieurs méthodes agnostiques existent. Cependant, dans le cadre de ce mémoire, les méthodes présentées sont les suivantes :

- La mesure de la force de l'interaction des variables (Variable interaction strength (IS)),
- L'importance des variables par permutation (Permutation feature importance - VIP),
- Le graphique de dépendance partielle (Partial Dépendance Plots - PDP),
- Le graphique de l'espérance conditionnelle individuelle (Individual Conditional Expectation (ICE) plots).

4.2 Mise en oeuvre de l'interpretabilité

4.2.1 Le poste Généraliste

4.2.1.1 Importance des variables

Le graphique ci-dessous présente l'importance des variables explicatives dans chacun des différents modèles de prédiction pour le poste Généraliste.

L'objectif est de pouvoir mesurer l'apport marginal en termes de pouvoir de prédiction, de chacune des variables présentes dans les différents modèles. Pour une variable explicative donnée, cette métrique se traduit par la perte de pouvoir de prédiction, mesurée ici par la baisse du MAE, lors de la permutation des modalités de cette variable dans le modèle.

Cette permutation, qui déforme la relation entre la variable explicative et l'observation de la variable à expliquer, permet de mesurer l'évolution de la qualité de la prédiction introduite en neutralisant l'effet de cette variable explicative dans le modèle.

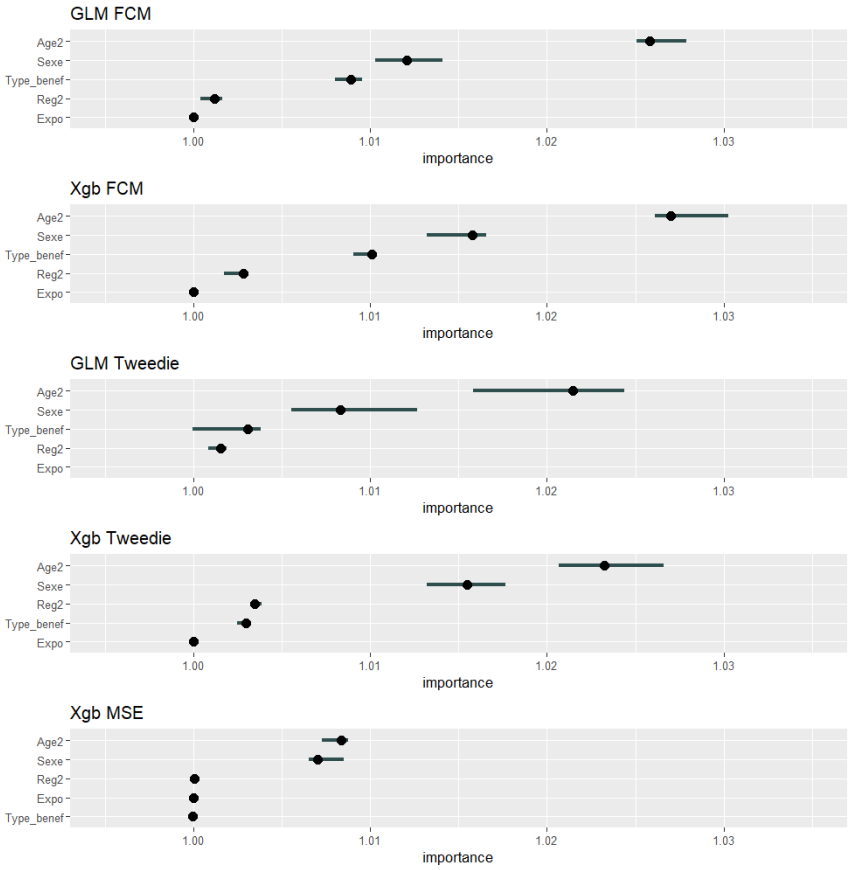


FIGURE 4.1 – Importance des variables pour le poste Généraliste

En analysant le graphique, on constate une importance beaucoup plus marquée de la variable Classe d'âge dans tous les différents modèles. En effet l'absence de celle-ci provoque une baisse moyenne du pouvoir de prédiction de 2,6% dans le modèle ***GLM FCM***, 2,7% dans le modèle ***XGB FCM*** ; 2,2% dans le modèle ***GLM Tweedie***, 2,4% dans le modèle ***XGB Tweedie*** et 0,9% dans le modèle ***XGB MSE***. Ces baisses sont plutôt stables selon les approches Fréquence coût moyen (2,6% vs 2,7%) et Tweedie (2,2% vs 2,4%). Cependant, pour le modèle ***XGB MSE***, on observe une baisse significativement faible. Cela montre que ce modèle utilise moins cette variable pour segmenter.

Concrètement, ces observations traduisent le fait que l'âge est la variable qui permet de mieux segmenter les assurés sur le poste Généraliste pour les différents modèles. Ainsi, si l'assureur devait considérer un axe privilégié de segmentation, le choix judicieux, au regard de cette analyse, serait d'accorder plus d'importance à cette variable dans la tarification.

La deuxième variable la plus importante dans tous les différents modèles est la variable Sexe. On trouve par la suite les variables Type de bénéficiaire et Classes de régions à des niveaux d'importance moindre.

Aussi, on observe que le modèle ***XGB MSE*** n'utilise pas les variables Groupes de région et Type de bénéficiaire dans sa segmentation. Cela se traduit par une importance nulle de ces variables dans la qualité de la prédiction.

Cette dernière observation renforce l'hypothèse selon laquelle ce modèle ne serait pas très adapté aux données pour le poste généraliste.

En outre, on observe que les modèles de type GLM ont des niveaux d'importance de variable moins élevés que ceux de leurs équivalents Gradient Boosting. Cela traduit aussi la capacité des modèles de type Gradient Boosting à dépendre beaucoup plus de ces variables et donc à segmenter plus finement par rapport à celles-ci.

Aussi, la stabilité des niveaux d'importance observée selon les approches utilisées est très intéressante. En effet, cela à tendance à montrer que pour les données de ce poste, sous les mêmes hypothèses de loi, les deux modèles XGBoost et GLM ont tendance à segmenter de la même manière avec un avantage pour le modèle XGBoost comme observé précédemment.

4.2.1.2 Force d'interaction des variables

Le graphe ci-dessous présente la contribution de chaque variable à l'ensemble des interactions de chaque modèle.

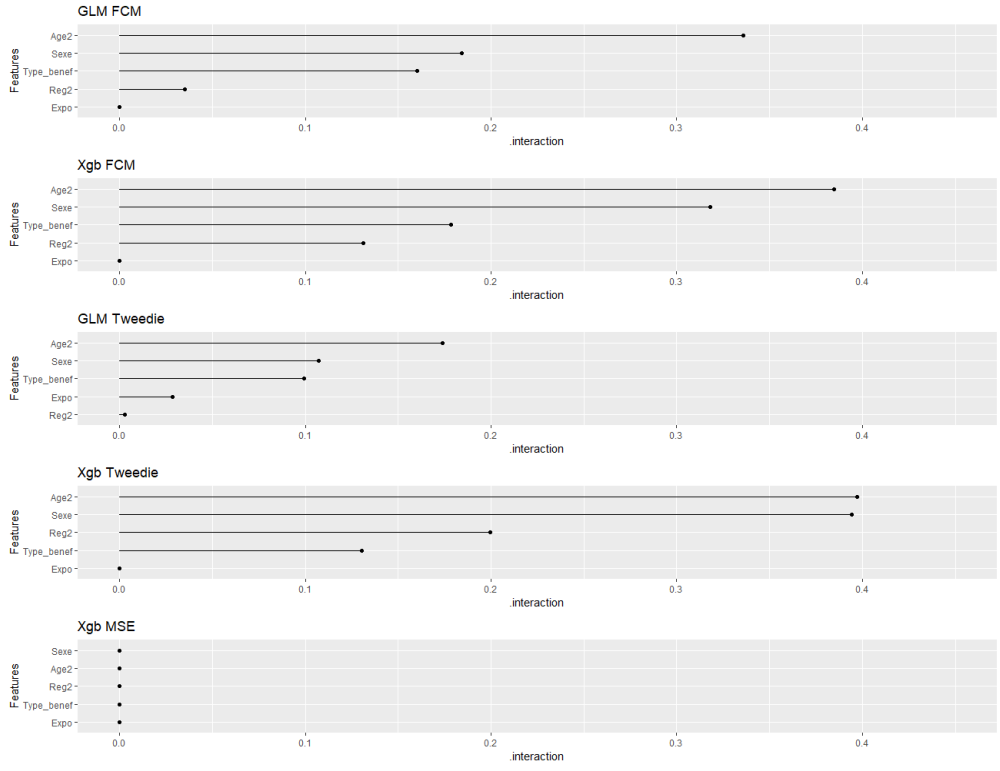


FIGURE 4.2 – Forces d'interaction pour le poste Généraliste

L'analyse de ces graphiques permet de constater que pour le poste Généraliste, la variable Classe d'âge est celle qui interagit le plus avec les autres variables. C'est notamment le cas pour les modèles ***GLM FCM***, ***XGB FCM***, GLM Tweedie et ***XGB Tweedie***.

Concrètement, cela signifie qu'en dehors de l'apport individuel de cette variable, un supplément de pouvoir de prédiction est obtenu dans ces différents modèles en croisant cette variable avec les autres variables explicatives.

Cette information supplémentaire permettrait d'isoler au sein d'une Classe d'âge donnée, les individus plus ou moins à risque par rapport à l'ensemble de la classe en considérant des interactions importantes entre l'Age et d'autres variables.

On observe aussi que pour les modèles GLM, les niveaux des interactions entre les variables sont moins importants que ceux de leurs équivalents Gradient Boosting. Cela est principalement dû à la forme paramétrique des modèles GLM qui impose une contrainte de linéarité qui n'est pas observé avec le modèle XGB. Cela traduit aussi la capacité des modèles de Gradient Boosting à capter des interactions complexes et donc plus d'informations pour la segmentation lorsque les hypothèses de loi sont suffisamment correctes.

Ces interactions sont bien marquées dans les modèles ***XGB Tweedie*** et ***XGB FCM*** et quasi-inexistante dans le modèle ***XGB MSE***. Cela s'explique par le fait que ce modèle est en effet très proche du modèle linaire Gaussien comme on le montre en 2.3.3.1.

Ces observations supplémentaires viennent renforcer l'hypothèse selon laquelle la variable Classe d'âge est la variable la plus importante dans l'ensemble pour ce poste.

4.2.1.3 Graphiques de dépendance partielle et ICE

Dans la continuité de l'analyse de l'importance des variables ainsi que des interactions des variables pour le poste Généraliste, notre analyse de la dépendance partielle ainsi que des graphiques ICE va se porter sur la variable Classe d'âge qui correspond à la variable explicative la plus importante en termes d'importance dans la prédiction (A.1.4).

Pour chacun des modèles construits, les boxplots ci-dessous présentent les distributions des prédictions, en fonction des modalités de la variable Classe d'âge (graphique ICE) et les courbes en rouge permettent quant à elles, d'observer l'évolution de la prédiction moyenne en fonction des modalités de la variable explicative Classe d'âge (graphique PDP représenté par la courbe en rouge)

Le graphique ci-dessous présente les graphiques ICE et PDP pour les modèles ***GLM FCM*** et ***XGB FCM***.

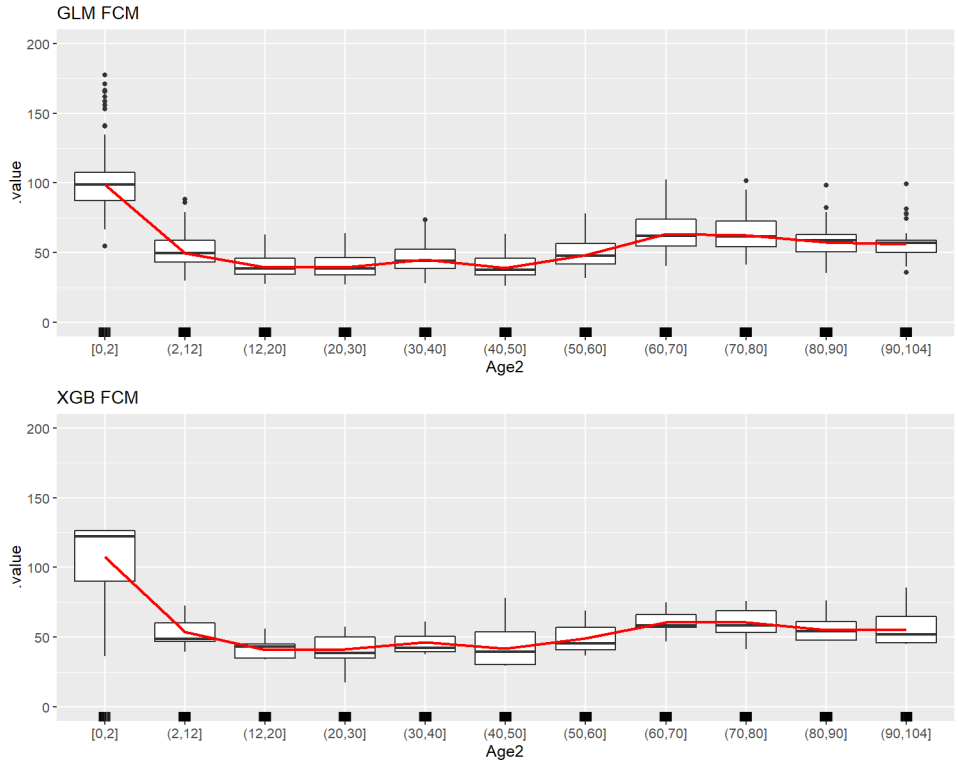


FIGURE 4.3 – Graphes des Dépendances partielles et ICE pour les modèles Fréquence - Coût moyen du poste Généraliste

Le graphique ICE permet d'observer les boites des boxplots représentant les distributions marginales des prédictions des bénéficiaires en fonction des différentes classes d'âge. Cette observation permet de comparer les niveaux de risque des différentes classes d'âge.

On se rend en effet compte, en analysant le graphique ci-dessus, que les boxplots représentant la distribution marginale des prédictions des bénéficiaires de la classe d'âge [0, 2] sont beaucoup plus au-dessus des autres boxplots dans les deux modèles. Cela traduit la significativité de la différence entre les prédictions de cette classe avec celles des autres groupes. Cette différence significative montre le caractère beaucoup plus risqué de cette classe d'âge par rapport aux autres classes. Par ailleurs, ces boxplots permettent de comparer les différentes classes d'âge entre elles en fonction de leurs niveaux de risque.

On note aussi qu'on observe un nombre beaucoup plus important de valeurs extrêmes sur le modèle ***GLM FCM***. Ce résultat vient renforcer l'hypothèse selon laquelle les modèles XGBoost sont beaucoup moins sensibles à la variation des données.

Le graphique PDP permet quant à lui, d'observer que la courbe de la prédiction moyenne des bénéficiaires par Classe d'âge a la même tendance pour ces deux modèles. Ce résultat montre que les deux modèles ont des segmentations très similaires par rapport à l'âge ce qui est plutôt

attendu puisque les hypothèses de loi faites sur les deux modèles sont identiques.

De même que pour les modèles Fréquence coût moyen, le graphique ci-dessous présente les graphiques ICE et PDP pour le modèle GLM Tweedie, ***XGB Tweedie*** et ***XGB MSE***.

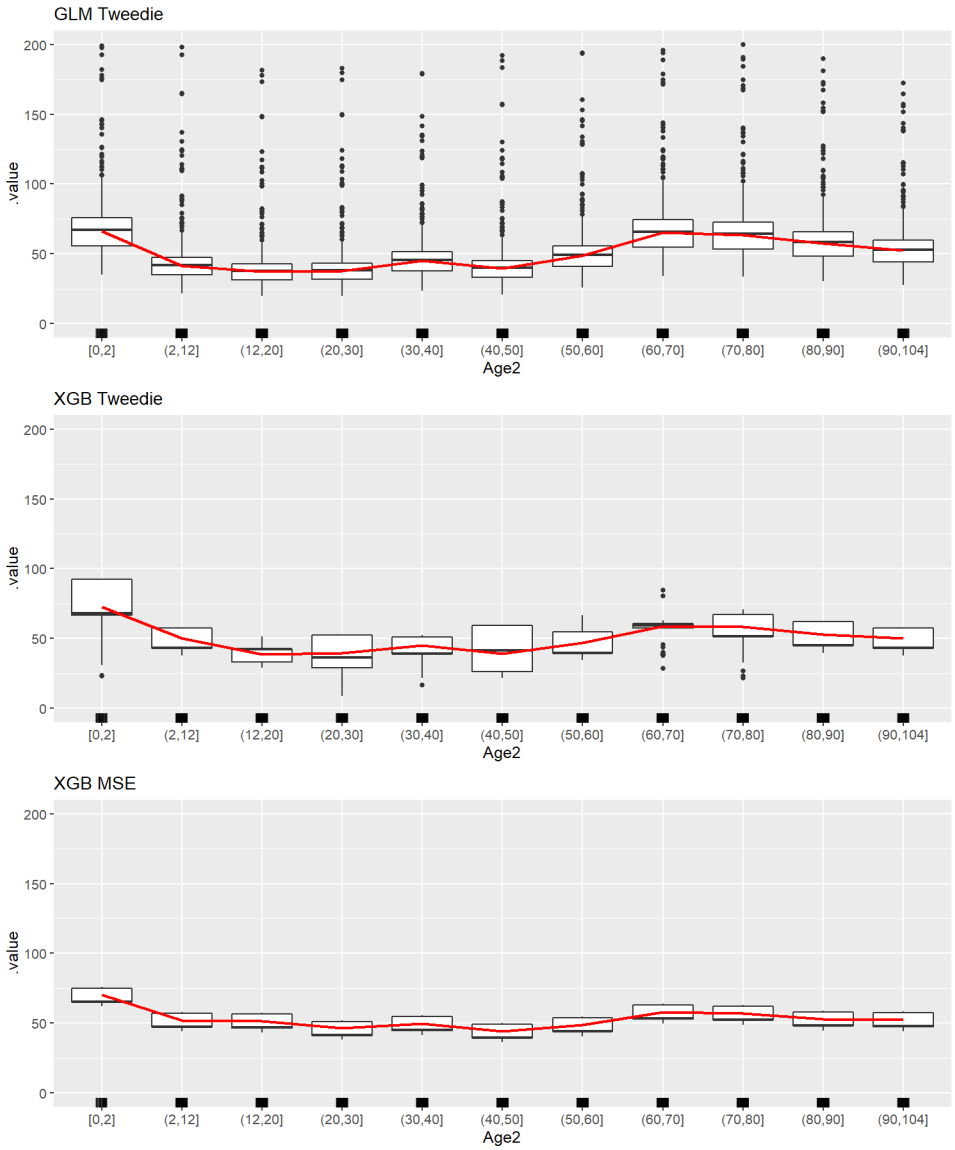


FIGURE 4.4 – Graphes des Dépendances partielles et ICE pour les modèles Tweedie du poste Généraliste

Par le biais du graphique ICE, on se rend aussi compte que le boxplot représentant la distribution marginale des prédictions des bénéficiaires de la classe d'âge [0, 2] est beaucoup plus au-dessus des autres boxplots dans les ***XGB Tweedie*** et ***XGB MSE***. Cela est moins vrai

dans le modèle ***GLM Tweedie*** où les classes d'âge (60 ;70] et (70 ;80] sont plus risquées. On peut ainsi comparer les classes entre elles selon la distribution de leurs prédictions pour en déduire leurs niveaux relatifs de risque.

En outre, l'analyse du graphique PDP permet d'observer que la courbe de la prédiction moyenne des bénéficiaires par Classe d'âge a la même tendance pour ces trois modèles. Cependant, le modèle ***XGB MSE*** présente une segmentation beaucoup moins marquée. Cette observation vient une fois de plus renforcer l'hypothèse selon laquelle ce modèle segmente très peu et n'est pas très adapté aux données de ce poste.

Ces observations qui sont semblables à celles faites pour les modèles Fréquence coût moyen viennent renforcer les résultats précédemment cités.

D'autre part, on en déduit de tous ces résultats que sur les données de ce poste, à hypothèse de loi équivalente, les modèles XGBoost sont beaucoup plus robustes aux données. En effet, ces modèles, ont des prédictions beaucoup moins volatiles et donc beaucoup moins sensibles à la déformation du portefeuille tout en gardant la même structure de segmentation que leurs équivalents GLM.

4.2.2 Le poste Optique

4.2.2.1 Importance des variables

Le graphique ci-dessous présente l'importance des variables explicatives dans chacun des différents modèles de prédiction pour le poste Optique.

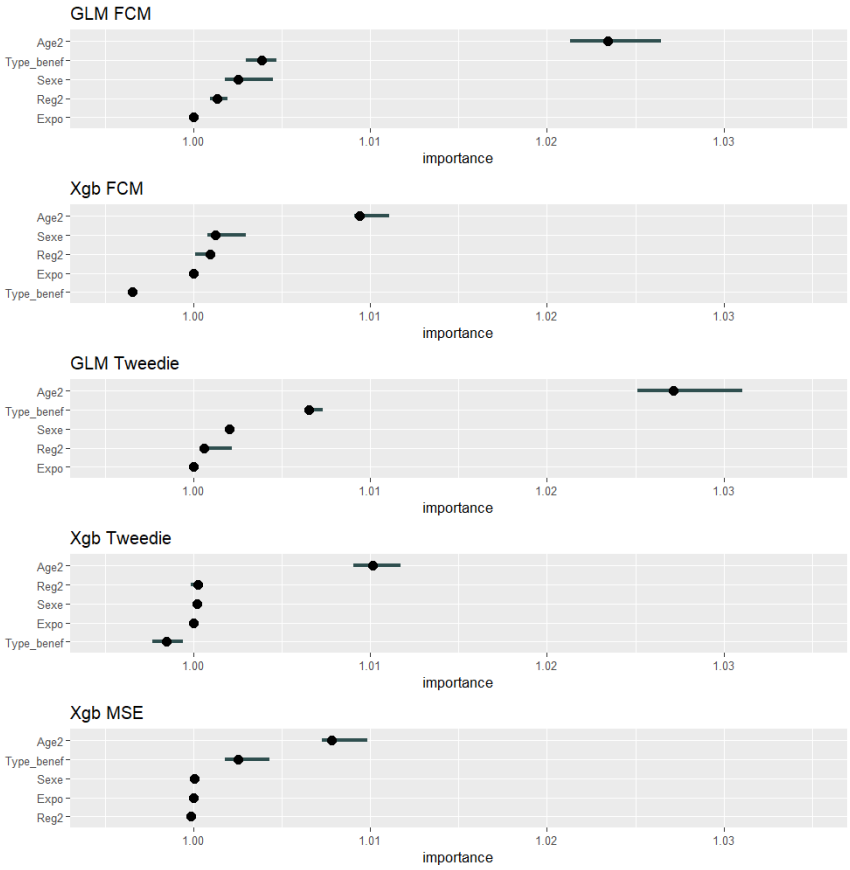


FIGURE 4.5 – Importance des variables pour le poste Optique

Pour ce poste aussi, on constate une importance beaucoup plus marquée de la variable Classe d'âge dans tous les différents modèles. Cependant, cette importance est beaucoup moins marquée dans les modèles XGBoost par rapport au modèle GLM. En effet, l'absence de la variable classe d'âge provoque une baisse moyenne du pouvoir de prédiction de 2,4% dans le modèle ***GLM FCM***, 1,1% dans le modèle ***XGB FCM*** ; 2,7% dans le modèle ***GLM Tweedie***, 1,2% dans le modèle ***XGB Tweedie*** et 0,9% dans le modèle ***XGB MSE***. Cela montre une dépendance beaucoup plus forte de l'âge, pour les modèles GLM.

Aussi, ces observations montrent que cette variable est celle qui permet de mieux segmenter les assurés sur le poste Optique pour les différents modèles. Ainsi, si l'assureur devait considérer un axe privilégié de segmentation, le choix judicieux, au regard de cette analyse, serait une fois de plus l'âge.

On observe par ailleurs que la variable Type de bénéficiaire a tendance à pénaliser la prédiction dans les modèles ***XGB FCM*** et ***XGB Tweedie*** tandis que l'effet contraire est observé dans leurs équivalents GLM. Ce résultat plutôt inattendu signifie que cette variable a une influence moins bonne qu'un bruit complètement aléatoire dans ces deux modèles XGBoost.

Cette variable doit donc être retirée du modèle.

En outre, pour les variables Sexe et Classes de région, on observe une importance du même ordre dans les modèles ***GLM FCM***, ***GLM Tweedie*** et ***XGB FCM***. Cependant, on observe une importance quasiment nulle dans les modèles ***XGB Tweedie*** et ***XGB MSE***. Cela montre que ces modèles dépendent très peu de ces variables.

En définitive, on observe sur ce poste que les modèles GLM ont tendance à utiliser plus de variables dans leurs segmentations que les modèles XGBoost. Elles ont donc tendance à segmenter un peu plus que les modèles XGBoost ce qui est en phase avec les résultats obtenus au deuxième niveau de processus d'évaluation du modèle (3.11)

Cependant, ces résultats doivent aussi être analysés avec beaucoup de recul. En effet, la connaissance du métier montre que sur ce poste, il existe une périodicité de deux ans sur la fréquence des sinistres qui est due à des contraintes contractuelles. Cette structure de la fréquence des sinistres est assez atypique et n'est pas toujours en adéquation avec la loi de poisson. Les modèles XGBoost qui sont beaucoup plus sensibles aux hypothèses faites sur les modèles ont tendance à être moins efficaces. En effet, comme évoqué en 2.3.2.3, les modèles XGBoost utilisent plusieurs méthodes supplémentaires d'optimisation (régularisation, traitement de données creuses, ***Shrinkage***, etc.) pour optimiser au mieux le modèle aux données, ce qui aura tendance à faire diverger le modèle en cas d'hypothèses non vérifiées.

4.2.2.2 Force d'interaction des variables

Le graphe ci-dessous présente la contribution de chaque variable à l'ensemble des interactions de chaque modèle pour le poste Optique.

L'analyse de ces graphiques permet de constater que pour le poste Optique, la variable Classe d'âge est celle qui interagit le plus avec les autres variables. C'est notamment le cas pour les modèles ***XGB FCM***, ***GLM Tweedie*** et ***XGB Tweedie***. Cependant, dans le modèle ***GLM FCM***, son interaction avec les autres variables est beaucoup moins marquée.

Par ailleurs, on observe que sur ce poste, les interactions des modèles XGBoost ne sont pas plus marquées que celles de leurs équivalents GLM. Ce résultat est en phase avec les observations de l'analyse de l'importance des variables et s'explique par l'adéquation aux données qui est moins bonne pour les modèles XGBoost du fait de l'hypothèse de poisson trop forte dans le cadre des données de ce poste comme cité précédemment.

Aussi, pour le poste optique, on observe des interactions nulles dans le modèle XGBoost MSE. Cela s'explique une fois de plus par le fait que ce modèle est très proche du modèle linaire Gaussien comme on le montre en 2.3.3.1.

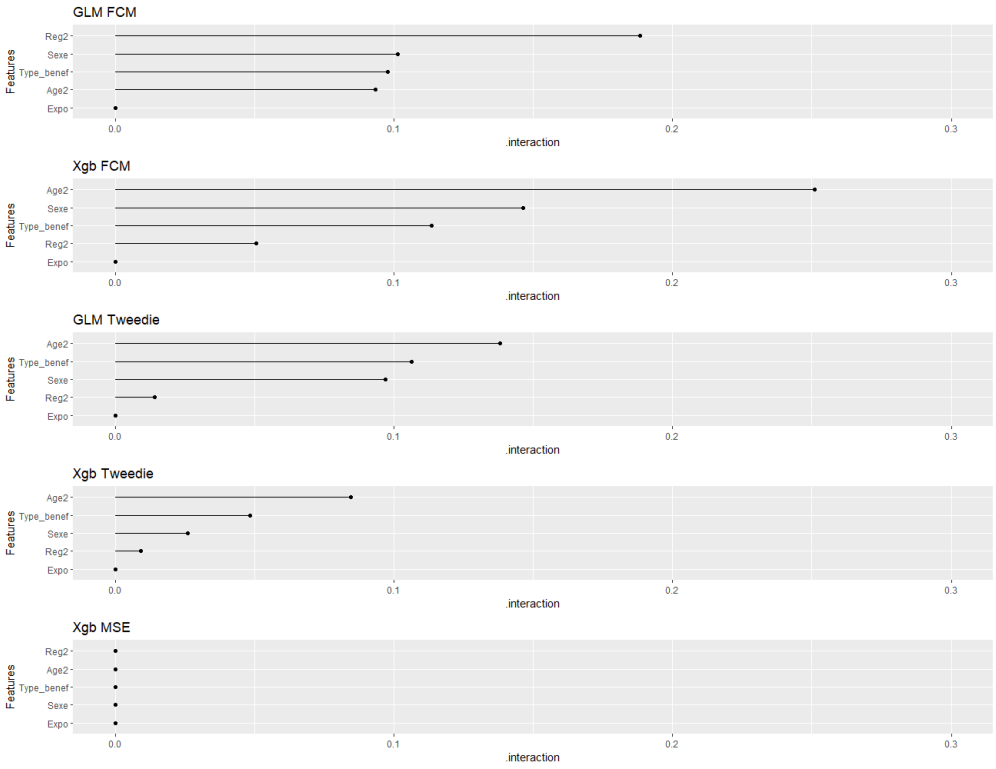


FIGURE 4.6 – Force d'interaction pour le poste Optique

4.2.2.3 Graphiques de dépendance partielle et ICE

De même que pour le poste optique et dans la continuité de l'analyse de l'importance des variables ainsi que des interactions des variables pour le poste Généraliste, notre analyse de la dépendance partielle ainsi que des graphiques ICE va se porter sur la variable Classe d'âge qui correspond à la variable explicative la plus importante en termes d'importance dans la prédiction (A.1.4).

Le graphique ci-dessous présente les graphiques ICE et PDP pour les modèles ***GLM FCM*** et ***XGB FCM***.

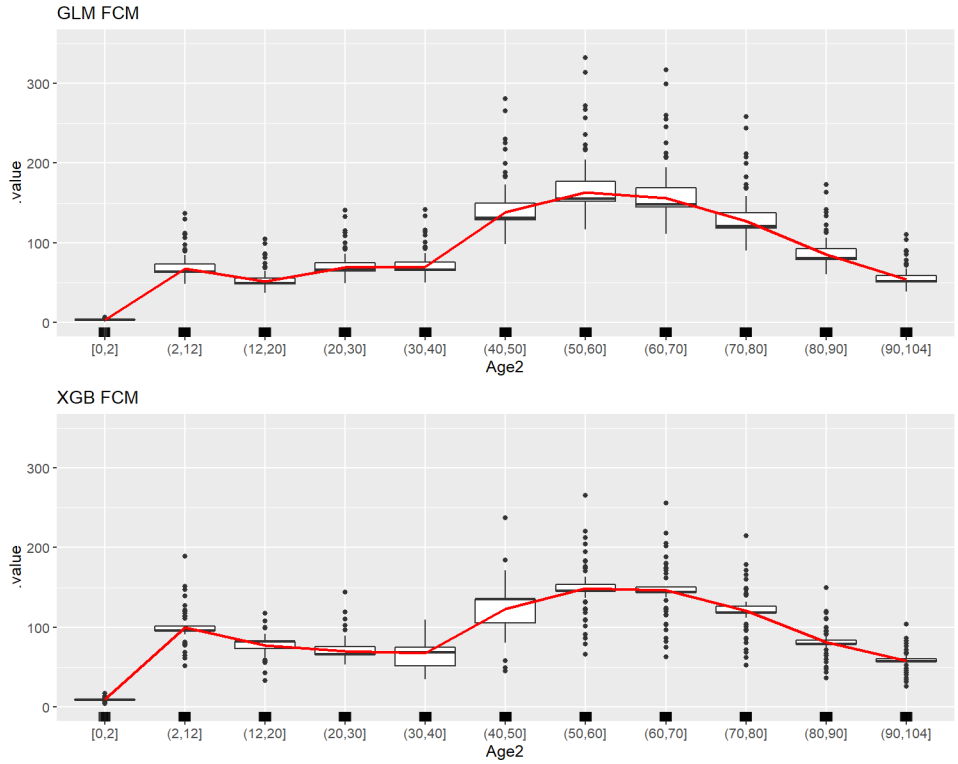


FIGURE 4.7 – Graphes des Dépendances partielles et ICE pour les modèles Fréquence - Coût moyen du poste Généraliste

Pour rappel, Le graphique ICE permet d'observer les boites des boxplots représentant les distributions marginales des prédictions des bénéficiaires en fonction des différentes classes d'âge.

L'analyse des boxplots du graphique ci-dessus permet de se rendre compte que les bénéficiaires de la classe d'âge [0, 2] sont beaucoup plus en dessous des autres boxplots dans les deux modèles. Cela montre que cette classe d'âge est significativement moins risquée que les autres classes. Ce résultat est complètement en ligne avec la connaissance métier sur ce poste. En effet les enfants entre 0 et 2 ans sont rarement concernés par des prestations d'optique. Ce graphique permet ainsi de comparer les niveaux de risque des différentes classes entre elles.

On note par ailleurs que le nombre de valeurs extrêmes est très important dans les deux modèles considérés. Ce résultat montre une hétérogénéité des prédictions qui serait due d'une part, à l'influence des autres variables explicatives dans la prédiction et d'autre part, au manque de stabilité de ces modèles.

Le graphique PDP permet quant à lui, d'observer des différences sur les courbes de la prédiction moyenne des bénéficiaires par Classe d'âge pour ces deux modèles. Son analyse permet d'observer des différences dans la manière de segmenter pour les âges entre 2 ans en 30

ans. Toutefois, les résultats de l'analyse des données (3.17) montrent que le modèle GLM est beaucoup plus en phase avec les résultats empiriques obtenus précédemment.

De même que pour les modèles Fréquence coût moyen, le graphique ci-dessous présente les graphiques ICE et PDP pour le modèle ***GLM Tweedie***, ***XGB Tweedie*** et ***XGB MSE***.

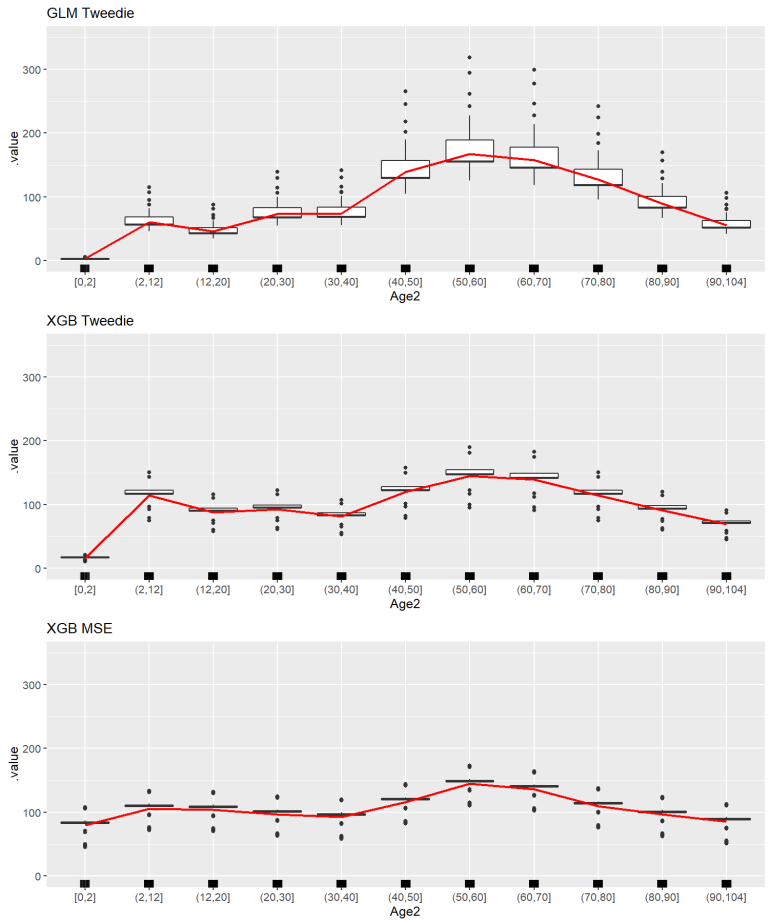


FIGURE 4.8 – Graphes des Dépendances partielles et ICE pour les modèles Tweedie du poste Généraliste

A travers l'analyse du graphique ICE, on se rend compte que les prédictions par Classe d'âge des modèles ***XGB Tweedie*** et ***XGB MSE*** sont fortement homogènes. Ce résultat vient renforcer les résultats observés sur l'importance des variables ainsi que sur les interactions des variables qui montrent que ces deux modèles sont moins adaptés aux données dans la segmentation du poste Optique.

On note par ailleurs que pour le modèle ***GLM Tweedie***, le boxplot représentant la distribution marginale des prédictions des bénéficiaires de la classe d'âge [0, 2] est très en dessous des autres boxplots ce qui est en phase avec les résultats observés aussi bien sur l'analyse

préliminaire des données que sur la prédiction des modèles Fréquence - coût moyen.

D'autre part, l'analyse du graphique PDP permet d'observer que la courbe de la prédiction moyenne des bénéficiaires par Classe d'âge a la même tendance pour ces trois modèles. Cependant, la segmentation est beaucoup plus marquée sur le modèle ***GLM Tweedie***.

En définitive, l'ensemble de ces résultats montrent une segmentation et une adéquation aux données beaucoup plus intéressante sur les modèles GLM. Toutefois, il faut garder à l'esprit que sur ce poste, certaines hypothèses de loi sont très peu réalistes ce qui a tendance à être pénalisant pour les modèles XGBoost.

4.3 Conclusion de l'interprétabilité

Au terme de la mise en œuvre des méthodes d'interprétabilité, plusieurs résultats importants sont à retenir.

Dans un premier temps, cette analyse nous a permis d'avoir un aperçu sur l'importance des variables dans les différents modèles. C'est ainsi qu'on a pu observer que pour les deux postes étudiés, l'âge est la variable la plus utilisée pour la segmentation. Par ailleurs, on a aussi observé l'effet des interactions sur la prédiction dans les différents modèles. Par la suite, l'analyse des graphiques PDP et ICE ont permis d'observer comment les prédictions sont faites sur les différents modèles.

Toutes ces observations ont donc permis de comprendre un peu plus l'adéquation aux données et la qualité de la segmentation des différents modèles construits.

Par ailleurs, on a pu se rendre compte que même si le modèle XGBoost semble théoriquement être plus intéressant, celui-ci est très sensible à l'hypothèse de loi et donc à la fonction de perte utilisée. Ainsi, on a pu observer que celui-ci a une segmentation beaucoup plus intéressante par rapport au GLM sur le poste Généraliste, tandis que l'effet contraire est observé sur le poste Optique où l'hypothèse de Poisson pour la fréquence est beaucoup moins réaliste.

5 Conclusion Générale

L'objectif principal des travaux réalisés dans le cadre de ce mémoire a été la construction de modèles de prédiction efficients pour la modélisation de différents postes de dépense en assurance santé. Pour ce faire, l'approche utilisée a été de construire plusieurs modèles afin de les combiner pour aboutir à un modèle plus optimal.

Les modèles construits dans ce mémoire sont d'une part, différentes variantes des modèles GLM et d'autres part, différentes variantes des modèles de Gradient Boosting basées sur différentes fonctions de perte inspirées des modèles GLM.

Pour analyser les résultats obtenus, l'approche utilisée dans ce mémoire a été de construire un processus d'évaluation afin de donner une vision aussi bien globale que détaillée des performances de chaque modèle construit.

L'analyse des résultats obtenus à travers les différentes évaluations effectuées a permis de mettre en évidence plusieurs résultats importants :

- En premier lieu, l'on note les différences dans les résultats obtenus en fonction des métriques utilisées. Ainsi, il est nécessaire de choisir ou de construire rigoureusement la métrique d'évaluation des modèles que l'on construit, car celle-ci dépend fortement de l'objectif recherché.
- En ayant une vision par profil d'assurés de l'évaluation des modèles construits, l'on a pu observer que pour les deux postes analysés, aucun modèle n'a une supériorité absolue sur l'ensemble de la base de test. En outre, cette approche permet d'avoir une vision de la rentabilité de chaque profil de risque et permet ainsi, de réduire significativement les risques d'antisélection.
- Les résultats obtenus montrent aussi que le modèle optimal construit à partir des différents modèles de prédiction ajustés permet d'améliorer les résultats sur toutes les métriques d'évaluation et se présente ainsi comme une alternative crédible en termes de prédiction et de segmentation.
- L'utilisation des modèles de Gradient Boosting a permis d'avoir un gain significatif en termes de performance des modèles. C'est notamment le cas sur le poste Généraliste où ces modèles ont presque tout le temps été meilleurs que les modèles GLM en termes de performance. En outre, l'utilisation des fonctions de perte spécifiques aux problèmes assurantiels a permis d'avoir un gain significatif en termes de performance notamment sur le poste Optique où la fonction de perte classique s'est avérée peu efficace.

- Aussi, on note la stabilité et la capacité des modèles de Gradient Boosting à pourvoir s'adapter aux données. C'est notamment le cas lors de l'utilisation de la variable continue ***Age***, au détriment de la variable catégorielle ***Classe d'âges*** sur le poste Généraliste. Cela a fortement dégradé le modèle GLM mais cela fut moins le cas pour le modèle Gradient Boosting équivalent.
- Par ailleurs, l'analyse effectuée sur l'interprétabilité du modèle a permis d'avoir une vision plus détaillée sur la manière dont la segmentation est effectuée par nos différents modèles et plus particulièrement par nos modèles de Machine Learning. Cette analyse a aussi permis d'observer de très fortes similarités dans la manière dont les différents modèles effectuent leurs segmentations. Ces résultats permettent ainsi d'interpréter par segment, le modèle optimal qui a été construit.

  On a également pu constater que même si le modèle XGBoost semble théoriquement être plus intéressant, celui-ci est très sensible à l'hypothèse de loi et donc à la fonction de perte utilisée. Ainsi, on a pu observer que celui-ci a une segmentation beaucoup plus intéressante par rapport au GLM sur le poste Généraliste, tandis que l'effet contraire est observé sur le poste Optique où l'hypothèse de Poisson pour la fréquence est beaucoup moins réaliste.

En définitive, tous ces résultats montrent l'efficience du modèle optimal qui a été construit dans ce mémoire. Cependant, il est bien de noter que la base de données utilisée dans le cadre de ce mémoire reste assez restreinte. Par conséquent, pour mesurer efficacement la stabilité des résultats obtenus ainsi que l'apport marginal des méthodes de Gradient Boosting, l'approche utilisée dans ce mémoire doit être appliquée à une base de données plus grande, avec plus de variables explicatives. Ce qui sera en tout état de cause de plus en plus le cas au vu des possibilités qu'offre le Big Data quant à l'utilisation de sources de données externes.

Glossaire

GLM Generalized linear model; modèle linéaire généralisé.

ICE Individual Conditional Expectation.

MAE Mean Average Error.

MSE Mean Squared Error.

O/P Observation / Prédiction.

PDP Partial Dependence Plot.

RMSE Root Mean Square Error .

XGB XGBOOST; Extrem Gradient Boosting .

A Annexes

A.1 Les méthodes agnostiques d'interprétabilité

Les méthodes d'interprétabilité de Machine Learning sont dites agnostiques lorsque ces méthodes permettent de séparer l'interprétabilité du modèle à ce modèle. Ils s'opposent aux méthodes qui sont dites spécifiques et qui sont des outils d'interprétabilité spécifiques à un modèle unique ou à un groupe de modèles. Ces outils dépendent fortement du fonctionnement et des capacités du modèle ou du groupe de modèle spécifique auxquels ils sont associés.

A.1.1 Le graphique de dépendance partielle (Partial Dépendance Plots - PDP)

Le graphique de dépendance partielle (PDP) de **Friedman** [1] est une méthode qui permet de visualiser l'effet marginal qu'une ou deux variables explicatives ont sur les prédictions d'un modèle de Machine Learning. Plus précisément, la méthode PDP est une méthode de représentation de la distribution partielle d'une variable ou deux variables explicatives.

Cette méthode permet de montrer si la relation entre la cible et une caractéristique est linéaire, monotone ou plus complexe. En l'occurrence, lorsque l'on applique cette méthode à un modèle de régression linéaire, le PDP obtenu est une courbe linéaire.

Plus formellement, en notant f la méthode de Machine Learning, x_s la variable explicative pour laquelle l'on souhaite obtenir le PDP et x_c l'ensemble des autres variables explicatives alors :

$$f(x_s) = E_{x_c}[f(x_s, x_c)] = \int f(x_s, x_c) d\mathbb{P}(x_c) \tag{A.1}$$

Cette fonction $f(x_s)$ est estimée en calculant la moyenne ci-dessous, également connue sous le nom de méthode de Monte Carlo, sur les données d'apprentissage :

$$\hat{f}(x_s) = \frac{1}{n} \sum_{i=1}^{n} f(x_s, x_c^i)] \tag{A.2}$$

Conceptuellement, cette méthode consiste à maintenir la variable explicative d'intérêt x_s constante et à trouver les prédictions sur toutes les autres combinaisons de l'ensemble des autres variables explicatives x_c. Par la suite, la méthode considère la moyenne de toutes les prédictions

1. Friedman 2001.

obtenues sur l'ensemble des données d'apprentissage. Après avoir généré des moyennes pour chaque valeur de l'ensemble de définition de x_s, on obtient les graphiques PDP.

A.1.2 Individual Conditional Expectation (ICE) plots

Le graphique ICE est une extension du Partial Dependence Plot (PDP), décrit précédemment. Visuellement, les graphiques ICE désagrègent les résultats des PDP classiques.

Plutôt que de tracer l'effet partiel moyen d'une variables explicatives sur la prédiction, la méthode consiste à tracer les courbes de la dépendance de la prédiction à cette variable explicative pour chaque instance de cette variable séparément. On obtient ainsi un graphique par instance de la variable. Chacun reflète la prédiction en fonction de la variable explicative x_s, conditionnelle à un x_c.

Plus formellement, dans la méthode ICE, pour chaque instance dans $\{(x_s^i, x_c^i)\}_{i=1}^N$, la courbe $\hat{f}_s^i$ est tracée en fonction de x_s^i, avec x_c^i fixé.

A.1.3 Importance des variables par permutation (Permutation feature importance - VIP)

Le principe de la méthode (Permutation feature importance) est de mesurer l'importance d'une variable explicative dans un modèle de Machine Learning en mesurant l'augmentation de l'erreur de prédiction du modèle, après avoir permuté les valeurs de la variable. Cette permutation permet de s'affranchir d'une éventuelle corrélation existante entre la variable explicative et la variable à prédire.

Ainsi, une variable est "importante" si la permutation de ses valeurs augmente l'erreur du modèle. En effet, cela voudrait dire que le modèle s'est appuyé sur la variable pour effectuer la prédiction. De manière analogue, une variable est dite "non importante" si la permutation de ses valeurs laisse l'erreur du modèle inchangée. Dans ce cas, cela voudrait dire que le modèle a ignoré la variable pour effectuer la prédiction.

Cette méthode a été initialement introduite par **Breiman** [2] pour des algorithmes de forêts aléatoires. Poursuivant cette idée, **Fisher**, **Rudin** et **Dominici** [3] ont proposé une version

2. Breiman 2001.
3. Fisher, Rudin et Dominici 2019.

agnostique de cette méthode. L'algorithme de cette version se présente comme suit :

Algorithme 3 : Permutation feature importance

Soient m le modèle de Machine Learning et D la base de données à J colonnes ;
Calculer la perte initiale l du modèle m sur la base de données D à l'aide d'une fonction de perle $\mathcal{L}$;
for $j = 1$ *à* J **do**
  for $k = 1$ *à* K **do**
    Permuter aléatoirement les valeurs de la colonne j de la base de données D pour générer une autre version des données nommée $\widetilde{D}_{k,j}$;
    Calculer le score $s_{k,j}$ du modèle m sur la nouvelle base $\widetilde{D}_{k,j}$;
  end
  Calculer l'importance i_j de la variable explicative f_j définie par :
  $i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j}$;
end

A.1.4 Variable Interaction strength

Les variables explicatives d'un modèle de prédiction ont tendance à collaborer dans la phase de prédiction, ce qui indique la présence d'interactions entre elles. Différentes méthodes de mesure de l'interaction entre les variables explicatives d'un modèle de Machine Learning ont été proposées dont plusieurs sont des méthodes basées sur des outils statistiques.

Une façon d'estimer la force de l'interaction entre des variables est de mesurer dans quelle mesure la variation de la prédiction dépend de l'interaction de ces variables. Cette méthode, appelée H-statistique et introduite par Friedman et Popescu est basée sur la dépendance partielle définie précédemment.

En considérant la variable explicative x_j et x_{-j} l'ensemble des autres variables explicatives en déhors de de x_j et en considerant que x_j et x_{-j} n'ont pas d'interactions, on peut décomposer la dépendance partielle de la manière suivante :

$$f(x) = PD_j(x_j) + PD_k(x_{-j})$$

où f est la dépendance partielle jointe de toutes les variables , $PD_j(x_j)$ la dépendance partielle de la variable x_j et $PD_k(x_{-j})$ la dépendance partielle de toutes les autres variables à l'exception de x_j.

Dans ces conditions, le niveau de la variance expliquée par l'interaction (différence entre la PD observée et la PD sans interaction) est utilisé comme statistique pour mesurer la force de l'interaction. La statistique est de 0 s'il n'y a pas d'interaction du tout et de 1 s'il y'a dépendance totale.

Mathématiquement, la H-statistique proposée par Friedman et Popescu pour l'interaction entre les caractéristiques j et toutes les autres variable est définie par :

$$H_j^2 = \sum_{i=1}^{n} [\hat{f}(x^i) - PD_j(x_j^i) + PD_k(x_{-j}^i)] / \sum_{i=1}^{n} \hat{f}(x^i)$$

Table des figures

Liste des tableaux

Bibliographie

A Colin Cameron et Pravin K Trivedi (1986). « Econometric models based on count data. Comparisons and applications of some estimators and tests ». In : *Journal of applied econometrics* 1.1, p. 29-53.

Aaron Fisher, Cynthia Rudin et Francesca Dominici (2019). « All Models are Wrong, but Many are Useful : Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. » In : *Journal of Machine Learning Research* 20.177, p. 1-81.

Bent Jørgensen (1987). « Exponential dispersion models ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 49.2, p. 127-145.

Diane Lambert (fév. 1992). « Zero-Inflated Poisson Regression, With An Application to Defects in Manufacturing ». In : *Technometrics* 34. doi : 10.1080/00401706.1992.10485228.

Esbjörn Ohlsson et Björn Johansson (2010). *Non-life insurance pricing with generalized linear models*. T. 174. Springer.

Harald Cramér (1946). « Mathematical methods of statistics ». In : *Princeton U. Press, Princeton* 500.

Insee, Tableaux de l'économie française Édition 2020, Dépenses de santé (2018). url : https://www.insee.fr/fr/statistiques/4277750?sommaire=4318291.

J. A. Nelder et R. W. M. Wedderburn (1972). « Generalized Linear Models ». In : *Journal of the Royal Statistical Society. Series A (General)* 135.3, p. 370-384. issn : 00359238. url : http://www.jstor.org/stable/2344614.

Jerome H Friedman (2001). « Greedy function approximation : a gradient boosting machine ». In : *Annals of statistics*, p. 1189-1232.

— (2002). « Stochastic gradient boosting ». In : *Computational statistics & data analysis* 38.4, p. 367-378.

John Mullahy (1986). « Specification and testing of some modified count data models ». In : *Journal of econometrics* 33.3, p. 341-365.

Leo Breiman (2001). « Random forests ». In : *Machine learning* 45.1, p. 5-32.

Michael Rothschild et Joseph Stiglitz (nov. 1976). « Equilibrium in Competitive Insurance Markets : An Essay on the Economics of Imperfect Information* ». In : *The Quarterly Journal of Economics* 90.4, p. 629-649. issn : 0033-5533. doi : 10.2307/1885326. url : https://doi.org/10.2307/1885326.

Roel Henckaerts et al. (avr. 2019). « Boosting insights in insurance tariff plans with tree-based machine learning ». In :

State of Health in the EU, France, Profils de santé par pays (2019). url : https://ec.europa.eu/health/sites/health/files/state/docs/2019_chp_fr_french.pdf.

Yoav Freund, Robert E Schapire et al. (1996). « Experiments with a new boosting algorithm ». In : *icml*. T. 96. Citeseer, p. 148-156.