INSTITUT DES ACTUAIRES

ESSEC BUSINESS SCHOOL

Mémoire présenté le Mardi 17 décembre 2019
en vue de l'obtention du titre d'Actuaire de l'Institut des Actuaires
par **Clara ADICEOM**

suite à son stage effectué dans le cadre de la **filière Actuariat ESSEC-ISUP**

(Intitulé du mémoire)

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

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus Membres présents du jury de l'Institut des Actuaires :

Membres présents du jury de la filière :
- *Marie KRATZ*
-
-

Membres présents du jury de l'Institut des Actuaires
- *Yi RONG*
- Emmanuel DUBREUIL
-
-
-

Entreprise : Axa
Directeur du mémoire en entreprise : Mohammed Halimi

Invité :

Signature :

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

Signature du responsable entreprise :

Clara ADICEOM

Mémoire présenté le : 17 décembre 2019

devant l'ESSEC Business School pour l'obtention du titre d'actuaire à l'Institut des Actuaires

Par : Clara ADICEOM

Titre : Optimisation de la stratégie de majoration des primes de contrats d'assurance habitation au terme.

Confidentialité : ☐ NON OUI ☒ (Durée : ☐ 1 an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Tuteur académique :

Nom : Marie KRATZ

Signature :

.....

Membres présents du jury de l'Institut des Actuaires

.....

.....

.....

Entreprise : AXA France

Signature :

Directeur de mémoire en entreprise :
Nom : Mohamed HALIMI

Signature :

Invité :
Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Résumé

Mots clés

Multirisque habitation, Loi Hamon, Majorations, Optimisation sous contraintes, Lagrangien, Résiliation, Expected Loss Ratio, Crédit commercial, SHAP Values, Modèles Linéaires Généralisés, Gradient Boosting Machine

L'échéance d'un contrat d'assurance habitation est un moment déterminant pour l'assureur : pour amortir ses frais et rentabiliser le contrat, l'assureur pense ne pas avoir d'autres choix que de majorer la prime d'assurance chaque année. Il s'agit d'une étape délicate car cette augmentation peut éventuellement conduire le client à résilier son contrat qu'il va considérer trop cher par rapport à ce que proposerait la concurrence. En outre, la mise en application de la loi Hamon en 2014, facilitant les démarches de résiliation après un an de souscription, accentue le risque de départ du client.

L'objectif devient alors pour l'assureur de limiter au maximum la résiliation des contrats qui lui sont le plus profitables en ajustant le niveau de majoration pour chaque contrat arrivant à leur terme. Autrement dit, comment minimiser le taux de résiliation du portefeuille sous contrainte de la profitabilité totale ?

Ce mémoire s'articule autour de trois étapes. La première consiste à élaborer un modèle de résiliation ainsi qu'un modèle de crédit commercial à l'aide de modèles linéaires généralisés, ce qui permettra d'estimer la ressource espérée de chaque contrat en fonction de la majoration. La deuxième étape se focalisera sur l'optimisation des majorations sous contrainte pour minimiser le taux de résiliation du portefeuille par la méthode du lagrangien. Enfin, la troisième étape s'intéressera à la réalisation d'un reverse engineering par Gradient Boosting Machine pour prédire les majorations optimales obtenues lors de la deuxième étape. Dans ce cadre, un travail d'interprétation des prédictions sera effectué au moyen des SHAP values.

Abstract

Keywords

Household Insurance, Hamon law, Increased premium, Constrained optimization, Lagrangian, Termination, Expected Loss Ratio, Discount, SHAP Values, Generalized Linear Models, Gradient Boosting Machine

The anniversary date of a household insurance contract is a decisive moment for the insurer : in order to amortize the fees and make the contract profitable, the insurer considers he has no choise but increase the insurance premium every year. This is a critical step as this increase in price can potentially encourage the client to put an end to his contract, considering the fact he might find a cheaper deal on the market. Besides, the implementation of the Hamon law in 2014, which makes easier the cancellations' procedures after one year of subscription, will emphasize the risk of leaving for each client.

Then, the insurer's goal is to contain as much as possible the cancellations for the most profitable contracts by ajusting the increase in price for each of them while they are approaching their anniversary date. In other words, how should the insurer minimize the cancellation rate of the portfolio under profitability constraint ?

This study is built around three main steps. The first one involves the creation of cancellation and discount models, using generalized linear models, in order to estimate the earning expectancy of each contract depending on the increase in price. The second step aims to optimize the increases in price under constraint in order to minimize the cancellation rate of the whole portfolio with the Lagrangian method. Finally, the third step will focus on the reverse engineering by Gradient Boosting Machine to predict the optimal increases obtained in the second step. In this framework, an interpretation work will be done using SHAP values.

Synthèse

L'objectif de ce mémoire est de proposer une méthodologie d'optimisation de la stratégie de majoration des primes de contrats d'assurance habitation arrivant à leur terme.

En effet, lors de la souscription d'un nouveau contrat, l'assureur prend à sa charge des frais d'acquisition, de souscription et de gestion afin de pouvoir proposer un tarif compétitif sur le marché très concurrentiel. De ce fait, du point de vue de l'assureur, le contrat n'est pas rentable la première année, mais il le devient sur le long terme puisque chaque année, une majoration est appliquée en fonction notamment de l'historique des sinistres mais aussi de l'inflation, et qui viendra amortir au fur et à mesure les frais attenants au contrat. Or, en 2014, la loi Hamon est mise en application, permettant d'assouplir et de faciliter les démarches de résiliation dès un an de souscription. Ainsi, tout client peut souscrire un contrat à un prix intéressant et l'arrêter au bout d'un an pour en souscrire un nouveau moins cher chez un concurrent, sans que l'assureur ait pu dégager des bénéfices. Il apparaît donc primordial d'étudier les comportements des clients afin d'ajuster le niveau de majoration de leur contrat, de telle sorte qu'ils considèrent la majoration suffisamment raisonnable pour ne pas résilier, mais aussi que l'assureur soit certain de faire des bénéfices sur les années à venir.

L'idée est donc d'optimiser les niveaux de majoration pour chaque contrat arrivant à leur terme, de telle sorte que le taux de résiliation moyen du portefeuille soit minimal, sous contrainte de rentabilité globale, cette dernière étant évaluée par l'Expected Loss Ratio (ELR).

Pour ce faire, nous disposons des données de la branche Multirisques Habitation (MRH) d'AXA France sur les années 2017 et 2018 afin de constituer une base de travail. L'horizon d'étude ne doit pas être trop lointain, car le comportement des clients évolue sensiblement dans le temps. Une base de données a ainsi été construite par "image de terme", c'est-à-dire que chaque ligne de cette base est associée à un terme (et donc à une année) concernant un contrat, lui-même rattaché à un client. Une telle architecture permet d'étudier l'évolution des contrats, et notamment leur résiliation tarifaire au terme, année après année. Cette base nous renseigne sur les caractéristiques de chaque contrat et du client associé. Elle indique également les primes, les majorations et les rabais appliqués.

Afin de répondre à notre objectif d'optimisation, l'étude se décompose en 3 phases. La première consiste à élaborer d'une part un modèle prédisant la probabilité de résiliation d'un contrat donné, et d'autre part un modèle prédisant l'espérance du pourcentage de crédit commercial (c'est-à-dire de rabais) qui serait appliqué à la prime du contrat. Ainsi, combinés au modèle interne de prime pure, ces deux modèles vont permettre d'estimer la ressource espérée pour chaque contrat en fonction de la majoration qui leur est appliquée. De cette manière, il sera possible, en fonction des majorations de tous les contrats, d'évaluer le taux de résiliation moyen du portefeuille ainsi que son ELR. Ce qui nous amène à la deuxième phase : l'optimisation des majorations pour minimiser le taux de résiliation du portefeuille sous contrainte de l'ELR. Enfin, la troisième et dernière phase répond au besoin opérationnel d'automatiser et rationaliser le calcul des majorations optimales pour chaque

contrat, en réalisant un reverse engineering. Et pour expliquer les prédictions obtenues, un travail d'interprétation est effectué au moyen de la théorie de SHAP values.

Le plan d'action énoncé, que nous détaillerons par la suite, est résumé en figure 1.

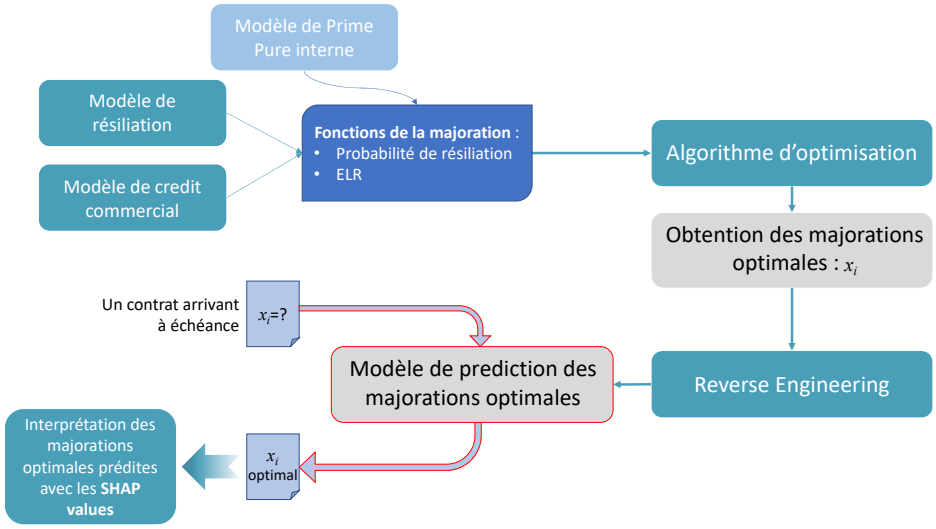


FIGURE 1 – Schéma de la méthodologie d'optimisation

1ère phase : Modélisation de la probabilité de résiliation et de l'espérance de pourcentage de crédit commercial

1.A. Modélisation de la probabilité de résiliation

L'objectif de cette étape est de prédire la probabilité qu'un client résilie son contrat à cause d'une augmentation de prime au terme. Deux approches sont envisagées : une régression logistique pénalisée (Generalized Linear Model, GLM), qui a l'avantage d'être un modèle interprétable, ou bien une modélisation par Gradient Boosting Machine (GBM). Dans les deux cas, une recherche sur grille a été réalisée pour déterminer dans le premier cas les variables explicatives à retenir dans le modèle pour maximiser l'indice de Gini tout en gardant la variable de majoration prédominante dans le modèle, et dans le second cas les paramètres du GBM qui maximisent l'AUC. Le modèle GLM et le modèle GBM choisis indiquent que la variable explicative associée à la majoration est significative dans la prédiction, ce qui est une bonne chose puisque par la suite, l'optimisation sera effectuée en ajustant le niveau de majoration de chaque contrat.

En comparant les résultats sur la base d'entraînement et la base de test des meilleurs modèles de résiliation GLM et GBM identifiés, il apparaît que le GBM soit plus performant que le GLM au sens de l'indice de Gini, mais il est moins stable. En effet, du fait de sa complexité, le GBM a tendance à réaliser un sur-apprentissage. Ainsi, nous optons pour le modèle de résiliation obtenu par régression logistique, plus stable, que nous garderons pour la suite de l'étude.

Toutefois, un point d'attention est à porter sur l'analyse de l'élasticité-majoration de la résiliation : puisque AXA France n'effectue pas de price-test au terme, il n'a pas encore été possible de comparer l'élasticité du modèle avec ce qui est observé dans la réalité. Cela devra faire l'objet d'un travail complémentaire.

1.B. Modélisation de l'espérance du pourcentage de crédit commercial

L'objectif de cette phase est de déterminer, pour un contrat donné, le pourcentage de crédit commercial par rapport à la prime commerciale après majoration qui serait accordé à un client. Une telle modélisation se fait en deux étapes :

— 1ère étape : Modélisation de la probabilité d'application du crédit commercial
A nouveau, une régression logistique pénalisée est réalisée, et le meilleur modèle est déterminé par une recherche sur grille. Par ailleurs, la variable explicative associée à la majoration est toujours significative.

— 2ème étape : Modélisation du pourcentage de crédit commercial
Dans le cas où un crédit commercial est appliqué, il est nécessaire de déterminer ce que ce rabais représente en terme de pourcentage de la prime commerciale post majoration. Pour cela, nous réalisons une première étude de la distribution des pourcentages de crédit commerciaux afin de déterminer si cette dernière suit une loi appartenant à une famille exponentielle et ainsi déterminer si une modélisation GLM est envisageable. Nous choisissons alors de faire une forte approximation en considérant que les pourcentages suivent une loi log-normale. Cela permet d'effectuer une régression linéaire sur le logarithme du pourcentage et remonter ainsi à la prédiction du pourcentage de crédit commercial. Au regard des prédictions obtenues, les performances du modèle sont suffisamment satisfaisantes pour considérer que l'approximation est acceptable. Un point d'attention est à porter toutefois sur les pourcentages des crédit commerciaux importants, qui ne peuvent pas être modélisés de cette manière.

Le produit des deux prédictions nous permet d'obtenir l'espérance du pourcentage de crédit commercial qui serait appliqué à un contrat en fonction de ses caractéristiques et celle du client.

2ème phase : Optimisation des majorations

Le modèle de prime pure ainsi que les modèles de résiliation et de crédit commercial développés en phase 1 permettent de déterminer pour le portefeuille le taux de résiliation moyen ainsi que l'ELR total en fonction des différentes majorations appliquées aux primes des contrats qui le composent. Cela permet de poser le problème d'optimisation suivant : minimiser le taux de résiliation moyen du portefeuille sous contrainte de l'ELR.

Pour le résoudre, nous avons recours à la théorie du Lagrangien. Or deux difficultés se posent :

— Première difficulté : le grand volume de majorations à optimiser, ce qui rend la résolution du problème d'optimisation complexe et lourde en terme de calculs.
Pour y remédier, le problème sur le portefeuille est reformulé de telle sorte à obtenir plusieurs problèmes d'optimisation individuels, sous une contrainte différente et sous des hypothèses fortes d'indépendance entre les comportements des clients, aboutissant ainsi à autant de Lagrangiens que de majorations à optimiser. C'est alors le multiplicateur de Lagrange, commun à tous ces problèmes d'optimisation individuels qui relie ces derniers au problème global.

— Seconde difficulté : l'absence de convexité des Lagrangiens correspondants aux problèmes individuels en fonction de la majoration. Cela a pour conséquence de proposer uniquement des solutions en coin.
Pour pallier ce problème, un argument a été rajouté à chaque Langrangien afin de rajouter une légère convexité.

Par ailleurs, l'absence de convexité de la probabilité de résiliation en fonction de la majoration induit que le problème dual n'est pas équivalent au problème primal. Ainsi, la résolution du problème primal nous donne accès à des optimums locaux suivant la valeur du multiplicateur de Lagrange choisi, mais pas à un optimum global.

En faisant varier la valeur du multiplicateur de Lagrange, nous avons alors la possibilité de tracer la frontière efficiente du taux de résiliation en fonction de l'ELR du portefeuille, et de positionner la stratégie actuelle d'AXA

France par rapport à cette frontière caractérisant l'ensemble des stratégies optimales atteignables. Cette frontière est illustrée en figure 2.

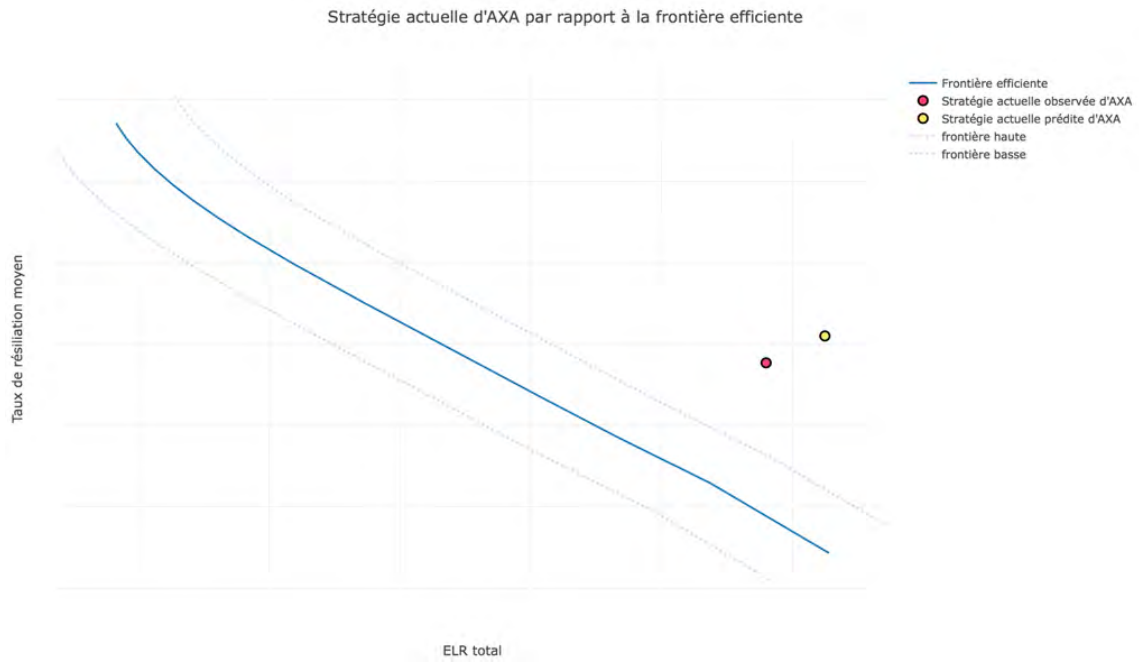


FIGURE 2 – Stratégie actuelle d'AXA France par rapport à la frontière efficiente

Au regard de cette figure, une marge de progression de la politique de majoration de l'assureur est possible pour rejoindre un point de cette frontière. Mais finalement, le choix du meilleur multiplicateur de Lagrange, et donc la détermination des meilleures majorations optimales se fait en vérifiant que les KPIs (que sont le taux de résiliation du portefeuille, l'ELR total et le chiffre d'affaires) répondent bien au cahier des charges : le taux de résiliation du portefeuille doit diminuer par rapport à la situation actuelle, l'ELR ne doit pas augmenter, et le chiffre d'affaires doit au moins être maintenu.

Les majorations optimales ainsi obtenues sont en moyenne inférieures aux majorations appliquées actuellement sur les contrats, et certaines majorations optimales sont même négatives, ce qui contribuerait à embellir l'image de marque de l'assureur. Ces minorations concernent principalement les contrats rentables dont les détenteurs sont sensibles à la majoration, et qui présentent un historique de sinistres léger. Il apparaît donc que l'optimisation propose une autre vision de l'application de la majoration, plus sensible à l'ELR des contrats, davantage orientée vers le devenir des contrats.

3ème phase : Reverse engineering

Dans un souci opérationnel, effectuer tous les mois le travail d'optimisation des majorations n'est pas raisonnable : cela serait trop complexe et trop chronophage. C'est la raison pour laquelle nous réalisons un reverse engineering qui va permettre de développer un modèle de prédiction des majorations optimales en fonction des variables explicatives dont les actuaires disposent au moment du calcul de la majoration.

En étudiant la distribution des majorations optimales, on remarque que cette dernière ne suit pas une loi appartenant à une famille exponentielle. Il est alors exclu de développer un modèle GLM. C'est pourquoi nous optons pour un modèle GBM, dont les paramètres ont à nouveau été déterminés grâce à une recherche sur grille, et dont les performances de prédiction se sont révélées satisfaisantes.

Toutefois, à la différence d'un modèle GLM totalement transparent et interprétable grâce à une formule générique, le GBM lui constitue une "boîte noire". Il s'agit en effet d'un modèle complexe, non auditable tel quel, ce qui signifie que nous ne pouvons pas expliquer directement la valeur de la prédiction. Cela est regrettable car une telle opacité empêcherait l'assureur d'utiliser ce modèle.

Cependant, Lundberg et Lee [6] proposent une solution pour expliquer de tels modèles complexes, directement inspirée de la théorie des jeux : les SHAP values (SHapley Additive exPlanations, ou l'explication incrémentale par les valeurs de Shapley). Les SHAP values permettent alors, pour une prédiction de majoration optimale donnée, d'identifier les variables qui auront contribué à l'augmentation ou à la diminution de la valeur prédite, et dans quelle proportion par rapport à la moyenne des prédictions.

Bilan

Finalement, ce mémoire montre qu'il est possible d'optimiser la stratégie de majoration des primes de contrats d'assurance habitation au terme. L'étude a d'ailleurs montré des résultats très encourageants : une telle optimisation diminuerait le taux de résiliation des contrats tout en améliorant en théorie la rentabilité du portefeuille ainsi que l'image de marque de l'assureur sans pour autant dégrader le chiffre d'affaires. Bien sûr, l'atteinte exacte de la frontière efficiente en pratique reste utopique, car l'assureur ne maîtrise pas un certain nombre de facteurs, et notamment le comportement des clients en terme de résiliation, ainsi que celui des agents généraux qui appliquent les crédits commerciaux. Toutefois, cette méthode d'optimisation permettrait de rendre davantage efficace la stratégie de l'assureur en la rapprochant au maximum de la frontière efficiente.

Synthesis

The objective of this thesis is to propose a methodology for optimizing the strategy of increasing the premiums of expiring home insurance contracts.

Indeed, when a new contract is subscribed, the insurer pays acquisition, subscription and management fees in order to be able to offer a competitive rate in the highly competitive market. As a result, from the insurer's point of view, the contract is not profitable in the first year, but it becomes profitable in the long term since each year, an increase in price is applied depending on the history of claims but also on inflation, and which will gradually absorb the costs associated with the contract. However, in 2014, the Hamon law is applied, making it possible to make the termination procedures more flexible and easier from one year of subscription. Thus, any customer can take out a contract at an attractive price and stop it after one year to subscribe a new one at a lower price with a competitor, with no chance for the insurer to make a profit. It therefore seems essential to study the customers' behaviour in order to adjust the level of increase in price of their contract, so that they consider the increase reasonable enough not to terminate, but also that the insurer is certain to make a profit in the coming years.

The idea is therefore to optimize the levels of increase in price for each contract at their anniversary date, so that the average cancellation rate of the portfolio is minimum, under the constraint of overall profitability, which is being evaluated by the Expected Loss Ratio (ELR).

To accomplish this, we have access to the data from AXA France's household insurance business for the years 2017 and 2018 in order to build a database. The study horizon should not be too distant, as customer behaviour changes significantly over time. A database has thus been built by "term image", i.e. each line of this database is associated with a term (and therefore a year) concerning a contract, itself attached to a customer. Such an architecture makes it possible to study the evolution of contracts, and in particular their cancellation at the end of the contract, year after year. This database gives us information on the characteristics of each contract and the associated customer. It also indicates the premiums, increases and discounts applied.

In order to meet our optimization goal, the study is divided into 3 stages. The first one is to develop a model that predicts the probability of termination of a given contract, and a model that predicts the expectation of the percentage of commercial credit (i.e. discounts) that would be applied to the premium of the contract. Thus, combined with the internal pure premium model, these two models will enable us to estimate the expected resource for each contract according to the increase of premium applied to them. This way, it will be possible, depending on the increases in all contracts, to assess the average termination rate of the portfolio and its ELR. This brings us to the second stage : the optimization of increase in price to minimize the cancellation rate of the portfolio under the ELR constraint. Finally, the third and final stage responds to the operational need to automatize and rationalize the calculation of optimal increases for each contract, by carrying out reverse engineering. And to explain the predictions obtained, an interpretation work is performed using the SHAP values theory.

The stated action plan, which we will detail later, is summarized in figure 3.

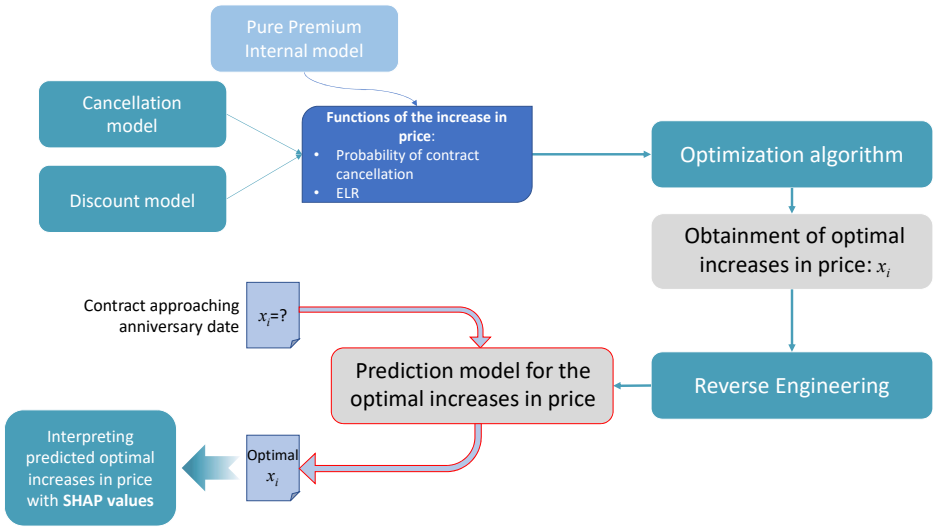


FIGURE 3 – Scheme of the optimization methodology

1st stage : Modelling the probability of cancellation and the expected percentage of commercial credit

1.A. Modelling the probability of cancellation

The objective of this step is to predict the probability that a client will terminate his contract due to a premium increase at the anniversary date of the contract. Two approaches are considered : a penalized logistic regression (Generalized Linear Model, GLM), which has the advantage of being an interpretable model, or a Gradient Boosting Machine (GBM) model. In both cases, a grid search was conducted to determine in the first case the explanatory variables to be used in the model in order to maximize the Gini index while keeping the premium variable predominant in the model, and in the second case the GBM parameters that maximize the AUC. The GLM and GBM models selected indicate that the explanatory variable associated with the increase in price is significant in the prediction, which is a good thing since the optimization will be performed by adjusting the level of the increase in price for each contract.

By comparing the results on the training and the test databases of the best identified GLM and GBM termination models, it appears that GBM is more efficient than GLM in the sense of the Gini index, but it is less stable. Indeed, due to its complexity, the GBM tends to over-learn. Thus, we are opting for the termination model obtained by logistic regression, which is more stable and which we will keep for the rest of the study.

However, one point of attention should be paid to the analysis of the elasticity-increase in price of termination : since AXA France does not carry out a price-test at the end of the term, it has not yet been possible to compare the elasticity of the model with what is observed in reality. This will require further work.

1.B. Modeling the expectation of the percentage of discount

The purpose of this stage is to determine, for a given contract, the percentage of discount in relation to the commercial premium after the increase that would be allowed to a customer. Such modelling is done in two steps :

— 1st step : Modelling the probability of discount application.
Again, a penalized logistic regression is performed, and the best model is determined by a grid search. Furthermore, the explanatory variable associated with the increase in price is still significant.

— 2nd step : Modeling the percentage of discount
In the case where a discount is applied, it is necessary to determine what this discount represents in terms of percentage of the post-increase commercial premium. To do this, we carry out a first study of the distribution of commercial credit percentages in order to determine if it follows a law belonging to an exponential family and thus determine if the development of a GLM model is possible. We then choose to make a strong approximation by considering that the percentages follow a log-normal law. This allows a linear regression on the logarithm of the percentage and thus it is possible to go back to the prediction of the percentage of commercial credit. Considering the predictions obtained, the model's performance is sufficiently convincing to consider the approximation to be acceptable. However, a point of attention should be paid to the percentages of large commercial credits, which cannot be modelled in this way.

The product of the two predictions allows us to obtain the expectation of the percentage of commercial credit that would be applied to a contract based on its characteristics and those of the customer.

2nd stage : Optimization of increases in price

The pure premium model as well as the termination and commercial credit models developed in stage 1 enable us to determine the average cancellation rate and the total ELR for the portfolio based on the various increases applied to the premiums of the contracts that compose it. This raises the following optimization problem : minimizing the average cancellation rate of the portfolio under constraint of the ELR.

To solve it, we use the Langrangian theory. However, two difficulties arise :

— The first difficulty is the large volume of increases to be optimized, which makes the resolution of the optimization problem complex and computationally heavy.
To overcome this, the problem on the portfolio is reformulated in such a way as to obtain several individual optimization problems, under a different constraint and under strong assumptions of independence between client behaviours, thus resulting in as many Lagrangians as increases to be optimized. It is then the Lagrange multiplier, common to all these individual optimization problems, that links them to the global problem.

— The second difficulty is the lack of convexity of the Lagrangians corresponding to the individual problems according to the increase. As a result, only edge solutions are available.
To overcome this problem, an argument was added to each Langrangian to add a slight convexity.

Moreover, the lack of convexity of the probability of cancellation as a function of the increase in price implies that the dual problem is not equivalent to the primal problem. Thus, solving the primal problem gives us access to local optimums according to the value of the Lagrange multiplier chosen, but not to a global optimum.

By shifting the value of the Lagrange multiplier, we then have the opportunity to plot the efficient frontier of the cancellation rate according to the ELR of the portfolio, and to situate AXA France's current strategy compared to this frontier, which characterizes all the optimal strategies that can be achieved. This boundary is illustrated in figure 4.

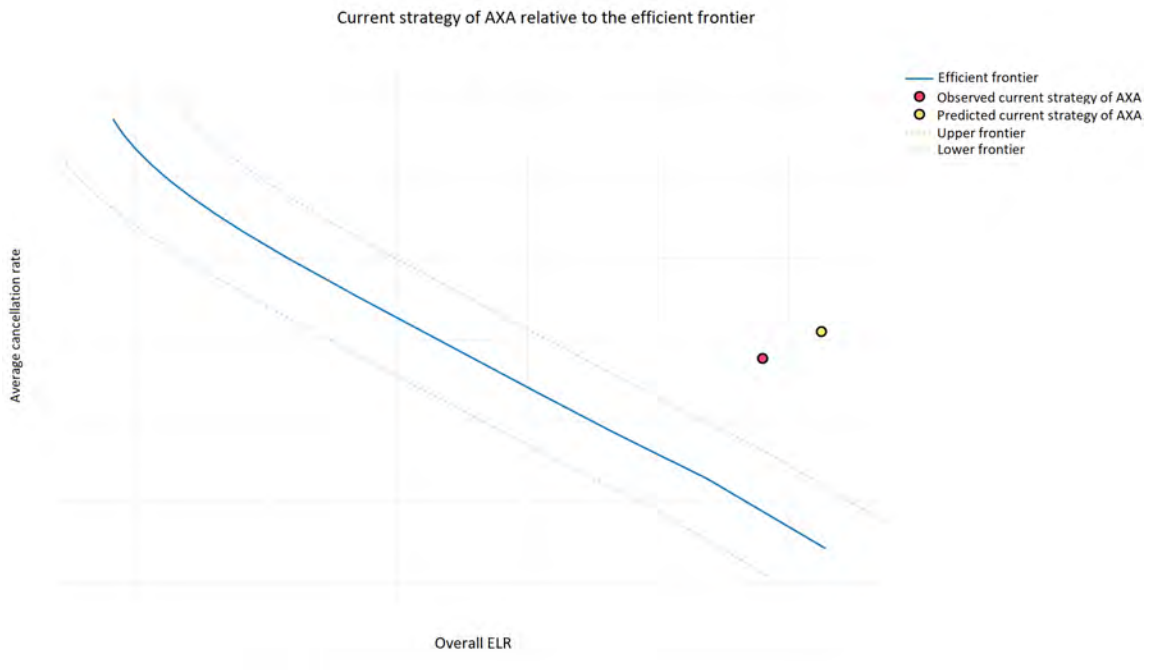


FIGURE 4 – AXA France's current strategy compared to the efficient frontier

Looking at this figure, we can see a range of progress in the insurer's increase policy to reach a point on this frontier. But finally, the choice of the best Lagrange multiplier, and therefore the determination of the best optimal increases in price, is made by checking that the KPIs (which are the portfolio cancellation rate, the total ELR and the revenues) meet the specifications : the portfolio cancellation rate must decrease compared to the current situation, the ELR must not increase and the revenues must at least be maintained.

The optimal increases obtained are on average lower than the increases currently applied to contracts, and some optimal increases are even negative, which would help to enhance the insurer's brand image. These decreases in price mainly affect profitable contracts whose holders are sensitive to the increase and which have a slight claims history. It therefore appears that optimization offers another vision of the application of the increase in price, more sensitive to the ELR, more oriented towards the future of the contracts.

3rd stage : Reverse engineering

From an operational point of view, it is not reasonable to carry out the work of optimising the increases in price every month : this would be too complex and time-consuming. This is why we are carrying out reverse engineering, which will enable us to develop a model for predicting optimal increases based on the explanatory variables available to actuaries at the time of calculating the increases in price.

By studying the distribution of optimal increases, we notice that it does not follow a law belonging to an exponential family. It is then not possible to develop a GLM model. This is why we opt for a GBM model, whose parameters have again been determined by grid search, and whose prediction performance has been found to be satisfactory.

However, unlike a fully transparent GLM model that can be interpreted using a standard formula, the GBM is a "black box". It is indeed a complex model, not verifiable as such, which means that we cannot directly explain the value of the prediction. This is unfortunate because such opacity would prevent the insurer from

using this model.

However, Lundberg and Lee suggest a solution to explain such complex models, directly inspired by the game theory : the SHAP values (SHapley Additive exPlanations, or the incremental explanation by Shapley values). SHAP values then allow, for a given optimal increase prediction, the variables that will have contributed to the increase or decrease of the predicted value, and in which proportion compared to the average of the predictions.

Conclusion

Finally, this paper shows that it is possible to optimize the strategy of increasing premiums for household insurance contracts at their anniversary date. The study showed very encouraging results : such an optimization would reduce the cancellation rate of contracts while theoretically improving the profitability of the portfolio as well as the insurer's brand image without degrading the revenues. Of course, the exact reach of the efficient frontier in practice remains an utopian goal, because the insurer does not control a certain number of factors, in particular the behaviour of customers in terms of cancellation, as well as the behaviour of the general agents who apply trade credits. However, this optimization method would make the insurer's strategy more effective by bringing it as close as possible to the efficient frontier.

Remerciements

Certains souriront en lisant cette page, car ils se souviendront à quel point leur aide m'a été précieuse.

Dans le cadre de la rédaction de ce mémoire, j'ai été accompagnée par des personnes brillantes, qui ont su me challenger tout au long de l'étude, mais aussi me rassurer et me soutenir pendant mes moments de doute, et je n'aurai pas la prétention de dire que j'y serais arrivée seule.

C'est pourquoi je tiens à remercier Thomas GAUTHRON, responsable politique technique IARD Dommages et Responsabilité Civile Particuliers d'AXA France, de m'avoir accueillie au sein de son équipe et pour sa bienveillance.

Je remercie naturellement Anne-Laure LE GALLO, responsable de l'équipe actuariat Multirisques Habitation et Plaisance, qui a su me faire confiance et m'a suggéré cet inspirant sujet de mémoire.

Un grand merci à Mohamed HALIMI, mon maître de stage, pour sa disponibilité, nos discussions éclairantes, son expertise sur l'ensemble des sujets techniques que j'ai pu aborder, pour n'avoir jamais cessé de m'encourager et avec qui j'ai eu tant plaisir à travailler.

Je remercie Charles PARTINGTON, pour son aide dans la construction de la base de données et son regard expert et critique qui m'a permis de préciser davantage les hypothèses et les conclusions de ce mémoire.

Je tiens également à remercier Chae-In KIM pour ces précieux conseils et ses idées aussi originales qu'astucieuses, ainsi que Trieu LE QUOC pour ses travaux d'analyse de marché.

Mes remerciements à Joanna CHARDON, responsable de l'équipe "Center of Pricing Excellence" d'AXA France, ainsi qu'à Clément LADIER et Serge BEN YAMIN, pour leur expertise et leur éclairage notamment sur la théorie de l'optimisation.

Je souhaite aussi remercier Yu TANG et Sébastien LE GALLO pour leur relecture en tant que spécialistes de l'optimisation.

Je remercie l'équipe pédagogique de la filière actuariat de l'ESSEC, et plus particulièrement ma tutrice académique Marie KRATZ pour son accompagnement dans ma réflexion.

Enfin, un grand merci à mes parents pour leur soutien inconditionnel depuis 1993, à mes soeurs et à mes proches qui n'ont jamais douté de moi.

Table des matières

Introduction

Les contrats d'assurance habitation représentent une part importante du marché de l'assurance IARD (Incendie, Accidents, Risques Divers), et comptent souvent, avec l'assurance automobile, parmi les premiers contrats souscrits par les particuliers chez un assureur avant d'investir dans d'autres produits comme la santé, ou la prévoyance. Cela fait de **l'assurance Multirisque Habitation (MRH) un pilier stratégique dans la conquête de nouveaux clients**. Il est connu que le marché IARD est fortement concurrentiel : les assureurs redoublent d'efforts pour proposer des primes d'affaires nouvelles attractives, et doivent surtout faire **preuve d'ingéniosité pour parvenir à garder leurs assurés en portefeuille**. Car l'assurance MRH, par nature, possède un système de "fidélisation" particulier, et tout assuré qui se serait intéressé à cet aspect ne manquera pas de le pointer du doigt : la prime d'assurance croît avec l'ancienneté du contrat, et ce, même en absence de sinistre. En effet, chaque année, au terme du contrat, une majoration est appliquée à la prime commerciale. Elle est propre à chaque contrat et se justifie notamment par l'évolution du risque, de l'inflation, mais surtout elle permet d'amortir des frais d'acquisition, de souscription et de gestion que l'assureur prend en charge la première année afin de proposer un prix compétitif et contracter une affaire nouvelle. Ainsi, mécaniquement, un contrat MRH n'est pas rémunérateur la première année, c'est pourquoi sa rentabilité doit se construire sur le long terme.

Afin de fluidifier le marché de l'assurance et de faciliter les démarches de résiliations des assurés dès un an de souscription, la loi Hamon (aussi connue sous le nom de "loi Consommation") mise en place en 2014, a eu un impact significatif sur les comportements des clients, amplifiant de ce fait les taux de résiliations. Il devient donc impératif pour les assureurs de réagir et piloter autrement leur portefeuille.

C'est le cas d'AXA France, pour qui le portefeuille MRH indique aujourd'hui plus de résiliations que d'affaires nouvelles. Or, le départ d'un client en assurance habitation risque d'engendrer la rupture d'autres contrats AXA si ce dernier est multidétenteur. Par conséquent, contenir l'hémorragie du portefeuille MRH apparaît être un enjeu majeur pour l'assureur.

L'objectif de ce mémoire réalisé au sein d'AXA France est, dans un premier temps, d'**évaluer la stratégie actuelle de l'entreprise concernant la majoration des primes d'assurance au terme, et dans un second temps, de proposer une méthodologie d'optimisation de cette stratégie afin de contenir les résiliations des contrats les plus rentables au sens de l'Expected Loss Ratio (ELR)**. En d'autres termes, **comment minimiser le taux de résiliation du portefeuille sous contrainte de la profitabilité totale en jouant sur les niveaux de majorations ?**

Cette étude s'articule autour de six chapitres. Le premier présente plus en détail le contexte de l'étude. Il explique la nécessité de procéder à un travail d'optimisation sur les majorations appliquées au terme et expose le plan d'action préconisé pour le réaliser.

Le deuxième chapitre s'intéresse à la modélisation de la probabilité de résiliation tarifaire au terme pour

chaque contrat. Après avoir construit une base de données adéquate, les performances d'un modèle linéaire généralisé et d'un Gradient Boosting Machiche seront comparées afin de choisir le modèle le plus pertinent et le plus robuste.

Ensuite, le chapitre 3 se focalise sur une modélisation inédite : celle du crédit commercial, c'est-à-dire le rabais que les agents généraux se permettent d'appliquer aux primes d'assurance afin de retenir le client si ce dernier manifeste un désaccord sur le prix. Cela sera réalisé au moyen de modèles linaires généralisés.

Sur la base des modèles construits aux chapitres 2 et 3, la ressource espérée des contrats pourra être estimée en fonction de la majoration qui leur sera appliquée. Ces estimations sont ainsi exploitées dans le quatrième chapitre qui détaille la méthodologie adoptée pour l'optimisation sous contraintes afin de déterminer les majorations qui minimiseront le taux de résiliation du portefeuille tout en s'assurant que ce dernier reste rentable et satisfait un chiffre d'affaires suffisant. Pour ce faire, la théorie du Lagrangien est mise à profit et adaptée au problème. Dans le cadre de ce chapitre, la notion de frontière efficiente, inspirée du mémoire d'actuariat d'A. De Larrard (2016) [3], est abordée et permet de positionner la stratégie actuelle d'AXA par rapport à l'ensemble des stratégies optimales accessibles, et ainsi d'évaluer le manque à gagner par le travail d'optimisation. L'idée de ce chapitre est de proposer une autre manière d'appréhender le calcul de la majoration au terme, davantage orientée vers le devenir des contrats plus que sur leur historique.

Le cinquième chapitre traite du reverse engineering, l'objectif étant de pouvoir prédire plus simplement tous les mois les majorations optimales obtenues au chapitre 4 en vue d'une implémentation opérationnelle. Ce processus de rétro-ingénierie est effectué au moyen d'un Gradient Boosting Machine, et les prédictions obtenues seront interprétées grâce aux SHAP values, concept développé par Lundberg et Lee (2017) [7] directement inspiré de la théorie des jeux en économie.

Le sixième et dernier chapitre aborde les voies d'améliorations possibles en vue d'une application concrète de cette étude au niveau opérationnel.

Enfin, le lecteur notera que, pour des raisons confidentialité, certaines variables demeureront anonymes et les échelles de certains axes de graphiques seront masquées dans l'ensemble du mémoire.

Chapitre 1

Contexte : stratégie d'optimisation de la majoration des contrats Multirisques Habitation au terme

Sommaire

Préambule

Ce premier chapitre pose le cadre de l'étude. Il explique pourquoi l'optimisation des majorations est aujourd'hui nécessaire compte tenu du contexte du marché assurantiel et de la performance d'AXA France. Après avoir expliqué et évalué la stratégie actuelle de majoration au terme d'AXA, nous signalerons ses limites et proposerons une nouvelle approche rationnelle de calcul de ces majorations : il s'agira alors de définir ce que nous souhaitons maximiser ou minimiser au moyen de l'optimisation des majorations pour chaque contrat arrivant à échéance, et sous quelles contraintes. Sur la base de ce nouvel objectif, nous détaillerons quelle stratégie nous déciderons d'adopter, quels moyens nous allons mettre en place pour la réaliser, et sur quels types de contrats et de clients nous allons focaliser l'étude dans un premier temps.

1.1 AXA France et le contexte assurantiel

AXA France, entité du Groupe AXA, première marque mondiale d'assurance, est agréée pour exercer des activités d'assurance de dommages, destinées aux particuliers comme aux professionnels. L'entreprise propose un large panel de produits d'assurance IARD [1], dont les principaux sont l'assurance automobile des particuliers ou entreprises, l'assurance multirisques dommages aux biens (habitation, agricole, professionnelle, immeuble, entreprise), l'assurance de responsabilité civile, l'assurance de construction ou encore l'assurance santé (sur le périmètre des frais de soin, l'hospitalisation, l'optique et le dentaire).

1. Incendie, Accidents et Risques Divers

AXA France présente en 2018 un chiffre d'affaires total de 6 003 millions d'euros [2], pour un résultat net de 869 millions d'euros. Il faut noter que ce chiffre d'affaires est en légère baisse par rapport à 2017, qui est notamment due à une diminution du volume de contrats en assurance Auto et Multirisques Habitation (MRH).

Cette évolution n'est pas anodine et pourrait s'expliquer en partie par l'arrivée de la loi Hamon [3] en mars 2014. Alors qu'auparavant l'assuré ne pouvait résilier son assurance que dans les deux mois qui précédaient l'échéance de son contrat (autrement ce dernier était renouvelé tacitement), la loi Hamon permet aux souscripteurs de pouvoir résilier à n'importe quel moment leur contrat d'assurance automobile ou habitation à partir d'un an. Ainsi, si nous nous référons aux volumes d'affaires nouvelles et des résiliations en assurance habitation chez AXA France comme le montre la figure 1.1, nous constatons bien une inversion de tendance de la courbe des résiliations relativement peu de temps après ce changement de législation. A partir de juillet 2015, le volume des résiliations dépasse le volume des souscriptions, appelées "Affaires nouvelles", puis l'écart se creuse nettement entre les deux courbes.

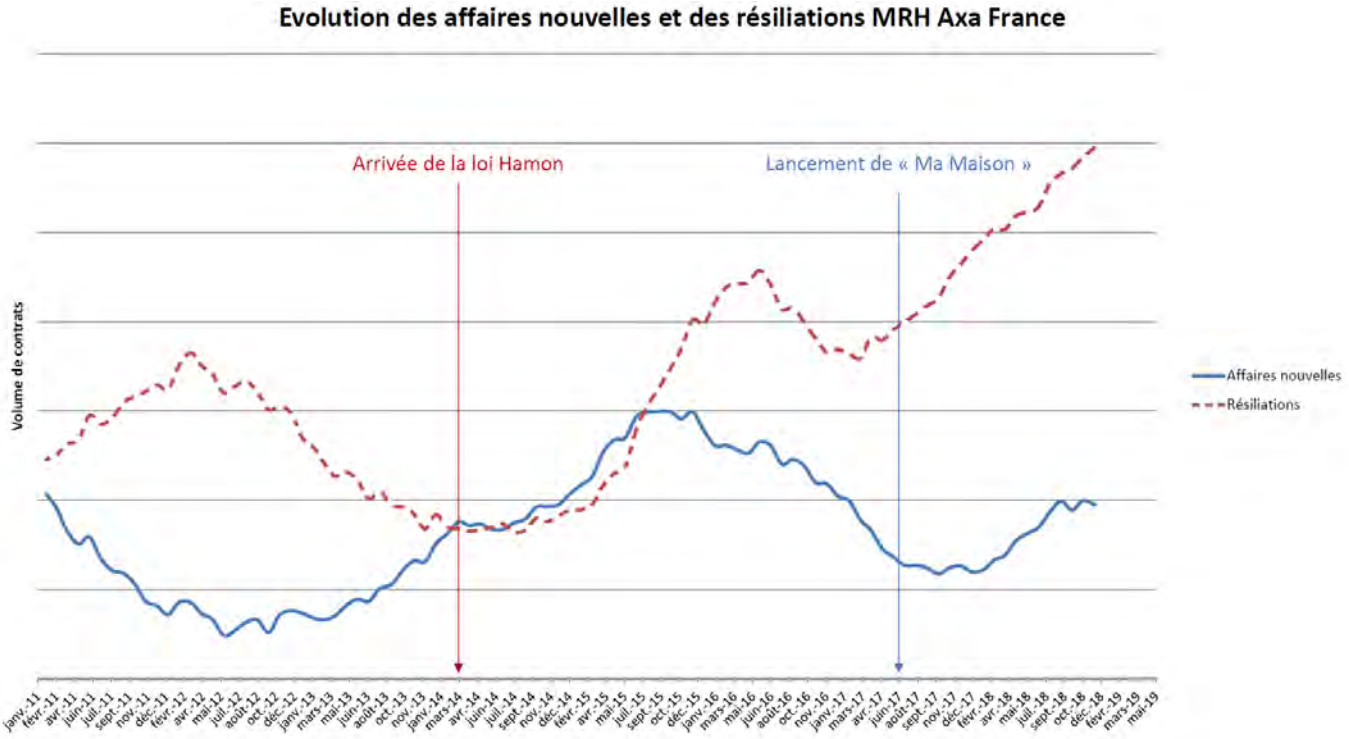


FIGURE 1.1 – Evolution depuis 2011 des volumes d'affaires nouvelles et des résiliations MRH d'AXA France

A chaque échéance de contrat, le calcul de la nouvelle prime doit donc prendre en compte la possible résiliation facilitée par la loi Hamon, au regard de la forte concurrence sur le marché IARD.

En outre, un autre évènement notable survenu ces dernières années est l'arrivée sur le marché IARD de "Ma Maison", le nouveau produit d'assurance habitation d'AXA France, dont la quasi totalité du périmètre des biens assurables de cette branche est couvert dès juin 2017 (d'ailleurs, quelques mois plus tard s'en suit l'apparition de "Mon Auto", pour l'assurance automobile). L'idée de ce nouveau produit est de proposer des garanties de base (socle) auxquelles peuvent s'ajouter en option un certain nombre de garanties selon le souhait du souscripteur. De ce fait, la prime est mieux ajustée au risque associé au bien assuré. Avec ce nouveau produit MRH, le volume d'affaires nouvelles augmente à nouveau et vient stabiliser pour un temps l'apport net [4] négatif, sans pour autant arriver à réduire cet écart.

2. D'après le *Rapport sur la Solvabilité et la Situation Financière 2018* [1]
3. Cf. Annexe 2
4. c'est la différence entre le nombre d'affaires nouvelles et le nombre de résiliations

Une question se pose alors : quelles stratégies adopter pour réduire l'écart entre les volumes d'affaires nouvelles et de résiliations ?

Plusieurs axes d'amélioration sont possibles :

— **Augmenter le nombre de souscriptions** : en ajustant la politique marketing ou en proposant des tarifs d'affaires nouvelles encore plus intéressants ;

— **Contenir le nombre de résiliations** :

  — En améliorant l'expérience et la satisfaction client, dans la gestion de leurs sinistres par exemple, afin de gagner leur confiance ;

  — En imaginant un système de fidélisation avantageux pour le client qui viendrait compenser sur le plan moral l'augmentation de sa prime chaque année ;

  — En optimisant la valeur de la majoration à appliquer à chaque contrat afin de retenir au mieux le client.

Ainsi, dans le cadre de ce mémoire, nous nous concentrerons sur ce troisième point : **l'optimisation du taux de majoration à l'échéance du contrat qui répondrait au compromis entre la diminution du taux de résiliation du portefeuille et la rentabilité du portefeuille**.

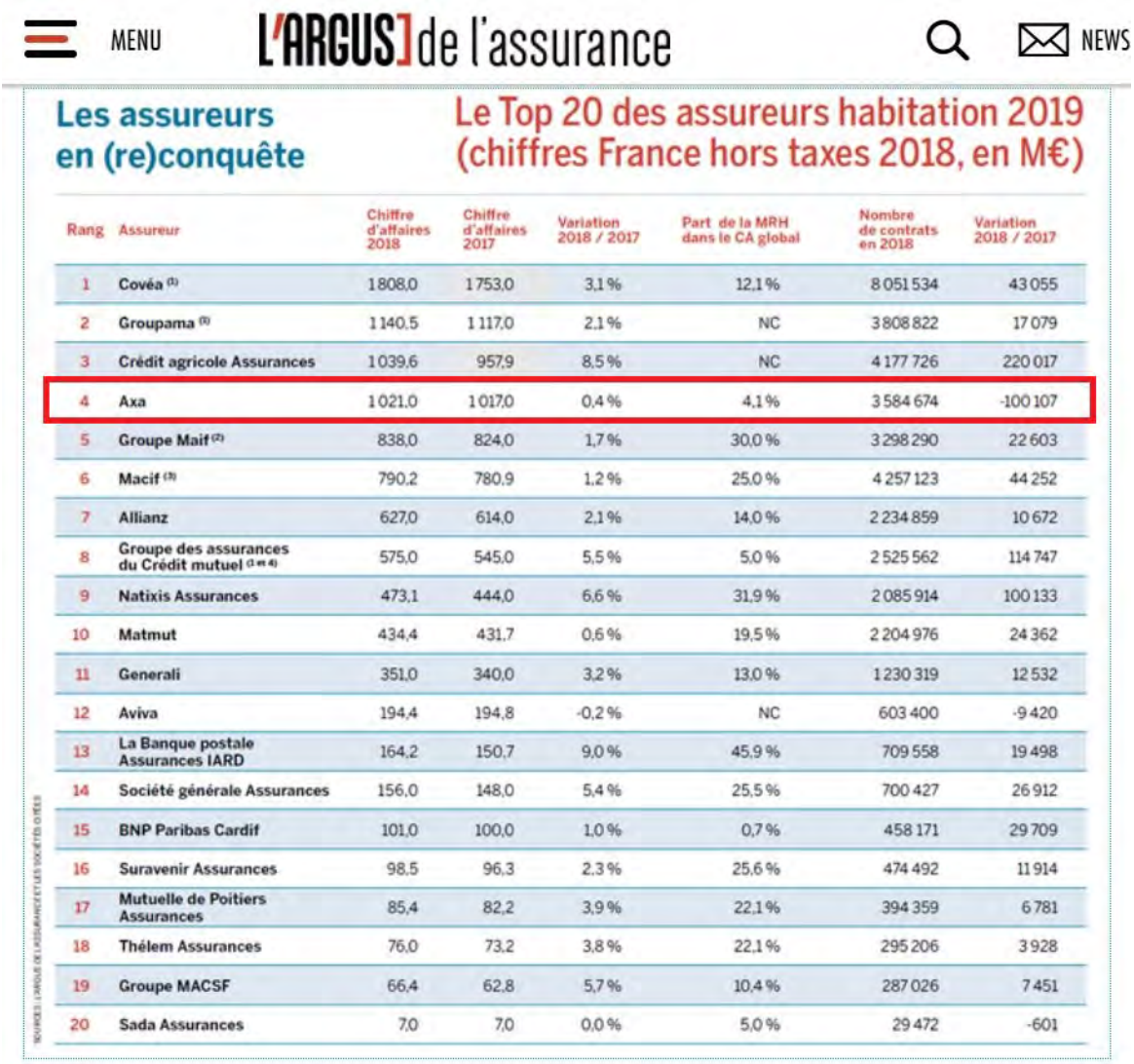
1.2 Terme : enjeux et stratégie actuelle pour l'application de la majoration

L'assurance habitation est historiquement peu ou pas profitable pour les assureurs, mais cache des disparités importantes de profitabilité, les générations récentes de contrats ayant une profitabilité dégradée. En effet, bien que la prime d'assurance MRH soit calculée à partir de l'espérance du coût que représente l'assuré (cette espérance de coût s'appelle la prime pure), l'assureur doit amortir dès la première année des frais d'acquisition importants. Ces frais d'acquisition correspondent notamment à l'effort financier que réalise l'assureur afin de proposer des tarifs d'affaires nouvelles cohérents avec le marché, c'est-à-dire des tarifs pas trop éloignés des offres des assureurs concurrents. Alors, la rentabilité de ce contrat et l'amortissement des frais (d'acquisition mais aussi de souscription et de gestion) doivent se développer sur le long terme : à chaque date d'anniversaire de la souscription du contrat, une majoration est appliquée sur la prime commerciale, et ce, même si aucun sinistre n'est survenu. Cette politique de majoration a été originellement mise en place d'une part pour revenir progressivement à l'équilibre économique (qui n'était pas atteint la première année), et d'autre part pour couvrir l'effet de l'inflation vis-à-vis de l'indice FFB (Fédération Française du Bâtiment) du coût de la construction [5]. Ainsi, au fil des années, la prime commerciale devient supérieure à la prime pure augmentée des frais et l'assureur peut alors en tirer un bénéfice.

Actuellement, AXA France présente un portefeuille Multirisques Habitation avec un apport net négatif, comme l'indique *L'Argus de l'assurance*, figure 1.2, page 6 (cf. encadré en gras sur le tableau), c'est-à-dire qu'il y a plus de résiliations que de nouveaux contrats en portefeuille. Bien que le chiffre d'affaires de la MRH soit important et permette à AXA de se placer en quatrième position du classement des assureurs habitation en 2019 [6], le portefeuille de contrats habitation reste en danger puisque son volume diminue d'année en année, ce qui aura un impact négatif au long terme sur sa rentabilité. Outre la problématique d'augmenter le nombre de souscriptions, il faut contenir l'hémorragie du portefeuille en retenant au maximum les clients, et de préférence retenir ceux les plus rentables.

5. L'indice FFB du coût de la construction est calculé à partir du prix de revient d'un immeuble de rapport de type courant à Paris. Il enregistre les variations de coût des différents éléments qui entrent dans la composition de l'ouvrage. Ce calcul ne prend pas en compte la valeur des terrains. L'objet initial de cet indice est l'indexation des polices d'assurance. Source : `https://www.ffbatiment.fr/federation-francaise-du-batiment/le-batiment-et-vous/en_chiffres/indices-index/Chiffres_Index_FFB_Construction.html`

6. Remarque : AXA France a perdu une place dans ce classement par rapport à 2019



MENU L'ARGUS de l'assurance NEWS

Les assureurs en (re)conquête

Le Top 20 des assureurs habitation 2019 (chiffres France hors taxes 2018, en M€)

Rang	Assureur	Chiffre d'affaires 2018	Chiffre d'affaires 2017	Variation 2018 / 2017	Part de la MRH dans le CA global	Nombre de contrats en 2018	Variation 2018 / 2017
1	Covéa [1]	1808,0	1753,0	3,1 %	12,1 %	8 051 534	43 055
2	Groupama [1]	1140,5	1117,0	2,1 %	NC	3 808 822	17 079
3	Crédit agricole Assurances	1039,6	957,9	8,5 %	NC	4 177 726	220 017
4	Axa	1021,0	1017,0	0,4 %	4,1 %	3 584 674	-100 107
5	Groupe Maif [2]	838,0	824,0	1,7 %	30,0 %	3 298 290	22 603
6	Macif [3]	790,2	780,9	1,2 %	25,0 %	4 257 123	44 252
7	Allianz	627,0	614,0	2,1 %	14,0 %	2 234 859	10 672
8	Groupe des assurances du Crédit mutuel [3 et 4]	575,0	545,0	5,5 %	5,0 %	2 525 562	114 747
9	Natixis Assurances	473,1	444,0	6,6 %	31,9 %	2 085 914	100 133
10	Matmut	434,4	431,7	0,6 %	19,5 %	2 204 976	24 362
11	Generali	351,0	340,0	3,2 %	13,0 %	1 230 319	12 532
12	Aviva	194,4	194,8	-0,2 %	NC	603 400	-9 420
13	La Banque postale Assurances IARD	164,2	150,7	9,0 %	45,9 %	709 558	19 498
14	Société générale Assurances	156,0	148,0	5,4 %	25,5 %	700 427	26 912
15	BNP Paribas Cardif	101,0	100,0	1,0 %	0,7 %	458 171	29 709
16	Suravenir Assurances	98,5	96,3	2,3 %	25,6 %	474 492	11 914
17	Mutuelle de Poitiers Assurances	85,4	82,2	3,9 %	22,1 %	394 359	6 781
18	Thélem Assurances	76,0	73,2	3,8 %	22,1 %	295 206	3 928
19	Groupe MACSF	66,4	62,8	5,7 %	10,4 %	287 026	7 451
20	Sada Assurances	7,0	7,0	0,0 %	5,0 %	29 472	-601

FIGURE 1.2 – Capture d'écran du site de l'Argus de l'assurance concernant la performance des assureurs habitation en France

L'enjeu est donc de déterminer le juste niveau d'augmentation de primes des contrats les moins profitables arrivant au terme (c'est-à-dire à l'échéance), tout en limitant le nombre de résiliations.

Aujourd'hui, chez AXA France, cette hausse de prime sur les contrats en portefeuille est évaluée selon différents critères, et en particulier la sinistralité et la rentabilité du contrat.

En fonction de ces critères, la majoration à appliquer est calculée suivant plusieurs étapes.

1ère étape, la majoration ELR (Expected Loss Ratio) :

L'*Expected Loss Ratio* est calculé pour chaque contrat du portefeuille. Il s'agit d'un indicateur de rentabilité du contrat dont le calcul sera explicité plus tard dans le Chapitre 2. Selon la tranche d'ELR à laquelle appartient le contrat, un certain nombre de points de majoration sera appliqué à la prime hors taxe du contrat.

2ème étape, la majoration pour sinistralité :

En fonction du nombre de sinistres sur des historiques plus ou moins longs et selon la catégorie du client, un certain nombre de points de majorations supplémentaires s'ajoute à la majoration globale. Une méthode différente de calcul de ces points est appliquée si les sinistres sont nombreux et coûteux. Cette dernière méthode se base notamment sur le type de bien assuré, le profil de l'assuré, les types de sinistres et l'ancienneté des sinistres.

3ème étape, la majoration pour mise en demeure :

Le cas échéant, les points de majorations à ajouter aux précédents dépendent de la date et du nombre de mises en demeure observés pour ce contrat.

Bien sûr, pour chaque règle, il existe des exceptions qui permettent de limiter les majorations afin que les nouvelles primes ne soient pas excessives pour les clients les plus fidèles.

La majoration totale est donc la somme de ces différents points de majoration, et est appliquée à la prime hors taxe du client.

1.3 Pourquoi développer un modèle d'optimisation pour le terme ?

Nous constatons ainsi que la technique de calcul de la majoration au terme est une succession de sommes de points de majoration prédéterminés en fonction des caractéristiques du contrat et dont les valeurs dépendent d'une décision basée sur des études empiriques et qui répond à la stratégie de l'entreprise de vouloir majorer plus ou moins certains contrats tout en satisfaisant une contrainte de ressource suffisante. Par exemple, c'est en regardant l'impact des majorations sur la rentabilité des contrats et sur le chiffre d'affaires global que ces dernières sont ajustées sur avis d'experts.

Aujourd'hui il n'existe pas de modèle général d'optimisation de la majoration des contrats et il n'est pas possible de savoir si la méthode actuelle de majoration est la plus "efficace". Le développement d'un tel modèle d'optimisation prenant en compte l'ensemble des contraintes et des données disponibles permettrait certainement d'explorer des solutions non testées à ce jour, car l'étude serait réalisée en mode séquentiel, c'est-à-dire qu'elle s'adapterait à la composition globale du portefeuille mois après mois.

Que faut-il entendre par "efficace" ? La notion d'efficacité dépend de l'objectif que nous nous fixons.

1.4 Définition de l'objectif de l'optimisation de la majoration

Notre objectif est d'**améliorer l'apport net sans détériorer la rentabilité du portefeuille**. Ainsi, deux possibilités s'offrent à nous : augmenter le nombre d'affaires nouvelles et/ou diminuer le nombre de résiliations. Dans le cadre de notre étude, nous décidons de nous concentrer sur la résiliation. Nous souhaitons alors déterminer la **majoration optimale** telle qu'elle **minimise le taux de résiliation du portefeuille sous contrainte de la profitabilité totale (ELR)**.

L'ELR, ou *Expected Loss Ratio*, est le ratio du coût attendu des sinistres sur la somme des primes commerciales des contrats en portefeuille et se calcule selon la formule :

$$ELR = \frac{\sum\limits_{i} PP_i}{\sum\limits_{i} PC_{HT,i}} \tag{1.1}$$

avec PP_i la prime pure du contrat i et $PC_{HT,i}$ la prime commerciale hors taxe appliquée au contrat i après l'application de la majoration.

Rappelons que la prime pure PP pour un contrat donné reflète le coût que va représenter l'assuré pendant l'année à venir. En effet, en assurance, le cycle de production est inversé : le contrat est vendu à un certain prix appelé "prime commerciale" (PC). Cette prime commerciale devra couvrir les frais de gestion du contrat ainsi que le coût des potentielles prestations à venir qui n'est pas encore connu à la date de souscription. Une tarification *a priori* basée sur des méthodes statistiques est alors réalisée en se basant sur un modèle de prime pure (PP) qui détermine l'espérance du coût que représente un assuré en fonction de ses différentes caractéristiques et des garanties qu'il aura souscrites.

Ainsi, pour que le portefeuille de contrat soit profitable, c'est-à-dire pour qu'il génère de la marge, il faut naturellement que la somme des primes commerciales hors taxes soit supérieure à la somme des primes pures de chaque contrat augmentées des frais à payer, autrement dit il faut que l'ELR soit bien inférieur à 1. Dans le cas contraire, le portefeuille générerait des pertes. Plus l'ELR sera petit, plus le portefeuille sera profitable, et inversement, plus ce ratio sera grand, moins le portefeuille générera de profits.

Le profit généré par le portefeuille dépend à la fois du volume de contrats, de la prime commerciale de chaque contrat ainsi que de la sinistralité réelle (estimée par la prime pure). Et pour les contrats ayant plus d'un an, la prime commerciale dépend de la majoration qui leur est appliquée. Or, plus la majoration augmente, plus le risque de résiliation augmente. Toutefois, le risque de résiliation pour une même majoration varie d'un client à l'autre, et nous aurons l'occasion de le constater au prochain chapitre dans le cadre de l'étude des statistiques descriptives. Ainsi, les différents niveaux de majorations ne sont peut-être pas appliqués aux bons contrats, résiliant les contrats les plus rentables et gardant les moins rentables.

La figure 1.3 page 9 schématise la **frontière efficiente** [7], c'est-à-dire le lieu des points où le taux de résiliation est minimal pour une profitabilité donnée. L'état d'un portefeuille de contrats peut être signalé sur ce schéma par un point en fonction de son ELR et de son taux de résiliation, et ne peut figurer qu'au-dessus ou sur la frontière efficiente. La région en dessous de la frontière est inaccessible, puisque le frontière représente une optimalité qui ne peut être surpassée.

Nous supposons aujourd'hui que la performance actuelle d'AXA France ne se trouve pas sur cette frontière efficiente et qu'il est alors possible de la rejoindre en améliorant la profitabilité (il y a une meilleure profitabilité lorsque l'ELR se rapproche de zéro) et en gardant le même taux de résiliation, ou bien en diminuant le taux de résiliation tout en maintenant le même niveau d'ELR. Nous vérifierons cette hypothèse dans le cadre de notre

7. Schéma inspiré du mémoire d'actuariat d'A. De Larrard [3]

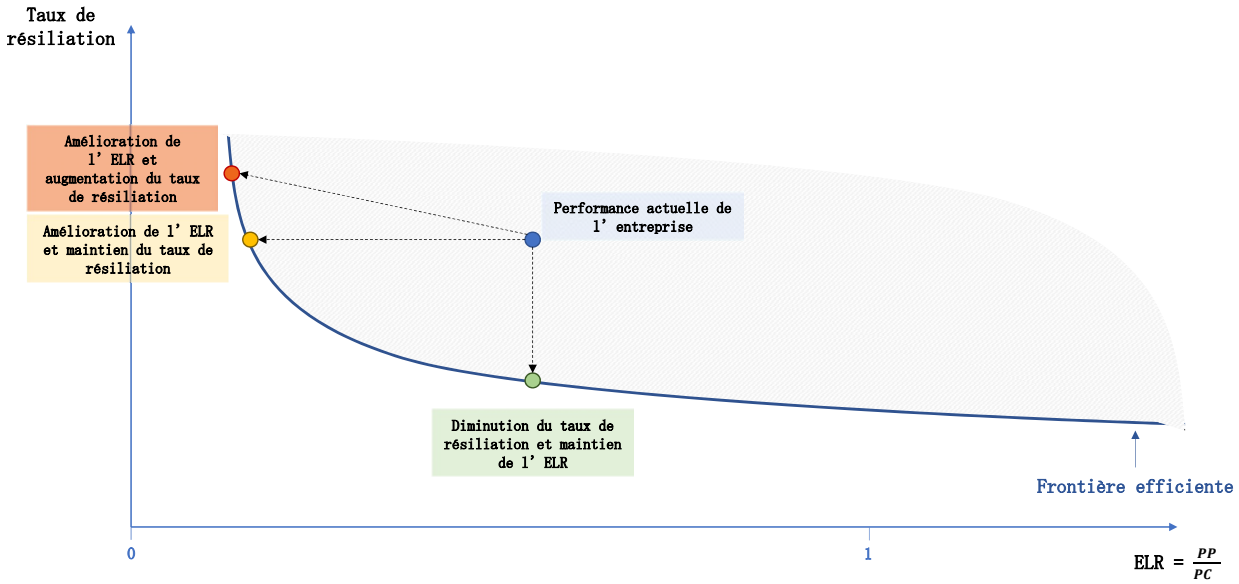


FIGURE 1.3 – Schéma de la frontière efficiente

modèle d'optimisation et proposerons une solution pour calculer les majorations optimales pour chaque contrat afin de se rapprocher au maximum de cette frontière.

1.5 Plan d'action

Rappelons que notre objectif est de déterminer la majoration optimale pour minimiser le taux de résiliation sous contrainte de rentabilité.

Nous souhaitons donc dans un premier temps **modéliser la résiliation due au tarif** (c'est-à-dire une résiliation qui n'est pas due à un décès par exemple, à un déménagement ou à une résiliation décidée par AXA), *en supposant que cette résiliation tarifaire dépend de la majoration.*

Ensuite, la rentabilité future d'un contrat ne dépend pas uniquement de la majoration appliquée car les agents généraux peuvent appliquer des rabais préventifs (avant l'envoi de l'avis d'échéance ou sur demande du client) appelés "**crédit commercial**" comme geste commercial, afin de retenir les clients qu'ils souhaitent garder dans leur portefeuille. Cela constitue donc un manque à gagner pour AXA, appelé "**leakage**", qui viendra donc réduire la profitabilité du contrat. Toutefois, ce levier commercial est utile pour éviter qu'un client ne résilie son contrat, ce qui est d'autant plus important si le client en question est multidétenteur (c'est-à-dire s'il possède plusieurs contrats d'assurance AXA) : en effet, il est fort probable qu'en décidant de résilier son contrat habitation, le multidétenteur décide de résilier la totalité (ou partie) de ses contrats AXA. Nous décidons donc de réaliser un **modèle de crédit commercial** en deux étapes : d'abord la modélisation de la **probabilité d'application de crédit commercial** pour un contrat, puis la modélisation du **pourcentage de crédit commercial** par rapport à la prime commerciale toutes taxes comprises pour le contrat correspondant.

Sur la base de ces modèles de résiliation et de crédit commercial, nous pourrons obtenir, pour chaque contrat et selon différentes majorations, la probabilité de résiliation, la prime pure pondérée par la probabilité que le contrat reste en portefeuille, ainsi que le chiffre d'affaires espéré de ce contrat pondéré à nouveau par la probabilité qu'il reste dans le portefeuille. Il sera ensuite possible de procéder à l' **optimisation sur les**

majorations pour **minimiser le taux de résiliation du portefeuille sous la contrainte de globale de l'ELR** (rentabilité) :

$$\begin{array}{ll} \min\limits_{x_1,\dots,x_N} & \frac{1}{N}\sum_{i=1}^{N} f_i(x_i) \\ \text{sous contrainte} & ELR_{tot} \leq ELR_{max} \end{array} \tag{1.2}$$

avec :

- N le nombre de contrats dont la majoration est à optimiser ;
- x_i la majoration hors taxes appliquée au contrat i ;
- f_i la fonction de résiliation, qui prédit la probabilité de résiliation d'un contrat i en fonction de la majoration x_i. A noter que la fonction de résiliation est unique à chaque contrat et qu'elle est obtenue grâce au **modèle de résiliation** ;
- ELR_{tot}, l'ELR du portefeuille calculé selon la formule :

$$\begin{aligned} ELR_{tot} &= \frac{\sum\limits_{i}^{N} PP_{espérée,i}}{\sum\limits_{i}^{N} PC_{HT,n,espérée,i}} \\ &= \frac{\sum\limits_{i}^{N} PP_i \times (1 - f_i(x_i))}{\sum\limits_{i}^{N} PC_{HT,n-1,i}(1+x_i)(1-f_i(x_i))(1-\%CC_i(x_i)\times \text{Proba}_{CC,i}(x_i))} \end{aligned} \tag{1.3}$$

  L'ELR que nous considérons est le rapport entre la somme des primes pures espérées de chaque contrat et la somme des primes commerciales espérées hors taxe de chaque contrat pour l'année après l'application de la majoration au terme. Nous utilisons l'expression "prime pure espérée" car, bien que la prime pure soit déjà une espérance de coût des sinistres, dans notre ELR nous pondérons cette prime pure avec la probabilité de rétention du contrat. En effet, à la date où nous calculons cet ELR, nous ne savons pas exactement si un client va résilier ou non son contrat à l'issue de la majoration. Nous devons donc estimer l'ELR que nous obtiendrions après la majoration, en prenant en compte les potentielles résiliations de contrats. De même, nous pondérons chaque prime commerciale avec la probabilité de rétention, car un contrat résilié ne rapporte plus de chiffre d'affaires.

  L'ELR total prend donc notamment en compte la probabilité de résiliation de chaque contrat f_i, la prime commerciale hors taxes $PC_{HT,n-1,i}$ avant la majoration, le pourcentage de crédit commercial $\%CC_i$ obtenu par le **modèle de coût de crédit commercial** ainsi que la probabilité d'application du crédit commercial $\text{Proba}_{CC,i}$ sur ce contrat obtenu par le **modèle de probabilité de crédit commercial** ;
- ELR_{max}, la valeur maximale que ne doit pas dépasser l'ELR total du portefeuille.

En outre, nous nous assurerons que le chiffre d'affaires à l'issue de l'optimisation n'a pas trop été dégradé.

Nous réaliserons ensuite sur les résultats de cette optimisation un *Reverse engineering*[8] afin de déterminer les variables qui prédiraient le mieux la majoration optimale pour chaque contrat. L'objectif étant, pour les années à venir, de déterminer pour chaque contrat, en fonction de ses caractéristiques, la majoration optimale à appliquer grâce au modèle obtenu par le *Reverse engineering* sans avoir à réaliser à nouveau une optimisation en tant que telle. Le *Reverse engineering* a surtout un intérêt opérationnel : puisque que le travail d'optimisation est long et complexe, il n'est pas envisageable de réaliser ce travail chaque mois. Il est donc nécessaire de trouver un moyen plus direct et rapide de calculer la majoration optimale pour chaque contrat.

8. Le *Reverse engineering* (ou Rétro-ingénierie), est un processus d'étude d'un modèle pour en comprendre le fonctionnement interne à partir des données d'entrée et de sortie.

La figure 1.4 résume le processus de modélisation et d'optimisation que nous allons suivre.

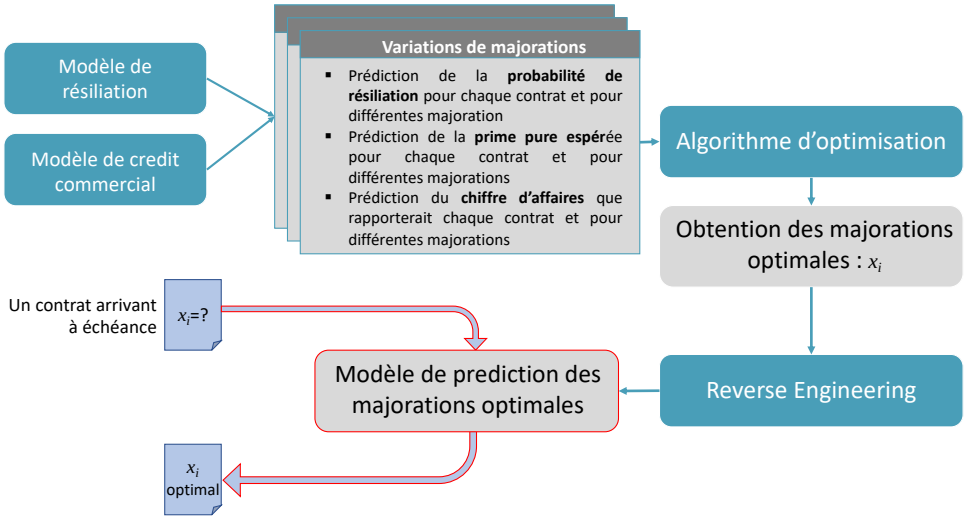


FIGURE 1.4 – Plan d'action schématisé pour l'obtention de majorations optimales

L'étude sera réalisée à partir d'une base de données que nous aurons pris le soin de construire selon une architecture dite par "image de terme", où chaque ligne de la base sera associée à une échéance d'un contrat (cf. détail en section 2.1.1), à partir de différentes sources d'informations que nous détaillerons dans le Chapitre 2.

1.6 Définition du périmètre d'étude

Dans le cadre de notre étude, nous nous concentrons dans un premier temps sur un certain périmètre du portefeuille MRH suffisamment grand pour que l'étude soit fiable, et sur lequel les données sont facilement accessibles. Si l'étude d'optimisation sur ce périmètre procure des résultats satisfaisants, il sera alors envisageable d'élargir ce périmètre afin de généraliser l'optimisation sur l'ensemble du portefeuille MRH. Nous définissons ci-dessous le périmètre d'étude :

En terme de profil du souscripteur : Nous décidons de limiter notre périmètre d'étude en retirant de notre base les contrats "Propriétaires non occupants" et les contrats "Etudiants", dont la tarification des affaires nouvelles est particulière et dont le comportement des souscripteurs en terme de résiliation sont particuliers. Par exemple, nous observons une saisonnalité dans la résiliation des contrats étudiants : le taux de résiliation augmente significativement entre juin et septembre, ce qui correspond aux vacances scolaires, et donc généralement à une période propice aux déménagements des étudiants.

Dans l'idéal, il faudrait réaliser par la suite la même étude d'optimisation sur chacun de ces profils pour plus de justesse.

En terme de géographie : AXA France vend des contrats en France métropolitaine et en outre-mer. Toutefois, les méthodes de tarification des risques MRH en outre-mer sont différentes, notamment sur l'aspect climatique. De ce fait, nous allons restreindre notre périmètre à la France métropolitaine.

En terme d'historique : Nous souhaitons avoir une base de données la plus complète possible et notamment avoir accès à l'historique des majorations appliquées aux contrats. En outre, il faut considérer le fait que le

comportement des clients évolue avec le temps. Ainsi, prendre un historique trop long n'aurait pas de sens. Cela est d'autant plus vrai qu'avec l'arrivée de la loi Hamon en 2014, les clients ont changé significativement leur comportement en ayant la possibilité de résilier beaucoup plus facilement leurs contrats au bout d'un an. De plus, l'essor du digital et des comparateurs d'offres d'assurance sur internet ces dernières années a également un impact sur la tendance des clients à résilier un contrat d'assurance.

Ainsi, pour capter au mieux le comportement actuel des clients AXA, nous décidons de réaliser notre étude sur deux années complètes les plus récentes : 2017 et 2018. Cela nous éloigne d'ailleurs suffisamment de l'apparition de la loi Hamon qui aurait pu, par effet de nouveauté, biaiser notre analyse. En 2017 et 2018 nous nous assurons que les clients ont un comportement stable et ne sont pas perturbés par l'apparition d'une loi impactant leur consommation de produits d'assurance.

Conclusion du chapitre

Nous avons constaté que le volume des résiliations est supérieur au volume d'affaires nouvelles pour les contrats multirisques habitation proposés par AXA France, rendant de ce fait l'apport net négatif depuis près de 4 ans. Notre objectif est donc de réduire cet écart, et pour ce faire nous décidons de résorber le taux de résiliation au terme des contrats, en ajustant le pourcentage de majoration tout en maintenant la profitabilité du portefeuille que nous estimons par l'ELR. Le problème d'optimisation devient clair : minimiser le taux de résiliation des contrats au terme sous contrainte de l'ELR.

Nous avons également introduit, dans le cadre de ce chapitre, la notion de frontière efficiente, qui constitue le lieu des points où le taux de résiliation est minimal pour un niveau d'ELR donné. L'idéal serait de minimiser le taux des résiliations et également d'améliorer la profitabilité du portefeuille. L'enjeu est donc de savoir comment se positionne la stratégie actuelle d'AXA par rapport à cette frontière qui matérialise les stratégies optimales, le but étant de rejoindre cette dernière.

Pour atteindre cet objectif, un plan d'action a été mis en place : une première étape consistera à élaborer un modèle de résiliation et un modèle de crédit commercial afin de simuler le comportement des clients face à différents niveaux de majoration. Nous pourrons ensuite, dans une deuxième étape, résoudre le problème d'optimisation des majorations. Enfin, à des fins opérationnelles, nous allons réaliser dans une troisième et dernière étape un reverse engineering sur la base des résultats de l'optimisation, afin d'obtenir de manière plus directe et rapide les majorations les plus adéquates pour chaque contrat.

Chapitre 2

Modélisation de la probabilité de résiliation au terme

Sommaire

Préambule

Notre problème d'optimisation demande de minimiser le taux de résiliation au terme des contrats en portefeuille. Il est donc nécessaire de modéliser le comportement du client, et, plus spécifiquement, déterminer la probabilité qu'un client décide de résilier son contrat lorsque ce dernier arrive à échéance. Cette probabilité dépendra naturellement du profil du client mais également des caractéristiques du contrat.

Il faudra à cette occasion déterminer ce qui différencie une résiliation tarifaire [1] liée au terme (celle qui nous intéresse dans le cadre de cette étude), d'une résiliation tarifaire non liée au terme. Car nous voulons prédire le comportement du client au moment de l'échéance du contrat, autrement dit au moment où le client réalise que la prime de son contrat a été majorée.

Cette étape de modélisation de la résiliation, très classique, est également une bonne occasion d'étudier les données dont nous disposons : des statistiques descriptives simples vont nous permettre de vérifier si notre étude a du sens et un réel intérêt stratégique.

2.1 Construction de la base de données

Pour réaliser nos modèles de résiliation et de crédit commercial, l'étape préliminaire indispensable est de construire une base de données robuste la plus complète possible contenant les informations des contrats MRH termés, sur plusieurs années, afin d'obtenir un maximum de caractéristiques concernant le contrat, le client associé et la gestion d'éventuels sinistres notamment. Nous préférons dans un premier temps sélectionner un

1. C'est-à-dire une résiliation pour cause d'une hausse de la prime.

large nombre de variables potentiellement explicatives à notre disposition, quitte à réaliser une sélection plus fine par la suite, après analyse des statistiques descriptives.

2.1.1 Une architecture en "image de terme"

L'idée est de construire une base de données **par "image de terme"**, c'est-à-dire que chaque ligne de notre base est associée à un terme concernant un contrat, lui-même rattaché à un client. La raison d'une telle méthode de construction provient du fait que les bases de données AXA sont des "photographies" du portefeuille de contrats à une date donnée. Nous souhaitons donc capter l'évolution des contrats en comparant l'état du portefeuille à deux dates distinctes séparées d'un certain nombre de mois que nous établirons plus tard. La base de données possédera alors autant de lignes que de termes observés pendant une période donnée (c'est-à-dire une fenêtre d'observation sur plusieurs années) pour chacun des contrats du portefeuille MRH. La figure 2.1 schématise le principe de l'architecture de notre base.

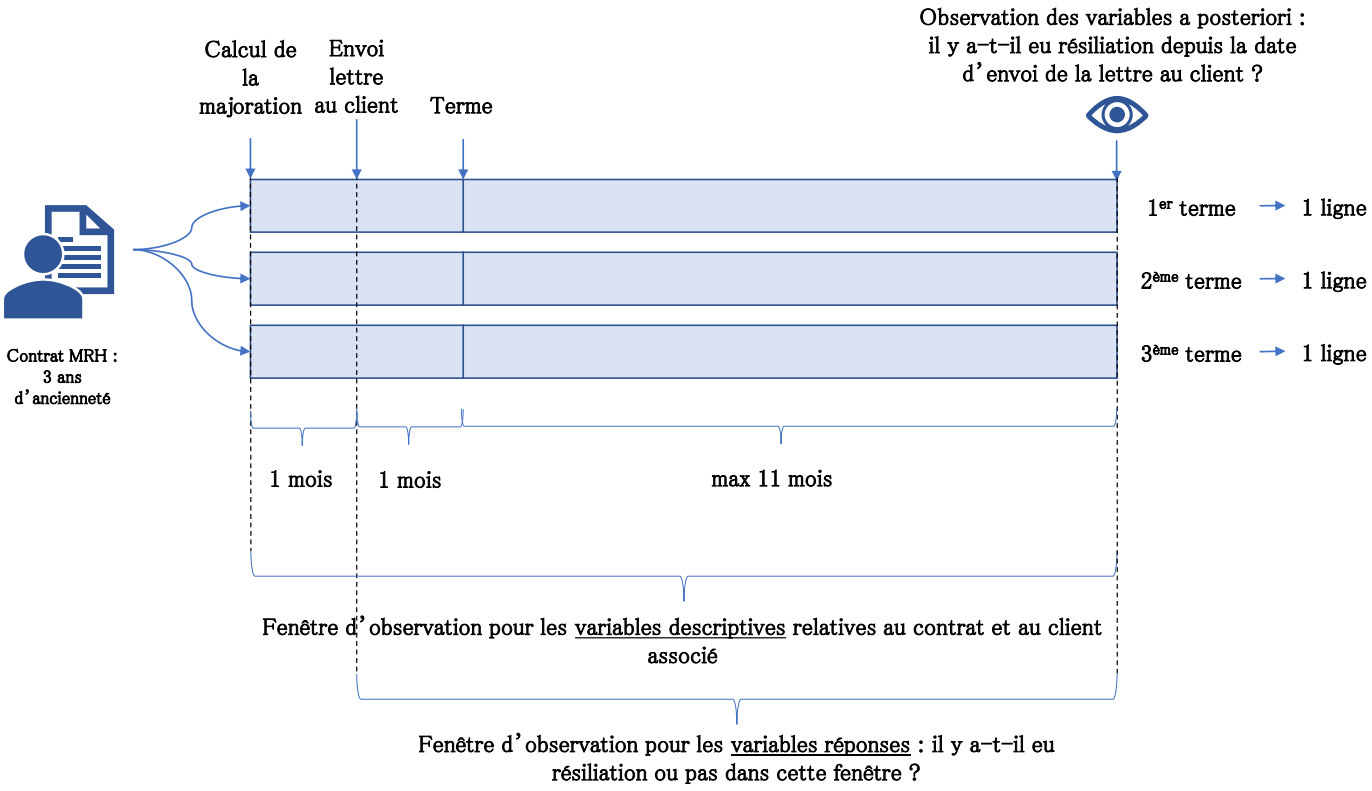


FIGURE 2.1 – Construction de la base des résiliations par image de terme

Prenons ainsi l'exemple d'un contrat i associé à un client C. Ce contrat a une ancienneté de 3 ans (nous avons donc pris ici une fenêtre d'observation d'au moins trois ans car nous observons 3 lignes, c'est-à-dire 3 images de terme). L'objectif est d'observer la réaction du client sur chaque ligne, c'est-à-dire chaque année.

Pour chaque image de terme, il faut choisir une fenêtre d'observation des "**variables explicatives**". Notons d'ailleurs que nous entendons par "variables explicatives" toutes les variables candidates qui selon nous expliqueraient en partie la résiliation tarifaire d'un contrat au terme. Un travail d'étude sur la significativité de ces variables sera entrepris dans la suite de l'étude pour garder celles les plus pertinentes.

Pour l'analyse des statistiques descriptives de notre base de données, nous choisissons une fenêtre d'observation suffisamment grande pour pouvoir observer les tendances de résiliation sur une année complète : sachant que le calcul de la majoration pour un contrat se fait 2 mois avant la date d'anniversaire du contrat, le début de l'observation se fera donc 2 mois avant le terme du contrat en question. Enfin, la fin de la période d'observation est fixée à 11 mois après le terme de telle sorte que la fin de l'observation coïncide avec l'envoi de la lettre d'information de la majoration pour le terme suivant (s'il existe). Nous avons donc pour chaque ligne une période d'observation totale de 13 mois qui encadre la date de terme du contrat concerné.

Pour réaliser notre modèle de résiliation, étant donné que nous souhaitons capturer les résiliations tarifaires dues au terme, nous allons réduire la fenêtre d'observation autour du terme : le début de l'observation sera toujours 2 mois avant la date d'échéance du contrat, mais nous ajustons la fin de l'observation à 4 mois après le terme en considérant qu'une résiliation tarifaire dans cette fenêtre est déclenchée par la majoration au terme. Nous vérifierons la pertinence de ce choix de fenêtre avec l'analyse des statistiques descriptives.

Aux "variables explicatives" s'ajoutent les **variables réponses**, qui indiquent s'il y a eu une résiliation pendant la période observée. Il est toutefois important de noter que cette période d'observation des variables réponses diffère de la fenêtre d'observation des variables descriptives. En effet, tandis que le calcul de la majoration est réalisé 2 mois avant le mois du terme, le client ne reçoit la lettre d'information concernant sa majoration qu'un mois avant le terme (dans le cadre de la loi Chatel [2]). Il peut donc prendre sa décision de résilier ou non au plus tôt un mois avant le terme. Ainsi la période d'observation des variables réponses se limite-t-elle à 1 mois avant le terme et 11 mois après le terme. Nous supposons donc que le client peut réagir à sa majoration sur une durée d'un an : nous pouvons en effet envisager le cas où un client réalise que sa prime est trop chère 5 mois après avoir été informé de sa majoration.

En outre, il faut discerner la **date d'émission** d'une résiliation, c'est-à-dire le jour où le client prend la décision de résilier son contrat, de la **date d'effet** de la résiliation qui indique le jour où le contrat prend réellement fin. Ces deux dates de résiliation sont intéressantes pour étudier le comportement des clients en observant le temps qui s'écoule entre ces deux dates.

Nous distinguons ainsi quatre variables réponses :

- **REP_ResilEmise** : indique si une résiliation a été émise ou non pendant la période d'observation ;
- **REP_ResilEmise_bis** : indique, dans le cas d'une résiliation émise, le motif de cette résiliation.
- **REP_ResilEffet** : indique si une résiliation prend effet ou non pendant la période d'observation ;
- **REP_ResilEffet_bis** : indique, dans le cas où une résiliation prendrait effet, le motif de cette résiliation.

Concernant le motif d'une résiliation, il s'agit de déterminer si cette résiliation est tarifaire ou non, et si elle est tarifaire (à savoir si elle est liée au terme ou pas, par exemple : une résiliation pour cause de déménagement, décès ou sur décision d'AXA n'est pas tarifaire, alors qu'une résiliation "Hamon" ou "Majoration" sera considérée comme tarifaire). Nous considérerons que la résiliation est liée au terme si elle apparaît dans la fenêtre [1 mois avant le terme ; 4 mois après le terme]. Cette fenêtre a été choisie en observant la distribution des résiliations par rapport au nombre de jours qui séparent le terme de l'émission de la résiliation. Sur la figure 2.2 nous pouvons constater un pic d'émission de résiliations à l'échéance du contrat, et que le volume des résiliations tarifaires se stabilise à partir de 4 mois après le terme.

2. Cf. Annexe 1

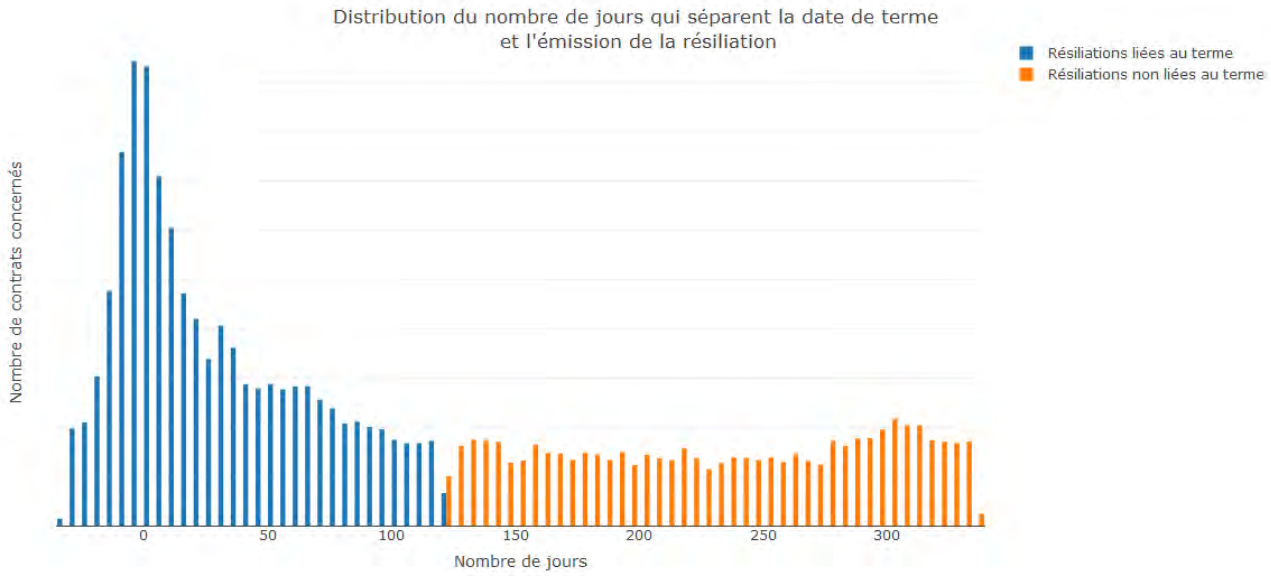


FIGURE 2.2 – Représentation de la répartition des résiliations en 2017

Par la suite, nous garderons comme **variable cible**, pour l'élaboration du modèle de résiliation, la date d'**émission de résiliation**, étant donné que nous nous intéressons au comportement du client et donc au moment où le client prend la décision de résilier. **Notre objectif premier est de déterminer si la majoration au terme a un impact ou non sur la décision du client de résilier**.

2.1.2 Processus itératif de construction de la base de données

Tous les contrats n'ont bien entendu pas la même date d'anniversaire, ainsi, selon les mois de terme, les fenêtres d'observations varient et suivent l'encadrement [-2 mois avant le terme ; +11 mois (ou 4 mois) après le terme]. La collecte des variables descriptives va se faire de manière itérative : des bases intermédiaires répertoriant les contrats concernés par une année et un mois de terme donnés vont être créées, puis elles vont être concaténées afin d'obtenir une base contenant autant de lignes que de termes réalisés sur une période de plus d'un an.
La figure 2.3 [3] synthétise le processus de construction de la base de données, et plus particulièrement, la partie en bleu détaille la méthode itérative par mois de terme.

A chaque début d'itération sont indiquées **4 variables d'entrées** indiquant le mois du terme, l'année du terme, ainsi que les dates de début et de fin d'observation, soit [-2 mois avant le terme ; +11 mois (ou 4 mois) après le terme].

Sont ensuite récupérées des informations issues de la **base de données des contrats MRH** : pour chaque contrat, nous récupérons les différentes caractéristiques du contrat, la date d'affaire nouvelle (i.e. de souscription), les garanties souscrites, la localisation du bien assuré... Cela nous permet notamment d'obtenir la date d'échéance du contrat (ou date d'anniversaire), ce qui rend possible la sélection des contrats termés uniquement durant le mois qui nous intéresse pour la base intermédiaire. Par ailleurs, une variable "RésilEmise" est créée, prenant la modalité "OUI" si une résiliation a été émise avant la date de début d'observation mais que la date d'effet a lieu après, et la modalité "NON" s'il n'y a pas encore eu de résiliation à la la date de début d'observation. Cela va nous aider par la suite à sélectionner plus facilement les lignes qui nous intéressent, selon si l'on souhaite faire notre modélisation sur les résiliations émises ou sur les prises d'effet des résiliations.

3. Il faut comprendre par le verbe "termer" (jargon professionnel), le fait que le contrat arrive à l'échéance, et par le mot "topage" l'occurrence ou non d'un évènement (dans notre cas, l'occurrence ou non d'une résiliation).

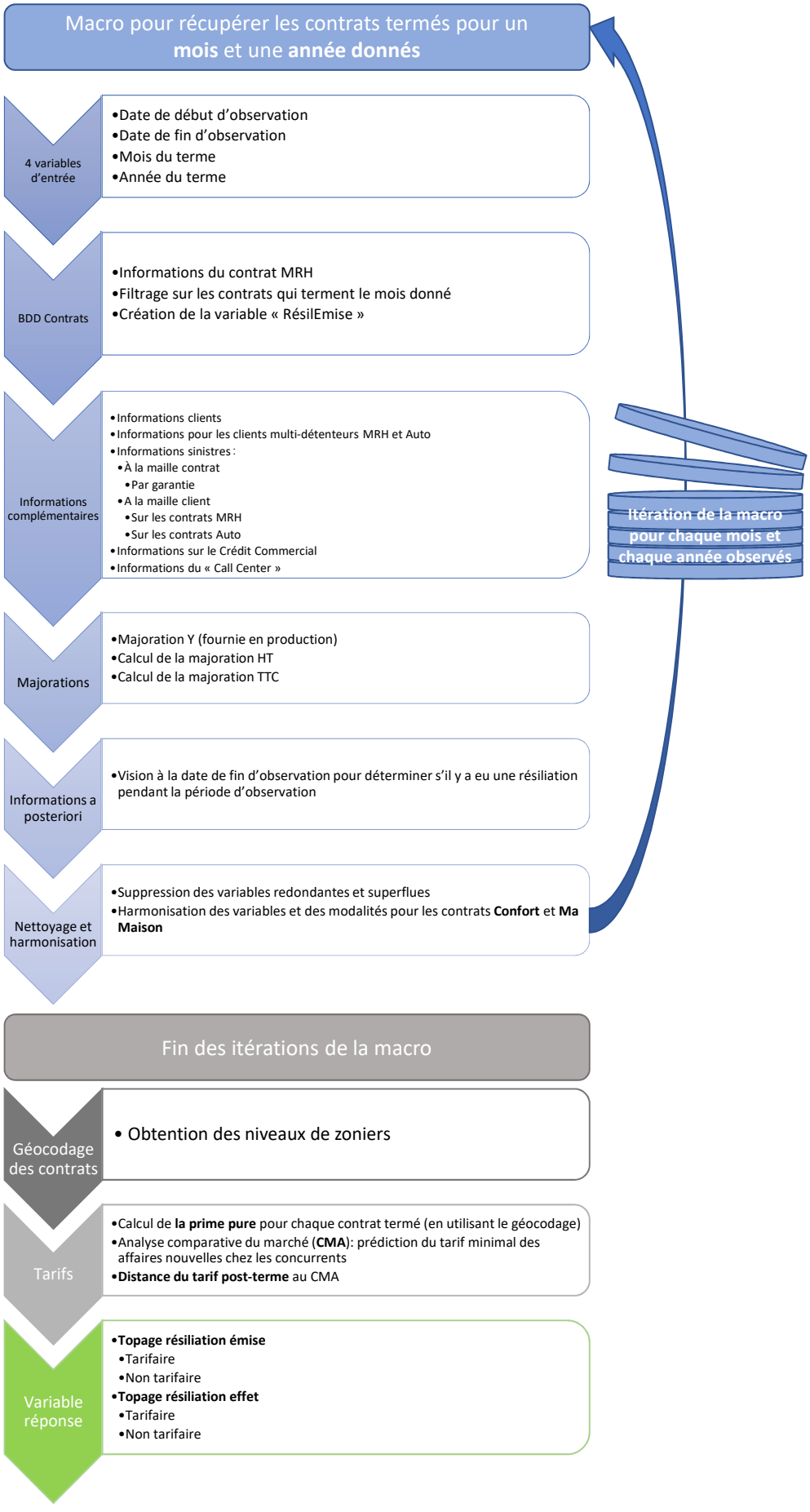


FIGURE 2.3 – Processus itératif de la construction de la base des résiliations

Puis sont ajoutées des **informations complémentaires**, notamment des renseignements sur le client (son âge, nombre d'enfants...), est-ce qu'il possède plusieurs contrats chez AXA (multi-détenteur) : en habitation, auto ou autre, et des informations sur ses autres contrats habitation et auto s'il en possède. Nous complétons par les informations sinistres pour le contrat en question, garantie par garantie. Nous renseignons également pour chaque client les informations sinistres liées à tous leurs contrats MRH et auto. Les données concernant la gestion des sinistres enrichissent la base avec les informations du "Call center" indiquant le nombre d'appels et le lieu de réception des appels. Enfin, nous ajoutons le montant du "Crédit Commercial" avant et après l'impact de la majoration, qui indique le rabais qui a été affecté sur la prime par l'agent dans le cas où le client serait venu faire une réclamation sur le prix de son contrat.

Les **majorations** assignées à chaque contrat sont également récupérées. Nous distinguons la *majoration Y* qui est le pourcentage de majoration brut calculé à partir des caractéristiques du client et de ses sinistres passés, la *majoration Hors Taxes* calculée notamment à partir de la majoration Y et la *majoration Toutes Taxes Comprises* calculée aussi à partir de la majoration Y.

Nous rajoutons en outre des **informations *a posteriori***, observées à la fin de la fenêtre d'observation, qui nous permettent notamment de déterminer s'il y a eu une résiliation ou non pendant l'intervalle d'observation. Cela va nous permettre de construire nos variables réponses.

Enfin, la dernière étape du processus itératif est le **nettoyage des données et l'harmonisation** : les variables redondantes et superflues sont supprimées. De plus, étant donné que le portefeuille MRH contient des contrats *Confort* (ancienne gamme de contrats d'assurance habitation) et des contrats *Ma Maison* (nouvelle gamme depuis 2017), les variables concernant les garanties choisies ainsi que leurs modalités peuvent différer. Il est donc nécessaire d'harmoniser les noms des variables et des modalités.

Le nombre d'itérations dans la construction de la base est égal au nombre de mois que nous souhaitons étudier.

2.1.3 Finalisation de la base de données

Une fois la phase d'itération terminée, les bases intermédiaires sont assemblées pour ne faire qu'une et nous pouvons donc calculer pour chaque ligne la prime pure associée, un estimateur d'un prix compétitif provenant de la concurrence, ainsi que les variables réponses.

2.1.3.1 Calcul de la prime pure

La prime pure pour un contrat donné reflète le coût que va représenter l'assuré pendant l'année à venir. Il s'agit de l'espérance de coût de la sinistralité pour cet assuré.

L'intérêt de calculer la prime pure dans la base de résiliation est d'avoir une idée de la rentabilité de chaque contrat, notamment en calculant par la suite l'*Expected Loss Ratio*. Par exemple, il sera intéressant de regarder si les contrats qui sont résiliés sont rentables ou non. De plus, dans le cadre de l'optimisation future, la prime pure est également un élément essentiel pour déterminer la rentabilité du portefeuille.

Dans le cadre de la construction de notre base de données, nous utilisons un modèle de prime pure conçu en interne et adapté au renouvellement annuel du contrat, basé sur certaines caractéristiques du client, les garanties et les options souscrites. A noter que ce calcul nécessite des "niveaux de zonier" indiquant des niveaux de risque par garantie en fonction de la localisation du bien à assurer. Ces "niveaux de zonier" internes et propres à AXA sont récupérés avec un algorithme de "géocodage".

Connaissant la prime commerciale qui a été appliquée à chaque contrat, il est alors aisé de calculer l'*Expected Loss Ratio* selon la formule :

$$ELR_i = \frac{PP_i}{PC_{HT,i}} \tag{2.1}$$

2.1.3.2 Calcul de l'estimateur d'un prix compétitif provenant de la concurrence

La loi Hamon autorise le client à résilier son contrat à n'importe quel moment à partir d'une année de souscription. Nous pouvons donc imaginer qu'un souscripteur expert, souhaitant minimiser le coût de son assurance pour la même utilité, se renseigne et compare, à l'issue du terme, le nouveau tarif de son contrat d'assurance habitation avec les tarifs proposés par la concurrence pour des contrats équivalents en termes de garanties afin de se diriger vers la solution la moins coûteuse.

Dans le cadre d'une analyse de marché comparative (communément appelée *CMA* chez AXA, ie *Competitive Market Analysis*), AXA France a la possibilité d'obtenir tous les mois les tarifs d'une autre filiale du groupe AXA en France sur un échantillon de contrats représentatifs du portefeuille MRH d'AXA France. Il nous est alors possible d'entraîner sur cette base un *Generalized Linear Model* (GLM, ou *Modèle Linéaire Généralisé*, méthode que nous détaillons en annexe 4, page 104) qui sera capable de prédire, pour n'importe quel contrat du portefeuille AXA France, le tarif qui serait proposé par l'autre entité AXA en affaire nouvelle, suivant les garanties du contrat. De cette prédiction, il est alors possible de calculer la "distance" du prix AXA France après majoration par rapport au prix concurrent :

$$Distance\ CMA = \frac{Prix\ AXA\ France\ TTC\ post\ majoration}{Prix\ concurrence} \tag{2.2}$$

Si cette distance est supérieure à 1, il n'est pas exclu d'envisager que le client parte s'assurer à la concurrence. Et il est d'ailleurs fort probable que le client parte s'il avait effectivement cette information.

A défaut de ne pas pouvoir observer les tarifs affichés par le reste de la concurrence, cette méthode permet toutefois d'envisager le cas de figure où le tarif proposé par AXA au terme ne serait plus compétitif.

Le modèle utilisé pour la prédiction du prix de la concurrence a été sélectionné sur le critère du meilleur indice de Gini [4]. Plusieurs modèles GLM, différents en termes de variables explicatives employées, ont été entraînés sur la même base de données et testés en *Cross Validation* [5]. Un indice de Gini moyen est alors calculé à partir de ceux obtenus sur chaque *fold* [6] de la *Cross Validation*. Nous choisissons le modèle qui présente le plus grand indice de Gini moyen.

2.1.3.3 Ajout des variables réponses

La dernière étape de la construction de la base est d'indiquer pour chaque ligne les variables réponses, c'est-à-dire s'il y a eu une résiliation (émise ou prise d'effet) pendant la période d'observation, et si oui pour quel motif. Ces variables réponses sont construites grâce aux dates de résiliation récupérées *a posteriori*, *ie* à la fin de la période d'observation.

2.2 Analyse des statistiques descriptives

Avant d'élaborer nos modèles, la détermination de statistiques univariées et multivariées va nous permettre de mieux connaître la population observée, de détecter d'éventuelles tendances et informations importantes dans notre base de données, ce qui permettra de mieux diriger notre étude par la suite.

Une première difficulté survient : la base de données ainsi construite est très volumineuse, il y a plus de 5 millions d'observations et plus de 300 variables. Il ne serait ni judicieux ni efficace de faire des statistiques sur toutes ces variables. Une sélection préalable de ces dernières doit être réalisée pour limiter notre analyse aux variables les plus significatives, c'est-à-dire celles qui expliquent le mieux l'émission de résiliation. Par ailleurs,

4. Il s'agit d'un indicateur de la qualité de la segmentation des variables réponses. Cf. annexe 5
5. La Cross Validation (ou Validation Croisée) permet de déterminer la qualité moyenne de la prédiction et de vérifier la stabilité du modèle en réalisant l'apprentissage sur différents échantillons. Cf. annexe 4.
6. Il s'agit d'un "pli" de la Cross Validation. Plus de détails en annexe 4.

nous étudierons également les variables retenues suite à une discussion menée avec les experts, celles qui selon nous pourraient expliquer la résiliation tarifaire au terme (comme le profil locataire ou propriétaire de l'assuré, ou le nombre de sinistres associés au contrat).

Pour ce faire, nous allons réaliser une sélection de variables par *Gradient Boosting Machine*.

2.2.1 Sélection par *Gradient Boosting Machine* des variables pour la description de la base

2.2.1.1 Méthode du *Gradient Boosting Machine*

La méthode de *Gradient Boosting Machine* (GBM, ou boosting d'arbres de régression [7]), est une méthode supervisée de machine learning basée sur les arbres de décision. Nous rappelons la méthode du GBM et sa théorie en annexe 3, page 99.

Le GBM, bien qu'il s'agisse d'une "boîte noire", est capable d'entraîner un modèle de classification pour une variable réponse catégorielle, et va pouvoir indiquer quelles variables explicatives contribuent le plus à la prédiction de la variable cible en règle générale. A noter que le GBM prend en compte les corrélations linéaires entre les différentes variables explicatives, ce qui permet d'identifier les variables qui contribuent bien différemment à la prédiction de la variable cible.

2.2.1.2 Résultats du *Gradient Boosting Machine*

Nous décidons d'effectuer un GBM sur notre base de données en choisissant comme variable cible la **Résiliation émise tarifaire liée au terme** puisque que c'est cette résiliation que nous allons tenter de modéliser par la suite. A l'issue d'un *grid search* [8], le meilleur modèle GBM nous donne l'importance des différentes variables explicatives (en figure 2.4), c'est-à-dire le pouvoir explicatif d'une variable sur sa capacité à prédire la variable réponse :

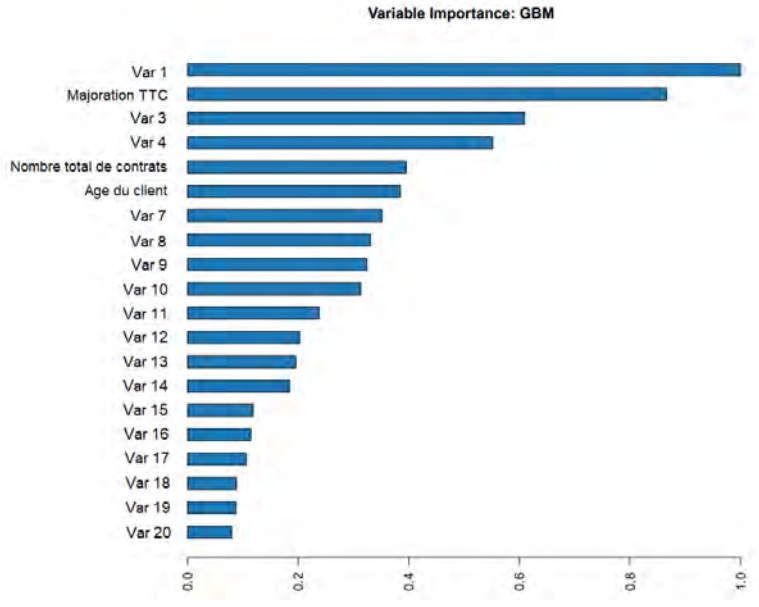


FIGURE 2.4 – Résultat de la sélection de variables par GBM

Nous retrouvons que le pourcentage de majoration TTC est fortement explicatif de la résiliation tarifaire au terme, ce qui est rassurant pour la suite de notre étude puisque nous souhaitons minimiser le taux de résiliation

7. A noter que cette méthode ainsi que ses résultats seront détaillés plus loin dans ce chapitre en section 2.5
8. Cf. explication du Grid search en annexe 3, page 99

en jouant sur la majoration. En outre, nous observons un certain nombre de variables pertinentes, notamment le nombre de contrats détenus par le client ainsi que son âge. Nous verrons dans l'analyse des statistiques descriptives comment ces variables influent sur la majoration.

2.2.2 Statistiques univariées : une résiliation pour quel motif ?

Selon les cas, les contrats ne sont pas résiliés pour les mêmes raisons. La figure 2.5 représente la répartition des motifs de résiliation. En 2017, un pourcentage non négligeable des contrats en portefeuille a été résilié, parmi lesquels la moitié ne répondaient pas à une augmentation du tarif, mais plutôt à des circonstances comme un déménagement [9], le décès du souscripteur, ou encore la décision d'AXA France de résilier ce contrat pour cause de non paiement de prime.

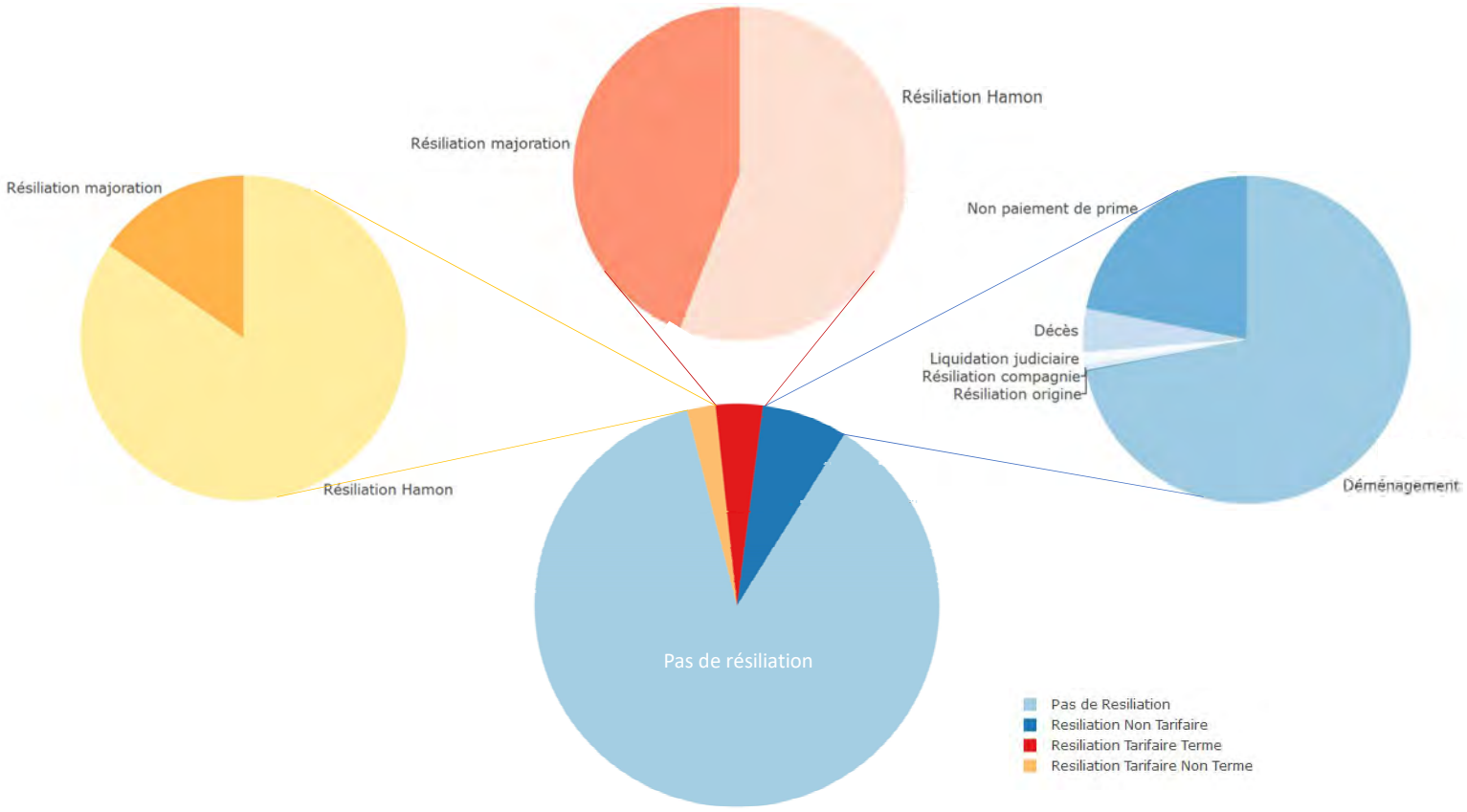


FIGURE 2.5 – Représentation de la répartition des résiliations en 2017

L'autre moitié des résiliations est indiquée comme tarifaire, soit pour cause de majoration, soit dans le cas d'une résiliation Hamon (la résiliation du contrat est dans ce cas signalée à AXA France par la nouvelle entité qui assurera le bien). Comme expliqué dans la section 2.1.1, nous considérons que la résiliation tarifaire est une réaction au terme si elle est émise dans une fenêtre [1 mois avant le terme ; 4 mois après le terme]. Ce type de résiliation représente un peu plus de la moitié des résiliations tarifaires. Nous remarquons d'ailleurs que cette fenêtre est pertinente car presque la moitié des résiliations tarifaires terme sont dues à la majoration, contre $1/6^{\text{ième}}$ pour les résiliations tarifaires non liées au terme.

9. A noter que la résiliation due à un déménagement peut aussi être liée au tarif, mais nous n'avons pas de moyen de le vérifier. Cela pourra faire l'objet d'une étude ultérieure.

2.2.3 Statistiques bivariées : évolution du taux de résiliation tarifaire liée au terme

2.2.3.1 Impact de la majoration

Compte tenu du sujet de notre étude, la première dimension à étudier est le taux de majoration : comment évolue le taux de résiliation tarifaire terme en fonction du taux de majoration appliqué au contrat ? La figure 2.6 confirme bien une augmentation du taux de résiliation tarifaire terme lorsque la majoration augmente. Il faut toutefois être prudent sur la significativité de la pente de la courbe : pour des raisons de confidentialité, l'échelle a volontairement été retirée. Ainsi nous retiendrons surtout la tendance croissante du taux de résiliation avec la majoration, ce qui renforce notre hypothèse que la résiliation due au terme est en partie expliquée par la majoration.

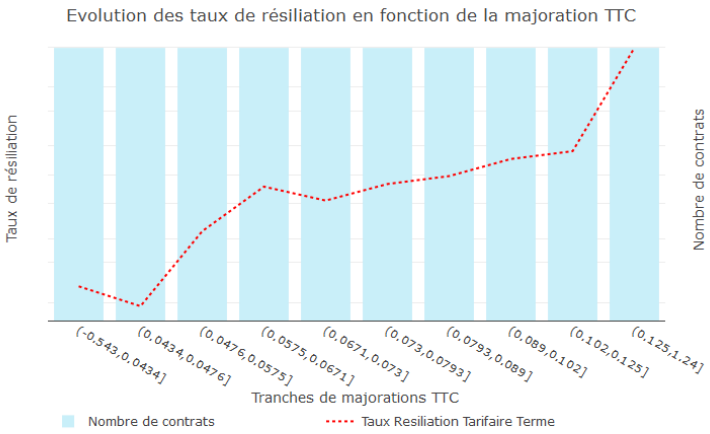


FIGURE 2.6 – Taux de résiliation terme en fonction de la majoration

2.2.3.2 Impact de l'*Expected Loss Ratio* (ELR)

Nous pouvons supposer qu'un souscripteur connaisseur et rationnel est conscient de la rentabilité de son contrat. En effet, s'il n'a pas eu de sinistres ou bien s'il sait qu'il est peu susceptible d'être sinistré, il pense de ce fait qu'il rapporte plus d'argent à l'assureur par rapport à ce que ce dernier pourrait lui indemniser. Ainsi, au vu de la majoration qui lui sera appliquée au terme et des prix très compétitifs en affaires nouvelles proposés pas la concurrence, il ne serait pas illogique que le client résilie son contrat pour souscrire chez un concurrent : en partant du principe que les concurrents établissent leurs tarifs sur la base de la prime pure du client (qui ne doit pas tant différer d'un assureur à un autre), leurs tarifs d'affaires nouvelles seront forcément plus intéressants qu'une prime AXA majorée.

L'ELR reflète la rentabilité du contrat : un contrat est d'autant plus rentable si son ELR est proche de zéro. La figure 2.7 indique bien une tendance des contrats en portefeuille à résilier davantage quand l'ELR est bas.

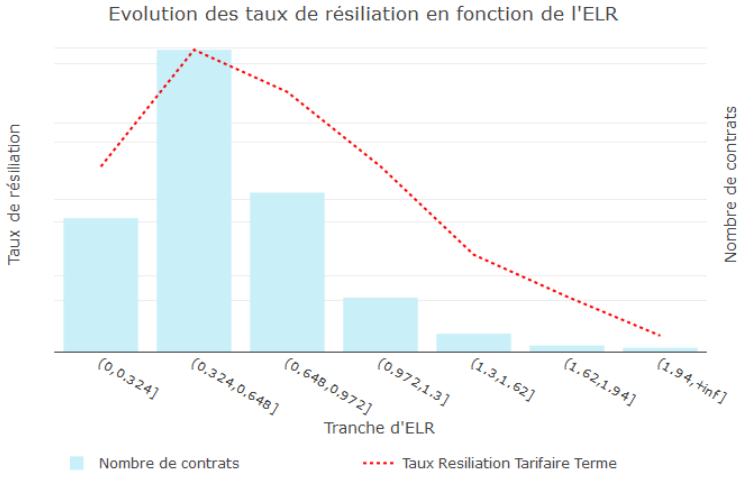
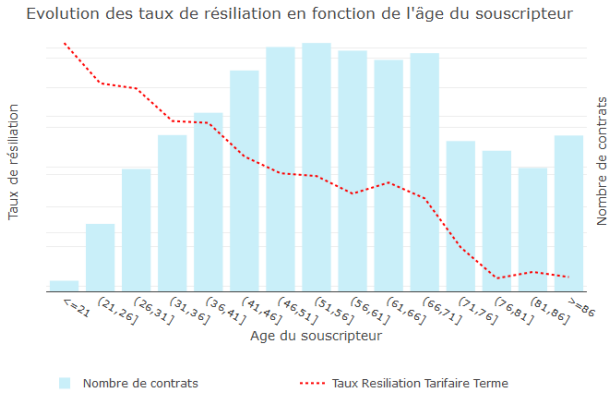


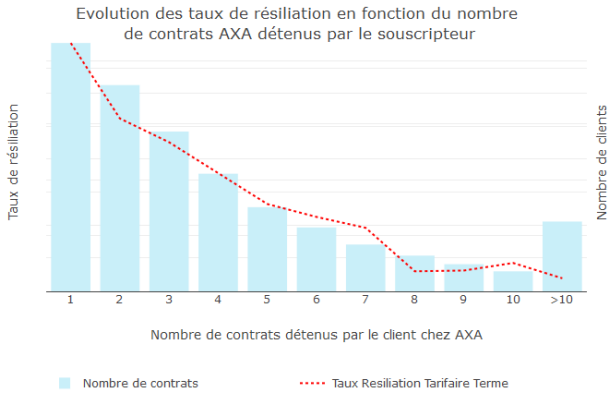
FIGURE 2.7 – Taux de résiliation terme en fonction de l'ELR avant application de la majoration

2.2.3.3 Impact de l'âge du client et du nombre de contrats AXA

Il est également intéressant de constater sur la figure 2.8 (a) que, plus le client est jeune, plus il aura tendance à résilier. Cela pourrait s'expliquer, d'une part du fait que les jeunes soient en grande partie des locataires d'appartement, et ce type de profil est par nature plus sensible au prix (nous le verrons d'ailleurs en figure 2.9), d'autre part par le fait qu'un jeune souscripteur soit plus renseigné sur les prix proposés par la concurrence (on peut citer notamment l'essor des comparateurs d'assurances sur internet qui facilitent le sondage des tarifs sur le marché) et en connaissance de cause décide de se faire assurer ou non chez un concurrent plus intéressant en terme de prime. Nous pouvons aussi envisager qu'AXA France est l'assureur historique des clients les plus seniors du portefeuille, lesquels sont confortables avec le fait de rester chez le même assureur et qui ont développé une relation de confiance avec l'agent général chez qui ils ont souscrit.



(a) Taux de résiliation terme en fonction de l'âge



(b) Taux de résiliation terme en fonction du nombre total de contrats

FIGURE 2.8 – Taux de résiliation terme en fonction de l'âge du client ou de son nombre de contrats

En outre, la figure 2.8 (b) indique que plus un client détient de contrats dans le groupe AXA, moins il a de chance de résilier. Ce phénomène pourrait s'expliquer par pénibilité administrative de résilier tous les contrats chez AXA et souscrire chez d'autres assureurs. Par ailleurs, nous constatons que la multidétention protège de la résiliation.

2.2.3.4 Impact du profil du client

Le comportement de l'assuré varie sensiblement selon qu'il soit locataire ou propriétaire, et s'il occupe un appartement ou une maison. La figure 2.9 indique que les propriétaires ont une probabilité de résilier inférieure à celle des locataires. Cela s'explique peut-être par une sensibilité au prix plus importante chez les profils locataires, et étant donné qu'en règle générale ces profils déménagent plus fréquemment que les propriétaires. Cette tendance soulève également la question des résiliations dues aux changements de domicile. En effet, bien que les résiliations non tarifaires (dont les motifs de déménagement) aient été retirées de notre périmètre d'étude, il n'est pas impossible que, le jour de l'émission de la résiliation, l'agent général renseigne par exemple dans la base de données le motif d'une résiliation "Hamon" par défaut s'il n'a pas connaissance du déménagement de l'assuré.

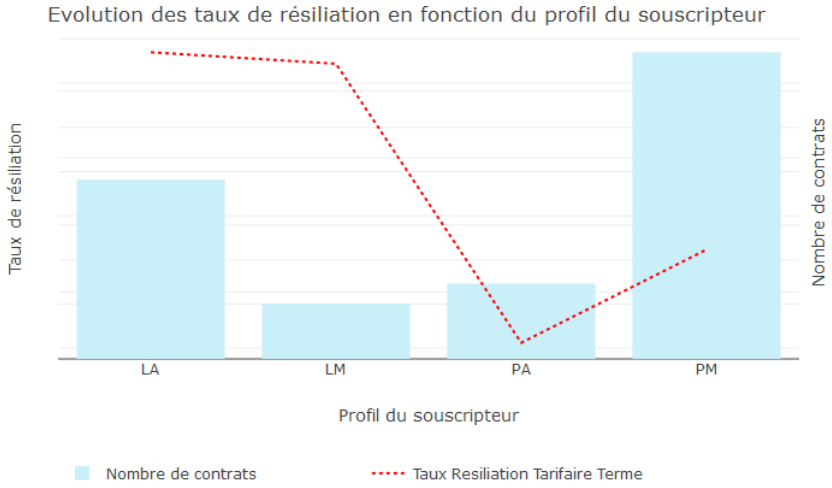


FIGURE 2.9 – Taux de résiliation terme en fonction du profil du client

2.2.3.5 Prédictions du *Competitive market analysis* (CMA)

Afin de vérifier la pertinence de la prédiction du tarif affaire nouvelle de la filiale d'AXA France, il est souhaitable de savoir pour chaque contrat de la base de données si le tarif **à l'issue du terme** est plus compétitif ou non par rapport à la prédiction du tarif concurrent (cf. distance CMA, définie par la formule (2.2)). Il sera alors possible d'observer la répartition des résiliations sur les contrats AXA France plus ou moins compétitifs :

Compétitivité d'AXA par rapport à la concurrence	Pas de résiliation	Résiliation	Part du portefeuille MRH d'AXA
Moins compétitif	96 %	4 %	89 %
Plus compétitif	97 %	3 %	11 %

TABLE 2.1 – Compétitivité des tarifs AXA France versus la part de résiliation

Il apparaît que parmi les cas où AXA France est moins compétitif par rapport à la concurrence [10], il y a en proportion plus de résiliations (1 point supplémentaire en proportion de résiliations par rapport au cas où AXA France est plus compétitif).

Il faut toutefois prendre ces chiffres avec précaution : étant donné que le modèle CMA a été entraîné sur une base composée de seulement 1600 individus, la valeur estimée du prix minimal proposé par la concurrence reste une indication. Toutefois, le modèle CMA ne donne pas de résultats totalement absurdes, mais ils ne sont pas non plus spectaculaires, ce qui laisse penser que la distance entre le tarif proposé par AXA et le tarif

10. Filiale d'AXA France

concurrentiel ne sera peut-être pas aussi significative que nous le pensions pour la prédiction de la probabilité d'un client de résilier son contrat. Cette hypothèse sera confirmée ou non plus tard par le modèle de résiliation.

2.2.3.6 Influence du comportement de l'agent général

Afin de mieux comprendre l'impact des agents généraux et d'introduire la dimension agent dans ses analyses, la direction de l'offre a développé une segmentation de ces derniers en 4 groupes :

- **Groupe 1** : ces agents ont une bonne dynamique commerciale et observent une croissance de chiffre d'affaires en assurance Auto et MRH. Cependant ils présentent de mauvais résultats techniques.
- **Groupe 2** : ces agents sont orientés rentabilité et ont une bonne maîtrise des coûts. Toutefois leur dynamique commerciale est à renforcer.
- **Groupe 3** : ces agents voient la croissance de leur chiffre d'affaires ralentir. Ils allouent trop de crédit commercial (*ie* de rabais commerciaux) et n'optimisent pas assez l'utilisation de leur budget compte tenu de la structure de leur production.
- **Groupe 4** : la production de ces agents est en perte de croissance et ces derniers ont du mal à retenir leurs clients. Ils présentent également de mauvais résultats techniques.

Remarque : Certains agents généraux ne sont associés à aucune catégorie. Cela provient du fait qu'il y ait de nouveaux agents ou bien qu'il y ait eu des transferts ou fusions de portefeuilles de clients entre agents depuis la réalisation de la dernière segmentation.

Nous constatons donc sur la figure 2.10 que le taux de résiliation est important chez les agents du groupe 4, ce qui est cohérent avec la définition même de cette catégorie (ces agents ont du mal à retenir leurs clients). En outre, le taux de résiliation est minimal chez les agents du groupe 2 (peut-être grâce à une bonne gestion de leur crédit commercial pour piloter leur portefeuille) et chez les agents du groupe 3 (qui doivent sur-consommer du crédit commercial justement pour retenir leurs clients).

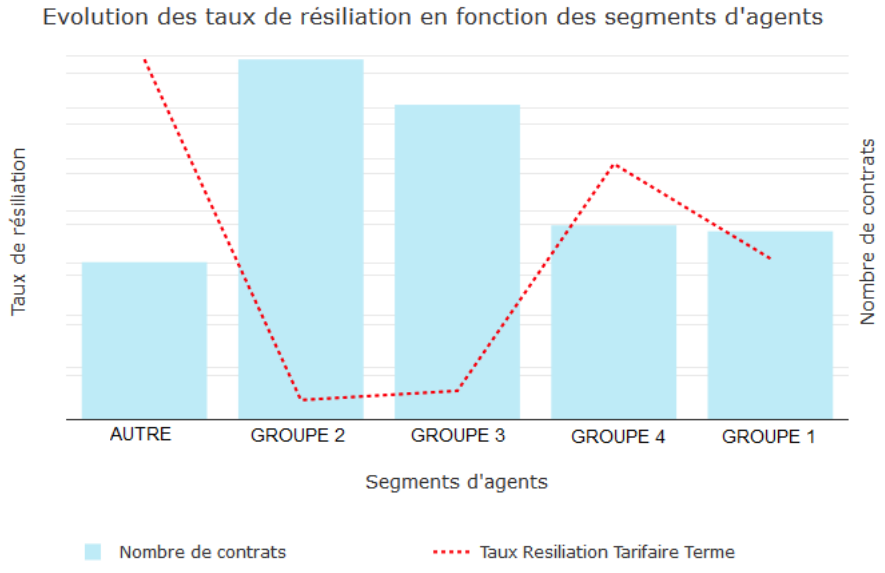


FIGURE 2.10 – Taux de résiliation terme en fonction de la catégorie d'agents

2.2.3.7 Observation de la saisonnalité du taux de résiliation liée au terme

Rappelons que nous avons retiré de notre périmètre d'étude les propriétaires non occupants ainsi que les étudiants. Nous savons notamment que les étudiants souscrivent et résilient en grande majorité pendant l'été, période des grandes vacances scolaires. Toutefois, pour les autres assurés, nous souhaitons déterminer si la

résiliation tarifaire liée à la majoration est plus importante selon certaines périodes de l'année. La figure 2.11 indique que non : par rapport au taux de résiliation global, la résiliation due au terme est relativement stable.

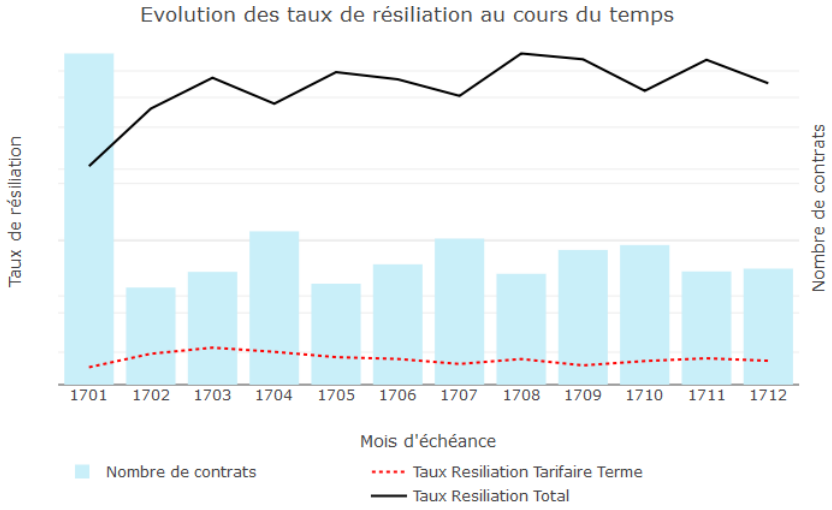


FIGURE 2.11 – Taux de résiliation terme au cours de l'année

2.2.4 Etude des corrélations linéaires et V de Cramer

Pour compléter notre étude descriptive des variables qui constituent notre base de données, il est utile d'observer les corrélations linéaires et les associations entre elles. Notre base étant constituée de variables mixtes, nous allons comparer les variables continues entre elles en établissant leurs corrélations, puis comparer les variables catégorielles entre elles en déterminant leurs V de Cramer. Les calculs de la corrélation ainsi que du V de Cramer sont rappelés en annexe 7. Cette étape est d'autant plus importante que le logiciel de modélisation linéaire généralisée utilisé est une version *beta* qui ne prend pas encore en compte les effets des variables corrélées pour la modélisation logistique.

A la vue des matrices de corrélations et de V de Cramer, la base de données présente des variables redondantes étant donné que certaines informations sur les clients et les contrats se recoupent (montants de primes et informations sur les sinistres notamment). Ainsi, les variables continues présentent une forte corrélation en valeur absolue, et les variables catégorielles un V de Cramer important.

2.3 Modélisation linéaire généralisée de la résiliation

L'objectif du modèle de résiliation est de déterminer la probabilité pour chaque contrat de résilier pour une cause tarifaire liée au terme. Il s'agit en quelque sorte de modéliser une fonction complémentaire à la demande : en fonction de la majoration qui lui sera appliquée, le client souhaite-t-il rester assuré chez AXA ou non ? Bien sûr, cette fonction de résiliation est croissante avec la majoration (et nous l'avons constaté lors de l'analyse des statistiques descriptives) mais elle n'est pas unique : elle varie suivant les caractéristiques du client et de son contrat. Nous pouvons citer par exemple l'âge du client, le fait qu'il soit multidétenteur ou pas, ou encore le prix estimé de son contrat chez la concurrence.

Appelons $Y = (Y_1, ...Y_n)'$ le vecteur colonne représentant les différentes valeurs observées de notre variable réponse, où chaque Y_i est l'indicatrice d'une résiliation tarifaire liée au terme. Y_i ne peut prendre que deux valeurs : **1** si le contrat est résilié pour cause tarifaire liée au terme, **0** sinon. Appelons également $X_i = (1, X_{i,1}, ..., X_{i,p})'$ le vecteur colonne représentant les p variables explicatives pour la variable réponse Y_i. X est alors une matrice de taille $n \times (p+1)$ dont les lignes sont les vecteurs lignes X_i'.

$Y|X$ suit donc une loi de Bernoulli, et prend la valeur 1 avec la probabilité $\pi(x)$ dépendant des variables explicatives.

Nous cherchons alors à modéliser la probabilité conditionnelle suivante pour chaque contrat i :

$$\pi_i(x) = \mathbb{P}(Y_i = 1|X_i = x) \tag{2.3}$$

2.3.1 Préparation de la base de données

Pour réaliser notre modèle de résiliation, il n'est pas judicieux d'utiliser la base de données exhaustive que nous avons élaborée. En effet, cette base n'a pas seulement été construite pour le modèle de résiliation mais aussi pour le modèle de crédit commercial que nous réaliserons par la suite. En outre, cette base est très volumineuse et possède un très grand nombre de variables, dont certaines ne sont pas pertinentes pour le modèle de résiliation.

Ainsi, la base initiale sera modifiée de la façon suivante afin de rendre la modélisation plus efficace :

— Les contrats résiliés pour cause tarifaire liée au terme sont signalés par un indicateur 1, tandis que tous les autres contrats présentent l'indicateur 0. Cet indicateur sera la variable réponse de notre modèle de résiliation.

— Nous avons retiré certaines variables qui engendrent des corrélations linéaires ou des V de Cramer importants (coefficient de corrélation supérieur à 0.60 en valeur absolue, et V de Cramer supérieur à 0.30).

— Sont également retirées les variables *a posteriori* (c'est-à-dire les variables indiquant une information obtenue après la décision du client de résilier son contrat ou non) comme les crédits commerciaux accordés aux clients en réaction au terme par exemple.

— Les variables non tarifaires, comme le genre du client, sont retirées, en accord avec la législation en vigueur.

2.3.2 Generalized Linear Model (GLM, ou Modèle Linéaire Généralisé)

Nous souhaitons dans un premier temps modéliser la probabilité de résiliation tarifaire due au terme par une régression linéaire généralisée, méthode que nous expliquons en détail en annexe 4. Cette méthode a l'avantage de générer, après un apprentissage, des prédictions faciles à interpréter grâce aux paramètres qui sont des coefficients multiplicateurs. Le GLM a également un grand intérêt opérationnel : il est possible de construire aisément un programme de prédiction à partir des variables explicatives, avec des coefficients obtenus et selon une fonction de lien appropriée.

Dans le cadre de notre étude, nous utilisons un nouveau logiciel de modélisation GLM : *Akur8*. Développé par l'entreprise du même nom [11], il s'agit d'une version beta adoptée récemment par les actuaires d'AXA France, après avoir été validée par plusieurs séries de tests en comparant les résultats ainsi que l'ergonomie du logiciel par rapport à Emblem [12]. Ce logiciel réalise un GLM pénalisé [13]. Pour des raisons de confidentialité, nous ne communiquons pas le type de pénalisation employé par *Akur8*. Cela n'empêche pas les actuaires non-vie d'AXA France à avoir largement recours à l'assistance d'un tel logiciel par souci d'efficacité.

2.3.3 La régression logistique pour modéliser la probabilité de résiliation

Comme $Y_i|X_i$ suit une loi de Bernoulli de paramètre $\pi_i(x)$, son espérance est donc égale à la probabilité $\pi_i(x)$:

11. L'entreprise Akur8 s'est développée au sein de l'incubateur Kamet, détenu par le groupe AXA, et spécialisé dans le développement de start ups en *InsurTech*.

12. Logiciel de modélisation GLM anciennement utilisé par les équipes d'actuariat non-vie chez AXA.

13. Dans le cadre d'un GLM, la pénalisation permet de sélectionner les variables les plus significatives et/ou de limiter la valeur de leur coefficient. Plus de détails en annexe 4.

$$\mathbb{E}[Y_i|X_i = x] = \pi_i(x) \tag{2.4}$$

Rappelons que la fonction de densité d'une variable aléatoire Y_i suivant une loi de Bernoulli, de probabilité de succès p_i est de la forme :

$$f(y_i, p_i) = p_i^{y_i}(1-p_i)^{1-y_i} = (1-p_i)\exp\left(y_i \ln\frac{p_i}{1-p_i}\right) \tag{2.5}$$

Cette densité appartient bien à une famille exponentielle en remarquant que :

$$\theta_i = \ln\frac{p_i}{1-p_i}\ ,\ a(\theta) = \ln(1-e^{\theta})\ ,\ \phi = 1 \tag{2.6}$$

avec θ le paramètre canonique, et ϕ le paramètre de dispersion tels que décrits en annexe 4.

Le principe de la régression est de modéliser l'espérance $\mathbb{E}[Y_i|X_i]$, c'est-à-dire la probabilite $\pi_i(x)$. Dans le cadre d'une loi de Bernoulli, il est possible de réaliser un modèle linéaire généralisé de la forme :

$$g(\pi_i(x)) = X_i\beta = \beta_0 + X_{i,1}\beta_1 + ... + X_{i,p}\beta_p \tag{2.7}$$

avec β le vecteur colonne correspondant aux $p+1$ coefficients du modèle et g la fonction de lien canonique :

$$g(\pi_i(x)) = \theta_i = logit(\pi_i(x)) = \ln\left(\frac{\pi_i(x)}{1-\pi_i(x)}\right) \tag{2.8}$$

Il s'agit d'une **régression logistique**, puisqu'en effet nous obtenons la fonction de répartition de la loi logistique :

$$\pi_i(X) = \frac{e^{X_i\beta}}{1-e^{X_i\beta}} \tag{2.9}$$

En supposant les (Y_i, X_i) indépendants et identiquement distribués, pour tout $i \in 1, ..., n$ (il est en effet supposé dans notre étude que les clients ne se concertent pas pour prendre leur décision de résiliation, et donc que chaque décision est indépendante. Notons qu'il s'agit d'une grosse hypothèse : en effet, les clients partagent facilement leurs avis et habitudes de consommation sur les blogs et réseaux sociaux.), la log-vraisemblance peut s'écrire de la façon suivante :

$$\begin{array}{rl} \mathcal{L}_n(Y|X,\beta) = & \ln\prod_{i=1}^{n}(g^{-1}(X_i\beta)^{Y_i}(1-g^{-1}(X_i\beta))^{1-Y_i}) \\ = & \sum_{i=1}^{n} Y_i \ln(g^{-1}(X_i\beta)) + (1-Y_i)\ln(1-g^{-1}(X_i\beta)) \end{array} \tag{2.10}$$

L'objectif est donc de trouver β qui maximise cette log-vraisemblance, c'est-à-dire β solution de l'équation donnée par la condition du premier ordre :

$$\frac{\partial\mathcal{L}_n(Y|X,\beta)}{\partial\beta} = 0 \tag{2.11}$$

2.3.4 Modélisation GLM de la résiliation sur Akur8

2.3.4.1 Choix du modèle le plus pertinent

Le logiciel Akur8 est capable de réaliser une telle régression logistique pénalisée. Nous lui proposons un jeu varié de variables explicatives non corrélées susceptibles d'expliquer la résiliation tarifaire due au terme. La pénalisation va permettre de sélectionner les variables qui expliquent le mieux cette résiliation. Un *grid search* ("Recherche sur grille") est automatiquement réalisé par le logiciel sur des combinaisons et des nombres

différents de variables. La figure 2.12 indique ainsi la valeur du Gini [14] moyenne obtenue par "validation croisée stratifiée" [15] en fonction du nombre de variables. En outre, pour un nombre donné de variables, les valeurs de Gini de différentes combinaisons sont affichées.

FIGURE 2.12 – Résultat du Grid Search pour le GLM

Assurément, la cross validation permet de déterminer la qualité moyenne de la prédiction du modèle et permet de vérifier que ce dernier est suffisamment stable.

Nous choisissons le modèle qui réalise le meilleur compromis entre un nombre raisonnable de variables utilisées (car plus les variables sont nombreuses dans le modèle, plus le calcul de la prédiction peut devenir lourd), une valeur du Gini qui doit être la plus élevée possible, ainsi qu'un ordre pertinent des variables significatives. En effet, étant donné que nous souhaitons par la suite étudier l'impact de la majoration sur la résiliation, il est souhaitable de choisir un modèle avec une majoration ayant un impact le plus important possible. En ce qui concerne le nombre de variables, nous répondons au principe de parcimonie, également appelé *Rasoir d'Occam*, qui stipule que "les multiples ne doivent pas être utilisés sans nécessité", autrement dit, il vaut mieux choisir les hypothèses les plus simples, car rajouter de la complexité n'améliora pas tant le résultat.

Ainsi, notre choix se porte sur un modèle à 13 variables, dont la valeur du Gini moyen est suffisamment élevée, et dont la majoration, toutes taxes comprises, est la variable la plus significative. Ce modèle est indiqué par la flèche "best model" sur la figure 2.12.

Il est important toutefois de noter que notre "best model" présente un Gini de près de 3 points inférieur à ceux des modèles à 28 variables. Or, ces modèles n'indiquent pas la majoration comme variable la plus significative, car d'autres variables légèrement corrélées à la majoration se sont ajoutées. Le Gini augmente donc mécaniquement puisque l'ajout d'une variable supplémentaire à un modèle de régression permet d'expliquer davantage la variable réponse et permet donc d'améliorer la prédiction.

14. Il s'agit d'un indicateur de la qualité de la segmentation des variables réponses. Cf. Annexe 5
15. Technique d'évaluation de la qualité et de la stabilité de la prédiction, détaillée en annexe 8.

2.3.4.2 Pertinence des coefficients de la régression logistique

Le modèle choisi indique, pour chaque variable, des valeurs de coefficients pour chaque catégorie [16]. Pour les variables initialement continues, les catégories sont ordonnées.

Par ailleurs, nous précisons que, tandis que la qualité de chaque modèle proposé par Akur8 est évaluée par cross validation, les coefficients β du modèle choisi sont eux estimés sur la totalité de la base de données.

Nous nous assurons que, pour chaque variable, la tendance des coefficients suit la tendance du taux de résiliation. En effet, la différence entre deux coefficients associés à des catégories différentes détermine si un individu de la première catégorie a plus ou moins de chance de résilier par rapport à un individu de la deuxième catégorie. Cela se démontre facilement avec la notion d'*odds ratio* pour la régression logistique [17].

L'*odds* est la chance pour un individu x d'obtenir la réponse $Y = 1$, et est défini par :

$$odds(x) = \frac{p(x)}{1 - p(x)}, \text{ avec } p(x) = \mathbb{P}(Y = 1|X = x) \ . \tag{2.12}$$

L'*odds ratio*, c'est-à-dire le rapport des chances entre deux individus x_1 et x_2 est donné par :

$$OR(x_1, x_2) = \frac{odds(x_1)}{odds(x_2)} = \frac{\frac{p(x_1)}{1-p(x_1)}}{\frac{p(x_2)}{1-p(x_2)}} \ . \tag{2.13}$$

Sachant que, dans le cadre d'une régression logistique, la probabilité prédite pour un individu x à partir des coefficients β est donnée par la relation :

$$\text{logit} p_\beta(x) = \beta X \tag{2.14}$$

$$\iff \frac{p_\beta(x)}{1 - p_\beta(x)} = e^{\beta X} \tag{2.15}$$

nous pouvons en déduire l'odds ratio pour deux individus uniquement différenciés par des catégories différentes sur une variable, en supposant que β_1 est le coefficient associé à la catégorie à laquelle appartient l'individu x_1, et β_2 celui correspondant à la catégorie dont dépend l'individu x_2 :

$$OR(x_1, x_2) = \frac{odds(x_1)}{odds(x_2)} = \frac{\frac{p(x_1)}{1-p(x_1)}}{\frac{p(x_2)}{1-p(x_2)}} = e^{\beta_1 - \beta_2} \tag{2.16}$$

Ainsi, si $\beta_1 > \beta_2$, alors $OR(x_1, x_2) > 1$, ce qui indique que l'individu x_1 a plus de chance d'obtenir la réponse $Y = 1$ que l'individu x_2.

En ce qui concerne la variable "Majoration TTC", nous observons dans les statistiques descriptives qu'un individu appartenant à une tranche élevée de majoration a plus de chance de résilier qu'un individu d'une tranche faible de majoration. Le coefficient de la tranche élevée doit donc être plus élevé que celui de la tranche faible : la tendance des coefficients doit donc être la même que celle du taux de résiliation. C'est effectivement ce nous observons sur la figure 2.13.

Cette même figure présente aussi la qualité de la prédiction du taux de résiliation par rapport à celui observé par tranche de majoration.

16. En effet, le logiciel ne réalise pas directement de régression sur les variables continues. Il faut les catégoriser en amont.
17. D'après le cours de Laurent Rouvière, de l'Université de Rennes [10]

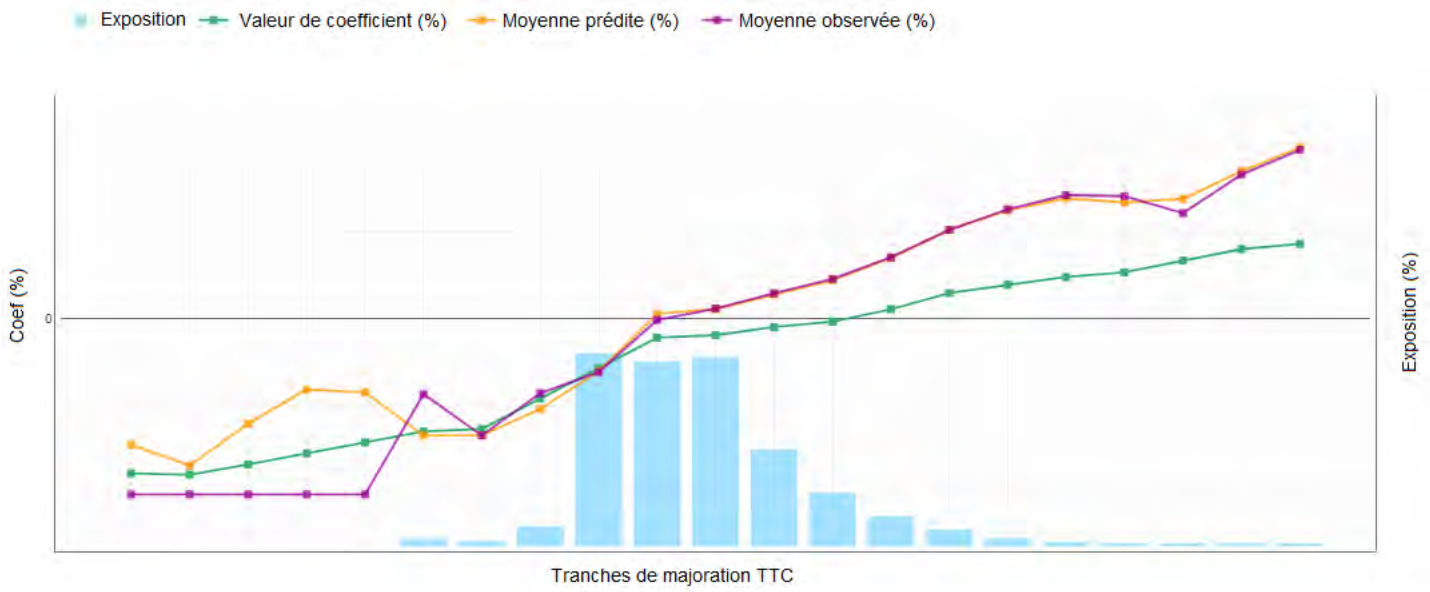


FIGURE 2.13 – Tendance des coefficients associés à la variable "Majoration TTC" après modélisation GLM

En outre, Akur8 nous permet de nous assurer de la stabilité des coefficients dans le temps. C'est une autre manière de vérifier la robustesse de notre modèle et de le valider. La figure 2.14 présente les coefficients obtenus si la modélisation n'avait été réalisée que sur les données de 2017 ou que sur celles de 2018. Sur ces deux années, la tendance est sensiblement la même, nous en déduisons que le modèle est stable vis-à-vis de la majoration TTC.

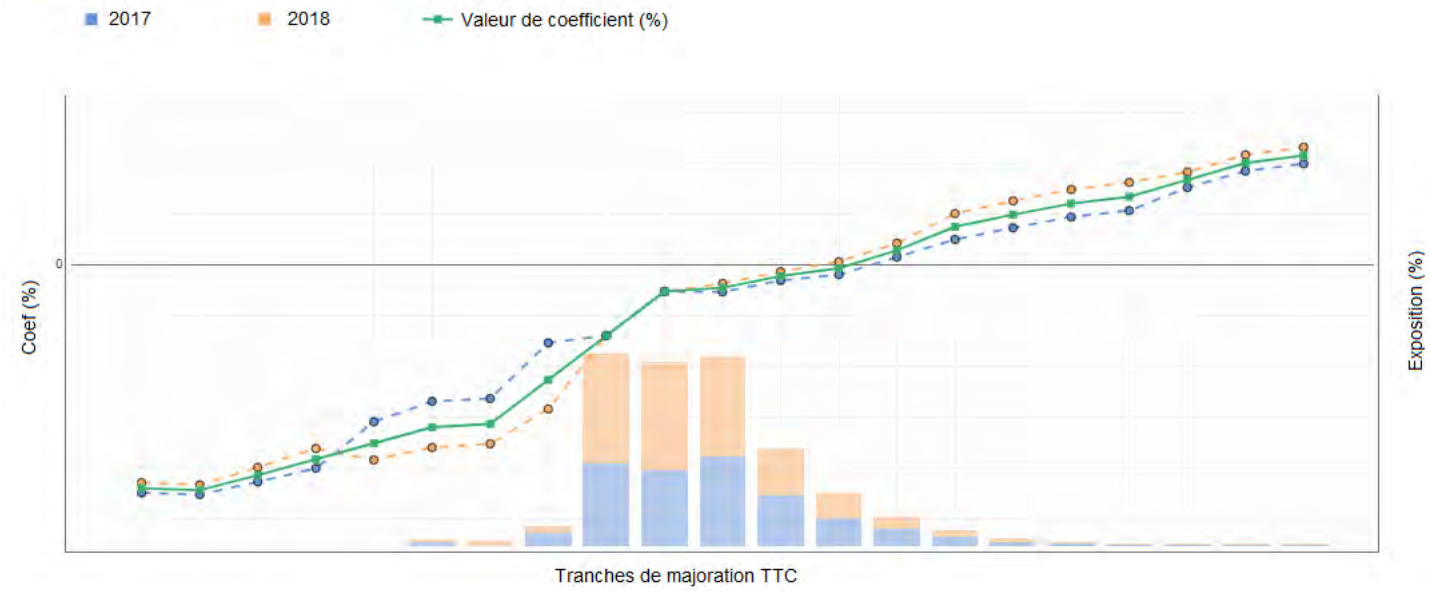


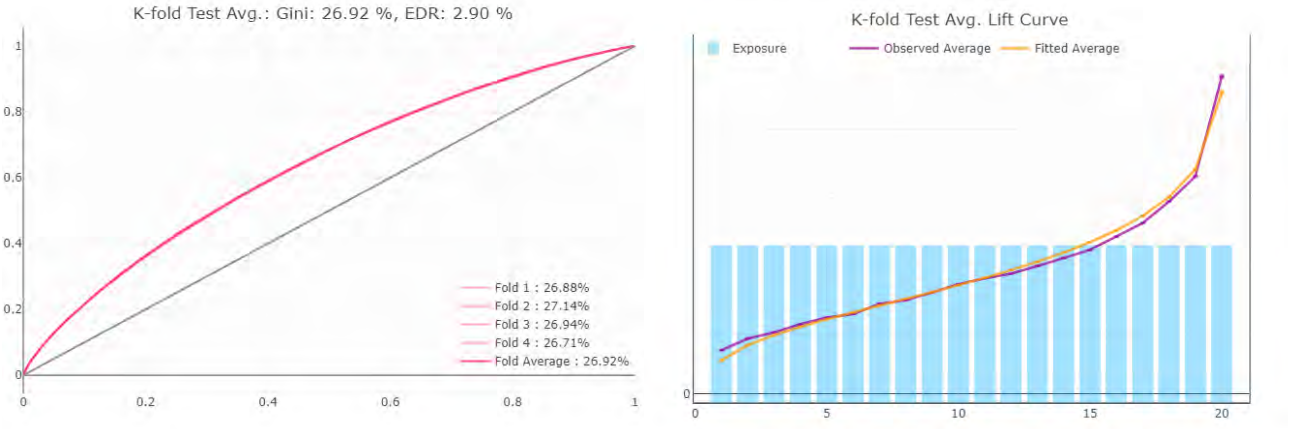
FIGURE 2.14 – Stabilité dans le temps des coefficients associés à la variable "Majoration TTC"

2.3.4.3 Extrapolation des coefficients pour les majorations négatives

AXA France applique des majorations majoritairement positives, ce qui induit forcément une prédiction discutable sur les tranches de majorations négatives. Toutefois, étant donné que nous souhaitons savoir si l'application de majorations négatives sur certains contrats améliorerait l'optimisation du taux de résiliation dû au terme sous contrainte de l'ELR, nous décidons d'extrapoler de façon presque linéaire les coefficients sur ces tranches de majoration, en maintenant leur tendance croissante. Bien entendu, nous n'avons aucun moyen actuellement de vérifier si cette tendance imposée est légitime, cela nécessiterait un test à part entière. La figure 2.13 représente bien ces coefficients extrapolés.

2.3.4.4 Analyse de la qualité du modèle

La moyenne des Gini sur le 4-fold [18] du modèle choisi est de 26,92%. Cette valeur est relativement élevée par rapport aux autres modèles proposés par Akur8, ce qui nous satisfait. De plus, la *lift curve* représentée en figure 2.15 (b) qui affiche, pour chaque quantile de la base de données, les valeurs moyennes observées (en violet) et prédites (en orange) du taux de résiliation tarifaire liée au terme, indique que ces valeurs ne sont pas très éloignées. Malgré la légère sous-estimation du taux de résiliation pour les premiers quantiles et la légère sur-estimation pour les derniers quantiles, la petite différence entre les valeurs prédites et observées n'est pas problématique car dans l'ensemble, la tendance de la lift-curve est bien conservée.



(a) Courbes ROC du modèle de résiliation

(b) Lift curve du modèle de résiliation

FIGURE 2.15 – Graphiques indicateurs de la qualité de la prédiction

Quant à la figure 2.15 (a), elle illustre les courbes ROC [19] sur chaque fold réalisé. Ces courbes sont bien superposées, et d'ailleurs, les valeurs de Gini correspondantes sont très proches. Cela indique que la qualité de la prédiction par le modèle est stable.

2.3.4.5 Linéarisation des coefficients pour la majoration

Comme nous l'avons signalé précédemment, Akur8 ne réalise pas directement de régression sur les variables continues : une catégorisation est réalisée en amont sur ces variables. C'est pourquoi nous avons un coefficient par catégorie pour chaque variable. Cela pose problème pour la variable "Majoration TTC". En effet, la fonction de probabilité de résiliation par rapport à la majoration (variable d'intérêt) est alors constante par morceaux et discontinue. Or nous souhaitons étudier par la suite l'élasticité à la majoration de notre modèle, ce qui ne peut

18. Il est souhaitable d'évaluer la stabilité du modèle en utilisant la méthode de *cross validation* qui va, à partir de la base de données, partager les données en k échantillons (ou k-plis, k-folds en anglais), chaque échantillon servant tour à tour de test après avoir généré l'apprentissage sur les k-1 autres échantillons.
19. Le principe de la courbe ROC est rappelé en annexe 5.

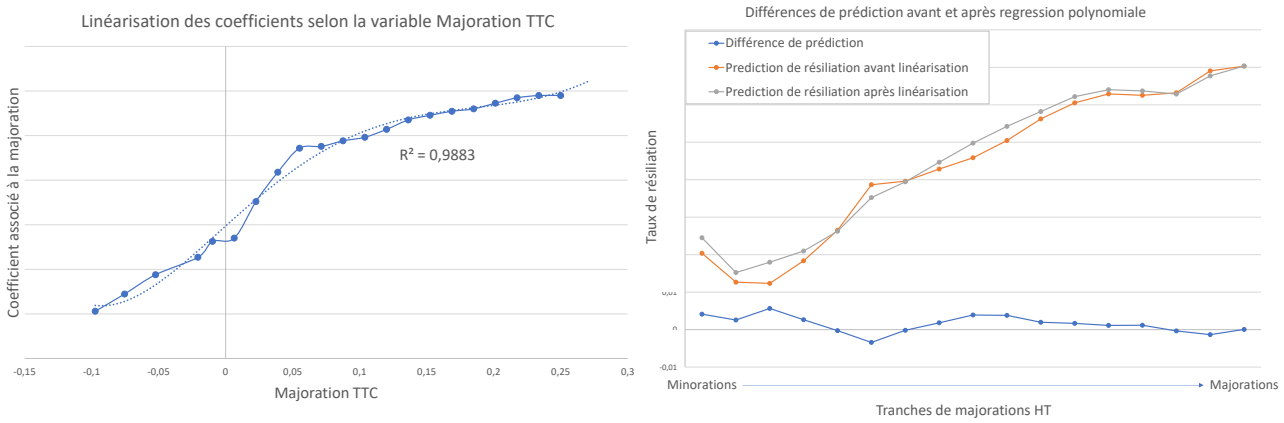
pas être réalisé si la probabilité de résiliation reste la même pour deux niveaux de majorations très proches. Par ailleurs, cela n'est pas non plus idéal pour le futur travail d'optimisation, qui ne pourra proposer que des plages de majorations en guise d'optimum.

Pour pallier cet inconvénient, nous décidons de réaliser une régression polynomiale sur les coefficients obtenus par Akur8. Il s'agit basiquement d'une régression linéaire multiple que nous réalisons sur Excel : il s'agit d'expliquer la valeur Y_i d'un coefficient en fonction des $X_{i,p} = X_i^p$, X_i étant la majoration d'une observation i :

$$Y_i = a_0 + a_1 \cdot X_{i,1} + \ldots + a_{n-1} \cdot X_{i,n-1} + a_n \cdot X_{i,n} + \varepsilon_i \tag{2.17}$$

Nous obtenons un polynôme de degré 5. Nous évaluons la qualité de cette régression par le coefficient de détermination R^2 qui exprime le rapport entre la variance expliquée par le modèle et la variance totale. Plus R^2 est proche de 1, plus la variance du modèle explique la variance totale, et meilleure est la qualité de la régression.

Excel nous signale un R^2 de 0,9883 comme nous l'indique la figure 2.16 (a), ce qui indique que la régression est très satisfaisante. Par ailleurs, la figure 2.16 (b) représente la différence de prédictions avant et après avoir effectué la linéarisation polynomiale des coefficients. Sur les différentes tranches de majoration, la différence fluctue autour de zéro, et la racine de la moyenne des erreurs au carré est égale à 2×10^{-3}, ce qui reste petit par rapport aux taux de résiliation observés.



(a) Régression polynomiale sur les coefficients de majoration TTC (b) Différence de prédiction avant et après linéarisation

FIGURE 2.16 – Résultats de la régression polynomiale sur les coefficients GLM de majoration

Enfin, si nous nous concentrons sur l'allure de la fonction polynomiale, qui s'accorde sur la tendance des coefficients d'Akur8 en figure 2.16 (a), nous remarquons qu'elle suit globalement bien les variations de pente dessinées par les coefficients. Cela est important car l'élasticité est représentée par la pente de cette fonction. Ainsi, nous voyons par exemple que la pente des coefficients est maximale entre 0% et 5% de majoration, et c'est également le cas pour la fonction polynomiale. Nous pouvons donc être sereins lorsque nous étudierons l'élasticité sur cette fonction, car nous savons que l'élasticité de la fonction polynomiale n'est pas éloignée de celle du modèle.

2.4 Evaluation de l'élasticité du modèle GLM

2.4.1 Définition de l'élasticité de la résiliation

L'étude de l'impact de la majoration sur le taux de résiliation équivaut en quelque sorte à une étude de la fonction de demande par rapport à la majoration. Il nous est donc possible, à partir de la fonction de résiliation que nous avons modélisée, d'évaluer la sensibilité des clients à l'augmentation de leur prime P. Usuellement, cela se détermine avec l'élasticité-prix de la demande $\epsilon(P)$:

$$\epsilon(P) = -\frac{\partial d(P)}{d(P)} \frac{P}{\partial P} \tag{2.18}$$

avec d la fonction de demande. Or, dans le cadre de notre étude, nous nous intéressons plutôt à l'impact de la variation de majoration sur la demande. Nous définissons donc l'élasticité-majoration de la demande :

$$\epsilon(x) = -\frac{\Delta d(x)}{d(x)} \frac{x}{\Delta x} \tag{2.19}$$

avec x la majoration TTC. Nous prenons la majoration TTC puisque c'est bien cette majoration que le client perçoit. Selon cette définition, l'élasticité ϵ est bien positive, puisque la demande d est décroissante avec la majoration. Elle exprime donc de combien de pourcent la demande diminue lorsque la majoration augmente de 1% par rapport à sa valeur initiale [20].

De manière équivalente, étant donné que la fonction de résiliation f est complémentaire à la fonction de demande, nous définissons l'élasticité-majoration vis-à-vis de la résiliation :

$$\epsilon_f(x) = +\frac{\Delta f(x)}{f(x)} \frac{x}{\Delta x} \ . \tag{2.20}$$

Ici, $\epsilon_f(x)$ est aussi positive, puisque le taux de résiliation augmente lorsque la majoration augmente. Elle exprime le pourcentage d'évolution de la résiliation lorsque la majoration varie de 1%. Ainsi, si $|\epsilon_f(x)| > 1$, alors la résiliation est élastique au prix, c'est-à-dire que l'individu est sensible à la variation de majoration ; si $|\epsilon_f(x)| < 1$, alors la résiliation est relativement inélastique au prix, et dans ce cas, l'individu est peu sensible à la variation de majoration.

2.4.2 Détermination de l'élasticité de la résiliation modélisée

Pour calculer $\epsilon_{\widehat{f}}(x)$, l'élasticité-majoration TTC de la résiliation modélisée $\widehat{f}$, nous "choquons" la majoration de chacun des contrats i d'un pourcentage γ et observons l'impact sur la fonction de résiliation. La majoration choquée x_{choc} est telle que :

$$x_{choc,i} = x_i(1 + \gamma_i), \quad \text{avec } \gamma_i \times 100 \sim \mathcal{N}(0, 1) \tag{2.21}$$

et les γ_i sont indépendants identiquement distribués.

Nous obtenons une distribution normale centrée en zéro des chocs de majoration que nous arrondissons au centième près.

Plus concrètement, à partir de notre base de données qui répertorie des contrats i ayant subi une majoration x_i, nous allons comparer la probabilité de résiliation $\widehat{f_i}$ de chaque contrat i lorsque ce dernier est majoré de x_i et lorsque qu'il est majoré de $x_{choc,i}$:

20. La majoration est elle-même exprimée en pourcentage

Numéro de contrat	Majoration initiale	Prédiction de résiliation avec majoration initiale	Majoration choquée	Prédiction de résiliation avec majoration choquée
1	x_1	$\widehat{f}_1(x_1)$	$x_{choc,1} = x_1(1+\gamma_1)$	$\widehat{f}_1(x_{choc,1})$
...	...	...	...	...
i	x_i	$\widehat{f}_i(x_i)$	$x_{choc,i} = x_i(1+\gamma_i)$	$\widehat{f}_i(x_{choc,i})$
...	...	...	...	...
N	x_N	$\widehat{f}_N(x_N)$	$x_{choc,N} = x_N(1+\gamma_N)$	$\widehat{f}_N(x_{choc,N})$

Issu de la base de données

TABLE 2.2 – Application des chocs de majoration sur les contrats de la base

Nous calculons ensuite, l'élasticité de la résiliation estimée à la majoration selon la formule :

$$\epsilon_{\widehat{f}_i}(x_i) = \frac{\widehat{f}_i(x_{choc,i}) - \widehat{f}_i(x_i)}{\widehat{f}_i(x_i)} \times \frac{x_i}{x_{choc,i} - x_i} \ . \tag{2.22}$$

Puis nous calculons la moyenne des élasticités obtenues sur chaque tranche de majoration pour obtenir le résultat illustré en figure 2.17.

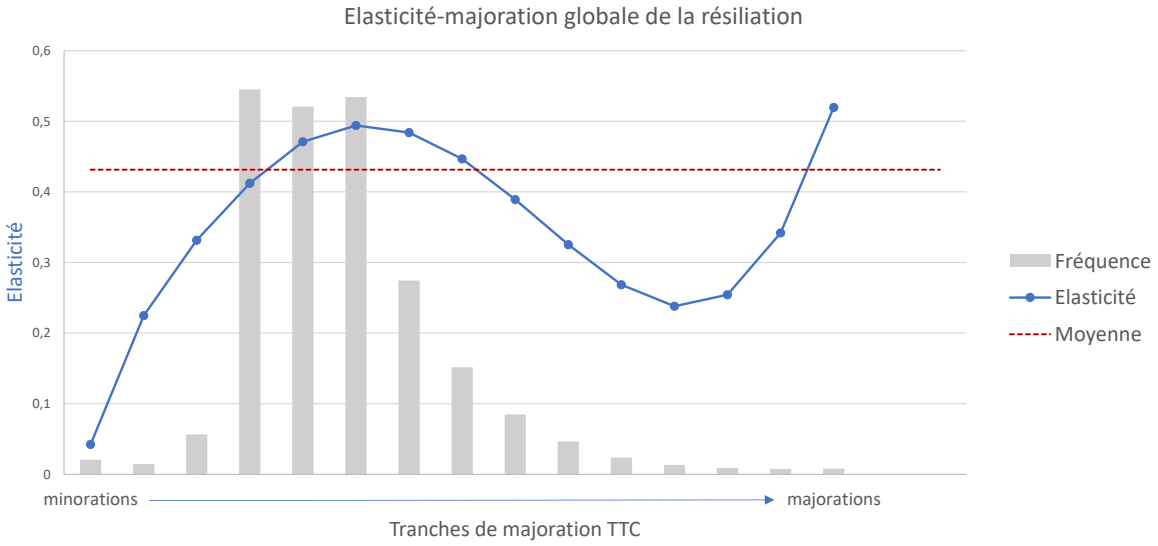


FIGURE 2.17 – Elasticité-majoration de la résiliation modélisée

Selon notre modèle de résiliation, les clients seraient peu sensibles à la variation de majoration quand cette dernière est faible voire négative. Cette sensibilité à la variation augmente lorsque la majoration augmente et atteint son apogée vers des majorations moyennes avant de diminuer légèrement. Nous remarquons également que les clients sont très sensibles à la variation de majoration lorsque cette dernière est déjà extrême.

Cette allure de l'élasticité-majoration n'est pas intuitive mais l'objectif est de pouvoir comparer l'élasticité de notre modèle avec l'élasticité réellement observée au terme afin d'établir la qualité de notre modélisation.

2.4.3 Comparaison avec l'élasticité réelle de la résiliation

Pour vérifier la pertinence de notre modèle, il faudrait pouvoir comparer l'élasticité-majoration du modèle avec l'élasticité-majoration réelle. Pour ce faire, il serait nécessaire de mettre en place en production un "Price test" qui choquerait les majorations prévues de la même manière qu'au paragraphe précédent, indépendamment

du profil du client, puis observer l'élasticité moyenne sur chaque tranche de majoration, avant de la comparer avec l'élasticité de notre modèle pour vérifier qu'elles sont semblables. Or, actuellement, du fait des contraintes informatiques et de communication aux agents sur la branche habitation, il n'est pas possible en pratique pour les équipes actuarielles de réaliser un vrai « price test » avec choc aléatoire. A ce stade, seuls des proxy sont possibles. Lorsque cette contrainte sera levée et lorsque les données seront disponibles, il sera important de refaire cette analyse.

2.5 Modélisation de la résiliation par Gradient Boosting Machine

Il est aussi possible de modéliser la probabilité de résiliation grâce au *Gradient Boosting Machine* (GBM). Comme nous l'avons énoncé plus tôt dans l'analyse des statistiques descriptives, le GBM est une technique ensembliste, ce qui rend le modèle moins gourmand en mémoire (en comparaison avec les forêts aléatoires par exemple) et ce qui conduit à des prédictions souvent plus rapides. Le GBM est une méthode puissante d'apprentissage supervisé, capable de fonctionner sur une base avec des données qualitatives et quantitatives, et qui a montré son efficacité en termes de prédiction dans de nombreuses études et défis. Toutefois, elle nécessite un ajustement précis des paramètres énoncés plus haut afin d'éviter le sur-apprentissage, et il est difficile de visualiser l'itinéraire d'optimisation du GBM. Enfin, cet algorithme est réputé pour la qualité de ses prédictions.

Nous rappelons en détail la méthode du *Gradient Boosting* en annexe 3. En pratique, nous allons réaliser le GBM sur Python, grâce au package "*xgboost*".

2.5.1 Préparation de la base de données

Pour le GBM, la base de données utilisée est la même que celle utilisée pour le GLM.

2.5.2 Gradient Boosting pour la classification binaire

Nous souhaitons prédire la probabilité de résilier d'un individu pour une cause tarifaire au terme à partir de la base de données dont nous disposons : $\{\boldsymbol{X}_i, Y_i\}_{i=1,\dots,n}$, avec X_i représentant le vecteur des variables explicatives de l'observation i , et Y_i la variable réponse associée qui, dans le cadre du GBM, prend la valeur 1 s'il y a eu une résiliation tarifaire au terme, -1 sinon. Selon l'article de Friedman, *Greedy Function Approximation : A Gradient Boosting Machine* [5], la fonction de perte pour un gradient boosting appliqué à une classification binaire est donnée par la log-vraisemblance de la loi binomiale négative :

$$L(Y,F) = \log\left(1+e^{-2YF}\right), \quad \text{avec } F(X) = \frac{1}{2}\log\left[\frac{\mathbb{P}(Y=1|X)}{\mathbb{P}(Y=-1|X)}\right], \quad Y \in \{-1;1\} \tag{2.23}$$

avec F une fonction associant les variables explicatives X_i à la variable réponse Y_i.

Nous définissons la pseudo-réponse $\widetilde{Y}_i$ comme le gradient négatif en X_i :

$$\widetilde{Y}_i = -\left[\frac{\partial L(Y_i, F(X_i))}{\partial F(X_i)}\right]_{F(X)=F_{m-1}(X)} = \frac{2Y_i}{1+\exp(2Y_i F_{m-1}(X_i))} \tag{2.24}$$

Nous pouvons déterminer le facteur multiplicatif ρ_m pour la réduction d'erreurs :

$$\rho_m = \arg\ \min_{\rho} \sum_{i=1}^{n} L(Y_i, F_{m-1}(X_i) + \rho h(X_i, a_m)) \tag{2.25}$$

$$= \arg\ \min_{\rho} \sum_{i=1}^{n} \log\left(1+\exp(-2Y(F_{m-1}(X_i) + \rho h(X_i, a_m)))\right) \tag{2.26}$$

avec le vecteur a_m représentant les paramètres du m-ième arbre de décision CART modélisant la pseudo-réponse $\widetilde{Y}_i$ par $h(X_i, a_m)$, et avec pour initialisation :

$$F_0(X) = \frac{1}{2} \log \frac{1+\bar{Y}}{1-\bar{Y}} \tag{2.27}$$

Au bout de M itérations, nous obtenons une approximation F_M de F^*, c'est-à-dire la fonction qui explique le mieux Y_i à partir des variables X_i. Nous pouvons à partir de là déterminer la probabilité conditionnelle qu'un individu résilie son contrat pour cause tarifaire au terme :

$$p_+(X_i) = \widehat{\mathbb{P}}(Y_i = 1|X_i) = \frac{1}{1+e^{-2F_M(X_i)}} \tag{2.28}$$

2.5.3 Choix des meilleurs paramètres par *Grid Search* (Recherche sur grille)

Il n'est pas évident de déterminer les paramètres du GBM qui maximisent la qualité de la prédiction. Alors, pour choisir les meilleurs paramètres, nous réalisons un *Grid Search* (cf. explications en annexe 3) sur plusieurs paramètres tels que la profondeur maximale des arbres, la réduction minimale de l'impureté pour générer un nouveau noeud ou nouvelle feuille dans un arbre, ou encore la fraction de variables à considérer pour chaque itération d'arbre. Une cross validation sur un *5-folds* est réalisée sur chaque combinaison A_i de paramètres. Par exemple, A_1 présente la combinaison de paramètres [21] suivante : une présélection aléatoire d'un échantillon des variables à hauteur de 40% pour la construction des arbres de décision ; une réduction minimale de 30% de l'impureté pour qu'une nouvelle feuille de l'arbre soit créée ; et une profondeur maximale des arbres égale à 3 (c'est-à-dire au maximum 3 créations de noeuds). Les autres A_i se verront attribuer d'autres valeurs pour chacun de ces paramètres.

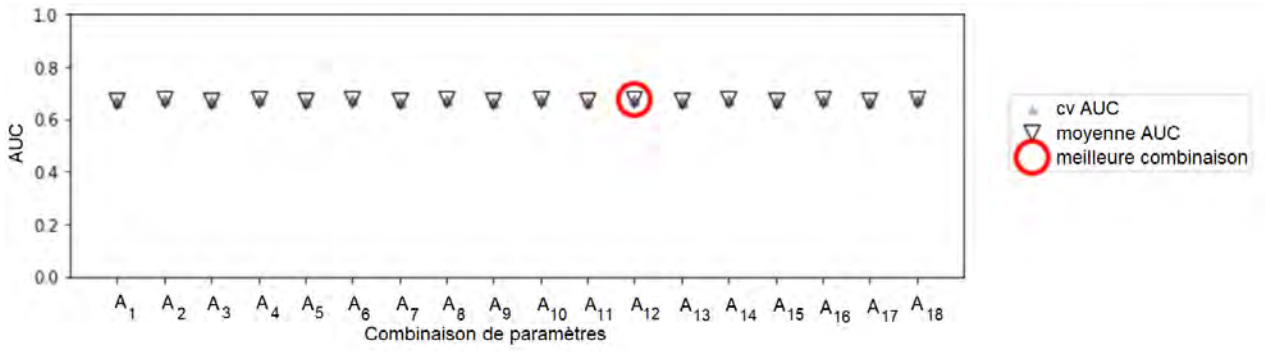


FIGURE 2.18 – Résultat du Grid Search pour le GBM

Parmi toutes les combinaisons A_i ainsi testées, nous choisissons celle qui maximise l'AUC, c'est-à-dire l'aire sous la courbe ROC (cf. annexe 5). Les résultats sont représentés en figure 2.18. Sur cette illustration, les petits triangles pleins indiquent la valeur de l'AUC pour un fold de la cross validation : puisqu'il s'agit d'un 5-folds, il y a donc 5 petits triangles par combinaison. Les triangles à l'envers indique pour chaque combinaison la valeur moyenne de l'AUC sur les 5-folds. Le cercle indique alors quelle est la moyenne la plus élevée, ce qui nous permet de repérer la combinaison des meilleures valeurs de paramètres correspondante.

21. Pour plus de détails sur les paramètres, et sur la définition de l'impureté, cf. annexe 3

2.5.4 Résultats du GBM

Avec cette combinaison optimale, nous obtenons la courbe ROC illustrée en figure 2.19, qui indique une valeur de Gini de 0,38 sur la base d'entraînement, et de 0,34 sur la base de test : ce sont des valeurs satisfaisantes compte tenu de la faible exposition des résiliations tarifaires dues au terme sur la base de données. Les courbes ROC sont alors quasiment confondues, signe d'une bonne stabilité du modèle.

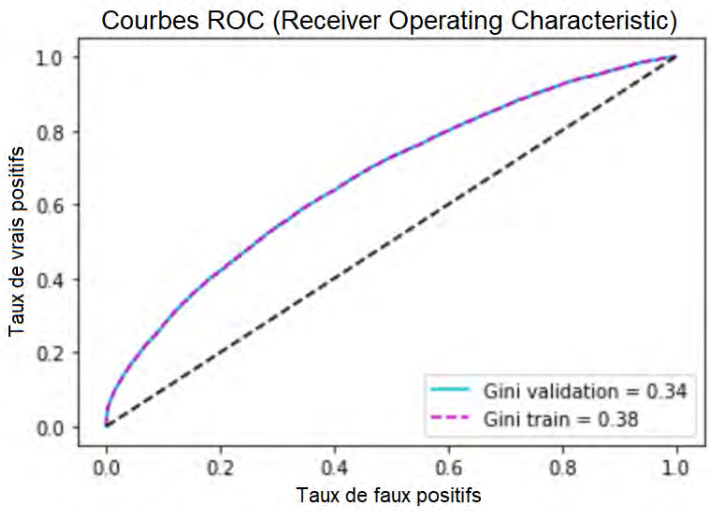


FIGURE 2.19 – Courbe ROC associée au GBM entraîné avec les meilleurs paramètres

Il est également possible, à partir de ce GBM, de repérer les variables les plus significatives : nous obtenons donc la figure 2.4, que nous avions justement vue lors de notre analyse exploratoire des données. Rappelons que les variables affichées sont suffisamment décorrélées, car l'algorithme du GBM répère les corrélations et utilise des variables suffisamment décorrélées pour converger le plus rapidement vers la solution optimale. Cela implique que les variables affichées ne sont pas redondantes et contribuent chacune largement à l'explication de la probabilité de résiliation.

2.6 Comparaison des modèles GLM et GBM

Nous disposons de deux modèles de résiliation au terme dont la qualité de prédiction est satisfaisante pour chacun. Alors lequel choisir pour réaliser la suite de notre étude ? Le GLM a l'avantage de fournir des résultats interprétables : ils peuvent être exprimés comme une fonction des variables explicatives. Le GBM quant à lui réalise un apprentissage rapide et présente des performances de prédiction supérieures à celles du GLM. C'est d'ailleurs ce que qu'indique la figure 2.20 : elle représente la valeur du Gini sur la base d'entraînement et la base de test (validation) pour chaque modèle.

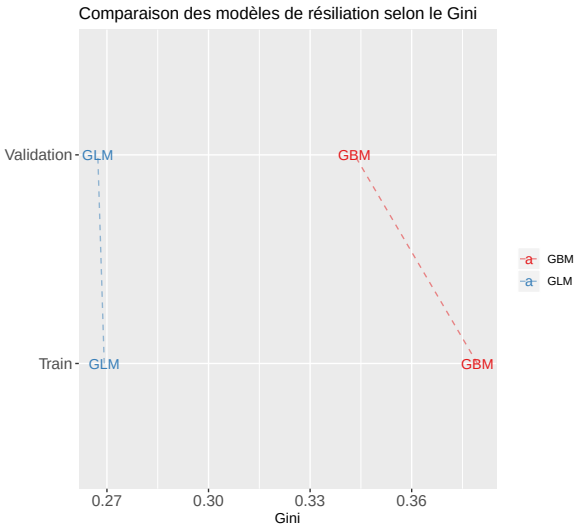


FIGURE 2.20 – Comparaison des performances de prédiction au sens du Gini pour le GLM et le GBM

Nous constatons que, bien que le GBM soit plus performant que le GLM au sens du Gini sur la base d'entraînement et sur la base de validation, il est toutefois moins stable : la performance sur la validation est nettement dégradée par rapport à l'entraînement. Cela provient du fait que le GBM, du fait de sa complexité, réalise un léger sur-apprentissage, ce qui implique qu'il généralise moins bien les prédictions sur de nouvelles données par rapport au GLM.

Ainsi, nous préférons utiliser un modèle stable, c'est-à-dire suffisamment généralisable, et interprétable. C'est pourquoi nous utiliserons le modèle GLM dans la suite de l'étude.

Conclusion du chapitre

Les statistiques descriptives montrent que le niveau de majoration a un impact sur le taux de résiliation tarifaire au terme : plus la majoration est importante, plus la probabilité qu'un client résilie son contrat augmente. L'analyse descriptive révèle également que, suivant leur profil et les caractéristiques de leur contrat, les clients présentent des comportements différents. Cela indique que notre étude a bien un intérêt stratégique : pour retenir au mieux les clients, la marge de manœuvre de l'assureur réside, du moins en partie, dans l'établissement d'une majoration adéquate en fonction du client et de son contrat.

La probabilité de résiliation au terme pour chaque contrat en portefeuille a ensuite pu être modélisée, de telle sorte que la majoration TTC tienne une place importante parmi toutes les variables explicatives, car c'est bien cette variable qui devra être déterminée par l'algorithme d'optimisation. Par ailleurs, la modélisation GLM est préférable à une modélisation GBM, d'une part puisque la prédiction est plus stable de cette manière, et d'autre part puisqu'un tel modèle est totalement transparent et interprétable.

La modélisation ainsi obtenue présente une qualité de prédiction satisfaisante, toutefois elle est à utiliser avec précaution. En effet, l'absence de price-test ne permet pas pour l'instant de valider l'élasticité-majoration du modèle, et donc de savoir si ce dernier capte correctement la sensibilité réelle du client à la majoration. Ainsi, pour obtenir une prédiction plus juste, cette étape devra être réalisée dès que les résultats du price-test au terme seront disponibles. Mais pour l'instant, en première approximation, nous nous contenterons de ce premier modèle pour élaborer l'algorithme d'optimisation.

Chapitre 3

Modélisation du crédit commercial

Sommaire

Préambule

Nous avons vu, dans le cadre du chapitre précédent, comment estimer le taux de résiliation selon les caractéristiques du client, de son contrat et la majoration qui lui serait appliquée. Or, dans le cadre de notre problème d'optimisation, nous devons également estimer l'ELR du portefeuille après l'application de la majoration au terme, *ie* le rapport entre les primes pures des différents contrats et les cotisations hors taxe.

Considérer simplement les cotisations comme le produit de la prime commerciale hors taxe avec la majoration est insuffisant, car en réalité, lorsque le client prend connaissance de sa majoration, il va chercher à prendre contact avec son agent général ou son courtier afin d'obtenir une justification de sa majoration et éventuellement demander un geste commercial sous peine de résilier son contrat et souscrire une affaire nouvelle chez un concurrent sans doute moins chère à iso-risque. En effet, les assureurs en France ont des stratégies de conquête du marché IARD similaires : présenter des tarifs en affaires nouvelles très compétitifs dans un contexte ultra concurrentiel afin d'engendrer de nouvelles souscriptions, puis appliquer chaque année une majoration sur ces contrats pour les rendre enfin rentables.

Ainsi, si l'agent général (ou le courtier) souhaite garder le client majoré dans son portefeuille, deux choix s'offrent à lui :

— Soit il propose un rabais en euros sur la nouvelle prime commerciale : c'est le crédit commercial ;

— Soit il propose un "remplacement" de contrat, c'est-à-dire qu'il va modifier voire retirer certaines garanties du contrat afin d'en diminuer la prime commerciale.

Dans tous les cas, remplacement et crédit commercial viennent diminuer la ressource prévue initialement par l'assureur dans le cadre de la majoration, et donc augmenter l'ELR, même si ces leviers commerciaux permettent de limiter la hausse du taux de résiliation.

3.1 Statistiques descriptives

Pour vérifier nos hypothèses sur l'impact de la majoration au terme sur les usages des remplacements par les agents généraux et les courtiers, nous traçons la répartition du nombre de contrats remplacés par rapport à la date de terme 0. La figure 3.1 indique effectivement une forte concentration de remplacements aux alentours de la date du terme. Nous pouvons donc à juste titre penser que la plupart de ces remplacements ont pour objet la diminution de la prime commerciale.

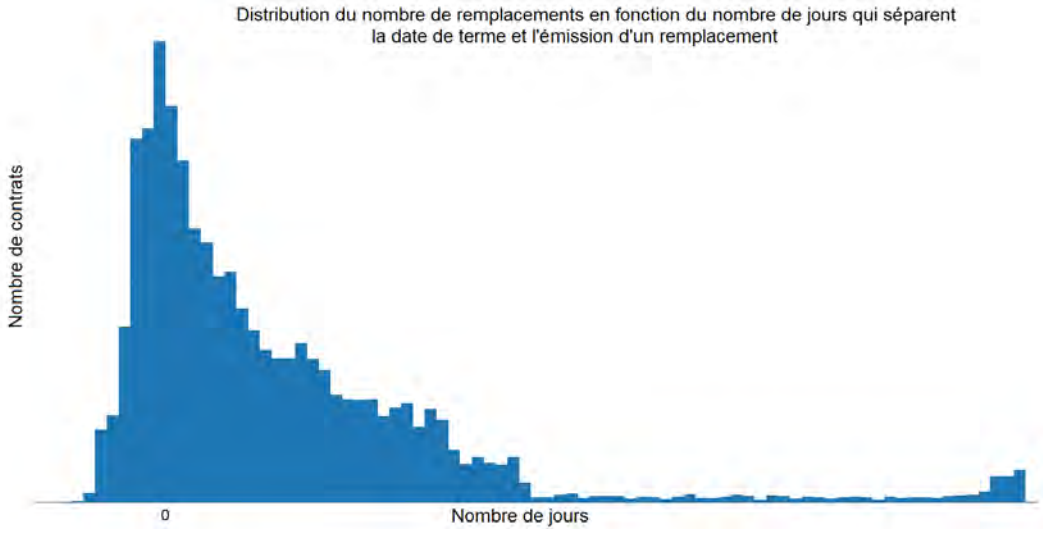


FIGURE 3.1 – Répartition des remplacements par rapport à la date de terme

Toutefois, la prise en compte des remplacements dans notre estimation de l'ELR est compliquée opérationnellement parlant : en effet, les motivations exactes de ces remplacements sont inconnues et leur impact tarifaire est difficile à prédire. De plus, tous les remplacements ne diminuent pas la prime : certains clients augmentent leur niveau de garantie volontairement ce qui engendre un accroissement de la prime. C'est le cas par exemple de clients souhaitant assurer plus de capital, ou bien encore des clients voulant assurer une piscine nouvellement construite sur leur terrain. Pour ces raisons, nous décidons de ne pas modéliser la probabilité de remplacement.

En revanche, si nous retirons de notre base de données tous les contrats qui sont remplacés, nous sommes capables d'observer l'usage des crédits commerciaux par les agents généraux, en commençant par déterminer quel type d'agent utilise le plus fréquemment ce levier commercial. La figure 3.2 indique effectivement que ce sont les agents du groupe 3 qui ont le plus recours au rabais, que ce soit lors d'une affaire nouvelle ou au terme d'un contrat. Nous remarquons que les deuxièmes plus gros utilisateurs de crédits commerciaux sont les agents du groupe 4, et cela a un impact sur leur capacité à retenir leurs clients (rappelons que ce segment d'agents présente le plus fort taux de résiliations tarifaires dues au terme). Enfin, ce sont les agents du groupe 1 qui utilisent en proportion moins de rabais : leur bonne gestion de portefeuille leur permet de ne pas avoir régulièrement recours à ce geste commercial.

Il est également intéressant de voir sur la figure 3.3 la distribution de pourcentage de crédit commercial par rapport à la prime commerciale. Quelque soit le segment, la distribution est quasiment la même. Nous analyserons d'ailleurs cette distribution en détail dans la partie 3.3.4 afin de voir si elle se rapproche d'une loi connue et déterminer si nous pouvons prédire ce pourcentage de rabais avec un GLM.

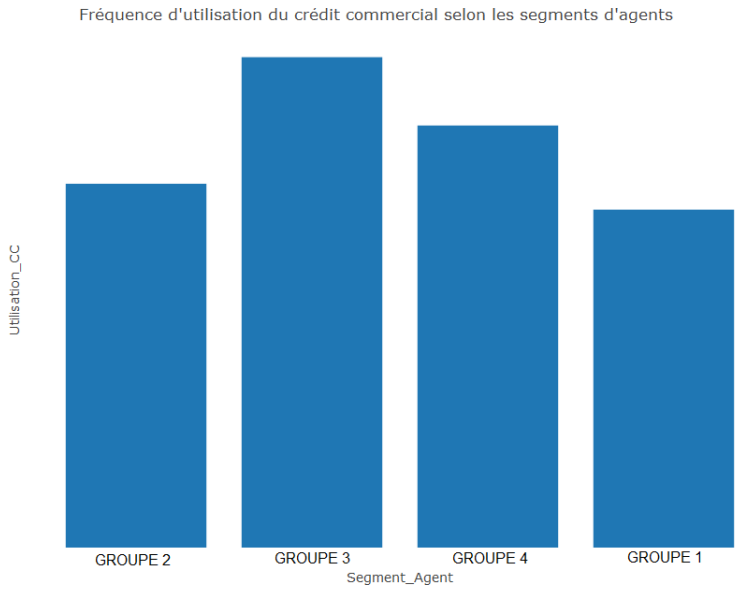


FIGURE 3.2 – Fréquence d'usage du crédit commercial selon les agents

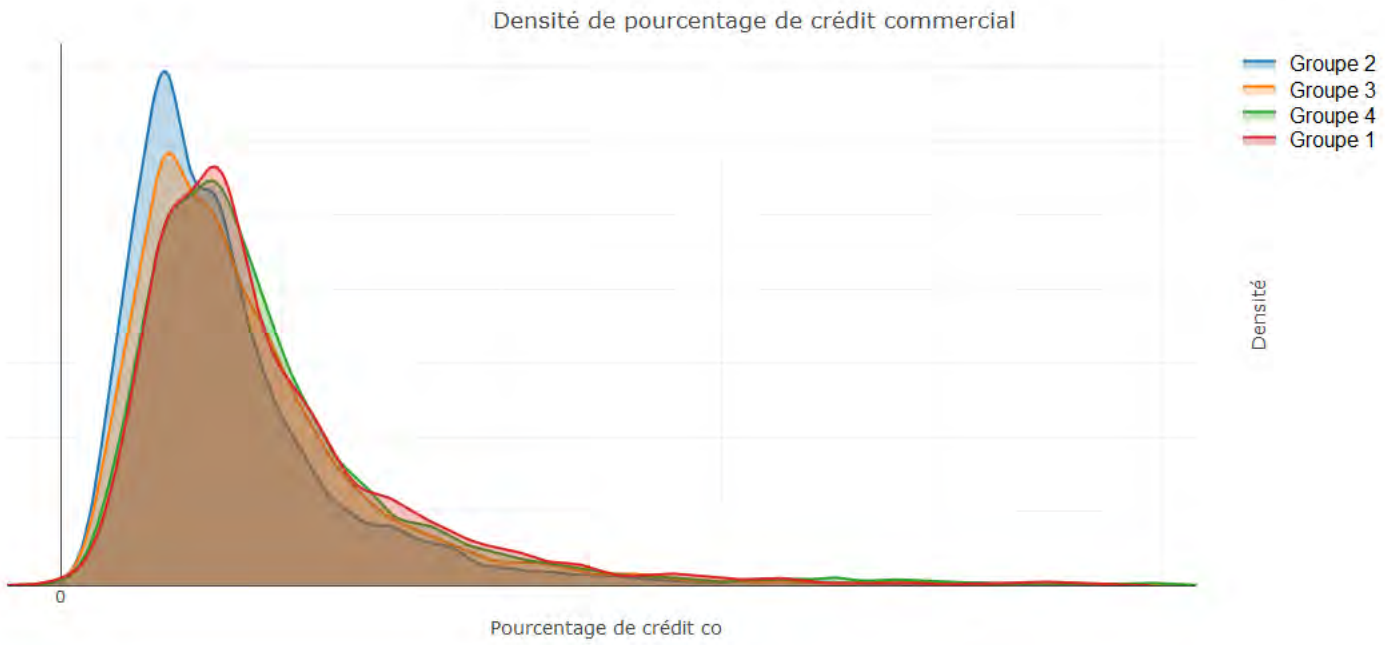


FIGURE 3.3 – Pourcentage de crédit commercial selon les agents

3.2 Modélisation de la probabilité d'attribution de crédit commercial

L'objectif est de repérer l'occurrence d'un contrat qui n'a pas été remplacé et qui a obtenu un rabais commercial après l'annonce de la majoration de son contrat. Alors, pour modéliser la probabilité d'attribution du crédit commercial, nous souhaitons réaliser un modèle logistique par GLM sur le logiciel Akur8. En effet, forts de notre constat au chapitre 2, le GLM représente des performances satisfaisantes tout en étant suffisamment généralisable et interprétable.

3.2.1 Base de données utilisée

Nous utilisons quasiment la même base de données utilisée lors de la modélisation de la résiliation tarifaire au terme (puisqu'elle donne une vision très complète des informations de chaque contrat et que les corrélations trop importantes ont été retirées) à quelques différences près :

— Nous rajoutons des variables explicatives *a posteriori* de la majoration, comme le montant de crédit commercial éventuellement appliqué, ainsi que l'indicatrice d'un remplacement après l'annonce de la majoration ;

— La variable réponse Y devient l'indicatrice de l'application du crédit commercial sur un contrat non remplacé : elle prend la valeur 1 si le contrat concerné n'a pas été remplacé et s'il a reçu un rabais, 0 sinon.

3.2.2 Modélisation GLM de la probabilité d'application de crédit commercial sur Akur8

3.2.2.1 Choix du modèle le plus pertinent

Tout comme au chapitre précédent, nous choisissons sur la base du Grid Search le modèle qui réalise le meilleur compromis entre qualité de prédiction (au sens du Gini) et un nombre raisonnable de variables. Par ailleurs, nous n'avons pas d'exigence particulière sur les variables les plus significatives tant qu'elles nous paraissent pertinentes

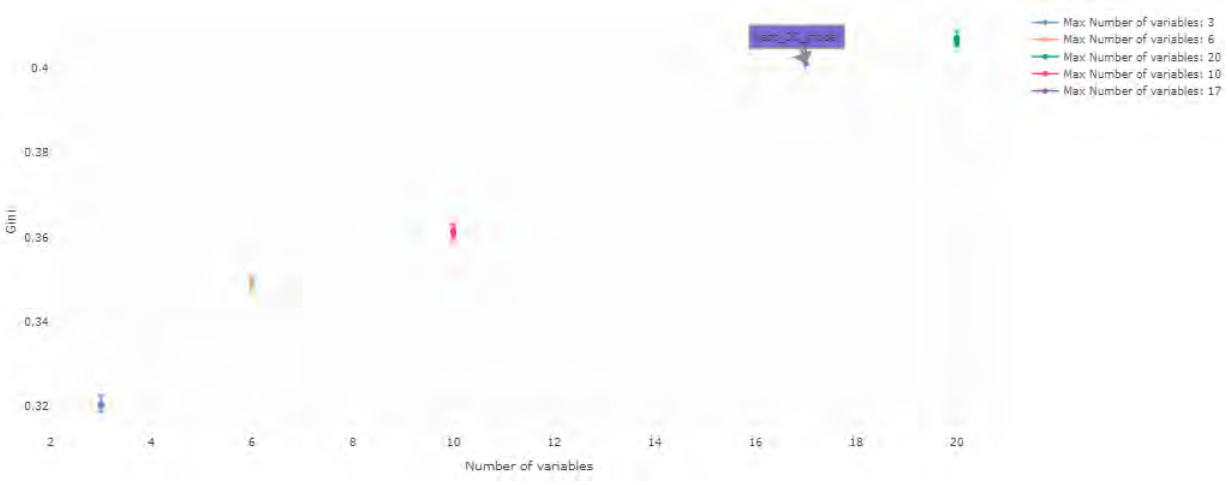


FIGURE 3.4 – Résultat du Grid Search pour le GLM modélisation la probabilité d'application de crédit commercial

Nous choisissons le modèle à 17 variables présentant le meilleur Gini, indiqué par la flèche "best_CC_model" sur la figure 3.4. Nous considérons en effet que rajouter 3 variables à ce modèle pour gagner à peine 0,01 point de Gini est superflu.

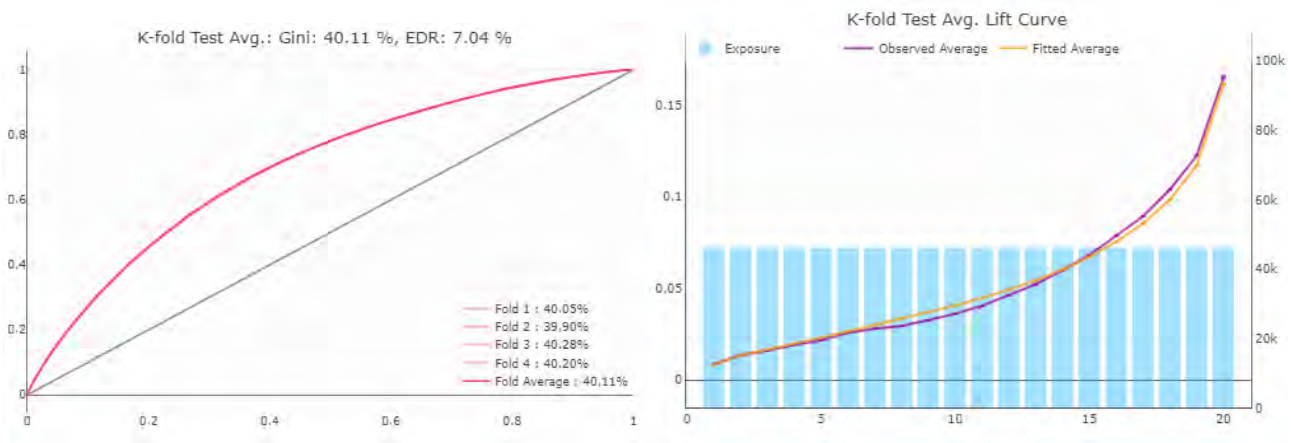
Dans ce modèle, c'est la variable du canal de distribution qui est la plus significative, suivie du logarithme du rapport entre le prix du contrat après l'application de la majoration (que nous appelons "prix post-terme") sur l'estimation du prix proposé par la concurrence [1] pour une affaire nouvelle sur le même risque (CMA [2]) : $\log\left(\frac{\text{prix AXA post terme}}{CMA}\right)$. Ces deux variables font sens puisque d'une part, la proximité d'un client avec son agent général ou son courtier favorise la réalisation d'un geste commercial par rapport à un client ayant souscrit sur internet qui ne saura pas forcément vers qui s'orienter pour obtenir un rabais sur sa prime ; d'autre part, la

1. Filiale d'AXA France
2. Competitive Market Analysis. Cf. chapitre 2 pour en savoir plus sur la manière dont le CMA a été estimé

différence entre le tarif AXA et celui de la concurrence, moins cher en affaire nouvelle, est un argument largement employé par les clients pour exiger une diminution de leur prime majorée.

3.2.2.2 Analyse de la qualité du modèle

Sur le modèle sélectionné, la moyenne des indices de Gini sur le 4-fold est de 41,11%, ce qui est une valeur satisfaisante. Les courbes ROC associées à chacun des folds représentées en figure 3.5 (a) sont bien superposées, ce qui indique une bonne stabilité du modèle. Par ailleurs, la lift-curve en figure 3.5 (b) signale une bonne qualité de la prédiction de manière générale. Bien que nous sur-estimons ou sous-estimons légèrement par endroit la probabilité d'application du crédit commercial, la tendance de la lift-curve observée est bien représentée par celle prédite.



(a) Courbes ROC du modèle de probabilité de Crédit Co.

(b) Lift curve du modèle de probabilité de Crédit Co.

FIGURE 3.5 – Graphiques indicateurs de la qualité de la prédiction de la probabilité de Crédit Co.

Nous nous assurons également, pour chaque variable dont les catégories sont ordonnées, que la tendance des coefficients a la même allure que la probabilité d'application de crédit commercial. Par exemple, nous constatons sur la figure 3.6 que c'est bien le cas pour la variable du logarithme du rapport prix AXA post-terme sur l'estimation du tarif proposé par la concurrence (CMA).

Sur ce graphique, nous constatons que, lorsque la prime AXA post-terme est supérieure à l'estimation du CMA (cas du logarithme positif), plus l'écart entre la prime AXA et l'estimation du tarif de la concurrence se creuse, moins le contrat a de chance d'obtenir un rabais. Nous pouvons interpréter la diminution de la probabilité d'application de crédit commercial avec cet écart par le fait que, si l'écart est grand, cela signifie que l'agent devra faire un gros effort commercial pour s'aligner aux prix de la concurrence, engendrant une consommation importante de son budget de crédit commercial.

En outre, nous remarquons sur cette même figure des difficultés de prédictions sur les queues de distribution de cette variable. Cela provient directement de la faible exposition de ces catégories extrêmes, qui ne permet pas un apprentissage du modèle suffisant. Enfin, il est important de remarquer que pour les logarithmes négatifs, c'est-à-dire les cas où la prime AXA est plus faible que les tarifs concurrents, la probabilité d'application de crédit commercial pour ces contrats est maximale. Cela semble être un non sens : pourquoi appliquer un rabais si le prix proposé par AXA est déjà le moins cher sur le marché ? La raison est intrinsèque à la modélisation : Akur8 va effectuer des regroupements, garder une tendance des coefficients cohérente avec les observations, et ne va pas accorder d'importance aux catégories présentant une faible exposition par rapport aux autres. Dans la pratique, cela ne nous pose pas d'inconvénient puisque ces cas sont rares et que la qualité globale de la prédiction est bonne, comme l'ont indiqué la courbe ROC et la lift-curve en figure 3.5.

Enfin nous pouvons également nous assurer de la stabilité de notre modèle en observant l'évolution de la

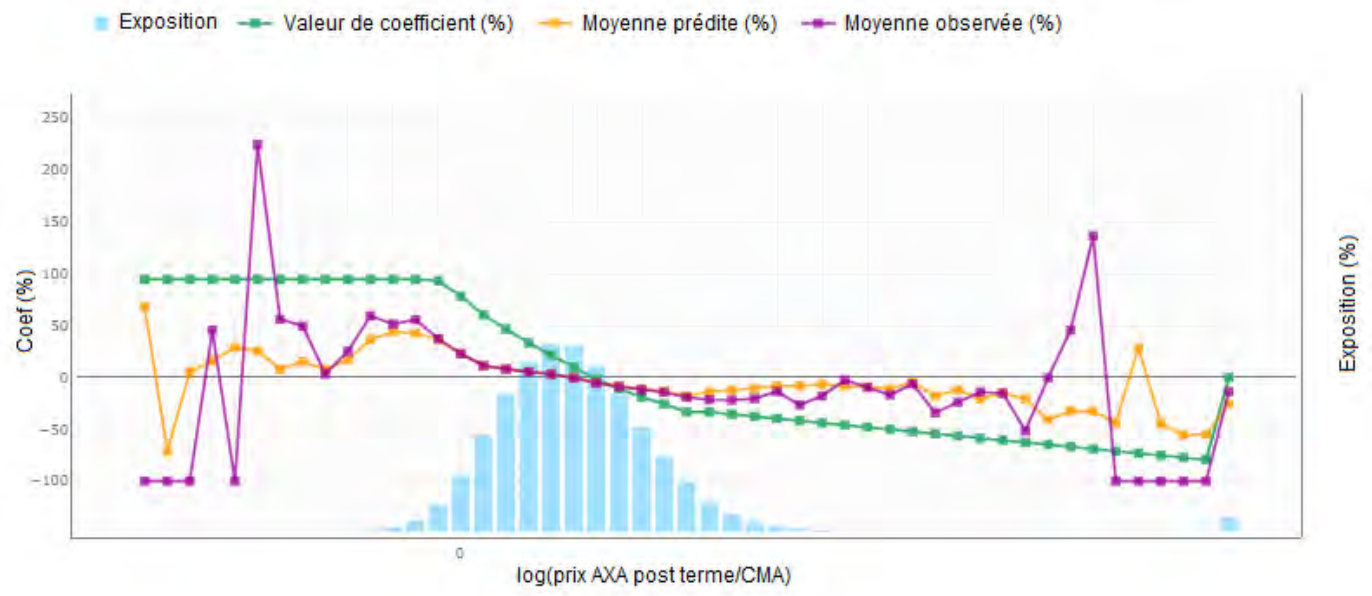


FIGURE 3.6 – Tendance des coefficients associés au logarithme du rapport de la prime AXA sur le CMA après modélisation GLM

valeur des coefficients dans le temps en figure 3.7 : relativement aux années 2017 ou 2018, les coefficients se confondent, sauf justement sur la queue gauche de la distribution, lorsque le logarithme est négatif. Le nombre réduit d'observations sur ces catégories empêche une généralisation suffisante de l'apprentissage du modèle, rendant instable la prédiction sur ces catégories.

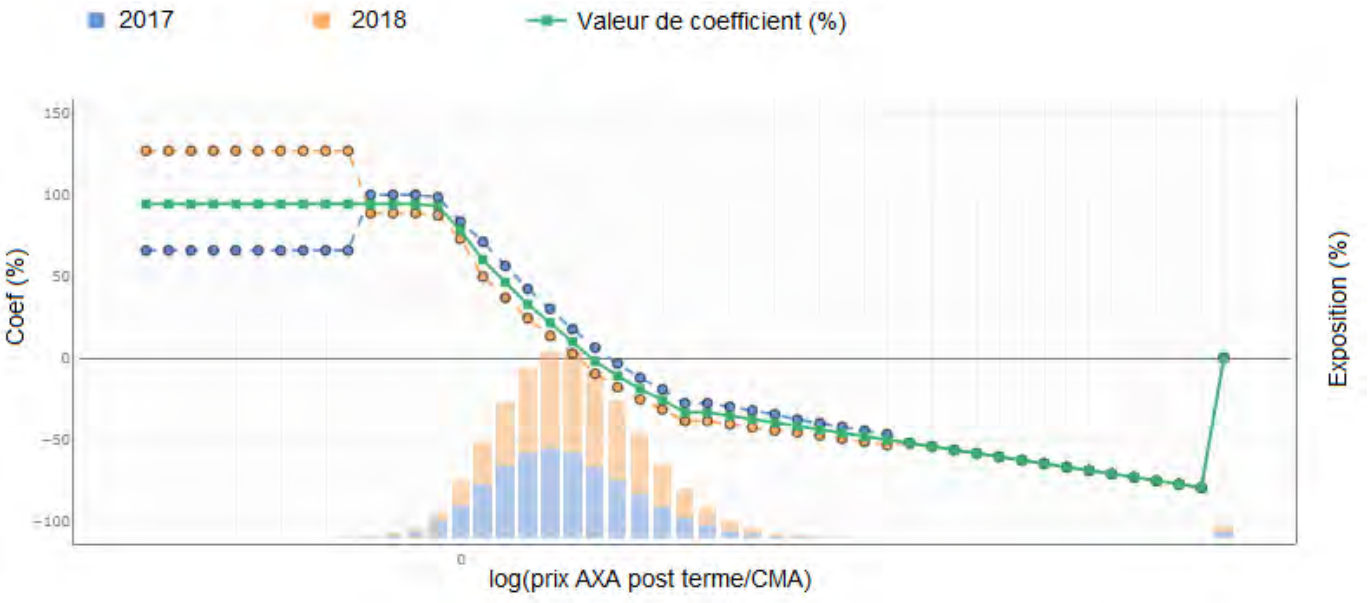


FIGURE 3.7 – Stabilité dans le temps des coefficients associés au logarithme du rapport de la prime AXA sur le CMA après modélisation GLM

3.2.2.3 Conclusion sur la modélisation de la probabilité de crédit commercial

Nous disposons à présent d'un modèle de probabilité d'application de crédit commercial satisfaisant en termes de précision, et interprétable, ce qui nous permettra par la suite de prédire cette probabilité pour un contrat selon ses caractéristiques.

Or, plus que la probabilité, nous souhaitons connaître l'espérance de crédit commercial qui sera attribuée à un contrat afin d'obtenir une meilleure estimation de ce qu'il générera en termes de chiffre d'affaires.

3.3 Modélisation du pourcentage de crédit commercial attribué par rapport à la prime commerciale

L'objectif de cette partie est de modéliser le pourcentage de crédit commercial appliqué sur la prime post-majoration d'un contrat.

3.3.1 Base de données utilisée

Nous partons de la base de données que nous avions utilisée pour la modélisation de la probabilité d'application du crédit commercial, et nous gardons uniquement les observations présentant un montant de crédit commercial strictement positif.

La variable réponse Y que nous considérons à présent est le pourcentage de crédit commercial appliqué par rapport à la prime commerciale après application de la majoration :

$$Y = \frac{\text{Montant de crédit commercial en euros}}{PC_{n-1,HT} * \text{majoration}_{HT}} \tag{3.1}$$

Il s'agit donc d'une variable continue à valeurs dans $\mathbf{R}_+$. Il nous faut donc connaître la distribution de Y pour savoir s'il nous est possible de réaliser un modèle GLM, c'est-à-dire s'assurer que cette variable réponse suit une loi appartenant à une famille exponentielle.

3.3.2 Analyse de la distribution du pourcentage de crédit commercial

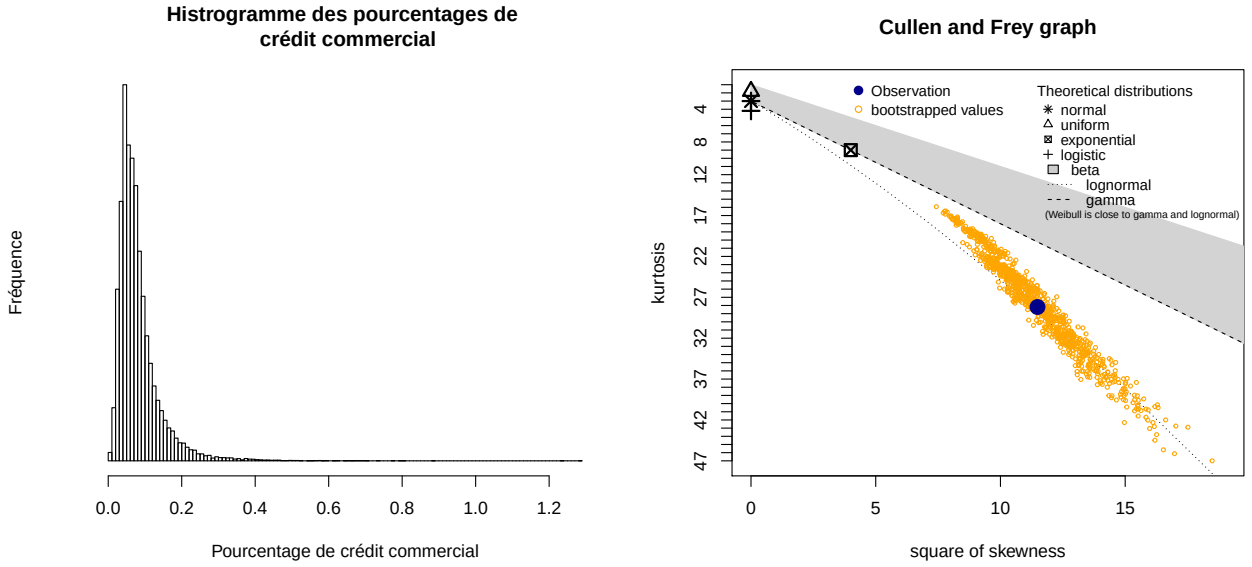
Sur la base de données considérée, nous obtenons la distribution empirique illustrée en figure 3.8 (a). Cette distribution présente un skewness positif et une valeur de kurtosis supérieure à 3, ce qui se rapproche de la distribution d'une loi log-normale. Nous allons donc vérifier si cette hypothèse sur la distribution est juste.

Une première manière de déterminer la loi de cette distribution est de repérer la distribution sur le graphique de Cullen et Frey, élaboré par Delignette-Muller et Dutang [4]. L'arrière plan de ce graphique en figure 3.8 (b) indique, pour chaque loi usuelle, l'évolution du kurtosis en fonction du carré du skewness. Par exemple, si nous étudions une variable suivant une loi uniforme, sa distribution empirique sera indiquée par un point qui se superposerait au petit triangle. Or l'estimation du skewness et du kurtosis sur une seule distribution n'est pas robuste. Pour pallier l'incertitude liée à ces estimations, il nous est alors possible de réaliser des "bootstraps" sur cette distribution, c'est-à-dire un échantillonage aléatoire avec remise à partir de la distribution initiale, et estimer à nouveau le skewness et le kurtosis sur chacun de ces échantillons. Les résultats sont indiqués par de petits cercles sur le graphique.

Concernant notre distribution du pourcentage de crédit commercial, la distribution empirique ainsi que les bootstraps associés suivent la ligne indiquant une loi log-normale.

La loi log-normale est donc une bonne candidate et il ne serait donc pas exclu de poser :

$$Y \sim \text{Log-}\mathcal{N}(\mu, \sigma^2) \tag{3.2}$$



(a) Distribution du pourcentage de crédit commercial

(b) Graphique de Cullen et Frey

FIGURE 3.8 – Détermination de la distribution du pourcentage de crédit commercial

avec

$$\mathbb{E}[Y] = e^{\mu+\sigma^2/2} \tag{3.3}$$

$$\mathbb{V}ar[Y] = \left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2} \tag{3.4}$$

Cela serait une bonne nouvelle car nous pouvons alors poser la variable aléatoire $Z = \log(Y)$ qui suit une loi normale de mêmes paramètres que la loi log-normale associée :

$$Z \sim \mathcal{N}(\mu, \sigma^2) \tag{3.5}$$

Toutefois il est souhaitable de réaliser un test de Shapiro-Wilk sur Z afin de s'assurer que sa distribution n'est effectivement pas différente d'une loi normale, ce qui permettra par la même occasion de valider ou non le fait que Y suit la loi log-normale associée.

Le test de Shapiro-Wilk permet de tester l'hypothèse nulle selon laquelle l'échantillon $z_1, ..., z_n$ est issu d'une variable aléatoire normalement distribuée. Ce test est basé sur la statistique W définie par (selon Rakotomalala [9]) :

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i \left(z_{(n-i+1)} - z_{(i)}\right)\right]^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2} \tag{3.6}$$

avec

- $z_{(i)}$ la série des z_i ordonnés ;
- $\left[\frac{n}{2}\right]$ la partie entière de $\frac{n}{2}$;
- a_i des constantes obtenues à partir de la moyenne et de la matrice des variances covariance des quantiles d'un échantillon de taille n.

En quelque sorte, la statistique W peut être interprétée comme un indicateur de la corrélation entre les quantiles théoriques de la loi normale et des quantiles empiriques obtenus à partir de la distribution observée.

En fonction de la taille de l'échantillon et du risque α (que nous choisissons égal à 5%), il existe une valeur seuil W_{crit} [3] en-dessous de laquelle l'hypothèse nulle est rejetée.

Lorsque nous effectuons le test de Shapiro-Wilk sur Z, nous obtenons une valeur de W égale à 0.98912, pour une p-valeur associée inférieure à $2.2.10^{-16}$. Nous devons donc rejeter l'hypothèse nulle : empiriquement Z ne suit pas une loi normale. Effectivement, lorsque nous affichons la distribution de Z sur le graphique de Cullen et Frey en figure 3.9, Z suit davantage la loi logistique : le kurtosis de sa distribution est estimée à 4,26, ce qui indique que les queues de distribution sont plus lourdes que celles d'une loi normale.

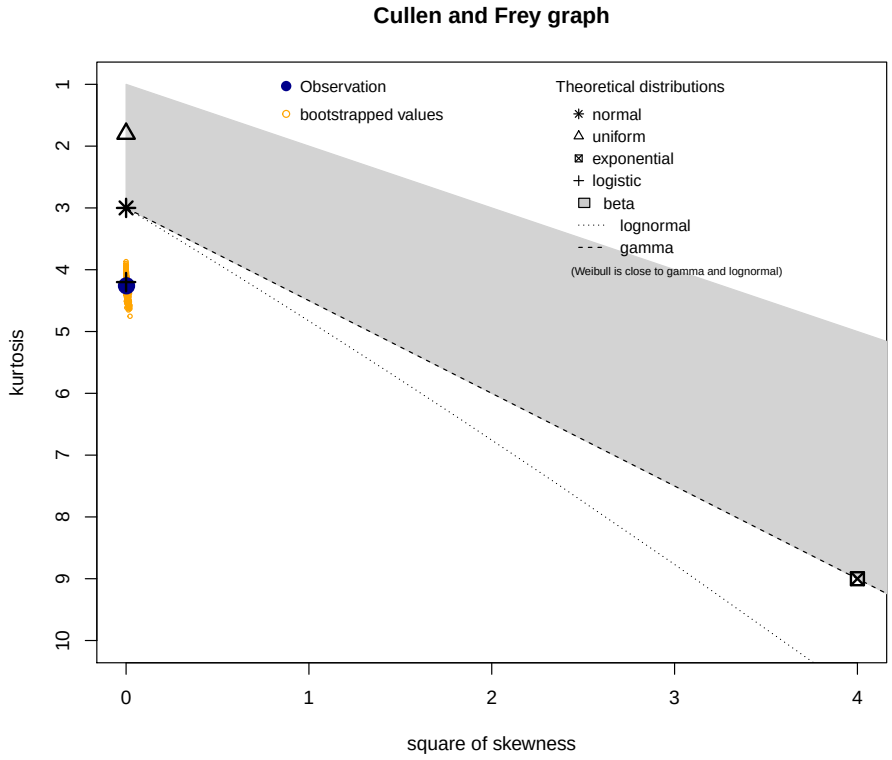


FIGURE 3.9 – Graphique de Cullen et Frey pour la distribution de Z

Or, la loi logistique n'appartient pas à une famille exponentielle, et sachant que le GLM ne peut modéliser que des lois appartenant à une telle famille, nous ne pourrons pas *a priori* effectuer une régression linéaire généralisée sur Z. Certes, un GBM peut s'affranchir de la forme de la distribution, or l'intérêt de réaliser un GLM est bien d'obtenir une fonction dépendant de plusieurs variables qui prédira facilement le pourcentage de crédit commercial à attribuer pour chaque contrat. Nous avons besoin de cette fonction pour pouvoir par la suite tester plusieurs scénarios dans le cadre de l'optimisation.

Si toutefois nous comparons la distribution de Z avec la loi normale sur la figure 3.10, nous remarquons que les tendances sur le *Q-Q plot*, le *P-P plot* et la fonction de répartition sont similaires. Bien que nous remarquions sur le *Q-Q plot* l'existence de queues lourdes pour notre distribution de Z, nous choisissons, pour des raisons opérationnelles dans la suite de l'étude, de faire une grande approximation en considérant malgré tout que Z suit une loi normale, et donc que Y suit une loi log-normale. Nous vérifierons plus tard, à l'issue du GLM, si les prédictions sont suffisamment satisfaisantes, ce qui nous permettra de justifier que l'approximation est bien raisonnable.

3. La valeur de W_{crit} se lie dans la table de Shapiro-Wilk.

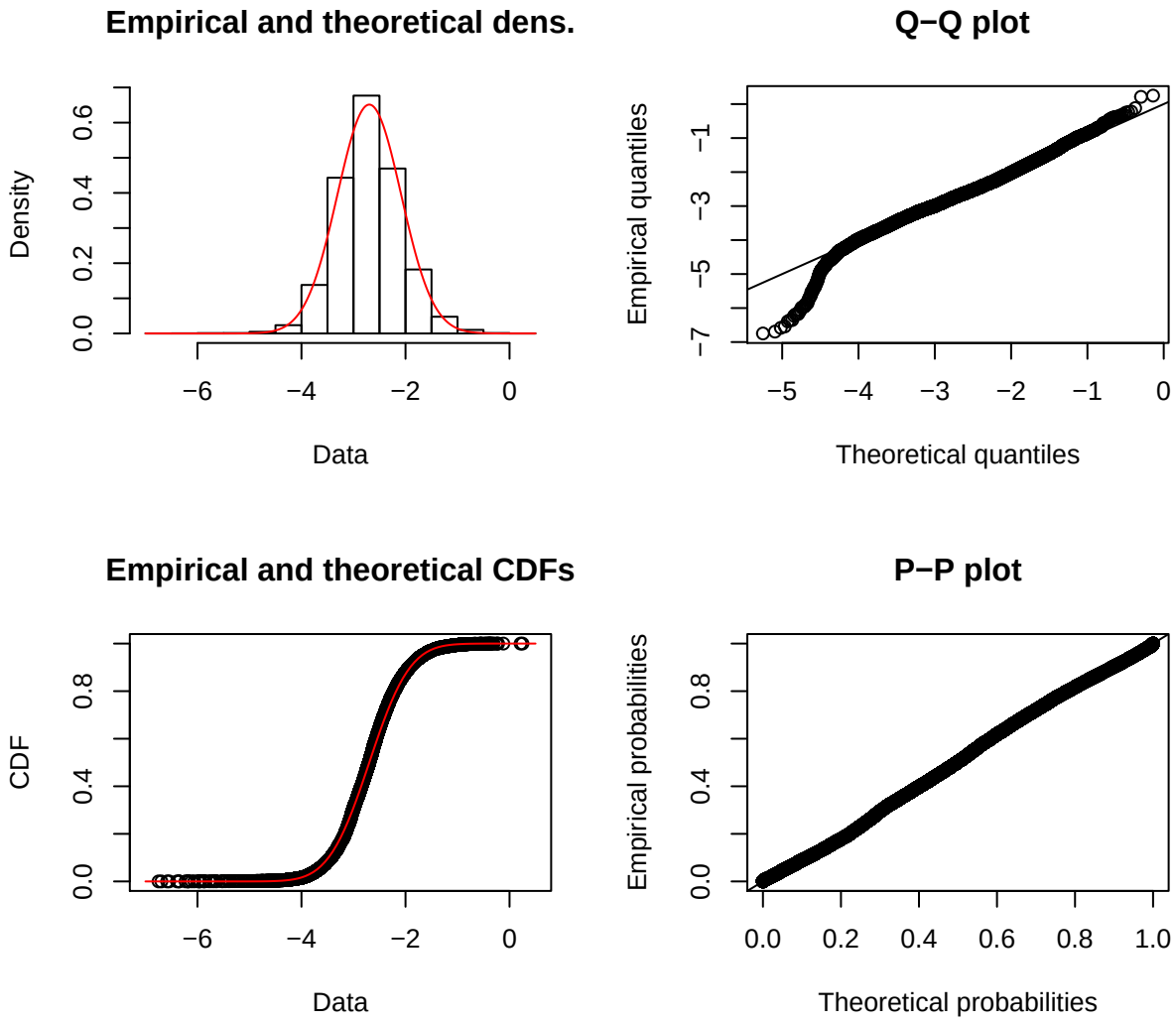


FIGURE 3.10 – Comparaison de la distribution empirique de Z avec la loi normale

3.3.3 Lien entre l'estimation de $\mathbb{E}[Z|X]$ et $\mathbb{E}[Y|X]$

Sous les hypothèses du paragraphe précédent, et puisque la loi normale appartient à une famille exponentielle, nous pouvons réaliser un GLM pour estimer l'espérance conditionnelle du logarithme du pourcentage de crédit commercial Z. Nous avons donc, en utilisant la fonction de lien canonique de la loi normale (c'est-à-dire l'identité), la relation suivante :

$$\mathbb{E}[Z|X] = \mathbb{E}[\log(Y)|X] = X\beta \tag{3.7}$$

de telle sorte que

$$Z = \mathbb{E}[Z|X] + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{3.8}$$

avec X la matrice de taille $n \times (p+1)$ dont la première colonne n'est composée que de 1 et dont chaque ligne représente une observation, et $\beta = (\beta_0, ..., \beta_p)$ le vecteur colonne correspondant aux $p+1$ paramètres du modèle.

Or, pour obtenir $\mathbb{E}[Y|X]$, nous ne pouvons pas directement prendre l'exponentielle de $\mathbb{E}[Z|X]$. Comme nous l'avons vu au paragraphe précédent, nous avons la relation :

$$\mathbb{E}[Y] = e^{\mu+\sigma^2/2} \tag{3.9}$$

or

$$\mu = \mathbb{E}[Z|X] \tag{3.10}$$
$$\sigma^2 = \mathbb{V}ar[Z] = \mathbb{V}ar[\varepsilon] \tag{3.11}$$

Nous pouvons donc en déduire :

$$\mathbb{E}[Y|X] = \exp\left[X\beta + \frac{\sigma^2}{2}\right] \tag{3.12}$$

Nous devons donc estimer de façon non biaisée σ^2 par s^2 à partir des résidus de la régression :

$$s^2 = \frac{\sum_{i=1}^{n}[\log(Y_i) - X_i\beta]^2}{n-p-1} \tag{3.13}$$

pour finalement obtenir un estimateur de $\mathbb{E}[Y|X]$:

$$\widehat{\mathbb{E}[Y|X]} = \exp\left[X\beta + \frac{s^2}{2}\right] \tag{3.14}$$

3.3.4 Modélisation GLM du pourcentage de crédit commercial

Nous voulons réaliser un GLM sur Z, c'est-à-dire le logarithme du pourcentage de crédit commercial. Puisque nous avons supposé que Z suivait une loi normale, cela revient à réaliser une régression linéaire. Etant donné qu'Akur8 ne nous permet pas encore de faire un tel modèle [4], nous effectuons la régression linéaire sur R.

De plus, nous souhaitons comparer deux modèles afin de choisir celui qui aura le meilleur pouvoir de prédiction. Ils se différencient par des variables explicatives différentes, choisies selon deux méthodes distinctes : la première par Elastic Net, la seconde par GBM.

3.3.4.1 GLM avec sélection des variables explicatives par Elastic Net

L'idée de la pénalisation est de sélectionner uniquement les variables qui contribuent le plus à l'explication de la variable réponse Z, c'est-à-dire le logarithme du pourcentage de crédit commercial. Cependant, utiliser uniquement la pénalisation Lasso ou Ridge (cf. annexe 4) ne garantie pas les meilleures prédictions. Parfois, une combinaison des deux peut aboutir à de meilleures performances : c'est l'Elastic Net.

Il s'agit en réalité d'un problème d'optimisation : quels sont les β_i qui minimisent l'erreur quadratique sous la combinaison de contraintes Lasso et Ridge ? La formulation de ce problème s'écrit :

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n}\left(Z_i - \sum_{j=0}^{p}\beta_j Z_{i,j}\right)^2 + \lambda_1\sum_{j=0}^{p}|\beta_j| + \lambda_2\sum_{j=0}^{p}\beta_j^2 \tag{3.15}$$

avec les multiplicateurs de Lagrange $\lambda_1 \leq 0$ et $\lambda_2 \leq 0$, et qui peut aussi se réécrire sous la forme :

4. En effet, le logiciel Akur8 est encore en cours de développement, et de ce fait ne propose pour l'instant que certains types de modélisations.

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(Z_i - \sum_{j=0}^{p} \beta_j Z_{i,j} \right)^2 + \lambda \left[\alpha \sum_{j=0}^{p} |\beta_j| + (1-\alpha) \sum_{j=0}^{p} \beta_j^2 \right] \tag{3.16}$$

avec $\lambda \leq 0$ et $\alpha \in [0;1]$ pour nous ramener à un problème d'optimisation à une seule contrainte au lieu de deux. Par ailleurs, il s'agit bien l'effet du Lasso qui retirera de manière optimale les variables qui n'apportent pas suffisamment d'éclairage sur la variable réponse.

Quelle valeur d'α choisir ? Nous allons faire varier α de 0 à 1 par pas réguliers, et nous réaliserons à chaque valeur une cross validation 10-folds avec le GLM pénalisé correspondant. Nous noterons enfin pour chacune l'erreur quadratique moyenne (MSE) associée et choisirons l'α qui la minimise. Les résultats sont détaillés dans le tableau ci-dessous :

$\boldsymbol{\alpha}$	**MSE**
0	0,3055303
0,1	0,3046473
0,2	0,3052499
0,3	0,3050955
0,4	0,3039155
0,5	0,3045055
0,6	0,3051975
0,7	0,3046806
0,8	0,3042719
0,9	0,3042645
1	0,3046494

TABLE 3.1 – Performance des GLM pénalisés Elastic Net en fonction d'α

Le tableau 3.1 indique que le MSE est minimisé lorsqu'α est égal à 0,4. Nous retenons donc les variables qui ont été sélectionnées par l'Elastic Net correspondant et réalisons un simple GLM non contraint à partir de ces variables. Nous obtenons alors un **critère d'information d'Akaike** (AIC) égal à 210882,8 et un **Critère d'information bayésien** égale à 210936,6. L'AIC et le BIC sont des indicateurs de la qualité de la prédiction GLM, et leur calcul est détaillé en annexe 4. Plus la valeur de l'AIC ou du BIC sera petite 209721.7, meilleure sera la qualité de prédiction.

3.3.4.2 GLM avec sélection des variables explicatives par GBM

Nous pouvons sélectionner les variables qui expliquent le mieux le logarithme du pourcentage de crédit commercial en amont grâce à un GBM. Nous réalisons d'abord un grid search en faisant notamment varier la profondeur des arbres et le taux d'apprentissage à chaque itération, et nous choisissons la combinaison de paramètres qui minimise l'erreur de prédiction au carré.

A partir des variables sélectionnées, parmi lesquelles figure la majoration TTC, nous réalisons un GLM présentant un **AIC** égal à 209695,4 et un **BIC** égal à 209721,7.

3.3.4.3 Choix du meilleur GLM et résultats

Le modèle sélectionné est celui qui minimise l'AIC et le BIC : il s'agit donc du modèle qui sélectionne les variables par GBM. Ce modèle présente toutefois une erreur quadratique moyenne légèrement supérieure à celle du GLM obtenu à l'issue de l'Elastic Net : $MSE_Z = 0,3119289$. Quel est l'impact du MSE sur la prédiction ? Rappelons que cet indicateur est calculé pour la prédiction de $Z = log(Y)$, nous avons donc :

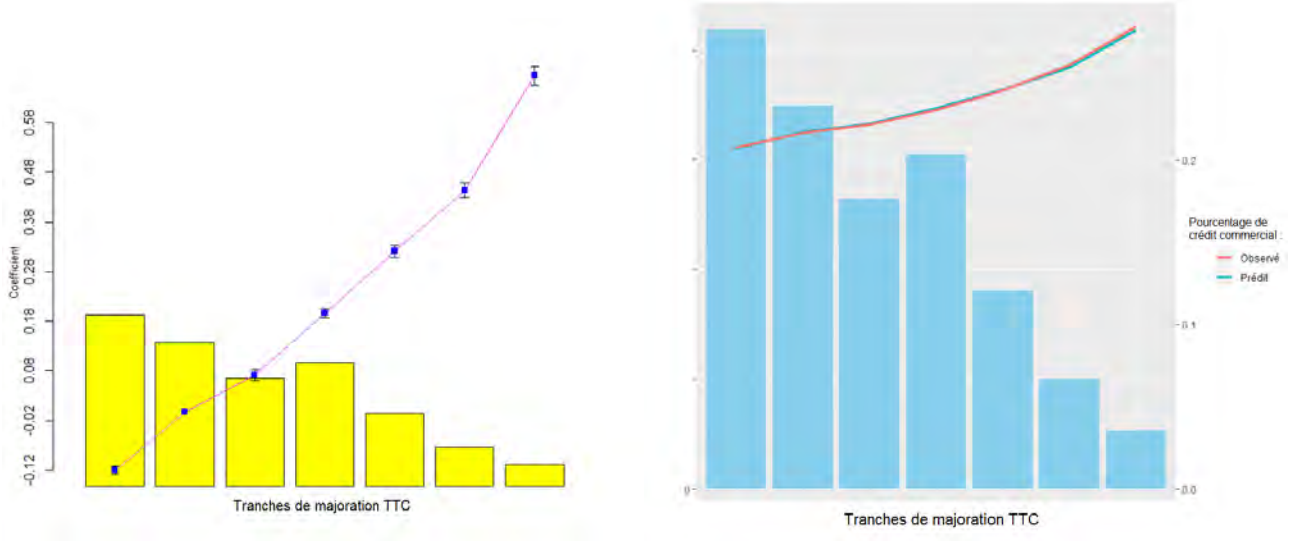
$$\text{MSE}_Z = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Z_i} - Z_i)^2 \tag{3.17}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\log(\widehat{Y_i}) - \log(Y_i))^2 \tag{3.18}$$

Il faut donc calculer en complément l'erreur quadratique moyenne directement sur les pourcentages de crédits commerciaux :

$$\text{MSE}_Y = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y_i} - Y_i)^2 = 0,0033673 \tag{3.19}$$

A l'échelle du pourcentage, cette erreur est suffisamment petite pour justifier que notre approximation de distribution log-normale de Y est raisonnable. La figure 3.11 (b), qui trace la moyenne, par tranche de majoration, du pourcentage de crédit commercial prédit versus l'observé, indique effectivement que la prédiction n'est pas mauvaise. Nous remarquons également que l'allure des coefficients présentée en figure 3.11 (a) suit la même tendance que les pourcentages de crédits commerciaux par tranche. En effet, dans le cadre d'une régression linéaire, plus la valeur d'un coefficient augmente, plus il contribue à l'augmentation de la valeur prédite de la variable réponse Z, et puisque la fonction exponentielle est croissante, l'augmentation du coefficient contribue donc à l'augmentation de la valeur prédite de $Y = e^Z + \frac{s^2}{2}$. En outre, nous remarquons que les intervalles de confiance pour chaque coefficient sont relativement petits, mais augmentent lorsque l'exposition des observations diminue dans une tranche de majoration.

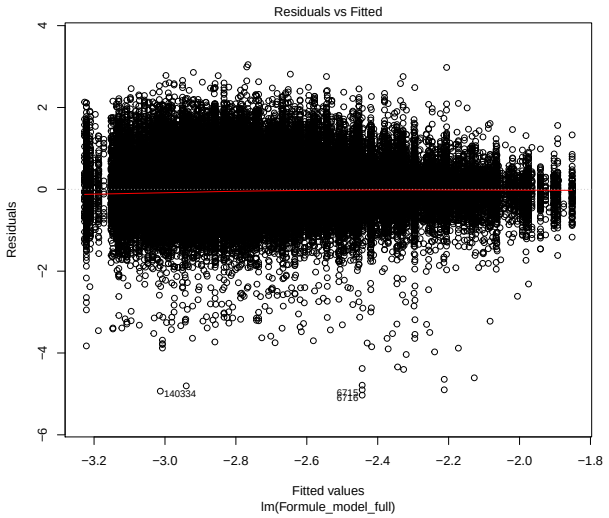


(a) Allure des coefficients pour la majoration TTC pour la prédiction de Z

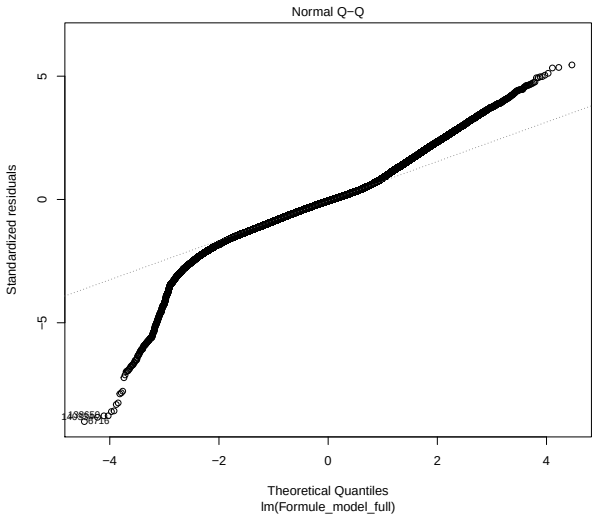
(b) Prédiction VS observation de Y par tranche de majoration TTC

FIGURE 3.11 – Résultats du GLM

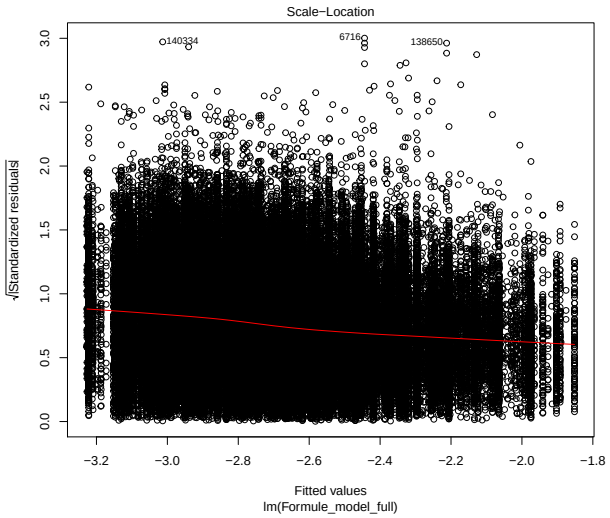
Il est également souhaitable d'étudier les résidus pour compléter notre évaluation de la qualité de la régression sur Y.



(a) Résidus en fonction des valeurs prédites



(b) Q-Q plot des résidus



(c) Envergure de la répartition des résidus

FIGURE 3.12 – Etude des résidus pour la prédiction de Z

La figure 3.12 (a) indique la valeur des résidus en fonction des valeurs prédites correspondantes. Nous remarquons alors que les valeurs des résidus ne sont pas corrélées aux variables réponses puisqu'ils sont à peu près répartis uniformément autour d'une ligne horizontale. Cette ligne, illustrée en rouge sur la figure, indique la moyenne des résidus en fonctions de la valeur prédite de la variable réponse : elle est très proche de zéro, ce qui permet de confirmer l'hypothèse $\mathbb{E}[\varepsilon] = 0$ nécessaire à la validité d'une modélisation linéaire. L'horizontalité de cette ligne indique également qu'il n'y a pas de dépendance entre les résidus et la prédiction, et donc que la variable réponse est bien linéairement dépendante des variables explicatives.

La figure 3.12 (b) quant à elle affiche le Q-Q plot des résidus. Nous remarquons que, plus nous nous éloignons de la médiane, plus les résidus s'éloignent de la ligne pointillée en diagonale indiquant une distribution normale. Cela signifie que les résidus ne sont pas linéairement distribués. Or cela n'est pas réellement problématique :

en effet, l'hypothèse de normalité du bruit n'est pas indispensable, elle est nécessaire pour s'assurer que la distribution de chaque estimateur des coefficients de la régression linéaire conditionnellement à X suit une loi normale, mais dans le cas où ε n'est pas gaussien, alors les estimateurs des coefficients conditionnellement à X suivent de manière asymptotique une distribution normale. Cette information est utile si nous souhaitons calculer les intervalles de confiance des coefficients, mais dans le cadre de notre étude

Enfin, la figure 3.12 (c) représente la racine carré des résidus standardisés [5] en fonction de la valeur de la prédiction. Elle permet de vérifier l'hypothèse homoscedasticité sur nos données. La ligne rouge indique la moyenne de la racine carrée résidus en fonctions de la valeur prédite de la variable réponse : étant donnée qu'elle est quasiment horizontale, cela signifie que la variance des résidus ne dépend pas trop de la variable réponse, validant ainsi grossièrement l'hypothèse d'homoscedasticité.

Conclusion du chapitre

La modélisation de l'espérance de crédit commercial appliquée à chaque contrat se fait en deux étapes : la première par un modèle logistique indiquant la probabilité qu'un rabais soit accordé pour un contrat donné, la seconde par un modèle prédisant le pourcentage de crédit commercial qui serait appliqué dans le cas où un tel geste commercial aurait lieu. La majoration TTC fait partie des variables les plus significatives dans ces deux modèles, et de ce fait l'algorithme d'optimisation sera sensible à l'intégration d'un tel modèle de rabais.

En outre, bien que la distribution du pourcentage de crédit commercial ait été approximée par une loi log-normale, les performances de prédiction nous satisfont suffisamment pour que nous décidions d'utiliser ces résultats dans la suite de notre étude. Il faut en effet signaler qu'introduire le rabais dans notre calcul de l'ELR permet de ne pas surestimer le chiffre d'affaires, et, par conséquent, l'optimisation héritera de cette vison prudente.

5. Résidus standardisés : c'est la valeur des résidus divisés par l'écart type.

Chapitre 4

Optimisation de la majoration au terme

Sommaire

Préambule

Nous disposons à présent d'un modèle de résiliation au terme et d'un modèle de crédit commercial. Nous sommes donc capable de prédire pour chaque contrat la probabilité que le client le résilie ainsi que la rentabilité et le chiffre d'affaires apporté par ce contrat. Finalement, nous avons toutes les informations nécessaires pour répondre à notre problématique d'optimisation du taux de résiliation moyen du portefeuille sous contrainte de l'ELR global.

La qualité de la solution (si elle existe) au problème d'optimisation dépend de la façon dont ce dernier est posé. Nous allons donc dans un premier temps rappeler notre programme d'optimisation et déterminer de quelle manière le résoudre. Pour ce faire, nous nous baserons sur la théorie du Lagrangien et du problème dual, rappelée en annexe 6.

4.1 Programme d'optimisation

Rappelons notre objectif : il s'agit de minimiser le taux moyen des résiliations sur le portefeuille sous contrainte de maintenir l'ELR total du portefeuille sous un certain seuil :

$$\begin{array}{rl} \min\limits_{x_1,\dots,x_N} & \frac{1}{N}\sum_{i=1}^{N} f_i(x_i) \\ \text{s.c.} & ELR_{tot} \leq ELR_{max} \end{array} \tag{4.1}$$

avec :

- N le nombre de contrats dont la majoration est à optimiser ;
- x_i la majoration hors taxes appliquée au contrat i ;
- f_i la fonction de résiliation, qui prédit la probabilité de résiliation d'un contrat i en fonction de la majoration x_i ;

- ELR_{tot}, l'ELR du portefeuille calculé selon la formule :

$$ELR_{tot} = \frac{\sum_{i}^{N} PP_i \times (1 - f_i(x_i))}{\sum_{i}^{N} PC_{HT,n-1,i}(1 + x_i)(1 - f_i(x_i))\,(1 - \%CC_i(x_i) \times \text{Proba}_{CC,i}(x_i))} \tag{4.2}$$

où

  - $PC_{HT,n-1,i}$ la prime commerciale du contrat i hors taxes avant la majoration (n étant l'année suivant la majoration) ;
  - $\text{Proba}_{CC,i}$ la probabilité qu'un crédit commercial soit accordé au contrat i ;
  - $\%CC_i$ le pourcentage de crédit commercial appliqué sur la prime commerciale, dans le cas où ce crédit commercial a été accordé.

- ELR_{max}, la valeur maximale que ne doit pas dépasser l'ELR total du portefeuille.

De plus, gardons à l'esprit qu'il faudra vérifier *a posteriori* l'impact de cette optimisation sur le chiffre d'affaires obtenu, et vérifier qu'il n'aura pas été dégradé par rapport à la situation actuelle d'AXA. En effet, nous ne voulons pas introduire une autre contrainte d'inégalité sur le chiffre d'affaires directement dans le problème d'optimisation car cela rendrait celui-ci trop complexe à résoudre avec les moyens que nous avons à disposition.

En effet, rappelons que nous souhaitons réaliser une optimisation à N variables (correspondant aux majorations des N contrats). Une contrainte implique, dans le cadre de la minimisation du Lagrangien, de déterminer la valeur d'un multiplicateur de Lagrange λ en plus des N variables, en passant par le problème dual (l'annexe 6 rappelle la méthode du problème dual et la manière dont les multiplicateurs de Lagrange sont choisis). Or, l'ajout d'une contrainte supplémentaire engendre l'apparition d'un autre multiplicateur de Lagrange qu'il faudra aussi déterminer. Cela va inévitablement solliciter davantage de mémoire pour que l'ordinateur puisque réaliser les calculs, et étant donné le nombre important de contrats (plusieurs milliers de contrats) sur lesquels nous allons réaliser l'optimisation, le temps de calcul risque d'être très long. C'est pourquoi nous souhaitons garder le problème d'optimisation simple, c'est-à-dire avec le minimum de contraintes possible, afin que sa résolution soit réalisable par nos ordinateurs. Et nous verrons par la suite que cette simplicité nous permettra tout de même de répondre au cahier des charges.

Enfin, remarquons que ce problème de minimisation de la moyenne du taux de résiliation est équivalent à minimiser la somme des taux de résiliation, ce qui sera plus simple à résoudre par la suite :

$$\begin{aligned} \min_{x_1,\dots,x_N} \quad & \textstyle\sum_{i=1}^{N} f_i(x_i) \\ \text{s.c.} \quad & ELR_{tot} \leq ELR_{max} \end{aligned} \tag{4.3}$$

4.1.1 Transformation de l'optimisation globale

Tel quel, ce problème d'optimisation présente une contrainte globale sur l'ELR non linéaire, ce qui rend la résolution du problème complexe et difficile en considérant un large nombre de contrats dont la majoration est à optimiser. Il est toutefois possible de modifier ce problème d'optimisation en transformant la contrainte globale sur l'ELR pour la rendre linéaire :

$$
\begin{aligned}
& ELR_{max} \geq ELR_{tot} \\
\Leftrightarrow \quad & ELR_{max} \geq \frac{\sum_{i}^{N} PP_i \times (1 - f_i(x_i))}{\sum_{i}^{N} PC_{HT,n-1,i}(1 + x_i)(1 - f_i(x_i))\,(1 - \%CC_i(x_i) \times \text{Proba}_{CC,i}(x_i))} \\
\Leftrightarrow \quad & 0 \geq \sum_{i}^{N} PP_i \times (1 - f_i(x_i)) \\
& \qquad - ELR_{max} \sum_{i}^{N} PC_{HT,n-1,i}(1 + x_i)(1 - f_i(x_i))\,(1 - \%CC_i(x_i) \times \text{Proba}_{CC,i}(x_i)) \\
\Leftrightarrow \quad & 0 \geq \sum_{i}^{N} g_i(x_i)
\end{aligned}
$$

Nous pouvons en déduire le Lagrangien :

$$L(x_1, ..., x_N, \lambda) = \sum_{i=1}^{N} f_i(x_i) + \lambda \sum_{i=1}^{N} g_i(x_i) \tag{4.4}$$

Nous supposons que, quelques soient $i, j \in [\![1; N]\!]$ avec $i \neq j$, les fonctions f_i et f_j sont indépendantes (hypothèse forte, puisque comme nous l'avons vu au chapitre 2, les clients peuvent se concerter et il n'est pas exclu que le comportement d'un client puisse influencer celui d'un autre client, notamment en termes de résiliation), g_i et g_j également, et il en est de même pour les majorations x_i et x_j. Ainsi, minimiser la somme de taux de résiliation $\sum_{i=1}^{N} f_i(x_i)$ sous la contrainte $\sum_{i=1}^{N} g_i(x_i)$ revient à minimiser chaque $f_i(x_i)$ sous la contrainte $g_i(x_i)$. Nous nous retrouvons donc avec N problèmes d'optimisation :

$$\forall i \in [\![1; N]\!], \quad \min_{x_i} f_i(x_i) \tag{4.5}$$

$$\text{s.c.} \quad g_i(x_i) \leq 0 \tag{4.6}$$

Les N Lagrangiens correspondants sont donc :

$$\forall i \in [\![1; N]\!], \quad L_i(x_i, \lambda) = f_i(x_i) + \lambda g_i(x_i) \tag{4.7}$$

et de ce fait nous avons :

$$L(x_1, ..., x_N, \lambda) = \sum_{i=1}^{N} L_i(x_i, \lambda) \tag{4.8}$$

Finalement, pour tout λ, minimiser le Lagrangien L revient à minimiser chacun des L_i et à les sommer :

$$\forall \lambda, \qquad \min_{x_1,...,x_N} L(x_1, ..., x_N, \lambda) = \min_{x_1,...,x_N} \left(\sum_{i=1}^{N} L_i(x_i, \lambda) \right) \tag{4.9}$$

$$= \sum_{i=1}^{N} \left(\min_{x_i} L_i(x_i, \lambda) \right) \tag{4.10}$$

* * *

Preuve :

Nous voulons montrer que si, pour tout $i \neq j$, L_i et L_j sont indépendants, de même pour x_i et x_j, alors :

$$\min_{x_1,\ldots,x_N} \left(\sum_{i=1}^{N} L_i(x_i, \lambda) \right) = \sum_{i=1}^{N} \left(\min_{x_i} L_i(x_i, \lambda) \right)$$

1ère étape, existence du minimum :

Pour tout i, L_i est continue sur un intervalle fermé borné $I = [x_{min}; x_{max}]$ par somme de fonctions continues sur cet intervalle I. En effet, f_i et g_i sont continues sur cet intervalle.

Or, pour toute fonction f continue sur un intervalle fermé borné [a, b], f est bornée sur [a, b] et atteint ses bornes au sein de l'intervalle [a, b]. C'est-à-dire que l'image d'un segment par une fonction continue est un segment.

Ainsi, il existe deux réels m et M appartenant à I tels que $L_i(I) = [m; M] : m$ (resp. M) est donc le minimum (resp. le maximum) de L_i sur I.

2ème étape, inégalité triviale :

Montrons que $\min_{x_1,\ldots,x_N} \left(\sum_{i=1}^{N} L_i(x_i, \lambda) \right) \geq \sum_{i=1}^{N} \left(\min_{x_i} L_i(x_i, \lambda) \right)$.

$$\forall i \text{ et } \forall x_i, \quad L_i(x_i) \geq \min_{x_i} L_i(x_i)$$

Donc :

$$\forall x_i, \quad \sum_{i=1}^{N} L_i(x_i) \geq \sum_{i=1}^{N} \min_{x_i} L_i(x_i)$$

et *a fortiori* :

$$\min_{x_1,\ldots,x_N} \left(\sum_{i=1}^{N} L_i(x_i) \right) \geq \sum_{i=1}^{N} \min_{x_i} L_i(x_i)$$

3ème étape, autre sens de l'inégalité :

Montrons que $\min_{x_1,\ldots,x_N} \left(\sum_{i=1}^{N} L_i(x_i, \lambda) \right) \leq \sum_{i=1}^{N} \left(\min_{x_i} L_i(x_i, \lambda) \right)$.

Nous pouvons sélectionner un N-uplet $(x_1^*, ..., x_N^*)$ tel que

$$\forall i, \quad L_i(x_i^*) = \min_{x_i} L_i(x_i)$$

Un tel N-uplet existe car les L_i sont indépendants (car les f_i et g_i le sont) et car les x_i sont aussi indépendants.

Nous obtenons alors :

$$\sum_{i=1}^{N} L_i(x_i^*) = \sum_{i=1}^{N} \min_{x_i} L_i(x_i)$$

Par ailleurs, nous pouvons écrire :

$$\begin{aligned}\min_{x_1,...,x_N} L(x_1,...,x_N) &= \min_{x_1,...,x_N} \sum_{i=1}^{N} L_i(x_i) \\ &\leq \sum_{i=1}^{N} L_i(x_i), \quad \forall i \text{ et } \forall x_i\end{aligned}$$

Donc en particulier :

$$\begin{aligned}&\min_{x_1,...,x_N} L(x_1,...,x_N) \leq \sum_{i=1}^{N} L_i(x_i^*) \\ \iff \quad &\min_{x_1,...,x_N} \left(\sum_{i=1}^{N} L_i(x_i) \right) \leq \sum_{i=1}^{N} \min_{x_i} L_i(x_i)\end{aligned}$$

* * *

Notons que ce qui relie tous ces Lagrangiens est le λ commun : c'est bien lui qui permet de retrouver le problème d'optimisation initial donné par l'équation (4.3). Ainsi, minimiser chaque Lagrangien L_i pour le même lambda revient à minimiser le Langragien L du problème initial.

Il est à présent légitime de se poser la question suivante : pouvons-nous trouver les majorations optimales $(x_1, ...x_N)$ résolvant le problème dual ?

Rappelons-le, le problème dual s'écrit :

$$\begin{aligned}\max_{\lambda} \quad & q(\lambda) \\ \text{s.c.} \quad & \lambda \geq 0, \quad \forall i = 1, ..., m\end{aligned} \tag{4.11}$$

avec $q(\lambda) = \min_{x_1,...,x_N} L(x_1, ..., x_N, \lambda) = \min_{x_1,...,x_N} \left(\sum_{i=1}^{N} f_i(x_i) + \lambda \sum_{i=1}^{N} g_i(x_i) \right)$.

Cela est possible si et seulement si les fonctions de résiliation f_i sont bien convexes. Or, les hypothèses dont nous disposons ne suffisent pas à affirmer une telle convexité pour chaque f_i. Lorsque nous traçons quelques fonctions du taux de résiliation correspondant à des individus distincts en fonction de la majoration, nous constatons effectivement sur la figure 4.1 que les fonctions de résiliation de 10 clients pris au hasard dans notre base ne sont pas convexes.

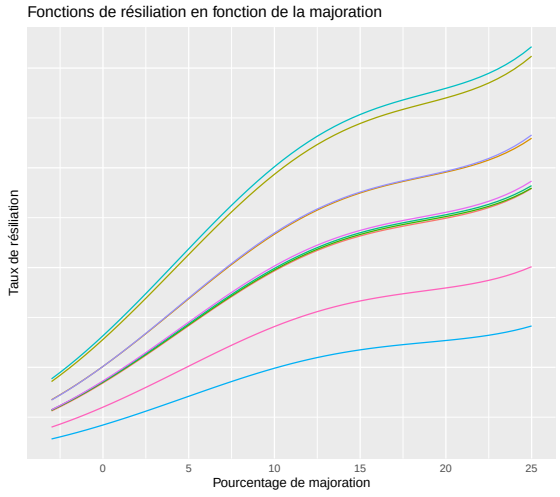


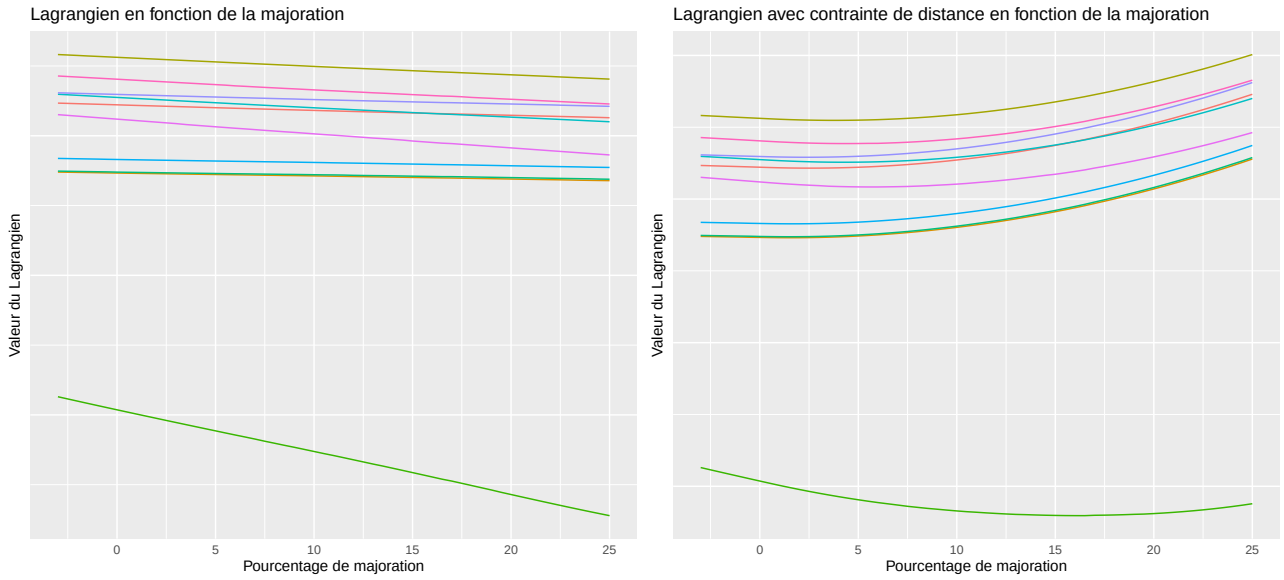
FIGURE 4.1 – Allure des fonctions de résiliation en fonction de la majoration TTC

Ce constat suffit à nous indiquer que le problème dual ne nous permettra pas de trouver une solution globale, quelle que soit la contrainte d'inégalité globale $ELR \leq ELR_{max}$, conformément à la théorie détaillée en annexe 6.

En revanche, il est possible de trouver des solutions optimales en faisant varier le poids de la contrainte : plus λ est grand, plus l'optimisation accorde de l'importance à la contrainte. Le choix de λ reposera alors sur l'impact de l'optimisation *a posteriori* sur les indicateurs clés de performance (que nous appellerons par la suite *KPIs*, Key Performance Indicators) que sont l'ELR total, le taux de résiliation moyen ainsi que le chiffre d'affaires. Notons que même si le chiffre d'affaires n'apparaît pas dans le problème d'optimisation, rien n'empêche de l'évaluer pour un λ donné. Cela nous permettra de s'assurer que le chiffre d'affaires n'est pas dégradé par le travail d'optimisation.

4.1.2 Convexification du Lagrangien

Un autre problème se pose : celui de la convexité des Lagrangiens L_i. En effet, quelque soit la valeur positive de λ, les L_i ont des allures concaves, et lorsque λ prend de grandes valeurs, les L_i varient presque linéairement avec la majoration. Nous le voyons par exemple sur la figure 4.2 (a) qui présente l'allure des Lagrangiens en fonction de la majoration pour 10 individus choisis au hasard dans la base de données. Cela est problématique car dans ce cas nous avons une solution en coin, et sur la figure, cela implique que ces 10 Lagrangiens atteignent leur minimum pour une majoration de 25%. Ce n'est évidemment pas le résultat que nous recherchons. Pour y remédier, il est possible d'ajouter de la convexité au Lagrangien en rajoutant une contrainte.



(a) Lagrangien en fonction de la majoration

(b) Lagrangien contraint en fonction de la majoration

FIGURE 4.2 – Observation de l'allure des Lagrangiens non pénalisés et pénalisés

Contrainte sur la distance de la majoration à une certaine majoration cible

L'idée est de rajouter une contrainte sur la distance de la majoration x_i par rapport à une majoration de

référence que nous choisissons. Le problème d'optimisation devient :

$$\forall i \in [\![1; N]\!], \quad \min_{x_i} f_i(x_i) \tag{4.12}$$

$$\text{s.c.} \quad g_i(x_i) \leq 0 \tag{4.13}$$

$$(x_i - \alpha)^2 = 0 \tag{4.14}$$

avec α une constante, du même ordre de grandeur que la majoration, que nous ajusterons. On a donc les Lagragiens suivants pour chaque individu i :

$$\forall i \in [\![1; N]\!], \quad L_i(x_i, \lambda) = f_i(x_i) + \lambda g_i(x_i) + \mu(x_i - \alpha)^2 \tag{4.15}$$

avec $\lambda \leq 0$ et μ les multiplicateurs de Lagrange. μ doit être positif dans notre cas particulier afin d'ajouter de la convexité au Lagrangien.

Nous précisons néanmoins que nous ne souhaitons pas absolument que la contrainte de distance sur la majoration soit satisfaite, et donc nous ne voulons pas que la valeur du multiplicateur de Lagrange μ associé à cette contrainte ne soit pas trop important, mais suffisamment élevé pour que cette contrainte apporte de la convexité au Lagrangien. Nous expliquons comment nous choisissons une valeur appropriée de μ dans le paragraphe suivant.

Choix des paramètres λ, μ et α

Comment choisir λ, μ et α ? La contrainte de distance à la valeur α permet d'indiquer une distribution de majorations optimales plus ou moins centrée autour de ce que nous souhaitons avoir comme majoration moyenne cible (qui n'est cependant pas égale à α), mais nous ne nous focalisons pas tant sur cette contrainte. Dans un cadre idéal, il faudrait considérer une multitude de combinaisons de μ, de λ et d'α, puis choisir ces paramètres tels que nous répondions au cahier des charges en terme de taux moyen de résiliation, d'ELR total, de chiffre d'affaires et de distribution de majorations. Dans la pratique, cette démarche est coûteuse en terme de capacité de calcul. C'est pourquoi nous faisons le choix de fixer les paramètres μ et α à certaines valeurs initiales que nous ajusterons éventuellement : nous lançons dans un premier temps l'optimisation pour une première valeur d'α égale à 5%, et une première valeur de μ égale à 1. A l'issue de cette optimisation, nous pouvons tracer l'évolution des KPIs en fonction de la valeur de λ et nous choisissons une valeur de λ qui satisfait les demandes du cahier des charges. Nous regardons ensuite la distribution des majorations associée à ce λ. Si la moyenne des majorations optimales s'approche bien de la moyenne cible, alors la distribution nous satisfait. Le cas échéant, nous ajustons la valeur d'α : plus alpha est petit, plus la distribution est translatée vers la gauche. Une fois satisfait de la majoration moyenne, il est possible d'ajuster l'écart-type autour de la moyenne avec la valeur de μ : plus μ est grand, plus l'écart-type est réduit. A noter que μ ne doit pas être trop petit au risque de ne pas rendre assez convexe le Lagrangien et rester sur une solution en coin.

Avec les valeurs de λ, de μ et α bien choisies, on constate bien sur la figure 4.2 (b) qu'une convexité a été introduite sur chaque Lagrangien L_i, nous permettant de d'obtenir des majorations optimales différentes.

4.2 Base de données : les contrats termés

Nous allons réaliser notre étude d'optimisation sur une base de contrats concernés par le même mois de terme. Etant donné qu'il n'y a pas de saisonnalité sur le taux de résiliation, nous supposons que les contrats de ce mois de terme sont représentatifs de l'ensemble des contrats du portefeuille. Pour plus de justesse, il est

possible de réaliser plus tard l'étude sur tout le portefeuille de contrats, mais cela reste coûteux en temps machine.

Cette base présente la même structure que celle sur laquelle nous avons réalisé le modèle de résiliation, à quelques différences près :

— La base contient uniquement les informations disponibles 2 mois avant le terme relatives aux contrats ;

— La base ne contient pas encore l'information de la majoration, car nous allons jouer sur cette variable pour résoudre notre problème d'optimisation.

A partir de cette base, nous calculons les primes pures associées à chaque contrat, puis en faisant varier la majoration, nous calculons la probabilité de résiliation et le pourcentage de crédit commercial espéré grâce aux modèles élaborés aux chapitres 2 et 3 afin d'estimer le coût espéré de sinistres et le chiffre d'affaire rapporté par contrat.

Notons que nous décidons de faire parcourir la majoration sur une plage de valeurs démarrant avec un niveau de majoration négatif. En effet, il serait intéressant de voir si des minorations sur certains contrats seraient bénéfiques à la diminution du taux de résiliation tout en améliorant la rentabilité du portefeuille.

4.3 Résolution du problème d'optimisation

4.3.1 Evolution des KPIs en fonction de λ

Pour des valeurs de μ et d'α bien choisies, nous obtenons l'évolution des KPIs qui nous intéressent en fonction de λ. La figure 4.3 indique ainsi en train plein (cf. légende du graphique) le taux de résiliation moyen, l'ELR total et le chiffre d'affaires.

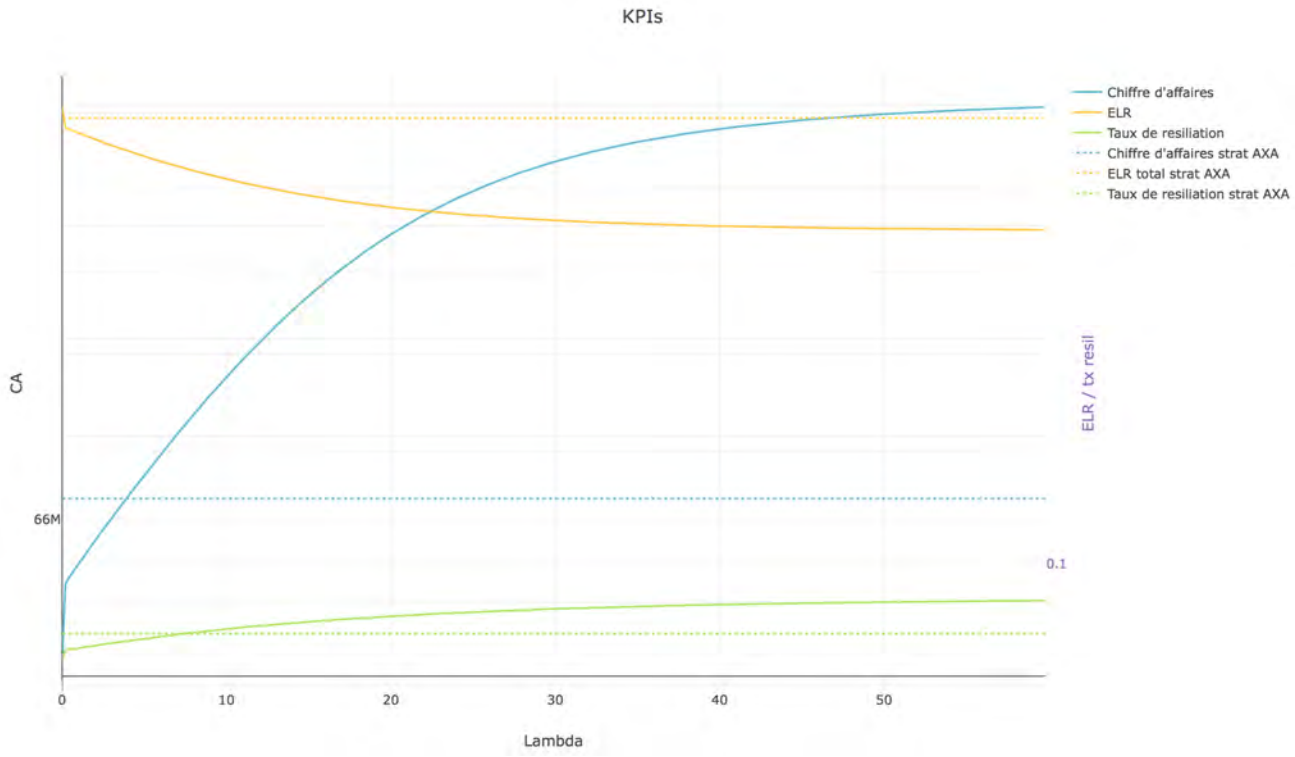


FIGURE 4.3 – Evolution des KPIs à l'issue de l'optimisation en fonction de lambda

En pointillé sont indiqués les mêmes KPIs correspondant à la stratégie actuelle d'AXA France. Pour comparer ce qui est comparable, ces niveaux ont été obtenus à partir de la base de contrats termés au même mois

mais de l'année précédente, diminuée des contrats qui ont été remplacés au cours de l'année, puisque pour ces contrats remplacés le risque assuré n'est plus le même. Sur cette base de comparaison nous savons également quels contrats ont été résiliés, et ainsi calculer un taux de résiliation.

Il est également intéressant de savoir à quel point la stratégie actuelle d'AXA est éloignée d'une situation optimale. La figure 4.4 présente donc la stratégie actuelle observée d'AXA (point rouge) par rapport à la frontière efficiente , c'est-à-dire le lieu des points (ELR_{total}, Taux de résiliation) optimaux qui répondent au problème d'optimisation selon différentes valeurs de λ. En faisant augmenter λ, nous parcourrons la frontière en partant du bas. En effet, rappelons que plus λ est grand, plus la contrainte de l'ELR total a d'importance au détriment de la diminution du taux moyen de résiliation. Cette figure est d'ailleurs cohérente avec la théorie illustrée en figure 1.3 au chapitre 1.

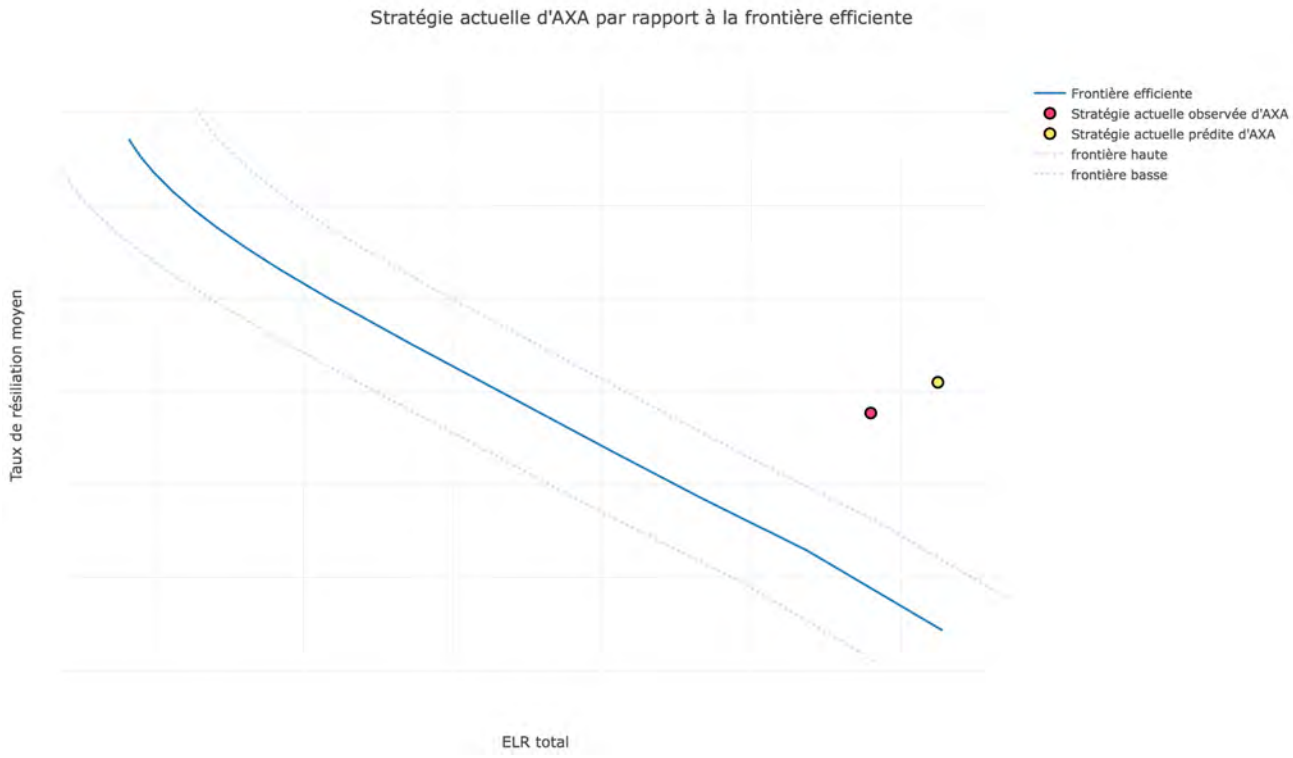


FIGURE 4.4 – Stratégie d'AXA France par rapport à la frontière efficiente

Sur la figure 4.4 est également indiquée par un point jaune la stratégie actuelle prédite d'AXA : il s'agit de l'estimation du taux de résiliation moyen des contrats et de l'estimation de l'ELR total, obtenues grâce à nos modèles de résiliation et de crédit commercial, avec les majorations qui ont effectivement été appliquées l'année précédente sur ces contrats. Cela indique à quel point notre prédiction est bonne. Nous observons ainsi que nous ne sommes pas trop éloignés de la réalité, bien que nous sur-estimons le taux de résiliation de 0,3 points et sur-estimons l'ELR de 0,9 points. Notre modèle indique donc une vision prudente.

Il est d'ailleurs important de noter que notre prédiction de la stratégie d'AXA ainsi que la frontière efficiente obtenue sont basés sur des estimateurs de prime pure, de probabilité de résiliation et de crédit commercial, obtenus par modélisation GLM. Il existe donc des intervalles de confiance de ces prédictions qu'il serait judicieux de reporter autour de la frontière efficiente. Toutefois, d'un point de vue opérationnel, de tels intervalles ne sont pas calculés car il n'y a pas d'intérêt à le faire en règle générale (par exemple dans le cadre de l'estimation de la prime pure). De plus, en pratique, il est difficile de calculer ces intervalles de confiance car il faudrait connaître, pour chaque modèle, les intervalles de confiance des estimateurs de chaque coefficient du GLM. Or, un logiciel

de modélisation tel que Akur8 ne fournit pas cette information.

Alors comment déterminer si la stratégie actuelle observée d'AXA est relativement loin de la frontière efficiente ? A défaut de pouvoir obtenir un intervalle de confiance, nous pouvons créer une marge autour de la frontière dont la distance est égale à la distance qui sépare la stratégie actuelle observée d'AXA de celle prédite. Cette marge est matérialisée par les lignes en pointillés sur la figure 4.4. Nous remarquons donc que la stratégie d'AXA observée est en dehors de cette marge et ainsi, nous pouvons considérer qu'elle est suffisamment loin de la frontière efficiente pour considérer que le travail d'optimisation a un sens.

4.3.2 Commentaire sur la frontière efficiente

Il est légitime de se demander pourquoi la stratégie actuelle d'AXA est relativement loin de la frontière optimale. Plusieurs théories pourraient expliquer cet écart :

— Nous pouvons tout d'abord penser à la méthode actuelle d'AXA France pour calculer les majorations à appliquer sur les contrats au terme. En effet, cette méthode n'est pas basée sur une optimisation mais plutôt sur des études empiriques sur certains critères comme l'ELR du contrat, ou encore sa sinistralité. Ces études permettent d'établir des règles de majorations mais ne garantissent pas l'obtention d'un état optimal ;

— Dans un environnement hyper concurrentiel, un geste commercial peut parfois être accordé sur les contrats du portefeuille d'un agent général ;

— Enfin, il est raisonnable de remettre en cause notre modèle, car il a été principalement construit avec une combinaison de GLM. Or la théorie des GLM se base notamment sur la loi des grands nombres et estime une moyenne. Ainsi, par définition, les GLM ne modélisent pas les cas extrêmes, ce qui peut expliquer un écart entre la réalité de la stratégie actuelle d'AXA et la frontière, qui ne prend pas en compte les valeurs extrême. A noter toutefois qu'une partie de cet effet est matérialisé par la marge en pointillés, et ne peut pas à elle seule expliquer la situation sous optimale de la stratégie actuelle.

Notre programme d'optimisation s'affranchit de ces règles établies et de cette dimension politique. En ce qui concerne les exceptions, l'optimisation ne les remet pas forcément en cause. En effet, il s'agit d'une nouvelle manière d'établir la majoration de manière rationnelle d'un point de vue économique. Et il n'est pas impossible qu'elle permette d'améliorer l'image de marque, notamment grâce à d'éventuelles minorations sur certains contrats.

4.3.3 Choix d'un λ approprié

Chaque point de la frontière efficiente matérialise un optimal pour différentes valeurs de lambda. Mais tous ces points ne répondent pas forcément à notre cahier des charges qui comprend :

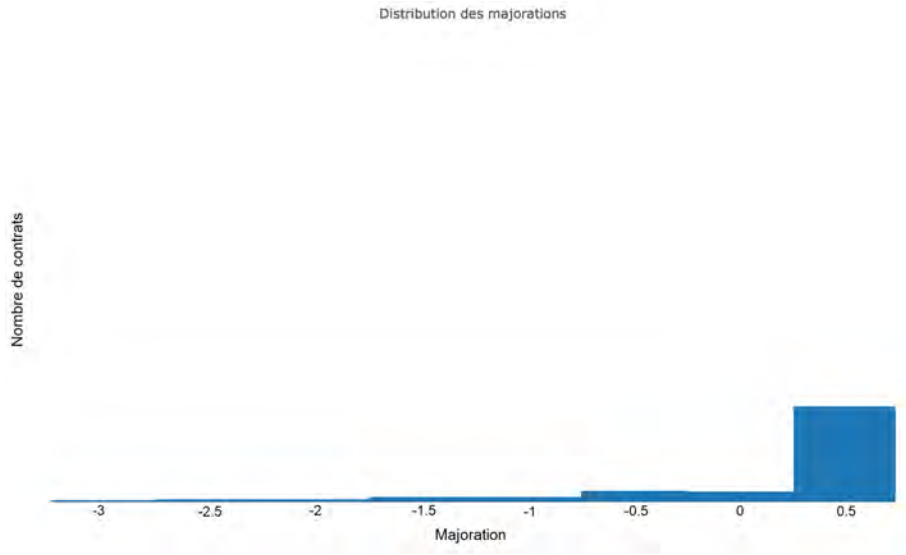
— La diminution du taux de résiliation moyen par rapport à la situation initiale d'AXA ;

— Le maintien voire la diminution de l'ELR total par rapport à l'ELR max donné par la situation initiale d'AXA ;

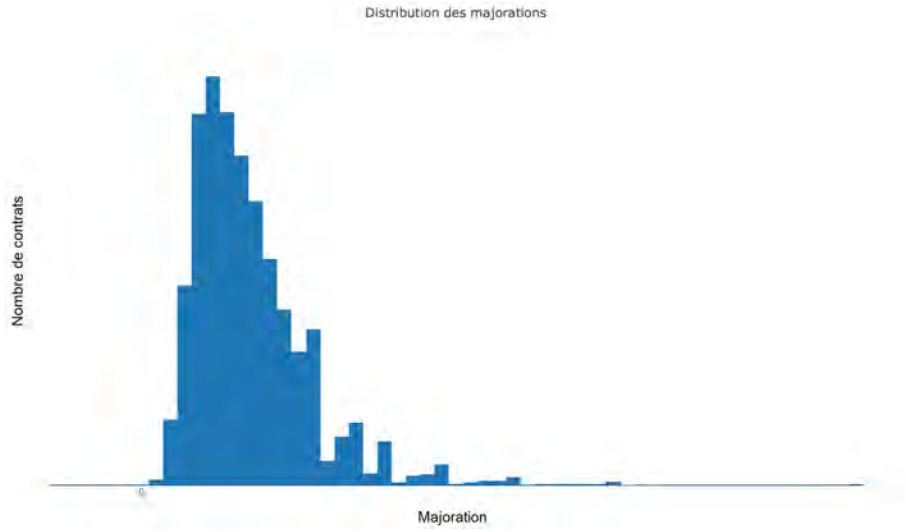
— Le maintien voire l'augmentation du chiffre d'affaires par rapport au chiffre d'affaires initial.

Un moyen de trouver une valeur de λ qui satisfait ces contraintes est de se référer à la figure 4.3, en comparant les niveaux de KPIs issus de l'optimisation avec le KPIs fournis par la stratégie actuelle d'AXA.

Ainsi, en choisissant un λ approprié, nous obtenons des majorations optimales dont la distribution est donnée en figure 4.5 (b).



(a) Queue de distribution : majorations négatives



(b) Distribution des majorations optimales

FIGURE 4.5 – Distribution des majoration optimales pour le λ choisi

L'allure de cette distribution nous satisfait de par la moyenne, et également par la concentration modérée des majorations autour de la moyenne. Notons la positivité du *skewness*[1], ce qui indique que des majorations modérées sont plus fréquentes que les majorations élevées.

En "zoomant" sur la queue de distribution à gauche, la figure 4.5 (a) indique la présence de majorations optimales négatives. Elles ne sont pas nombreuses par rapport au volume des majorations positives. Nous étudierons par la suite dans la section 4.4 ce qui caractérise les heureux contrats minorés.

4.4 Comparaison de la situation optimale avec la stratégie actuelle

Nous étudions ici ce qui caractérise la distribution des majorations optimales et ce qui la différencie de la distribution des majorations appliquées actuellement selon la stratégie d'AXA, ainsi que l'évolution des KPIs qui en découle.

1. Asymétrie de la distribution à gauche

4.4.1 Comparaison des distributions de majoration

La figure 4.6 indique l'allure des distributions de majoration avant et après l'optimisation. Les deux distributions sont très différentes. Déjà, la moyenne des majorations optimales hors taxe est de 4,4 points inférieure à la moyenne des majorations telle que le souhaite la stratégie actuelle d'AXA. De plus, si nous regardons comment ont évolué les majorations pour chaque contrat, nous constatons que seulement 14% des contrats se voient augmenter leur majoration après optimisation. La distribution des majorations optimales présente effectivement une concentration autour de la moyenne plus importante et une queue à droite bien plus fine par rapport à l'ancienne distribution, plus étalée.

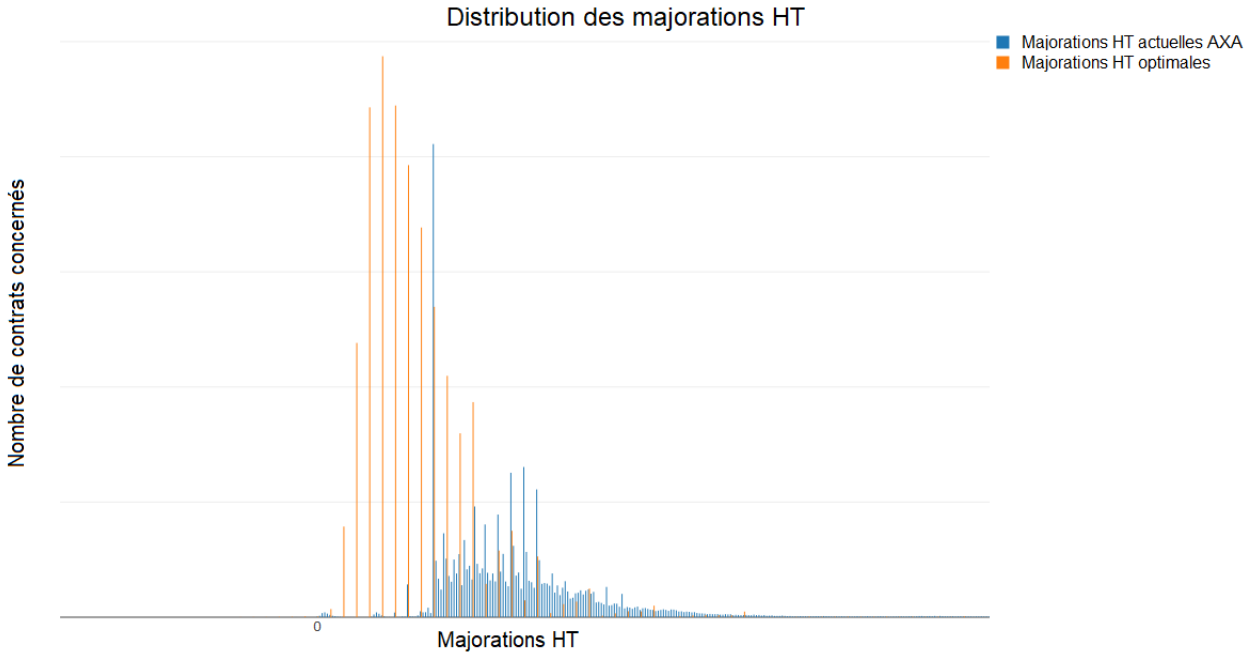


FIGURE 4.6 – Comparaison des distributions de majoration avant et après optimisation

4.4.2 Evolution des KPIs sur le mois étudié

Nous souhaitons savoir dans quelles proportions les KPIs se sont améliorés sur le mois étudié grâce à l'optimisation par rapport à la situation actuelle d'AXA France.

Ainsi, le tableau 4.1 indique la variation de la majoration, du taux de résiliation lié au terme et de l'ELR en points, tandis que le pourcentage d'évolution par rapport à la situation initiale est renseigné pour le chiffre d'affaires :

Δ Majoration HT	-4,4 points
Δ ELR	-2,9 points
Δ Taux de résiliation tarifaire liée au terme	-0,6 points
Evolution du chiffre d'affaires	0,0%

TABLE 4.1 – Evolution des KPIs après optimisation

Le taux de résiliation lié à la majoration au terme a été légèrement diminué. Cette diminution peut paraître anodine, néanmoins elle n'est pas négligeable si nous considérons le pourcentage déjà faible des résiliations tarifaires au terme, comparé aux résiliations non tarifaires. En revanche, la diminution de l'ELR n'est pas négligeable : cela indique que les résiliations concernent davantage les contrats non rentables puisque, comme l'indique le tableau, le chiffre d'affaires n'a pas évolué.

4.5 Etude des contrats aux majorations optimales extrêmes

4.5.1 Etude du segment des contrats minorés

Nous avons vu, en section 4.3.3 dans le cadre de l'optimisation, que certains contrats ont été minorés. Nous analysons ici ce qui caractérise le segment des contrats minorés, qui représente seulement 0,06% de la base de contrats à termer.

Pour ce faire, nous allons réaliser un arbre de décision [2] dont la variable réponse Y est l'indicatrice d'une majoration négative. Mais étant donné la faible occurrence de telles majorations, notre base de donnée est déséquilibrée. Ce déséquilibre ne sera pas sans conséquence sur notre analyse si nous ne pallions pas ce problème, car l'apprentissage ne prendra pas assez compte de la classe 1 de majorations négatives noyées dans le reste des données.

Rééquilibrage de la base de données avec l'algorithme SMOTE

A défaut de ne pouvoir récolter plus de données pour tenter de combler ce gap, une solution serait donc de réaliser un "*over sampling*", c'est-à-dire un "sur échantillonage", à partir des données que nous avons à disposition, afin d'obtenir des proportions de représentation de chaque classe plus équilibrées. Plus concrêtement, nous allons introduire un biais dans notre base de données en ajoutant de nouvelles observations que nous aurons nous même construites en utilisant la méthode du SMOTE, "*Synthetic Minority Over-sampling Technique*", qui va générer des observations synthétiques portant les étiquettes des classes sous représentées.

L'algorithme SMOTE se focalise sur un échantillon de notre base, composé d'une classe sous représentée (par exemple, dans notre cas, les majorations négatives), et récupère ses variables caractéristiques. A partir de ces informations, il est possible de créer un échantillon d'observations artificielles correspondant à des majorations négatives. Ces nouvelles observations sont créées de la manière suivante : un point de l'échantillon initial (c'est-à-dire une vraie observation) est choisi au hasard et les k-plus proches voisins, dans l'espace vectoriel composé des variables caractéristiques, de ce point seront considérés. Un vecteur est alors tracé entre le point de départ et l'un des k-plus proches voisins, puis la nouvelle "observation synthétique" sera trouvée en multipliant ce vecteur par une valeur entre 0 et 1. En quelque sorte, la nouvelle observation artificielle dont la majoration est négative est une combinaison d'observations réelles dont les majorations sont aussi négatives. Nous répétons ce procédé jusqu'à avoir un nombre total d'observations suffisant pour la classe minoritaire considérée.

Construction de l'arbre de décision

Sur la nouvelle base de données moins déséquilibrée grâce à l'application de l'algorithme SMOTE, nous réalisons un arbre de décision avec comme variable réponse Y, égale à 1 si la majoration est négative, égale à 0 sinon. Le résultat de l'apprentissage fournit l'arbre qui est illustré en figure 4.7. Il indique qu'un contrat dont les variables descriptives répondent aux critères menant à la dernière feuille en bas à droite ont plus de chance de présenter une majoration négative. Cela nous permet d'identifier ce qui caractérise les contrats minorés.

2. Le fonctionnement de l'arbre de décision est rappelé en annexe 3.

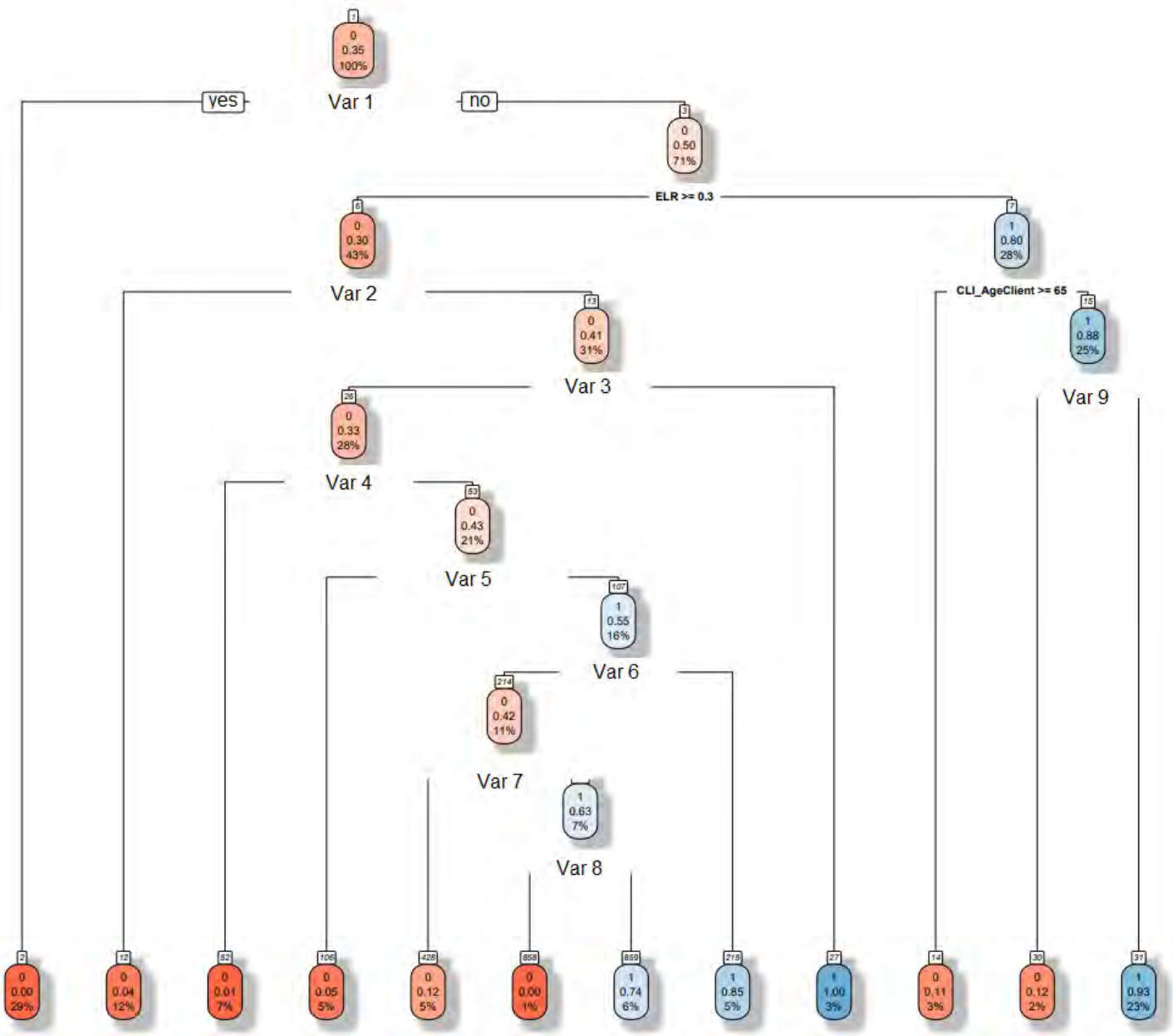


FIGURE 4.7 – Arbre de décision sur les majorations négatives VS positives

La matrice de confusion (dont le principe est rappelé en annexe 5) issue de la prédiction de cet arbre sur les données de test issues du SMOTE est la suivante :

		Observation	
		Majoration positive	Majoration négative
Prédiction	Majoration positive	1864	65
	Majoration négative	116	1024

TABLE 4.2 – Matrice de confusion

Cette classification présente donc une précision de 94%, ce qui est satisfaisant.

Nous avons volontairement anonymisé la plupart des variables descriptives qui apparaissent sur cet arbre, mais nous pouvons dire, par exemple, que les majorations négatives sont préférablement appliquées sur les

contrats rentables (ELR inférieur à 30%), ce qui est d'ailleurs rassurant. En outre, l'arbre indique qu'il faudrait minorer des clients âgés de moins de 65 ans, c'est-à-dire les non seniors. Cela est intéressant : la stratégie actuelle d'AXA "protège" les segments de population plus âgées que la moyenne du portefeuille en limitant leurs majorations alors que ce sont bien les plus jeunes qui sont sensibles à la majoration, comme nous l'avons vu lors des statistiques descriptives au chapitre 1. Il n'est d'ailleurs pas exclu de penser que les segments plus âgées deviendraient peut-être aussi sensibles à la majoration sans cette protection.

L'optimisation prend en compte cette sensibilité au prix et décide de favoriser les personnes sensibles à la variation de majoration et dont les contrats sont rentables afin de les garder en portefeuille.

Analyse du segment dont les contrats sont probablement minorés

Ce segment représente 2,36% des contrats de la base initiale (avant le SMOTE), et 1% des contrats de ce segment présentent effectivement une majoration négative. Nous allons donc comparer la majoration moyenne de ce segment pris dans sa totalité avant et après optimisation, ainsi que le taux de résiliation due au terme :

Δ Majoration HT	-6,9 points
Δ Taux de résiliation tarifaire liée au terme	-24,6 points

TABLE 4.3 – Evolution de la majoration et du taux de résiliation après optimisation

En effet, la majoration a été nettement diminuée, prédisant une grande diminution du taux de résiliation due au terme.

Nous allons également nous assurer que la sinistralité de ce segment n'est pas trop importante par rapport à celle de l'ensemble du portefeuille, car autrement, cela n'aurait aucun sens de minorer des contrats appartenant à un segment à forte sinistralité :

	Segment	Tout le portefeuille MRH
Fréquence moyenne de sinistres	102,90%	100%
Coût moyen des sinistres	84,79 €	100 €

TABLE 4.4 – Comparaison de la sinistralité de ce segment avec le portefeuille MRH en base 100

Ce segment présente une fréquence de sinistres légèrement plus élevée par rapport à l'ensemble du portefeuille MRH, mais le coût moyen de ces sinistres est tout de même moins important.

En conclusion, la minoration de contrats appartenant à ce segment est pertinente : il s'agit de contrats peu sinistrés, rentables et détenus par des clients sensibles à la majoration.

4.5.2 Etude du quantile 90% des majorations

Nous souhaitons maintenant étudier les contrats qui subissent une très forte majoration à l'issue de l'optimisation, et les comparer aux contrats fortement majorés selon la stratégie actuelle d'AXA. Nous décidons donc d'analyser les contrats dont les majorations sont supérieures au quantile 90%. La comparaison est résumée dans le tableau 4.5.

Les quantiles 90% des majorations avant et après optimisation sont très différents, ce qui induit que la moyenne des majorations au-delà du quantile 90% après l'optimisation est largement diminuée par rapport à

Δ $q_{90\%}$	-5,9 points
Δ Majoration moyenne HT	-11,1 points
Δ Taux de résiliation tarifaire au terme	-5,3 points
Δ ELR	+5,7 points
Δ Fréquence de sinistres	-16 points
Variation du coût moyen des sinistres	-31,6%

TABLE 4.5 – Comparaison des quantiles 90%

la stratégie AXA. Cela provient directement du fait que la distribution des majorations optimales est nettement différente de l'ancienne distribution. Concernant les contrats fortement majorés après l'optimisation, ils sont bien moins rentables comparés à la situation initiale, mais ils résilient aussi beaucoup moins. Toutefois, ils présentent une sinistralité plus faible en terme de fréquence et de coût, ce qui laisse penser que des contrats fortement sinistrés ne se retrouvent pas tous dans ce quantile et de ce fait ont une majoration moindre, ce qui ne paraît pas économiquement rationnel.

Nous réalisons également des arbres de décision, avant et après optimisation, sur ces contrats, afin de voir ce qui les caractérise et ce qui les différencie. La variable réponse Y prend alors la valeur 1 si la majoration est supérieure au quantile 90%, 0 sinon.

Après l'optimisation, les majorations importantes concernent les ELR élevés, mais n'ont effectivement pas de lien avec la sinistralité observée sur ces contrats, à la différence de la stratégie actuelle d'AXA France.

En conclusion, nous pouvons dire que l'optimisation est orientée vers le futur, puisque qu'elle se base davantage sur l'ELR, qui est une estimation de la profitabilité, plus que sur la sinistralité survenue dans le passé. L'optimisation considère donc que les coûts de sinistres constituent une somme définitivement perdue mais compensée par la mutualisation des risques, et qu'il faut préférablement se référer à la sinistralité future estimée.

4.5.3 Critique de l'optimisation

Ainsi, le résultat d'une telle optimisation paraît au premier abord surprenant : il montre que globalement, en majorant moins fortement la majorité du portefeuille, nous diminuons le taux de résiliation, nous diminuons l'ELR et nous maintenons le chiffre d'affaires. Mais ce n'est pas le seul avantage : l'image de marque va bénéficier de telles majorations modérées voire négatives.

S'agit-il d'une solution miracle ? Il faut bien sûr émettre quelques réserves. Rappelons notamment la dimension politique prépondérante des agents généraux ou des courtiers. Il ne serait pas impossible qu'un geste commercial soit accordé pour un portefeuille peu rentable où les majorations optimales seraient importantes. Le cas échéant, il faudrait compenser cette sous optimalité imposée avec les autres majorations, ce qui nous éloignerait mécaniquement de la situation optimale pour l'ensemble des contrats MRH. Par ailleurs il faut aussi penser à l'implémentation opérationnelle de cette optimisation : telle quelle, l'optimisation est un processus trop lourd pour être réalisée chaque mois. Une solution, détaillée dans le chapitre 5, serait de réaliser un *reverse engineering* qui consisterait à prédire les majorations optimales pour chaque contrat compte tenu de leurs caractéristiques ainsi que de celles des assurés correspondants. Cela est plus simple à effectuer tous les mois, mais les majorations obtenues seront des estimations des majorations optimales, ce qui nous éloignerait un peu de la situation optimale : si les KPIs seront légèrement améliorés, ils ne seront peut-être pas aussi bons que ceux prévus.

Conclusion du chapitre

Notre objectif est de minimiser le taux moyen des résiliations sous contrainte de l'ELR, et nous avons montré que cela revenait, pour N contrats et sous des hypothèses fortes d'indépendance des fonctions de résiliations, des majorations et des fonctions de crédits commerciaux, à résoudre N problèmes individuels plus simples d'optimisation sous contrainte. Le multiplicateur de Lagrange λ, qui est commun à tous, permet de lier ces N problèmes d'optimisation au problème initial.

Nous nous sommes cependant heurtés à un autre sujet : les fonctions de résiliation n'étant pas convexes, résoudre le problème dual n'équivaut pas à résoudre le problème primal. Or l'idée du problème dual est d'obtenir le multiplicateur de Lagrange λ qui permet de trouver l'optimal global. Ainsi, il a fallu faire preuve d'imagination pour espérer trouver des majorations qui satisfassent nos exigences : en faisant varier la valeur de λ, nous avons obtenu différents jeux de majorations optimales et nous avons pu obtenir l'évolution des KPIs en fonction de λ. De cette manière, nous avons également pu établir concrètement la position stratégique d'AXA France par rapport à la frontière efficiente et confirmer que la stratégie de majoration actuelle peut être améliorée. Nous avons alors déterminé la valeur de λ qui satisfait le mieux le cahier des charges en termes de taux de résiliation moyen, d'ELR et de chiffre d'affaires, et donc par transitivité le jeu de majorations optimales qui répond à nos objectifs.

A cette occasion, nous avons pu constater que la distribution des majorations optimales est plus concentrée autour de sa moyenne, elle-même inférieure à celle des anciennes majorations. D'ailleurs, certaines majorations optimales sont négatives : si elles sont peu nombreuses, elles favorisent les jeunes clients rentables qui, nous l'avons vu, ont plus facilement tendance à résilier leur contrat MRH. Les majorations négatives contribuent ainsi à l'image de marque et à la fidélisation, sans pour autant dégrader le chiffre d'affaires ni la profitabilité du portefeuille.

Le travail d'optimisation a fourni des résultats intéressants, mais il reste laborieux et complexe : le réaliser tous les mois ne serait pas approprié. Ainsi, l'élaboration d'un reverse engineering permettrait d'automatiser par la suite les calculs de majoration au terme sur la base de notre étude d'optimisation.

Chapitre 5

Reverse engineering

Sommaire

Préambule

Nous avons vu lors du chapitre précédent comment obtenir les majorations optimales pour chaque contrat en fonction de leurs caractéristiques et de celles du client. Or, opérationnellement, cette étude reste lourde et chronophage à réaliser tous les mois. C'est pourquoi il est judicieux de réaliser un *reverse engineering* (Ingénierie inverse) : l'idée est de réaliser un modèle de prédiction des majorations optimales en fonction des variables explicatives dont les actuaires disposent au moment du calcul de la majoration, deux mois avant le terme du contrat. Il sera alors possible chaque mois, à partir de ce modèle, de calculer la majoration optimale de chaque contrat sans repasser par l'étape optimisation, ce qui se révélera être un gain de temps et d'efficacité non négligeable.

5.1 Une distribution de majorations optimales originale

Dans le cadre du reverse engineering, l'objectif est d'aboutir à un modèle de prédiction des majorations optimales interprétable afin que le calcul de ces dernières soit facile et automatique.

Ainsi, réaliser le reverse engineering par un GLM serait l'idéal car la majoration optimale serait une fonction explicite de certaines variables explicatives. De plus, l'Autorité de Contrôle Prudentiel et de Résolution (ACPR) exige que les modèles de tarification soient transparents et facilement interprétables afin de pouvoir être audités sans difficulté. Par ailleurs, il est impératif que les actuaires maîtrisent et comprennent les tarifs qu'ils proposent.

Or, un problème se pose, la distribution des majorations optimales telle que nous l'avons vue sur la figure 4.5, page 65, ne ressemble a aucune loi connue : la positivité du skewness, le kurtosis élevé ainsi qu'une queue de distribution à gauche extrêmement fine (du fait des quelques majorations négatives) nous éloignent des lois usuelles. Le graphique de Cullen et Frey en figure 5.1 confirme d'ailleurs qu'aucune loi usuelle ne serait une bonne candidate pour décrire cette distribution.

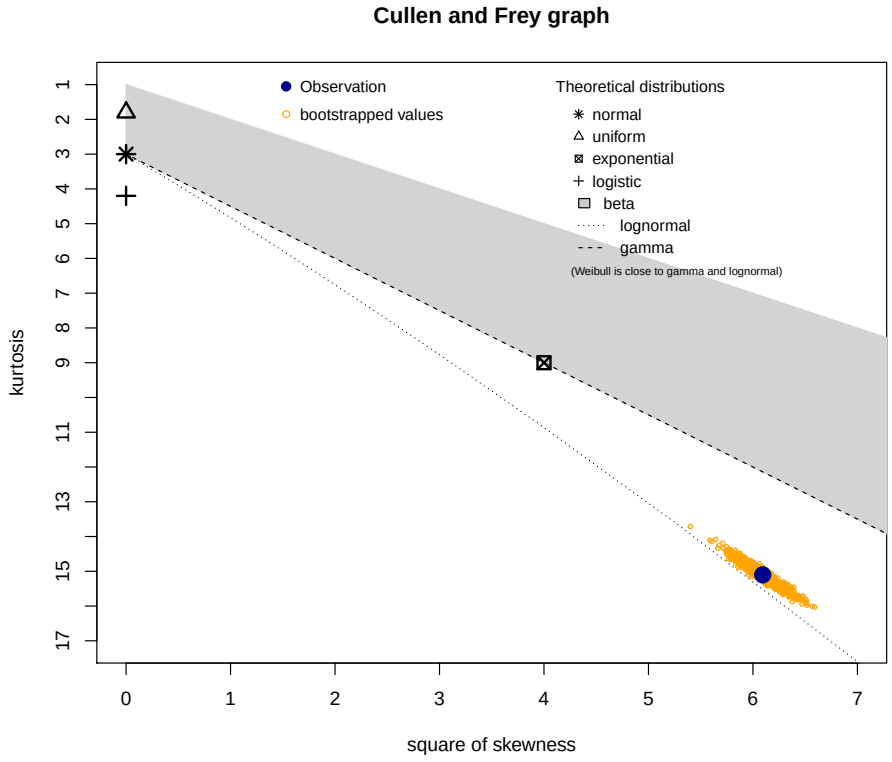


FIGURE 5.1 – Graphique de Cullen et Frey sur la distribution des majorations optimales

Ainsi, nous ne pourrons pas réaliser le reverse engineering par GLM puisque la distribution n'appartient pas à une famille exponentielle. En revanche, un GBM serait capable de faire une telle prédiction puisqu'il s'affranchit des hypothèses sur la forme de la distribution de la variable réponse. Toutefois, nous avons vus que le GBM est une "boîte noire" : le résultat issu d'une telle modélisation est difficilement interprétable et nous ne savons pas *a priori* quelles sont les variables qui ont contribué à l'augmentation de la valeur de la variable réponse, et dans quelle mesure.

Il existe cependant un moyen de contourner cette problématique : il est possible d'expliquer les résultats d'un GBM à l'aide de la théorie des *Shap Values*, que nous détaillerons plus tard dans ce chapitre, et qui va permettre de signaler les variables utilisées pour réaliser chaque prédiction ainsi que leur contribution à l'estimation de la variable réponse. Certes, cette méthode n'est pas aussi exhaustive par rapport à une fonction obtenue par GLM, néanmoins elle permet de justifier chaque majoration à partir des variables explicatives et vérifier que le valeur prédite est cohérente.

5.2 Réalisation du reverse engineering par GBM

5.2.1 Base de données utilisée

La base de données sur laquelle nous allons réaliser le reverse engineering par GBM est composée des contrats sur lesquels nous avons fait l'optimisation, car les majorations optimales dont nous disposons concerne ces contrats. La variable réponse Y est naturellement la majoration optimale associée à chaque contrat. Quant aux variables explicatives, nous retenons uniquement les variables *a priori*, c'est-à-dire les variables disponibles au moment du calcul de la majoration, deux mois avant le terme du contrat. De plus, parmi ces variables explicatives, nous décidons de garder celles qui sont actuellement utilisées par les actuaires pour le calcul de

la majoration. En effet, nous ne pouvons pas justifier au client l'augmentation de sa prime d'assurance si ce dernier possède une grande maison plutôt qu'une petite par exemple, car sa maison n'a pas changé depuis un an. Les caractéristiques du bien assuré doivent préférablement être prises en compte dans le tarif d'une affaire nouvelle, ce qui n'est pas le sujet de notre étude. En revanche, il n'est pas impossible que le risque associé à ce bien ait évolué : l'ELR est donc admissible en tant que variable explicative.

5.2.2 Modélisation GBM

Nous appliquons la même procédure que pour les précédents GBM réalisés dans ce mémoire, à une différence près : la variable réponse Y est une variable suivant une loi continue. Nous réalisons donc une régression plutôt qu'une classification, grâce à un boosting d'arbres de régression.

5.2.2.1 Gradient Boosting pour la régression

Toujours selon l'article de Friedman [5], la fonction de perte L utilisée par le package *XGBoost* sur Python pour un tel gradient boosting avec des arbres de régression est celle des moindres carrés :

$$L(Y, F) = \frac{(Y - F)^2}{2} \tag{5.1}$$

avec F une fonction associant les variables explicatives X_i à la variable réponse Y_i.
Nous définissons la pseudo-réponse $\widetilde{Y_i}$ comme le gradient négatif en X_i :

$$\widetilde{Y_i} = -\left[\frac{\partial L(Y_i, F(X_i))}{\partial F(X_i)}\right]_{F(X)=F_{m-1}(X)} = Y_i - F_{m-1}(X_i) \tag{5.2}$$

A partir de cette pseudo-réponse $\widetilde{Y_i}$, nous pouvons estimer, pour tout m, le vecteur a_m qui représente les paramètres du m-ième arbre de régression CART, $h(X_i, a_m)$, modélisant $\widetilde{Y_i}$.

L'algorithme débute en choisissant comme initialisation :

$$F_0(X) = \bar{Y}, \tag{5.3}$$

Puis, à chaque itération m, après avoir calculé la pseudo réponse, nous déterminons le facteur multiplicatif ρ_m qui va minimiser la fonction de perte :

$$\rho_m = \arg\ \min_{\rho} \sum_{i=1}^{n} L(Y_i, F_{m-1}(X_i) + \rho h(X_i, a_m)) \tag{5.4}$$

$$= \arg\ \min_{\rho} \sum_{i=1}^{n} \left[\widetilde{Y_i} - \rho h(X_i, a_m)\right]^2 \tag{5.5}$$

et finalement :

$$F_m(X_i) = F_{m-1}(X_i) + \rho_m h(X_i, a_m) \tag{5.6}$$

5.2.2.2 Choix optimal des paramètres du GBM par *Grid Search*

Nous réalisons le GBM sur Python, grâce au package "*xgboost*".

Tout comme dans la section 2.5.3 au chapitre 2, nous souhaitons déterminer les paramètres du GBM qui maximisent la qualité de la prédiction grâce à un *Grid Search*. Nous réalisons toujours une cross validation sur un *5-folds* sur chaque combinaison B_i de paramètres. Par exemple, B_1 présente la combinaison de paramètres

suivante : une présélection aléatoire d'un échantillon des variables à hauteur de 70% pour la construction des arbres de décision ; une réduction minimale de 30% de l'impureté pour qu'une nouvelle feuille de l'arbre soit créée ; et une profondeur maximale des arbres égale à 3 (c'est-à-dire au maximum 3 créations de noeuds). La valeur de ces paramètres change pour les autres combinaisons B_i.

Nous évaluons la performance des prédictions pour chaque combinaison selon deux critères :

— La **Mean Squared Error** (MSE) : c'est la moyenne des erreurs au carré, définie selon :

$$\text{MSE}(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

Plus petite sera la MSE, meilleure est la prédiction par le modèle.

— La **Variance expliquée** : indique à quel point le modèle explique la variance des observations. Elle est définie par :

$$\text{Variance Expliquée}(Y, \widehat{Y}) = 1 - \frac{\mathbb{V}ar(Y - \widehat{Y})}{\mathbb{V}ar(Y)}$$

La variance expliquée est comprise entre 0 et 1. Plus la variance expliquée est proche de 1, meilleure est la prédiction du modèle.

Nous obtenons donc en figure 5.2 les résultats du Grid Search selon la Mean Squared Error pour la prédiction des majorations optimales, et en figure 5.3 les résultats selon la variance expliquée.

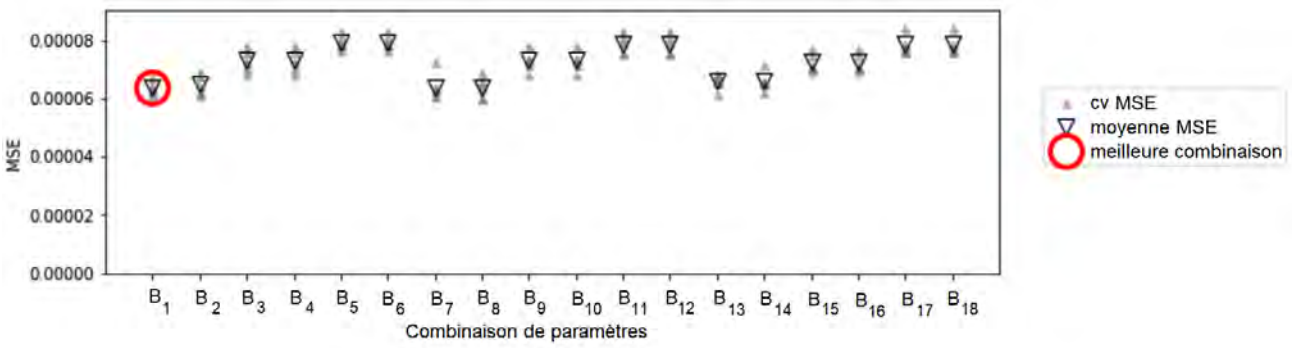


FIGURE 5.2 – Résultat sur Grid Search selon la MSE pour la prédiction des majorations optimales

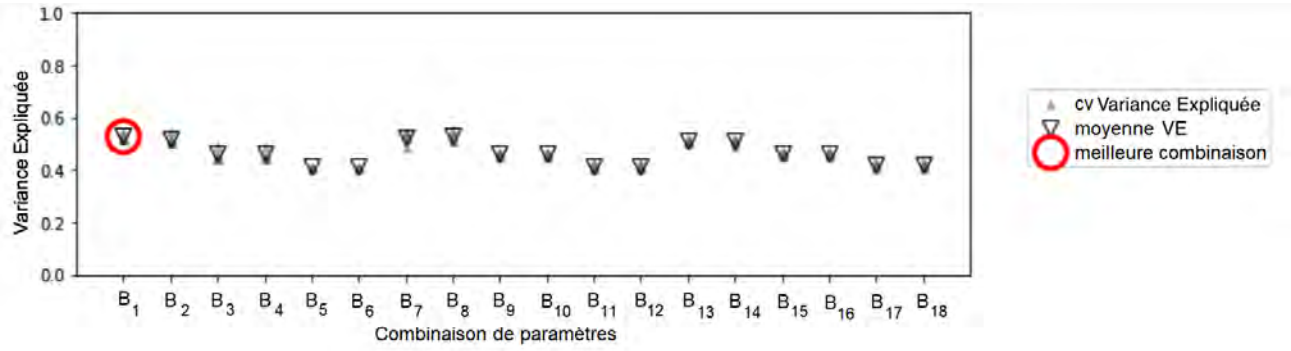


FIGURE 5.3 – Résultat sur Grid Search selon la variance expliquée pour la prédiction des majorations optimales

C'est la combinaison B_1 qui assure les meilleures performances de prédiction dans les deux cas : elle minimise la MSE, et elle maximise la variance expliquée.

Avec une telle combinaison, le modèle estime une majoration optimale moyenne à 3,847%, contre la majoration moyenne optimale observée de 3,848%, ce qui est très satisfaisant.

Nous considérons alors que nous avons trouvé un modèle suffisamment bon pour prédire les majorations optimales pour un contrat donné. Reste à présent à pouvoir l'interpréter afin de pouvoir l'exploiter en pratique et justifier les majorations obtenues. Pour ce faire, nous avons recours à la théorie des Shap Values.

5.3 Interprétation des résultats selon la théorie des Shap Values

Dans le paragraphe précédent, nous avons obtenu un modèle de gradient boosting f qui permet de prédire la majoration optimale Y à partir des variables explicatives observées X. Nous pouvons donc, pour tout X, estimer $\widehat{Y} = f(X)$. Or, le modèle f est difficile à interpréter : il n'existe pas de formule explicite qui indique l'importance des variables dans le calcul de la prédiction comme le ferait par exemple un GLM.

Nous connaissons en revanche la moyenne des prédictions : $\mathbb{E}[f(X)]$. A partir de cette moyenne, nous souhaitons savoir comment chaque variable observée de X contribue à se rapprocher de la prédiction $f(X)$: la figure 5.4 indique, pour un contrat donné, la distance entre la prédiction et la moyenne des prédictions. L'objectif est donc d'expliquer cet écart grâce aux variables explicatives observées.

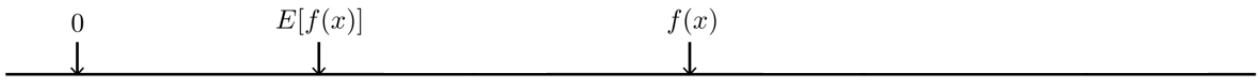


FIGURE 5.4 – Graphique indiquant l'éloignement de la moyenne des prédictions avec une prédiction donnée pour une observation X (Graphique inspiré d'une illustration de l'article de Lundberg et Lee : *Consistent Individualized Feature Attribution for Tree Ensembles* [6])

Prenons un exemple simple pour illustrer la démarche : supposons que nous ayons un contrat détenu par un souscripteur de moins de 25 ans, monodétenteur, n'ayant eu aucun sinistre mais son ELR est très élevé. Le modèle de prédiction indique une moyenne des prédictions des majorations optimales à 3,8%. Or, pour ce contrat, la majoration optimale prédite est par exemple $f(X) = 5\%$. Quelles sont les variables qui ont le plus joué dans l'augmentation de la majoration optimale par rapport à la majoration optimale moyenne ? Lundberg et Lee proposent une méthode répondant à cette problématique : il s'agit de la théorie des *SHAP values*.

5.3.1 Théorie des SHAP Values

Lundberg et Lee, dans leur article *A Unified Approach to interpreting Model Predictions* [7], proposent une méthode pour interpréter des modèles de prédiction complexes, comme le Gradient Boosting Machine, souvent considérés comme des "boîtes noires" car nous ne savons pas exactement quelles variables ont plus ou moins eu une influence sur l'estimation de la variable réponse. Lundgerg et Lee partent du constat que, bien qu'ils présentent de très bonnes performances de prédictions, les modèles de *deep learning* sont trop souvent complexes et ne sont donc pas assez transparents pour pouvoir être exploités dans certains cas concrets. C'est exactement ce que nous observons dans le domaine de l'assurance non-vie : nous sommes aujourd'hui capables de réaliser des modèles avec un grand pouvoir de prédiction, mais leur manque de transparence nous empêche de les utiliser en pratique pour élaborer nos tarifs. En effet, comment justifier auprès du client le prix de son assurance habitation ? L'assureur doit être en mesure de renseigner au client les facteurs qui ont eu un impact sur l'élaboration de son tarif.

Une telle censure de modèle est dommage à la fois pour l'assureur, qui est limité dans la modélisation de ses risques, mais aussi pour le souscripteur, qui de ce fait va peut-être payer une prime plus chère que celle qui

serait obtenue avec un modèle complexe, car le risque associé à son bien aura été surestimé.

L'idée proposée par Lundberg et Lee est d'expliquer le modèle "boîte noire" par un autre modèle plus transparent : il s'agit de la **méthode SHAP** (*SHapley Additive exPlanations*, ou l'explication incrémentale par les valeurs de Shapley [1]). La SHAP va attribuer, pour chaque prédiction $\widehat{Y}$, une importance à chaque variable explicative. Plus concrètement, pour chaque estimation, nous serons en mesure de connaître quelles seront les variables qui auront contribué à l'augmentation ou à la diminution de la valeur de $\widehat{Y}$, et dans quelle proportion par rapport à une valeur de base.

La méthode SHAP appartient à un groupe de méthodes appelées **"Méthodes de contributions incrémentales des variables caractéristiques"** (ou **"*Additive feature attribution methods*"**). Parmi ces méthodes, le LIME par exemple, qui va approximer localement le modèle que nous souhaitons expliquer par un autre modèle plus transparent. Cette méthode est détaillée dans l'article *A Unified Approach to interpreting Model Predictions* [7]. Mais la méthode SHAP est la seule de ces méthodes à définir de manière unique la contribution de chaque variable explicative à la valeur de la variable réponse.

5.3.1.1 Méthodes de contributions incrémentales des variables caractéristiques

Plus connues sous le nom de *Additive feature attribution methods*, ces méthodes vont expliquer **localement**, par un modèle simple, un modèle complexe et difficile à interpréter.

Soit f le modèle complexe, et nous notons $f(X) = \widehat{Y}$ la prédiction que nous souhaitons interpréter à partir des variables explicatives X par un modèle explicatif g.

Les modèles explicatifs utilisent souvent des variables explicatives simplifiées, obtenues à partir des variables explicatives d'origine et d'une *fonction de mappage* : pour tout X, nous pouvons déterminer une fonction de mappage h_X telle que $X = h_X(X')$. Notons que chaque fonction de mappage h_X est spécifique au vecteur X.

L'objectif est de trouver un modèle g qui explique localement, c'est-à-dire au voisinage de X, $f(X)$:

$$\forall z' \text{ tel que } z' \approx X' \text{ nous voulons } g(z') \approx f(h_X(z')) \tag{5.7}$$

Les **méthodes de contributions incrémentales des variables caractéristiques** répondent à cette exigence locale en proposant un modèle explicatif g qui est une fonction linéaire de variables binaires :

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{5.8}$$

$$\approx f(h_X(z')) \tag{5.9}$$

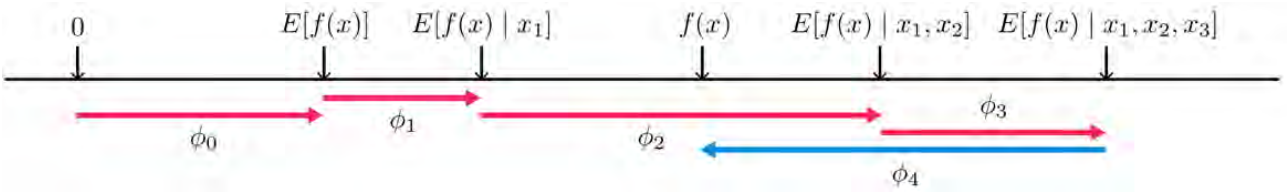
$$\approx f(X) \tag{5.10}$$

où $z' \in \{0,1\}^M$, et M le nombre de variables explicatives (ou caractéristiques) simplifiées.

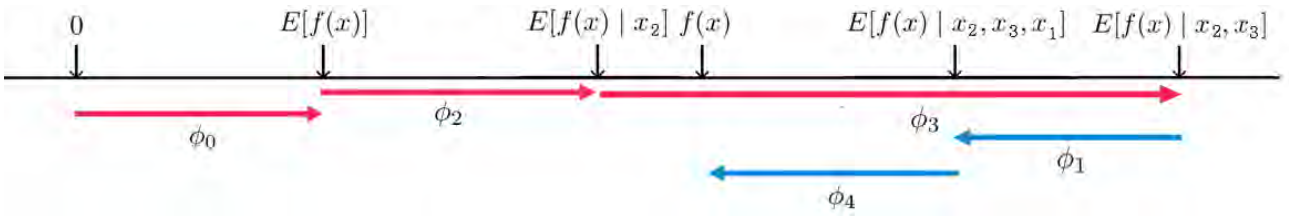
Un tel modèle assigne une contribution ϕ_i à chaque variable observée, et la somme de ces ϕ_i avec une valeur de base ϕ_0 approxime la prédiction de Y par f, c'est-à-dire $f(X)$. Notons que les ϕ_i ainsi obtenus sont propres à chaque prédiction.

1. Dans le cadre de la théorie des jeux, la valeur de Shapley, dont le nom est emprunté au mathématicien économiste Lloyd Shapley, donne une répartition des gains aux joueurs qui ont coopéré pour maximiser un gain global. Pour plus de détails, le lecteur peut se référer à l'essai original de Shapley : *A Value for N-person Games* [11].

Lundberg et Lee définissent ainsi chaque SHAP value ϕ_i : il s'agit de la variation de valeur de l'espérance conditionnelle de $f(X)$ lorsque nous rajoutons la condition par rapport à la variable X_i. La figure 5.5 (a) montre la manière dont sont obtenues les SHAP values pour un modèle f à 4 variables explicatives : en partant de l'espérance des prédictions, c'est-à-dire $\mathbb{E}[f(X)]$, nous obtenons la valeur de base ϕ_0. Puis, nous souhaitons voir à quel point la variable X_1 apporte de l'information sur la valeur de $f(X)$. Pour cela nous calculons la différence $\mathbb{E}[f(X)|X_1] - \mathbb{E}[f(X)] = \phi_1$. En ajoutant une à une les variables restantes, nous sommes capables d'obtenir des valeurs de ϕ_i.



(a) Obtention des SHAP values pour un ordre donné d'incrémentation des variables



(b) Obtention des SHAP values pour un ordre différent d'incrémentation des variables

FIGURE 5.5 – Exemple d'attribution des SHAP values pour un modèle complexe f à 4 variables (Inspiré du graphique qui figure dans l'article *Consistent individualized feature attribution for tree ensembles* [6])

Sur la figure 5.5 (a) nous remarquons ainsi que les SHAP values ϕ_1, ϕ_2 et ϕ_3 contribuent à augmenter la valeur de la prédiction par rapport à $\mathbb{E}[f(X)]$, tandis que ϕ_4 permet de diminuer cette somme de contributions afin de retrouver la valeur de la prédiction $f(X)$.

Or, les variables X_i ne sont pas forcément indépendantes : les corrélations entre les variables ou les effets d'interaction peuvent influer sur la valeur des ϕ_i suivant l'ordre d'ajout des variables X_i. Par exemple, la figure 5.5 (b) présente les SHAP values expliquant la même prédiction $f(X)$ qu'en figure 5.5 (a), mais avec un ordre d'ajout de variables différent, ce qui conduit à des valeurs de ϕ_i différentes. En effet, si nous supposons, pour l'exemple, que les variables X_1 et X_3 interagissent : X_1 sans X_3 contribue à augmenter la valeur de la prédiction (figure (a)), et X_3 sans X_1 contribue encore plus à augmenter la valeur de prédiction (figure (b)). Mais lorsque X_1 et X_3 interviennent toutes les deux, l'effet d'interaction vient diminuer la somme de ces deux contributions prises seules. C'est pourquoi, suivant l'ordre d'ajout de ces variables, les SHAP values ϕ_i peuvent changer afin de respecter ces effets d'interaction : si X_3 intervient après X_1 (figure (a)), alors ϕ_3 sera positif mais moins important que sur la figure (b) ; si X_1 intervient après X_3, comme ϕ_3 est trop important, alors ϕ_1 sera négatif pour diminuer la contribution de ces deux variables.

Ainsi, l'ordre d'ajout des variables est important car il conduit à des SHAP values différentes. L'idée alors proposée par Lundberg et Lee est de considérer tous les ordres possibles d'ajouts de variables et de moyenner sur chaque SHAP value ϕ_i obtenue. Pour ce faire, ils s'inspirent des travaux de Lloyd Shapley pour le calcul des valeurs de Shapley.

5.3.1.2 Le calcul des valeurs de Shapley selon la théorie des jeux

Nous pouvons estimer les coefficients ϕ_i en les calculant comme des valeurs de Shapley : une telle méthode de calcul permettra de prendre en compte l'ensemble des ordres possibles d'ajouts de variables lors du calcul des SHAP values.

Les travaux de Lloyd Shapley s'incrivent dans la théorie des jeux. Dans son article *A Value for N-person Games* [11], L. Shapley explique notamment, lorsque N joueurs gagnent à un jeu d'argent en collaborant, comment répartir de la manière la plus juste les gains obtenus. En effet, certains joueurs ont contribué plus de d'autres à la victoire, et méritent ainsi une plus grande part du gain.

La répartition du gain la plus juste est obtenue en considérant les gains qui auraient été obtenus pour le même jeu suivant les différentes combinaisons de joueurs possibles parmi les N joueurs.

L'analogie avec les SHAP values commence à apparaître. Nous rappelons que notre objectif est d'expliquer la valeur de la prédiction $f(X)$, qui n'est autre que le gain dans l'article de Shapley, par les variables explicatives X_i, qui sont les joueurs selon Shapley.

Pour plus de renseignements sur les valeurs de Shapley dans le cadre de la théorie des jeux, le lecteur est invité à lire *A Value for N-person Games* [11].

Pour obtenir la valeur de Shapley ϕ_i qui donne l'importance de la contribution de la variable i au calcul de la prédiction, l'idée est d'observer l'impact de l'absence de cette variable sur la prédiction de Y. Plus simplement, nous partons du principe que l'importance de la variable i est caractérisée par la différence de prédiction pour une observation donnée entre le modèle prenant en compte cette variable, et le modèle ne la prenant pas en compte :

$$\text{Importance de } i = f_{\text{avec i}} - f_{\text{sans i}} \tag{5.11}$$

Or l'effet du retrait de la variable i sur la prédiction dépend également des autres variables. En effet, il se peut par exemple que cette variable i soit corrélée avec une autre, et dans ce cas il est vraisemblable que l'absence de l'une engendre une augmentation (en valeur absolue) de la contribution de l'autre. C'est pourquoi il faut considérer les différents sous-ensembles possibles $S \subseteq F \setminus \{i\}$, avec F l'ensemble de toutes les variables explicatives, et entraîner à nouveau le modèle f, sur l'ensemble de la base de données restreinte aux variables incluses dans S.

Pour chaque sous-ensemble S, il sera alors possible de calculer la différence de prédiction pour une observation donnée :

$$f_{S\cup\{i\}}(X_{S\cup\{i\}}) - f_S(X_S)$$

avec X_S le vecteur des valeurs observées des variables présentes dans S.

Il est ensuite nécessaire de pondérer cette différence par le nombre de combinaisons possibles des sous-ensembles S parmi les variables qui composent $F \setminus \{i\}$.

ϕ_i est alors obtenue par la somme pondérée de ces différences de prédiction, pour une observation donnée, sur tous les sous-ensembles de variables S possibles :

$$\phi_i(f, X) = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}(X_{S\cup\{i\}}) - f_S(X_S)\right] \tag{5.12}$$

avec :

- f_S le modèle complexe entraîné sur les variables appartenant à S ;

- $f_{S\cup\{i\}}$ le modèle complexe entraîné sur les variables appartenant à S et la variable i

La pondération $\frac{|S|!(|F|-|S|-1)!}{|F|!}$ peut être expliquée de la façon suivante :

- Il y a $|F|$ tailles différentes de S, allant de 0 à $|F|-1$.
- Le même poids est attribué à chaque taille possible de sous-ensemble S, c'est-à-dire un poids égal à $\frac{1}{|F|}$.
- De plus, il y a $\begin{pmatrix} |F|-1 \\ |S| \end{pmatrix} = \frac{(|F|-1)!}{|S|!(|F|-|S|-1)!}$ sous-ensembles possibles de taille $|S|$ parmi les variables de $F \setminus \{i\}$.
- Il faut donc pondérer les différences de prédiction pour chaque S par $\frac{1}{|F|}\frac{1}{\begin{pmatrix} |F|-1 \\ |S| \end{pmatrix}}$, ce qui est notamment égal à $\frac{|S|!(|F|-|S|-1)!}{|F|!}$.

Enfin, il est possible de retrouver l'équation (5.8) si nous posons $\phi_0 = f_{\varnothing}(\{\varnothing\})$, c'est-à-dire le modèle trivial qui ne contient aucune variable explicative.

5.3.1.3 Obtention des SHAP values à partir des valeurs de Shapley et d'un modèle de prédiction

Dans la pratique, comment obtenons-nous les valeurs $f_S(X_S)$ (telles que décrites dans l'équation (5.12)) à partir du modèle f ?

Rappelons que X_S est le vecteur des valeurs observées des variables présentes dans S. La taille de X_S est inférieure ou égale à la taille du vecteur X.

Définissons à présent Z' le **vecteur des variables explicatives simplifiées** que nous utiliserons en input dans l'équation (5.12) : Z' est un vecteur aux composantes binaires de même taille que le vecteur des variables explicatives X, tel que :

$$\forall i, \quad Z'_i = \begin{cases} 1 & \text{si } X_i \in S \\ 0 & \text{sinon.} \end{cases} \tag{5.13}$$

Nous définissons alors une fonction de mappage appropriée h_X tel que :

$$h_X(Z') = Z_S, \quad \text{avec } Z_S \text{ de même taille que } X \quad \text{t.q. } \forall i,\ Z_{S,i} = \begin{cases} X_i & \text{si } X_i \in S \\ \text{manquant} & \text{sinon.} \end{cases} \tag{5.14}$$

Nous avons donc :

$$f_S(X_S) = f_S(Z') = f(h_X(Z')) = f(Z_S) \tag{5.15}$$

Une telle fonction de mappage permet de faire le lien entre f_S et f, et de s'assurer que, si une variable est absente de la base de données sur laquelle le modèle est entraîné, alors aucune contribution ne lui sera attribuée dans le calcul de la prédiction, conformément à l'équation (5.8).

Or, la plupart des modèles de prédictions (comme les arbres de décision ou les Gradient Boostings) ne peuvent pas gérer l'absence de variables explicatives. En effet, prenons l'exemple d'un arbre de décision que nous souhaitons interpréter avec les SHAP values et supposons que cet arbre de décision sollicite dans ses noeuds 4 variables explicatives X_1, X_2, X_3 et X_4 pour prédire $f(X)$. Si nous retirons une de ces variables explicatives,

l'arbre de décision ne sera pas capable de prendre une décision au niveau du noeud qui nécessite l'observation de cette variable manquante. Dans ce cas de figure, il devient alors impossible d'obtenir une valeur de $f(Z_S)$. Il faut donc approximer cette prédiction par son espérance conditionnellement aux variables explicatives que nous connaissons, c'est-à-dire aux variables contenues dans S :

$$f_S(Z') = f(h_X(Z')) = f(Z_S) \approx \mathbb{E}[f(Z)|Z_S] \tag{5.16}$$

Une façon de calculer cette espérance est d'intégrer $f(Z_S)$ sur l'ensemble des variables n'appartenant pas à S. Remarquons qu'il s'agit d'un calcul très lourd, surtout si nous devons considérer tous les sous-ensembles S possibles. L'article *A Unified Approach to interpreting Model Predictions* [7] propose des méthodes pour rendre ce calcul plus efficace et plus rapide, comme l'échantillonnage des valeurs de Shapley, ou encore le noyau SHAP (Kernel SHAP).

Si nous revenons à notre exemple d'arbre de décision, calculer cette espérance $\mathbb{E}[f(Z)|Z_S]$ revient à calculer la moyenne des prédictions sur les branches issues du noeud sollicitant la variable manquante, sachant les observations des autres variables disponibles, c'est-à-dire appartenant à S.

Ainsi, l'équation (5.12) devient :

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[\mathbb{E}[f(Z)|Z_{S\cup\{i\}}] - \mathbb{E}[f(Z)|Z_S]\right] \tag{5.17}$$

Si nous revenons à notre exemple à 4 variables X_1, X_2, X_3 et X_4 et que nous souhaitons déterminer ϕ_1, nous avons donc $2^{|F\setminus\{i\}|} = 2^3$ ordres différents d'apparition de variables, auxquels correspond un certain rang d'apparition de la variable X_1 dans le calcul des SHAP values, comme le présente le tableau 5.1 :

	X_1 apparaît en 1er	X_1 apparaît en 2ème	X_1 apparaît en 3ème	X_1 apparaît en dernier
S	$\{\varnothing\}$	$\{X_2\}$	$\{X_2,\ X_3\}$	$\{X_2,\ X_3,\ X_4\}$
			$\{X_2,\ X_4\}$	$\{X_2,\ X_4,\ X_3\}$
		$\{X_3\}$	$\{X_3,\ X_2\}$	$\{X_3,\ X_2,\ X_4\}$
			$\{X_3,\ X_4\}$	$\{X_3,\ X_4,\ X_2\}$
		$\{X_4\}$	$\{X_4,\ X_2\}$	$\{X_4,\ X_2,\ X_3\}$
			$\{X_4,\ X_3\}$	$\{X_4,\ X_3,\ X_2\}$
$\lvert S\rvert$	0	1	2	3

TABLE 5.1 – Les différents sous-ensembles S possibles pour le calcul de ϕ_1 dans un modèle à 4 variables

Notons que ce tableau rend compte de différents ordres d'apparition des variables X_i, et certaines combinaisons se répètent (par exemple $\{X_2,\ X_3\}$ et $\{X_3,\ X_2\}$). Or l'espérance conditionnelle ne prend pas en compte l'ordre de ces variables, et donc nous nous affranchissons de l'ordre dans les sous-ensembles S que nous considérons.

Ainsi, dans notre exemple, il n'existe qu'une seule combinaison possible de sous-ensembles de taille 3 : $\{X_2,\ X_3,\ X_4\}$. Ce seul sous-ensemble S se verra attribuer un poids de $\frac{1}{|F|} = 0,25$. Tandis qu'il existe 3 combinaisons pour les sous-ensembles de taille 2 : $\{X_2,\ X_3\}$, $\{X_3,\ X_4\}$, et $\{X_2,\ X_4\}$, donc chacun de ces 3 sous-ensembles S se verra attribuer un poids de $\frac{1}{|F|*3} = \frac{0.25}{3}$.

5.3.1.4 Propriétés des SHAP Values

En plus de leur capacité à justifier de manière intuitive les sorties d'un modèle complexe, les SHAP values présentent des propriétés qui rendent légitime l'usage de telles valeurs pour expliquer un modèle quel qu'il soit : il s'agit de la propriété de précision locale, de la propriété des variables manquantes et de la propriété de cohérence.

Propriété 1 : Précision locale

L'approximation locale du modèle complexe f par un modèle g au voisinage d'une observation X donnée nécessite au moins l'égalité de la sortie du modèle f pour le vecteur X avec la sortie du modèle g pour le vecteur de variables simplifiées X' correspondant à X ($X = h_X(X')$) :

$$f(X) = g(X') = \phi_0 + \sum_{i=1}^{M} \phi_i X'_i \tag{5.18}$$

avec $\phi_0 = f(h_X(\mathbf{0}))$ qui représente la sortie du modèle f sans variable explicative.

Propriété 2 : Variables manquantes

Cette propriété met en avant l'intérêt de la fonction de mappage h_x : si le vecteur X' de variables simplifiées indique la présence ou non de variables explicatives, alors, compte tenu de l'équation (5.8), les variables absentes n'ont pas d'impact sur l'explication de $f(X)$:

$$\forall i, \quad X'_i = 0 \Rightarrow \phi_i = 0 \tag{5.19}$$

Propriété 3 : Cohérence

La cohérence reflète le fait que si le modèle f change en un modèle f' de telle sorte que la contribution d'une variable simplifiée Z'_i augmente ou reste la même, alors sa SHAP value correspondante ϕ_i ne doit pas diminuer.

Soit $f_S(Z') = f(h_X(Z'))$, et $Z' \setminus i$ indique que $Z'_i = 0$. Alors pour tous modèles f et f', si

$$\forall Z' \in \{0,1\}^M \quad f'_S(Z') - f'_S(Z' \setminus i) \geq f_S(Z') - f_S(Z' \setminus i) \tag{5.20}$$

alors

$$\phi_i(f', X) \geq \phi_i(f, X) \tag{5.21}$$

5.3.1.5 Calcul des SHAP values pour les modèles d'arbres

Selon les modèles f à expliquer, le calcul des SHAP values peut être plus ou moins complexe. Dans le cadre des modèles utilisant des arbres de décisions (comme de gradient boosting dans notre cas) l'estimation de $\mathbb{E}[f(Z)|Z_S]$ n'est pas évidente et souvent très coûteuse en temps machine. Une manière triviale de réaliser ce calcul serait de faire une récursion pour chaque variable mais cela prendrait un temps trop important qui augmenterait de manière exponentielle si le nombre de variables augmente.

Dans leur article complémentaire *Consistent individualized feature attribution for tree ensembles* [6], Lundberg et Lee proposent un algorithme plus efficace et rapide pour déterminer les SHAP values. Nous ne nous attarderons pas sur l'architecture de cet algorithme détaillé dans l'article en question. Toutefois, nous l'utilisons dans notre étude par le biais du package "*shap*".

5.3.1.6 Compréhension du comportement global du modèle grâce aux SHAP values

Pour un modèle, les SHAP values sont calculées pour une observation donnée. Il est toutefois possible d'avoir une idée du comportement global du modèle en observant, pour chaque variable, certaines statistiques sur l'ensemble des SHAP values obtenues. Nous détaillerons davantage ce point en pratique, lors de l'interprétation par les SHAP values des résultats du GBM pour la prédiction de la majoration optimale.

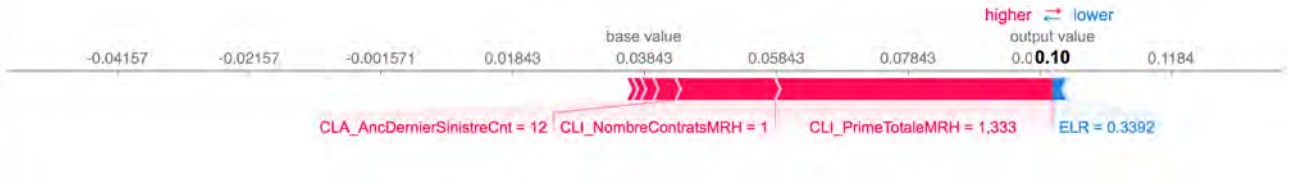
5.3.2 Interprétation des résultats du GBM avec les SHAP values

Nous avons vu, en section 5.2.2, que nous étions capables de prédire la majoration optimale à appliquer à chaque contrat à partir de leurs caractéristiques. Nous pouvons donc à présent obtenir les SHAP values pour n'importe quelle prédiction que nous avons obtenue.

5.3.2.1 Interprétation des prédictions individuelles

Nous nous intéressons par exemple à deux prédictions distinctes, correspondant à deux contrats différents : l'un reçoit une majoration optimale de 10%, l'autre une majoration optimale de 2%. La figure 5.6 représente pour chacun de ces contrats un *diagramme de forces* correspondant.

Un *diagramme de forces* illustre la contribution des variables explicatives à la valeur de prédite grâce aux SHAP values par rapport à une valeur de base. Les flèches pointant vers la droite correspondent aux variables qui participent à l'augmentation de la prédiction et la longueur de ces flèches correspond à la valeur absolue de la SHAP value associée. Tandis que les flèches qui pointent vers la gauche sont responsable de la diminution de la prédiction, et de même leur longueur correspond à la valeur absolue de la SHAP value. Un tel diagramme permet en un coup d'oeil de savoir quels facteurs ont le plus participé à l'éloignement de la prédiction par rapport à la valeur moyenne des prédictions.



(a) Diagramme de forces des SHAP values pour une majoration optimale prédite de 10%

(b) Diagramme de forces des SHAP values pour une majoration optimale prédite de 2%

FIGURE 5.6 – Diagrammes de forces représentant les SHAP values des prédictions de majorations optimales pour deux contrats différents

Le diagramme 5.6 (a) correspond à la prédiction de 10%, tandis que le (b) correspond à celle de 2%. Ces prédictions sont repérées sur un axe de valeurs croissantes, là où les flèches pointant à droite et celles pointant à gauche se rencontrent.

Le premier point remarquable est le fait que la valeur de base est la même pour ces deux prédictions. Cela est normal, puisque cette valeur est égale à la moyenne de l'ensemble des prédictions que nous étudions : 3,8%. Ces diagrammes indiquent également les noms des variables qui expliquent le mieux la valeur de la prédiction.

Dans le cas (a), c'est la prime totale payée par le client en contrats MRH (1333) qui fait augmenter de 4,2% la majoration optimale, soit plus de la moitié de la différence entre la prédiction de 10% et la moyenne à 3,8%. Ensuite, le fait que ce client détienne un seul contrat MRH augmente encore la prédiction de 1,5%. Par ailleurs, ce client a déjà été sinistré, mais il y a de cela 12 ans : cela implique une petite augmentation de la majoration de 0,3%. L'ELR quant à lui est plutôt bon (0,339), ce qui vient diminuer de 0,2% la valeur de la majoration optimale.

En ce qui concerne le cas (b), le fait que ce client détienne seulement un contrat MRH augmente la valeur

de la prédiction de 0,2%. Notons déjà que, comparé au cas précédent, la mono-détention en MRH n'a pas le même impact : ici l'augmentation de la majoration optimale est plus modérée. Ensuite, la prime totale payée par le client en contrats MRH (246) vient diminuer de 1,3% la valeur de la majoration optimale. L'ELR quant à lui (égal à 0,335) diminue également la prédiction mais à hauteur de 0,2%. Enfin, ce contrat a une ancienneté de 2 ans, ce qui implique à nouveau une diminution de la prédiction de 0,2% également.

Dans les deux cas, il y a encore d'autres variables qui rentrent en jeu dans l'évolution de la prédiction, mais elles ont un petit impact par rapport à celles que nous venons de citer.

A ce stade, il est légitime de se poser la question suivante : pourquoi imposer une majoration optimale plus importante à un contrat dont la prime est déjà très élevée par rapport à la moyenne, plutôt qu'à un contrat dont la prime est faible ?

Rappelons que le modèle que nous étudions prédit la majoration optimale pour chaque contrat, c'est-à-dire la majoration qui minimise le taux de résiliation du portefeuille sous contrainte de rentabilité (ELR) et de chiffre d'affaires. Or, pour arriver à cette optimum, nous avons élaboré un modèle de résiliation par GLM (cf. chapitre 2), et à cette occasion nous avons pu repérer les variables les plus significatives dans la prédiction de la probabilité de résiliation au terme. Il s'avère que la majoration TTC ainsi que l'ELR ressortent en premier parmi ces variables. En revanche, ni le prix du contrat MRH, ni même la prime totale des contrats MRH détenus par le client ne ressortent. C'est pour cette raison que l'optimisation joue notamment sur cette variable pour satisfaire les contraintes d'ELR et de chiffre d'affaires puisqu'elle a un impact limité sur la probabilité de résiliation du contrat.

Il est possible de voir les diagrammes de forces pour toutes les prédictions, afin d'observer éventuellement le comportement du modèle sur certains segments de contrats. Il suffit de concaténer tous les diagrammes de forces individuels pris verticalement.

La figure 5.7 représente par exemple les diagrammes de force d'un échantillon de 500 prédictions classées de deux manières : la première, en (a), classe les prédictions par majorations optimales décroissantes ; la seconde, en (b), les classe par ELR croissant. C'est d'ailleurs cette figure 5.7 (b) qui est intéressante car elle montre une cohérence dans notre étude. En effet, plus l'ELR associé à un contrat est grand, c'est-à-dire moins le contrat est rentable, plus la majoration optimale aura tendance à être grande pour satisfaire notre contrainte globale d'ELR sur notre portefeuille.

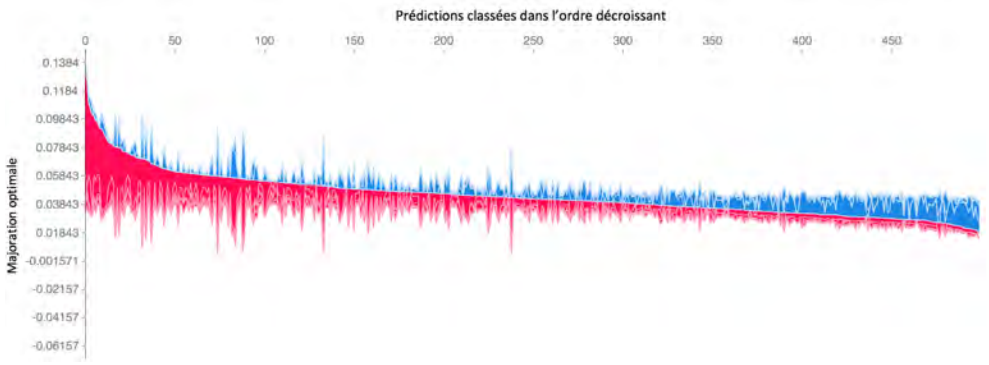
5.3.2.2 Interprétation globale du modèle

Si nous sommes capables d'expliquer les prédictions une à une, il n'y a pas cependant de formule générique qui relie les valeurs observées des variables aux prédictions comme le ferait un GLM. Toutefois, les SHAP values obtenues sur l'ensemble des prédictions permettent de nous donner une idée générale du comportement du modèle et d'observer certaines interactions.

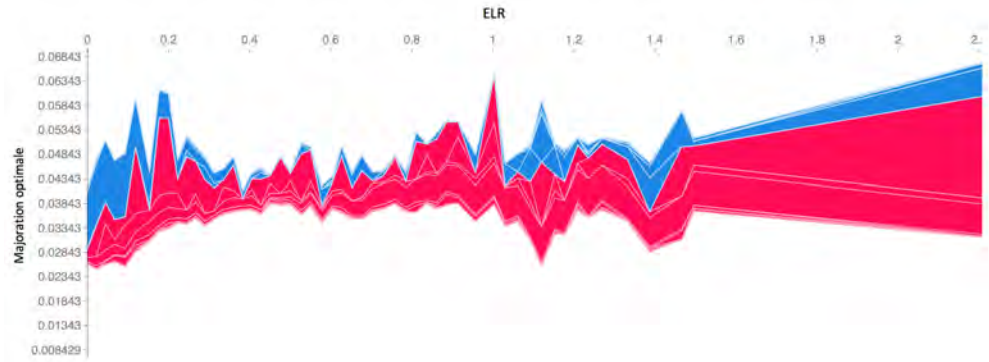
1ère observation possible : l'importance des variables

En règle générale, dans le modèle que nous considérons, quelles sont les variables qui ont le plus d'impact sur les prédictions ?

Une manière de répondre à cette question est de réaliser la moyenne des valeurs absolues des SHAP values pour chaque variable. Soit N le nombre de prédictions, et $\phi_i^{(j)}$ indiquant la SHAP values de la i-ième variable pour la j-ième prédiction :



(a) Diagramme de forces pour les prédictions de majorations optimales classées dans l'ordre décroissant



(b) Diagramme de forces pour les mêmes prédictions classées par ELR croissants

FIGURE 5.7 – Diagrammes de forces pour un échantillon de 500 prédictions

$$\text{Importance variable } i = \frac{1}{N}\sum_{j=1}^{N} |\phi_i^{(j)}| \tag{5.22}$$

Pour l'ensemble des variables considérées, il est alors possible de tracer un diagramme de barres pour repérer d'un seul coup d'oeil les variables les plus importantes. Bien que similaire au graphique de l'importance des variables qui peut être fourni à l'issu d'un GBM, le diagramme de barres des SHAP values est différent puisque le calcul de l'importance des variables est différent [2].

La figure 5.8 nous présente ainsi l'importance des variables au sens des SHAP values, autrement-dit l'impact moyen des variables sur la détermination de la valeur de la prédiction de la majoration optimale.

Cette figure nous indique notamment que notre modèle de prédiction assigne une importance prépondérante à la prime totale MRH que paie le client, au nombre total de contrats MRH qu'il détient, à son ELR et à l'ancienneté du dernier sinistre associé à ce contrat. Cela semble cohérent car ces variables permettent notamment de répondre aux contraintes de notre problème d'optimisation sans risquer d'augmenter le taux de résiliations au terme.

2ème observation possible : les éventuels effets d'interaction entre les variables

Il est possible de compléter cette première observation par l'étude d'éventuels effets d'interaction entre les variables. Cela peut être réalisé au moyen de graphiques des dépendances des SHAP values.

2. Dans le cas des graphiques d'importance des variables traditionnels, l'importance des variables est calculée au sens de l'erreur quadratique : une variables est d'autant plus importante qu'elle réduit l'impureté dans les noeuds des arbres de décision.

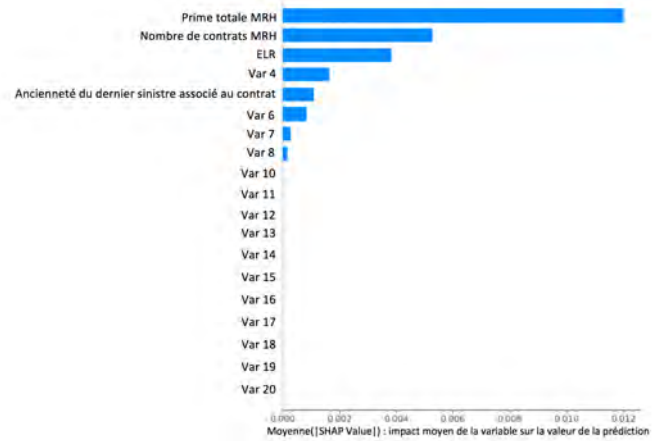


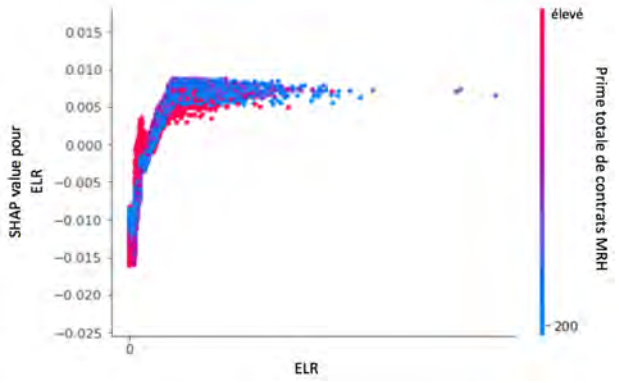
FIGURE 5.8 – Importance des variables selon le modèle de prédiction au sens des SHAP values

De tels graphiques indiquent en ordonnées, pour une variable donnée, l'ensemble des SHAP values correspondantes pour une valeur observée de cette variable (en abscisses). De cette manière, il devient possible d'observer la variation de l'importance de la contribution de cette variable en fonction de sa valeur observée. De plus, la dispersion verticale des points pour une valeur donnée en abscisse indique la présence d'interactions avec d'autres variables. Ces effets sont rendus visibles grâce à l'échelle de couleur bleue-rouge associant à chaque point la valeur d'une autre variable qui présente une interaction importante avec la première variable.

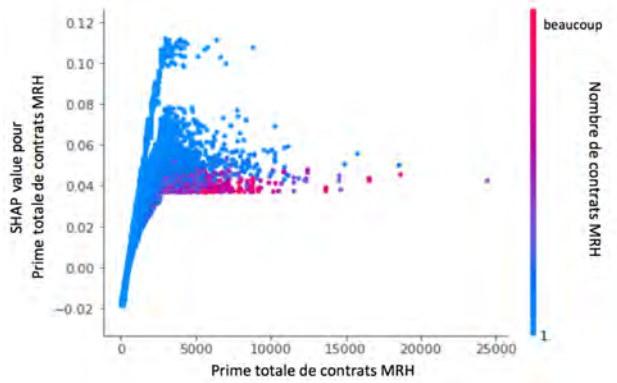
La figure 5.9 représente les graphiques des dépendances des SHAP values pour les 4 variables importantes que nous avons repérées précédemment en figure 5.8.

- **En figure 5.9 (a) :** Les SHAP values de l'ELR augmentent quasi-linéairement lorsque ce dernier augmente, jusqu'à ce qu'il atteigne la valeur 1. Pour les ELR supérieurs à 1 nous constatons davantage de dispersion verticale et donc d'interactions, qui ne sont pas seulement liées à la prime totale des contrats MRH des clients, puisque nous ne constatons pas de tendance claire dans les couleurs des points.
- **En figure 5.9 (b) :** Les SHAP values de la prime totale MRH des clients augmentent linéairement avec l'augmentation de cette dernière pour les petites valeurs. Puis nous constatons de l'interaction avec le nombre total de contrats MRH. Compte tenu des couleurs des points, nous remarquons qu'un nombre important de contrats MRH limite l'augmentation des SHAP values en question. Cela est cohérent avec notre volonté de protéger les multi-détenteurs MRH d'une trop grande majoration.
- **En figure 5.9 (c) :** Les SHAP values associées au nombre total de contrats MRH sont maximales pour les mono-détenteurs. De plus, Ces SHAP values sont quasi nulles lorsque la prime totale MRH est petite. Nous pouvons donc clairement en déduire qu'il existe une interaction entre le nombre total de contrats MRH et la prime totale MRH.
- **Enfin, en figure 5.9 (d) :** Les SHAP values correspondant à l'ancienneté du dernier sinistre indiquent que si un contrat n'a jamais connu de sinistre, alors la SHAP value sera négative pour venir réduire la majoration. Il est également remarquable que, quelque soit l'ancienneté du dernier sinistre, à partir du moment où il y en a eu un, une SHAP value positive qui ne dépend justement pas de l'ancienneté du sinistre mais de la prime totale MRH : plus cette dernière est importante, plus la SHAP value associée à l'ancienneté de sinistre sera grande.

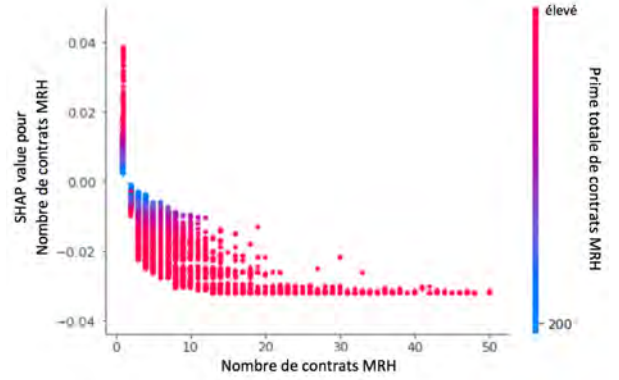
Ces graphiques de dépendances des SHAP values nous permettent d'avoir conscience de ces interactions, et d'ajuster notre modèle de prédiction en y ajoutant les termes d'interaction les plus significatifs.



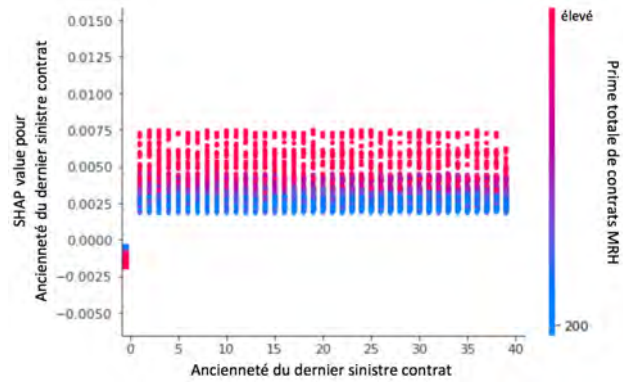
(a) SHAP Values de l'ELR en fonction de lui-même versus la Prime totale MRH



(b) SHAP values de la Prime totale MRH en fonction d'elle-même versus le nombre de contrats MRH



(c) SHAP values du Nombre de contrats MRH en fonction de lui-même versus la Prime totale MRH



(d) SHAP values de l'Ancienneté du dernier sinistre en fonction d'elle-même versus la Prime totale MRH

FIGURE 5.9 – Graphiques des dépendances des SHAP values

Conclusion du chapitre

L'objectif de ce chapitre était de montrer qu'il est possible de prédire les majorations optimales en fonction des caractéristiques de chaque contrat sans repasser par l'étude d'optimisation, trop lourde et trop chronophage pour pouvoir être utilisée tous les mois au niveau opérationnel. Pour ce faire, nous avons réalisé un reverse engineering sur les résultats de l'optimisation : l'idée était de construire un modèle de prédiction des majorations optimales à partir des caractéristiques des contrats actuellement utilisées pour réaliser le calcul de la majoration au terme.

Nous avons été capables de réaliser ce modèle au moyen d'un gradient boosting. Une telle méthode de machine learning s'affranchit de la loi de distribution de la variable réponse, ce qui nous satisfait car nous n'avons pas pu rapprocher la distribution des majorations optimales à une loi usuelle connue, nous empêchant donc d'avoir recours à un GLM. Une recherche sur grille a été réalisée afin de déterminer les meilleurs paramètres du GBM qui minimisent l'erreur quadratique et maximise la variance expliquée par le modèle. Le meilleur modèle GBM nous a alors fourni une précision de prédiction satisfaisante.

L'inconvénient du GBM est que, de prime abord, un tel modèle n'est pas interprétable aussi facilement qu'un GLM, car il ne présente pas de formule explicite conduisant à l'obtention de la prédiction en fonction des valeurs observées des variables explicatives. Nous avons donc proposé une alternative qui nous permet de lever le voile sur cette "boîte noire", une méthode qui vient interpréter chaque prédiction fournie par le modèle GBM complexe : il s'agit de la méthode des SHAP values. Cette méthode, qui tire son origine de la théorie des jeux en économie, attribue une contribution à chaque variable pour la prédiction de la majoration optimale. De cette manière, il est possible de connaître quelle variable a le plus contribué à augmenter ou diminuer la majoration

optimale.

Les SHAP values vont ainsi permettre aux actuaires de comprendre pourquoi une telle majoration optimale est appliquée, mais aussi de justifier ces niveaux de majorations auprès des autorités de contrôle et des clients.

Il est vrai que cette nouvelle approche ne se rapproche en rien des usages actuels en actuariat. Toutefois ce n'est pas parce que cette alternative n'a jamais été appliquée auparavant dans le monde de l'assurance non vie qu'il ne faut pas l'utiliser. Cette méthode mérite notre attention et peut trouver une application opérationnelle dans la détermination de la majoration au terme tant qu'elle reste transparente, que l'actuaire maîtrise son calcul et comprend totalement l'origine des résultats.

Chapitre 6

Prochaines étapes et voies d'amélioration

Sommaire

Préambule

Au travers de l'élaboration d'un modèle de résiliation et d'un modèle de crédit commercial, nous avons pu résoudre le problème d'optimisation sous contraintes pour déterminer les majorations optimales des primes afin de limiter la résiliation tarifaires des contrats MRH arrivant leur terme. Les résultats obtenus sont encourageants et indiquent notamment que la stratégie de majoration actuelle d'AXA France n'est pas la plus efficiente. Nous avons alors proposé un modèle de prédiction de majorations optimales basé sur des variables tarifaires qui pourrait être mis en place au niveau opérationnel tous les mois.

Ainsi, l'objectif de ce dernier chapitre n'est pas de remettre en cause l'étude réalisée, mais plutôt de proposer des voies d'améliorations, des pistes de réflexion qui viendraient compléter cette étude et qui permettraient une meilleure implémentation du calcul des majorations optimales au terme opérationnellement parlant.

6.1 Amélioration des modèles de résiliation et de crédit commercial

6.1.1 Voies d'amélioration du modèle de résiliation

Nous avons vu dans le chapitre 2 comment nous avions procédé pour créer le modèle de résiliation : sur la base des observations des années 2017 et 2018, et en se focalisant sur les variables explicatives les plus significatives, nous avons réalisé une régression logistique avec un GLM. Ce modèle pourrait être amélioré, notamment :

— En ajoutant des **effets d'interactions** : par exemple, suivant leur profil (propriétaires d'appartement ou de maison, ou locataires d'appartement ou de maison), les comportements des clients peuvent être différents. Ajouter des interactions avec cette variable de profil pourrait améliorer la prédiction de résiliation tarifaire au terme et rendre le modèle plus fiable.

— En **introduisant une pondération sur le type de logement assuré** lors du calcul de la moyenne du taux de résiliation : en effet, la résiliation d'un contrat concernant l'assurance d'une maison a plus d'impact sur le portefeuille que la résiliation d'un contrat d'assurance d'appartement, puisque l'expérience montre qu'un contrat pour une maison dure en moyenne plus longtemps que pour un appartement. Ainsi,

pour une même ancienneté de contrat, le manque à gagner est plus important lors de la résiliation d'un contrat pour une maison que pour un appartement. Dans ce cadre, il faudrait introduire la notion de valeur de contrat et de valeur client, qui estime la ressource apportée par un contrat sur les années à venir. La pondération sera plus importante pour les contrats à forte valeur. Cela permettrait par la suite d'ajuster l'optimisation.

— En introduisant une nouvelle variable qui serait l'**indicatrice du passage à la centaine de la prime après l'application de la majoration** : le passage à la centaine peut potentiellement avoir un impact important sur la réaction du client en accentuant sa volonté de résilier son contrat. Il faudrait dans un premier temps déterminer si oui ou non cette variable est réellement significative, et si oui, entraîner un nouveau modèle GLM avec cette variable. Toutefois, si cette variable devait être introduite dans le modèle, cela engendrerait l'apparition de discontinuités dans la fonction de résiliation, ce qui serait problématique par la suite lors de la résolution du problème d'optimisation qui nécessite la continuité des fonctions. Une idée pour prendre en compte cet effet de passage à la centaine serait de garder les résultats de l'optimisation telle que nous l'avons faite (avec le modèle de résiliation initial, sans la variable de passage à la centaine), puis de repérer les contrats qui montrent effectivement un passage à la centaine suite à la majoration optimale. Pour ces contrats dont la majoration n'est pas trop grande, nous pourrions regarder la probabilité de résiliation si la prime était maintenue juste en dessous de la centaine supérieure (par exemple une prime de 270 € qui serait augmentée à 310 € avec la majoration optimale serait finalement maintenue à 299 €) en utilisant le nouveau modèle de résiliation qui prend en compte le passage à la centaine, et ensuite observer l'impact sur le taux moyen de résiliation du portefeuille, le chiffre d'affaires et l'ELR total.

— **En étendant le périmètre d'étude aux étudiants et aux propriétaires non occupants**, qui ont des comportements différents et pour lesquels il faudrait réaliser une régression logistique qui leur serait propre.

— Enfin, un **price test sur les contrats termés** doit être réalisé avant même de considérer utiliser cette étude au niveau opérationnel. En effet, afin de s'assurer que notre modèle de résiliation est fidèle à la sensibilité réelle du client à l'augmentation de sa prime, il est nécessaire de comparer l'élasticité-majoration de la résiliation du modèle avec celle observée dans la réalité. En cas d'écart trop grand, il sera nécessaire d'ajuster le modèle de résiliation.

6.1.2 Voies d'amélioration du modèle de crédit commercial

Rappelons que le modèle de crédit commercial a été construit en deux étapes afin d'estimer l'espérance du pourcentage de crédit commercial qui serait appliquée à un contrat arrivant à son terme par rapport à la prime initiale : d'abord par une régression logistique (GLM) déterminant la probabilité d'application du rabais, puis par le bais d'une régression linaire qui estime l'espérance du logarithme du pourcentage de crédit commercial dans le cas échéant.

Tout comme pour le modèle de résiliation, certaines améliorations peuvent être apportées à ces modèles :

— En ajoutant des **effets d'interactions**, notamment sur les profils des clients ou sur les réseaux de distribution.

— **En étendant le périmètre d'étude aux étudiants et aux propriétaires non occupants**.

— **En améliorant le modèle de Competitive Market Analysis** (CMA), car comme nous l'avons vu au chapitre 3, la distance du prix au CMA est une variable significative dans le modèle de régression logistique.

Nous pourrions également vérifier si la distribution du pourcentage de crédit commercial se rapproche d'une loi Gamma-Pareto, auquel cas il serait possible de réaliser un GLM car cette loi appartient à une famille

exponentielle. Cela résoudrait le problème de modélisation des queues lourdes qui apparaissent dans cette distribution.

En outre, comme nous l'avons remarqué au chapitre 3, si le client réclame une diminution de sa prime d'assurance suite à une majoration, l'agent général peut soit appliquer un rabais, soit proposer de "remplacer" le contrat en diminuant le nombre d'options et garanties ce qui aura mécaniquement pour effet de venir diminuer le coût de l'assurance. Nous pourrions considérer la création d'un **modèle de remplacement**, de la même manière que nous avons créé le modèle de crédit commercial. Nous estimerions dans un premier temps la probabilité qu'un contrat soit remplacé, puis de combien sa prime serait diminuée le cas échéant.

Toutefois, un tel modèle reste compliqué à mettre en place : en effet, tous les remplacements ne consistent pas à réduire le nombre de garanties. Par exemple, un contrat peut être remplacé si le client souhaite ajouter des garanties ou encore s'il déménage, et dans ce cas la prime peut augmenter. Il sera donc nécessaire d'identifier particulièrement les remplacements hors déménagement qui revoient la prime à la baisse.

6.2 Amélioration de l'optimisation des majorations au terme

Sur la base des modèles de résiliation et de crédit commercial perfectionnés, il serait alors possible de réaliser à nouveau le travail d'optimisation.

Dans le cadre de notre étude, nous avons résolu le problème d'optimisation sur un mois, car compte tenu de notre périmètre, nous avons jugé qu'il n'y avait pas de saisonnalité dans les taux de résiliation tarifaire. Toutefois, il ne serait pas inutile de réaliser ce travail sur les 12 mois de l'année pour plus de précision, et observer l'impact de cette optimisation complète sur le taux de résiliation moyen, l'ELR total et le chiffre d'affaires.

Par ailleurs, si nous élargissons le périmètre aux étudiants et aux propriétaires non occupants, des effets de saisonnalité peuvent apparaître, plus particulièrement pour les étudiants qui auront davantage tendance à résilier pendant l'été. Dans ce cas, l'optimisation devra bien considérer le mois de terme des contrats.

6.3 Ajustement du reverse engineering

Le travail de reverse engineering a pour principal intérêt de faciliter tous les mois la détermination des majorations optimales pour les contrats arrivant à leur terme. Dans le chapitre 5, nous avons été confronté au sujet de la distribution des majorations optimales, laquelle ne se rapproche pas de lois usuelles connues, nous empêchant de ce fait de réaliser un GLM pour prédire les majorations optimales grâce à une formule générique. Nous avons donc décidé de réaliser le reverse engineering avec un GBM et nous avons proposé une méthode alternative pour expliquer les prédictions obtenues avec la théorie des SHAP values.

Cependant, nous comprenons que malgré tout, l'utilisation d'un GLM pour la prédiction des majorations optimales serait plus simple pour une implémentation opérationnelle immédiate, car cela aboutirait à une formule générique transparente et interprétable. Or, nous avons vu au chapitre 4 qu'il nous était éventuellement possible d'influencer plus ou moins la forme de la distribution des majorations optimales suivant les contraintes du problème d'optimisation. L'idée serait d'ajuster le poids de la contrainte de distance à une majoration α (telle que définie au chapitre 4) afin de modifier le skewness et le kurtosis de la distribution, de telle sorte que cette dernière se rapproche d'une loi exponentielle modélisable par un GLM, tout en respectant les contraintes du cahier des charges sur le taux de résiliation moyen, l'ELR total et le chiffre d'affaires.

Conclusion du chapitre

Les voies d'amélioration proposées permettraient d'affiner la prédiction des résiliations et du crédit commercial, ce qui ajustera la résolution du problème d'optimisation. Plus les modèles seront robustes et précis, plus nous pourrons avoir confiance en la prédiction des majorations optimales, sans pour autant remettre en cause la méthodologie proposée tout au long de cette étude.

Conclusion

Afin de contenir l'hémorragie du portefeuille MRH d'AXA France, où le nombre de contrats résiliés dépasse le nombre d'affaires nouvelles, ce mémoire vise à trouver une solution pour limiter les résiliations tarifaires au terme des contrats les plus rentables en jouant sur les niveaux de majoration. Il s'agit donc d'un problème d'optimisation dont l'objectif est de minimiser le taux de résiliation tarifaire au terme du portefeuille sous contrainte de rentabilité totale évaluée par l'ELR.

Les statistiques descriptives ont en effet révélé que la majoration a un impact significatif sur la probabilité de résiliation tarifaire et sur l'application du crédit commercial. De ce fait, cela a du sens de manipuler les majorations pour minimiser le taux de résiliation sous contrainte de profitabilité.

Ce mémoire a donc proposé une méthodologie d'optimisation de ces majorations, en passant par des étapes de modélisation de la probabilité de résiliation tarifaire au terme et de l'espérance de crédit commercial.

Il a été montré que le modèle linéaire généralisé pénalisé était suffisamment performant pour modéliser la probabilité de résiliation, et était surtout plus robuste et stable que la méthode de gradient boosting machine. Cependant, pour être exhaustif dans l'étude et avant même de considérer implémenter ce modèle au niveau opérationnel, il conviendra de réaliser dans la suite des travaux un price-test au terme, afin d'évaluer l'élasticité-majoration de la résiliation du modèle et la comparer aux observations résultants du price-test.

Par ailleurs, l'espérance du pourcentage de crédit commercial appliquée aux primes de contrat a été modélisée en deux étapes, également sur la base de modèles linéaires généralisés. La première consiste en la modélisation de la probabilité d'application du crédit commercial, par une régression logistique. La seconde étape concerne l'estimation du pourcentage de crédit commercial par rapport à la prime qui serait accordé le cas échéant. Cette deuxième étape a nécessité une analyse préalable de la distribution des pourcentages de crédit commercial : il a été décidé d'approximer cette dernière par une distribution log-normale, ce qui permet par la suite de modéliser aisément le logarithme du pourcentage de rabais par une régression linéaire, et les variables explicatives les plus pertinentes ont été sélectionnées par un GBM. La combinaison de ces deux étapes aboutit à des prédictions satisfaisantes et l'analyse des résidus a montré que l'approximation sur la distribution des pourcentages de rabais était raisonnable.

A l'issue de ces premières modélisations et à partir du modèle de prime pure déjà établit par AXA France, il a été possible d'estimer la ressource espérée de chacun des contrats en portefeuille en fonction de leurs caractéristiques, celles du client, et surtout en fonction de la majoration de chaque contrat. De là, une formule générique du taux de résiliation du portefeuille a été établie en fonction des majorations appliquées aux différents contrats, et de même pour l'ELR total du portefeuille, qui représente la contrainte de rentabilité du portefeuille. A partir de ces éléments, le travail d'optimisation a pu être réalisé.

Plusieurs difficultés sont apparues lors de la résolution du problème d'optimisation sous contrainte. La première concerne le grand volume de majorations à optimiser pour chaque mois. Or la résolution d'un problème d'optimisation sur un grand nombre de variables devient vite complexe et nécessite une grande puissance de calcul. Il a donc été nécessaire de reformuler le problème sans perdre de vue l'objectif final : minimiser le taux de résiliation du portefeuille sous contrainte de l'ELR total. L'idée a été de transformer ce problème global en plusieurs problèmes d'optimisation individuels, sous une contrainte différente et sous des hypothèses fortes d'indépendance entre les comportements des clients, aboutissant ainsi à autant de Lagrangiens que de majorations à optimiser. La seconde difficulté réside dans le fait que les Lagrangiens individuels ainsi définis n'étaient pas convexes en fonction de la majoration, ce qui provoque des solutions en coin. Une contrainte supplémentaire de convexité a dû être rajoutée. De plus, l'absence de convexité de la probabilité de résiliation en fonction de la majoration pour chaque contrat induit que le problème dual n'est pas équivalent au problème primal. Ainsi, la résolution du problème primal nous donne accès à des optimums locaux suivant la valeur du multiplicateur de Lagrange choisi, mais pas à un optimum global. Finalement, le choix du meilleur multiplicateur de Lagrange, et donc la détermination des meilleures majorations optimales a été fait en vérifiant que les KPIs (que sont le taux de résiliation du portefeuille, l'ELR total et le chiffre d'affaires) répondent bien au cahier des charges.

Cette recherche de majorations optimales a également permis de matérialiser la frontière efficiente du taux de résiliation en fonction de l'ELR, et de positionner la stratégie actuelle d'AXA France par rapport à cette frontière caractérisant l'ensemble des stratégies optimales atteignables. Il apparaît alors clairement une marge de progression possible de la politique de majoration de l'assureur pour rejoindre un point de cette frontière.

Les résultats obtenus à l'issue de cette étude sont très satisfaisants, inédits et encourageants : **l'application des majorations optimales obtenues diminuerait le taux de résiliation moyen, améliorerait la rentabilité du portefeuille ainsi que l'image de marque de l'assureur, sans pour autant dégrader le chiffre d'affaires**.

Ces majorations optimales sont distribuées de manière très différente en comparaison avec la stratégie actuelle d'AXA France : en moyenne, elles sont inférieures aux majorations appliquées actuellement sur les contrats, et certaines majorations optimales sont même négatives, ce qui contribuerait a embellir l'image de marque de l'assureur. Ces minorations concernent principalement les contrats rentables dont les détenteurs sont sensibles à la majoration (notamment les clients les moins seniors), et qui présentent un historique de sinistres léger. En outre, en étudiant les majorations les plus importantes, il apparaît que les majorations optimales sont davantage orientées vers le devenir des contrats : ce ne sont pas en effet les contrats les plus sinistrés qui subissent les plus grosses majorations, mais plutôt ceux qui présentent une rentabilité estimée défavorable pour l'assureur. L'optimisation propose alors une autre vision de l'application de la majoration, plus sensible à l'ELR des contrats.

S'il demeurera toujours un écart entre la théorie et la pratique, notamment puisqu'il est difficile de prévoir et d'influencer les méthodes commerciales des agents généraux et des courtiers concernant les gestes commerciaux, ce travail d'optimisation permet toutefois d'améliorer la stratégie actuelle d'AXA France en l'approchant de la frontière efficiente.

Ensuite, en vue d'une implémentation opérationnelle de ces majorations optimales, il convient de dire qu'il n'est pas aisé de pratiquer un travail d'optimisation tous les mois. C'est pourquoi un reverse engineering a été réalisé au moyen d'un gradient boosting machine afin de prédire ces majorations à partir de certaines caractéristiques des contrats et des clients. Toutefois, l'usage du gradient boosting machine pose un problème de transparence : les prédictions obtenues doivent pouvoir être expliquées pour être utilisées à des fins tarifaires. Les SHAP values, issues de la théorie des jeux, permettent ainsi d'attribuer, pour chaque prédiction, une contri-

bution (positive ou négative) à chaque variable explicative, justifiant ainsi, à qui souhaite le savoir, le niveau de majoration suggéré. Cette méthode innovante d'interprétation de modèles mérite d'être davantage exploitée car elle lève le voile sur des modèles "boîtes noires" qui ne pouvaient pas être utilisés jusqu'à lors pour la tarification de contrats.

Enfin, ce mémoire répond à la problématique de majoration au terme des contrats MRH pour l'année en cours. Il peut toutefois être envisagé, en guise d'ouverture, de considérer une optimisation des majorations sur les années suivantes, à un horizon plus ou moins lointain, en prenant en compte la notion de valeur client pour chaque contrat considéré. Il serait également pertinent de prendre en compte l'évolution des anciennes majorations pour un contrat donné (si cette information est disponible), afin d'éviter l'application à répétition de majorations importantes à un client déjà suffisamment rentable et susceptible de résilier.

Annexes

Sommaire

Annexe 1 : Article de la loi Chatel

LOI n° 2005-67 du 28 janvier 2005 tendant à conforter la confiance et la protection du consommateur

TITRE 1er : FACILITER LA RÉSILIATION DES CONTRATS TACITEMENT RECONDUCTIBLES

Article 1

Le titre III du livre Ier du code de la consommation est complété par un chapitre VI ainsi rédigé :

« Chapitre VI »

« Reconduction des contrats »

« Art. L. 136-1. - Le professionnel prestataire de services informe le consommateur par écrit, au plus tôt trois mois et au plus tard un mois avant le terme de la période autorisant le rejet de la reconduction, de la possibilité de ne pas reconduire le contrat qu'il a conclu avec une clause de reconduction tacite. »

« Lorsque cette information ne lui a pas été adressée conformément aux dispositions du premier alinéa, le consommateur peut mettre gratuitement un terme au contrat, à tout moment à compter de la date de reconduction. Les avances effectuées après la dernière date de reconduction ou, s'agissant des contrats à durée indéterminée, après la date de transformation du contrat initial à durée déterminée, sont dans ce cas remboursées dans un délai de trente jours à compter de la date de résiliation, déduction faite des sommes correspondant, jusqu'à celle-ci, à l'exécution du contrat. A défaut de remboursement dans les conditions prévues ci-dessus, les sommes dues sont productives d'intérêts au taux légal. »

« Les dispositions du présent article s'appliquent sans préjudice de celles qui soumettent légalement certains contrats à des règles particulières en ce qui concerne l'information du consommateur. »

Annexe 2 : Article de la loi Hamon

Article L113-15-2 du code des assurances

Créé par la LOI n° 2014-344 du 17 mars 2014 - art. 61 (V)

Pour les contrats d'assurance couvrant les personnes physiques en dehors de leurs activités professionnelles et relevant des branches définies par décret en Conseil d'Etat, l'assuré peut, à l'expiration d'un délai d'un an à compter de la première souscription, résilier sans frais ni pénalités les contrats et adhésions tacitement reconductibles. La résiliation prend effet un mois après que l'assureur en a reçu notification par l'assuré, par lettre ou tout autre support durable.

Le droit de résiliation prévu au premier alinéa est mentionné dans chaque contrat d'assurance. Il est en outre rappelé avec chaque avis d'échéance de prime ou de cotisation.

Lorsque le contrat est résilié dans les conditions prévues au premier alinéa, l'assuré n'est tenu qu'au paiement de la partie de prime ou de cotisation correspondant à la période pendant laquelle le risque est couvert, cette période étant calculée jusqu'à la date d'effet de la résiliation. L'assureur est tenu de rembourser le solde à l'assuré dans un délai de trente jours à compter de la date de résiliation. A défaut de remboursement dans ce délai, les sommes dues à l'assuré produisent de plein droit intérêts au taux légal.

Pour l'assurance de responsabilité civile automobile définie à l'article L. 211-1 et pour l'assurance mentionnée au g de l'article 7 de la loi n 89-462 du 6 juillet 1989 tendant à améliorer les rapports locatifs et portant modification de la loi n 86-1290 du 23 décembre 1986, le nouvel assureur effectue pour le compte de l'assuré souhaitant le rejoindre les formalités nécessaires à l'exercice du droit de résiliation dans les conditions prévues au premier alinéa du présent article. Il s'assure en particulier de la permanence de la couverture de l'assuré durant la procédure.

Un décret en Conseil d'Etat précise les modalités et conditions d'application du présent article.

NOTA : Loi n° 2014-344 du 17 mars 2014 art. 61 II : Ces dispositions s'appliquent aux contrats conclus ou tacitement reconduits à compter de la publication du décret mentionné au dernier alinéa de l'article L. 113-15-2 du code des assurances.

Annexe 3 : Théorie du Gradient Boosting Machine

La méthode de *Gradient Boosting Machine* [1] (ou boosting d'arbres de régression [2]), est une méthode supervisée de machine learning basée sur les arbres de décision CART (*Classification and Regression Tree*).

Théorie des arbres de décision CART [3]

Les arbres de décision permettent de résoudre des problèmes de régression ou de classification. Cette méthode d'apprentissage supervisée repose sur une série successives de tests conditionnels. L'objectif est alors de séparer les observations $\boldsymbol{X_i}$ étudiées en groupes homogènes en terme de variables réponses $\boldsymbol{Y_i}$, à partir de leurs différentes variables explicatives.

Ainsi, la construction d'un arbre débute par un premier test qui constitue un *noeud racine* : l'arbre CART va sélectionner la variable $X_{i,j}$ et une valeur bien choisie de cette variable qui sépare le mieux l'espace des variables réponses en deux sous-espaces les plus homogènes possibles et les plus "éloignés" l'un de l'autre (au sens de l'impureté de Gini dans un problème de classification). De ce noeud racine naissent deux branches correspondant aux deux réponses possibles. Selon l'input, l'arbre va orienter l'observation dans une des deux branches, laquelle engendrera soit un nouveau noeud (et donc un nouveau test conditionnellement au test précédent) soit un noeud final (aussi appelé *feuille*) qui correspondra à une décision, c'est-à-dire à l'attribution d'une variable réponse (étiquette) à cette observation.

Le choix optimal de la variable $X_{i,j}$ ainsi que de sa valeur a_j à chaque noeud dépend de la réduction maximale de l'impureté en passant de l'échantillon du noeud père N aux échantillons des deux noeuds fils N_g (noeud de gauche) et N_d, (noeud de droite). Soit $I(N)$ une fonction de mesure de l'impureté du noeud N par rapport à la classe cible. L'objectif est alors de résoudre le problème d'optimisation suivant :

$$\arg \max_{j,a_j} \Delta I(N)$$

avec

$$\begin{aligned}\Delta I(N) &= I(N) - \mathbb{E}[I(N_{gd})] \\ &= I(N) - \left(\frac{|N_g|}{|N|} I(N_g) + \frac{|N_d|}{|N|} I(N_d)\right)\end{aligned}$$

Dans le cadre d'un problème de classification, la mesure de l'impureté utilisée par les arbres CART est l'indice de Gini, dont le principe est rappelé en annexe 5. Si nous partitionnons le noeud N en k groupes $C_1, ..., C_k$, et que nous considérons la probabilité $p_i \approx \frac{|C_i|}{N}$ qu'un élément de N appartienne au groupe C_i, alors la mesure de l'impuretée selon l'indice de Gini est calculée comme suit :

$$I_G(N) = \sum_{i=1}^{k} p_i(1-p_i) = 1 - \sum_{i=1}^{k} p_i^2$$

$$I_G(N) = \sum_{\substack{i,j=1,\\ i\neq j}}^{k} p_i p_j$$

1. D'après l'ouvrage *Le Machine learning avec Python* [8].
2. A noter que ce même algorithme est aussi utilisable pour des problèmes de classification.
3. D'après le cours du CNAM sur les arbres de décisions [2].

Ainsi $I_G(N) = 0$ si le noeud N est homogène, c'est-à-dire que tous les éléments de N appartiennent au même groupe.

L'arbre de décision partitionne l'espace des observations en un certain nombre de régions égal au nombre de feuilles. Pour un problème de classification, l'étiquette attribué à une feuille correspond à l'étiquette majoritaire de cette région.

La construction de l'arbre de décision se termine suivant un critère d'arrêt que l'utilisation impose. Cela peut être un critère de volume minimal d'observations dans chaque feuille par exemple, ou bien un critère de précision. Cette méthode a l'avantage d'être facile à comprendre, à la fois pour l'apprentissage et pour la prédiction. A noter que si l'arbre est trop profond (beaucoup de noeuds) alors précision augmente mais le modèle risque de sur-apprendre, ce qui aboutira à de mauvaises performances en terme de prédiction.

Théorie du Gradient Boosting Machine

Principe général

Revenons au Gradient Boosting : cette méthode consiste en la construction successive d'arbres de décision, de telle sorte que l'arbre suivant corrige les erreurs engendrées par l'arbre précédent (principe du *boosting*), et l'optimisation itérée pour réduire l'erreur est réalisée par descente de gradient.

Les arbres de décision utilisés dans ce processus sont peu profonds (maximum 5 niveaux en général), résultant ainsi en un modèle plus puissant en termes de rapidité de prédiction. Le gradient boosting repose sur l'idée que la combinaison de plusieurs petits arbres de décision différents aux performances individuelles moindres sur l'ensemble des observations résulte en un algorithme d'une performance largement supérieure aux performances individuelles.

Elaboration de l'algorithme d'optimisation

Selon l'article de Friedman, *Greedy Function Approximation : A Gradient Boosting Machine* [5], nous supposons qu'il existe une fonction F^* qui explique le mieux la variable réponse Y par les p variables explicatives $(X_1, ..., X_p)$ retenues dans le vecteur $\boldsymbol{X}$:

$$F^* \in \arg\ \min_F\ \mathbb{E}_X\left[\mathbb{E}_Y[L(Y, F(\boldsymbol{X}))|\boldsymbol{X}]\right]$$

avec L une fonction de perte qui, dans le cadre d'un algorithme de classification, peut être égale à la log-vraisemblance de la loi binomiale négative $\log(1+e^{-2YF(\boldsymbol{X})})$. La fonction de perte est une mesure de la différence entre la valeur observée de Y et la valeur estimée de Y, c'est-à-dire $F(\boldsymbol{X})$. L'objectif est donc de trouver F qui va minimiser cette fonction de perte.

Nous souhaitons trouver une approximation $\widetilde{F^*}$ de la fonction F^*. Pour ce faire, l'idée est dans un premier temps d'exprimer la fonction F par un modèle paramétrique $F(\boldsymbol{X}, \boldsymbol{P})$, avec $\boldsymbol{P} = (P_1, P_2, ...)$ un ensemble fini de paramètres tels que $P_i = \{\beta_i, \boldsymbol{a}_i\} \in \mathbb{R} \times \mathbb{R}^{k_i}$. Le modèle paramétrique s'écrit alors sous une forme additive :

$$F(\boldsymbol{X}, \{\beta_m, \boldsymbol{a}_m\}_1^M) = \sum_{m=1}^{m} \beta_m h(\{\boldsymbol{X}, \boldsymbol{a}_m\})$$

avec h un modèle paramétrique dont les performances sont faibles. Dans notre cas, h est un arbre de décision CART, dont les paramètres caractéristiques sont donnés par $\boldsymbol{a}_i = (a_{1,i}, a_{1,i}, ..., a_{k_i,i})$.

Nous nous ramenons donc à un problème d'optimisation paramétrique :

$$\begin{aligned}\boldsymbol{P^*} &= \arg\ \min_{P}\ \Phi(\boldsymbol{P}) \\ &= \arg\ \min_{P}\ \mathbb{E}_X\left[\mathbb{E}_Y[L(Y, F(\boldsymbol{X}, \boldsymbol{P}))|\boldsymbol{X}]\right]\end{aligned}$$

d'où

$$F^*(\boldsymbol{X}) = F(\boldsymbol{X}, \boldsymbol{P^*})$$

La solution d'un tel problème d'optimisation paramétrique s'écrit sous une forme additive :

$$\boldsymbol{P^*} = \sum_{m=0}^{M} \boldsymbol{p}_m$$

avec $\boldsymbol{p}_0$ un vecteur d'initiation et $\{\boldsymbol{P}_m\}_1^M$ les "*boosts*", autrement dit les incréments successifs permettant de réduire les erreurs afin de s'approcher de $\boldsymbol{P^*}$. Ces derniers sont calculés par descente de gradient.

Cela revient donc à résoudre le modèle non paramétrique suivant :

$$\begin{aligned}& F^*(\boldsymbol{X}) \in \arg\ \min_{F}\ \Phi(F) \\ \iff & F^*(\boldsymbol{X}) \in \arg\ \min_{F}\ \mathbb{E}_X\left[\mathbb{E}_Y[L(Y, F(\boldsymbol{X}))|\boldsymbol{X}]\right] \\ \iff & F^*(\boldsymbol{X}) \in \arg\ \min_{F}\ \mathbb{E}_Y[L(Y, F(\boldsymbol{X}))|\boldsymbol{X}]\end{aligned}$$

La solution d'un tel problème d'optimisation s'écrit sous la forme :

$$F^*(\boldsymbol{X}) = \sum_{m=0}^{M} f_m(\boldsymbol{X})$$

avec $f_0(\boldsymbol{X})$ un choix initial (peut être la moyenne $\bar{Y}$ des variables réponse Y), et $f_m(\boldsymbol{X})$ les "*boosts*" obtenus par descente de gradient et exprimés selon :

$$f_m(\boldsymbol{X}) = -\rho_m g_m(\boldsymbol{X})$$

avec

$$g_m(\boldsymbol{X}) = \left[\frac{\partial \mathbb{E}_Y[L(Y, F(\boldsymbol{X}))|\boldsymbol{X}]}{\partial F(\boldsymbol{X})}\right]_{F(\boldsymbol{X})=F_{m-1}(\boldsymbol{X})}$$

et

$$F_{m-1}(\boldsymbol{X}) = \sum_{i=0}^{m-1} f_i(\boldsymbol{X})$$

g_m exprime le gradient, et va indiquer la direction de plus grande pente vers le minimum de l'espérance de la fonction de perte $\mathbb{E}_Y[L(Y, F(\boldsymbol{X}))|\boldsymbol{X}]$. En fait, l'idée derrière la descente de gradient est d'atteindre le plus rapidement possible le minimum, c'est-à-dire atteindre ce minimum en un nombre minimum d'itérations.

Si L et $\frac{\partial L(Y,F(\boldsymbol{X})}{\partial F(\boldsymbol{X})}$ sont continues, alors nous pouvons intervertir l'espérance et la dérivation :

$$g_m(\boldsymbol{X}) = \mathbb{E}_Y \left[\frac{\partial L(Y, F(\boldsymbol{X}))}{\partial F(\boldsymbol{X})} | \boldsymbol{X} \right]_{F(\boldsymbol{X})=F_{m-1}(\boldsymbol{X})}$$

Enfin le facteur ρ_m est déterminé de telle sorte que :

$$\begin{aligned} \rho_m &= \arg \min_{\rho} \mathbb{E}_{Y,X}[L(Y, F_{m-1}(\boldsymbol{X}) - \rho g_m(\boldsymbol{X}))] \\ &= \arg \min_{\rho} \mathbb{E}_{Y,X}[L(Y, F_m(\boldsymbol{X}))] \end{aligned}$$

L'approximation $\widetilde{F^*} = F_M(\boldsymbol{X})$ de F^* peut alors se faire sur un échantillon $\{\boldsymbol{X}_i, Y_i\}_{i=1,\dots,n}$ de manière empirique :

1ère étape : Initialisation

L'initialisation de $f_0(\boldsymbol{X}_i)$ dépend du type de classification ou de régression.

2ème étape : Itération pour m allant de 1 à M

La réduction d'erreur au fur et à mesure des itérations se fait par descente de gradient. Nous définissons donc le gradient en $\boldsymbol{X_i}$ à partir de l'échantillon observé :

$$g_m(\boldsymbol{X}_i) = \left[\frac{\partial L(Y_i, F(\boldsymbol{X}_i))}{\partial F(\boldsymbol{X}_i)} \right]_{F(\boldsymbol{X})=F_{m-1}(\boldsymbol{X})}$$

Ce gradient n'est défini que sur les $\boldsymbol{X_i}$ disponibles. Alors pour généraliser le gradient négatif (puisqu'il s'agit bien d'une descente de gradient) à toutes les valeurs de $\boldsymbol{X}$ possibles, nous pouvons l'estimer avec le modèle paramétrique $h(\boldsymbol{X}, \boldsymbol{a}_m)$ tel que :

$$\boldsymbol{a}_m = \arg \min_{\boldsymbol{a},\beta} \sum_{i=1}^{n} [-g_m(\boldsymbol{X}_i) - \beta h(\boldsymbol{X}, \boldsymbol{a})]^2$$

Une fois les paramètres $\boldsymbol{a}_m$ estimés, nous pouvons résoudre l'optimisation suivante :

$$\rho_m = \arg \min_{\rho} \sum_{i=1}^{n} L(Y_i, F_{m-1}(\boldsymbol{X}_i) + \rho h(\boldsymbol{X}, \boldsymbol{a}_m))$$

La mise à jour de l'approximation de F^* devient, à la m-ième itération :

$$\widetilde{F^*}(\boldsymbol{X}) = F_m(\boldsymbol{X}) = F_{m-1}(\boldsymbol{X}) + \rho_m h(\boldsymbol{X}, \boldsymbol{a}_m)$$

Paramètres du Gradient Boosting

Trois paramètres sont à définir pour réaliser un apprentissage GBM :

— La **profondeur maximale** des arbres,

— le *learning state*, qui détermine l'**intensité de la correction des erreurs** de l'arbre précédent à chaque itération de construction d'arbre ;

— et le **nombre d'arbres** M à construire au total.

Augmenter ces trois paramètres [4], toutes choses égales par ailleurs, complexifie le modèle et améliore la précision, mais augmente le risque de sur-apprentissage. Alors choisir de manière optimale ces paramètres peut

4. Il est aussi possible de rajouter d'autres types de paramètres, comme le nombre maximal de variables à considérer dans un arbre CART par exemple.

être fait grâce à un *Grid Search* (ou recherche sur grille) : à chaque paramètre nous allons assigner un certain nombre de valeurs possibles, puis le *Grid Search* va réaliser le modèle en testant toutes les combinaisons possibles des paramètres et va retenir les paramètres correspondants au modèle qui maximise l'AUC (aire sous la courbe ROC, ie *Receiver Operating Characteristic*[5]).

A partir de ce meilleur modèle, il est possible de visualiser l'importance de chacune des variables explicatives, c'est-à-dire à quel point elles contribuent à la décision que prennent les arbres de décision. Ainsi, si l'importance d'une variable vaut 1, cela signifie que cette variable prédit le mieux la variable réponse, et si au contraire l'importance vaut 0, cela indique que cette variable n'est pas du tout utilisée.

5. L'AUC est également un indicateur de la qualité de la segmentation des variables réponses. Cf. Annexe 5

Annexe 4 : Théorie du modèle linéaire généralisé (GLM)

Soit $Y = (Y_i, ...Y_n)'$ le vecteur colonne représentant les différentes valeurs observées d'une variable réponse et $X_i = (1, X_{i,1}, ..., X_{i,p})'$ le vecteur colonne représentant les p variables explicatives pour la variable réponse Y_i. X est alors une matrice de taille $n \times (p+1)$ dont les lignes sont les vecteurs lignes X_i'.

Rappels sur les modèles linéaires gaussiens

Rappelons que le principe de la **régression linéaire** est de modéliser l'espérance conditionnelle de la variable réponse $\mathbb{E}[Y|X]$ comme une combinaison linéaire des variables explicatives caractérisée par une fonction que nous appellerons h :

$$\mathbb{E}[Y|X] = h(X) = X\beta \tag{6.1}$$

avec $\beta = (\beta_0, ..., \beta_p)$ le vecteur colonne correspondant aux $p+1$ paramètres du modèle qui seront à estimer.

De ce fait, la variable Y s'écrit alors sous la forme d'un modèle linéaire gaussien :

$$Y = \mathbb{E}[Y|X] + \varepsilon = h(X) + \varepsilon \tag{6.2}$$

si les hypothèses suivantes sont vérifiées :

- ε est un bruit aléatoire gaussien [6] tel que :
  - $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ $\forall i$ (hypothèse d'homoskedasticité : la variance des résidus est constante) [7]
  - $Cov(\varepsilon_i, \varepsilon_j) = 0$, $\forall i \neq j$
- la matrice X est de plein rang, *ie* rang(X)=$p+1$.

L'objectif est alors de retrouver la fonction h, c'est-à-dire d'estimer les paramètres β_i en fonction des observations (Y, X). Cela est possible si les hypothèse sus-citées sont vérifiées, grâce à la méthode des moindres carrés :

$$\widehat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \|Y - X\beta\|_2^2 \tag{6.3}$$

Nous obtenons alors $\widehat{\beta} = (X'X)^{-1}X'Y$ qui est un estimateur unique et sans biais de variance $\widehat{\sigma^2} = \frac{1}{n-p-1}\left\|Y - X\widehat{\beta}\right\|^2$.

Rappels sur les modèles linéaires généralisés

Un **modèle linéaire généralisé** (ou *Generalized Linear Model*, GLM) est proche du modèle de régression linéaire dans la mesure où il va chercher à expliquer la variable réponse Y au moyen d'une combinaison linéaire des variables explicatives X qui sera ensuite transformée par une fonction de lien, ce qui permet de notamment de relâcher les hypothèses imposant la forme linéaire de h, que le bruit soit gaussien et de travailler avec de l'hétéroskedasticité.

Un modèle est un modèle linéaire généralisé s'il vérifie des hypothèses suivantes :

- $Y|X = x \sim \mathbb{P}_{\theta,\phi}$ appartient à une **famille exponentielle**

6. Cette hypothèse de normalité du bruit n'est pas indispensable : elle est nécessaire pour s'assurer que la distribution de chaque estimateur des coefficients de la régression linéaire conditionnellement à X suit une loi normale, mais dans le cas où ε n'est pas gaussien, alors les estimateurs des coefficients conditionnellement à X suivent de manière asymptotique une distribution normale

7. Dans le cas contraire, nous sommes dans un cas d'hétéroskedasticité. Il est possible de revenir au cas d'homoskedasticité en effectuant une transformation sur les Y_i et X_i en les standardisant par rapport à la variance.

- Pour une certaine fonction g bijective appelée **fonction de lien** nous avons : $g(\mathbb{E}[Y|X]) = X\beta$

Le GLM présente donc trois composantes : la distribution de la variable réponse Y conditionnellement à X, le prédicteur linéaire β, et la fonction de lien g.

La distribution de $Y|X$ ***:***

Nous appelons *famille exponentielle* un modèle statistique $(\Omega, \mathcal{F}, (\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi>0})$ si les probabilités $\mathbb{P}_{\theta,\phi}$ admettent une densité f par rapport à une mesure dominante [8] (dite de référence) telle que :

$$f_{\theta,\phi}(y) = c_\phi(y) \exp\left(\frac{y\theta - a(\theta)}{\phi}\right) \tag{6.4}$$

avec :

— θ le paramètre canonique ;

— ϕ le paramètre de dispersion, considéré comme un paramètre de nuisance. ϕ est toujours strictement positif ;

— $a(.)$ est une fonction de classe C^2 et convexe ;

— $c_\phi(.)$ est une fonction ne dépendant pas de θ.

Ainsi, une famille exponentielle de lois peut regrouper différentes distributions comme la loi Normale, la loi Gamma ou encore la loi de Bernoulli.

Le prédicteur linéaire β ***:***

Soit β, un vecteur de $p+1$ paramètres. Nous appelons prédicteur linéaire le vecteur à n composantes η tel que $\eta = X\beta$. Les paramètres de β sont estimés par maximisation de la vraisemblance qui provient d'une famille exponentielle de lois.

La fonction de lien g ***:***

La fonction de lien g permet de créer une relation fonctionnelle entre le prédicteur linéaire η et l'espérance conditionnelle de la variable aléatoire Y :

$$g(\mathbb{E}[Y|X]) = \eta = X\beta \tag{6.5}$$

g doit être monotone et différentiable. Par ailleurs, la fonction de lien qui associe $\mathbb{E}[Y|X]$ au paramètre naturel θ est appelée **fonction de lien canonique**. En effet, une propriété associée aux familles de lois exponentielles indique que si une variable aléatoire Y est distribuée selon une loi appartenant à une famille exponentielle, alors $\mathbb{E}[Y] = a'(\theta)$. Ainsi, dans le cas de notre modèle linéaire généralisé nous avons :

$$g(\mathbb{E}[Y|X]) = g(a'(\theta)) = X\beta \tag{6.6}$$

En choisissant $g = (a')^{-1}$ (fonction de lien canonique) nous avons la relation suivante :

$$\theta = X\beta \tag{6.7}$$

Le GLM pénalisé

Dans la pratique, il n'est pas évident de discerner parmi une multitude de variables celles qui sont inutiles à l'explication de la valeur de la variable réponse à prédire, de celles qui sont les plus significatives. La pénalisation

8. Le modèle statistique est dit un modèle dominé s'il existe une mesure positive μ sur $(\Omega, \mathcal{F})$ telle que pour tout $\theta \in \Theta, \phi > 0$ $\mathbb{P}_{\theta,\phi}$ est absolument continue par rapport à μ.

du GLM permet de lever cette contrainte de choix en déterminant quelles seraient les variables explicatives à éliminer en rendant nuls certains coefficients de l'estimation de β, ou en limitant leur valeur. Ainsi, l'estimation du vecteur β est telle que :

$$\widehat{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^{p+1}} \left\{ \frac{1}{n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_q^q \right\} \tag{6.8}$$

avec $\lambda > 0$ appelé paramètre de régularisation, et $q \geq 1$ le paramètre d'optimisation.

Selon les valeurs de q nous obtenons certaines pénalités usuelles :

— Si $q = 1$, nous obtenons une régression Lasso. Ce type de régression permet notamment de faire de la sélection de variables en assignant un coefficient égal à zéro à celles qui ne seront pas gardées. Cela permet notamment d'enlever les effets de corrélation en supprimant l'une des variables corrélées ;

— Si $q = 2$, nous obtenons une régression Ridge. Une telle régression limite l'amplitude de la valeur des estimateurs des coefficients en pénalisant les grandes valeurs de coefficients. Cela permet un meilleur contrôle de la variance de l'estimateur.

Pour la plupart des modèles GLM que nous allons réaliser, nous utilisons un nouveau logiciel appelé *Akur8*, qui réalise une pénalisation (gardée confidentielle par le développeur) afin de sélectionner les variables les plus pertinentes pour expliquer la variable réponse.

Indicateurs pour la sélection des modèles

Plusieurs indicateurs de performance des GLM sont à notre disposition. Ils sont surtout utiles pour comparer les modèles entre eux plus que pour l'évaluation individuelle des modèles.

Le critère d'information d'Akaike (AIC)

L'AIC permet d'évaluer le niveau de perte d'informations en fonction du nombre de variables explicatives dans le modèle. Pour cela, son calcul se base sur la fonction de vraisemblance du modèle. L'idée sur laquelle repose l'AIC est le principe de parcimonie (ou rasoir d'Occam), c'est-à-dire trouver le modèle qui satisfait le mieux le compromis entre le nombre de variables explicatives et la perte minimale d'informations par le modèle. L'AIC se calcule selon la formule :

$$AIC = 2k - 2\ln(L)$$

avec k le nombre de paramètres à estimer dans le modèle, et L le maximum de la fonction de vraisemblance du modèle.

Sur différents modèles entraînés, le meilleur modèle au sens du critère d'information d'Akaike est celui présentant l'AIC le plus faible.

Le critère d'information bayésien (BIC)

Le BIC dérive de l'AIC, mais en plus de dépendre du nombre de variables explicatives du modèle, il dépend également de la taille de l'échantillon. Le BIC se calcule de la manière suivante :

$$BIC = -2\ln(L) + k\ln(N)$$

avec k le nombre de paramètres à estimer dans le modèle, L le maximum de la fonction de vraisemblance du modèle, et N la taille de l'échantillon.

De même, le meilleur modèle est celui qui minimise le critère de BIC.

L'indice de Gini

La sélection du modèle peut aussi se faire grâce à l'indice de Gini (détail en Annexe 5) : plus ce dernier est proche de 1, meilleur est le modèle.

Annexe 5 : ROC, AUC et indice de Gini [9]

La courbe ROC : *Receiver-Operator Characteristic*

Dans le cadre d'un algorithme de classification, il est possible d'évaluer la qualité de la prédiction de l'étiquette de classe grâce à la courbe ROC, qui repose sur le principe du *un contre tous*, lequel détermine, pour chaque classe, si un individu appartient ou non à cette classe. En effet, pour prédire l'étiquette associée à une observation, une fonction de décision détermine à partir d'un certain seuil, si l'observation peut être associée à raison à la classe considérée. Quatre cas de figure se distinguent alors :

— **TP** : *True Positives*, *ie* les vrais positifs. C'est le nombre d'observations appartenant à la classe d'intérêt qui sont effectivement prédites comme appartenant à la classe d'intérêt.

— **TN** : *True Negatives*, *ie* les vrais négatifs. C'est le nombre d'observations n'appartenant pas à la classe d'intérêt qui sont effectivement prédites comme n'appartenant pas à la classe d'intérêt.

— **FP** : *False Positives*, *ie* les faux positifs. C'est le nombre d'observations n'appartenant pas à la classe d'intérêt mais qui sont pourtant prédites comme appartenant à la classe d'intérêt.

— **FN** : *False Negatives*, *ie* les faux négatifs. C'est le nombre d'observations appartenant à la classe d'intérêt mais qui sont prédites comme n'appartenant pas à la classe d'intérêt.

Nous pouvons schématiquement représenter ces résultats de prédiction dans une matrice de confusion :

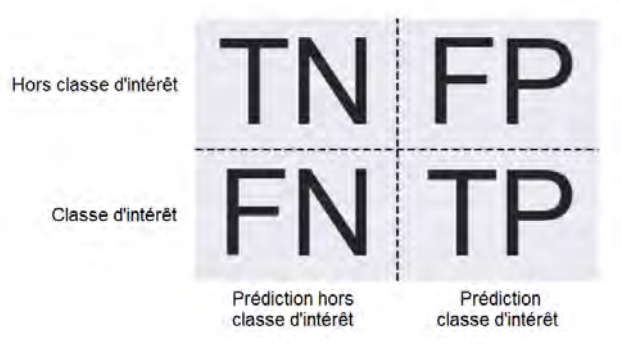


FIGURE A.1 – Matrice de confusion

Nous pouvons alors définir les mesures suivantes :

- La **sensibilité**, ou **taux de vrais positifs** : c'est le nombre de vrais positifs sur l'ensemble des observations positives :

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

- La **spécificité** : c'est le nombre de vrais négatifs sur l'ensemble des observations négatives :

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

- Le **taux de faux positifs** : c'est le nombre de faux positifs sur l'ensemble des observations négatives :

$$\text{TFP} = 1 - \text{Spécificité} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

Suivant la valeur de seuil employée dans la fonction de décision, ces taux évoluent, ce qui permet d'obtenir la **courbe ROC** qui trace le **taux des vrais positifs en fonction des faux positifs** quand la valeur du seuil de décision varie. En effet, le fait de diminuer cette valeur de seuil va identifier davantage d'individus comme

9. D'après *Le Machine Learning avec Python* [8]

positifs, ce qui augmentera inévitablement à la fois le nombre de vrais positifs et le nombre de faux positifs, et donc les taux associés.

La figure A.2 est un exemple d'une telle courbe. Nous distinguons d'abord la bissectrice qui illustre le fait que le modèle associé n'est pas capable de mieux distinguer les faux des vrais positifs par rapport au hasard, et une courbe au dessus de la bissectrice indiquant une capacité de discrimination meilleure que le hasard.

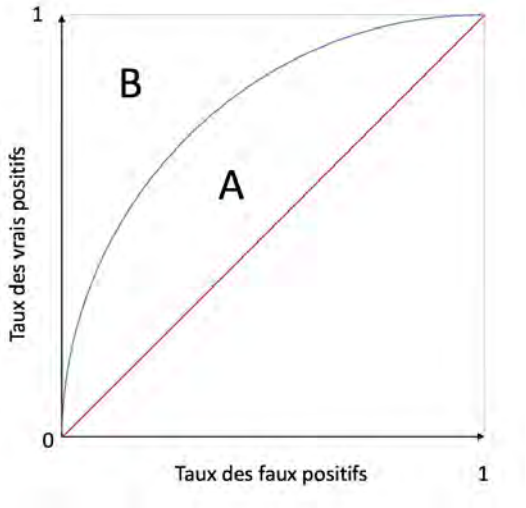


FIGURE A.2 – Exemple d'une courbe ROC

Plus cette courbe sera proche de l'axe des ordonnées et du haut de la figure, formant alors un angle droit dans le coin supérieur gauche du graphique, meilleur sera le modèle en terme de classification entre les positifs et les négatifs. En effet, cela correspond au cas où il n'y a aucun faux positif.

L'AUC : *Area Under the Curve*

L'AUC, ou *Area Under the Curve*, ou encore **l'aire sous la courbe ROC**, est un indicateur de performance et est égal à la probabilité que le modèle classe en positifs plutôt qu'en négatifs les individus réellement positifs. C'est un indicateur de la qualité de la bonne classification des individus positifs, sans pour autant juger de la qualité de la prédiction des individus négatifs. Ainsi, l'AUC est surtout **utile pour comparer des modèles de classification** plus que pour évaluer un modèle seul.

Sur la figure A.2, l'AUC pour la courbe bleue est égale à $2A + B$.

L'AUC peut prendre des valeurs comprises entre 0,5 (pire des cas où le modèle ne classifie pas mieux que le hasard) et 1 (meilleur des cas où le modèle ne classe aucun faux positif).

L'indice de Gini

L'indice de Gini est également un indicateur de performance du modèle et provient directement de l'AUC et se calcule selon la formule :

$$2\text{AUC} - 1 \tag{6.9}$$

Il s'agit en fait du double de l'aire entre la courbe ROC et la bissectrice. Dans le cas de la figure A.2, l'indice de Gini pour la courbe bleue est égal à $2A$.

L'indice de Gini peut prendre des valeurs comprises entre 0 (pire des cas) et 1 (meilleur des cas).

Annexe 6 : Problème d'optimisation sous contraintes [10]

Problème primal, définition du Lagrangien et de la fonction duale

Considérons le problème primal d'optimisation suivant sous contraintes d'inégalités :

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.c.} \quad & g_i(x) \leq 0, \quad i = 1, ..., m \end{aligned} \tag{6.10}$$

Notons $f^* = f(x^*)$ la valeur optimale de la fonction de décision f sous la contrainte d'inégalité, obtenue au minimum global x^*.

Nous définissons le Lagrangien du problème d'optimisation par la fonction $L : \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}$ telle que :

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \tag{6.11}$$

avec λ_i le multiplicateur de Lagrange associé à la contrainte $g_i(x) \leq 0$, et λ le vecteur $(\lambda_1, ..., \lambda_m)$.

Nous définissons également la fonction duale $q : \mathbb{R}^{\geqslant} \longrightarrow \mathbb{R}$ telle que :

$$q(\lambda) = \inf_{x \in \mathbb{R}^n} L(x, \lambda) \tag{6.12}$$

$$= \inf_{x \in \mathbb{R}^n} \left(f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \right) \tag{6.13}$$

A noter que si L n'a pas de borne inférieure en x, alors $q(\lambda) = +\infty$. De plus, q présente deux propriétés importantes pour la résolution du problème d'optimisation :

1. ***q* est une fonction concave**, et ce, même si le Lagrangien n'est pas convexe ;
2. Quelque soit $\lambda \geq 0$, alors la valeur de la fonction duale sera toujours inférieure ou égale à la valeur optimale de la fonction de décision : $\boldsymbol{q(\lambda) \leq f^*}$

Preuve du 1 : Quelque soit x, la fonction $\lambda \mapsto L(x, \lambda)$ est linéaire. Elle est donc concave et convexe en λ. Or le lieu des points du minimum d'une fonction concave est concave. Ce résultat repose sur le fait que l'intersection de domaines concaves est concave. En effet, soit h une fonction telle que pour tout $y \in \mathcal{A}$, $h(x, y)$ est concave en x, et posons $v(x) = \inf_{y \in \mathcal{A}} h(x, y)$, alors pour tout $\alpha \in [0, 1]$ et x_1, x_2 appartenant au domaine de définition de v :

$$\begin{aligned} v(\alpha x_1 + (1-\alpha)x_2) &= \inf_{y \in \mathcal{A}} h(\alpha x_1 + (1-\alpha)x_2, y) \\ &\geq \inf_{y \in \mathcal{A}} (\alpha h(x_1, y) + (1-\alpha)h(x_2, y)) \qquad \text{car } h \text{ est concave} \\ &\geq \inf_{y \in \mathcal{A}} (\alpha h(x_1, y) + \inf_{y \in \mathcal{A}} ((1-\alpha)h(x_2, y)) \end{aligned}$$

Alors nous obtenons :

$$v(\alpha x_1 + (1-\alpha)x_2) \geq \alpha v(x_1) + (1-\alpha)v(x_2)$$

Ainsi v est bien concave.

10. D'après le cours des Mines de Jean-Philippe Vert [12]

Preuve du 2 : Soit $\bar{x}$ n'importe quel point x satisfaisant la contrainte d'inégalité $\forall i\ g_i(\bar{x}) \leq 0$. Alors :

$$\begin{aligned} \sum_{i=1}^{m} \lambda_i g_i(\bar{x}) &\leq 0, \qquad \forall \geq 0 \\ \Rightarrow L(\bar{x}, \lambda) = f(\bar{x}) + \sum_{i=1}^{m} \lambda_i g_i(\bar{x}) &\leq f(\bar{x}) \\ \Rightarrow q(\lambda) = \inf_{x \in \mathbb{R}^n} L(x, \lambda) \leq L(\bar{x}, \lambda) &\leq f(\bar{x}) \qquad \forall \bar{x} \end{aligned}$$

Problème dual

Rappelons que le problème primal s'écrit :

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.c.} \quad & g_i(x) \leq 0, \quad i = 1, ..., m \end{aligned}$$

Le problème dual associé s'écrit alors :

$$\begin{aligned} \max_{\lambda} \quad & q(\lambda) \\ \text{s.c.} \quad & \lambda_i \geq 0, \quad \forall i = 1, ..., m \end{aligned}$$

Il s'agit d'un problème d'optimisation convexe puisque nous cherchons à maximiser une fonction objective q concave.

Dualité faible

Notons d^* la valeur optimale de la fonction objective du problème dual q. Etant donné que pour tout $\lambda \leq 0$, $q(\lambda) \leq f^*$, alors l'inégalité de la dualité faible est toujours vraie si d^* et f^* ont des valeurs finies :

$$d^* \leq f^* \tag{6.14}$$

La différence $f^* - d^*$ s'appelle un saut de dualité. Cela implique que le problème dual fournit une borne inférieure à la valeur optimale de la fonction objective f, ce qui peut être une information utile.

Dualité forte

Le problème d'optimisation présente une dualité forte lorsque :

$$d^* = f^* \tag{6.15}$$

Dans ce cas, résoudre le problème primal équivaut à résoudre le problème dual et le point x^* vérifiant cette égalité est un optimum global du problème primal. Cela est rendu possible si le problème primal est convexe (c'est-à-dire que la fonction objective f et la contrainte g sont convexes et de classe C^1), et si les conditions de Slater sont vérifiées, c'est-à-dire s'il existe x^* qui vérifie strictement la contrainte d'inégalité :

$$\exists\ x^* \text{ tel que } \forall i \quad g_i(x^*) < 0 \tag{6.16}$$

Dans ce cadre de dualité forte, nous remarquons également que les conditions de Karush-Kuhn-Tucker (KKT) sont des conditions nécessaires et suffisantes pour que le point optimum x^* soit un optimum global.

Conditions de Karush-Kuhn-Tucker (KKT)

Les conditions nécessaires KKT d'optimalité :

Hypothèses : f et g_i, pour tout $i = 1, ..., m$ sont de classe C^1.

Soit x^* une solution du problème d'optimisation 6.10, avec x^* un point régulier. Alors il existe $\lambda^* \in \mathbb{R}^m$ tel que :

$$\forall i \quad \lambda_i^* g_i(x^*) = 0, \quad g_i(x^*) \leq 0, \quad \lambda_i^* \geq 0 \tag{6.17}$$

$$\nabla f(x^*) + \sum_i^m \lambda_i^* \nabla g_i(x^*) = 0 \tag{6.18}$$

Les conditions nécessaires et suffisantes KKT d'optimalité :

Si nous rajoutons aux hypothèses le fait que f et g_i, pour tout $i = 1, ..., m$ soient convexes, alors les conditions citées précédemment deviennent nécessaires et suffisantes : si ces conditions sont vérifiées, alors x^* est un optimum global, et nous sommes dans un cas de dualité forte.

Annexe 7 : Corrélation et V de Cramer

Calcul de la corrélation

L'étude de la corrélation entre deux variables X et Y continues permet de déterminer si ces dernières sont liées linéairement ou non. Cette mesure donne une information sur l'intensité de la relation ainsi que le sens de la relation linéaire entre les variables. Le coefficient de corrélation $\rho_{X,Y}$ est une valeur comprise entre -1 et 1 et se calcule selon la formule :

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \rho_{Y,X}$$

Si $\rho_{X,Y}$ est proche de 1, alors la relation linéaire entre X et Y est positive et forte. S'il est proche de -1, la relation linéaire est négative et forte. Enfin, s'il est proche de zéro, alors les variables sont dites décorrélées et il n'y a pas de relation linéaire entre les deux variables.

Calcul du V de Cramer

La mesure de l'intensité de la relation entre deux variables catégorielles se fait grâce au calcul du V de Cramer, qui prend des valeurs entre 0 et 1. Soit X une variable catégorielle présentant k catégories, et Y une autre variable de même nature avec r catégories. Si l'on dispose d'un échantillon de taille N, alors le V de Cramer se calcule de la façon suivante :

$$\forall i \in [\![1;k]\!] \qquad \text{et} \qquad \forall j \in [\![1;r]\!]$$

$$V = \sqrt{\frac{\chi^2}{N.\min(k-1, r-1)}} \qquad \text{avec} \qquad \chi^2 = \sum_{i,j} \frac{\left(N_{i,j} - \frac{N_i N_j}{N}\right)^2}{\frac{N_i N_j}{N}}$$

avec $N_{i,j}$ le nombre de fois où le couple (X_i, Y_j) a été observé.

Lorsque le V de Cramer est inférieur à 0.10, alors la relation entre X et Y est considérée faible voire nulle. Si le V de Cramer est supérieur à 0.30, alors la relation est forte.

Annexe 8 : Validation croisée stratifiée [11]

Au début de la modélisation d'un algorithme de classification, il est de coutume de diviser de manière aléatoire la base de données en deux parties : une base d'apprentissage (*train*) et une base de test, moins grande que celle de *train*.

Or, l'apprentissage de l'algorithme de classification peut très bien dépendre de cette partition et peut donc être instable si cette partition change. C'est pourquoi il est souhaitable d'évaluer la stabilité du modèle en utilisant la méthode de *cross validation* qui va, à partir de la base de données, partager les données en k échantillons, chaque échantillon servant tour à tour de test après avoir généré l'apprentissage sur les k-1 autres échantillons. Ci-dessous, en figure A.3 un exemple de cross validation à 5 plis :

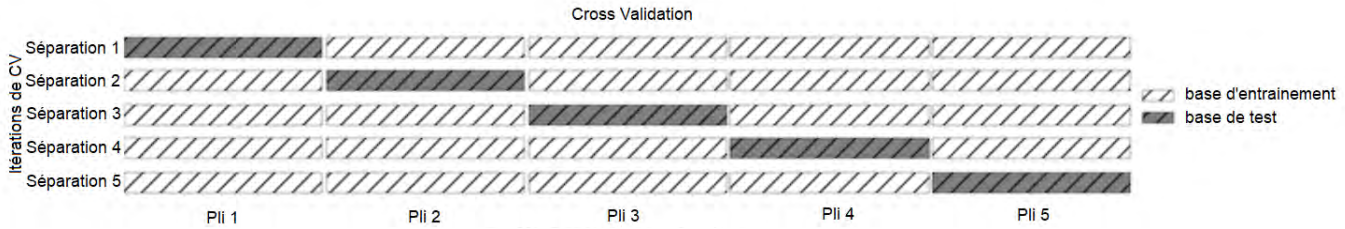


FIGURE A.3 – Schéma du principe de la validation croisée

La validation croisée stratifiée, ou *cross validation* permet de déterminer à la fois la qualité moyenne de prédiction et permet aussi de déterminer la stabilité d'un modèle (en regardant si les précisions données pour chaque pli (communément appelé *fold*) de test sont assez proches.

Définition de la validation croisée

Etant donné un jeu D de n observations, et un nombre K , nous appelons validation croisée la procédure qui consiste à :

— Partitionner D en K parties de tailles sensiblement similaires : $D_1,\ D_2,\ ...,\ D_K$;

— Pour chaque valeur de $k = 1, ..., K$:

- entraîner un modèle sur $\cup_{l \neq k} D_l$
- évaluer ce modèle sur D_k

Chaque partition de D en deux ensembles D_k et $\cup_{l \neq k} D_l$ est appelée "fold" (ou pli) de la validation croisée. Chaque observation étiquetée (c'est-à-dire faisant référence à une variable réponse donnée) du jeu D appartient à un unique jeu de test, et à $(K1)$ jeux d'entraînement. Ainsi, cette procédure génère une prédiction par observation de D .

Il faut toutefois se prémunir du fait que certains plis regroupent à eux seuls une unique étiquette de variable réponse, sans représenter les autres types d'étiquettes, car cela aurait un impact négatif sur l'évaluation de la stabilité de notre modèle. Nous utilisons pour cela la validation croisée stratifiée.

Définition de la validation croisée stratifiée

Une validation croisée stratifiée est une cross validation faisant en sorte que la proportion des différentes étiquettes de variables réponse soit la même dans chacun des plis, autrement dit la moyenne des étiquettes des observations est sensiblement la même dans chacun des K sous-ensembles D_k. Ainsi, dans le cas d'un problème

11. D'après *Le Machine Learning avec Python* [8]

de classification, cela signifie que la proportion d'exemples de chaque classe est la même dans chacun des D_k . Cette proportion est aussi la même que dans le jeu de donnée D complet.

Glossaire

Affaire nouvelle : Il s'agit d'une nouvelle souscription de contrat.

Apport net : Différence entre le volume d'affaires nouvelles et le volume de contrats résiliés.

Competitive Market Analysis : En fraçais : analyse de marché concurrentiel. Cette analyse consiste à étudier la concurrence et comment un acteur du marché se positionne en fonction des autres.

Crédit commercial : Il s'agit d'un rabais commercial qu'un agent général peut accorder à un contrat afin d'en réduire la prime commerciale.

Expected Loss Ratio : Rapport de la prime pure prédite sur laprime commerciale. L'expected loss ratio est une estimation du rapport des sinistres sur la prime commerciale.

Indice FFB : L'indice de la Fédération Française du Bâtiment du coût de la construction est un indice trimestriel calculé à partir du prix de revient d'un immeuble de rapport de type courant à Paris. Il enregistre les variations de coût des différents éléments qui entrent dans la composition de l'ouvrage. Ce calcul ne prend pas en compte la valeur des terrains.

Prime pure : La prime pure représente le risque probable de sinistre, et matérialise l'espérance de coût des sinistres sur l'année à venir.

Reverse engineering : En français : rétro-ingénierie. C'est l'examen et l'analyse en détail d'un processus afin d'en comprendre le fonctionnement et éventuellement le reproduire.

Terme : Dans le contexte de cette étude, le terme signifie l'échéance annuelle du contrat.

Liste des acronymes

AIC : Akaike information criterion. En français : Critère d'Information d'Akaike. Pour plus d'informations, cf. annexe 4.

AUC : Aera Under the Curve. En français : Aire sous la courbe. Pour plus d'informations, cf. annexe 5.

BIC : Bayesian Information Criterion. En français : Critère d'Information Bayésien. Pour plus d'informations, cf. annexe 4.

CMA : Competitive Market Analysis. En français : analyse de marché concurrentiel.

ELR : Expected Loss Ratio.

GBM : Gradient Boosting Machine. Pour plus d'informations, cf. annexe 3.

GLM : Generalized Linear Model. En français : modèle linéaire généralisé. Pour plus d'informations, cf. annexe 4.

IARD : Incendie, Accidents, Risques Divers

MRH : MultiRisque Habitation

MSE : Mean Squared Error. En français : erreur quadratique moyenne.

PP : Prime Pure

ROC : Reciever-Operator Characteristic. Pour plus d'informations, cf. annexe 5.

SHAP : SHapley Additive exPlanations. En français : explication incrémentale par les valeurs de Shapley. Pour plus d'informations, cf. chapitre 5.

Table des figures

Bibliographie

[1] AXA France IARD: *Rapport sur la Solvabilité et la Situation Financière 2018*, 2019. https://entreprise.axa.fr/content/dam/axa-fr-convergence/transverse/informations_financieres/AXA_France_IARD-Rapport_2018_sur_la_solvabilite_et_la_situation_financiere.pdf.

[2] Crucianu Michel, Ferecatu Marin et Thome Nicolas: *Cours sur les arbres de décision*. CNAM, http://cedric.cnam.fr/vertigo/Cours/ml2/coursArbresDecision.html, 2018.

[3] De Larrard Alexandre: *Commercial price optimization strategies in car insurance*. Mémoire d'actuariat réalisé chez AXA Global Direct, 2016.

[4] Delignette-Muller Marie Laure, Dutang Christophe *et al.*: *fitdistrplus : An R package for fitting distributions*. Journal of Statistical Software, 64, 2015.

[5] Friedman Jerome H.: *Greedy Function Approximation : A Gradient Boosting Machine*. Annals of Statistics, 29 :1189–1232, 2000.

[6] Lundberg Scott, Erion Gabriel et Lee Su-In: *Consistent Individualized Feature Attribution for Tree Ensembles*, 2018. https://arxiv.org/pdf/1802.03888.pdf.

[7] Lundberg Scott et Lee Su-In: *A Unified Approach to Interpreting Model Predictions*, 2017. https://arxiv.org/pdf/1705.07874.pdf.

[8] Müller Andreas C. et Guido Sarah: *Le Machine Learning avec Python*. Editions First, Paris, France, 2018.

[9] Rakotomalala Ricco: *Tests de normalité, Techniques empiriques et tests statistiques*, 2011. https://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf, visité le 2019-07-01.

[10] Rouvière Laurent: *Cours de régression logistique avec R*. Université de Rennes, https://perso.univ-rennes2.fr/system/files/users/rouviere_l/poly_logistique_web.pdf.

[11] Shapley L.S.: *A Value for N-person Games*. Defense Technical Information Center. https://books.google.fr/books?id=QzW6tgAACAAJ.

[12] Vert Jean-Philippe: *Nonlinear Optimization : Duality*. Ecole des Mines de Paris, http://members.cbio.mines-paristech.fr/~jvert/teaching/2006insead/slides/4_duality/duality.pdf, 2006.