Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 10/03/2022

Par : **Ilias Kamal**

Titre : **Proposition de leviers d'amélioration du taux de transformation en assurance multirisque habitation**

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

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

Membres présents du jury de la filière

Nom : Pierre Picard

Membres présents du jury de l'Institut des Actuaires

Entreprise : Allianz France

Signature :

Directeur du mémoire en entreprise :

Nom : Omar Bouattour
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

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Le marché de l'assurance multirisque habitation est un marché fortement concurrentiel. L'acquisition d'affaires nouvelles constitue donc un enjeu déterminant pour les compagnies présentes sur ce marché. Par conséquent, ces dernières cherchent d'autant plus à accroître leur taux de transformation. Aussi appelé taux de conversion, cet indicateur mesure la part de devis convertis en contrats. L'objectif de ce mémoire est de proposer plusieurs leviers d'amélioration du taux de transformation sur des catégories rentables de prospects venant du digital.

Les deux premières parties décrivent le contexte et les éléments théoriques de l'étude. Elles sont suivies d'une troisième partie consacrée à l'élaboration de différents modèles de conversion. Ces modèles visent à prédire la probabilité de conversion des devis d'origine digitale. Nous comparons les performances d'une régression logistique et de deux algorithmes de *machine learning* basés sur les arbres de décision. Le meilleur modèle est choisi sur la base de métriques de classification et d'une étude des probabilités prédites. Outre ces modèles élaborés pour le cas général de la conversion des devis digitaux, deux autres modèles plus spécifiques sont entraînés pour l'étude des leviers. Le premier est un modèle visant à capter davantage l'effet du prix, tandis que le second est un modèle spécifique à la conversion des devis transmis en agence.

La quatrième partie porte sur l'identification de profils cibles, dont nous souhaiterions augmenter le taux de transformation. L'indicateur utilisé pour construire les profils est le *Projected* S/C (PSC). Cet indicateur fournit une mesure individuelle et prospective de la rentabilité technique des contrats à l'affaire nouvelle. Des arbres de régression sont utilisés pour segmenter les devis en une liste de profils caractérisés par une valeur de PSC. Une valeur d'équilibre du PSC est choisie afin de sélectionner les profils rentables à cibler.

Dans la cinquième et dernière partie de ce mémoire, nous nous appuyons sur les modèles de conversion afin d'étudier trois leviers d'amélioration du taux de transformation des cibles. Nous étudions en premier lieu l'impact du prix affiché sur le devis. Pour cela, nous dérivons analytiquement d'un des modèles l'élasticité-prix de la conversion. Le second levier proposé est la distribution de mois offerts sur la prime. Après avoir simulé l'impact de ces offres promotionnelles sur la conversion chez les différents profils, nous effectuons une réallocation ciblée des mois gratuits sur le portefeuille. Enfin, nous explorons le levier du choix de l'agent destinataire du devis. Nous définissons plusieurs catégories d'agents susceptibles de recevoir les devis, et estimons les probabilités de conversion associées pour chaque devis. Nous mesurons alors les impacts sur le portefeuille d'une réorientation des devis cibles vers l'agent qui a la plus grande probabilité de les convertir.

Mots-clés : Taux de transformation, multirisque habitation, régression logistique, arbres de décision, machine learning, profil rentable, digital, multi-accès, élasticité.

Abstract

In a highly competitive market such as home insurance, acquiring new business is a major challenge. As a result, the companies operating in this market are all the more interested in increasing their conversion rate. This indicator measures the proportion of quotes converted into contracts. The objective of this dissertation is to propose several levers which enable to improve the conversion rate of profitable categories of prospects coming from the internet.

The first two sections present the context and theoretical tools of the study. Then, the third section is dedicated to the elaboration of different conversion models. These models aim at predicting the conversion probability of digital quotes. We compare the performances of a logistic regression and two decision tree-based machine learning algorithms. The best model is chosen on the basis of classification metrics and a study of the predicted probabilities. In addition to these general conversion models, two other models are more specifically designed for the application of the levers. The first one is designed to capture the price effect, while the second one predicts specifically the conversion rate of the quotes which are sent to an agency.

The fourth section deals with the identification of target profiles on which the levers will be applied. The indicator used to build the profiles is the Projected S/C (PSC). This indicator provides an individual and prospective metric of the technical profitability of new contracts. Regression trees are used to segment the quotes into a list of profiles associated to a PSC value. A threshold is then chosen in order to select the profitable profiles to target.

In the last section, we use the conversion models to study the effect of three levers for improving the conversion rate of targets. Firstly, we study the impact of the quote price. In order to do so, we derive analytically from one of the models the price elasticity of the conversion. The second lever is the distribution of free months. After simulating the impact of these promotional offers on the conversion of different profiles, we perform a targeted reallocation of the free months in the portfolio. Finally, we explore the leverage of choosing the agent receiving the quote. We define several categories of agents and estimate the associated conversion probabilities for each quote. We then assess the impact on the portfolio of redirecting the target quotes to the agent who is most likely to convert them.

Key words : Conversion rate, home insurance, logistic regression, decision trees, machine learning, profitable profile, digital, multiaccess, price elasticity.

Note de synthèse

Le marché de l'assurance habitation est très compétitif. De ce fait, l'acquisition d'affaires nouvelles est un enjeu essentiel pour les compagnies présentes sur ce secteur. Un indicateur clé pour suivre les affaires nouvelles est le taux de transformation. Aussi appelé taux de conversion, il représente la part de devis transformés en affaires nouvelles :

$$\text{Taux de transformation} = \frac{\text{Nombre de devis transformés en contrats}}{\text{Nombre de devis}}$$

L'objectif de ce mémoire est de proposer des leviers permettant d'améliorer le taux de transformation sur des cibles rentables. Cette étude se place dans le cadre de la stratégie de distribution multi-accès d'Allianz. Dans l'univers multi-accès, le prospect peut souscrire en utilisant le canal de son choix parmi le réseau physique des agents généraux, les plateformes téléphoniques et le web. Nous nous concentrons sur les devis d'origine digitale, c'est-à-dire issus d'une simulation tarifaire effectuée sur le site allianz.fr ou un comparateur en ligne.

Les modèles de conversion :

L'objectif des modèles de conversion est de prédire la probabilité qu'un devis soit transformé en contrat. Ils utilisent des variables explicatives et prédisent en sortie la variable réponse Y :

$$Y = \begin{cases} 1 & \text{si le devis est converti} \\ 0 & \text{sinon} \end{cases}$$

La probabilité que nous cherchons à prédire est la probabilité d'appartenance du devis i à la classe des devis convertis :

$$\pi_i = P(Y_i = 1)$$

Le taux de transformation d'un échantillon de N devis est quant à lui donné par la proportion de devis convertis :

$$\tau = \frac{\#(Y = 1)}{N} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

Après un traitement minutieux des données, la base d'étude est séparée en un échantillon d'entraînement et un échantillon de test. Trois modèles sont ajustés sur l'échantillon d'entraînement : un GLM *logit* aussi appelé régression logistique, un *Random Forest* et un *Gradient Boosting*.

Résultats de la modélisation :

Les modèles sont évalués sur l'échantillon de test. Les métriques utilisées sont dérivées de la matrice de confusion. Cette matrice dépend d'un seuil de décision, que nous fixons égal à la valeur du taux de transformation sur la base d'étude. Les trois modèles obtiennent de bonnes performances en termes de classification. La métrique considérée en priorité est l'AUC, qui ne dépend pas du seuil de décision. Les trois modèles obtiennent de bons scores AUC, assez proches, compris entre 76% et 79%. Le *Gradient Boosting* est le modèle qui obtient la meilleure valeur d'AUC : 78,7%.

Dans un second temps, nous nous assurons de la bonne capacité des modèles à prédire la probabilité d'appartenance aux classes. Pour cela, nous découpons les observations par tranches de probabilités prédites et traçons le diagramme de fiabilité des trois modèles. Ce diagramme

confronte les probabilités moyennes de chaque tranche aux taux de conversion observés. Nous effectuons également la comparaison des probabilités prédites aux taux observés par modalité. Ces analyses révèlent un mauvais ajustement du modèle *Random Forest*, malgré ses bons résultats de classification. Les probabilités prédites par le GLM et le *Gradient Boosting* s'ajustent quant à elles assez bien à l'observé. En définitive, la supériorité du *Gradient Boosting* sur toutes les métriques d'évaluation nous conduit à retenir ce modèle.

En complément des modèles précédents, nous entraînons deux modèles spécifiques en vue de l'application des leviers. Le premier est un GLM *logit* visant à capter plus précisément l'influence du tarif proposé sur la conversion. Nous supposons que cet effet diffère en fonction du segment auquel appartient le prospect. Nous effectuons donc un croisement de la prime avec les principales variables qualitatives dans le modèle. Le second modèle est un *Gradient Boosting* visant à prédire la probabilité de transformation des devis digitaux envoyés en agence. Des variables liées à l'agent sont intégrées. Ces modèles atteignent sur leurs échantillons de test respectifs des scores d'AUC de 75,9% et 78,3%. Ils montrent ainsi de bonnes performances de classification, et prédisent par ailleurs assez fidèlement la probabilité de conversion.

Identification des profils rentables :

Une fois les modèles de conversion élaborés, nous cherchons à identifier des cibles rentables parmi les devis. L'indicateur de rentabilité retenu est le PSC, défini par :

$$PSC = \frac{EUL}{AP}$$

où

- EUL (*Expected Ultimate Loss*) est la prime pure projetée à l'ultime
- AP (*Actual Price*) est le montant hors taxe payé par l'assuré

Le PSC est une prédiction du S/C définie à l'échelle du contrat. C'est un indicateur de la rentabilité prospective à l'affaire nouvelle. L'enjeu est de définir une valeur seuil de PSC d'équilibre permettant de juger si un contrat est rentable. Une vision complète de la rentabilité devrait intégrer des aspects supplémentaires, tel que la durée de vie et la multi-détention. Le PSC est ainsi un indicateur limité. Nous tentons de surmonter ces limites en prenant en compte le niveau de prime et l'ancienneté dans le portefeuille, pour le choix du seuil de PSC. Ces indicateurs sont disponibles à l'échelle de quatre macro-segments :

- les locataires d'appartement
- les propriétaires d'appartement
- les locataires de maison
- les propriétaires de maison

Les écarts observés entre ces quatre populations nous conduisent à calculer quatre seuils de PSC. Cette différentiation est motivée par la différence de dilution des coûts fixes entre les quatre segments. L'ancienneté dans le portefeuille est quant à elle prise en compte dans l'ajout d'une tolérance par rapport aux seuils fixés. Cette tolérance est d'autant plus large que l'ancienneté moyenne dans le portefeuille est élevée.

La segmentation des devis en différents profils est effectuée à l'aide d'arbres de régression du PSC. Cette segmentation se fait sur les variables suivantes :

- la qualité juridique et le type de logement
- l'âge de l'assuré
- la CSP de l'assuré

- le nombre de pièces du bien
- l'étage du bien

Nous obtenons 31 profils, chacun associé à une valeur caractéristique de PSC qui traduit sa rentabilité. 17 profils sont ciblés. Ils représentent 44% des devis étudiés. Les cibles regroupent l'ensemble des propriétaires d'appartement, cinq des sept profils locataires d'appartement, un des huit profils locataires de maison et trois des huit profils propriétaires de maison. Une fois les cibles déterminées, nous passons à l'étude des leviers d'amélioration de leur taux de transformation.

Levier tarifaire

Un premier levier essentiel permettant d'agir sur le taux de transformation est le prix. Pour utiliser ce levier, nous cherchons à connaître la sensibilité du prospect au prix. Nous introduisons l'élasticité-prix de la probabilité de conversion, définie par la variation relative de la probabilité de conversion face à une variation de la prime du devis :

$$\varepsilon(p) = -\frac{\Delta\pi(p)/\pi(p)}{\Delta p/p}$$

Nous utilisons le GLM centré sur les effets du prix. En manipulant la formule du *logit*, nous établissons une formule fermée de l'élasticité-prix d'un devis faisant intervenir la prime, la probabilité de conversion estimée par le modèle ainsi que les coefficients du modèle liés au prix :

$$\hat{\varepsilon}(p_i) = -Bp_i(1 - \hat{\pi}(p_i))$$

Les valeurs d'élasticité obtenues sont positives. Avec la convention d'écriture choisie, cela signifie que la probabilité de conversion diminue lorsque le tarif augmente. Par ailleurs, l'élasticité est globalement croissante avec le prix. Par conséquent, elle est la plus élevée chez les profils propriétaires de maison, et la plus faible chez des profils de locataires et propriétaires d'appartement. Cela signifie que pour un niveau de prix de départ plus élevé, les prospects sont globalement plus sensibles à une variation du prix.

Les valeurs d'élasticité obtenues avec cette méthode sont globalement trop faibles pour être utilisées dans le cadre d'une optimisation tarifaire. En effet, l'élasticité moyenne mesurée s'élève à 0,92. Nous utilisons une seconde approche d'estimation de l'élasticité, qui consiste à appliquer différents chocs de prime au modèle. Nous obtenons des valeurs d'élasticité similaires, qui crédibilisent les résultats de la méthode analytique. Nous émettons l'idée que la faible élasticité mesurée est potentiellement valable seulement pour de faibles variations de prix, ou qu'elle découle d'une incapacité du modèle à capter correctement la sensibilité au prix.

Les mois offerts

Le deuxième levier étudié est la distribution de mois offerts. Nous mesurons en premier lieu l'impact sur la probabilité moyenne de conversion des profils, lors du passage d'un état sans aucun mois gratuit à un état avec des mois gratuits proposés à tous les devis du profil. La variation relative induite est la plus élevée chez les profils de propriétaires de maison ou d'appartement. Elle atteint au maximum près de 60% chez les profils cibles de retraités propriétaires de maison, ou d'un appartement principal au rez-de-chaussée ou dernier étage. Les plus faibles impacts sont mesurés chez des profils de locataires d'appartement.

Nous souhaitons illustrer le levier par l'application d'une redistribution ciblée des mois gratuits, en conservant le volume d'offres distribuées dans notre portefeuille de devis. Pour cela, nous cherchons à prioriser les profils a priori les plus sensibles à ce levier. Les profils cibles à prioriser sont ceux susceptibles d'apporter le plus grand volume d'affaires nouvelles supplémentaires. En priorisant de la sorte, nous parvenons à améliorer de 0,74 points le taux de transformation sur l'ensemble des devis ciblés. Par ailleurs, ce gain sur les cibles surcompense la perte d'affaires nouvelles sur les profils hors cible, liée à la redistribution de leurs mois gratuits. Ainsi, nous améliorons le mix entre profils cibles et hors cible parmi les affaires nouvelles, tout en améliorant le taux de conversion global du portefeuille de devis. En définitive, seuls deux profils cibles, les plus volumineux, ont hérité des mois gratuits hors cible lors de réallocation. Ces deux profils ayant des primes assez faibles, nous pouvons supposer qu'ils n'auraient en réalité pas consommé l'intégralité du budget de mois gratuits. Nous aurions donc probablement pu aller encore plus loin dans l'utilisation du levier.

Levier du choix de l'agent

Le dernier levier que nous étudions est l'optimisation du choix de l'agent destinataire du devis. Nous imaginons qu'un devis puisse être envoyé à différents agents A_1, ..., A_n. Le levier consiste à orienter le devis vers l'agent qui a la plus grande probabilité de convertir ce devis en contrat. Nous nous appuyons sur l'un des modèles construits, qui intègre deux variables agent. La combinaison des modalités de ces deux variables définit 12 catégories d'agent, affectant la probabilité de conversion des devis. Nous nous apercevons que le meilleur choix d'agent est propre à chaque devis. Il s'agit tout de même dans un cas sur deux de la catégorie d'agent d'ancienneté comprise entre 13 et 20 ans, dont au moins 2 ans au sein du protocole multi-accès.

Nous cherchons ensuite à déterminer les profils les plus sensibles au choix de l'agent. Pour cela, nous mesurons l'écart de probabilité moyenne de conversion par profil entre deux états :

- état 1 : chaque devis est orienté vers son moins bon agent
- état 2 : chaque devis est orienté vers son meilleur agent

Nous constatons que les profils les plus sensibles à ce changement sont les profils de propriétaires de maison. Leur probabilité moyenne va jusqu'à doubler pour les retraités propriétaires d'une résidence principale.

Nous simulons une réorientation des devis cibles, consistant à leur attribuer systématiquement l'agent qui leur offre la meilleure probabilité de conversion. Cette réorientation apporte un gain de 234 affaires nouvelles, en augmentant le taux de transformation de 16% sur la population ciblée. La répartition entre profils cibles et hors cible en devient plus favorable, passant ainsi d'un rapport 38/62 à un rapport 42/58. La limite principale de ce résultat est l'absence de contrainte appliquée. En effet, le modèle ne nous permet pas de mettre en oeuvre la contrainte géographique d'attribution des devis. Nous aurions également pu envisager une contrainte sur la quantité de devis envoyés à chaque agent.

Conclusion

Grâce aux divers éléments traités dans ce mémoire, nous avons pu proposer et étudier l'efficacité de trois leviers d'amélioration du taux de transformation. L'étude du premier levier (le levier tarifaire) peut être améliorée notamment par la prise en compte de la concurrence, puis poursuivie par une optimisation tarifaire. En actionnant les deux autres leviers, nous sommes parvenus à accroître le volume espéré d'affaires nouvelles sur des segments rentables, sans toucher au prix. Ces travaux offrent ainsi des résultats et des perspectives intéressantes dans le cadre du développement de la stratégie multi-accès.

Executive summary

The home insurance market is very competitive. As a result, new business acquisition is a critical issue for companies operating in this sector. A key indicator for tracking new business is the conversion rate. It represents the share of quotes that are converted into new business :

$$\text{Conversion rate} = \frac{\text{Number of quotes turned into contracts}}{\text{Number of quotes}}$$

The objective of this dissertation is to propose levers to improve the conversion rate on profitable targets. This study is based on Allianz's multiaccess distribution strategy. In the multiaccess universe, the prospect can subscribe using the channel of his choice among the physical network of general agents, the phone and the web. We focus on digital quotes, i.e. quotes which originated on the allianz.fr website or an online comparator.

The conversion models :

The objective of conversion models is to predict the probability of a quote being transformed into a contract. These models use explanatory variables and predict the output Y :

$$Y = \begin{cases} 1 & \text{if the quote is converted} \\ 0 & \text{else} \end{cases}$$

The probability we are trying to predict is the probability of quote i belonging to the class of converted quotes :

$$\pi_i = P(Y_i = 1)$$

The conversion rate of a sample of N estimates is given by the proportion of converted estimates :

$$\tau = \frac{\#(Y = 1)}{N} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

After carefully preprocessing the database, it is divided into a training sample and a test sample. Three models are fitted to the training sample : a GLM, also known as logistic regression, a Random Forest and a Gradient Boosting.

Results of the modeling :

The models are evaluated on the test sample. The metrics used are derived from the confusion matrix. This matrix depends on a decision threshold, which we set equal to the value of the conversion rate on the database. The three models show good performances in terms of classification. The metric considered in priority is the AUC, which does not depend on the decision threshold. The three models obtain good AUC scores, quite close, between 76% and 79%. The Gradient Boosting is the model that obtains the best AUC score : 78.7%.

Then, we check if the models can correctly predict the probability associated with the classes. In order to do so, we cut the observations into bins of predicted probabilities and plot the reliability diagram of the three models. This diagram compares the average probabilities of each bin with the observed conversion rates. We also compare the predicted probabilities to the observed rates by modality. These analyses reveal a poor fit of the Random Forest model, despite its good classification results. The probabilities predicted by the GLM and the Gradient

Boosting fit the observed rates fairly well. Finally, the superiority of the Gradient Boosting over all the evaluation metrics leads us to retain this model.

In addition to the previous models, we train two specific models for the application of the levers. The first one is a logistic regression designed to capture more precisely the effect of the quote price on the conversion. We assume that this effect differs depending on the segment of the prospect. We therefore cross the premium with the main qualitative variables in the model. The second model is a Gradient Boosting model that aims at predicting the conversion probability of digital quotes sent to agencies. Variables related to the agent are included. These models reach AUC scores of 75.9% and 78.3% on their test samples. They thus show good classification performances, and besides predict fairly accurately the conversion probability.

Profitable profiles identification :

Once the conversion models have been developed, we try to identify profitable targets among the quotes. The profitability indicator used is the PSC, defined as :

$$PSC = \frac{EUL}{AP}$$

where :

- EUL (Expected Ultimate Loss) is the ultimate pure premium
- AP (Actual Price) is the pre-tax amount paid by the client

The PSC is a prediction of the loss ratio defined at the contract level. It is an indicator of the prospective profitability of new business. The challenge is to define a PSC threshold in order to judge whether a contract is profitable. A complete vision of profitability should integrate additional aspects, such as the lifetime of the contract and whether the client is a multi-holder. The PSC is thus a limited indicator. We attempt to overcome these limitations by taking into account the premium level and the age of the contracts in the portfolio, when selecting a PSC threshold. These indicators are available at the level of four macro segments :

- apartment renters
- apartment owners
- house tenants
- house owners

The gap between these four populations leads us to calculate four PSC thresholds. The differentiation applied is motivated by the difference in fixed cost dilution between the four segments. The average age in the portfolio is taken into account by adding a tolerance margin to the thresholds : the higher the average age in the portfolio, the wider the margin.

The segmentation of the quotes into different profiles is carried out by using regression trees on the PSC. This segmentation is done on the following variables :

- occupation status and type of property
- age of the insured
- profession of the insured
- number of rooms in the property
- floor of the property

We obtain 31 profiles, each one associated with a characteristic PSC value which reflects its profitability. 17 profiles are targeted, which account for 44% of the quotes. The targets include all apartment owners, five of the seven apartment renter profiles, one of the eight home renter

profiles, and three of the eight home owner profiles. Once the targets have been determined, we move on to the levers allowing to improve their conversion rate.

Price :

A first essential lever to act on the conversion rate is the price. To use this lever, we try to know the prospect's sensitivity to the price. We introduce the price elasticity of the conversion probability, defined as the relative variation in the conversion probability caused by a relative variation in the quote premium :

$$\varepsilon(p) = -\frac{\Delta\pi(p)/\pi(p)}{\Delta p/p}$$

We use the GLM centered on price effects. By manipulating the logit formula, we derive a closed-form formula for the price elasticity of a quote involving the premium, the model-estimated conversion probability and the price-related coefficients of the model :

$$\hat{\varepsilon}(p_i) = -\hat{B}_i p_i (1 - \hat{\pi}(p_i))$$

The obtained elasticity values are positive. With the chosen writing convention, this means that the conversion probability decreases when the premium increases. Moreover, the elasticity is globally increasing with the price. Consequently, it is highest for homeowner profiles, and lowest for renter and apartment owner profiles. This means that for a higher starting price, prospects are globally more sensitive to a price variation.

The elasticity values obtained with this method are globally too low to be used in a price optimization. Indeed, the average elasticity is 0.92. We try another approach consisting in estimating the elasticity by performing different premium shocks using the model. We obtain similar elasticity values, which supports the results of the analytical method. We suggest that the low elasticity is potentially valid only for small price changes, or that it arises from an inability of the model to correctly capture the price sensitivity.

Free Months

The second lever studied is the distribution of free months. We first measure the impact on the average conversion probability of the profiles, when moving from a state without any free months to a state with free months offered to all the profile's quotes. The relative variation induced can reach nearly 60% for some profiles. It is highest for house and apartment owner profiles. The smallest impacts are measured for apartment renter profiles.

We then proceed to apply a targeted redistribution of free months, while maintaining the number of offers distributed in our quote portfolio. In order to do so, we seek to prioritize the profiles that are most sensitive to this lever. The target profiles to be prioritized are those likely to bring the most additional contracts. By doing so, we manage to improve the conversion rate by 0.74 points on the targeted quotes. In addition, this gain on targets overcompensates for the loss of new business on non-target profiles, linked to the redistribution of their free months. Thus, we manage to improve the mix between target and non-target profiles among new business, while improving the overall conversion rate of the portfolio. In the end, only two target profiles, the largest ones, inherited the free months upon reallocation. Since these two profiles had relatively low premiums, we can assume that they would not have actually

consumed the entire free months budget. So we probably could have gone even further in using the leverage.

Choosing the agent

The last studied lever consists in optimizing the choice of the agent receiving the quote. We imagine that a quote can be sent to different agents A_1, ..., A_n. The lever consists in directing the quote to the agent that has the highest probability of converting this quote into a contract. We rely on one of the constructed models, which incorporates two agent variables. The modalities of these two variables define 12 agent categories. We show that the best agent is specific to each quote. However, it is half of the time the agent category with seniority between 13 and 20 years, including at least 2 years within the multi-access protocol.

We then try to find out which profiles this lever has the most impact on. To do so, we measure the average conversion probability difference per profile between two states :

- state 1 : each quote is sent to its worst agent
- state 2 : each quotation is sent to its best agent

We observe that the most sensitive profiles to this change are the homeowner profiles. Their average conversion probability is even doubled for retired homeowners.

Finally, we simulate a redirection of the target quotes, consisting in systematically assigning them to the agent that offers them the best chance to subscribe. This redirection generates 234 new contracts, thus increasing the conversion rate of the target population by 16%. The distribution between target and non-target profiles also improves, going from a 38/62 ratio to a 42/58 ratio. The main limitation of this result is the lack of constraint. Indeed, the model does not allow us to implement the geographical allocation constraint. We could also have considered a constraint on the quantity of quotes sent to each agent.

Conclusion

Thanks to the various elements discussed in this dissertation, we have been able to propose and study the effectiveness of three levers for improving the conversion rate. The study of the first lever (the price) can be improved by taking into account the competition, and continued by a price optimization. By activating the other two levers, we were able to increase the expected new business volume on profitable segments, without acting on the price. This work thus shows some interesting results and perspectives in the context of the development of the multiaccess strategy.

Remerciements

Avant toutes choses, je tiens à remercier l'ensemble des personnes ayant directement ou indirectement contribué à la réalisation de ce mémoire. Je remercie avant tout mon maître de stage Omar Bouattour pour le suivi et les conseils précieux qu'il a pu me prodiguer tout au long de l'élaboration de ce mémoire. Il fut d'une aide précieuse dans les moments les plus délicats.

Je tiens également à remercier les membres de la squad Pilotage du Multi-accès d'Allianz France : Christine De Gandt, David Jaouen et Mamihasina Rakotobe pour m'avoir accueilli et donné de leur temps pour me permettre de mener à bien ce mémoire. Je salue et remercie également ma co-stagiaire Aicha Korovaev pour la qualité de nos échanges et l'entraide que nous avons pu développer.

Mes remerciements vont à l'équipe pédagogique de la filière Actuariat de l'ENSAE Paris pour la qualité des enseignements prodigués, qui m'ont permis d'acquérir des connaissances notamment mises à profit dans la rédaction de ce mémoire. Je souhaite remercier en particulier monsieur Pierre Picard pour ses conseils et remarques éclairantes liées à mon sujet.

Merci enfin à mes parents et mes proches pour leurs encouragements et leur soutien.

Table des matières

Note concernant la confidentialité

Les éléments chiffrés confidentiels de ce mémoire ont été modifiés.

Introduction

Le marché de l'assurance habitation représente la seconde source de chiffre d'affaires de l'assurance de biens des particuliers. Ce marché est investi par un grand nombre d'acteurs, allant des compagnies aux mutuelles sans oublier les bancassureurs et les assurtechs. De plus, grâce à internet, les consommateurs ont plus facilement accès à un panel d'offres et n'hésitent pas à faire jouer la concurrence. Tout ceci rend le marché de l'assurance habitation très compétitif. Dans un tel contexte, l'acquisition d'affaires nouvelles constitue un enjeu particulièrement déterminant pour les compagnies.

Misant sur la multiplicité des canaux de contact des clients potentiels, Allianz France a déployé un modèle de distribution multi-accès sur son produit d'assurance habitation. Le client peut entrer en contact avec la compagnie non seulement par une agence, mais également par téléphone et internet. De plus, il peut basculer d'un canal à l'autre. Le multi-accès est un levier d'accès à de nouveaux clients, par le biais du digital. Le nombre de devis réalisés en ligne est en effet conséquent. L'objectif pour la compagnie est alors de convertir la plus grande part de ces devis. Cet objectif est traduit par l'augmentation du taux de transformation. Aussi appelé taux de conversion, cet indicateur correspond à la proportion de devis concrétisés en affaires nouvelles.

Néanmoins, la croissance du taux de transformation ne doit pas se faire au détriment de la rentabilité. En effet, il est dans l'intérêt de la compagnie de maîtriser la qualité de ses affaires nouvelles, en cherchant à cibler les profils les plus rentables pour une croissance du taux de transformation.

L'objectif de ce mémoire est de proposer plusieurs leviers d'amélioration du taux de transformation. Le produit étudié est le produit d'assurance multirisque habitation d'Allianz France. Nous ciblerons les profils de prospects les plus rentables, parmi les devis issus d'une simulation tarifaire en ligne. La question de l'augmentation du volume d'affaires nouvelles est souvent traitée sous l'angle de la tarification. Le prix est en effet le levier naturel, il s'agit donc du premier levier que nous proposerons. Cependant, nous le compléterons par deux autres leviers visant à augmenter le taux de transformation des cibles sans toucher au prix : la distribution d'offres promotionnelles et le choix de l'agent.

La première étape est la construction d'une base servant de support à l'étude. Une fois la base construite, le reste de l'étude s'articulera en trois grandes étapes :

En premier lieu, il est nécessaire de développer un ou plusieurs modèles de conversion. Ces modèles chercheront à prédire le plus fidèlement possible la probabilité de conversion des devis issus du digital. Cette partie nécessite la plus grande attention. En effet, elle conditionne la bonne application des leviers.

Dans un second temps, nous chercherons à cibler les profils rentables de la base d'étude. Nous devrons pour cela élaborer une méthodologie de segmentation des devis en profils, associés à différents niveaux de rentabilité. Une fois les profils construits, il nous faudra établir un critère de sélection des profils à cibler pour l'application des leviers.

Enfin, nous pourrons simuler l'impact des différents leviers sur la conversion chez les différents profils. Cette dernière partie permettra de répondre à la problématique, en mettant en application les modèles de conversion élaborés et le ciblage des profils. Nous analyserons l'effet des différents leviers sur la probabilité de transformation des profils ciblés.

1 Contexte et mise en place de l'étude

Dans cette première partie, nous présentons plus précisément le contexte et la mise en place des travaux. Après avoir précisé le cadre général et les objectifs de l'étude, nous détaillons la construction de la base et les premières analyses des données.

1.1 Contexte

1.1.1 Le marché de l'assurance habitation en France

L'assurance de biens et responsabilité (ABR) offre aux particuliers et aux entreprises une protection contre la perte financière engendrée par un dommage à leurs biens ou causé à autrui. En 2020, ce secteur représentait 60,1 milliards d'euros, soit près de 30% des cotisations d'assurance versées en France.

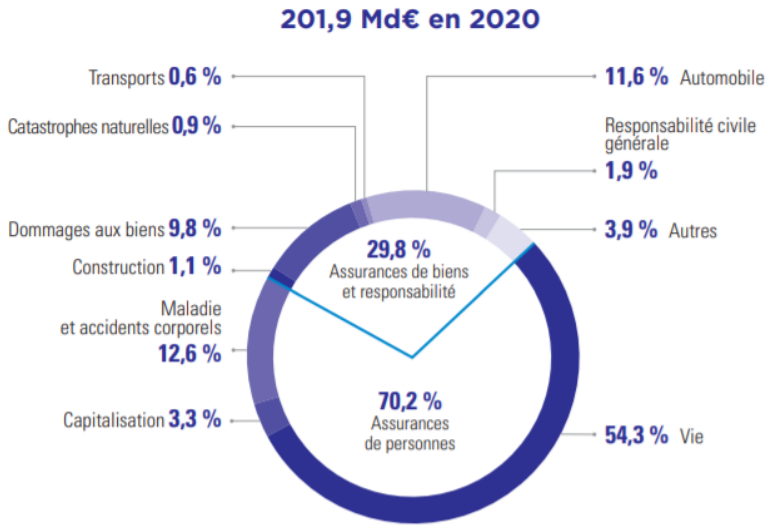

FIGURE 1 – Cotisations de l'assurance en France en 2020 (*source : Bilan 2020 de la FFA*[1])

L'assurance habitation est une branche de l'assurance de biens et responsabilités destinée à couvrir les locaux, leur contenu et la responsabilité civile des occupants. Sur le marché des particuliers, elle représente 29% des cotisations d'assurance de biens et responsabilité (11,3 milliards d'euros en 2020), d'après la Fédération française de l'assurance [1].

Le marché de l'assurance habitation est un marché très mature. La concurrence y est très rude, en raison du nombre important d'acteurs présents sur le marché. Il est donc difficile d'augmenter ses parts de marché tout en étant rentable. En 2015, l'entrée en vigueur de la loi Hamon a offert aux assurés la possibilité de résilier à tout moment leur contrat au-delà d'un an de souscription, en respectant un préavis d'un mois et sans frais. De plus, c'est le nouvel assureur qui se charge de la résiliation du contrat de l'assuré. Ce changement a eu pour effet d'accroître les taux de résiliation, mais également d'offrir aux assureurs l'opportunité de capter de nouvelles affaires.

1. rebaptisée "France Assureurs" en janvier 2022

1.1.2 Le produit Allianz Habitation

Le contrat Allianz Habitation est un contrat d'assurance multirisque habitation (MRH) commercialisé par Allianz France à destination des particuliers. Il couvre les dommages aux locaux d'habitation et la responsabilité civile. Le produit comporte :

- un socle de garanties obligatoires
- des garanties facultatives, déclinées en deux à trois niveaux de couverture
- des renforts de garanties ainsi que des options sur-mesure

Les différentes garanties sont résumées dans le tableau 1 :

Obligatoires	**Facultatives**	**Renforts et options**
incendie	vol/vandalisme	remplacement à neuf
tempête	bris de glace	remboursement d'emprunt
grêle	assistance	pertes pécuniaires
neige		dégâts électriques
attentats		installations extérieures
dégâts des eaux		protection juridique
catastrophes naturelles		assistance voyage
responsabilité civile		scolaire
défense pénale et recours		

Table 1 – Garanties du produit Allianz Habitation

Le contrat est conclu pour une période d'un an puis renouvelé automatiquement chaque année. Il est résiliable à tout moment à partir d'un an dans le cadre de la loi Hamon.

1.1.3 La stratégie Multi-accès

Le modèle traditionnel de la distribution de produits d'assurance repose sur les intermédiaires physiques : les agents généraux et les courtiers. Les agents généraux sont des professionnels indépendants en partenariat avec une société d'assurance qu'ils représentent au sein de leur périmètre géographique. Les courtiers sont mandatés par les clients. Ils leur proposent le contrat le plus adapté à leurs besoins parmi des contrats de plusieurs sociétés d'assurance. Les agents et les courtiers sont rémunérés à la commission.

Depuis les années 1980, un autre modèle a progressivement émergé : le direct. Celui-ci repose sur la vente à distance, sans intermédiaire physique. Initialement porté par la vente par téléphone, le direct a ensuite intégré la vente en ligne. Si la souscription entièrement en ligne reste aujourd'hui minoritaire, le digital a transformé les usages en rendant quasiment systématique l'utilisation d'internet pour consulter les tarifs, notamment sur les comparateurs.

Désormais, les clients souhaitent être en mesure d'entrer en contact avec l'assureur par différents moyens et à n'importe quel moment. Afin de s'adapter à ces nouvelles exigences, Allianz France a mis en place un modèle de distribution multi-accès. Celui-ci permet aux prospects d'utiliser le canal de leur choix parmi le réseau physique des agents généraux, les plateformes téléphoniques et le web. De plus, ils peuvent basculer d'un canal à un autre et ainsi constituer le parcours de leur choix parmi les différentes possibilités illustrées par la figure 2 :

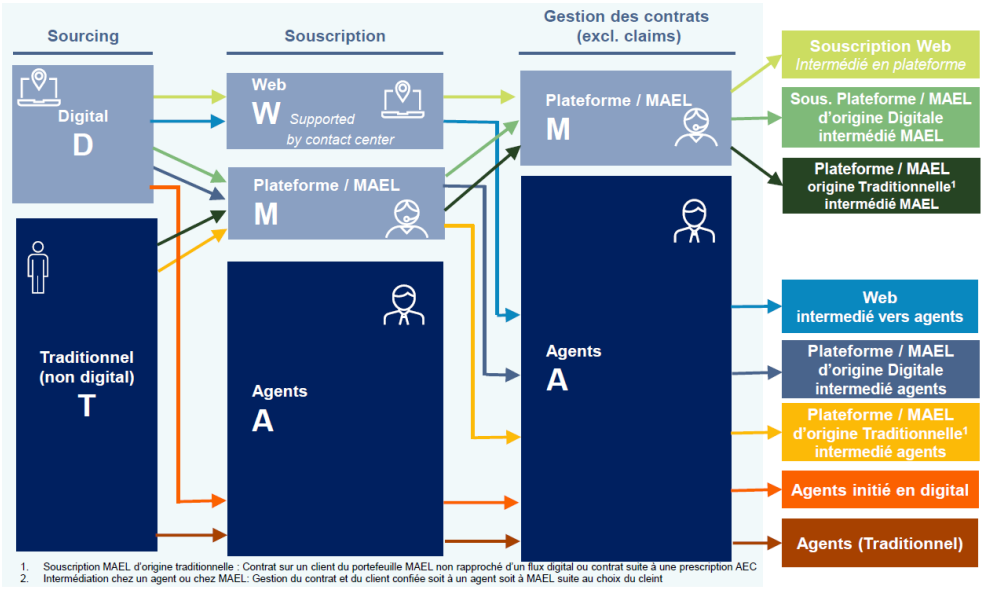

FIGURE 2 – Schéma des parcours multi-accès

Explication de la figure 2 : La première étape d'une souscription est l'acquisition du prospect ou *sourcing*. Il s'agit du point d'entrée. Traditionnellement, le prospect entre en contact avec un agent. Dans le multi-accès, le prospect peut également venir du digital, par le biais d'un comparateur ou du site allianz.fr. Une fois entré, l'étape suivante est celle de la souscription. Le prospect peut souscrire en agence, au téléphone auprès d'un conseiller d'une plateforme [2], ou directement sur internet. Une fois en portefeuille, les clients confient la gestion de leur contrat aux plateformes téléphoniques ou aux agents. En définitive, ces trois étapes forment huit parcours possibles.

Zoom sur la partie digitale :

Nous nous intéressons dans ce mémoire aux prospects issus du bloc de *sourcing* digital. Initialement, le prospect effectue une demande de tarif en ligne directement sur le site allianz.fr ou sur un comparateur.

Sur allianz.fr :

Sur allianz.fr, le prospect commence par renseigner ses informations principales et obtient un premier tarif en moins d'une minute : il s'agit du parcours Fast Quote. Après avoir visualisé ce tarif, il peut choisir entre les actions suivantes :

- être appelé par un conseiller d'une plateforme téléphonique
- rencontrer un agent
- recevoir le tarif par email
- affiner son tarif

En choisissant d'affiner son tarif, le prospect bascule sur un parcours plus long visant à récolter toutes les informations nécessaires : c'est le parcours Normal Quote. Le prospect est invité à renseigner davantage d'informations sur son bien, puis il obtient un tarif plus précis et peut choisir ses garanties. Il dispose alors à nouveau du choix entre être appelé par un conseiller de la plateforme téléphonique, rencontrer un agent, recevoir le devis par email ou souscrire directement en ligne.

2. MAEL (Mon Allianz en Ligne) est le nom du regroupement des plateformes téléphoniques d'Allianz

Sur les comparateurs :

Sur les comparateurs, les prospects visualisent les tarifs proposés par plusieurs assureurs en fonction des informations qu'ils ont saisies. En sélectionnant le tarif Allianz, ils peuvent choisir entre recevoir leur devis par email, être contactés par téléphone ou poursuivre sur le site d'Allianz en Normal Quote.

1.1.4 Les spécificités du digital

Le digital présente plusieurs particularités par rapport au modèle de distribution traditionnel. La première concerne les coûts d'acquisition. De manière générale, une affaire initiée en digital coûte plus cher en raison du coût de référencement sur les comparateurs et les moteurs de recherche.

Une autre différence concerne la probabilité de souscription des prospects. Dans le modèle traditionnel, le prospect se déplace en agence. Cet effort traduit un certain degré d'intention de souscrire un contrat. Intuitivement, la probabilité de souscription des prospects arrivés en agence est donc plus élevée que celle du digital, où le prospect consulte simplement les tarifs sur internet pour éventuellement souscrire par la suite. En outre, les deux populations n'ont pas le même niveau d'information. Les prospects du digital ont facilement accès aux prix de la concurrence, notamment grâce aux comparateurs. Par conséquent, ils sont généralement plus sensibles au prix que la clientèle traditionnelle.

Notons enfin une différence en ce qui concerne la probabilité de résiliation. Celle-ci est en effet plus élevée sur le digital, notamment car cette clientèle est davantage susceptible de résilier pour faire jouer la concurrence. Ceci est problématique, étant donné que les coûts d'acquisition, élevés pour le digital, ne sont pas amortis si la durée du contrat en portefeuille est trop courte. La clientèle traditionnelle est quant à elle moins regardante sur le prix. Une autre raison est que la clientèle traditionnelle est davantage multi-équipée, ce qui limite également le phénomène de résiliation.

1.2 Objectifs de l'étude

L'un des principaux objectifs de la compagnie est d'augmenter son nombre d'affaires nouvelles, tant que cela n'impacte pas son niveau de rentabilité. Cela est d'autant plus difficile à réaliser dans un marché concurrentiel tel que celui de l'assurance habitation. Grâce à son volet digital, l'univers multi-accès permet de démultiplier les demandes de devis. La part de devis convertis en affaires nouvelles peut être suivie à l'aide d'un indicateur appelé taux de transformation, ou taux de conversion. Celui-ci est défini ainsi :

$$\text{Taux de transformation} = \frac{\text{Nombre de devis transformés en contrats}}{\text{Nombre de devis}}$$

L'objectif de ce mémoire est de chercher des leviers permettant d'améliorer le taux de transformation sur le digital, en ciblant uniquement une population déjà rentable. En effet, il n'est pas dans l'intérêt de la compagnie de chercher à attirer davantage les moins bons profils de prospects. Pour répondre à cette problématique, notre étude s'organisera autour de trois objectifs intermédiaires :

• Nous chercherons en premier lieu à élaborer un modèle de conversion. Ce modèle nous permettra de prédire la probabilité de conversion, en fonction des caractéristiques du devis

mais également de celles de l'interaction digitale du prospect avec les comparateurs et le site internet d'Allianz. Au préalable, il faudra construire une base de données contenant les variables nécessaires à la modélisation et servant de support à l'étude.

• Le deuxième objectif sera de parvenir à cibler des profils dont nous souhaitons améliorer le taux de transformation. Pour cela, il faudra tout d'abord convenir d'un indicateur reflétant la rentabilité des devis. Il faudra ensuite construire des profils liés à différents niveaux de rentabilité et cibler les bons profils.

• Enfin, nous proposerons des leviers qui permettent d'améliorer le taux de transformation sur les profils ciblés. Nous chercherons à évaluer l'efficacité de ces leviers sur les devis de notre base d'étude.

1.3 Construction de la base de données

Le premier défi majeur rencontré lors de la réalisation de ce mémoire a été la construction d'une base de données servant de support à l'étude. Les travaux ont été effectués à l'aide du logiciel SAS. Nous détaillons les différentes étapes de construction dans cette section.

1.3.1 Périmètre de l'étude

Tout d'abord, il nous a fallu fixer un périmètre et une période d'observation. Le point de départ d'une potentielle souscription d'origine digitale est une simulation tarifaire effectuée par le prospect sur le site allianz.fr ou un comparateur. Les premières informations issues de cette simulation constituent un flux digital. Lorsque le prospect a suffisamment avancé dans son parcours internet, un devis est généré et le moteur tarifaire lui affiche un prix. Le périmètre de notre étude est formé par les devis digitaux, c'est-à-dire les devis que nous parviendrons à lier à un flux digital.

Le choix de la période d'observation a été fait en tenant compte de plusieurs contraintes. Pour étudier le phénomène de conversion, il est préférable de sélectionner une période récente et assez courte. En effet, nous souhaiterions éviter une mutation du marché qui pourrait impacter le taux de transformation. Cependant, nous souhaitons disposer d'un volume de données suffisant. Il y a donc un arbitrage à effectuer quant à la longueur de la période d'observation. Enfin, il est nécessaire de disposer d'une période de recul pour laisser aux devis les plus récents le temps d'être convertis en contrats. L'étude de la période de développement des devis, que nous détaillons plus loin, révèle le besoin d'observer un recul de deux mois. En raison de ces contraintes, nous avons choisi une période d'observation s'étendant sur deux ans, de juillet 2019 à fin juin 2021.

1.3.2 Structure des données

Compte tenu du mécanisme du digital que nous avons présenté, nous sommes en présence de trois types de données, qui sont stockées dans trois familles de bases :

- les bases flux
- les bases devis
- les bases contrat

Les bases flux :

Les informations d'un flux digital sont stockées dans les bases flux dès qu'un prospect interagit avec le site allianz.fr, ou avec le prix Allianz proposé sur un comparateur. La clé commune à ces bases est le numéro de flux. Les bases de flux contiennent notamment les informations suivantes :

- le numéro et la date du flux
- les coordonnées du prospect (nom, prénom, adresse, numéro de téléphone *etc...*)
- le code état, le parcours, la situation et le code de mise en relation, qui permettent de catégoriser le parcours digital du prospect

Les bases devis :

Les informations du devis sont stockées dans un second ensemble de bases de données : les bases devis. Nous aurons besoin de trois d'entre elles : les bases DEVIPERS, DEVIBIEN et DEVIGAR. La clé commune à ces bases est le numéro de devis. Certaines informations générales du devis sont communes aux différentes bases, par exemple :

- le numéro et la date du devis
- le type de contrat
- la date souhaitée de prise d'effet du contrat

Le reste des informations est réparti entre les différentes bases dont nous aurons besoin.

La base DEVIPERS contient les informations liées au prospect, tel que :

- l'état civil du prospect (nom, prénom, date de naissance, sexe, statut marital)
- la catégorie socio-professionnelle du prospect

La base DEVIBIEN contient les informations liées au logement à assurer. Il s'agit principalement des informations suivantes :

- le type de résidence (principale ou secondaire)
- la qualité juridique (propriétaire, locataire ou autre [3])
- le type d'habitation (appartement, maison ou autre [4])
- l'étage (rez-de-chaussée, dernier ou intermédiaire)
- le nombre de pièces du logement
- les équipements (piscine, alarme, télésurveillance, équipements d'énergies renouvelables, insert, cheminée) ou dépendances du logement
- l'activation de l'offre "petites surfaces", qui propose un prix spécifique pour les logements d'une ou deux pièces des étudiants, jeunes actifs ou enfants d'assurés Allianz.
- le code postal et le code IRIS [5] du bien.

La base DEVIGAR contient les informations liées aux garanties incluses dans le devis. Chaque ligne correspond à une garantie, si bien qu'un même contrat occupe plusieurs lignes de la base. Cette base contient essentiellement :

- la liste des garanties facultatives et des renforts demandés par le prospect
- les primes des différentes garanties

3. occupant à titre gratuit, usufruitier
4. lofts, mobile homes, chalets
5. Le code IRIS est un découpage infracommunale de l'INSEE correspondant approximativement à des zones de population de 2000 habitants.

Les bases contrat :

Les affaires nouvelles sont stockées dans un troisième ensemble de bases : les bases contrat. Ces bases nous serviront uniquement à identifier les devis transformés.

Structure de la base à construire pour l'étude :

L'objectif est de construire une base unique d'étude de la forme suivante : chaque ligne représente un devis et contient les informations du devis, de son flux ainsi qu'une variable indiquant si le devis a été transformé. L'obtention d'une telle structure nécessite le rapprochement de quatre environnements de bases de données :

- les bases flux
- les bases devis
- les bases contrat
- d'autres bases permettant d'enrichir la base d'étude

La construction de la base a nécessité un long travail en deux temps. Dans un premier temps, le socle de la base a été construit en rapprochant les bases de flux et de contrat avec la base de devis DEVIPERS, contenant les informations du prospect. Puis, ce socle a été enrichi par l'ajout de variables issues des autres bases. Ce processus est schématisé sur la figure 3 :

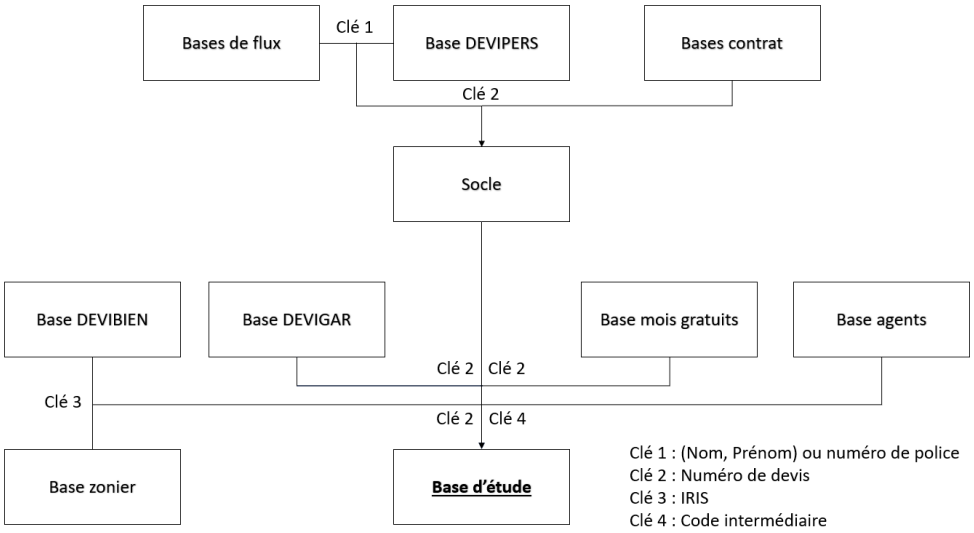

FIGURE 3 – Schéma de construction de la base d'étude

1.3.3 Sélection des flux

Le pivot de construction de notre base est le devis. Or, les bases de devis regroupent tous les types de devis, d'origine digitale ou non. Il est donc nécessaire d'identifier les devis d'origine digitale. Pour cela, il faut dans un premier temps sélectionner les flux inclus dans notre périmètre. Les devis seront ensuite rapprochés aux flux sélectionnés, pour que nous puissions identifier les devis d'origine digitale. La construction de la base d'étude commence donc véritablement par la sélection des flux.

Nous commençons par regrouper dans une seule base les informations utiles contenues dans les différentes bases flux. Puis, nous filtrons ces flux sur la période d'observation et nous conservons uniquement les flux contenant un nom et un prénom renseignés. Nous imposons cette

dernière contrainte afin d'être en mesure d'identifier le devis associé. Puis, nous excluons les types de flux suivants :

- les opérations de robots
- les tests effectués par des agents
- les flux d'un prospect qui est arrivé sur le site puis s'est arrêté sans effectuer d'action

1.3.4 Rapprochement des bases de flux et de devis

Une fois les flux sélectionnés, nous cherchons à identifier les devis d'origine digitale, par le biais d'une jointure entre flux et devis. Le rapprochement des bases de flux et de devis a constitué le plus grand défi de la construction de la base d'étude. En effet, il n'existe pas de clé de jointure permettant d'associer un devis au flux qui l'a généré. Nous effectuons donc la jointure sur les noms et prénoms des individus. Parmi les différentes bases constituant l'environnement des bases devis, ces informations sont stockées dans la base DEVIPERS. C'est donc cette base que nous rapprochons aux flux sélectionnés précédemment. Par ailleurs, les flux comportent parfois un champ "numéro de police" renseigné. Si tel est le cas, c'est qu'ils sont déjà liés à un devis qui porte ce numéro. Nous complétons donc le rapprochement des noms et prénoms par une seconde jointure sur les numéros de police.

Cependant, un même prospect, identifié par son nom et son prénom, effectue souvent plusieurs flux, et génère également plusieurs devis. En rapprochant les flux et les devis par les noms et prénoms, nous obtenons donc dans la jointure toutes les combinaisons possibles d'un flux et d'un devis de la même personne. Il est nécessaire d'appliquer un premier filtre temporel en sortie de la jointure, afin de retirer les combinaisons anachroniques d'un flux associé à un devis antérieur. En effet, un devis ne peut précéder son flux.

Après ce premier filtre, il peut rester des combinaisons d'un flux et d'un devis de la même personne mais qui ne correspondent pas à la même tentative. Il faut donc à présent exclure les combinaisons au-delà d'un certain délai entre le flux et le devis. Afin de choisir cette limite, nous traçons sur la figure 4 la fonction de répartition (tronquée à 70 jours) du délai entre les flux et les devis chez les combinaisons conservées par le premier filtre.

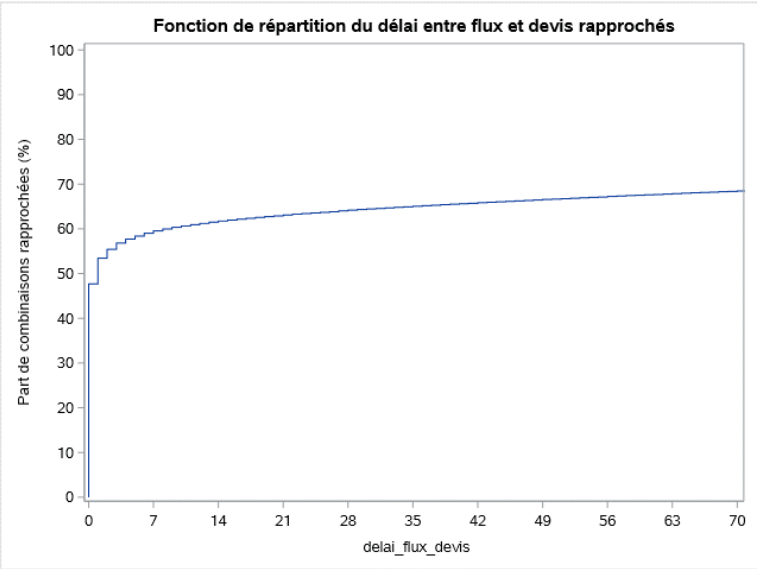

FIGURE 4 – Répartition du délai (en jours) entre les flux et les devis rapprochés

La figure révèle des délais trop longs pour une grande partie des combinaisons. Environ 30% des combinaisons restantes ont même un écart supérieur à 70 jours entre le flux et le devis.

Or, si le flux engendre un devis, cela a en général lieu le même jour. D'après la figure, c'est ce qui est observé sur la moitié des combinaisons. Toutefois, d'autres devis peuvent également être générés quelques jours après le flux, par exemple si l'agent contacte le prospect et édite un nouveau devis. Il ne faut donc pas filtrer trop tôt pour ne pas perdre cet ultime devis, qui est souvent celui converti. Nous fixons donc le délai maximal à 7 jours, qui correspond plus ou moins à l'abscisse à partir de laquelle la courbe commence à se stabiliser.

Enfin, en cas de doublons d'un même devis rattaché à plusieurs flux, nous procédons à un dédoublonnage selon des règles de priorisation des différentes combinaisons. Ces règles visent à conserver le flux le plus susceptible d'avoir généré le devis. Quatre règles sont appliquées successivement :

1. Comme nous l'avons évoqué, le flux comporte parfois un numéro de police. Si tel est le cas, c'est qu'il est déjà lié à un devis qui porte ce numéro. Nous privilégions dans ce cas cette combinaison.
2. Les flux comportent le plus souvent un identifiant de destinataire. Les devis comportent quant à eux toujours un numéro d'intermédiaire. Nous privilégions le flux dont l'identifiant de destinataire correspond à l'identifiant d'intermédiaire du devis, si un tel flux existe.
3. Nous privilégions ensuite le premier flux enregistré.
4. S'il reste plusieurs flux candidats après les trois règles précédentes, le flux est sélectionné aléatoirement

La combinaison conservée parmi les doublons est celle qui se place en tête après application de ces quatre règles.

1.3.5 Rapprochement des bases contrats et période de développement des devis

Les flux et devis que nous avons rapprochés constituent le socle de la base. Le rapprochement de ce socle aux bases contrats permet d'identifier les devis qui ont été transformés. Ce rapprochement est cette fois-ci immédiat et unique par le biais du numéro du devis.

Il faut cependant convenir de la durée laissée aux devis pour être convertis en contrats : c'est la période de développement. Il est préférable de prendre une période suffisamment longue pour ne pas manquer une partie des conversions. Toutefois, une période trop longue raccourcit la période d'observation et réduit donc le nombre de devis. En outre, il est discutable de qualifier de devis converti un devis ayant mis trop longtemps à générer une affaire nouvelle. Il faut donc faire un arbitrage.

Afin de déterminer la période de développement, nous avons observé le délai de conversion des devis des années 2019 et 2020 en affaires nouvelles. La figure 5 présente la fonction de répartition de ce délai :

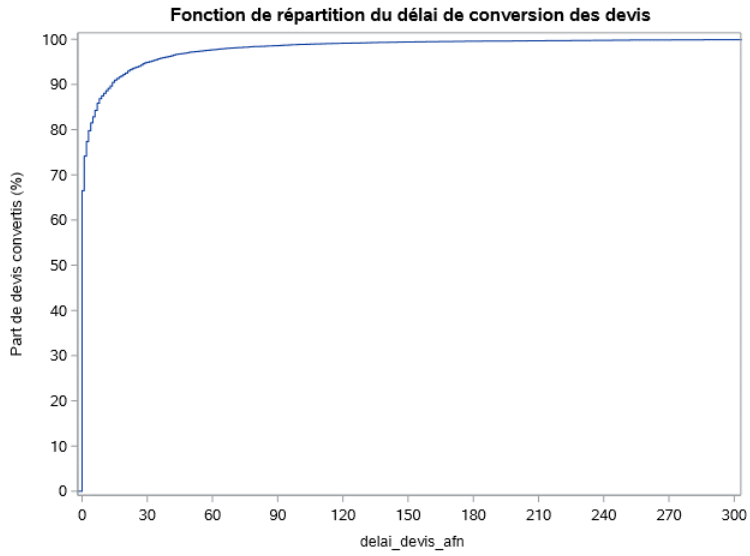

FIGURE 5 – Répartition des délais (en jours) de conversion des devis

La part des devis convertis atteint rapidement 95%, pour un délai situé entre 30 et 60 jours, avant de stagner. Nous décidons de fixer l'intervalle de développement à 60 jours, ce qui correspond à 98% des conversions sur la période. Le choix du 30 juin 2021 pour la fin de la période d'observation est la conséquence de ces deux mois de recul : nous observons les devis jusqu'au 30 juin 2021 et les affaires nouvelles jusqu'au 31 août 2021.

Pour déterminer les conversions, nous rapprochons donc aux devis les observations des bases contrat par leurs numéros, à condition que la date de souscription soit comprise dans l'intervalle $[t_{devis}; t_{devis} + 60j]$. Les devis conservés par la jointure sont les devis transformés. Nous les repérons en ajoutant une variable TOP_AFN indicatrice de la conversion :

$$TOP\_AFN = \begin{cases} 1 & \text{si le devis a été transformé} \\ 0 & \text{si le devis n'a pas été transformé} \end{cases}$$

1.3.6 Enrichissement et nettoyage de la base d'étude

Le rapprochement des bases contrat marque la fin de la construction du socle de la base. A présent, nous enrichissons celui-ci par l'ajout de variables utiles des autres bases dont nous disposons. Nous ajoutons en premier lieu les variables des bases DEVIBIEN et DEVIGAR pour compléter les informations du devis déjà présentes dans le socle. Puis, nous ajoutons des variables provenant d'autres bases :

Ajout des zoniers

Nous enrichissons la base d'étude par l'ajout du zonier commercial. Le zonier est une table de nombres entiers compris entre 1 et 30 et qui traduisent le niveau des différents risques sur une zone géographique. Les zones géographiques correspondent aux codes IRIS. Neuf risques sont quantifiés : incendie, catastrophe naturelle ou technologique, dégât des eaux, attentat, bris de glace, vol, responsabilité civile vie privée, responsabilité civile immeuble, dégâts électriques. Chaque risque est différencié entre les appartements et les maisons. Par conséquent, chaque zone IRIS est représentée par un 18-uplet de variables dans la base du zonier. Nous rapprochons cette base au socle par l'intermédiaire du code IRIS.

Ajout des variables agent

Nous disposons d'une base contenant des informations générales sur les différentes agences. En la rapprochant par le biais du code d'intermédiaire du devis, nous ajoutons au socle les variables suivantes :

- l'âge de l'agent
- l'ancienneté de l'agent
- l'ancienneté de l'agent dans le protocole du multi-accès particulier

Ajout d'une variable de mois gratuits

La compagnie conduit annuellement des campagnes promotionnelles prenant la forme de deux mois de cotisation offerts pour les nouveaux clients éligibles. L'offre s'applique sur le deuxième mois de la première année et le premier mois de la seconde année sous réserve de reconduction tacite du contrat. Par le biais d'un rapprochement des bases de mois gratuits grâce au numéro de devis, nous ajoutons cette information à notre base dans la variable TOP_MOISG :

$$TOP\_MOISG = \begin{cases} 1 & \text{si une offre de mois gratuits est appliquée} \\ 0 & \text{sinon} \end{cases}$$

Nettoyage

La dernière étape consiste en un nettoyage minutieux des variables de la base. Ce nettoyage inclut notamment :

- Le retrait des variables inutiles pour la suite ou faiblement renseignées
- Le retrait des lignes présentant des valeurs manquantes ou erronées (par exemple des âges, délais ou primes négatifs, des variables binaires à valeurs hors de 0,1 *etc...*)
- Le retrait d'observations appartenant à des catégories exceptionnelles, par exemple les chalets ou les occupants à titre gratuit.

Ces opérations de nettoyage conduisent à supprimer 13% des observations de la base d'étude. Nous obtenons finalement une base comportant 130 449 lignes et 170 variables. Celle-ci servira de support à notre étude.

1.4 Analyse descriptive

L'étude descriptive des variables de notre base de données est une étape incontournable et préliminaire à la construction des modèles. Elle nous permet d'analyser les caractéristiques de notre portefeuille de devis et d'identifier les variables pouvant jouer un rôle dans l'acte de conversion.

Nous commençons par résumer les principales caractéristiques de la base d'étude dans le tableau 2 :

Période d'observation	**Volume de devis**	**Volume d'affaires nouvelles**	**Taux de transformation**
01/07/2019 - 30/06/2021	130449	25122	19,26%

TABLE 2 – Caractéristiques principales de la base d'étude

Le taux de transformation global est de 19,3% sur l'ensemble de la période d'observation. La figure 6 représente son évolution mensuelle :

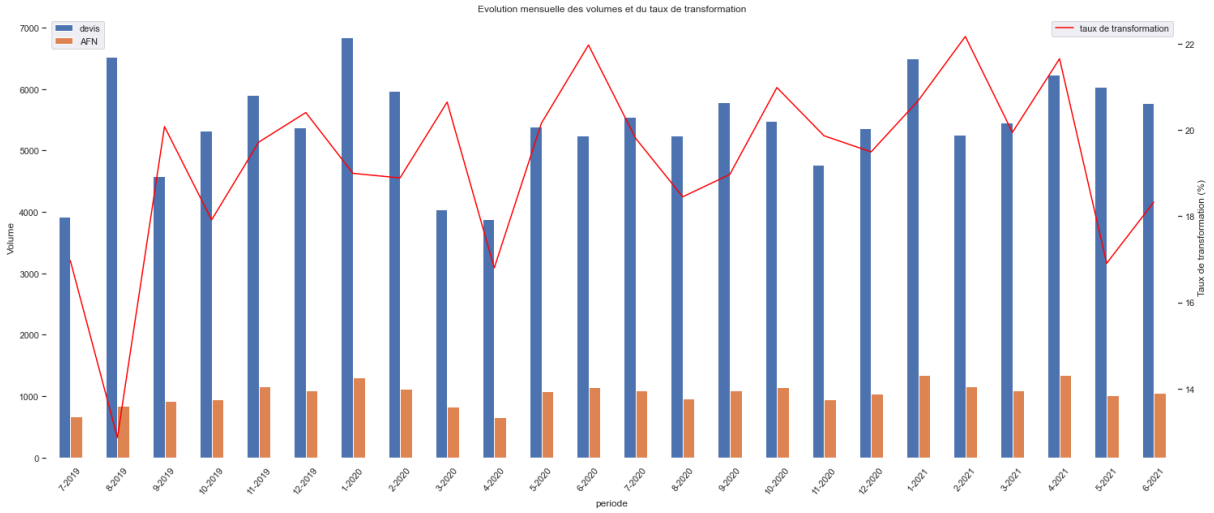

FIGURE 6 – Evolution mensuelle du taux de transformation

Cette figure révèle que le taux de transformation est assez stable, ce qui nous rassure sur le choix de deux années d'historique. Il n'y a donc a priori pas eu de déviation notable du marché.

Nous présentons à présent les analyses univariées du taux de transformation. Pour chacune des variables susceptibles d'entrer dans la modélisation du taux de transformation, un graphique présentant la fréquence des modalités et le taux de transformation associé a été tracé. Nous présentons ici l'analyse des variables principales. Les barres bleues représentent les effectifs des modalités. Les fréquences associées sont notées en bleu. Les points noirs représentent le taux de transformation de chaque modalité. Ces graphiques donnent une première idée de l'impact des variables sur le taux de transformation. Néanmoins, seuls les modèles permettront d'isoler l'effet pur de chacune des variables.

a) Le parcours digital :

La première des variables que nous explorons est la variable $parcours\_deb\_fin$, qui indique les points de départ et d'arrivée du parcours digital. La composante départ indique si le flux a été initié en comparateur ou sur allianz.fr. La composante arrivée indique sur quelle plateforme s'est achevé le parcours du prospect : sur le comparateur, un Fast Quote ou un Normal Quote.

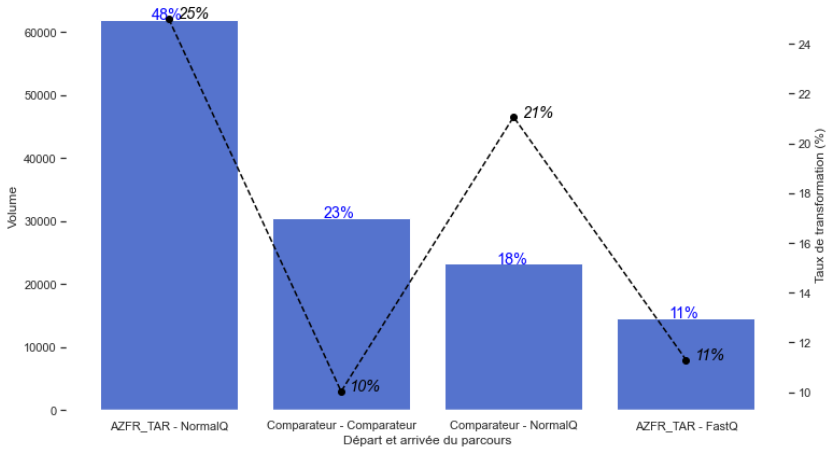

FIGURE 7 – Taux de transformation en fonction du parcours digital

Nous observons un fort écart de taux de transformation entre les différentes modalités. En effet, parmi les devis initiés sur le site allianz.fr, les devis Normal Quote sont deux fois plus transformés que les devis Fast Quote. De même, parmi les devis initiés sur un comparateur, ceux poursuivis sur allianz.fr (automatiquement en Normal Quote) sont deux fois plus transformés que les devis des prospects qui sont restés sur le comparateur. Nous en déduisons un impact a priori significatif de cette variable sur la conversion. L'interprétation que nous pouvons formuler est la suivante : après avoir vu le prix sur le comparateur ou en Fast Quote, les prospects qui font l'effort de poursuivre en Normal Quote ont a priori davantage l'intention de souscrire.

b) La situation du flux :

La variable suivante est la variable *situation*, qui permet de catégoriser le flux digital parmi les différents scénarios possibles dans le multi-accès :

- "Etre appelé" si le prospect a cliqué sur le bouton "être appelé" et a rempli les informations nécessaires pour cette mise en relation.
- "Rencontrer un conseiller" si le prospect a cliqué sur le bouton "rencontrer un conseiller" et a rempli les informations nécessaires pour cette mise en relation.
- "Souscription" si le prospect a cliqué sur le bouton "Souscrire en ligne" et a rempli les informations nécessaires pour cela.
- "Recevoir devis" si le prospect a cliqué sur le bouton "recevoir le devis par email".
- "Proposition sans tarif" : dans certains cas, le tarif est calculé mais n'est pas affiché sur internet. A la place, le prospect est prévenu qu'il sera contacté par un agent.
- "Parcours interrompu" si le prospect a abandonné avant d'atteindre les situations ci-dessus.

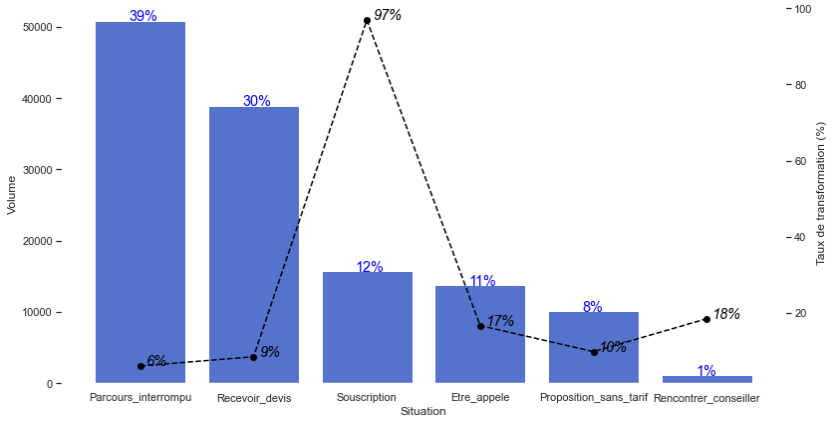

FIGURE 8 – Taux de transformation en fonction de la situation du flux

La majorité des devis correspond à un parcours interrompu, qui est assez logiquement la modalité au taux de transformation le plus faible. Le taux de transformation est un peu plus élevé chez les prospects demandant à recevoir le devis par mail et les propositions sans tarif. Les prospects se retrouvant dans ces trois situations ont logiquement moins d'intention à souscrire, mais il est intéressant de constater que le taux de transformation n'est pas nul pour autant.

Les situations "rencontrer un conseiller" et "être appelé", c'est-à-dire les situations de demande de mise en relation, ont un taux de transformation plus élevé. Néanmoins, nous constatons que seuls 1% des devis sont issus d'une demande de rencontre avec un conseiller. Nous

justifions a priori cela par le fait que les prospects utilisant le digital le font rarement avec l'objectif de se rendre ensuite en agence, ou alors ils ne formulent pas cette intention par une demande de mise en relation.

Enfin, nous constatons que les flux en situation "Souscription" sont presque systématiquement convertis. Par conséquent, nous ferons preuve de vigilance quant à l'utilisation ultérieure de cette variable.

c) Le type de destinataire :

La variable *type_ destinataire* indique si le flux a été envoyé en agence, en plateforme ou si le prospect a choisi de souscrire en ligne sans intermédiaire.

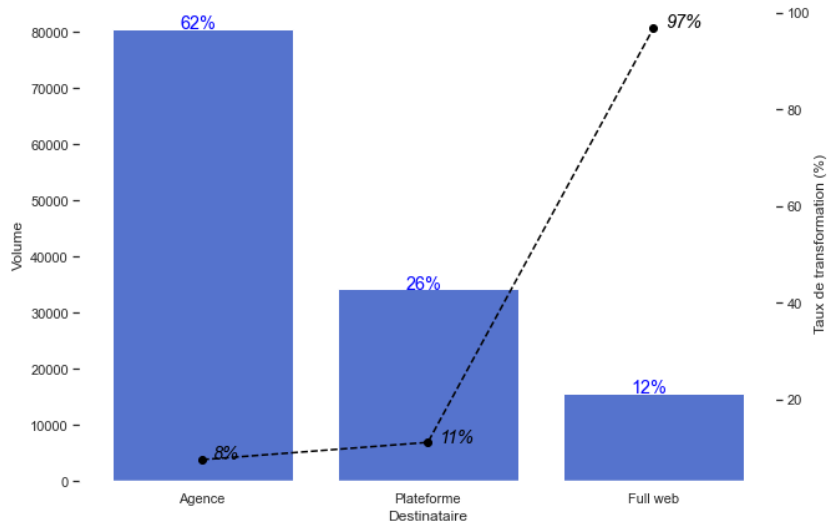

FIGURE 9 – Taux de transformation en fonction du type de destinataire du flux

Près des deux tiers des devis d'origine digitale sont envoyés en agence, tandis qu'un quart est envoyé en plateforme. Les devis qui se poursuivent vers une souscription en ligne représentent les 12% restants.

La variable *type_ destinataire* est liée à la variable *situation*. Les taux de transformation observés découlent de ce lien. En effet, les flux de type "rencontrer conseiller", "proposition sans tarif", "parcours interrompu" et une partie des flux "recevoir un devis" sont envoyés aux agents. La forte proportion de devis "parcours interrompus" contribue ainsi à baisser le taux de transformation des agents. Les flux "être appelé" et le reste des flux "recevoir un devis" sont envoyés aux plateformes téléphoniques. Ces flux "recevoir un devis" tirent ainsi le taux de transformation des plateformes vers le bas. La modalité "Full web" reprend quant à elle la catégorie de flux "Souscription".

d) La qualité juridique et le type d'habitation :

La figure 10 décrit les variations du taux de transformation en fonction du croisement entre la qualité juridique et le type d'habitation :

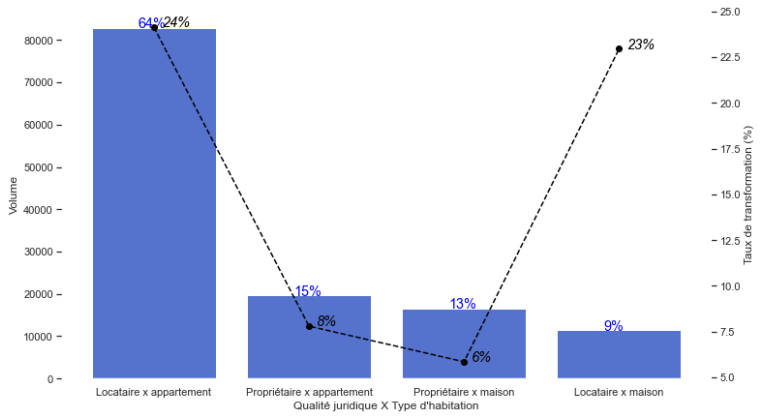

FIGURE 10 – Taux de transformation en fonction de la qualité juridique et du type d'habitation

Près des deux tiers des devis de la base d'étude proviennent de locataires d'appartement. Leur taux de transformation est élevé : il s'élève à 24%. Nous constatons un fort écart de transformation entre locataires et propriétaires, aussi bien sur les appartements que les maisons. En effet, les locataires transforment trois fois plus que les propriétaires. Le type d'habitation semble a priori moins discriminant que la qualité juridique.

e) La prime :

La figure 11 décrit les variations du taux de transformation en fonction de la prime inscrite sur le devis. Cette dernière a été découpée en 10 tranches d'effectifs assez proches.

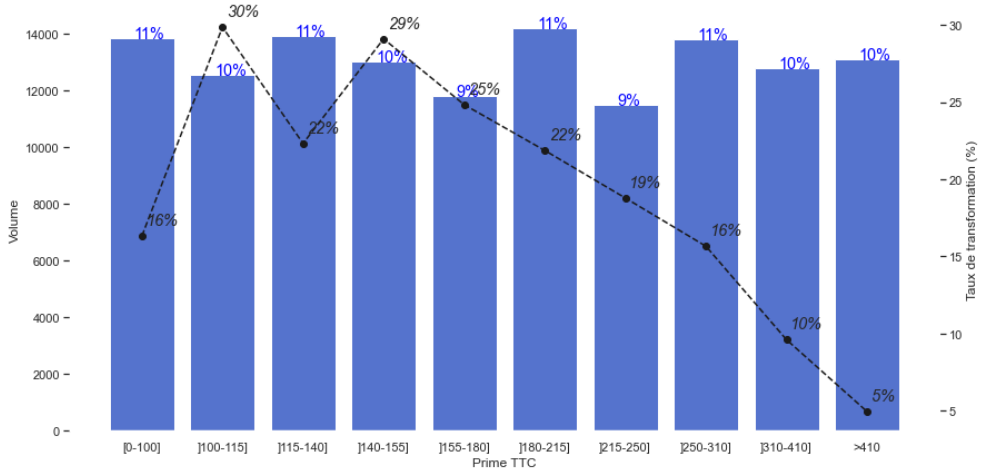

FIGURE 11 – Taux de transformation en fonction de la prime

La figure révèle une variation assez marquée du taux de transformation en fonction de la prime. Ce dernier évolue en dents de scie sur les quatre premières tranches. Cette évolution peut être liée au positionnement tarifaire. En effet, la première tranche correspond à un segment très concurrentiel, ce qui explique son taux de transformation plus faible. Le creux sur la troisième tranche révèle a priori un positionnement tarifaire moins compétitif. A partir de 155€, le taux de transformation entame une décroissance à mesure que la prime augmente.

f) Le nombre de pièces du bien :

La figure 12 décrit les variations du taux de transformation en fonction du nombre de pièces du bien :

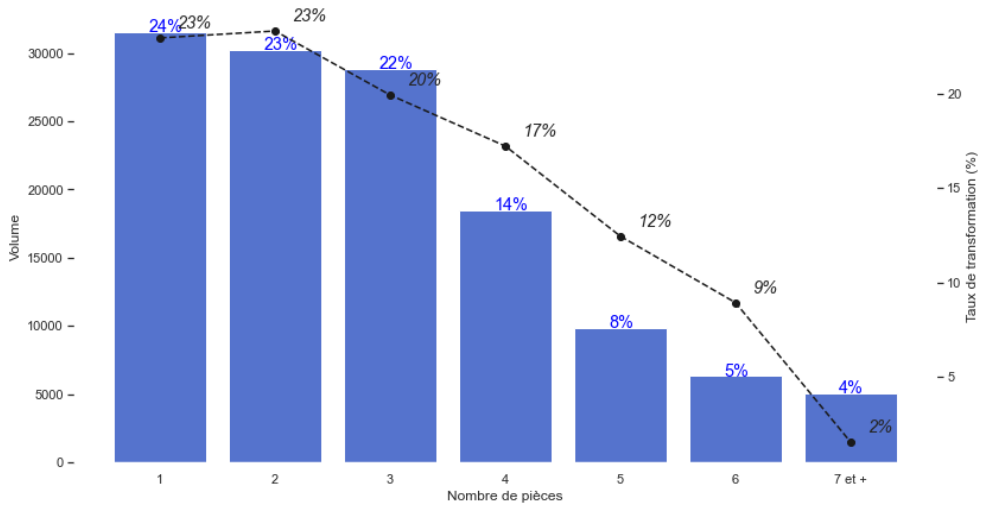

FIGURE 12 – Taux de transformation en fonction du nombre de pièces

Nous observons une décroissance à la fois du nombre des devis et du taux de transformation avec le nombre de pièces du bien. Cette évolution est a priori liée aux deux précédentes dans une certaine mesure. En effet, les logements de moins de trois pièces sont des biens plutôt standards, correspondant majoritairement à des locataires d'appartement et de faibles primes. Cela peut expliquer leur taux de conversion élevé. A l'inverse, les biens avec un nombre de pièces plus élevé sont souvent des maisons aux primes plus élevées.

g) Le délai entre le devis et le début du contrat :

La dernière variable que nous présentons est le délai entre la date du devis et la date de début de contrat souhaitée.

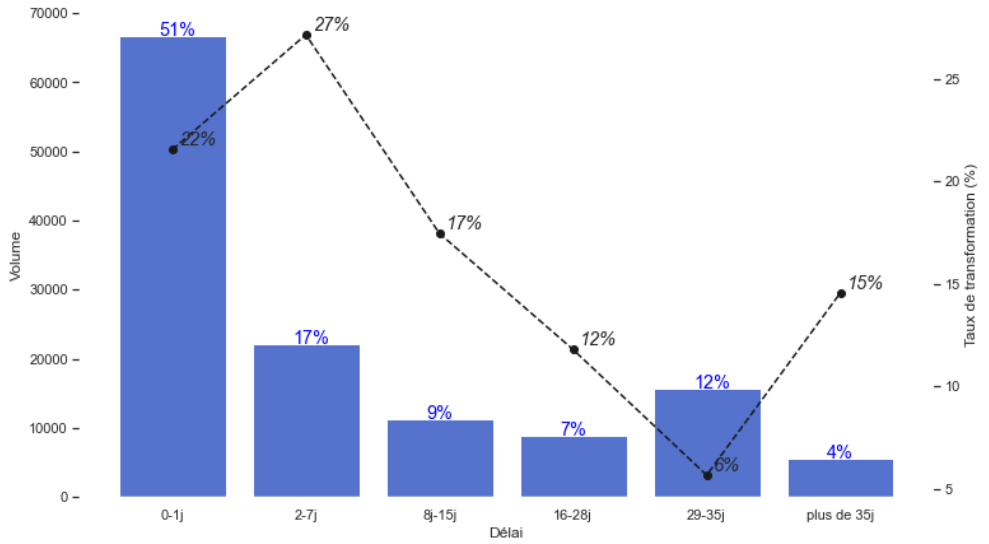

FIGURE 13 – Taux de transformation en fonction du délai avant la prise d'effet du contrat

Nous observons deux pics d'effectifs. Le premier correspond aux prospects souhaitant être assurés le jour même, ce qui représente la moitié des devis. Le second pic se trouve aux alentours d'un mois après la date de devis. Ce pic est dû à la loi Hamon. En effet, dans le cas d'une reprise loi Hamon, le délai entre le devis et la prise d'effet du contrat est automatiquement fixé à environ un mois.

Le taux de transformation varie fortement entre les différentes modalités. Il est maximal la première semaine, puis diminue à mesure que le délai s'allonge jusqu'à un mois. Il remonte néanmoins au-delà de 35 jours. Cela paraît contre-intuitif, car nous pensions qu'un plus grand délai rendait la conversion moins crédible étant donné qu'il laisse plus de chance au prospect de changer d'avis. Cette remontée peut cependant être due au faible effectif de la mesure.

La base de données étant construite, il est maintenant possible de passer à l'étape de modélisation du taux de transformation. La partie suivante permet de présenter les fondements théoriques des notions qui seront utilisées lors de cette étape.

2 Eléments théoriques de la modélisation

Nous exposons dans cette partie les fondements théoriques des outils utilisés lors de l'étape de modélisation.

2.1 Caractérisation du problème

Avant toute chose, il convient de caractériser le phénomène à modéliser. Notre objectif est de construire un modèle qui prédit la probabilité qu'un devis issu du digital soit transformé en contrat. La variable d'intérêt est la variable $Y = TOP\_AFN$. Celle-ci possède deux modalités :

$$Y = \begin{cases} 1 & \text{si le devis est converti} \\ 0 & \text{sinon} \end{cases}$$

Nous sommes donc a priori face à un problème de classification binaire. La probabilité de transformation d'un devis i correspond à la probabilité d'appartenance de ce devis à la classe des devis convertis :

$$\pi_i = P(Y_i = 1)$$

Le taux de transformation d'un échantillon de N devis est quant à lui donné par la proportion de devis convertis :

$$\tau = \frac{\#(Y = 1)}{N} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

Néanmoins, le besoin de notre étude diffère quelque peu de l'objectif premier d'une classification. Pour un devis donné, ce n'est pas la conversion que nous souhaitons prédire, mais la probabilité de conversion. Autrement dit, ce ne sont pas les classes mais la probabilité d'appartenance à la classe $Y = 1$ que nous souhaitons prédire *in fine*. Par conséquent, nous suivons la méthodologie de résolution d'un problème de classification, à deux points de vigilance près. D'une part, les algorithmes de classification choisis doivent intégrer une estimation des probabilités d'appartenance aux classes. D'autre part, il nous faudra nous assurer de la qualité d'ajustement des probabilités modélisées. Il est en effet possible qu'un modèle présente de bonnes performances de classification mais ne prédise pas les probabilités des classes avec une précision suffisante.

Compte tenu de ces éléments, nous avons retenu deux catégories de modèles de classification permettant de répondre à notre problème : la régression logistique et les algorithmes de *machine learning* basés sur les arbres de décision.

2.2 Les mesures du lien entre les variables

L'étude du lien entre les variables constitue une étape essentielle de la modélisation. Trois mesures de l'association entre des variables ont été considérées pour notre étude :

- le coefficient de corrélation de Pearson
- le test d'indépendance du khi-2
- le V de Cramer

Le coefficient de corrélation de Pearson

Il s'agit de la mesure classique de la corrélation linéaire entre deux variables numériques.

La covariance entre deux variables X et Y est définie par :

$$\mathrm{cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

Le coefficient de corrélation de Pearson entre deux variables X et Y de variances finies s'obtient alors en normalisant la covariance par le produit des écart-types des deux variables :

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

L'estimation du coefficient de corrélation sur un échantillon $\{(x_1, y_1), ..., (x_n, y_n)\}$ est donnée par la formule suivante :

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Ce coefficient est compris entre -1 et 1. Plus le coefficient est proche de 1 (respectivement -1) et plus il existe une corrélation positive (respectivement négative) entre les deux variables. Au contraire, si le coefficient est proche de 0, il n'existe pas de lien linéaire entre les deux variables.

Le test d'indépendance du khi-2 :

Le test du khi-2 permet de mesurer l'indépendance entre deux variables catégorielles. Les effectifs des différentes combinaisons de modalités des deux variables considérées sont comptabilisés dans un tableau à double entrée, appelé tableau de contingence. Ce tableau permet de calculer la statistique du khi-deux :

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_i \cdot n_j}{n}\right)^2}{\frac{n_i \cdot n_j}{n}}$$

où :

— n_{ij} est le nombre d'observations vérifiant $X = x_i$ et $Y = y_j$
— n_i est le nombre d'observations vérifiant $X = x_i$
— n_j est le nombre d'observations vérifiant $Y = y_j$
— n est l'effectif total

Sous l'hypothèse d'indépendance des deux variables, la statistique de test suit asymptotiquement une loi du khi-deux à k degrés de liberté. Ce nombre k est égal au produit $(l-1)\cdot(c-1)$ où (l, c) sont respectivement le nombre de lignes et de colonnes du tableau de contingence. Pour conclure sur l'indépendance des échantillons, la statistique calculée est comparée à une valeur critique issue d'une table de la distribution du khi-deux à k degrés de liberté, pour le niveau de confiance choisi.

Le V de Cramer

Le test du khi-deux possède certains inconvénients : il n'est pas normé, il est sensible à la taille de l'échantillon et il nécessite l'utilisation d'une table. Le V de Cramer est une mesure dérivée du khi-2 qui permet de résoudre ces problèmes. Cet indicateur s'obtient en normalisant la statistique du khi-2 par sa valeur maximale théorique $\chi_{\max}$. Celle-ci correspond à la valeur

du khi-2 obtenue si le tableau de contingence avait une seule case non nulle par ligne ou par colonne. Le V de Cramer s'écrit :

$$V\_Cramer = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} = \sqrt{\frac{\chi^2}{n \cdot (\min(l,c) - 1)}}$$

Ainsi, contrairement au khi-2, le V de Cramer est normé : il prend ses valeurs dans [0,1]. Plus sa valeur est proche de 1 et plus le lien entre les variables est fort. Inversement, le V de Cramer vaut 0 pour des variables qui n'ont aucun lien.

2.3 La base d'apprentissage et la base de test

Une exigence fondamentale à laquelle le modèle doit répondre est qu'il doit être capable de se généraliser à de nouvelles données, indépendantes des données sur lesquelles il a été entraîné. Par conséquent, il ne faut pas évaluer les performances d'un modèle en utilisant les données qui ont servi à son apprentissage. D'une part, une telle évaluation ne donnerait aucune information sur la capacité de généralisation du modèle. D'autre part, chercher les meilleures performances sur les données d'apprentissage favorise les chances d'obtenir un modèle trop spécifique à cet échantillon. Un tel modèle serait alors incapable d'effectuer de bonnes prédictions sur un nouvel échantillon inconnu. On parle dans ce cas d'*overfitting* ou surapprentissage.

Par conséquent, l'approche classique consiste à séparer la base de données en deux sous-bases disjointes : la base d'apprentissage (*training set*) et la base de test (*test set*). La première sert à entraîner les différents modèles, tandis que la seconde est uniquement utilisée pour les évaluer et comparer leurs performances. Le ratio de découpage entre base d'apprentissage et base de test est issu d'un compromis. D'une part, le modèle a besoin d'apprendre sur un échantillon d'apprentissage suffisamment grand, sans quoi il serait trop simple et aurait de piètres performances (on parle d'*underfitting*). Mais d'autre part, plus l'échantillon d'apprentissage est grand, plus celui de test se réduit, et moins l'évaluation du modèle est fiable. Les proportions choisies sont en général de l'ordre de 70 à 80% des données consacrées à l'apprentissage et 20 à 30% consacrées à l'évaluation. Par ailleurs, ce découpage doit impérativement être aléatoire pour ne pas biaiser l'entraînement du modèle.

2.4 La validation croisée

Lors de la phase d'entraînement, nous pouvons être amenés à vouloir évaluer le modèle afin de mieux calibrer ses hyperparamètres. Cependant, les données de test ne doivent pas être utilisées pour cela, car le modèle est encore en phase d'apprentissage. Une solution est de découper la base d'apprentissage initiale en une base d'entraînement et une base de validation. L'échantillon de validation est alors utilisé pour l'évaluation intermédiaire du modèle dans le but d'améliorer ses hyperparamètres. L'échantillon de test est toujours uniquement utilisé pour l'évaluation finale du modèle.

Néanmoins, le redécoupage de la base d'entraînement en deux parties réduit encore davantage la part des données consacrée à l'apprentissage. De plus, la mesure peut être biaisée par le découpage. La technique de validation croisée k-fold constitue alors une solution plus efficace. Son principe est décrit ci-après :

Principe de la validation croisée k-fold :

Les données ont été séparées en un échantillon d'apprentissage et un échantillon de test. Dans la méthode de validation croisée k-fold, l'échantillon d'apprentissage D est divisé en k sous-échantillons disjoints $D_1, ..., D_k$. Puis, pour chaque j dans {1,...,k} :

- le sous-échantillon D_j est exclu de l'échantillon d'apprentissage
- un modèle est entraîné sur $D \setminus D_j$
- le sous-échantillon D_j est utilisé comme échantillon de test pour mesurer un score de performance du modèle.

A l'issu de ces itérations, nous obtenons ainsi k scores de performance, mesurés sur k échantillons de test différents. Le score de performance globale donné au modèle est alors la moyenne des k scores mesurés. La validation croisée permet donc d'optimiser l'utilisation des données à disposition pour calibrer le modèle. En effet, les données de l'échantillon d'apprentissage sont toutes utilisées à la fois pour l'apprentissage et l'évaluation, sans que le principe de séparation des données d'apprentissage et d'évaluation ne soit enfreint. Contrairement à la méthode classique, l'échantillon d'évaluation est mouvant, ce qui évite le biais de découpage. Le choix du nombre k d'échantillons doit être tel que les échantillons restent suffisamment représentatifs.

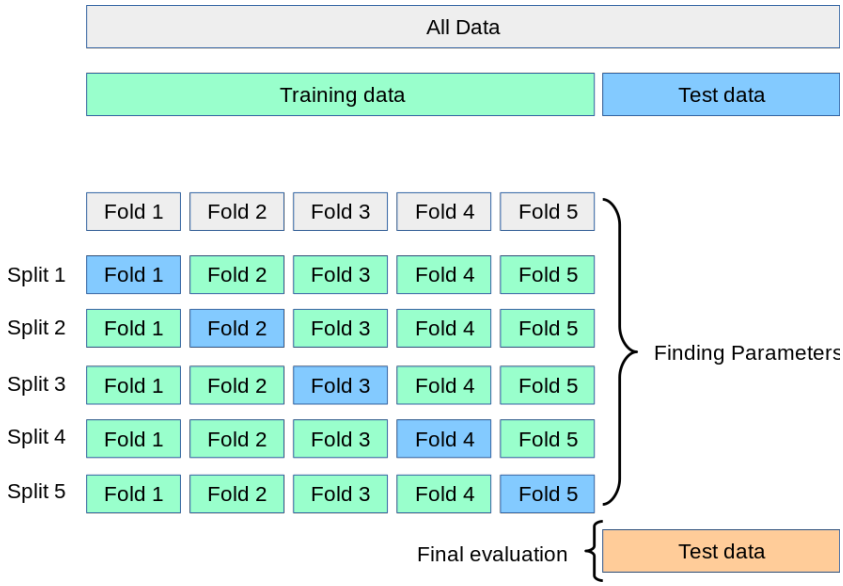

FIGURE 14 – Schéma d'une validation croisée 5-fold (*source : documentation scikit-learn*[2])

2.5 La régression logistique

2.5.1 La famille des modèles linéaires généralisés

Rappelons tout d'abord la définition des modèles linéaires. Ces modèles sont basés sur l'hypothèse d'existence d'une relation linéaire entre les variables explicatives X_1, ..., X_m et la variable à prédire Y :

$$Y|X = X\beta + \beta_0 + \varepsilon$$

où :

- Y est une variable aléatoire de variance supposée constante (hypothèse d'homoscédasticité)
- $X = (X_1, ..., X_k)$ est la matrice des observations des différentes variables explicatives
- ε est une variable aléatoire qui suit une loi normale centrée. Il s'agit du terme d'erreur dans le modèle

- $\beta = (\beta_1, ..., \beta_m)$ est le vecteur des coefficients associés aux variables explicatives et β_0 est le terme constant (*intercept*). Il s'agit des paramètres du modèle. Ils sont estimés par la méthode des moindres carrés.

,

La famille des modèles linéaires généralisés (GLM : *Generalized Linear Models*) est issue d'une généralisation moins restrictive du modèle linéaire classique. Ces modèles font l'hypothèse de la relation suivante :

$$E[Y|X] = g^{-1}(X\beta + \beta_0 + \epsilon)$$

Ainsi, ce n'est cette fois plus la variable Y, mais une transformation de son espérance qui est modélisée sous la forme d'une combinaison linéaire des variables explicatives. La fonction g, monotone, différentiable et inversible, est appelée fonction de lien. Les coefficients β et β_0 sont estimés par la méthode du maximum de vraisemblance.

Dans le cadre des modèles linéaires généralisés, la distribution de la loi conditionnelle de Y doit appartenir à la famille exponentielle. Celle-ci regroupe l'ensemble des fonctions de densité de la forme :

$$\ln f_Y(y) = \frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)$$

où :

- $a(.)$ est une fonction non nulle, dérivable sur $\mathbb{R}$
- $b(.)$ est une fonction trois fois dérivable sur $\mathbb{R}$ et de dérivée première inversible
- $c(.)$ est une fonction définie sur $\mathbb{R}^2$
- θ et ϕ sont les paramètres de moyenne et de dispersion.

Parmi les distributions appartenant à cette famille, on retrouve entre autres les distributions normale, exponentielle, Poisson, gamma ainsi que la distribution de Bernoulli qui va nous intéresser dans la suite.

2.5.2 La régression logistique

La régression logistique, ou modèle logit, est un cas particulier du modèle linéaire généralisé utilisé lorsque la variable à expliquer est binaire :

$$Y = \begin{cases} 1 & \text{de probabilité } P(Y=1) = \pi \\ 0 & \text{de probabilité } P(Y=0) = 1-\pi \end{cases}$$

La variable Y suit donc une loi de Bernoulli (appartenant à la famille exponentielle) de paramètre π. La fonction de lien associée à ce modèle est la fonction *logit* :

$$\begin{array}{rccc} g & : &]0;1[& \rightarrow & \mathbb{R} \\ & & x & \mapsto & ln(\frac{x}{1-x}) \end{array}$$

Elle est bijective et dérivable sur $]0;1[$. Son inverse est la fonction logistique :

$$\begin{array}{rccc} g^{-1} & : & \mathbb{R} & \rightarrow &]0;1[\\ & & x & \mapsto & \frac{1}{1+exp(-x)} \end{array}$$

Le modèle s'écrit alors :

$$E[Y|X] = \frac{1}{1 + exp(-(X\beta + \beta_0 + \epsilon))}$$

La quantité modélisée $E[Y|X]$ est une valeur numérique appartenant à $]0\,;1[$. Elle correspond à la probabilité que l'observation appartienne à la classe positive. En effet, la variable Y étant binaire :

$$E[Y|X] = 1 \cdot P(Y = 1|X) + 0 \cdot P(Y = 0|X) = P(Y = 1|X) = \pi$$

2.5.3 Estimation des coefficients

Les coefficients du modèle sont estimés par la méthode du maximum de vraisemblance, présentée rapidement ci-après. La probabilité conditionnelle de Y sachant X pour un individu ω peut être réécrite :

$$P(Y(\omega) = 1 \mid X(\omega))^{Y(\omega)} \times [1 - P(Y(\omega) = 1 \mid X(\omega))]^{1-Y(\omega)}$$

La vraisemblance d'un échantillon d'observations $(x_i, y_i)_{i=1,\dots N}$ s'écrit alors :

$$L(\beta_0, \beta) = \prod_{i=1}^{N} \left(\frac{1}{1 + \exp(-(\beta x_i + \beta_0))} \right)^{y_i} \left(1 - \frac{1}{1 + \exp(-(\beta x_i + \beta_0))} \right)^{1-y_i}$$

On définit la log-vraisemblance par :

$$l\left(\beta_0, \beta\right) = log(L\left(\beta_0, \beta\right))$$

La log-vraisemblance de l'échantillon s'écrit :

$$l(\beta_0, \beta) = \sum_{i=1}^{N} y_i \cdot log \left(\frac{1}{1 + \exp(-(\beta x_i + \beta_0))} \right) + (1 - y_i) \cdot log \left(1 - \frac{1}{1 + \exp(-(\beta x_i + \beta_0))} \right)$$

La maximisation de la vraisemblance est équivalente à la maximisation de la log vraisemblance. L'estimateur des coefficients du modèle est alors le vecteur solution du système du premier ordre suivant :

$$\begin{cases} \frac{\partial l(\beta_0,\beta)}{\partial \beta_0} = 0 \\ \\ \frac{\partial l(\beta_0,\beta)}{\partial \beta_j} = 0 \quad j = 1, \dots, m \end{cases}$$

En pratique, la résolution est faite de manière approchée par les logiciels à l'aide de méthodes itératives.

2.5.4 Validation du modèle

Test de Wald pour les coefficients :

Le test de Wald permet de tester l'hypothèse $H_0 : \beta_j = 0$ de nullité du coefficient associé à la variable j, contre l'hypothèse $H_1 : \beta_j \neq 0$. La statistique du test s'écrit :

$$W = \frac{\hat{\beta_j}^2}{\hat{V}(\beta_j)}$$

où $\hat{V}(\beta_j)$ est l'estimateur de la variance du coefficient. Cette statistique suit une loi du khi-2 à un degré de liberté. La p-value associée au test est :

$$p\text{-}value = P(W < \chi_1^2)$$

L'hypothèse nulle est rejetée si la p-value est inférieure au seuil α choisi. Dans ce cas, la variable X_j est jugée significative. Nous utiliserons les p-values du test de Wald afin d'éliminer récursivement du modèle la variable associée à la plus grande p-value, puis pour contrôler l'absence de variable superflue dans le modèle final, au seuil de 5%.

Validation du modèle

Il existe plusieurs critères, basés sur le maximum de vraisemblance, qui permettent de juger de la qualité d'ajustement d'une régression logistique. Ces derniers ne sont pas directement interprétables sans moyen de comparaison, c'est pourquoi nous les utiliserons pour comparer successivement les modèles construits jusqu'à obtenir le meilleur modèle. Les critères que nous utilisons sont les suivants :

• Le critère AIC :

Le critère d'information d'Akaike (AIC pour *Akaike Information Criterion*) utilise le maximum de vraisemblance en pénalisant le nombre m de variables explicatives. Il est défini par : $AIC = -2ln(L) + 2m$

• Le critère BIC :

Le critère d'information bayésien (BIC pour *Bayesian Information Criterion*), ou critère de Schwarz, prend en compte le nombre m de variables explicatives ainsi que le nombre n d'observations. Il est défini par : $BIC = -2ln(L) + mln(n)$

• Le log-loss :

Le log-loss est la moyenne de l'opposé de la log-vraisemblance des observations :

$$logloss = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right) \cdot \log\left(1 - p\left(y_i\right)\right)$$

où

— n est le nombre d'observations
— $y_i \in \{0; 1\}$ est la i-ème observation de la variable réponse
— $p_i = P(y_i = 1)$

Pour chacun de ces trois indicateurs, plus la valeur mesurée est faible, meilleur est le modèle. Notons que les méthodes présentées permettent de vérifier l'ajustement du GLM aux données d'entraînement. Or, au même titre que les modèles de *machine learning*, il est essentiel de vérifier que le modèle s'ajuste toujours à de nouvelles données pour le valider. Le modèle GLM retenu sera donc ensuite évalué sur l'échantillon de test.

2.5.5 Interprétation des coefficients

Il est possible d'interpréter facilement les coefficients de la régression logistique à l'aide des *odd ratios*, afin de connaître l'effet des variables sur la conversion d'après le modèle.

La cote (*odd*) d'un événement au sein d'un groupe est définie par le rapport entre les probabilités de succès et d'échec de l'événement :

$$C(x) = \frac{p(x)}{1 - p(x)}$$

Une cote de 2 signifie que l'événement a deux fois plus de chances de se produire au sein du groupe que de ne pas se produire : on dit que la cote est de deux contre un.

Le rapport de cotes (ou *odd ratio*) entre deux groupes est le rapport entre les *odds* des deux groupes :

$$OR = \frac{C(groupe_1)}{C(groupe_2)} = \frac{p_1(x)(1-p_2(x))}{p_2(x)(1-p_1(x))}$$

Un *odd ratio* égal à 2 signifie que l'événement a deux fois plus de chances de succès dans le groupe 1 que dans le groupe 2.

Soit une variable ayant été dichotomisée en variables binaires indicatrices de ses modalités. Considérons deux groupes : les observations x possédant la variable recodée $X_i = 1$ face aux observations x' possédant la modalité de référence (cette modalité n'est associée à aucun coefficient car elle est retirée du modèle). Supposons que toutes les autres variables sont identiques entre ces deux groupes, c'est-à-dire : $x - x' = (0, ..., 1, ..., 0)$. L'événement considéré est la conversion : $Y = 1$. Dans le cadre d'un modèle *logit*, l'*odd ratio* des deux groupes s'écrit alors :

$$OR(x, x') = \frac{p(x)(1-p(x'))}{p(x')(1-p(x))} = exp(\beta(x - x')) = exp(\beta_i)$$

Cet *odd ratio* mesure le rapport des chances de conversion des devis présentant la modalité i contre ceux présentant la modalité de référence, toutes autres variables identiques par ailleurs. Autrement dit, l'exponentielle du coefficient β_i d'un modèle *logit* mesure l'effet isolé de la variable X_i sur la conversion.

2.6 Les modèles basés sur les arbres de décision

2.6.1 Les arbres de décision

Un arbre binaire est composé d'un ensemble de noeuds engendrant chacun 0 ou 2 autres noeuds. Un noeud qui n'en engendre aucun autre est une feuille. Le premier noeud de l'arbre est appelé racine. Les arbres binaires peuvent être utilisés comme modèles prédictifs : on parle d'apprentissage par arbre de décision.

L'algorithme CART (*Classification and Regression Trees*) est un algorithme d'apprentissage supervisé permettant de construire des arbres de décision dans le cadre de problèmes de classification et de régression. La méthode consiste à répartir les observations de l'échantillon d'apprentissage dans des groupes en fonction des variables explicatives, de façon à obtenir les groupes les plus homogènes possibles et en minimisant une fonction de coût. Ces groupes correspondent aux différentes feuilles de l'arbre.

Prédiction avec un arbre CART :

Chaque noeud de l'arbre est associé à une règle de partition portant sur une des variables. Pour prédire la valeur associée à une nouvelle observation, il faut partir de la racine et parcourir les branches en fonction des différentes règles des noeuds, jusqu'à atteindre une feuille.

Dans le cadre d'une régression, la prédiction est la moyenne des labels des données d'apprentissage qui composent la feuille.

Dans le cadre d'une classification, la classe prédite est la classe majoritaire parmi les observations de la feuille. Le modèle fournit également une estimation des probabilités d'appartenance

de l'observation aux différentes classes. Cette estimation est en réalité la fréquence des différentes classes parmi l'échantillon d'observations associé à la feuille.

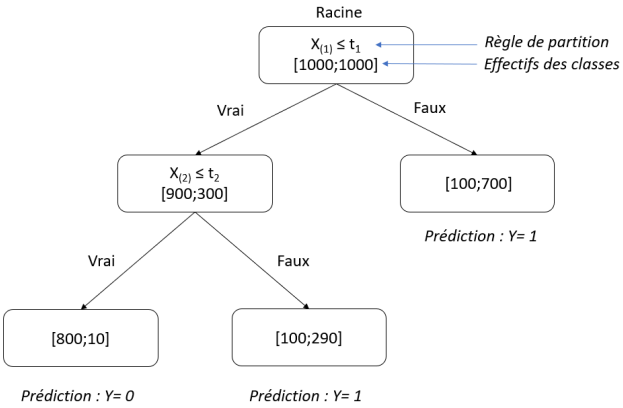

FIGURE 15 – Exemple d'arbre de classification binaire

Construction de l'arbre

La construction d'un arbre débute par la phase de croissance. Le point de départ est la racine, qui contient l'ensemble des observations. A chaque étape, l'algorithme doit choisir la variable et le seuil de division qui réalisent la meilleure partition possible de l'échantillon d'apprentissage. Cette partition divise l'échantillon associé au noeud en deux sous-échantillons disjoints. Dans un problème de classification, la qualité d'une partition est évaluée à l'aide d'une mesure d'homogénéité ou pureté. Une partition est pure si, après séparation, toutes les observations d'une branche appartiennent à la même classe. Dans ce cas, il n'est plus la peine de diviser l'échantillon et une feuille est créée. Si le noeud n'est pas pur, l'algorithme cherche à le diviser à nouveau en testant toutes les variables et tous les seuils possibles et en retenant la partition la plus pure. Pour une classification, la mesure utilisée par défaut est l'indice de Gini, défini pour le noeud Q_m par :

$$H\left(Q_m\right) = \sum_k p_{mk}\left(1 - p_{mk}\right)$$

où p_{mk} est la proportion d'observations de la classe k parmi toutes les observations du noeud Q_m. L'indice de Gini d'un noeud pur est ainsi nul.

Par défaut, la croissance de l'arbre se poursuit tant que les noeuds ne sont pas purs. Cependant, passé une certaine taille et à mesure que la croissance de l'arbre se poursuit, le risque d'entrer en surapprentissage s'accroît. En effet, l'ajout de règles de séparation jusqu'à obtenir des feuilles trop homogènes aboutit à un arbre qui n'est adapté qu'à l'échantillon utilisé pour son apprentissage.

Pour éviter cet écueil, une première solution est le choix d'un critère d'arrêt. On parle de pré-élagage. Le critère d'arrêt est en général lié à la profondeur de l'arbre, le nombre minimal d'observations dans les feuilles, ou encore le niveau de significativité de l'apport d'une partition supplémentaire (mesuré grâce à un test statistique). Une seconde solution est de laisser l'arbre croître entièrement avant de le réduire : c'est le post-élagage.

Intérêts et inconvénients des arbres de décision

Les arbres de décision possèdent l'avantage d'être facilement interprétables. En effet, il est possible d'observer directement les critères utilisés par l'arbre pour classer une observation. De

plus, ils ne requièrent pas de préparation excessive des données, ni d'hypothèses trop fortes. Cependant, ce sont des modèles instables. En effet, de légères variations dans l'échantillon d'apprentissage peuvent conduire à la construction d'arbres complètement différents. Par ailleurs, les arbres CART entrent facilement en surapprentissage.

Pour pallier ces faiblesses, les arbres de décision sont souvent combinés au sein d'algorithmes d'apprentissage ensembliste (*ensemble methods*). Il s'agit de méthodes basées sur l'agrégation de modèles, qui est une approche courante en apprentissage automatique. Cette approche est motivée par le constat que les différents modèles ne commettent en général pas exactement les mêmes erreurs. Ainsi, combiner les modèles au sein d'un modèle agrégé permet souvent d'obtenir un modèle dont les performances sont supérieures à celles de chacun des modèles qui le constituent. On distingue principalement deux familles de modèles d'apprentissage ensembliste : le *bagging* et le *boosting*. Les deux modèles que nous utilisons font chacun partie d'une de ces deux familles : il s'agit du *Random Forest* et du *Gradient Boosting*.

2.6.2 Le *bagging* et l'algorithme *Random Forest*

Le *bagging*

L'agrégation bootstrap, ou *bagging* (contraction de *Bootstrap Aggregation*), est un ensemble de méthodes d'agrégation de modèles basé sur le rééchantillonnage de la base d'apprentissage. Ces méthodes visent à réduire la variance globale en combinant des modèles qui sont séparément performants mais à forte variance. Le principe consiste à tirer B échantillons bootstrap, c'est-à-dire avec remise, de n observations de la base d'apprentissage. Chaque échantillon bootstrap sert à entraîner un modèle, qui est testé sur le reste de la base. On obtient ainsi B modèles qui sont finalement agrégés pour former le modèle final. Dans le cadre d'une classification, la prédiction donnée par le modèle agrégé est le résultat d'un vote à la majorité des modèles constitutifs.

L'algorithme *Random Forest*

La forêt aléatoire (*Random Forest*) est l'application du *bagging* aux arbres *CART*, à ceci-près que le rééchantillonnage est double. Tout d'abord, B échantillons *bootstrap* de n observations sont tirés. Pour chacun des B arbres, une combinaison de p variables est ensuite tirée parmi toutes les variables explicatives. Chacun des échantillons d'observations sert à entraîner un modèle, en utilisant exclusivement la combinaison de variables tirée. On obtient une "forêt" de B arbres qui ont été entraînés sur des observations et des variables différentes. Afin de prédire la classe d'une nouvelle observation, chaque arbre de la forêt fournit une prédiction. Le modèle renvoie alors la réponse majoritaire parmi les prédictions. De plus, l'algorithme *Random Forest* fournit une estimation des probabilités d'appartenance aux classes. Celle-ci est obtenue en moyennant les estimations des probabilités faites par chaque arbre.

Grâce à ce mode de construction, l'algorithme *Random Forest* est très efficace. Le double échantillonnage permet de garantir que chaque arbre est différent des autres, puisque chaque arbre a été entraîné non seulement sur un échantillon différent mais également avec une combinaison de variables différente. Par conséquent, les erreurs commises par les arbres ne coïncident pas et peuvent se compenser, et la variance du modèle s'en trouve réduite. De plus, l'algorithme ne nécessite pas de paramétrage poussé. En général, les arbres constitutifs n'entrent pas en surapprentissage et n'ont pas besoin d'être élagués. L'inconvénient principal de l'algorithme *Random Forest* est la difficulté d'interprétation de ses résultats. Contrairement à un arbre unique, il n'est en effet plus possible d'observer les partitions et les feuilles d'une agrégation de

plusieurs dizaines voire centaines d'arbres. Néanmoins, cet algorithme fournit une mesure de l'importance relative de chaque variable au sein du modèle.

2.6.3 Le *boosting* et l'algorithme *Gradient Boosting*

Tandis que le *bagging* est une méthode d'agrégation parallèle, le *boosting* est une méthode d'agrégation adaptative en série. Cette méthode consiste à entraîner l'un après l'autre des modèles faibles (ou *weak learner*) en faisant en sorte que chaque modèle soit plus attentif aux erreurs commises par le précédent. Un modèle faible est un algorithme dont les performances sont à peine supérieures à celles d'un modèle aléatoire. Dans les méthodes de *boosting*, combiner des modèles qui sont séparément faibles (en *underfitting*) permet de réduire le biais.

L'algorithme de *Gradient Boosting Decision Trees*, ou plus simplement *Gradient Boosting*, est un algorithme de *boosting* dont les *weak learners* sont des arbres de décision. Dans le cas d'une régression (cas de base), les étapes sont les suivantes :

• Le premier *weak learner* w_1 prédit systématiquement la valeur moyenne $\bar{y}$ des labels.

• Pour créer le second *weak learner* w_2, l'algorithme observe l'écart entre le vecteur $\hat{y_1}$ des prédictions de w_1 et le vecteur y des labels. Le second *weak learner* w_2 est entraîné pour prédire cet écart, qui est le résidu de la prédiction $\hat{y_1}$. Le *weak learner* w_2 est pondéré par une quantité α comprise entre 0 et 1 puis ajouté à w_1. Le résultat de cette opération est un modèle dont les prédictions sont plus proche du vecteur y.

• Ce schéma est répété jusqu'à la construction du dernier *weak learner* w_N où N est le nombre d'arbres choisi par l'utilisateur. Ainsi, à l'étape $k+1$, on ajoute αw_{k+1} au modèle de l'étape k, correspondant lui-même à la somme des *weak learners* précédents : $w_1 + \sum_{i=2}^{k} \alpha w_i$.

Ce processus est en fait une opération de descente de gradient (d'où le nom *Gradient Boosting*) qui fait converger petit à petit les prédictions du modèle agrégé vers le vecteur des prédictions parfaites, c'est à dire vers le minimum de la fonction de perte. Le pas α, constant, est appelé *learning rate*. Il s'agit d'un hyperparamètre fixé par l'utilisateur.

Le *Gradient Boosting* pour la classification

Dans le cas d'une classification, l'algorithme présente quelques changements. Il manipule toujours des valeurs continues, qu'il transforme à la fin en classes en les comparant au seuil de 0.5. Cependant, puisque les labels sont des valeurs binaires, l'algorithme travaille avec des *log-odds*. Les étapes sont les suivantes.

• La prédiction du premier *weak learner* ne correspond pas à la moyenne mais au *log-odds* des observations : $\hat{y_1} = log(\frac{\#\{Y=1\}}{\#\{Y=0\}})$. Cette quantité est convertie en probabilité par application de la fonction logistique.

• Le vecteur des résidus est calculé ainsi : $r_1 = \hat{y_1} - y$ où les labels y_i du vecteur y valent soit 0 soit 1.

• Puis le deuxième *weak learner* est entraîné sur les résidus et renvoie une prédiction résiduelle $\hat{r_2}$. La quantité $\hat{y_1} + \alpha\hat{r_2}$ est convertie en prédiction $\hat{y_2}$ par application de la fonction logistique, et de nouveaux résidus sont calculés à partir de cette prédiction.

• Ce processus est répété jusqu'au dernier *weak learner*. La prédiction finale est alors :

$$\frac{1}{1 + exp\left(-(\hat{y_1} + \alpha\hat{r_2} + ... + \alpha\hat{r_N})\right)}$$

Il s'agit de l'estimation par l'algorithme du vecteur des probabilités d'appartenance des observations à la classe positive. Le vecteur des prédictions de classe des observations est ensuite déduit de la comparaison de ces probabilités au seuil 0,5.

2.7 Outils d'évaluation de la performance

Les performances d'un modèle de classification peuvent être mesurées à l'aide de différentes métriques. La métrique la plus courante est le taux de bien classés (ou *accuracy*), défini par la part de prédictions correctes parmi toutes les observations. Cependant, cette mesure présente rapidement des limites. Dans notre cas de classes déséquilibrées, où 81% des instances ont pour réponse $Y = 0$, un modèle prédisant systématiquement cette réponse aurait une mesure d'*accuracy* de 81%, ce qui correspond à un très bon score. Par conséquent, nous utilisons d'autres mesures de performance plus fiables basées sur la matrice de confusion.

La matrice de confusion :

Dans le cadre d'une variable réponse binaire, le modèle peut fournir 4 types de prédictions pour une observation labellisée passée en entrée :

- Vrai positif : l'observation fait partie de la classe positive $Y = 1$ et la prédiction du modèle est conforme à la réalité.
- Vrai négatif : l'observation fait partie de la classe négative $Y = 0$ et la prédiction du modèle est conforme à la réalité.
- Faux positif : le modèle se trompe et prédit la classe $Y = 1$ pour l'observation, qui fait en réalité partie de la classe $Y = 0$. Cela correspond à une erreur de type I.
- Faux négatif : le modèle se trompe et prédit la classe $Y = 0$ pour l'observation, qui fait en réalité partie de la classe $Y = 1$. Cela correspond à une erreur de type II.

La matrice de confusion est un tableau des effectifs de ces quatre types de prédictions sur un échantillon labellisé. Elle se présente sous la forme suivante :

		Prédiction	
		Négatif	Positif
Réalité	Négatif	Vrai négatif (VN)	Faux positif (FP)
	Positif	Faux négatif (FN)	Vrai positif (VP)

TABLE 3 – Matrice de confusion

A partir de cette matrice de confusion, il est possible de calculer différentes mesures de la performance du modèle :

- Le rappel (*recall*, sensibilité ou encore taux de vrais positifs) donne la part des instances positives qui sont correctement prédites comme étant positives par le modèle. Il est défini par :

$$Rappel = \frac{VP}{VP + FN}$$

- La précision représente la part des instances classées comme étant positives qui le sont réellement. Elle est définie par :

$$Précision = \frac{VP}{VP + FP}$$

- La spécificité est analogue à la précision pour la classe négative : elle représente la part des instances classées négatives qui le sont réellement. Elle est définie par :

$$Spécificité = \frac{VN}{VN + FP}$$

- Le taux de faux positifs est défini par :

$$TFP = \frac{FP}{VN + FP} = 1 - spécificité$$

Remarquons que l'*accuracy* s'écrit : $\frac{VP+VN}{VP+VN+FP+FN}$.

Contrairement à l'*accuracy*, les mesures ci-dessus rendent compte du type de l'erreur commise par le modèle. La mesure à privilégier dépend du type d'erreur que nous souhaitons le plus éviter. En effet, chercher à augmenter la précision fait souvent diminuer le rappel, et réciproquement. Il faut donc chercher un compromis. Le F-Score combine le rappel et la précision à l'aide d'un paramètre β qui permet de contrôler l'importance relative du rappel par rapport à la précision :

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{\text{precision } \cdot \text{ rappel}}{(\beta^2 \cdot \text{ precision }) + \text{ rappel}}$$

Lorsque l'on souhaite donner autant d'importance au rappel qu'à la précision, il faut choisir $\beta = 1$. On obtient alors le F1-Score :

$$F_1 = 2 \cdot \frac{\text{precision } \cdot \text{ rappel}}{\text{precision } + \text{ rappel}} = \frac{\text{VP}}{\text{VP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

La courbe ROC et l'AUC :

La courbe ROC (*Receiver Operating Characteristic*) et l'AUC (*Area Under the ROC Curve*) sont des mesures indiquant à quel point un modèle de classification binaire parvient à séparer les classes. Elles permettent de comparer le comportement de différents modèles. La courbe ROC d'un modèle est obtenue en traçant le taux de vrais positifs contre le taux de faux positifs, pour différents seuils de décision. L'AUC correspond à l'aire sous la courbe ROC. La figure 16 présente un exemple de courbe ROC, associée à un AUC de 0,79 :

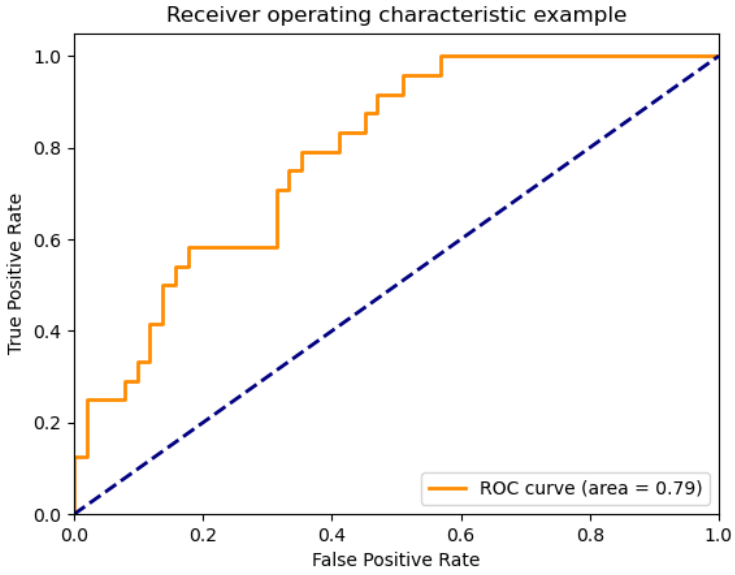

FIGURE 16 – Exemple de courbe ROC (*source : documentation scikit-learn*[2])

Toute courbe ROC passe par les points (0 ; 0) et (1 ; 1) correspondant respectivement à un seuil de décision de 100% et 0%, c'est-à-dire à un classifieur qui prédit respectivement toujours la classe négative et la classe positive. Un modèle parfait est un modèle qui prédit 100% de vrais positifs et aucun faux positif. La courbe de ce modèle passe donc par le point (0 ; 1), et l'AUC vaut 1. Un modèle aléatoire prédit autant de faux positifs que de faux négatifs. Sa courbe ROC correspondante passe donc par le point (0,5 ; 0,5), il s'agit de la première bissectrice. Un classifieur quelconque performe au moins aussi bien qu'un classifieur aléatoire. La courbe ROC des modèles évolue donc toujours entre la première bissectrice et le coin supérieur gauche de la figure, et l'AUC est compris entre 0,5 et 1. Plus l'AUC se rapproche de 1, meilleure est la performance du modèle.

2.8 Passage des classes aux probabilités

Pour savoir si un modèle de classification permet d'estimer correctement les probabilités, une bonne approche consiste à comparer les probabilités prédites à l'observé. Par exemple, considérons l'ensemble des observations pour lesquelles le modèle a prédit une probabilité autour de 30%. Si le modèle prédit correctement les probabilités, alors la part d'observations effectivement positives parmi cet ensemble doit être proche de 30%.

Cette observation peut être faite à l'aide d'un diagramme de fiabilité (*reliability diagram*, aussi appelé *calibration curve*). Ce diagramme permet de confronter les probabilités prédites aux taux réels d'observations positives parmi les données. Il doit si possible être réalisé à partir de données qui n'ont pas servi à l'entraînement. Les étapes de construction du diagramme sont les suivantes :

1. Le modèle de classification est appliqué aux observations afin d'estimer les probabilités d'appartenance à la classe positive.
2. Les observations sont séparées en groupes selon des tranches de la probabilité prédite.
3. Dans chaque groupe, le taux d'observations réellement positives, ainsi que la moyenne des probabilités prédites sont calculés.
4. Ces deux quantités sont placées sur un graphique sous la forme d'un nuage de points. Il s'agit du diagramme de fiabilité.

Si les probabilités sont correctement estimées, elles concordent avec les taux observés et le nuage de points forme une droite le long de la première bissectrice. Dans le cas contraire, les probabilités ne sont pas correctement estimées, quand bien même le modèle affiche de bonnes métriques de classification.

La figure 17 présente un exemple de diagramme de fiabilité, tracé pour un modèle fictif. Les faibles probabilités prédites par ce modèle semblent sous-estimées, tandis que les fortes probabilités semblent surestimées par rapport aux taux observés empiriquement.

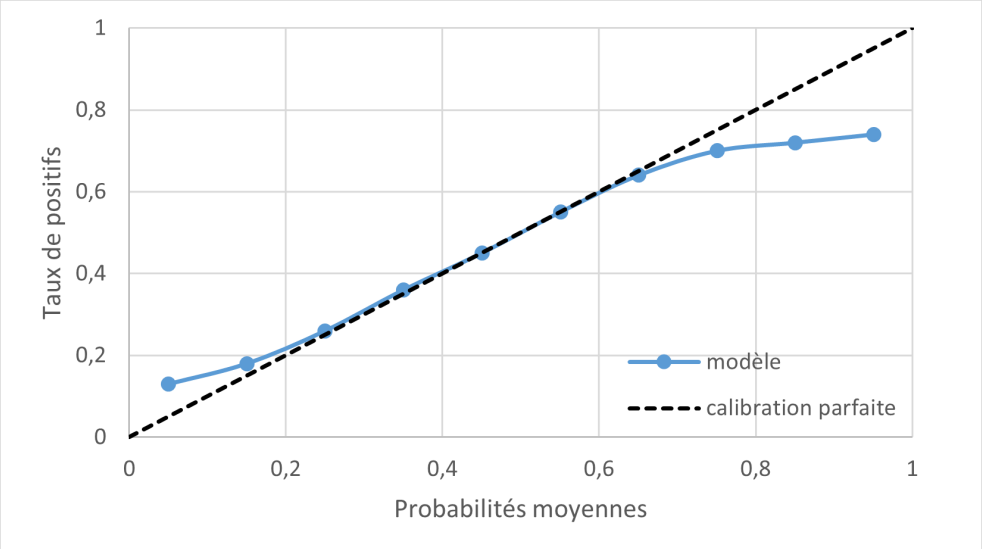

FIGURE 17 – Exemple de diagramme de fiabilité

3 Résultats de la modélisation

3.1 Préparation des données

L'essentiel du traitement des données a été opéré lors de la phase de construction de la base d'étude. Néanmoins, il faut encore opérer un traitement spécifique des données en vue de la modélisation. Il s'agit d'une étape cruciale car les performances des modèles en sont grandement dépendantes. C'est pourquoi, nous avons procédé à un traitement minutieux des variables avant l'étape de modélisation.

Découpage des variables quantitatives

Le découpage en tranches (ou *binning*) des variables continues possède plusieurs avantages. En général, cela peut permettre d'obtenir de meilleures performances de modélisation. D'un point de vue pratique, cela permet en outre de travailler avec un unique type de variables : les variables catégorielles. Le travail en sera ainsi facilité par la suite. Enfin, grâce au découpage en tranches, la présence éventuelle d'outliers parmi les variables n'est pas un problème.

Pour découper les variables continues, un premier découpage grossier est effectué afin de calculer les taux de transformation sur les tranches obtenues. En fonction du résultat obtenu, des tranches adjacentes sont regroupées si leurs taux de transformation sont proches ou si les effectifs sont trop réduits. Toutes les variables continues ont été découpées de cette manière, à l'exception des zoniers et de la prime. Pour ces dernières, un découpage par quantiles a été opéré. Les zoniers ont ainsi chacun été découpés en trois quantiles, représentant trois niveaux de risque (faible, modéré, élevé), tandis que la prime a été découpée en déciles.

Regroupement de modalités

Il est préférable d'éviter d'entrer des variables trop riches en modalités dans la modélisation. C'est pourquoi nous regroupons certaines modalités au sein des variables profession et garanties. Ces regroupements sont effectués sur la base de critères de proximité des modalités, d'effectifs faibles et de proximité des taux de transformation.

Etude des corrélations

L'analyse exploratoire et les traitements supplémentaires ci-dessus nous ont permis d'effectuer une première sélection de variables à conserver dans la modélisation. Parmi les 170 variables de la base de données, 27 ont été retenues à ce stade. Il convient à présent d'étudier le lien entre les différentes variables retenues, ainsi que leur lien avec la variable à prédire. D'une part, les variables conservées doivent être statistiquement liées à la variable à prédire. D'autre part, elles ne doivent pas être corrélées. En effet, une trop forte corrélation entre des variables peut empêcher la convergence du modèle de régression logistique, et peut nuire plus généralement à l'interprétation de l'influence des variables dans les modèles.

Les variables étant toutes catégorielles, la mesure de corrélation utilisée est le V de Cramer. La figure 18 présente la *heatmap* du V de Cramer pour chaque couple de variables présélectionnées.

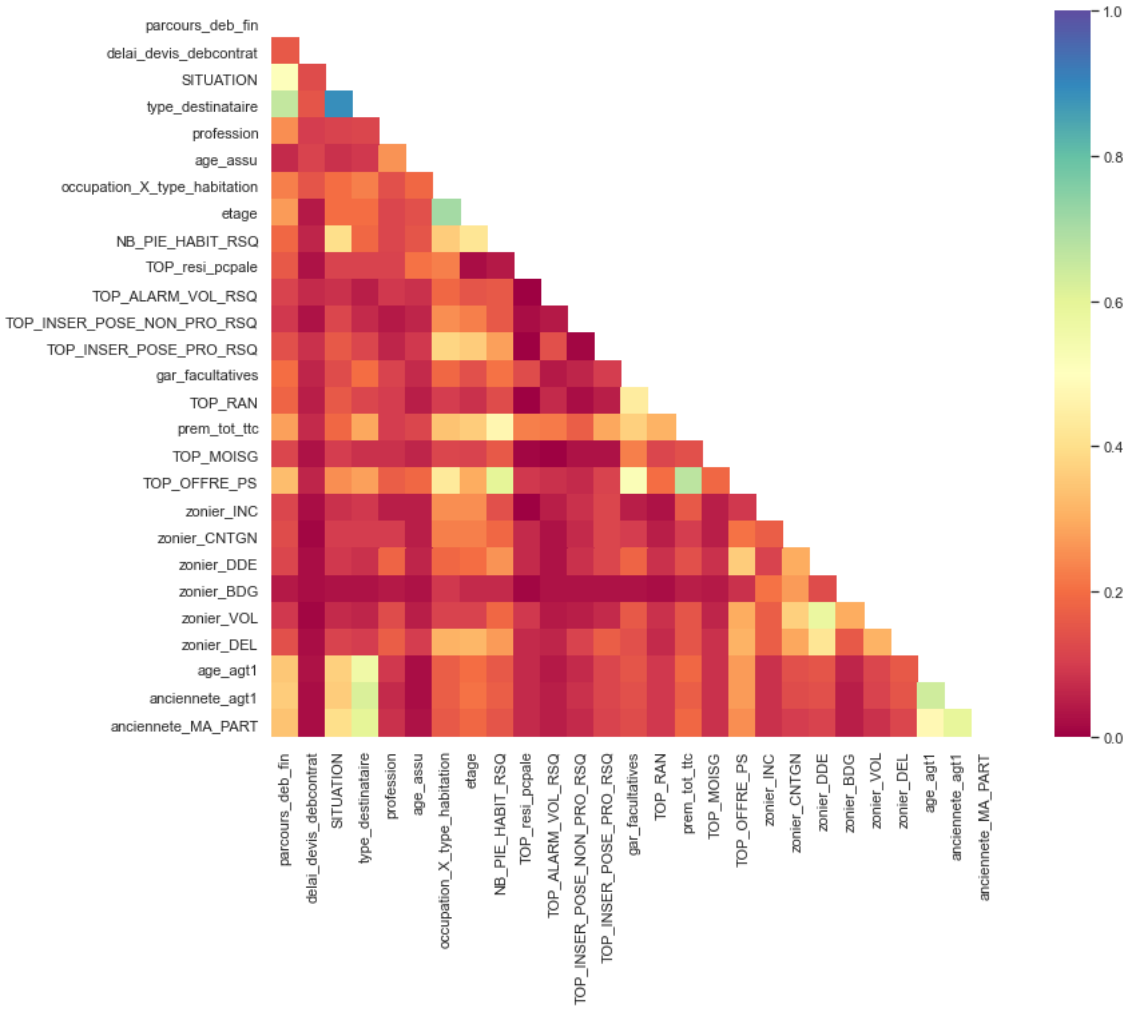

FIGURE 18 – *Heatmap* du V de Cramer entre les variables candidates

Nous fixons le seuil de décision à 0,5. Les couples de variables dont le V de Cramer est supérieur à ce seuil sont présentés dans le tableau 4 :

Variable 1	**Variable 2**	**V de Cramer**
Situation	Destinataire	0,89
Etage	Qualité juridique × Type d'habitation	0,71
Offre petites surfaces	Prime	0,67
Parcours	Destinataire	0,66
Ancienneté de l'agent	Age de l'agent	0,64
Ancienneté de l'agent	Destinataire	0,62
Ancienneté de l'agent en multi-accès	Destinataire	0,60
Ancienneté de l'agent	Ancienneté de l'agent en multi-accès	0,59
Zonier dégât des eaux	Zonier vol	0,58
Age de l'agent	Destinataire	0,56
Offre petites surfaces	Garanties	0,52
Parcours	Situation	0,51

TABLE 4 – Mesure du V de Cramer entre les variables

Ces couples sont trop fortement liés, c'est pourquoi nous décidons de retirer une des variables de chaque couple. Ainsi, nous retirons dans un premier temps les variables suivantes de la modélisation :

- l'offre petites surfaces
- le zonier dégâts des eaux
- l'étage

Avant de retirer les autres variables fortement corrélées, nous étudions le lien entre variables candidates et la variable à prédire. Le tableau 5 présente les variables dont le V de Cramer avec la variable à prédire est supérieur à 0,1 :

Variable	**V de Cramer**
Situation	0,73
Destinataire	0,73
Ancienneté de l'agent	0,33
Ancienneté de l'agent en multi-accès	0,33
Age de l'agent	0,20
Prime	0,19
Qualité juridique × Type d'habitation	0,19
Parcours	0,17
Délai devis/début de contrat	0,15
Nombre de pièces	0,13
Age de l'assuré	0,12
Profession	0,11
Etage	0,10

TABLE 5 – Mesure du V de Cramer entre les variables et la variable réponse

Nous remarquons que la situation et le type de destinataire ont un trop fort pouvoir explicatif de la variable TOP_AFN. L'analyse descriptive avait révélé que la modalité "Souscription" impliquait dans 97% des cas une conversion. Cette information est donc à exclure du modèle car elle comporte un biais. Il en est de même pour l'information du destinataire, qui lui est corrélée. Les variables agent sont également retirées car elles ne sont disponibles que pour une partie des devis (ceux envoyés en agence). Les autres valeurs observées pour le V de Cramer sont pour la plupart assez faibles, ce que nous expliquons par le déséquilibre entre les deux classes de la variable à prédire, qui rend les corrélations difficiles à déceler.

Par ailleurs, un examen plus approfondi des corrélations révèle que les variables d'équipement et de garanties comportent un biais. En effet, le choix de certaines garanties est quasiment exclusif aux devis Normal Quote. Les choix d'équipements ne sont quant à eux pas proposés de la même manière à tous les parcours. Par conséquent, nous décidons de retirer ces variables.

Les variables finalement retenues pour la modélisation figurent dans le tableau 6. Elles sont ensuite dichotomisées, c'est-à-dire décomposées en de nouvelles variables binaires, chacune associée à une modalité d'une variable d'origine.

Variable	Description	Nombre de modalités
TOP_resi_pcpale	résidence principale	1
TOP_MOISG	offre mois gratuits	1
parcours_deb_fin	parcours digital	4
delai_devis_debcontrat	délai avant le début de contrat	3
profession	CSP	6
age_assu	âge de l'assuré	5
occupation_X_type_habitation	Qualité juridique et occupation	4
NB_PIE_HABIT_RSQ	nombre de pièces	4
prem_tot_ttc	prime	10
zonier_INC	zonier incendie	3
zonier_CNTGN	zonier catastrophes naturelles et technologiques	3
zonier_BDG	zonier bris des glaces	3
zonier_VOL	zonier vol	3
zonier_DEL	zonier dégâts électriques	3

TABLE 6 – Liste des variables explicatives retenues pour la modélisation

Retrait des modalités de référence

Les variables ayant été dichotomisées, nous définissons pour chaque variable de départ une modalité de référence à retirer du modèle. Dans le cas du GLM, cette opération est nécessaire pour que le modèle puisse converger. En effet, le maintien de toutes les modalités qui composent les anciennes variables empêche l'inversion de la matrice puisque les prédicteurs forment une famille liée. Le choix des modalités de référence à retirer n'impacte pas les performances du modèle, mais modifie la valeur et l'interprétation de ses coefficients. Nous avons donc choisi les modalités qui représentent l'état de base des différentes variables. Les modalités choisies figurent dans le tableau 7 :

Variable	Modalité de référence
qualité juridique × occupation	Locataire d'appartement
parcours	AZFR - NormalQ
délai avant le début de contrat	7j ou moins
CSP	Employé/Ouvrier/Salarié
âge de l'assuré	25 ou moins
nombre de pièces	1-2
prime	0-100
zonier incendie	6-12
zonier Catnat	1-13
zonier bris de glace	6-18
zonier vol	1-13
zonier dégâts électriques	1-4

TABLE 7 – Liste des modalités de référence

Séparation des échantillons

Enfin, la base ainsi préparée pour la modélisation a été séparée aléatoirement en un échantillon d'apprentissage et un échantillon de test, selon les proportions suivantes : 75% pour l'apprentissage et 25% pour le test.

3.2 Ajustement d'une régression logistique

Une fois les données préparées pour la modélisation, nous entraînons nos modèles de conversion. Le premier modèle est une régression logistique. Celle-ci a été ajustée sur la base d'apprentissage, à l'aide de la librairie python *statsmodels*.

3.2.1 Sélection des variables

L'enjeu majeur de l'affinage d'un GLM est l'étape de sélection des variables. Nous partons du modèle saturé, c'est-à-dire comportant toutes les variables, et suivons une approche itérative de retrait d'une variable. A chaque itération, nous retirons la variable dont le coefficient présente la plus forte p-value, puis nous ajustons à nouveau le modèle sur la base d'apprentissage. Nous arrêtons cette procédure lorsque l'AIC ne s'améliore plus. Le tableau 8 présente les variables retirées à chaque étape et l'AIC du modèle réajusté :

Variable retirée	**AIC du nouveau modèle**
aucune	83024,8
zonier_CNTGN_17-29	83022,8
zonier_BDG_21-28	83020,9
zonier_VOL_14-23	83019,1
NB_PIE_HABIT_RSQ_5-6	83017,6
zonier_INC_19-23	**83016,8**
age_assu_26-30	83018,8

TABLE 8 – AIC des modèles réajustés après les retraits successifs d'une variable

Le retrait de la variable *age_assu_26-30* détériore l'AIC, c'est pourquoi nous arrêtons la procédure après le retrait de la variable *zonier_INC_19-23*.

3.2.2 Interprétation des coefficients

Les coefficients du modèle sont représentés sur la figure 19 :

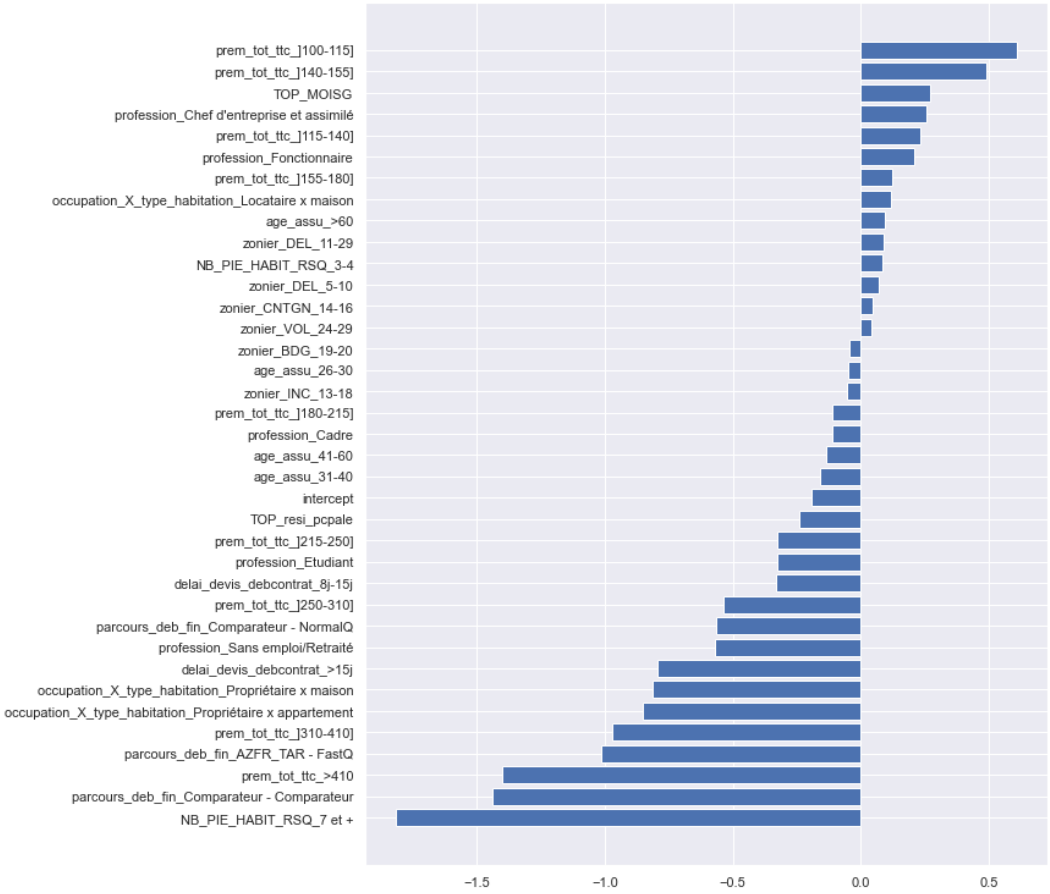

FIGURE 19 – Coefficients du GLM

Afin d'interpréter ces coefficients, nous calculons les *odd ratios*. La liste complète des *odd ratios* figure en annexe. Le tableau 9 en donne les plus extrêmes :

Variable	***Odd ratio***
prem_tot_ttc_]100-115]	1,84
prem_tot_ttc_]140-155]	1,63
TOP_MOISG	1,31
occupation_X_type_habitation_Propriétaire × maison	0,45
occupation_X_type_habitation_Propriétaire × appartement	0,43
prem_tot_ttc_]310-410]	0,38
parcours_deb_fin_AZFR_TAR - FastQ	0,36
prem_tot_ttc_>410	0,25
parcours_deb_fin_Comparateur - Comparateur	0,24
NB_PIE_HABIT_RSQ_7 et +	0,16

TABLE 9 – Liste des *odds ratios* les plus extrêmes

L'interprétation de ces *odd ratios* se fait par rapport aux modalités de référence. A la hausse, les *odd ratios* les plus forts concernent la prime et les mois gratuits. Ainsi, d'après le modèle, les chances qu'un devis soit converti sont :

- 1,84 fois plus élevées si sa prime est comprise entre 100 et 115€ plutôt qu'inférieure à 100€

- 1,63 fois plus élevées si la prime est comprise entre 140 et 155€, toujours en comparaison d'une prime inférieure à 100€.
- 1,31 fois plus élevées si des mois gratuits sont appliqués.

A la baisse, les *odd ratios* les plus forts concernent le nombre de pièces, le parcours, les plus fortes primes et la qualité juridique. Ainsi, les chances que le devis soit converti sont, d'après le modèle :

- 6 fois plus faibles si le bien possède 7 pièces ou plus, plutôt qu'une ou deux pièces
- 4,2 et 2,7 fois plus faibles si le prospect s'arrête respectivement sur le comparateur ou en Fast Quote, plutôt qu'en Normal Quote.
- 4,1 et 2,6 fois plus faibles si la prime est supérieure respectivement à 410 et 310€, plutôt qu'inférieure à 100€.
- 2,3 fois plus faibles pour un propriétaire de maison ou d'appartement par rapport à un locataire d'appartement.

Les *odd ratios* proches de 1 correspondent aux variables dont l'effet direct est jugé neutre par le modèle, relativement aux modalités de référence. Il s'agit de la plupart des zoniers ainsi que de la modalité d'âge 26-30 ans.

3.3 Ajustement de modèles *Random Forest* et *Gradient Boosting*

En complément du modèle GLM, des modèles *Random Forest* et *Gradient Boosting* sont ajustés sur la base d'apprentissage grâce à la librairie python de machine learning *scikit-learn*.

3.3.1 Choix des hyperparamètres

Avant d'entraîner les modèles, il faut les configurer en optimisant leurs principaux hyperparamètres. Pour cela, nous effectuons un *Grid Search*. Nous choisissons les hyperparamètres que nous souhaitons optimiser, ainsi qu'une liste de valeurs candidates pour chacun de ces hyperparamètres. Nous renseignons également une métrique d'évaluation. L'algorithme *Grid Search* construit alors un modèle pour chaque combinaison de valeurs des hyperparamètres et l'évalue sur la métrique spécifiée. De plus, nous effectuons cette recherche en validation croisée 5-fold. Ainsi, pour chaque combinaison d'hyperparamètres, un modèle est entraîné en 5 itérations, sur 5 échantillons différents. L'algorithme compare les scores de tous les modèles construits et retourne l'ensemble optimal d'hyperparamètres. Le modèle qui a été entraîné avec cette combinaison d'hyperparamètres est conservé.

Les algorithmes étant déjà assez gourmands en temps de calcul, l'application d'un *Grid Search* en validation croisée 5-fold peut rendre le temps de calcul excessif, en particulier pour le *Gradient Boosting*. Par conséquent, nous limitons le nombre d'hyperparamètres et de valeurs candidates à tester.

Les hyperparamètres que nous optimisons pour le *Random Forest* sont :

- la profondeur maximale des arbres : 4,6,8 ou 10
- le nombre d'arbres : 100, 200 ou 300.

La métrique à optimiser est l'AUC. Les différentes combinaisons aboutissent à des valeurs assez proches d'AUC, toutes supérieures à 0,75. De plus, les valeurs d'AUC sont très homogènes sur les différents folds, ce qui montre que les performances de l'algorithme ne sont a priori pas

trop dépendantes de l'échantillon. La combinaison d'hyperparamètres retenue pour le modèle est celle maximisant l'AUC moyen des 5 folds :

Profondeur max	Nombre d'arbres	AUC min	AUC max	AUC moyen
10	200	0,769	0,779	0,774

TABLE 10 – Hyperparamètres du modèle *Random Forest*

Le *Grid Search* a choisi la profondeur maximale proposée, ce qui suggère que les performances auraient potentiellement pu être encore améliorées en ajoutant des profondeurs supérieures. Néanmoins, nous ne souhaitons pas rendre le modèle trop lourd, et nous remarquons par ailleurs que les profondeurs plus faibles aboutissaient tout de même à des valeurs d'AUC très proches de l'AUC optimal.

Pour le *Gradient Boosting*, les hyperparamètres que nous optimisons sont :

- la profondeur maximale des arbres : 3, 4 ou 5
- le nombre d'arbres : 100 ou 200.
- le *learning rate* : 0,05 0,1 ou 0,2

Comme pour le *Random Forest*, les différentes combinaisons testées par le *Grid Search* ont des valeurs assez proches d'AUC moyen, toutes supérieures à 0,77. Les valeurs d'AUC sont également très homogènes entre les différents folds. La combinaison d'hyperparamètres retenue est :

Profondeur	Nombre d'arbres	Learning rate	AUC min	AUC max	AUC moyen
4	200	0.2	0,779	0,789	0,784

TABLE 11 – Hyperparamètres du modèle *Gradient Boosting*

3.3.2 Importance des variables

Les deux figures ci-dessous permettent de visualiser l'importance relative des variables pour les deux modèles :

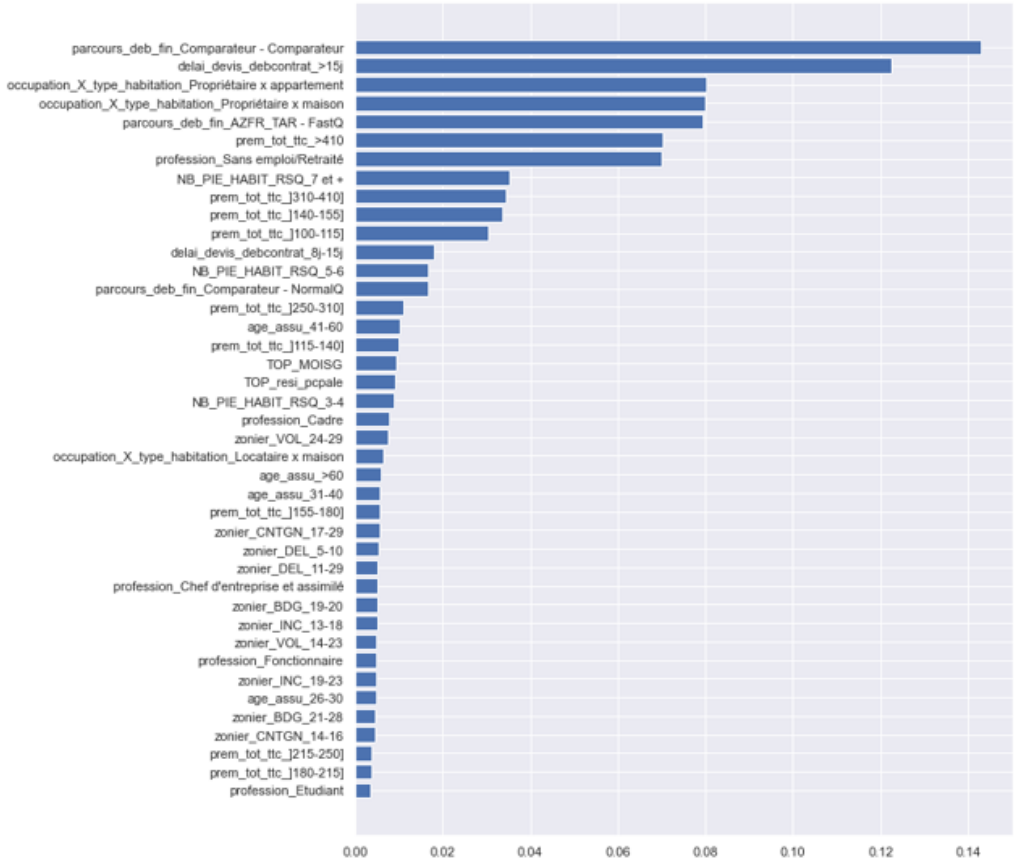

FIGURE 20 – Importance relative des variables dans le modèle *Random Forest*

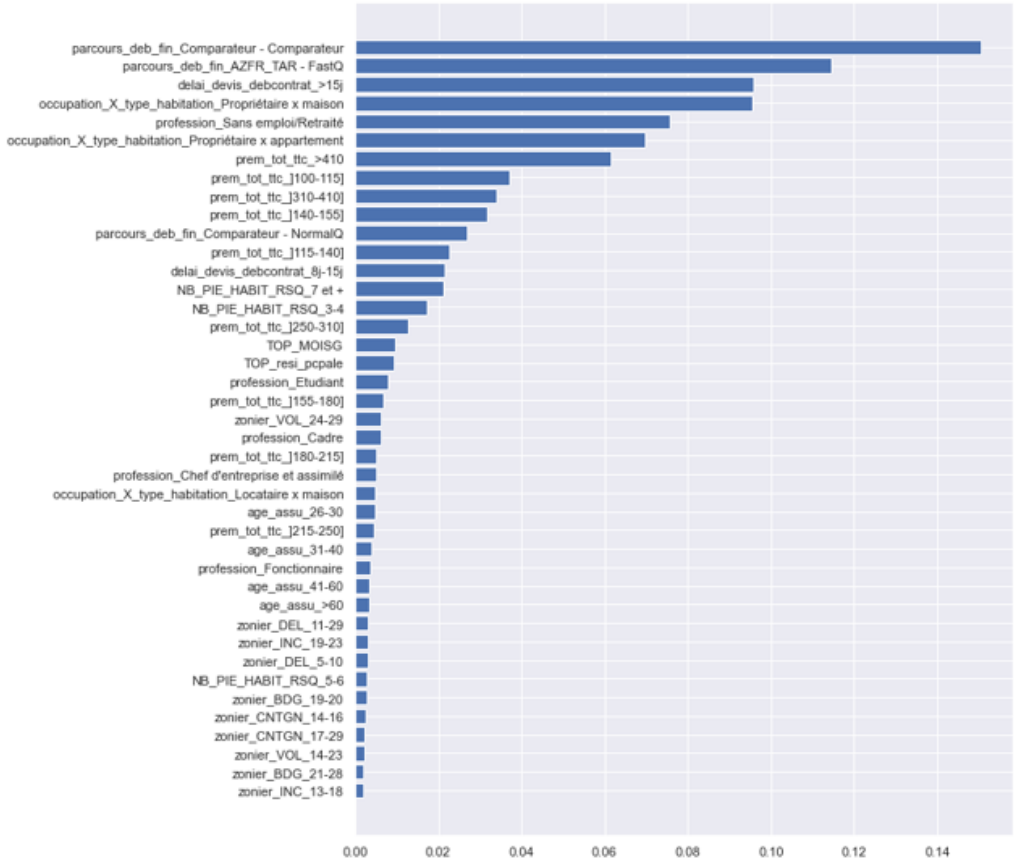

FIGURE 21 – Importance relative des variables dans le modèle *Gradient Boosting*

Nous constatons que les deux classements sont similaires, à quelques échanges près. Ainsi, les variables les plus importantes pour les deux algorithmes sont le parcours, le délai souhaité avant le début du contrat, et le croisement qualité juridique × type d'habitation. La catégorie des retraités et les différentes tranches de primes figurent également parmi les variables les plus importantes. Nous retrouvons ainsi la plupart des variables impactantes du GLM parmi les variables à forte importance pour le *Random Forest* et le *Gradient Boosting*, ce qui est rassurant.

Les variables les moins utilisées sont également similaires entre les deux modèles. Comme pour le GLM, il s'agit principalement des zoniers. D'autres modalités apparaissent parmi les variables les moins utilisées par le *Random Forest*, tel que les catégories des étudiants et fonctionnaires, la catégorie d'âge entre 26 et 30 ans et des catégories de primes intermédiaires.

3.4 Comparaison des résultats des modèles

Lors de l'entraînement, le GLM et les algorithmes de machine learning ont été calibrés sur l'échantillon d'apprentissage de manière à optimiser respectivement l'AIC et l'AUC. Nous comparons à présent les performances des différents modèles sur l'échantillon de validation, afin de sélectionner le meilleur d'entre eux.

3.4.1 Comparaison des métriques d'évaluation

L'évaluation à l'aide des métriques nécessite le choix d'un seuil de décision. Pour que les modèles parviennent à bien séparer les classes, il est préférable d'abaisser le seuil de décision de sa valeur par défaut, fixée à 50%. Nous ajustons ainsi le seuil de décision au taux de transformation de l'ensemble de la base d'étude, soit 19,3%. La figure 22 présente les matrices de confusion des modèles, normalisées sur les lignes :

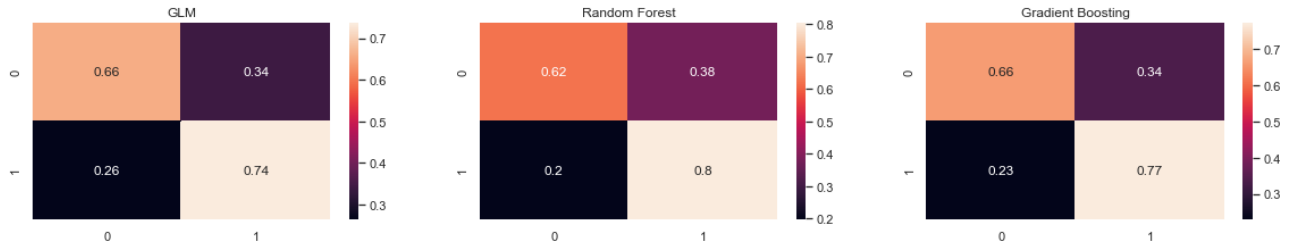

FIGURE 22 – Matrices de confusion au seuil de décision 19,3%

Le tableau 12 présente l'AUC et les métriques obtenues à partir des matrices de confusion, au seuil de décision fixé :

Métrique	**GLM**	***Random Forest***	***Gradient Boosting***
AUC	0,760	0,778	**0,787**
Rappel	0,74	**0,80**	0,77
Précision	0,68	0,68	**0,70**
F1	0,71	**0,74**	0,73
Log loss	0,420	0,417	**0,403**

TABLE 12 – Performances des trois modèles sur l'échantillon de test

Enfin, la figure 23 présente les courbes ROC des trois modèles :

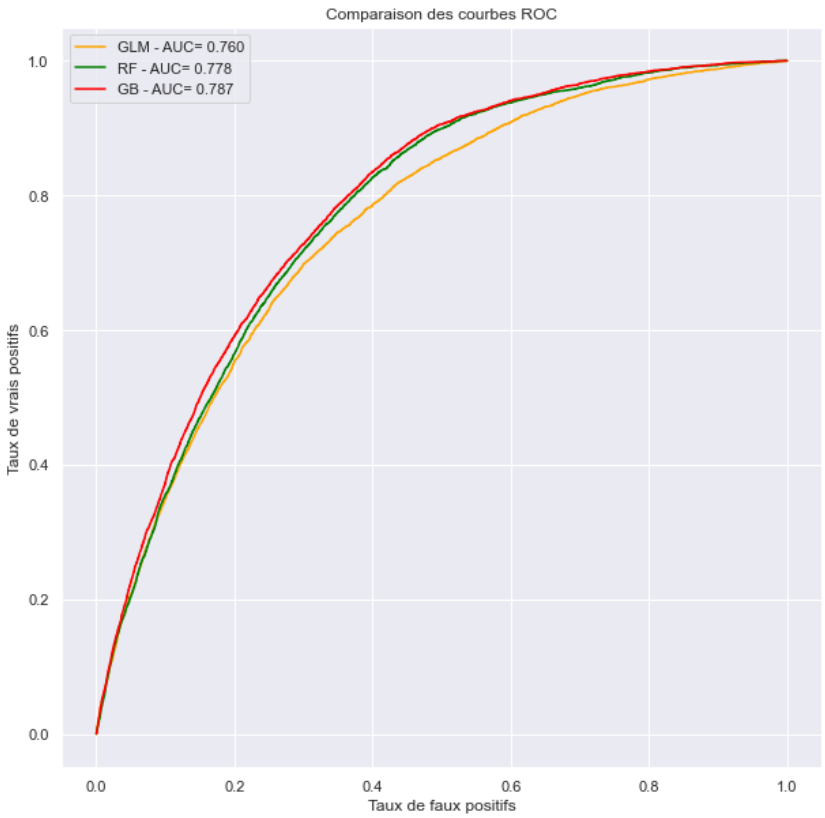

FIGURE 23 – Courbes ROC des trois modèles

Les différentes métriques observées témoignent de bonnes performances des trois modèles. Au seuil de décision fixé, c'est le modèle *Random Forest* qui présente le meilleur rappel, égal à 0,80. Cela signifie que sur 100 conversions, 80 sont effectivement détectées par le modèle. Le *Gradient Boosting* affiche quant à lui la meilleure précision, valant 0,70. Cela signifie que sur 70 prédictions de conversions faites par le modèle, 70 en sont réellement. Nous nous concentrons sur l'AUC, qui ne dépend pas d'un seuil de décision. Les trois modèles obtiennent un score d'AUC compris entre 0,76 et 0,79, ce qui correspond à de bons scores. Ces valeurs élevées d'AUC témoignent de la bonne capacité des modèles à distinguer les classes. Elles sont assez proches entre les trois modèles, mais c'est le modèle *Gradient Boosting* qui présente le meilleur score sur cette métrique. Le tracé des courbes ROC révèle que la courbe de ce modèle domine toujours les deux autres, bien que la courbe du *Random Forest* s'en écarte peu.

3.4.2 Capacité de prédiction de probabilités

D'après les métriques observées, les modèles montrent de bonnes performances en termes de classification. Néanmoins, il faut être prudent lors du passage des prédictions de classes à des prédictions de probabilités, en particulier pour les modèles basés sur les arbres. Il convient donc de vérifier à présent si les estimations de probabilité de conversion par les modèles sont correctement ajustées.

Nous visualisons dans un premier temps la distribution des probabilités prédites par les modèles sur la base de test :

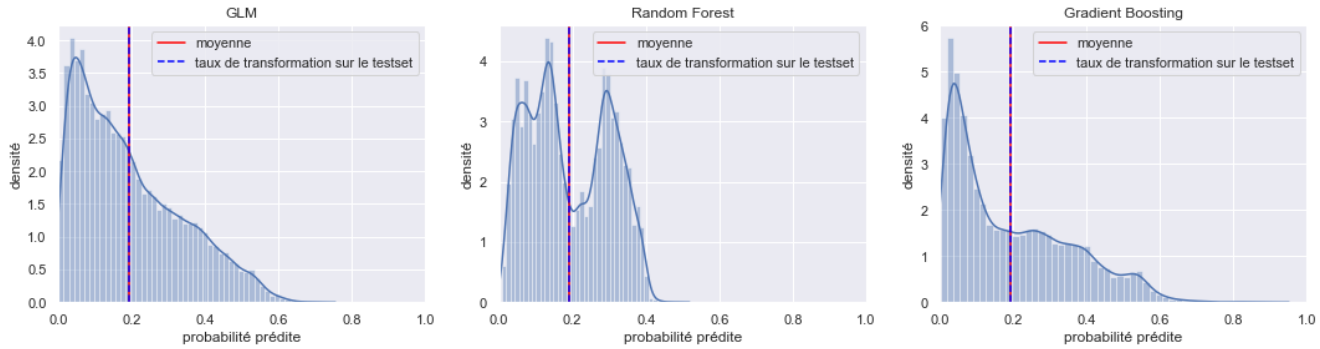

FIGURE 24 – Distribution des probabilités prédites par les trois modèles

Pour les trois modèles, nous observons que la moyenne des probabilités prédites sur l'échantillon de test est confondue avec le taux de transformation observé sur cet échantillon. L'estimation du taux de transformation n'est donc a priori pas biaisée. Les trois figures présentent un pic de densité autour de 5% de probabilité de conversion. Néanmoins, les trois distributions sont assez différentes. La distribution du *Random Forest* se distingue des deux autres par son support plus compact. En effet, les probabilités prédites par ce modèle semblent plafonner à 40%, tandis que les deux autres modèles ont des queues de distribution assez épaisses jusqu'à 60%. Par ailleurs, nous observons deux pics de densité exclusifs au *Random Forest*, aux alentours de 15 et 30%. Le modèle a donc tendance à plus souvent prédire une probabilité de conversion autour de 15 ou 30%, mais presque jamais au-delà de 40%.

Les distributions de probabilité prédite du GLM et du *Gradient Boosting* se ressemblent davantage, mais la première se distingue de la seconde par une diminution plus progressive après le pic de 5%.

Nous traçons à présent les diagrammes de fiabilité des trois modèles sur l'échantillon de test, à l'aide de la fonction *calibration_curve* de la librairie *scikit-learn*. Cette fonction découpe l'ensemble des probabilités prédites par le modèle en tranches, puis calcule la proportion d'observations appartenant réellement à la classe $Y = 1$ au sein de chaque tranche. Cela permet ensuite de tracer le diagramme de fiabilité confrontant la moyenne des probabilités prédites sur chaque tranche au taux de conversion observé sur la tranche. Nous découpons les probabilités en déciles du nombre d'observations, afin de ne pas biaiser l'analyse par une différence du volume d'observations dans les tranches. La figure 25 présente les diagrammes obtenus :

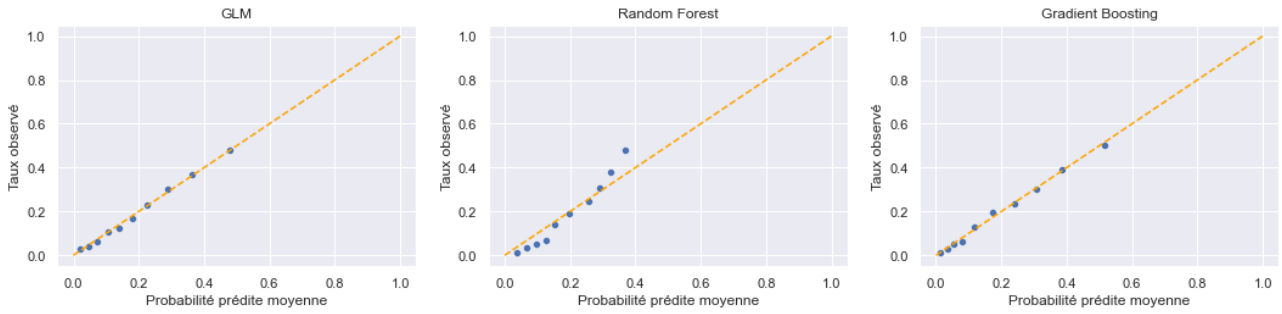

FIGURE 25 – Diagrammes de fiabilité

Plus le nuage de points est confondu avec la première bissectrice et meilleure est la calibration des probabilités prédites par le modèle. Visiblement, les probabilités sont correctement prédites par le GLM et le *Gradient Boosting*, mais pas par le modèle *Random Forest*. En effet,

la convexité de la courbe observée révèle que le modèle sous-estime la probabilité de conversion sur les tranches à fort taux de conversion et surestime cette probabilité sur les tranches à plus faible taux de conversion. Cette observation confirme les analyses de la figure 24 : le modèle prédit des probabilités à 30% pour beaucoup d'observations de tranches dont le taux est supérieur, ce qui fait qu'il les sous-estime. De même, il prédit des probabilités de l'ordre de 15% pour beaucoup d'observations de tranches dont le taux est en réalité pratiquement nul, c'est-à-dire qu'il les surestime.

Il est possible que l'origine de ces erreurs de prédiction des probabilités par le *Random Forest* tienne de la nature de cet algorithme [6]. En effet, les modèles de *bagging* ont parfois des difficultés à prédire correctement les probabilités aux frontières de la distribution, car ils moyennent des prédictions. Pour prédire les valeurs extrêmes de la distribution, il faudrait que tous les prédicteurs soient d'accord, ce qui est compromis par le bruit introduit par l'échantillonnage. Ces algorithmes peuvent alors avoir tendance à pousser les extrêmes de la distribution vers le centre, ce qui correspond a priori à ce que nous observons sur la figure 24 pour le *Random Forest*.

3.4.3 Backtesting

Ce que nous désignons ici par basktesting correspond à la comparaison, par modalité de la variable considérée, entre les probabilités prédites par les modèles sur l'échantillon de test et le taux de transformation véritablement observé.

Nous avons constaté lors du backtesting que le GLM et le *Gradient Boosting* s'ajustent bien aux taux de transformation observés sur la plupart des variables. Le modèle *Random Forest* s'ajuste quant à lui assez mal, comme l'analyse des courbes de calibration l'avait laissé présager.

Voici quelques exemples de graphiques du backtesting :

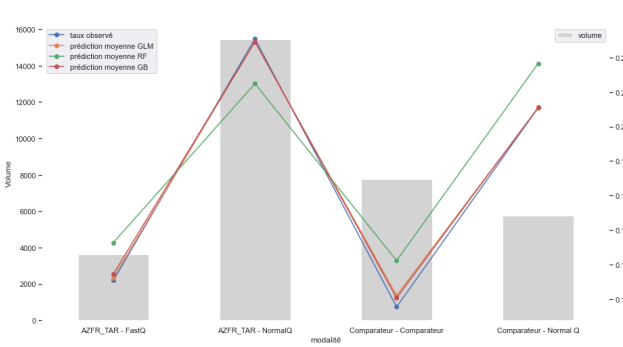

FIGURE 26 – Backtest sur le parcours

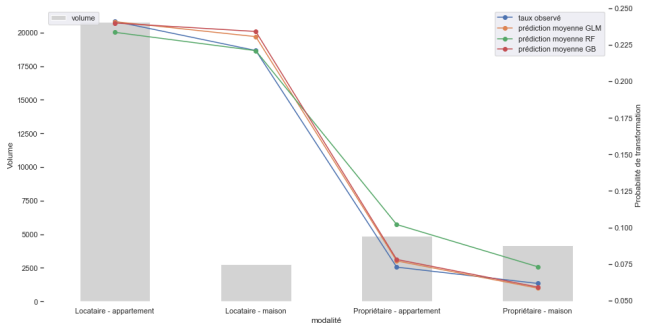

FIGURE 27 – Backtest sur le croisement qualité juridique × type d'habitation

6. (Niculescu-Mizil A. et Caruana R.) [3]

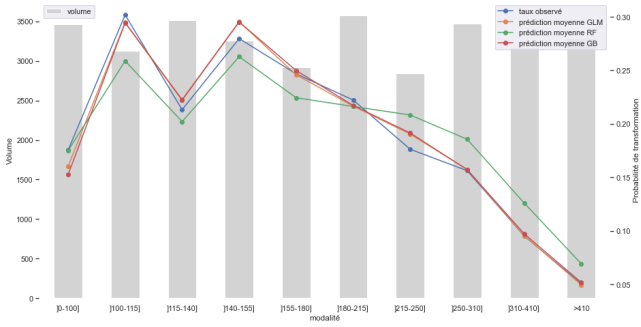

FIGURE 28 – Backtest sur la prime

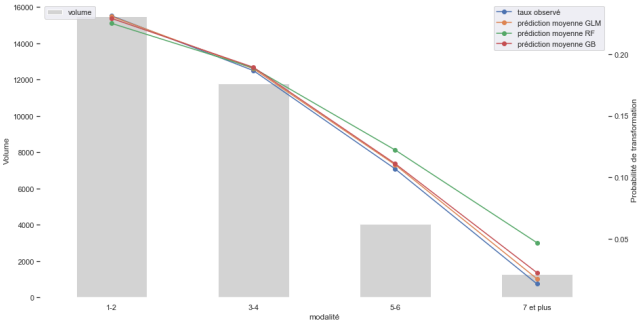

FIGURE 29 – Backtest sur le nombre de pièces

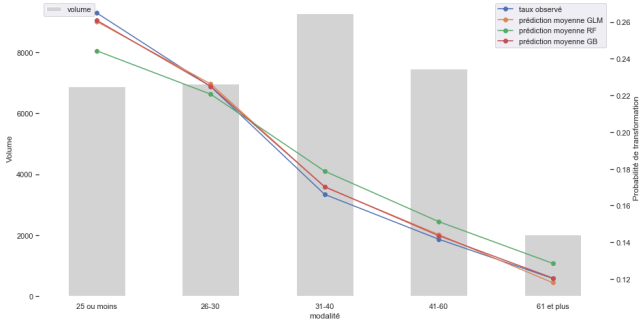

FIGURE 30 – Backtest sur l'âge de l'assuré

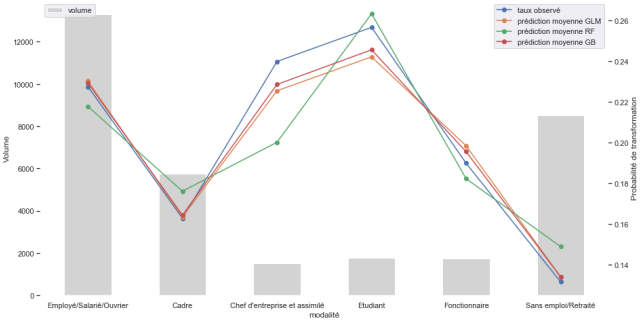

FIGURE 31 – Backtest sur la CSP de l'assuré

3.4.4 Choix du modèle

La comparaison des modèles sur l'échantillon de test a révélé que les trois modèles ont une bonne capacité de généralisation et de bonnes performances en termes de classification. Toutefois, le modèle *Random Forest* ne prédit pas correctement la probabilité de conversion, bien que les métriques d'évaluation soient toutes bonnes et parfois les meilleures sur ce modèle. Le *Gradient Boosting* s'ajuste tout aussi bien que le GLM aux taux observés. Cependant, il domine ce dernier sur toutes les métriques d'évaluation utilisées. Au vu de ces résultats, nous décidons de retenir le modèle *Gradient Boosting*.

3.5 Ajustement d'un GLM centré sur les effets de la prime

En complément des modèles précédents, nous souhaitons ajuster un modèle *logit* prenant davantage en compte les effets de la prime pour prédire la probabilité de transformation. Ce modèle "prix" nous servira à déterminer ultérieurement l'élasticité-prix des devis de manière analytique.

Nous reprenons la base de modélisation ayant servi à entraîner les trois modèles, mais nous conservons la prime en tant que variable continue, au lieu de la discrétiser. Le coefficient associé à la prime dans le *logit* nous permettra de mesurer l'effet simple de la prime sur la probabilité de transformation. Par ailleurs, nous supposons que la prime n'a pas les mêmes effets sur la probabilité de transformation du prospect en fonction du segment auquel celui-ci appartient. Afin de tenir compte de ces effets dans la modélisation, nous ajoutons des variables issues du croisement de la prime avec les variables qualitatives. Nous récapitulons les variables ajoutées au modèle dans le tableau 13 :

Variable	Effet simple	Effet croisé avec la prime
Résidence principale	•	
Mois gratuits	•	
Prime	•	
Parcours	•	
Délai devis - début du contrat	•	
Nombre de pièces	•	
Zoniers	•	
Qualité juridique	•	•
Type d'habitation	•	•
CSP	•	•
Age	•	•

TABLE 13 – Liste des variables du modèle GLM croisé

Le modèle est ajusté sur l'échantillon d'apprentissage, puis il est évalué sur l'échantillon de test. La courbe ROC du modèle est tracée sur la figure 32. Le modèle obtient un score AUC de 0,759, ce qui traduit de bonnes performances en termes de classification.

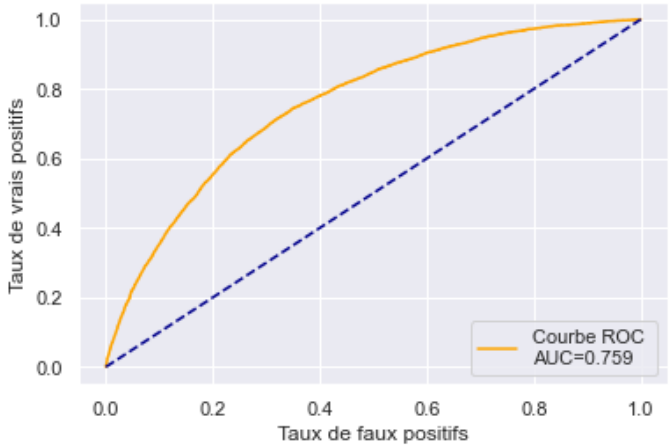

FIGURE 32 – Courbe ROC du modèle

La matrice de confusion est déterminée au seuil de décision de 19,3%. Elle est normalisée sur les lignes et présentée sur la figure 33 :

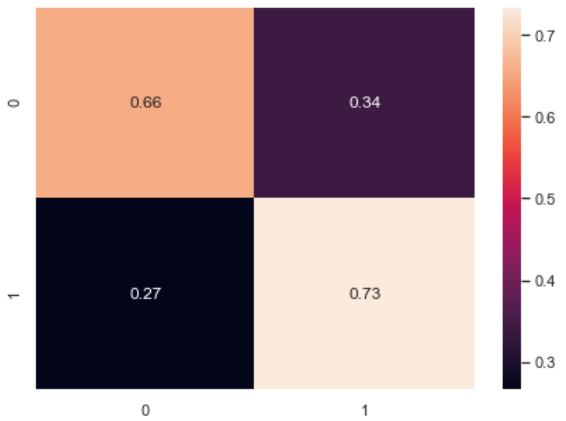

FIGURE 33 – Matrice de confusion au seuil de décision 19,3%

Puisque le modèle a été évalué sur le même échantillon de test et au même seuil de décision, nous pouvons comparer ses métriques déduites de la matrice de confusion avec celles du modèle GLM principal. Nous constatons que les valeurs des métriques sont quasiment identiques entre les deux modèles, ce qui témoigne de performances a priori équivalentes :

Métrique	**GLM principal**	**GLM "prix"**
AUC	0,760	0,759
Rappel	0,74	0,73
Précision	0,68	0,68
Score F1	0,71	0,71
Log loss	0,420	0,421

TABLE 14 – Métriques du modèle GLM prix

Enfin, le diagramme de fiabilité (figure 34) indique que le modèle est correctement ajusté pour l'estimation des probabilités de conversion. En effet, les points sont assez bien alignés le long de la première bissectrice :

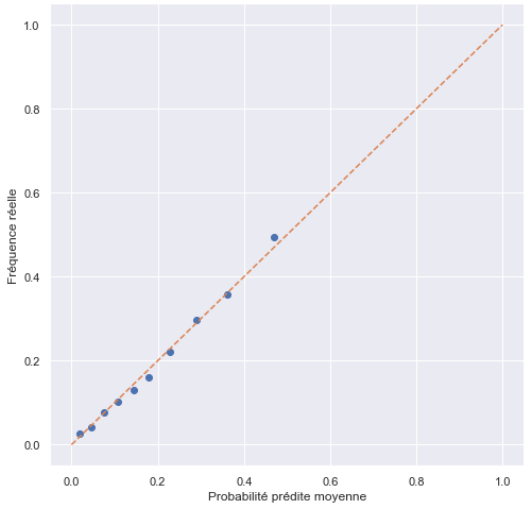

FIGURE 34 – Courbe de calibration du modèle

3.6 Ajustement d'un modèle agent

Nous souhaitons ajuster un dernier modèle prenant en compte les variables liées à l'agent dans la prédiction de la probabilité de transformation des devis envoyés en agence. Nous nous appuierons sur ce modèle lors de l'étude du levier "choix de l'agent".

La base de modélisation principale est filtrée pour conserver uniquement les devis envoyés en agence. Le taux de transformation s'en trouve réduit de 19,3% à 7,3%. Cette fois, les variables liées à l'agent sont ajoutées. Le modèle entraîné est un modèle *Gradient Boosting* dont les variables figurent dans le tableau suivant :

Variable	Nombre de modalités	Modalité de référence
Résidence principale	1	-
Mois gratuit	1	-
Parcours	3	AZFR - Normal Quote
Délai entre devis et début de contrat	3	7j ou moins
CSP	6	Employé/Salarié/Ouvrier
Age	3	25 ans ou moins
Qualité juridique × Type d'habitation	4	Locataire d'appartement
Nombre de pièces	4	1-2
Prime	9	0-100
Zonier incendie	3	6-12
Zonier catnat	2	1-13
Zonier bris de glaces	2	6-18
Zonier vol	3	1-13
Zonier dégâts électriques	3	1-4
Ancienneté de l'agent	3	<1 an
Ancienneté dans le protocole MA	2	<1 an

TABLE 15 – Variables du modèle agent

Les hyperparamètres sont optimisés à l'aide d'un *Grid Search* en validation croisée 5-fold. La combinaison d'hyperparamètres retenue est la suivante :

Profondeur max	Nombre d'arbres	Learning rate
3	100	0,2

TABLE 16 – Hyperparamètres du modèle agent

La figure 35 présente l'importance relative des variables dans le modèle :

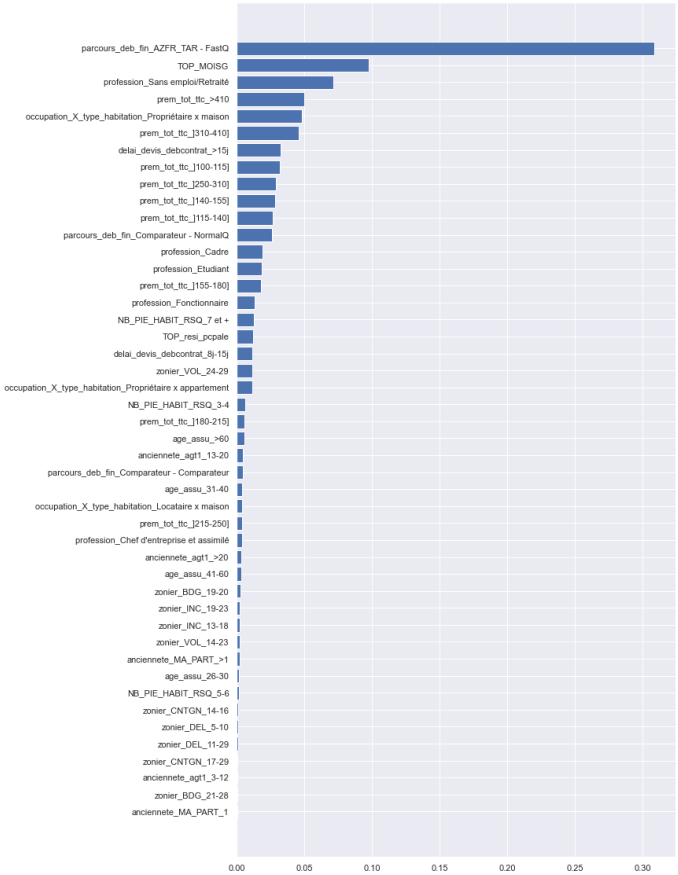

FIGURE 35 – Importance relative des variables dans le modèle agent

Le diagramme obtenu diffère sensiblement de celui du *Gradient Boosting* entraîné sur la base d'étude complète, ce qui traduit a priori les spécificités de la population des devis envoyés en agence. Tandis que plusieurs variables étaient fortement utilisées par le modèle général, la variable de parcours Fast Quote prédomine largement sur les autres variables du modèle agent. Les zoniers sont à nouveau les variables les moins utilisées par le modèle. Par ailleurs, nous constatons que les variables agent sont assez peu utilisées par le modèle.

Les métriques d'évaluation sur l'échantillon de test sont élevées, ce qui témoigne des bonnes performances du modèle. Elles ont été calculées pour un seuil de décision égal au taux de transformation sur l'ensemble de la base des devis envoyés en agence, soit 7.3%.

Métrique	**Mesure sur le *test set***
AUC	0,783
Recall	0,71
Precision	0,71
F1	0,71
Log loss	0,222

TABLE 17 – Métriques d'évaluation du modèle agent

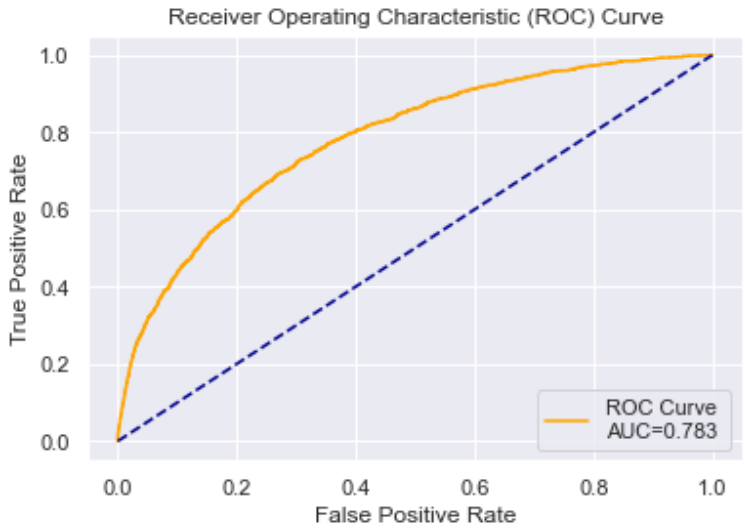

FIGURE 36 – Courbe ROC du modèle agent

Le diagramme de fiabilité est représenté en figure 37. Les points sont alignés le long de la première bissectrice, ce qui suggère que le modèle est correctement calibré pour prédire les probabilités de conversion :

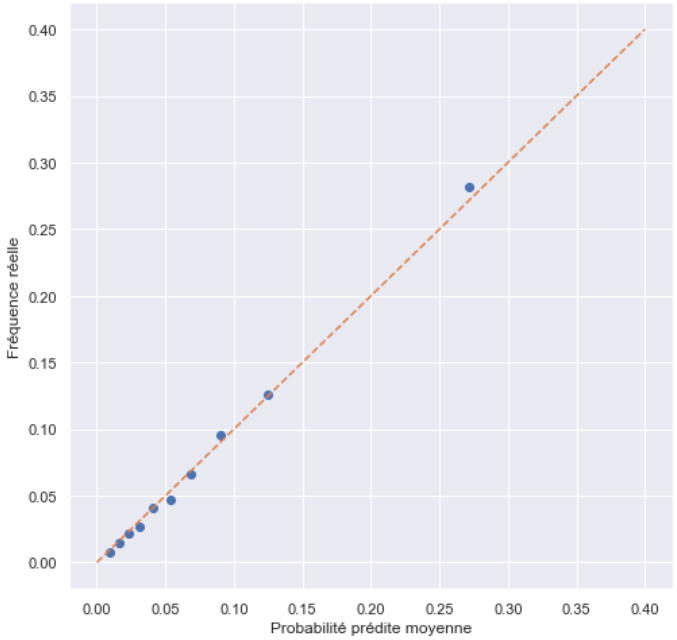

FIGURE 37 – Courbe de calibration des probabilités du modèle agent

Ce dernier modèle conclut l'étape de modélisation de la probabilité de conversion, qui est la première brique nous permettant d'étudier des leviers d'amélioration du taux de transformation. La seconde brique est l'identification des profils rentables, que nous traitons dans la partie suivante.

4 Identification des profils rentables

L'objectif de cette partie est d'identifier des cibles rentables parmi les devis de la base d'étude, sur lesquelles nous souhaiterions améliorer le taux de transformation. Le ciblage nécessite donc le choix d'un critère de rentabilité des devis. Dans cette optique, nous comparons plusieurs indicateurs de rentabilité.

4.1 Choix d'un indicateur de rentabilité

4.1.1 Le ratio de sinistralité et le ratio combiné

Les deux principaux indicateurs de rentabilité d'un portefeuille sont le ratio de sinistralité et le ratio combiné.

Le ratio de sinistralité, encore appelé S/C, S/P ou *Loss Ratio*, est le principal indicateur de rentabilité technique d'un portefeuille de contrats. Il mesure le rapport entre le coût des sinistres et le montant des primes acquises sur la même période :

$$S/C = \frac{charge\,des\,sinistres}{primes\,acquises}$$

La composante sinistre est l'estimation de la charge définitive des sinistres. Il s'agit donc d'une projection à l'ultime. Ainsi, cette composante comprend non seulement les indemnités versées aux assurés, mais également les montants provisionnés pour :

- les sinistres connus mais non réglés
- les sinistres survenus mais dont l'assureur n'a pas encore connaissance
- les réajustements du montant des sinistres passés

Le ratio combiné (*Combined Ratio*) est la somme du ratio de sinistralité et du ratio des coûts de l'assureur :

$$CoR = \frac{charge\,des\,sinistres\ +\ charge\,d'exploitation}{primes\,acquises}$$

La charge d'exploitation englobe les frais généraux (frais d'acquisition, d'administration et de gestion de sinistres) et les commissions versées à l'intermédiaire (agent ou courtier).

Le ratio combiné capte la rentabilité de l'activité en tenant compte des frais de fonctionnement. Le seuil de profitabilité d'un contrat correspond à la valeur de 100% du CoR. Un ratio combiné supérieur à ce seuil traduit une activité déficitaire. En dessous de 100%, l'activité est bénéficiaire.

Ces deux premiers indicateurs ne conviennent pas aux besoins de notre étude. D'une part, ils sont basés sur la sinistralité observée. Or, nous cherchons une mesure de la rentabilité prospective des prospects, au moment de l'affaire nouvelle. D'autre part, ce sont des indicateurs macroscopiques, qui n'ont de sens qu'à l'échelle d'un groupe de contrats. Or, nous cherchons une mesure de la rentabilité à l'échelle du contrat. Ces considérations nous amènent à nous intéresser au PSC, que nous définissons ci-après.

4.1.2 Le PSC

La meilleure estimation du coût espéré des sinistres à l'échelle d'un contrat correspond à sa prime pure. Ainsi, on définit le Projected S/C (PSC), ou *Expected Loss Ratio (ELR)*, par le ratio entre la prime pure à l'ultime calculée pour le contrat et la prime payée par l'assuré :

$$PSC = \frac{EUL}{AP}$$

où :

- *EUL (Expected Ultimate Loss)* est la prime pure projetée à l'ultime.
- *AP (Actual Price)* est le montant hors taxe payé par l'assuré

L'EUL est calculé par passage de la prime pure à l'ultime : une fois la prime pure modélisée, l'EUL est calculé en ajoutant à cette dernière une estimation des provisions pour sinistres à payer (PSAP). L'AP (*Actual Price*) correspond à la prime commerciale après application d'éventuelles modulations tarifaires.

Le PSC est une prédiction du ratio de sinistralité définie à l'échelle du contrat. Il s'agit donc d'un indicateur de la rentabilité technique du contrat, qui indique son niveau de sur- ou sous-tarification par rapport à son risque. Plus le PSC est élevé, plus le contrat est sous-tarifé et donc moins il est rentable pour l'assureur.

Contrairement au S/C observé, qui est un indicateur a posteriori et défini à l'échelle d'un groupe de contrats, le PSC est un indicateur prospectif et défini à l'échelle du contrat. Par conséquent, nous décidons de retenir cet indicateur, qui est plus adapté au cadre de notre étude.

4.1.3 Construction d'une base de PSC

En tant qu'indicateur prospectif et individuel, le PSC pourrait en théorie être calculé pour chaque devis enregistré. Toutefois, il n'est en pratique calculé que pour les affaires nouvelles. Par conséquent, cet indicateur n'était pas disponible pour tous les devis de la base d'étude. En revanche, nous avons pu récupérer les mesures de PSC à l'affaire nouvelle associées à près de 321 000 contrats souscrits entre janvier 2019 et août 2020 et issus des différents parcours multi-accès. Dans un premier temps, il a été envisagé de se restreindre à l'intersection entre ces contrats munis d'un PSC et les affaires nouvelles de la base d'étude. Cependant, cet échantillon est trop restreint (environ 16 000 observations) pour permettre de segmenter correctement les profils.

Nous avons donc décidé de construire une nouvelle base de données des contrats munis d'un PSC. Nous désignerons dans la suite cette base sous les noms de "base de PSC" ou "base secondaire". La segmentation des profils sera établie à partir de cette base de données secondaire, puis appliquée à la base principale contenant les devis étudiés. La majorité des contrats de la base de PSC ne sont pas des contrats d'origine digitale. Néanmoins, les modèles sous-jacents au calcul du PSC ne distinguent pas entre les affaires digitales et non digitales. Autrement dit, toute chose égale par ailleurs, une affaire digitale aura le même PSC qu'une affaire non digitale. Nous supposons donc que les profils construits à partir des contrats non digitaux peuvent être exploités pour des devis digitaux.

Nous avons reproduit la démarche de construction et de traitement de la base principale, afin d'intégrer les mêmes variables dans la base de PSC. Toutefois, nous n'avons pas répété certaines étapes qui avaient été effectuées spécifiquement dans le but de modéliser la conversion, tel que le découpage des variables qualitatives ou le retraitement de certaines modalités. Après nettoyage, la base de PSC obtenue comporte 270 271 contrats.

4.2 Choix d'un seuil d'équilibre du PSC

L'indicateur de rentabilité que nous retenons est le PSC. L'enjeu est de définir une valeur seuil sur cet indicateur, permettant de juger si un contrat est rentable. Ce seuil de PSC peut être choisi à partir du seuil de profitabilité du ratio combiné :

$$PSC_{seuil} = 1 - \text{ratio des dépenses}$$

Il suffirait donc a priori d'estimer le ratio des dépenses pour définir un niveau de PSC d'équilibre. Néanmoins, un tel critère ne prend pas entièrement en compte la rentabilité des prospects. En effet, le PSC mesure la rentabilité d'un contrat pris isolément, et uniquement à l'affaire nouvelle. Or, il convient de prendre également en compte la rentabilité future. Deux éléments jouent particulièrement sur celle-ci.

Le premier est la durée de vie. Dans le contexte concurrentiel du marché de l'assurance habitation, il est courant que les contrats ne soient pas rentables à l'affaire nouvelle. L'assureur mise alors sur la rétention des contrats en portefeuille et la revalorisation des primes à échéance, afin d'améliorer la rentabilité les années suivantes. Les contrats déficitaires à l'affaire nouvelle peuvent ainsi devenir profitables après quelques années, à condition qu'ils restent en portefeuille.

Un second aspect à prendre en compte est la multidétention. En effet, il est possible que certains profils de prospect ne soient pas rentables à l'affaire nouvelle en multirisque habitation, mais qu'ils soient en revanche susceptibles d'être multi-équipés. Il convient alors de considérer la rentabilité sur l'ensemble des contrats de l'assuré. La multidétention permet la fidélisation du client et réduit ainsi la probabilité de résiliation : un client multi-équipé résilie plus difficilement ses différents contrats. Il reste alors plus longtemps en portefeuille et devient plus rentable.

La prise en compte de la rentabilité de manière globale nécessite donc une modélisation des durées de vie et de la probabilité de multidétention. La construction d'un indicateur prenant en compte les multiples aspects de la rentabilité entre dans le cadre d'études sur la valeur client. Ceci dépasse le champ de notre étude et nous ne disposons pas de telles données. Nous disposons en revanche d'une vision moyenne de la prime des devis et de l'ancienneté des contrats en portefeuille sur quatre macro-segments homogènes :

- les locataires d'appartement
- les propriétaires d'appartement
- les locataires de maison
- les propriétaires de maison

Les primes moyennes des devis sont observées directement sur la base de PSC, tandis que l'ancienneté moyenne provient des données du recueil technique du portefeuille MRH. Nous tentons de surmonter les limites de l'utilisation du PSC en intégrant ces deux informations dans le choix du seuil d'équilibre.

Le tableau 18 récapitule le niveau moyen des primes des devis de la base d'étude :

Population	**Prime moyenne**
Locataires d'appartement	130€
Propriétaires d'appartement	180€
Locataires de maison	190€
Propriétaires de maison	310€

TABLE 18 – Prime moyenne sur les quatre macro-segments de la base d'étude

L'écart observé entre les primes moyennes suggère une différence de capacité d'absorption des coûts entre ces quatre macro-segments. En effet, les différents coûts (frais d'acquisition, de gestion, commissions) peuvent être décomposés en une partie fixe et une partie proportionnelle à la prime. La part fixe est d'autant plus élevée sur le digital, car elle englobe notamment les dépenses liées aux comparateurs, l'achat de mots-clés sur les navigateurs *etc...* Une prime plus élevée permet une meilleure absorption des coûts fixes, c'est pourquoi nous prenons la décision de fixer un seuil de PSC différent pour les quatre macro-segments. Le calcul des seuils prend en compte la différence de dilution des frais fixes de la manière suivante :

$$PSC_{seuil} = 1 - \left(\frac{\text{coûts fixes}}{\text{prime moyenne}} + \text{coûts proportionnels} \right)$$

L'application de cette formule mène à l'obtention des quatre valeurs de PSC d'équilibre suivantes :

Macro-segment	**Seuil de PSC**
Locataires d'appartement	60%
Propriétaires d'appartement	66%
Locataires de maison	66%
Propriétaires de maison	72%

TABLE 19 – Seuils d'équilibre du PSC choisis pour les quatre macro-segments

En complément des primes moyennes, le tableau 20 permet de visualiser l'ancienneté moyenne des contrats dans le portefeuille MRH :

Population	**Ancienneté moyenne**
Locataires d'appartement	4 ans
Propriétaires d'appartement	10 ans
Locataires de maison	5 ans
Propriétaires de maison	12 ans

TABLE 20 – Ancienneté moyenne sur les quatre macro-segments du portefeuille MRH

L'ancienneté moyenne dans le portefeuille est un bon indicateur de la fidélité des contrats. Nous la prendrons en compte lors du ciblage des profils en appliquant une tolérance par rapport aux seuils de PSC de la figure 19. Ainsi, sachant que les contrats de propriétaires de maison sont susceptibles de durer plus longtemps et ont des primes plus élevées, nous pouvons accepter d'être en perte la première année. Nous accepterons des profils dépassant le seuil de PSC d'équilibre associé. Au contraire, nous tolérons moins d'être initialement en perte sur des profils de locataires d'appartement de faibles primes et de courte durée.

4.3 Construction des profils à l'aide d'arbres de régression CART

4.3.1 Méthodologie

Nous cherchons à présent à segmenter les contrats de la base des PSC en différents profils associés à des niveaux de rentabilité.

Pour cela, nous avons choisi d'entraîner un arbre de régression du PSC sur les données. L'arbre cherche les variables et les seuils de séparation qui permettent de séparer les contrats

en fonction de leur PSC de la manière la plus discriminante. Chaque feuille constitue un profil, associé à un niveau de PSC caractéristique qui est le PSC moyen des observations de la feuille. L'avantage de l'arbre de décision est sa lisibilité : les critères définissant chaque profil sont directement lus sur l'arbre en parcourant le chemin de la racine à la feuille du profil.

Pour rappel, nous avons distingué quatre macro-segments dans l'échantillon. Nous décidons de segmenter séparément chacune de ces quatre populations. Ainsi, la base des PSC est séparée en quatre échantillons, de façon à ce qu'un arbre spécifique soit appliqué à chacun de ces quatre macro-segments. Les paramètres restant à choisir sont la liste des variables à inclure dans ces quatre segmentations ainsi que le nombre de profils souhaité.

Choix des variables

En plus de la qualité juridique et du type d'habitation, qui ont déjà été utilisés pour séparer les échantillons, il faut choisir les variables qui seront potentiellement sélectionnées par l'arbre pour constituer les profils. Celles-ci doivent représenter des caractéristiques générales et invariantes du prospect et de son bien. En outre, ces caractéristiques doivent pouvoir être plus ou moins ciblées dans la pratique. Par conséquent, les variables initialement retenues sont les suivantes :

- l'âge de l'assuré
- la profession de l'assuré
- le nombre de pièces du bien
- le type d'étage du bien
- les différents zoniers

Les premiers tests de segmentation effectués révèlent une utilisation prédominante des zoniers par les arbres. Ces variables sont plus discriminantes puisqu'elles sont directement liées au risque. Elles permettent de construire des profils plus hétérogènes, c'est-à-dire définis par des valeurs caractéristiques de PSC plus distantes. Toutefois, les zoniers sont des variables difficilement interprétables, c'est pourquoi nous décidons de les retirer du processus de segmentation. Les nouveaux profils que nous obtenons sont certes plus proches du point de vue des PSC, mais ils sont plus parlants.

Choix de la profondeur des arbres

La profondeur maximale des arbres est fixée à 3, afin d'obtenir au maximum 8 profils par segmentation. En effet, nous nous sommes aperçus qu'une plus grande profondeur ne permettait pas d'obtenir des valeurs de PSC plus intéressantes (c'est-à-dire plus éloignées les unes des autres). De plus, cela apportait une redondance dans la liste des variables utilisées par les arbres, puisque le nombre de variables disponibles est assez restreint.

4.3.2 Résultats de la segmentation

Segmentation des locataires d'appartement

Les locataires d'appartement (LA) forment la population majoritaire de la base de PSC. Cet échantillon est constitué de 148 959 observations. Lors de la segmentation, nous imposons une contrainte sur l'effectif minimal des feuilles, fixé à 1% de l'effectif total de l'arbre. L'arbre obtenu est présenté sur la figure 38 :

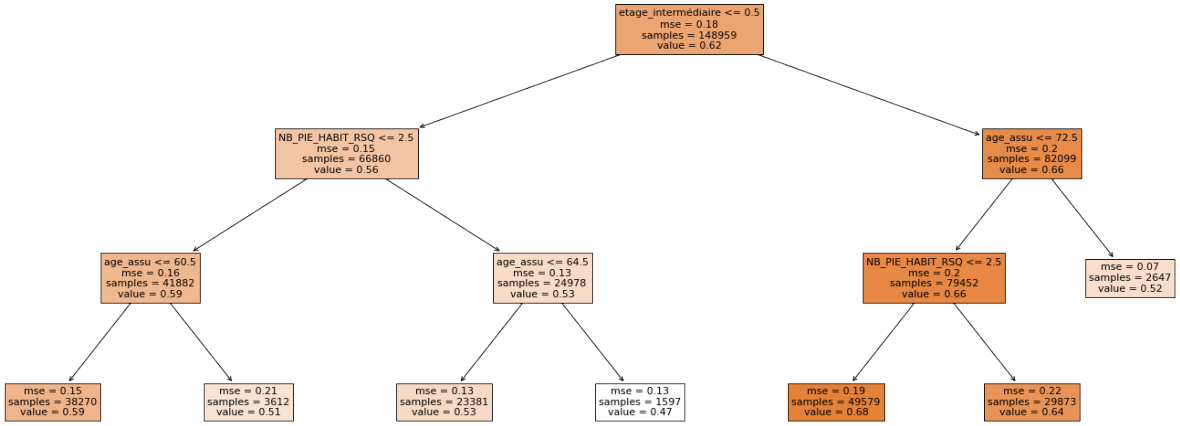

FIGURE 38 – Arbre de segmentation des locataires d'appartement

L'arbre produit 7 profils, en utilisant les variables de l'étage, du nombre de pièces et de l'âge de l'assuré. Le tableau 21 récapitule pour chaque profil ses critères d'appartenance, sa valeur de PSC caractéristique ainsi que sa part parmi l'effectif total des locataires d'appartements dans la base de PSC :

Profil	**Critère 1 : étage**	**Critère 2 : nombre de pièces**	**Critère 3 : âge**	**PSC**	**Proportion**
LA 1	RDC ou dernier	1-2	moins de 61 ans	59%	26%
LA 2	RDC ou dernier	1-2	61 ans ou plus	51%	2%
LA 3	RDC ou dernier	3 ou plus	moins de 65 ans	53%	16%
LA 4	RDC ou dernier	3 ou plus	65 ans ou plus	47%	1%
LA 5	intermédiaire	1-2	moins de 73 ans	68%	33%
LA 6	intermédiaire	3 ou plus	moins de 73 ans	64%	20%
LA 7	intermédiaire	-	73 ans ou plus	52%	2%

TABLE 21 – Critères de segmentation des locataires d'appartement

Nous constatons que les PSC caractéristiques des différents profils sont assez bons. En effet, les locataires d'appartements restant en moyenne assez brièvement dans le portefeuille, il est essentiel qu'ils soient rentables dès l'affaire nouvelle.

Segmentation des propriétaires d'appartement

L'échantillon des propriétaires d'appartement (PA) est constitué de 21 873 observations. L'effectif étant assez réduit, la contrainte sur l'effectif minimal des feuilles a été passée de 1% à 5% de l'effectif de la racine. L'arbre obtenu est présenté sur la figure 39 :

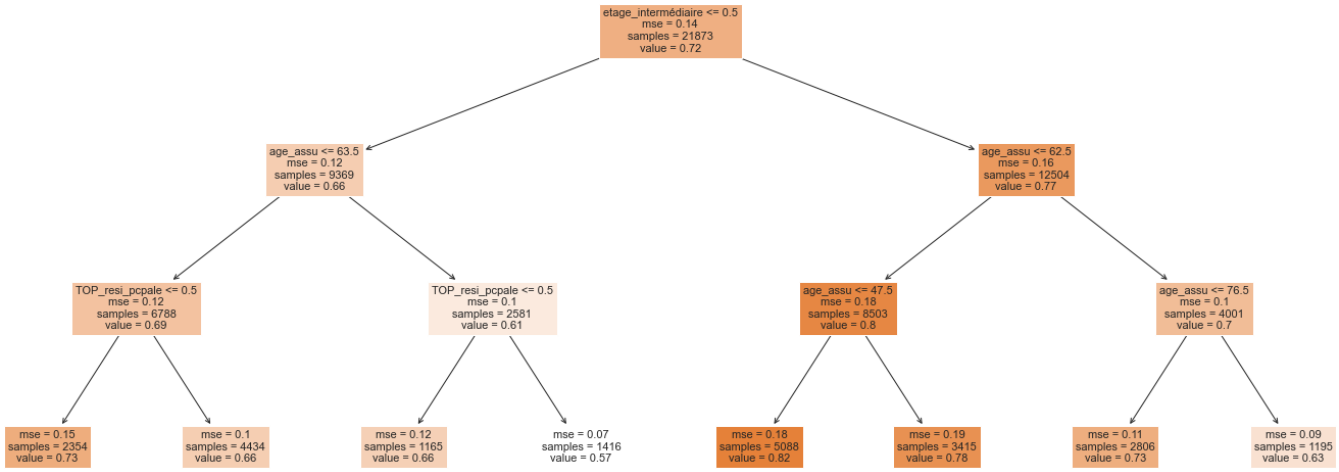

FIGURE 39 – Arbre de segmentation des propriétaires d'appartement

L'arbre produit 8 profils, présentés dans le tableau 22 :

Profil	**Critère 1 : étage**	**Critère 2 : âge**	**Critère 3 : résidence**	**PSC**	**Proportion**
PA 1	RDC ou dernier	moins de 64 ans	secondaire	73%	11%
PA 2	RDC ou dernier	moins de 64 ans	principale	66%	20%
PA 3	RDC ou dernier	64 ans ou plus	secondaire	66%	5%
PA 4	RDC ou dernier	64 ans ou plus	principale	57%	7%
PA 5	intermédiaire	moins de 48 ans	-	82%	23%
PA 6	intermédiaire	48-62 ans	-	78%	16%
PA 7	intermédiaire	63-76 ans	-	73%	13%
PA 8	intermédiaire	77 ans ou plus	-	63%	5%

TABLE 22 – Critères de segmentation des propriétaires d'appartement

Les valeurs de PSC de ces profils propriétaires d'appartement restent assez bonnes, bien qu'elles soient globalement supérieures à celles des profils locataires d'appartement. Nous constatons l'utilisation d'un nouveau critère : le type de résidence.

Segmentation des locataires de maison

L'échantillon des locataires de maison (LM) contient 36 739 observations. Cet effectif étant assez réduit, la contrainte sur l'effectif minimal des feuilles a là encore été placée à 5% de l'effectif de la racine. L'arbre obtenu est présenté sur la figure 40 :

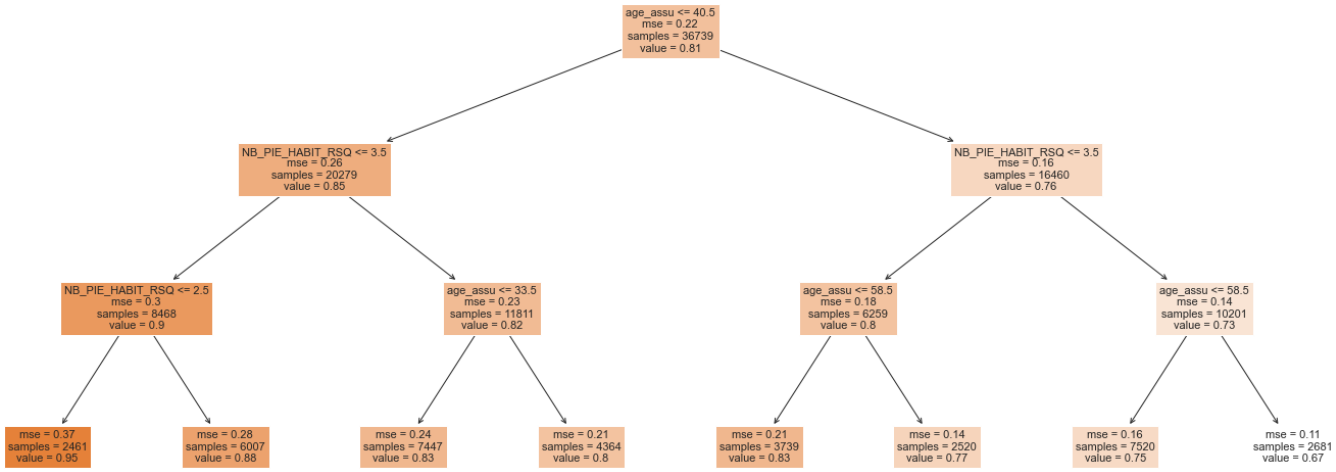

FIGURE 40 – Arbre de segmentation des locataires de maison

L'arbre produit 8 profils en utilisant les variables d'âge et du nombre de pièces. Ils sont présentés dans le tableau 23 :

Profil	**Critère 1 : âge**	**Critère 2 : nombre de pièces**	**PSC**	**Proportion**
LM 1	moins de 41 ans	1-2	95%	7%
LM 2	moins de 41 ans	3	88%	16%
LM 3	moins de 34 ans	4 ou plus	83%	20%
LM 4	34-40 ans	4 ou plus	80%	12%
LM 5	41-58 ans	1-2	83%	10%
LM 6	59 ans ou plus	1-2	77%	7%
LM 7	41-58 ans	3 ou plus	75%	21%
LM 8	59 ans ou plus	3 ou plus	67%	7%

TABLE 23 – Critères de segmentation des locataires de maison

Nous observons que les huit valeurs de PSC associées aux différents profils sont d'une part globalement plus élevées que celles des profils d'appartement, et d'autre part plus dispersées.

Segmentation des propriétaires de maison

Enfin, l'échantillon des propriétaires de maison (PM) contient 62 700 observations. Cet effectif est suffisant pour que la contrainte sur l'effectif minimal des feuilles soit placée à 1% de l'effectif total de l'échantillon. L'arbre obtenu est présenté sur la figure 41 :

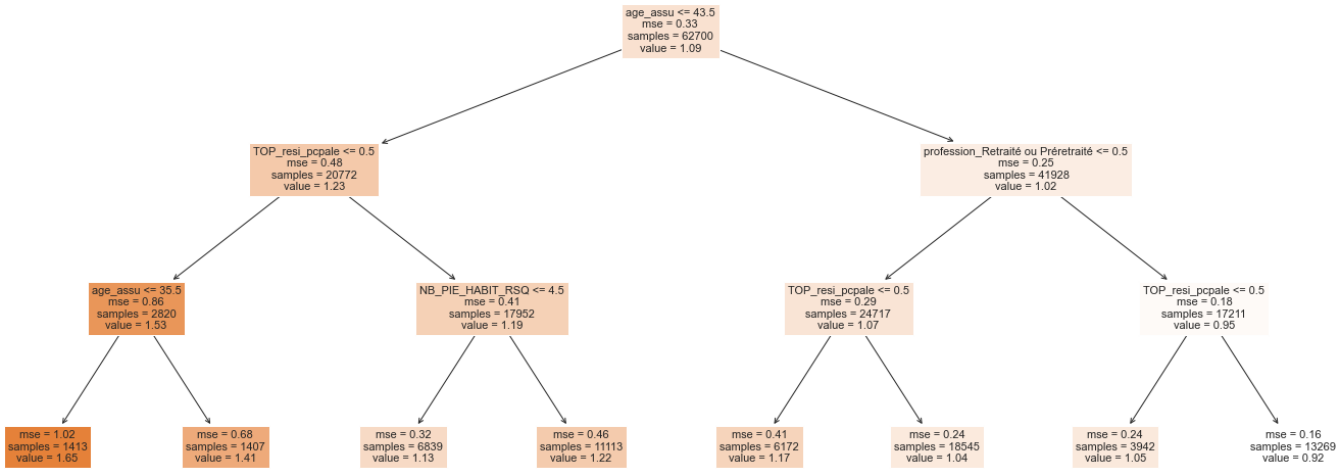

FIGURE 41 – Segmentation des propriétaires de maison

L'arbre produit 8 profils en utilisant les variables du type de résidence, de profession, d'âge et du nombre de pièces. Les profils figurent dans le tableau 24 :

Profil	**Critère 1 : âge**	**Critère 2 : résidence**	**Critère 3 : nombre de pièces**	**Critère 4 : CSP**	**PSC**	**Proportion**
PM 1	moins de 36 ans	secondaire	-	-	165%	2%
PM 2	36-43 ans	secondaire	-	-	141%	2%
PM 3	moins de 44 ans	principale	1-4	-	113%	11%
PM 4	moins de 44 ans	principale	5 ou plus	-	122%	18%
PM 5	44 ans ou plus	secondaire	-	actif	117%	10%
PM 6	44 ans ou plus	principale	-	actif	104%	30%
PM 7	44 ans ou plus	secondaire	-	retraité	105%	6%
PM 8	44 ans ou plus	principale	-	retraité	92%	21%

TABLE 24 – Critères de segmentation des propriétaires de maison

Les valeurs de PSC des profils construits contrastent avec les segmentations précédentes. En effet, 7 profils sur 8 ont un PSC moyen supérieur à 100%. Les PSC moyens des profils PM1 et PM2 dépassent même les 140%. Ces derniers ne représentent néanmoins à eux deux que 4% de l'échantillon. Les contrats des propriétaires de maison sont visiblement largement sous-tarifées à l'affaire nouvelle par rapport à leur risque. Toutefois, ce segment reste en moyenne assez longtemps en portefeuille. Ces valeurs de PSC, très élevées à l'affaire nouvelle, diminuent les années suivantes.

4.4 Choix des profils cibles

31 profils ont été construits à partir des données de la base des PSC. Chaque profil est la donnée d'un ensemble de critères et d'un niveau caractéristique de PSC. A présent, nous pouvons reproduire cette segmentation sur les devis de la base principale. Pour chacun des quatre macro-segments, nous présentons les caractéristiques des profils sur cette base, en les triant par ordre croissant du PSC. Les proportions indiquées sont cette fois-ci relatives à la base

principale. Nous choisissons ensuite les profils à cibler, en tenant compte des niveaux d'équilibre établis pour le PSC. Afin de faciliter la lecture, les profils ciblés sont signalés par une police bleue dans toute la suite.

Locataires d'appartement

Les devis locataires d'appartement (LA) représentent 64% de la base d'étude. Ils sont répartis au sein des 7 profils suivants :

Profil	Critère 1 : étage	Critère 2 : nombre de pièces	Critère 3 : âge	PSC	Part des LA	Part de la base
LA 4	RDC ou dernier	3 ou plus	65 ans ou plus	47%	0,3%	0,2%
LA 2	RDC ou dernier	1-2	61 ans ou plus	51%	0,8%	0,5%
LA 7	intermédiaire	-	73 ans ou plus	52%	0,5%	0,3%
LA 3	RDC ou dernier	3 ou plus	moins de 65 ans	53%	14%	9%
LA 1	RDC ou dernier	1-2	moins de 61 ans	59%	22%	14%
LA 6	intermédiaire	3 ou plus	moins de 73 ans	64%	23%	15%
LA 5	intermédiaire	1-2	moins de 73 ans	68%	40%	25%

TABLE 25 – Caractéristiques des profils locataires d'appartement

Le PSC d'équilibre calculé est de 60%. En raison de la faible ancienneté moyenne des contrats de locataires d'appartement, nous décidons de baser sur ce seuil le choix des profils à cibler, sans marge de tolérance. Ainsi, les profils ciblés sont les profils LA 1,2,3,4,7. Ils correspondent aux locataires d'appartement en RDC ou dernier étage, et aux retraités en étage intermédiaire. Les profils hors cible sont les profils LA 5 et 6. Il s'agit des locataires d'appartement en étage intermédiaire de moins de 73 ans.

Les profils cibles représentent 37% des devis locataires d'appartement, soit environ un quart de l'ensemble des devis présents dans la base d'étude. Ils sont surtout constitués des profils LA 3 et LA 1. En effet, les trois meilleurs profils, qui correspondent aux retraités, représentent à peine 1% des devis de la base. Les profils hors cible représentent les 63% restants des devis locataires d'appartement, soit 40% de l'ensemble des devis de la base d'étude.

Propriétaires d'appartement

Les propriétaires d'appartement représentent 15% de la base d'étude. Ils sont répartis en 8 profils :

Profil	Critère 1 : étage	Critère 2 : âge	Critère 3 : résidence	PSC	Part des PA	Part de la base
PA 4	RDC ou dernier	64 ans ou plus	principale	57%	2%	0,3%
PA 8	intermédiaire	77 ans ou plus	-	63%	1%	0,2%
PA 2	RDC ou dernier	moins de 64 ans	principale	66%	26%	4%
PA 3	RDC ou dernier	64 ans ou plus	secondaire	66%	1%	0,2%
PA 1	RDC ou dernier	moins de 64 ans	secondaire	73%	6%	1%
PA 7	intermédiaire	63-76 ans	-	73%	6%	1%
PA 6	intermédiaire	48-62 ans	-	78%	12%	2%
PA 5	intermédiaire	moins de 48 ans	-	82%	46%	7%

TABLE 26 – Caractéristiques des profils propriétaires d'appartement

Le seuil d'équilibre de PSC calculé est de 66%. Néanmoins, les propriétaires d'appartement ont une ancienneté moyenne de 10 ans en portefeuille. Nous considérons que, sur une si longue durée, même les contrats du profil PA 5 peuvent devenir rentables. Par conséquent, nous décidons de cibler les huit profils.

Locataires de maison

Les locataires de maison représentent 9% des devis de la base d'étude. L'application de la segmentation à ces profils aboutit au découpage suivant :

Profil	Critère 1 : âge	Critère 2 : nombre de pièces	PSC	Part des LM	Part de la base
LM 8	59 ans ou plus	3 ou plus	67%	3%	0,3%
LM 7	41-58 ans	3 ou plus	75%	18%	1,6%
LM 6	59 ans ou plus	1-2	77%	2%	0,2%
LM 4	34-40 ans	4 ou plus	80%	20%	1,7%
LM 3	moins de 34 ans	4 ou plus	83%	30%	2,7%
LM 5	41-58 ans	1-2	83%	5%	0,5%
LM 2	moins de 41 ans	3	88%	15%	1,3%
LM 1	moins de 41 ans	1-2	95%	7%	0,6%

TABLE 27 – Caractéristiques des profils locataires de maison

Le seuil d'équilibre calculé est de 66%. Compte tenu de l'ancienneté moyenne plutôt faible (5 ans) des locataires de maison en portefeuille, nous appliquons une moindre marge de tolérance autour de ce seuil. C'est pourquoi, l'unique profil ciblé est le profil LM 8. Il s'agit des locataires d'une maison de 3 pièces ou plus et âgés d'au moins 59 ans. Ce profil représente 3% des devis locataires de maison, et 0,3 % de l'ensemble de la base d'étude.

Propriétaires de maison

Enfin, nous appliquons la segmentation aux devis des propriétaires de maison, qui représentent 13% de la base d'étude :

Profil	Critère 1 : âge	Critère 2 : résidence	Critère 3 : nombre de pièces	Critère 4 : CSP	PSC	Part des PM	Part de la base
PM 8	44 ans ou plus	principale	-	retraité	92%	7%	1%
PM 6	44 ans ou plus	principale	-	actif	104%	28%	4%
PM 7	44 ans ou plus	secondaire	-	retraité	105%	2%	0,2%
PM 3	moins de 44 ans	principale	1-4	-	113%	19%	2%
PM 5	44 ans ou plus	secondaire	-	actif	117%	5%	0,6%
PM 4	moins de 44 ans	principale	5 ou plus	-	122%	36%	5%
PM 2	36-43 ans	secondaire	-	-	141%	2%	0,2%
PM 1	moins de 36 ans	secondaire	-	-	165%	1%	0,2%

TABLE 28 – Caractéristiques des profils propriétaires de maison

Hormis le profil PM 8, tous les PSC de profils sont supérieurs au seuil critique de 100%. Les contrats de propriétaires de maison sont visiblement déficitaires la première année. Néanmoins, ces profils survivent longtemps en portefeuille. Par ailleurs, des études effectuées par l'équipe montrent qu'ils ont un fort taux de multi-détention. Nous décidons de conserver en cible les

profils PM 8, 6 et 7, malgré la valeur d'équilibre de PSC établie à 82%. Ces profils correspondent aux prospects âgés d'au moins 44 ans, qui sont soit actifs et propriétaires d'une résidence principale, soit retraités et propriétaires d'une résidence secondaire. Ces catégories représentent un tiers des devis de propriétaires de maison, soit environ 5% de l'ensemble des devis.

Bilan du ciblage des profils :

En définitive, les profils que nous avons ciblés représentent 44 % des devis de la base d'étude. Il s'agit en majorité de profils de locataires d'appartement. En effet, les PSC caractéristiques des profils de maison sont trop élevés. Malgré la séparation des quatre populations lors de la construction des profils, la différenciation des seuils d'équilibre et enfin l'application d'une marge de tolérance, peu de profils de maisons ont pu être ciblés. Comme nous l'avons évoqué, le PSC ne reflète pas la vision globale de la rentabilité des contrats, en particulier sur les maisons. Ainsi, les critères de ciblage proposés ne sont pas parfaits. Néanmoins, ils fournissent une vision de la rentabilité intégrant de manière cohérente toutes les informations dont nous disposons.

5 Leviers d'amélioration du taux de transformation

Dans cette dernière partie, nous proposons d'agir sur trois leviers afin d'accroître le taux de transformation sur les profils que nous avons ciblés dans la partie précédente. Nous nous appuyons sur les différents modèles de conversion élaborés dans la partie 3.

5.1 Premier levier : l'optimisation tarifaire

Un premier levier essentiel permettant d'agir sur le taux de transformation est le prix. Une fois le prospect face au prix, s'il perçoit ce prix comme étant trop élevé alors il abandonnera et son devis ne sera pas transformé. Ainsi, nous cherchons à estimer la sensibilité du prospect au prix. Nous introduisons pour cela un indicateur microéconomique : l'élasticité-prix.

5.1.1 Définition de l'élasticité au prix

En micro-économie, l'élasticité-prix mesure la variation relative de la demande face à une variation du prix :

$$\varepsilon(p_0, p_1) = \frac{D(p_1) - D(p_0)}{D(p_0)} \cdot \frac{p_0}{p_1 - p_0} = \frac{\Delta D / D(p_0)}{\Delta p / p_0}$$

où ΔD désigne la variation de la demande et Δp désigne la variation du prix entre deux niveaux p_0 et p_1.

Dans notre étude, la demande correspond à la probabilité π de transformation du devis :

$$\varepsilon(p_0, p_1) = \frac{\Delta \pi(p) / \pi(p_0)}{\Delta p / p_0}$$

Pour la majorité des biens ou services, la demande diminue lorsque le prix augmente. L'élasticité est donc la plupart du temps négative. Une élasticité positive traduit une hausse de la demande avec le prix. Ce phénomène caractérise la catégorie des produits de luxe (c'est l'effet Veblen) ou de première nécessité (on parle de bien de Giffen). L'assurance habitation ne fait a priori pas partie de ces deux catégories. Nous décidons donc d'ajouter un signe négatif afin de travailler avec des valeurs positives d'élasticité :

$$\varepsilon(p_0, p_1) = -\frac{\Delta \pi(p) / \pi(p_0)}{\Delta p / p_0}$$

Il existe plusieurs méthodes d'estimation de l'élasticité. La méthode choisie dans ce mémoire est une méthode d'estimation analytique de l'élasticité. Celle-ci présente l'avantage d'être précise, rapide à implémenter et adaptée aux moyens dont nous disposons.

5.1.2 Méthode analytique de déduction de l'élasticité

Présentation de la méthode :

Cette méthode est inspirée du mémoire de Mehdi Boueddine [4]. Elle nécessite l'utilisation d'un modèle *logit* pour modéliser la probabilité de conversion. Il s'agit d'établir une formule fermée de l'élasticité en renversant la formule du modèle *logit*.

Comme nous l'avons vu, l'élasticité s'écrit :

$$\varepsilon(p) = -\frac{\Delta \pi(p) / \pi(p)}{\Delta p / p}$$

Supposons que la probabilité de transformation est dérivable par rapport à la prime. Alors, lorsque la variation de prix tend vers 0, l'élasticité devient :

$$\varepsilon(p) = -\frac{\delta\pi(p)/\pi(p)}{\delta p/p} = -\frac{\delta\pi(p)}{\delta p} \times p \times \frac{1}{\pi(p)}$$

Dans la section 3.5, nous avons mis en place un modèle GLM *logit* "prix". Celui-ci modélise la probabilité de transformation du devis i selon la formule suivante :

$$\hat{\pi}_i(p_i) = \frac{1}{1 + \exp(-(A_i + B_i \cdot p_i))}$$

où :

- $A_i = a_0 + a_1\,(i_1)\,I_{x_1=i_1} + \ldots + a_n\,(i_n)\,I_{x_n=i_n}$ est la combinaison linéaire des variables avec un effet non lié à la prime et de leurs coefficients respectifs dans le modèle
- $B_i = b_0 + b_1\,(j_1)\,I_{x'_1=j_1} + \ldots + b_m\,(j_m)\,I_{x'_m=j_m}$ est la combinaison linéaire des variables avec un effet lié à la prime et de leurs coefficients respectifs.

Cette probabilité de transformation est dérivable par rapport à la prime :

$$\frac{\delta\hat{\pi}_i(p_i)}{\delta p_i} = \frac{B_i \exp(-(A_i + B_i \cdot p_i))}{[1 + \exp(-(A_i + B_i \cdot p_i))]^2}$$

L'élasticité est alors :

$$\hat{\varepsilon}(p) = -\frac{\delta\hat{\pi}(p)}{\delta p} \times p \times \frac{1}{\hat{\pi}(p)} = -\frac{B_i \exp(-(A_i + B_i \cdot p_i))p_i}{1 + \exp(-(A_i + B_i \cdot p_i))}$$

ce que nous pouvons réécrire simplement :

$$\hat{\varepsilon}(p_i) = -Bp_i(1 - \hat{\pi}(p_i))$$

Ainsi, nous avons établi une formule fermée pour l'élasticité-prix de la probabilité de transformation de l'individu i. A l'aide de cette formule, nous pouvons calculer une valeur de l'élasticité pour chaque devis, c'est-à-dire une valeur d'élasticité individuelle. De manière générale, la probabilité de transformation baisse lorsque le tarif augmente. En supposant le coefficient B négatif, nous nous attendons donc à une croissance de l'élasticité mesurée avec la prime.

Résultats :

En appliquant la formule fermée précédente, nous calculons l'élasticité de chaque devis de la base d'étude à partir des coefficients du GLM "prix". La figure 42 présente la distribution des élasticités obtenues :

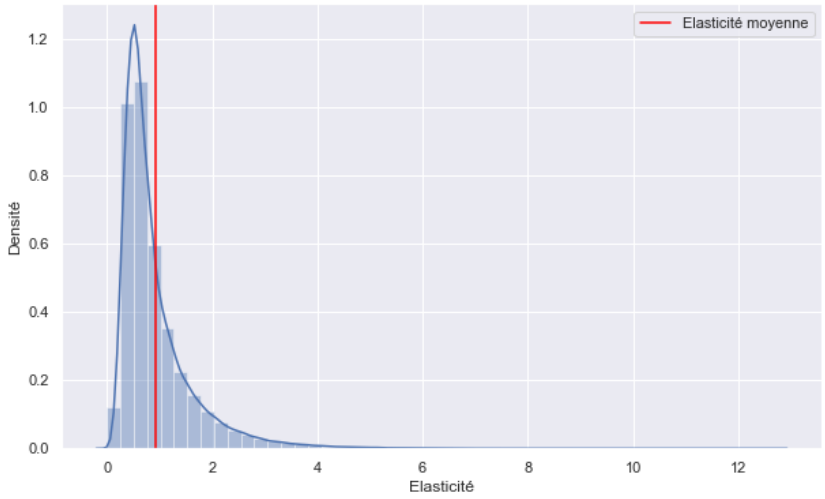

FIGURE 42 – Distribution des valeurs d'élasticité mesurées

Les valeurs d'élasticité mesurées sont positives. Etant donné la convention de signe adoptée, cela signifie que la probabilité de conversion diminue lorsque le tarif augmente. Malgré une queue de distribution assez épaisse au-delà de 1, l'élasticité moyenne vaut 0,92. La figure 43 permet de visualiser l'élasticité en fonction de la prime, pour les primes inférieures à 1000€ :

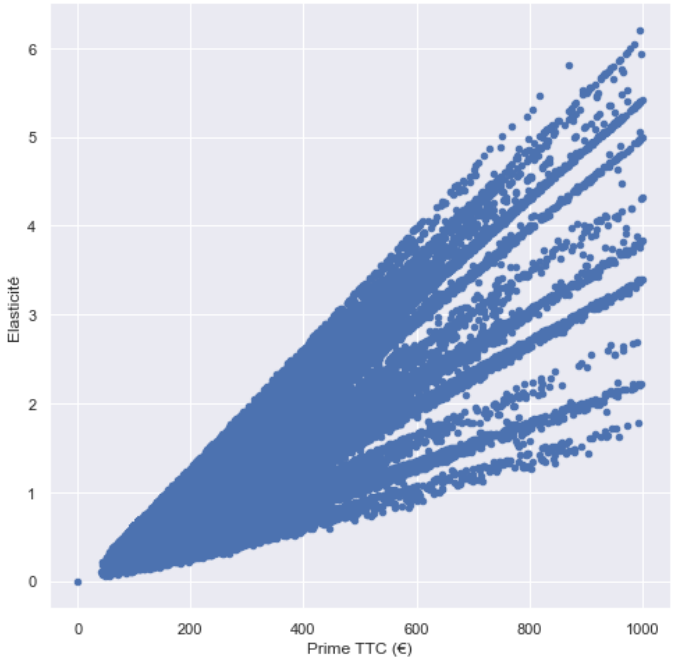

FIGURE 43 – Elasticité en fonction de la prime pour les primes inférieures à 1000€

Nous observons que l'élasticité croît globalement avec le tarif. Il ne s'agit néanmoins que d'une tendance générale, car nous observons une forte dispersion. Bien que des valeurs élevées d'élasticité soient mesurées sur une partie des primes les plus élevées, la plupart des primes du portefeuille sont faibles et réduisent la valeur moyenne de l'élasticité. Pour expliquer l'allure singulière observée sur la figure ci-dessus, nous observons sur la figure 44 le tracé des probabilités de transformation estimées par le modèle, face aux primes des devis :

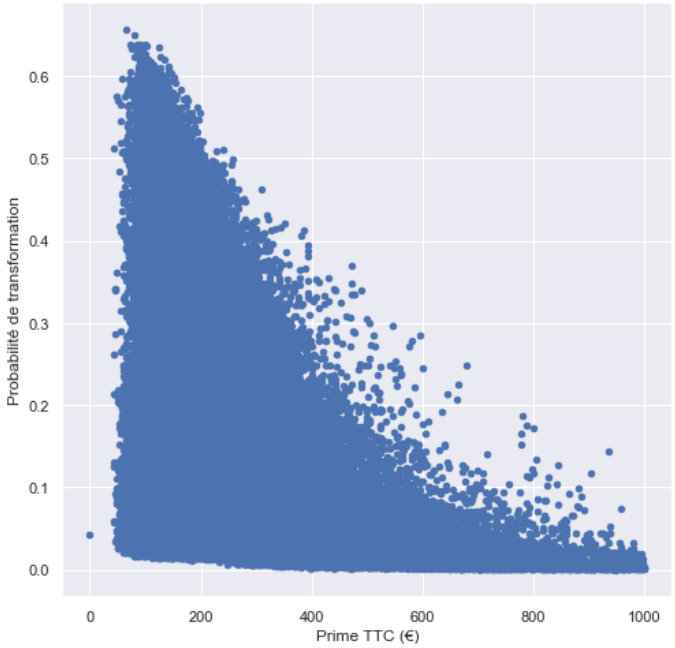

FIGURE 44 – Probabilité de conversion en fonction de la prime pour les primes inférieures à 1000€

Pour les faibles primes, la probabilité de transformation occupe tout l'intervalle $[0\,;0.65]$. Par conséquent, l'élasticité occupe également un large spectre de valeurs, étant donné la formule la liant à la probabilité de conversion. En revanche, lorsque la prime devient élevée, la probabilité de conversion tend vers 0. La formule de l'élasticité devient alors une fonction linéaire en la prime et dépendant de la valeur du coefficient B. Cela correspond à ce que nous observons sur la figure 43 : pour les faibles primes, l'élasticité est très dispersée, tandis que pour les primes plus élevées, l'élasticité suit des droites dont les pentes sont les différentes valeurs du coefficient B. Ces droites sont des iso-segments des modalités croisées dans le modèle.

La croissance obtenue de l'élasticité avec la prime peut avoir plusieurs interprétations. La première interprétation possible est que les prospects peuvent être sensibles à l'échelle absolue (en euros) du prix. Ainsi, nous pouvons imaginer qu'une variation de prime de 1% pour une prime de départ de 100 euros n'aura pas le même effet que pour une prime de départ de 500 euros.

Nous pouvons également proposer une interprétation de la croissance de l'élasticité avec la prime en lien avec la théorie économique des coûts de recherche (*search costs*). Un consommateur rationnel cherche le prix optimal. Pour cela, il cherche à comparer les différentes offres. Cette recherche a néanmoins un coût : l'énergie, le temps et les moyens dépensés forment des coûts de recherche. Dès lors, l'individu continuera à comparer les offres tant que le profit marginal tiré de la recherche dépasse le coût marginal de recherche engagé. Internet a considérablement réduit ces coûts, en particulier grâce aux comparateurs. Néanmoins, nous pouvons considérer qu'ils existent toujours. Considérons alors un locataire d'un petit appartement. Les différents prix qui peuvent lui être proposés seront plutôt faibles. Face au prix affiché sur son devis, il peut préférer ne pas engager de coûts de recherche supplémentaires en allant consulter d'autres prix. A l'inverse, un propriétaire d'une grande maison se retrouve face à un prix élevé. Le coût d'une recherche minutieuse des différentes offres du marché est alors davantage rentabilisé par la différence de prix que cette recherche peut lui procurer. Suite à cette recherche, ce prospect est mieux informé sur la concurrence et réagira plus fortement à une augmentation du prix.

Le tableau 29 présente les valeurs d'élasticité moyenne mesurées sur chaque profil :

Profil	Prime moyenne	Elasticité moyenne
PM 4	512	2,39
LM 4	343	1,59
PM 3	333	1,49
LM 3	298	1,40
PM 6	452	1,34
PM 1	238	1,08
PA 5	242	1,07
PA 2	239	1,03
LM 7	326	1,00
PM 8	452	0,97
LA 6	234	0,96
LA 3	230	0,96
PM 2	254	0,92
LM 2	208	0,85
LM 8	301	0,77
LM 1	167	0,68
PM 5	260	0,68
PA 6	228	0,65
LA 5	149	0,63
LA 1	147	0,62
LM 5	198	0,53
PA 8	236	0,52
PA 4	236	0,52
LA 4	209	0,51
PM 7	232	0,48
PA 7	203	0,45
LA 7	180	0,43
PA 1	125	0,40
LM 6	173	0,40
LA 2	140	0,34
PA 3	124	0,27

TABLE 29 – Elasticité par profil

Comme prévu, nous observons une certaine corrélation entre le niveau de prime moyenne des profils et l'élasticité moyenne. Les plus grandes valeurs d'élasticité sont globalement mesurées chez les profils de plus grandes primes moyennes, qui sont principalement des propriétaires de maison. A priori, ces profils sont donc les plus sensibles au prix. Parmi les cibles, les plus grandes valeurs moyennes d'élasticité (et les seules supérieures à 1) sont enregistrées chez les profils PM 6, PA 5 et PA 2. Il s'agit respectivement :

- des propriétaires de maison principale actifs et âgés de 44 ans ou plus
- des propriétaires d'appartement en étage intermédiaire, âgés de moins de 48 ans
- des propriétaires d'appartement principal en rez-de-chaussée ou dernier étage, âgés de moins de 64 ans.

Les plus faibles valeurs d'élasticité sont mesurées chez les profils de plus faibles primes moyennes, qui sont principalement des locataires et propriétaires d'appartement. A priori, une variation des primes aurait très peu d'influence sur la conversion de ces profils. Parmi les cibles, les plus faibles valeurs d'élasticité sont enregistrées chez les profils PA 3, LA 2 et PA 1. Il s'agit des propriétaires d'appartement secondaire en rez-de chaussée ou dernier étage, ainsi que des locataires d'appartement d'une ou deux pièces en rez-de-chaussée ou dernier étage âgés de 61 ans ou plus.

D'après les résultats obtenus, la plupart des profils ciblés sont en moyenne plutôt faiblement élastiques au prix. Néanmoins, les valeurs d'élasticité obtenues avec cette méthode analytique

paraissent faibles. Nous conservons donc une certaine prudence quant aux conclusions tirées de ces mesures.

5.1.3 Elasticité déduite de simulations du modèle

Afin de valider les résultats de la méthode analytique, nous essayons une seconde approche de calcul de l'élasticité. Cette approche plus classique est basée sur des simulations de revalorisation des primes à l'aide des modèles de conversion. Le modèle *logit* "prix" est utilisé pour simuler l'impact de différents chocs de primes sur la probabilité de conversion. Les chocs de prime testés sont les suivants : [-20% ; -15% ; -10% ; -5% ; -2% ; -1% ; 1% ; 2% ; 5% ; 10% ; 15% ; 20%]. Pour chaque taux r de revalorisation testé, les étapes sont les suivantes :

1. Appliquer une variation relative de la prime à tous les devis
2. Estimer, grâce au modèle, la nouvelle probabilité de conversion pour tous les devis
3. Déduire l'élasticité individuelle de chaque devis selon la formule

$$\hat{\varepsilon}(i) = -\frac{\hat{\pi}_1(i) - \hat{\pi}_0(i)}{r\hat{\pi}_0(i)}$$

Ce programme est également testé avec le modèle général de conversion, c'est-à-dire le modèle *Gradient Boosting*. Rappelons que la prime est découpée en tranches dans ce modèle. Par conséquent, les chocs n'ont d'effet sur la probabilité de conversion d'un devis que s'ils font basculer la prime d'une tranche à une autre.

Les valeurs d'élasticité moyennes obtenues pour les différents chocs sur les deux modèles sont présentées dans le tableau 30 et la figure 45 :

choc	**élasticité moyenne GLM "prix"**	**élasticité moyenne *Gradient Boosting***
-20%	1,06	0,22
-15%	1,02	0,01
-10%	0,99	-0,14
-5%	0,95	0,48
-2%	0,96	0,44
-1%	0,93	0,45
1%	0,92	-7,22
2%	0,93	-3,40
5%	0,89	-0,45
10%	0,87	0,16
15%	0,84	0,34
20%	0,82	0,61

TABLE 30 – Mesures d'élasticité par application d'un choc de prime

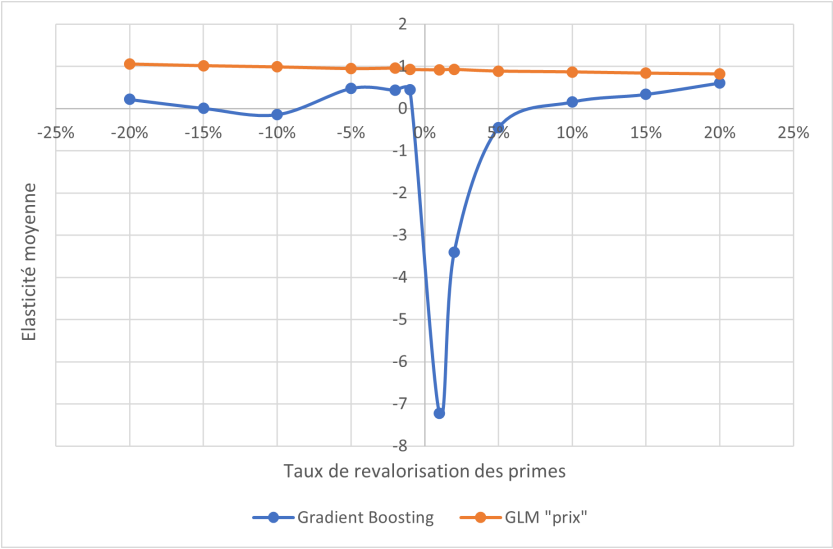

FIGURE 45 – Elasticité moyenne pour les différents chocs de primes

Les valeurs d'élasticité moyenne déduites du modèle *logit* "prix" sont très proches de la valeur d'élasticité moyenne obtenue de manière analytique. Cette seconde approche corrobore ainsi les résultats de la méthode analytique.

Par ailleurs, nous constatons que le modèle *Gradient Boosting* ne permet pas d'obtenir des résultats plus convaincants. L'élasticité mesurée avec ce modèle est plus faible et beaucoup moins stable. En particulier, nous observons un creux anormal d'élasticité pour les majorations de 1% et 2%. Nous supposons qu'il s'agit de l'effet d'un changement de tranche d'une part importante des devis. Etant donné que le modèle est assez opaque, il est difficile de déterminer plus précisément le phénomène en jeu.

5.1.4 Commentaire sur les valeurs d'élasticité obtenues

La simulation de chocs de primes donne du crédit aux valeurs d'élasticité mesurées analytiquement. Cependant, ces valeurs d'élasticité paraissent faibles pour la plupart des profils. En effet, des valeurs d'élasticité inférieures à 1 traduisent un scénario de demande faiblement élastique, c'est-à-dire qu'une augmentation ou une réduction de la prime a peu d'influence sur le taux de conversion des devis. Dans un tel scénario, l'augmentation des primes risque de toujours compenser la baisse du nombre de conversions, et provoquer l'augmentation de la marge. Un programme d'optimisation tarifaire préconiserait alors de fortes majorations de primes pour atteindre la marge optimale. Les résultats obtenus ne sont donc pas utilisables pour faire une optimisation tarifaire. Nous envisageons trois explications possibles quant aux faibles valeurs d'élasticité observées :

La première possibilité est que le phénomène de conversion soit réellement inélastique au prix. Néanmoins, cela nous paraît peu probable dans le contexte fortement concurrentiel de l'assurance habitation, qui plus est pour des devis d'origine digitale.

La seconde possibilité est que les résultats tirés du modèle ne soient valables que pour une faible variation de prix. Une explication à cela pourrait être que le modèle ne dispose pas d'observations avec un prix suffisamment différent toutes choses égales par ailleurs. S'il paraît raisonnable que des faibles variations de prime suscitent peu de réaction de la part du prospect, cela ne l'est plus en revanche pour d'importantes revalorisations. En particulier, si le tarif augmente trop et devient hors marché, le taux de conversion tendra quoi qu'il en soit vers 0. A

l'inverse, si le tarif Allianz devient très compétitif, la conversion devrait fortement augmenter sur tous les segments, sauf si le produit devient bas de gamme aux yeux des prospects. Pour utiliser les valeurs d'élasticité mesurées, il faudrait donc faire l'hypothèse d'une croissance suffisante de la sensibilité du prospect au prix au-delà d'un certain seuil de variation des primes.

Enfin, il est également possible que le modèle ne parvienne tout simplement pas à restituer correctement la sensibilité des prospects au prix.

5.1.5 Pistes d'amélioration pour l'étude de l'élasticité-prix

Dans cette section, nous donnons quelques pistes qui auraient peut-être pu nous permettre d'observer de plus grandes valeurs d'élasticité.

Affinage du modèle :

La première piste envisagée est celle d'un renforcement du modèle pour mieux capter la sensibilité des prospects au prix. En particulier, une dimension que nous n'avons pas pu intégrer à l'étude est la notion de compétitivité du tarif. Il s'agit d'une information essentielle à la fois pour l'estimation de la conversion et celle de la sensibilité au prix. Elle pourrait être intégrée au modèle sous la forme d'une variable d'écart aux tarifs des principaux concurrents, à caractéristiques équivalentes de devis. Muni de cette variable, le modèle pourrait tenir compte du fait que les fortes majorations ou réductions placent le tarif hors marché, lors de l'estimation de l'élasticité.

Test d'élasticité :

Une autre méthode consiste à mesurer empiriquement l'élasticité-prix en réalisant un test d'élasticité (ou *price test*). Plusieurs niveaux tarifaires sont choisis, chacun associés à un certain pourcentage de réduction de la prime (par exemple {-2%, -1%, 0%, 1%, 2%}). Sur une période donnée, les prospects se voient attribuer aléatoirement un des niveaux lors de leur simulation tarifaire. Une fois qu'un niveau est attribué à un prospect, celui-ci le conserve même s'il effectue un nouveau devis. Puisque l'attribution est aléatoire, la répartition des niveaux sur une modalité est sensiblement la même peu importe la modalité regardée. L'élasticité-prix peut ensuite être calculée pour chaque niveau tarifaire en prenant le niveau 0% comme point de référence :

$$\varepsilon(0\%, x\%) = -\frac{\tau(x\%) - \tau(0\%)}{\tau(0\%)} \cdot \frac{1}{x\%}$$

où $\tau(x\%)$ et $\tau(0\%)$ sont les taux de transformation respectifs des populations de niveaux tarifaires x% et 0%.

Le test d'élasticité permet d'éviter les biais liés à l'utilisation d'un modèle. Il serait très utile de comparer l'élasticité déduite des modèles à ces valeurs empiriques. Toutefois, sa mise en oeuvre pose des difficultés d'ordre opérationnel :

- la mise en oeuvre de ce test nécessite un certain investissement
- les modifications de prime peuvent être difficiles à justifier auprès des agents
- les écarts de prime ne doivent pas être trop importants entre deux prospects similaires.
- les modifications de prime doivent maintenir le produit concurrentiel et rentable. Cela limite assez fortement l'amplitude des écarts tarifaires applicables.

5.1.6 Méthodologie d'utilisation de l'élasticité-prix

Dans cette section, nous expliquons comment l'élasticité-prix pourrait être utilisée pour actionner le levier de l'optimisation tarifaire. Cette optimisation tarifaire n'a pas été effectuée dans le cadre de ce mémoire, mais nous la décrivons à titre de prolongement de l'étude menée.

Nous supposons que nous disposons de l'élasticité-prix individuelle de chaque devis. Si l'élasticité individuelle n'est pas disponible, il est souhaitable de disposer a minima d'une valeur d'élasticité par segment. Une première application de cette élasticité est l'estimation de la nouvelle probabilité de transformation du devis après une revalorisation de la prime à un taux $r(i)$ quelconque.

Considérons un devis caractérisé par :

- sa prime $p_0(i)$
- sa probabilité de transformation $\pi(p_0(i))$
- son élasticité $\varepsilon(i)$

Utilisation classique de l'élasticité :

Notons $p_1(i) = (1+r)p_0(i)$ la nouvelle prime du devis après une revalorisation au taux $r(i)$ souhaité et $\pi(p_1(i)) = \pi(p_0(i), r(i))$ la nouvelle probabilité de transformation après revalorisation. L'utilisation classique de l'élasticité consiste à supposer que la variation de la probabilité de transformation est proportionnelle à la valeur d'élasticité mesurée :

$$\pi(p_0(i), r(i)) = \pi(p_0(i)) \cdot (1 - \varepsilon(i)r(i))$$

Il s'agit de l'utilisation la plus courante de l'élasticité-prix : une variation de la prime de $r\%$ provoque une variation de la demande de $\varepsilon(i) \cdot r\%$. Cette approche suppose que le comportement du prospect est le même face à toutes les variations tarifaires.

Approche non linéaire :

Une alternative est de considérer que la sensibilité du prospect n'est pas la même face à différents taux de revalorisation. Cette approche est décrite dans le mémoire d'Omar Bouattour[5]. Nous en donnons ici le raisonnement sans entrer dans les détails du calcul.

Si la probabilité de transformation est dérivable par rapport au prix, l'élasticité au point p_1 s'écrit :

$$\varepsilon = -\frac{\delta\pi(p_1)}{\delta p_1}\frac{p_1}{\pi(p_1)}$$

En supposant que la probabilité de transformation est également dérivable par rapport au taux de revalorisation, un calcul mène à l'équation différentielle suivante :

$$\frac{\delta\pi(r)}{\delta r} = -\frac{\varepsilon}{1+r}\pi(r)$$

La résolution de cette équation différentielle permet d'aboutir à l'expression suivante pour la nouvelle probabilité de transformation :

$$\pi(p_0(i), r(i)) = \pi(p_0(i))(1 + r(i))^{-\varepsilon(i)}$$

Pour les faibles taux de revalorisation, nous retombons sur la formule de l'utilisation linéaire de l'élasticité. En revanche, lorsque la revalorisation devient importante, cette approche permet de modéliser un accroissement de la sensibilité des prospects au prix.

Optimisation tarifaire

L'élasticité-prix permet d'estimer la probabilité de transformation de chaque devis après une quelconque revalorisation. Elle permet de définir un arbitrage que doit opérer l'assureur entre la marge tarifaire et la probabilité de conversion. En réduisant le prix, l'assureur augmente ses bénéfices en cas de souscription du prospect, mais il réduit par la même occasion sa probabilité de réaliser l'affaire. Par conséquent, il cherche à maximiser sa marge espérée pour chaque catégorie de prospect. Dans un prolongement de l'étude, l'élasticité-prix pourrait donc être intégrée à un programme d'optimisation de la marge espérée de l'assureur.

Dans notre étude, les profils ciblés correspondent aux catégories de meilleures marges techniques (relativement à la prime). Pour l'optimisation, nous pourrions envisager plusieurs possibilités :

- Nous pourrions choisir d'appliquer uniquement des réductions de prime sur les profils cibles pour accroître leur volume de souscription. Les profils les plus intéressants pour une réduction seraient notamment les profils les plus sensibles au prix.
- Nous pourrions également choisir d'appliquer à la fois des réductions aux devis cibles et des majorations aux devis hors cible. L'accent serait alors mis sur l'amélioration du mix portefeuille, quitte à perdre des affaires nouvelles sur les profils hors cible.

Cependant, il est à noter que l'application d'une revalorisation de la prime modifie le PSC du devis. Pour rappel, les profils ont été initialement ciblés en raison de leur rentabilité. En réduisant la prime des catégories ciblées, les nouveaux clients de ces catégories risquent d'être moins rentables, ce qui limite l'intérêt de la réduction tarifaire. Nous voyons ainsi apparaître une boucle qui complexifie le problème. Pour éviter cet écueil, nous pourrions fixer une limite de réduction tarifaire à ne pas dépasser afin de maintenir la stabilité des profils.

5.2 Deuxième levier : les offres promotionnelles

Le deuxième levier étudié est l'utilisation d'offres promotionnelles, prenant ici la forme de deux mois offerts sur la prime. Comme décrit lors de la présentation de la base de données, les mois gratuits sont des promotions lancées au moment de campagnes annuelles par Allianz France et s'appliquent au premier mois des deux premières années du contrat. A l'aide du modèle général de conversion, nous étudions l'efficacité de cette promotion pour l'amélioration de la conversion sur les profils cibles.

5.2.1 Impacts d'une distribution complète de mois gratuits

Nous commençons par mesurer l'effet des mois gratuits sur les populations des différents profils en simulant une transition entre deux états :

- état 1 : aucun devis du profil étudié n'a de mois gratuit.
- état 2 : tous les devis du profil étudié ont des mois gratuits.

Afin d'évaluer l'impact du levier, nous mesurons la variation relative de la probabilité moyenne de conversion de chaque profil, lors du passage de l'état 1 à l'état 2. Plutôt que de partir de la répartition réelle des mois gratuits dans la base d'étude, nous avons choisi un état initial sans aucun mois gratuit afin d'éviter que la proportion de mois gratuits dans les profils n'influe sur la variation mesurée. En outre, le choix de ces deux états extrêmes permet d'estimer la limite supérieure de l'effet des mois gratuits. Les variations observées sont présentées dans le tableau 31 :

Profil	Part des devis	Probabilité moyenne état initial	Probabilité moyenne état final	Variation relative
PM 8	0,9%	5,7%	9,0%	+57,4%
PA 4	0,3%	6,3%	9,8%	+54,2%
PM 7	0,2%	12,4%	18,3%	+48,0%
PM 4	4,6%	3,6%	5,3%	+47,7%
PM 3	2,4%	5,6%	8,2%	+47,3%
PA 8	0,2%	6,7%	9,8%	+46,7%
PM 6	3,6%	4,4%	6,3%	+43,3%
LM 6	0,2%	22,7%	32,5%	+43,1%
LM 8	0,3%	16,7%	23,4%	+40,3%
PA 7	0,8%	7,6%	10,5%	+39,4%
PM 1	0,2%	14,4%	20,0%	+38,8%
LA 4	0,2%	15,5%	21,3%	+37,0%
PM 5	0,6%	14,0%	18,7%	+33,6%
PM 2	0,2%	14,6%	19,4%	+32,5%
PA 3	0,2%	10,5%	13,9%	+32,2%
LA 7	0,3%	15,8%	20,5%	+29,9%
LM 4	1,7%	17,3%	22,0%	+27,4%
LM 2	1,3%	30,9%	39,1%	+26,6%
PA 2	3,9%	7,1%	8,8%	+25,1%
LM 5	0,5%	26,2%	32,6%	+24,6%
LA 2	0,5%	18,3%	22,7%	+23,9%
LM 3	2,7%	20,8%	25,8%	+23,7%
PA 6	1,9%	7,3%	9,1%	+23,3%
LM 7	1,6%	18,6%	22,8%	+22,7%
PA 5	6,8%	7,3%	8,8%	+20,2%
PA 1	1,0%	12,9%	15,3%	+19,1%
LA 6	14,6%	20,0%	23,8%	+18,9%
LA 3	8,9%	23,6%	27,7%	+17,5%
LM 1	0,6%	33,4%	37,8%	+13,3%
LA 5	25,1%	24,5%	26,8%	+9,4%
LA 1	13,9%	26,7%	29,0%	+8,7%

TABLE 31 – Impacts sur la probabilité de conversion

Globalement, les profils où les impacts mesurés sont les plus forts sont les profils dont les primes moyennes sont les plus élevées. Il s'agit des propriétaires de maison ou d'appartement. Au contraire, les impacts les plus faibles d'un passage de 0 à 100% de devis avec mois gratuits sont mesurés chez les profils de plus faibles primes moyennes. Il s'agit principalement de locataires d'appartement. Ce constat semble cohérent. En effet, les mois gratuits peuvent être vus comme une remise sur la prime. En considérant la valeur en euros de la réduction, ils sont donc d'autant plus susceptibles d'avoir de l'effet sur le prospect que la prime de celui-ci est élevée au départ.

Nous constatons que les plus forts impacts sont enregistrés chez trois profils cibles : les profils PM 8, PA 4 et PM 7. Il s'agit des retraités propriétaires de maison ou d'un appartement

principal en rez-de-chaussée ou dernier étage. La probabilité de conversion moyenne augmente de moitié chez ces profils. Les cibles enregistrant le plus faible impact sont les profils LA 1 et LA 3. Il s'agit dans les deux cas de locataires d'appartement en rez-de-chaussée ou dernier étage, d'une part d'une ou deux pièces et âgés de moins de 61 ans, d'autre part de plus de deux pièces et âgés de moins de 65 ans. La probabilité de conversion moyenne de ces profils n'augmente que de 9 et 18%, mais cela s'avère tout de même remarquable étant donné que ces deux profils ont un volume de devis et un taux de conversion élevés.

L'observation des effectifs révèle que les profils aux plus forts impacts mesurés représentent une part négligeable de l'ensemble de la base. Ainsi, les 10 premiers profils en termes d'impact mesuré ne comptent que pour 13% des devis. Dans le cadre d'une optimisation des mois gratuits sur notre portefeuille actuel, l'amélioration du taux de conversion des cibles parmi ces profils risque d'être peu impactante. A l'inverse, les deux profils cibles aux plus faibles impacts sur la probabilité comptent pour près de 23% de la base. Si nous souhaitons optimiser la distribution des mois gratuits sur notre portefeuille actuel, le raisonnement en termes de probabilité de conversion n'est ainsi pas idéal.

A la place, nous considérons le volume attendu d'affaires nouvelles supplémentaires au sein du profil. Ce volume correspond au produit du volume de devis du profil par l'écart entre les probabilités moyennes de conversion initiale et finale sur le profil. En effet, en notant S et $S' = S + \Delta S$ les volumes initial et final espérés d'affaires nouvelles du profil, N le nombre de devis du profil et π_i et π'_i les probabilités initiale et finale de conversion du devis i, le volume attendu d'affaires nouvelles supplémentaires s'écrit :

$$E(\Delta S) = E(S' - S) = \sum_{i=1}^{N} \pi'_i \cdot 1 - \sum_{i=1}^{N} \pi_i \cdot 1 = N \cdot (\bar{\pi}' - \bar{\pi})$$

Les résultats figurent dans le tableau 32 :

Profil	Part des devis	Probabilité moyenne état initial	Probabilité moyenne état final	affaires nouvelles supplémentaires
LA 5	25,1%	24,5%	26,8%	753
LA 6	14,6%	20,0%	23,8%	722
LA 3	8,9%	23,6%	27,7%	479
LA 1	13,9%	26,7%	29,0%	416
LM 3	2,7%	20,8%	25,8%	173
LM 2	1,3%	30,9%	39,1%	138
PA 5	6,8%	7,3%	8,8%	134
LM 4	1,7%	17,3%	22,0%	107
PM 4	4,6%	3,6%	5,3%	102
PM 6	3,6%	4,4%	6,3%	88
LM 7	1,6%	18,6%	22,8%	86
PA 2	3,9%	7,1%	8,8%	86
PM 3	2,4%	5,6%	8,2%	82
PA 6	1,9%	7,3%	9,1%	44
PM 5	0,6%	14,0%	18,7%	39
PM 8	0,9%	5,7%	9,0%	38
LM 5	0,5%	26,2%	32,6%	38
LM 1	0,6%	33,4%	37,8%	33
PA 7	0,8%	7,6%	10,5%	32
PA 1	1,0%	12,9%	15,3%	31
LA 2	0,5%	18,3%	22,7%	28
LM 8	0,3%	16,7%	23,4%	26
LA 7	0,3%	15,8%	20,5%	21
LM 6	0,2%	22,7%	32,5%	21
LA 4	0,2%	15,5%	21,3%	16
PM 7	0,2%	12,4%	18,3%	15
PM 2	0,2%	14,6%	19,4%	14
PM 1	0,2%	14,4%	20,0%	12
PA 4	0,3%	6,3%	9,8%	12
PA 3	0,2%	10,5%	13,9%	9
PA 8	0,2%	6,7%	9,8%	8

TABLE 32 – Impacts en volume attendu d'affaires nouvelles supplémentaires

Comme prévu, les résultats sont assez différents. Les plus forts impacts en termes de volume correspondent globalement aux profils les plus volumineux. Le profil cible au plus fort impact est le profil LA 3. Même si un passage de 0 à 100% de devis avec mois gratuits n'augmente la probabilité de conversion que de 17% sur ce profil, cela représente près de 500 affaires nouvelles supplémentaires. Cette observation fait de ce profil la cible la plus intéressante pour l'allocation des mois gratuits. A l'inverse, l'allocation complète de mois gratuits aux cinq derniers profils cibles de ce tableau (LA 4, PM 7, PA 4, PA 3 et PA 8) apporterait en tout moins de 60 affaires nouvelles supplémentaires. Bien que le taux de conversion augmente fortement sur ces profils, ce ne sont donc pas des profils à privilégier pour l'attribution de mois gratuits.

5.2.2 Application du levier : réallocation des mois gratuits

Nous souhaitons appliquer le levier des mois gratuits à notre portefeuille de devis. Les analyses précédentes suggèrent un bon potentiel de croissance de la probabilité de transformation sur certains profils. Cependant, il n'est pas possible en pratique d'accorder des mois gratuits à tous les devis. Dans la réalité, une enveloppe budgétaire fixe est définie pour la distribution des offres promotionnelles. C'est pourquoi nous proposons ici un exemple d'utilisation du levier des mois gratuits tenant davantage compte de ce type de contrainte.

Nous partons de la répartition actuelle des mois gratuits parmi les devis de la base. Pour approximer la contrainte iso-budget, nous formulons une contrainte d'iso-volume de mois gratuits octroyés sur l'ensemble des devis de la base d'étude. L'idée est de conserver le même nombre de mois gratuits et de réallouer ces derniers en les distribuant exclusivement aux profils cibles. Autrement dit, les mois gratuits sont transférés des profils hors cible vers les profils cibles. A l'issue de cette réallocation, nous resimulons la conversion. Nous espérons ainsi parvenir à améliorer le taux de transformation chez les profils cibles et renforcer la part cible parmi les affaires nouvelles.

La réallocation est faite en priorisant les profils cibles sur lesquels le modèle prédit les plus forts impacts. Autrement dit, nous commençons par distribuer les mois gratuits aux devis du profil qui a enregistré le plus gros impact avant de passer au profil suivant, et ainsi de suite. Plutôt que l'impact sur la probabilité de conversion, la priorisation est portée sur l'impact en volume d'affaires nouvelles supplémentaires.

Les étapes du programme sont les suivantes :

1. les profils cibles sont triés par ordre décroissant d'impact.
2. Le nombre n de mois gratuits à redistribuer est déterminé. Il s'agit du nombre total de devis hors cible initialement pourvus de mois gratuits.
3. Ces n mois gratuits des devis hors cible sont désactivés et sont à redistribuer parmi les cibles.
4. Le premier profil cible est sélectionné. Tant qu'il reste des mois gratuits à distribuer :
   - 4.1 Le nombre m de devis cibles du profil qui ne possèdent pas encore de mois gratuits est déterminé.
   - 4.2 Si $n > m$ alors la variable de mois gratuits est activée chez les m devis. Sinon, les devis recevant un mois gratuit sont sélectionnés aléatoirement.
   - 4.3 Le nombre de mois gratuits restant à distribuer est diminué du nombre m de mois gratuits qui ont été distribués.
   - 4.4 On passe au profil suivant.

Le volume d'offres de mois gratuits parmi les devis hors cible est égal à 10 898. Ces 10 898 offres sont désactivées chez les devis hors cible et redistribuées parmi les profils cibles selon le programme détaillé ci-dessus. Les tableaux 33 et 34 permettent de visualiser les résultats de la réallocation :

	Devis	répartition des devis	taux de MG parmi les devis	répartition des MG	probabilité moyenne de conversion	volume attendu d'AFN	répartition des AFN
Total	130449	100%	15%	100%	19,28%	25151	100%
Cible	57218	44%	16%	45%	17,36%	9931	39,5%
Hors cible	73231	56%	15%	55%	20,78%	15219	60,5%

TABLE 33 – Caractéristiques du portefeuille avant la réallocation

	Devis	répartition des devis	taux de MG parmi les devis	répartition des MG	probabilité moyenne de conversion	volume attendu d'AFN	répartition des AFN
Total	130449	100%	15%	100%	19,32%	25205	100%
Cible	57218	44%	35%	100%	18,10%	10354	41,1%
Hors cible	73231	56%	0%	0%	20,28%	14851	58,9%

TABLE 34 – Caractéristiques du portefeuille après la réallocation

La réallocation permet d'apporter 423 affaires nouvelles sur les profils cibles, en augmentant la probabilité moyenne de conversion de 0,74 points. Il s'agit d'un résultat très encourageant. En partie compensé par la perte d'affaires sur les profils hors cible privés de mois offerts, le gain sur les cibles procure au global un gain de 54 affaires nouvelles. Ce gain traduit une augmentation du taux de transformation global prédit de 19,28% à 19,32%. En plus de cette hausse, la répartition des affaires nouvelles devient plus favorable, puisqu'elle passe d'un rapport de 39,5/60,5 à 41,1/58,9 entre cible et hors cible.

Grâce au levier des mois gratuits, nous parvenons ainsi à améliorer "gratuitement" le volume d'affaires nouvelles parmi les profils cibles. En effet, le volume de mois gratuits (assimilé au budget) est conservé. L'amélioration résulte exclusivement d'une meilleure allocation des mois gratuits. De plus, elle est accompagnée d'un renforcement de la part d'affaires cible au détriment des affaires hors cible. Nous obtenons ainsi une structure des affaires nouvelles plus favorable. En définitive, seuls les deux premiers profils cibles LA 1 et LA 3 ont hérité des mois gratuits hors cible, en raison de leur fort volume de devis initialement dépourvus de mois gratuits. Ces deux profils ayant des primes assez faibles, nous pouvons supposer qu'ils n'auraient en réalité pas consommé tout le budget de mois gratuits. Nous aurions donc probablement pu aller encore plus loin dans l'utilisation du levier.

5.3 Troisième levier : choix de l'agent

5.3.1 Modélisation du levier

Le troisième levier que nous imaginons est l'optimisation du choix de l'agent destinataire du flux. Ce levier ne concerne donc que les devis envoyés en agence. Les règles actuelles d'attribution des agents sont les suivantes :

- Pour être orienté vers un agent, il faut que le prospect ait cliqué sur "rencontrer un conseiller" ou "recevoir le devis par mail", que son tarif ne soit pas affiché ou qu'il soit reconnu parmi les clients existants.
- S'il est client, le flux sera transmis à son agent. Sinon, c'est l'agent le plus proche qui recevra son contact.

Imaginons qu'un devis puisse être envoyé à différents agents $A_1, ..., A_n$. Le levier que nous explorons consiste à orienter le flux vers l'agent qui a la plus grande probabilité de transformer le devis. Dans la section 3.6, nous avons construit un modèle agent à partir du sous-ensemble des devis envoyés en agence. Nous utilisons ce modèle pour étudier ce levier. Le modèle comporte deux variables agent :

- la variable ancienneté de l'agent, qui possède 4 modalités
- la variable ancienneté dans le multi-accès, qui possède 3 modalités

Le croisement de ces deux variables forme 12 modalités possibles. Le modèle agent permet ainsi de simuler la conversion en faisant varier les deux variables agent selon 12 combinaisons possibles. Nous imaginons donc la modélisation suivante : chaque devis peut être envoyé à 12 catégories d'agent $A_1, ..., A_{12}$, qui sont associées aux 12 modalités possibles du croisement :

	ancienneté en multi accès		
ancienneté	<1 an	1 an	2 ans ou plus
<3 ans	A_1	A_2	A_3
3-12 ans	A_4	A_5	A_6
13-20 ans	A_7	A_8	A_9
>20 ans	A_{10}	A_{11}	A_{12}

TABLE 35 – Catégories d'agent

Le critère géographique reste un critère important. Par conséquent, nous faisons l'hypothèse qu'il y a au moins un agent de chaque catégorie dans le rayon géographique de chaque prospect.

5.3.2 Identification de la meilleure et de la moins bonne catégorie d'agent

Le sous-échantillon des devis envoyés en agence contient 66 607 observations. Grâce au modèle agent, nous pouvons simuler la probabilité de conversion des devis en fonction de la modalité-agent choisie. Les autres variables des devis ne sont pas modifiées. Une probabilité de conversion est calculée pour chaque devis et chacune des 12 catégories d'agent. Nous obtenons alors une matrice des probabilités de conversion, de taille 66 607 $\times$ 12. Chaque ligne représente un devis et contient l'estimation des 12 probabilités de conversion en fonction de la catégorie d'agent choisie.

Nous souhaitons étudier l'impact du choix de la catégorie d'agent dans le phénomène de conversion. Pour cela, nous commençons par calculer la probabilité de conversion moyenne pour les différentes catégories d'agent. Les probabilités mesurées sont présentées par ordre décroissant dans le tableau 36 :

catégorie	**ancienneté agent**	**ancienneté MA**	**probabilité moyenne**
A_9	13-20 ans	2 ans ou plus	8,02%
A_8	13-20 ans	1 an	7,90%
A_7	13-20 ans	<1 an	7,80%
A_3	<3 ans	2 ans ou plus	7,37%
A_{12}	>20 ans	2 ans ou plus	7,20%
A_5	3-12 ans	1 an	7,04%
A_4	3-12 ans	<1 an	6,95%
A_2	<3 ans	1 an	6,95%
A_1	<3 ans	<1 an	6,86%
A_6	3-12 ans	2 ans ou plus	7,46%
A_{11}	>20 ans	1 an	5,68%
A_{10}	>20 ans	<1 an	5,60%

TABLE 36 – Probabilités moyennes des catégories d'agent

Nous observons un écart non négligeable entre les probabilités moyennes des différentes catégories. La catégorie A_9 possède la meilleure probabilité moyenne de conversion. Cette catégorie représente un agent cumulant une grande ancienneté à la fois dans l'exercice du métier et dans le protocole multi-accès. La catégorie A_{10} possède quant à elle la plus faible probabilité moyenne de conversion. Elle représente un agent en fin de carrière mais ayant fraîchement adhéré au protocole multi-accès.

Nous souhaitons vérifier si la catégorie A_9 est la meilleure pour chaque devis. Pour cela, nous identifions ligne par ligne la meilleure et la moins bonne des catégories d'agent pour le devis. La répartition des premières places entre les différentes catégories est la suivante :

catégorie	**ancienneté agent**	**ancienneté MA**	**meilleur candidat**
A_9	13-20 ans	2 ans ou plus	49,4%
A_7	13-20 ans	<1 an	25,0%
A_6	3-12 ans	2 ans ou plus	14,2%
A_8	13-20 ans	1 an	8,6%
A_{12}	>20 ans	2 ans ou plus	2,0%
A_5	3-12 ans	1 an	0,5%
A_4	3-12 ans	<1 an	0,3%

TABLE 37 – Proportion de "victoires" des catégories d'agent

Nous constatons qu'il n'existe pas de meilleure catégorie unique conférant à tous les devis le plus de chance d'être transformés. Au contraire, la meilleure catégorie varie d'un devis à l'autre. Cette pluralité est permise par la structure d'arbre du modèle *Gradient Boosting*. En effet, une telle structure permet l'interaction entre les variables liées à l'agent et les autres variables, ce qui explique que la meilleure modalité puisse différer d'un devis à l'autre. Néanmoins, les premières places ne sont pas si équilibrées pour autant. La catégorie A_9, est dans un cas sur deux celle qui donne le plus de chance au devis d'être converti. Par ailleurs, 6 des 12 catégories d'agent ne donnent jamais la meilleure probabilité sur l'échantillon.

De manière analogue, nous comptabilisons le nombre de dernières places de chaque catégorie d'agent. La répartition observée entre les différentes catégories est la suivante :

catégorie	**ancienneté agent**	**ancienneté MA**	**pire candidat**
A_{10}	>20 ans	<1 an	98%
A_7	13-20 ans	<1 an	0,6%
A_{12}	>20 ans	2 ans ou plus	0,5%
A_4	3-12 ans	<1 an	0,4%
A_6	3-12 ans	2 ans ou plus	0,2%
A_2	<3 ans	1 an	0,1%
A_9	13-20 ans	2 ans ou plus	0,1%
A_1	<3 ans	<1 an	0,1%
A_3	<3 ans	2 ans ou plus	0,02%

TABLE 38 – Proportion de "défaites" des catégories d'agent

Les résultats sont beaucoup moins partagés : la catégorie A_{10} donne la moins bonne probabilité de conversion dans 98% des cas. Cette catégorie est donc presque systématiquement la moins bonne.

D'après ces premières analyses, l'application du levier agent nécessite de calculer au préalable toutes les probabilités de conversion du devis à l'aide du modèle, comme nous l'avons fait, puis identifier l'agent qui maximise la probabilité. Le moins bon agent est moins incertain et nous pourrions envisager de l'identifier systématiquement à la catégorie A_{10}. Cependant, nous choisissons de le déterminer pour chaque devis, de la même manière que le meilleur agent.

5.3.3 Etude de l'impact de l'agent sur la probabilité de conversion

Afin d'estimer l'impact du levier, nous mesurons à présent pour chaque devis la variation relative de la probabilité de conversion lors d'un passage du pire au meilleur agent. Pour cela, nous mesurons pour chaque devis l'écart relatif de la plus basse à la plus haute probabilité de conversion. La distribution des écarts mesurés est la suivante :

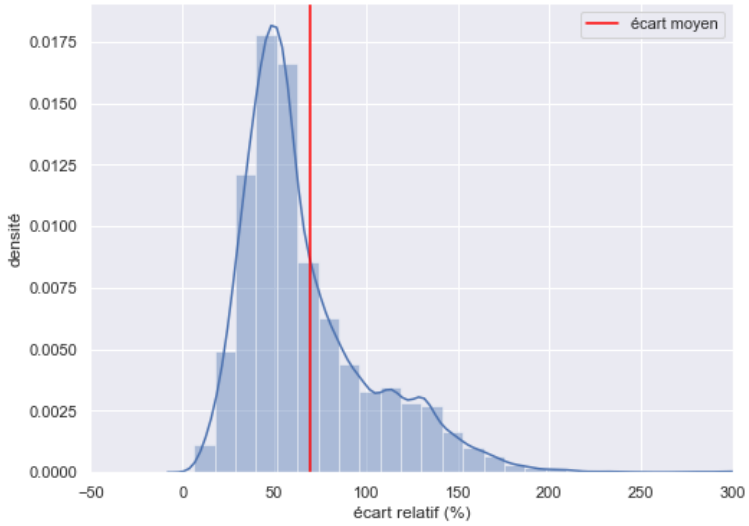

FIGURE 46 – Distribution des variations relatives de probabilité de conversion

En moyenne, la probabilité de transformation augmente de 69% lors du passage du moins bon au meilleur agent. Cet écart est concentré autour de 50%. La probabilité de conversion est même doublée dans près de 18% des cas. L'effet du passage du moins bon agent au meilleur semble donc significatif.

Nous cherchons à savoir pour quels profils cette opération aurait potentiellement le plus d'effet. Pour cela, nous mesurons la probabilité moyenne de conversion par profil dans deux états :

- état 1 : chaque devis est orienté vers son moins bon agent
- état 2 : chaque devis est orienté vers son meilleur agent

L'écart entre les deux probabilités moyennes est mesuré pour chaque profil, et présenté dans le tableau suivant dans l'ordre décroissant :

profil	probabilité moyenne dans l'état 1	probabilité moyenne dans l'état 2	évolution
PM 8	2,2%	4,6%	+106,1%
PM 4	1,7%	3,3%	+98,4%
PM 6	2,1%	4,0%	+89,2%
PM 7	5,9%	10,5%	+79,5%
PM 5	5,9%	10,3%	+73,7%
PM 1	5,5%	9,4%	+72,4%
PM 3	2,6%	4,5%	+71,1%
LM 8	5,2%	8,7%	+68,2%
PM 2	4,9%	8,3%	+67,8%
LA 7	6,1%	10,1%	+64,9%
PA 7	5,2%	8,5%	+64,5%
PA 4	3,9%	6,4%	+63,2%
LA 2	7,2%	11,6%	+60,7%
LM 4	4,1%	6,5%	+59,4%
LM 6	9,1%	14,4%	+58,2%
PA 8	5,4%	8,4%	+56,0%
LA 4	6,7%	10,3%	+54,8%
LM 7	5,2%	8,0%	+53,0%
LM 3	5,2%	7,8%	+51,6%
LA 5	7,5%	11,3%	+51,4%
PA 6	3,8%	5,7%	+50,6%
PA 5	3,0%	4,4%	+49,5%
LA 6	6,0%	8,8%	+47,8%
LA 1	8,1%	11,9%	+45,9%
PA 2	3,6%	5,2%	+45,4%
PA 3	12,1%	17,6%	+45,3%
LA 3	6,8%	9,7%	+43,9%
LM 2	10,3%	14,7%	+42,8%
PA 1	7,6%	10,8%	+42,1%
LM 1	10,3%	14,5%	+41,6%
LM 5	9,0%	12,8%	+41,6%

TABLE 39 – Evolution de la probabilité moyenne des profils entre le meilleur et le pire agent

Les profils dont les variations relatives sont les plus élevées sont tous les profils propriétaires de maison. Dans notre modélisation, ces profils sont donc les plus sensibles au choix de l'agent. La valeur la plus élevée est mesurée chez un profil cible : le profil PM 8. Ce profil correspond aux retraités propriétaires d'une résidence principale, et est également fortement sensible aux autres leviers. La probabilité moyenne de conversion de ce profil est doublée lors du passage du moins bon au meilleur agent pour chacun des devis. Les profils aux variations les plus faibles sont davantage diversifiés : nous retrouvons certains profils de locataires de maison ou d'appartement ainsi que des profils de propriétaires d'appartement. Les deux profils les moins sensibles à l'application du levier sont les profils LM 5 et LM 1. Il s'agit des locataires de maisons de moins de trois pièces. Parmi les cibles, les deux profils les moins sensibles sont les profils PA 1 et LA 3. Il s'agit des propriétaires d'appartement secondaire âgés de moins de 64 ans, ainsi que des locataires d'appartement de plus de deux pièces âgés de moins de 65 ans, tous deux en rez-de-chaussée ou dernier étage. La probabilité de conversion moyenne de ces profils augmente tout de même pratiquement de moitié sous l'effet du levier.

Ces résultats révèlent un bon potentiel de croissance de la probabilité de transformation lorsque les devis sont envoyés au meilleur agent. Par rapport au moins bon agent, cela peut a priori représenter une croissance moyenne de 41 à 106% de la probabilité de transformation, selon le profil considéré. Ces chiffres correspondent toutefois au cas limite où tous les devis d'un profil sont passés de leur moins bon à leur meilleur agent. Il faut s'attendre à des résultats plus modérés pour des devis initialement orientés vers une catégorie d'agent intermédiaire.

Par ailleurs, cette mesure d'impacts ne tient pas compte du volume de devis des profils. Dans l'optique d'une application du levier, il serait plus pertinent d'étudier le volume d'affaires nouvelles supplémentaires. Cela permettrait de prioriser plus efficacement les profils sur lesquels appliquer le levier.

5.3.4 Application du levier : réorientation des devis

Comme pour le levier des mois gratuits, nous souhaitons illustrer l'effet du levier agent en appliquant une réorientation des devis cibles. Cette réorientation consiste à attribuer systématiquement à chaque devis cible l'agent qui lui offre la meilleure probabilité de conversion.

Les statistiques initiales sont les suivantes :

	Devis	Répartition des devis	Cas d'orientation non optimale	Probabilité de conversion moyenne	Volume attendu d'affaires nouvelles	Répartition des affaires nouvelles
Total	66607	100%	58243	7,34%	4888	100%
Cible	26963	40%	23583	6,93%	1868	38,2%
Hors cible	39644	60%	34660	7,62%	3020	61,8%

Table 40 – Caractéristiques initiales du portefeuille agent

Les devis cibles sont orientés vers leur meilleur agent, puis les probabilités de transformation sont recalculées à l'aide du modèle. Les résultats sont les suivants :

	Devis	Répartition des devis	Cas d'orientation non optimale	Probabilité de conversion moyenne	Volume attendu d'affaires nouvelles	Répartition des affaires nouvelles
Total	66607	100%	34660	7,76%	5171	100%
Cible	26963	40%	0	7,98%	2152	41,6%
Hors cible	39644	60%	34660	7,62%	3020	58,4%

Table 41 – Caractéristiques du portefeuille agent après réorientation des devis

L'application du levier sur les 40% de devis cibles permet d'augmenter le taux de transformation de 6% sur l'ensemble de la base, et de 16% sur les profils cibles. Cela représente un gain de 234 affaires nouvelles. De plus, la répartition entre profils cibles et hors cible devient plus favorable, passant ainsi d'un rapport 38/62 à un rapport 42/58.

Limites de cette application :

Dans la pratique, il faudrait néanmoins tenir compte de plusieurs contraintes d'attribution des agents, à commencer par la contrainte géographique. Le modèle dont nous disposons ne nous permet pas de mettre en oeuvre cette contrainte géographique. Nous avons donc fait l'hypothèse d'existence d'au moins un agent de chaque catégorie à proximité du prospect. Nous aurions pu également envisager une contrainte sur la quantité de devis envoyés aux agents. En effet, nous pouvons imaginer que le taux de conversion évolue en fonction du nombre de devis traités par l'agent. La réalité est ainsi plus complexe que ce que nous avons pu modéliser. Enfin, notons que ce levier ne peut être appliqué qu'aux devis envoyés en agence. Cela représente la moitié des devis de notre base d'étude. Les résultats obtenus par ce levier seront donc plus ou moins dilués dans l'échantillon global en fonction de la proportion de devis envoyés en agence.

Conclusion générale

L'objectif de ce mémoire était de proposer des leviers d'amélioration du taux de transformation des devis d'origine digitale, pour le produit multirisque habitation d'Allianz.

Tout d'abord, nous avons cherché à élaborer un modèle de prédiction de la probabilité de transformation des devis. Le premier défi majeur lié à cet objectif a été la construction de la base d'étude. Une fois cela fait, nous avons entraîné plusieurs modèles de conversion sur l'échantillon d'apprentissage. Trois modèles différents ont été entraînés : une régression logistique, un *Random Forest* et un *Gradient Boosting*. Les trois modèles de conversion ont principalement utilisé la prime, les mois gratuits, le parcours web, le nombre de pièces, la qualité juridique et le délai souhaité avant la prise d'effet du contrat. L'étude des *odd ratios* de la régression logistique a révélé que la probabilité de conversion est particulièrement sensible à ces variables. Par rapport à une prime inférieure à 100€ et toutes choses égales par ailleurs, un devis a ainsi 1,8 fois plus de chances d'être converti si sa prime est comprise entre 100 et 115 euros, et 4,1 fois moins de chances d'être converti si elle est supérieure à 410 euros. La probabilité de conversion est 1,3 fois plus élevée si des mois gratuits sont proposés, et 4,2 fois plus faible si le prospect s'arrête sur le comparateur plutôt que de continuer sur le parcours complet du site allianz.fr.

Les trois modèles ont montré de bonnes performances de classification sur l'échantillon de test. Ils affichent des valeurs d'AUC comprises entre 76% et 79%. Cependant, le modèle *Random Forest* ne prédit pas correctement la probabilité de conversion et a donc été écarté. En définitive, la comparaison des différentes métriques d'évaluation entre le GLM et le *Gradient Boosting* nous a conduit à choisir ce second modèle. Par anticipation de l'analyse des leviers, nous avons également élaboré un modèle *logit* captant les effets de la prime en différenciant les segments, ainsi qu'un modèle *Gradient Boosting* spécifique à la conversion des devis envoyés en agence. Ces deux modèles supplémentaires ont également atteint de bonnes performances de classification, en affichant respectivement un AUC de 76% et 78%, ainsi qu'une bonne estimation des probabilités de conversion.

Nous avons ensuite cherché à identifier des cibles rentables, dont nous souhaitions améliorer le taux de transformation. Pour cela, nous avons cherché un indicateur permettant de mesurer la rentabilité des devis. Nous avons ainsi introduit le PSC. Il constitue une prédiction du ratio de sinistralité à l'échelle du contrat. Des arbres de régression du PSC ont permis de segmenter les devis en 31 profils, associés à un niveau de PSC caractéristique. Afin de choisir les profils à cibler, il nous a alors fallu convenir d'un seuil d'équilibre du PSC. Dans l'idéal, nous aurions souhaité que ce seuil reflète entièrement la rentabilité à l'échelle de la vie du contrat, voire à l'échelle du prospect. Il aurait fallu pour cela disposer d'autres informations qui ne sont pas prises en compte par le PSC, tel qu'une prévision de la durée de vie et de la multi-détention. Nous avons tenté de surmonter ces limites en tenant compte de l'information dont nous disposions quant au niveau de prime et à l'ancienneté moyenne des contrats sur quatre macro-segments. Nous avons ainsi défini quatre valeurs seuils pour les propriétaires et locataires d'appartement et de maison. Les profils dont le PSC caractéristique est inférieur à ces valeurs seuils ont été ciblés, à une marge de tolérance près liée à l'ancienneté moyenne en portefeuille sur les macro-segments.

En définitive, 17 profils ont été ciblés parmi les 31 profils construits. Parmi les profils d'appartement, tous les propriétaires et cinq des sept profils de locataires, principalement des locataires en rez-de-chaussée ou dernier étage de moins de 65 ans, ont été ciblés. En ce qui concerne les maisons, les profils ciblés sont les locataires d'une maison d'au moins 3 pièces et âgés de plus de 59 ans, ainsi que les propriétaires âgés de plus de 44 ans qui sont soit actifs et propriétaires d'une résidence principale, soit retraités et propriétaires d'une résidence secondaire.

Enfin, nous nous sommes appuyés sur nos modèles de conversion afin de tester trois leviers d'amélioration du taux de transformation. Le premier levier que nous avons étudié est le prix du contrat. Pour cela, nous avons cherché à calculer une élasticité au prix. Grâce à un des modèles *logit*, nous avons pu obtenir analytiquement une valeur d'élasticité pour chaque devis. Nous en avons déduit un classement des profils par leur niveau moyen d'élasticité. Parmi les cibles, les profils a priori les plus sensibles à une variation de prix sont les propriétaires d'une maison principale âgés de 44 ans ou plus et actifs, tandis que les moins sensibles à une variation de prix sont les propriétaires d'appartement secondaire en rez-de-chaussée ou dernier étage, âgés d'au moins 64 ans. Toutefois, les valeurs d'élasticité mesurées sont globalement trop faibles, c'est pourquoi nous conservons une certaine prudence vis-à-vis des résultats. Dès lors, une piste d'amélioration envisagée serait d'ajouter d'autres variables liées au prix dans le modèle, notamment un indice de compétitivité du tarif. Par ailleurs, il aurait été intéressant de prolonger l'étude de ce levier par une optimisation tarifaire utilisant l'élasticité-prix.

Le deuxième levier que nous avons exploré est la distribution de mois gratuits. A l'aide de simulations à partir du modèle de conversion, nous avons identifié les profils sur lesquels ce levier semble le plus efficace, d'abord au niveau d'une hausse de la probabilité de conversion puis au niveau du volume d'affaires nouvelles supplémentaires apportées. Parmi les profils ciblés, le plus fort impact sur la probabilité de conversion a été mesuré chez les retraités propriétaires d'une maison, ou d'un appartement principal au rez-de-chaussée ou dernier étage. Pour ces profils, la probabilité de conversion a augmenté de moitié lors du passage d'un état sans mois gratuit à un état où des mois gratuits sont appliqués à tous les devis du profil. En priorisant les profils cibles susceptibles d'apporter le plus d'affaires nouvelles, nous avons ensuite simulé une réallocation ciblée des mois gratuits dans le portefeuille, à iso-volume d'offres promotionnelles distribuées. Cette réallocation a abouti à un accroissement du taux de transformation de 4,2% sur les cibles. Malgré la perte d'affaires hors cible, l'ensemble du portefeuille de devis a tout de même vu son taux de transformation augmenter de 0,2% (0,04 points, soit 54 affaires nouvelles), tout en évoluant vers une répartition plus favorable des affaires nouvelles entre les profils cibles et hors cible.

Le dernier levier que nous avons étudié est le choix de l'agent. Plus précisément, nous avons modélisé le fait d'orienter le devis vers l'agent qui a le plus de chance de le transformer. A l'aide du modèle de conversion spécifique aux devis envoyés en agence, nous avons identifié les profils sur lesquels ce levier est le plus efficace. Parmi les cibles, il s'agit des trois profils propriétaires de maison. Le passage du moins bon au meilleur agent pour chacun des devis du profil, toutes choses égales par ailleurs, a ainsi notamment doublé la probabilité de conversion des retraités propriétaires d'une maison principale. Puis, nous avons appliqué ce levier en simulant une réorientation des devis cibles vers leur meilleur agent respectif. Cette réorientation a permis d'augmenter le taux de conversion des profils cibles de 16%, et celui de l'ensemble du portefeuille de 6%. Là encore, la hausse du taux de transformation des cibles a rendu la structure des affaires nouvelles plus favorable. La poursuite du travail sur ce levier consisterait à complexifier la modélisation des agents et intégrer les différentes contraintes liées à l'orientation des devis.

En définitive, les résultats de notre étude peuvent être considérés comme assez satisfaisants. En complément du levier tarifaire traditionnellement étudié, nous avons proposé deux autres leviers d'amélioration du taux de transformation. En actionnant ces derniers, nous sommes parvenus à accroître le volume espéré d'affaires nouvelles sur des segments rentables, sans modifier le prix. Cette étude offre ainsi des perspectives intéressantes pour l'entreprise, dans le cadre du développement du portefeuille multi-accès. Les travaux effectués pourraient apporter une aide à la décision, en étant repris et adaptés plus précisément aux objectifs du décideur. En intégrant une vision plus complète de la rentabilité, ce dernier pourrait par exemple cibler en priorité des

segments rentables, avec un taux de transformation assez bas et un bon potentiel de réaction aux leviers. Dans le prolongement des travaux menés, il serait notamment intéressant d'étudier l'efficacité d'une combinaison des différents leviers. Il pourrait également être intéressant de mener une étude comparative sur les devis issus du réseau traditionnel.

Références

[1] *L'assurance française : Données clés 2020.* Fédération française de l'assurance.

[2] *Documentation de la librairie python scikit-learn.* Consultable sur `https://scikit-learn.org/stable/user_guide.html`.

[3] Niculescu-Mizil A. and Caruana R. *Predicting Good Probabilities With Supervised Learning.* Department Of Computer Science, Cornell University.

[4] Boueddine M. *Assurance Automobile : analyse de l'impact d'une variation du tarif sur le comportement des assurés lors de l'acte de souscription et de résiliation*, 2013. Mémoire d'actuariat, EURIA.

[5] Bouattour O. *Proposition d'une approche de la tarification affaire nouvelle prenant en compte le comportement du prospect*, 2019. Mémoire d'actuariat, Université Paris-Dauphine.

[6] Becker J.-R. *La distribution de l'assurance à l'ère digitale : évolution ou révolution ?*, 2012. Thèse MBA ENASS.

[7] Rakotomalala R. *Pratique de la Régression Logistique, Régression Logistique Binaire et Polytomique.* Université Lumière Lyon 2.

[8] Kurama V. *Gradient Boosting in classification : not a black box anymore !* Article de blog, consultable sur `https://blog.paperspace.com/gradient-boosting-for-classification/`.

[9] Li C. *A gentle introduction to Gradient Boosting.* College of Computer and Information Science, Northeastern University, Support de cours.

[10] Burnel G. *Orientation tarifaire à l'échéance anniversaire : quelle stratégie pour l'automobile ?*, 2017. Mémoire d'actuariat, ENSAE Paris.

[11] Robert T. *Modélisation du taux de transformation et élasticité au prix*, 2019. Mémoire d'actuariat, ISUP.

[12] Sourisseau J. *Modélisation du taux de transformation et de l'élasticité prix en affaire nouvelle pour l'assurance automobile*, 2012. Mémoire d'actuariat, CEA.

[13] Chambolle C. *La théorie du consommateur, Introduction à la microéconomie*, 2017. Ecole Polytechnique, Support de cours.

Table des figures

Liste des tableaux

Annexe 1 : Sortie du GLM principal

Variable	**coefficient**	**écart type**	**p-value**	**IC inf**	**IC sup** [7]
TOP_resi_pcpale	-0.2419	0.041	0.000	-0.322	-0.162
TOP_MOISG	0.2710	0.025	0.000	0.223	0.319
parcours_deb_fin_AZFR_TAR - FastQ	-1.0121	0.036	0.000	-1.082	-0.942
parcours_deb_fin_Comparateur - Comparateur	-1.4378	0.028	0.000	-1.493	-1.382
parcours_deb_fin_Comparateur - NormalQ	-0.5654	0.024	0.000	-0.613	-0.518
delai_devis_debcontrat_8j-15j	-0.3304	0.032	0.000	-0.393	-0.268
delai_devis_debcontrat_>15j	-0.7925	0.026	0.000	-0.844	-0.741
profession_Cadre	-0.1139	0.027	0.000	-0.166	-0.062
profession_Chef d'entreprise et assimilé	0.2563	0.041	0.000	0.175	0.337
profession_Etudiant	-0.3268	0.039	0.000	-0.402	-0.251
profession_Fonctionnaire	0.2081	0.040	0.000	0.130	0.286
profession_Sans emploi/Retraité	-0.5681	0.025	0.000	-0.618	-0.518
age_assu_26-30	-0.0477	0.026	0.062	-0.098	0.002
age_assu_31-40	-0.1605	0.026	0.000	-0.211	-0.110
age_assu_41-60	-0.1372	0.028	0.000	-0.193	-0.081
age_assu_>60	0.0953	0.049	0.054	-0.002	0.192
occupation_X_type_habitation_Locataire x maison	0.1163	0.032	0.000	0.053	0.180
occupation_X_type_habitation_Propriétaire x appartement	-0.8535	0.035	0.000	-0.922	-0.785
occupation_X_type_habitation_Propriétaire x maison	-0.8112	0.048	0.000	-0.905	-0.718
NB_PIE_HABIT_RSQ_3-4	0.0852	0.021	0.000	0.044	0.126
NB_PIE_HABIT_RSQ_7 et +	-1.8143	0.129	0.000	-2.067	-1.562
prem_tot_ttc_>410	-1.4005	0.063	0.000	-1.525	-1.276
prem_tot_ttc_]100-115]	0.6072	0.037	0.000	0.534	0.681
prem_tot_ttc_]115-140]	0.2323	0.039	0.000	0.157	0.308
prem_tot_ttc_]140-155]	0.4883	0.038	0.000	0.414	0.563
prem_tot_ttc_]155-180]	0.1243	0.041	0.003	0.044	0.205
prem_tot_ttc_]180-215]	-0.1120	0.041	0.007	-0.193	-0.031
prem_tot_ttc_]215-250]	-0.3251	0.044	0.000	-0.412	-0.239
prem_tot_ttc_]250-310]	-0.5351	0.044	0.000	-0.622	-0.448
prem_tot_ttc_]310-410]	-0.9724	0.051	0.000	-1.072	-0.873
zonier_INC_13-18	-0.0524	0.020	0.007	-0.091	-0.014
zonier_CNTGN_14-16	0.0450	0.021	0.030	0.004	0.086
zonier_BDG_19-20	-0.0431	0.021	0.040	-0.084	-0.002
zonier_VOL_24-29	0.0421	0.023	0.064	-0.002	0.087
zonier_DEL_11-29	0.0888	0.028	0.001	0.035	0.143
zonier_DEL_5-10	0.0686	0.024	0.004	0.022	0.115
intercept	-0.1916	0.055	0.001	-0.300	-0.083

7. IC inf et IC sup désignent les bornes inférieure et supérieure de l'intervalle de confiance à 95% de l'estimateur du coefficient

Annexe 2 : *Odd ratios* du GLM principal

Variable	Odd ratio	IC inf	IC sup [8]
prem_tot_ttc_]100-115]	1,835	1,706	1,975
prem_tot_ttc_]140-155]	1,630	1,512	1,756
TOP_MOISG	1,311	1,250	1,376
profession_Chef d'entreprise et assimilé	1,292	1,192	1,401
prem_tot_ttc_]115-140]	1,262	1,170	1,361
profession_Fonctionnaire	1,231	1,139	1,331
prem_tot_ttc_]155-180]	1,132	1,045	1,227
occupation_X_type_habitation_Locataire × maison	1,123	1,054	1,197
age_assu_>60	1,100	0,998	1,212
zonier_DEL_11-29	1,093	1,035	1,154
NB_PIE_HABIT_RSQ_3-4	1,089	1,045	1,134
zonier_DEL_5-10	1,071	1,023	1,122
zonier_CNTGN_14-16	1,046	1,004	1,089
zonier_VOL_24-29	1,043	0,998	1,090
zonier_BDG_19-20	0,958	0,919	0,998
age_assu_26-30	0,953	0,907	1,002
zonier_INC_13-18	0,949	0,913	0,986
prem_tot_ttc_]180-215]	0,894	0,825	0,969
profession_Cadre	0,892	0,847	0,940
age_assu_41-60	0,872	0,825	0,922
age_assu_31-40	0,852	0,810	0,895
TOP_resi_pcpale	0,785	0,725	0,851
prem_tot_ttc_]215-250]	0,722	0,663	0,788
profession_Etudiant	0,721	0,669	0,778
delai_devis_debcontrat_8j-15j	0,719	0,675	0,765
prem_tot_ttc_]250-310]	0,586	0,537	0,639
parcours_deb_fin_Comparateur - NormalQ	0,568	0,542	0,596
profession_Sans emploi/Retraité	0,567	0,539	0,596
delai_devis_debcontrat_>15j	0,453	0,430	0,477
occupation_X_type_habitation_Propriétaire × maison	0,444	0,405	0,488
occupation_X_type_habitation_Propriétaire × appartement	0,426	0,398	0,456
prem_tot_ttc_]310-410]	0,378	0,342	0,418
parcours_deb_fin_AZFR_TAR - FastQ	0,363	0,339	0,390
prem_tot_ttc_>410	0,246	0,218	0,279
parcours_deb_fin_Comparateur - Comparateur	0,237	0,225	0,251
NB_PIE_HABIT_RSQ_7 et +	0,163	0,127	0,210

8. IC inf et IC sup désignent les bornes inférieure et supérieure de l'intervalle de confiance à 95% de l'estimateur de l'*odd ratio*

Annexe 3 : Sortie du GLM prix

Variable	coefficient	écart type	p-value	IC inf	IC sup [9]
TOP_resi_pcpale	-0.1784	0.041	0.000	-0.259	-0.098
TOP_MOISG	0.2386	0.024	0.000	0.191	0.286
TOP_Propriétaire	-1.0125	0.063	0.000	-1.136	-0.890
prime	-0.0067	0.000	0.000	-0.007	-0.006
parcours_deb_fin_AZFR_TAR - FastQ	-0.9643	0.035	0.000	-1.034	-0.895
parcours_deb_fin_Comparateur - Comparateur	-1.4938	0.028	0.000	-1.548	-1.440
parcours_deb_fin_Comparateur - NormalQ	-0.5882	0.024	0.000	-0.635	-0.541
delai_devis_debcontrat_8j-15j	-0.3243	0.032	0.000	-0.387	-0.262
delai_devis_debcontrat_>15j	-0.7893	0.026	0.000	-0.841	-0.738
profession_Cadre	-0.2031	0.055	0.000	-0.312	-0.094
profession_Chef d'entreprise et assimilé	-0.1552	0.081	0.057	-0.315	0.004
profession_Etudiant	-0.3900	0.037	0.000	-0.463	-0.317
profession_Sans emploi/Retraité	-0.8616	0.058	0.000	-0.975	-0.748
age_assu_31-40	-0.2473	0.051	0.000	-0.348	-0.147
age_assu_41-60	-0.5596	0.056	0.000	-0.670	-0.450
age_assu_>60	-0.3185	0.101	0.002	-0.516	-0.121
NB_PIE_HABIT_RSQ_3-4	0.1018	0.020	0.000	0.063	0.141
NB_PIE_HABIT_RSQ_7 et +	-1.7817	0.129	0.000	-2.035	-1.529
zonier_INC_13-18	-0.0542	0.019	0.005	-0.092	-0.016
zonier_CNTGN_14-16	0.0395	0.021	0.056	-0.001	0.080
zonier_BDG_19-20	-0.0435	0.021	0.038	-0.085	-0.002
zonier_VOL_24-29	0.0589	0.023	0.009	0.015	0.103
zonier_DEL_11-29	0.0842	0.028	0.002	0.030	0.138
zonier_DEL_5-10	0.0677	0.024	0.004	0.022	0.114
*profession_Cadre * prime*	0.0004	0.000	0.098	-7.95e-05	0.001
*profession_Chef d'entreprise et assimilé * prime*	0.0020	0.000	0.000	0.001	0.003
*profession_Fonctionnaire * prime*	0.0011	0.000	0.000	0.001	0.001
*profession_Sans emploi/Retraité * prime*	0.0016	0.000	0.000	0.001	0.002
*age_assu_31-40 * prime*	0.0007	0.000	0.005	0.000	0.001
*age_assu_41-60 * prime*	0.0023	0.000	0.000	0.002	0.003
*age_assu_>60 * prime*	0.0023	0.000	0.000	0.001	0.003
*TOP_Propriétaire * prime*	0.0006	0.000	0.021	8.2e-05	0.001
*TOP_Maison * prime*	0.0003	0.000	0.008	7.63e-05	0.001
intercept	0.9648	0.061	0.000	0.846	1.084

9. IC inf et IC sup désignent les bornes inférieure et supérieure de l'intervalle de confiance à 95% de l'estimateur du coefficient

Annexe 4 : Résultats des procédures *Grid Search*

Modèle *Random Forest* :

Profondeur	Nombre d'arbres	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	AUC moyen	Rang
4	100	0,760	0,749	0,746	0,753	0,754	0,753	10
4	200	0,759	0,747	0,746	0,751	0,754	0,751	11
4	300	0,759	0,746	0,745	0,750	0,753	0,751	12
6	100	0,767	0,756	0,755	0,760	0,763	0,760	7
6	200	0,767	0,756	0,755	0,759	0,763	0,760	8
6	300	0,767	0,756	0,755	0,759	0,761	0,759	9
8	100	0,774	0,764	0,763	0,767	0,770	0,768	5
8	200	0,774	0,764	0,763	0,767	0,771	0,768	4
8	300	0,774	0,763	0,763	0,767	0,771	0,768	6
10	100	0,779	0,768	0,770	0,773	0,777	0,773	3
10	**200**	**0,779**	**0,769**	**0,770**	**0,773**	**0,777**	**0,774**	**1**
10	300	0,779	0,769	0,770	0,773	0,777	0,774	2

Modèle *Gradient Boosting* :

Learning rate	Profondeur	Nombre d'arbres	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	AUC moyen	Rang
0,05	3	100	0,776	0,767	0,767	0,771	0,775	0,771	18
0,05	3	200	0,782	0,772	0,774	0,776	0,783	0,777	16
0,05	4	100	0,781	0,773	0,774	0,777	0,783	0,778	15
0,05	4	200	0,786	0,777	0,778	0,781	0,788	0,782	9
0,05	5	100	0,784	0,775	0,777	0,780	0,786	0,780	14
0,05	5	200	0,787	0,778	0,779	0,782	0,789	0,783	5
0,1	3	100	0,782	0,772	0,774	0,776	0,783	0,777	17
0,1	3	200	0,785	0,776	0,778	0,779	0,785	0,781	13
0,1	4	100	0,785	0,777	0,777	0,780	0,787	0,781	10
0,1	4	200	0,788	0,779	0,780	0,781	0,789	0,783	4
0,1	5	100	0,787	0,778	0,780	0,782	0,789	0,783	7
0,1	5	200	0,788	0,778	0,780	0,782	0,789	0,783	3
0,2	3	100	0,785	0,777	0,778	0,779	0,786	0,781	11
0,2	3	200	0,787	0,778	0,781	0,781	0,787	0,783	8
0,2	4	100	0,788	0,779	0,781	0,782	0,789	0,784	2
0,2	**4**	**200**	**0,788**	**0,779**	**0,781**	**0,782**	**0,789**	**0,784**	**1**
0,2	5	100	0,787	0,778	0,780	0,781	0,789	0,783	6
0,2	5	200	0,785	0,775	0,777	0,779	0,787	0,781	12

Nous constatons que les différents modèles construits avec les différents ensembles d'hyperparamètres obtiennent des performances similaires, aussi bien dans le cas du *Random Forest* que dans celui du *Gradient Boosting*. La calibration des hyperparamètres n'a donc pas été si déterminante. De plus, les scores diffèrent peu entre les différents folds. Ces observations mettent en évidence la robustesse de ces algorithmes.