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

le 07 Septembre 2022

Par : CISSOUMA Zié Daouda Joël
Titre : Etude du comportement du *cross-sell* afin d'identifier les facteurs le favorisant

Confidentialité : Oui 2 ans

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

***Membres présents du jury de l'Institut des Actuaires :***
Alexandre YOU
Gabrielle TERRE
Signature :

***Entreprise :***
AXA FRANCE
Signature :

***Membre présent du jury de l'EURIA :***
Pierre AILLIOT

***Directeurs de mémoire en entreprise :***
MOHEE GALI MANJULA
Théophile ROBERT
Signature :

***Invité :***

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

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

**Résumé**

Dans l'environnement concurrentiel du secteur de l'assurance, AXA cherche à aider ses agents à multi-équiper ses clients. Le multi-équipement ou *cross-sell* en anglais est une technique marketing consistant à profiter de l'occasion d'être avec le client à un instant T pour lui faire plusieurs propositions de produits (complémentaires ou indépendants) à la fois afin de répondre au mieux à son besoin.

Les bases utilisées pour la réalisation de cette étude résultent de la fusion des bases Auto et Multirisques habitation. Les produits utilisés sont respectivement les produits « Mon Auto » et « Ma Maison ». Cette démarche est gagnante pour le client et l'entreprise car l'un est conseillé et équipé pour répondre à son besoin de manière efficace et l'autre permet à l'entreprise d'augmenter son chiffre d'affaires.

Dans ce mémoire, nous traiterons du *cross-sell* Auto et MRH (Multi Risques Habitation). Un client réalise un *cross-sell* Auto MRH lorsqu'il souscrit d'abord un contrat Auto et ensuite un contrat MRH dans les 3 mois. L'objectif est d'encourager les agents AXA à augmenter leur taux de *cross-sell* en leur proposant des profils. Une probabilité de réalisation de *cross-sell* est ainsi estimée à partir de différents modèles notamment la régression logistique, les arbres de décision, le modèle XGBOOST. Une étude comparative est alors effectuée pour déterminer le modèle qui répondrait le mieux à nos attentes.

Une fois le modèle choisi, un exemple d'application est proposé pour aider l'agent à utiliser nos résultats.

**Mots clefs:** *cross-sell*, Cross-devis, *Gradient Boosting*, GLM, taux de transformation

**Abstract**

In the competitive environment of the insurance industry, AXA seeks to help its agents to multi-equip its customers. *cross-sell*ing is a marketing technique of taking advantage of the opportunity to be with the customer at an instant T to make several product proposals (complementary or independent) at the same time, in order to best meet their needs.

The databases used for this study result from the merger of the Auto and Multi-risk Home databases. The products used are respectively "Mon Auto" and "Ma Maison". This approach is a win-win situation for the client and the company, as the client is advised and equipped to meet his needs in an efficient manner and the company can increase its turnover.

In this dissertation, we will deal with *cross-sell* Auto and MRH. A customer makes an Auto MRH *cross-sell* when he first signs a contract Auto and then an MRH contract within 3 months. The objective is to encourage AXA agents to increase their *cross-sell* rate by offering them profiles. The probability of *cross-sell*ing is estimated using various models, notably logistic regression, decision trees and the XGBOOST model. A comparative study is then carried out to determine the model that best meets our expectations.

Once the model is chosen, an example of application is proposed to help the agent to use our results.

**Keywords:** *cross-sell*, Cross-devis, Gradient Boosting, GLM, transformation rate

# Remerciements

Je tiens tout d'abord à remercier l'équipe P&C *Business Analytics* pour leurs conseils, leur sympathie et leur disponibilité.

Je remercie particulièrement Manjula MOHEE GALI, manager de l'équipe, tutrice en entreprise de m'avoir accompagné tout au long de cette étude, pour son encadrement rigoureux, sa disponibilité et son attention portée à ce mémoire.

Un grand merci à Théophile ROBERT pour sa sympathie, sa disponibilité, sa rigueur et ses précieux conseils sur la partie actuarielle de cette étude.

Je remercie également Mona ARMANTERAS pour sa patience et ses conseils de rédaction.

Merci à Théo Mallet, Ralitsa JEANNE, Anne-Sophie HERLING, Sarah BRICHET pour leur disponibilité et leur bonne humeur.

Une profonde gratitude envers mon tuteur alternance, Franck VERMET pour ses conseils ainsi qu'à Marine HABART, ma tutrice mémoire pour sa disponibilité et l'intérêt porté à cette étude.

# Note de synthèse

Dans un environnement concurrentiel, cherchant à maximiser son chiffre d'affaires, AXA France décide d'aider ses agents à multi-équiper ses clients. Le multi-équipement ou *cross-sell* en anglais est une technique marketing consistant à profiter de l'occasion d'être avec le client à un instant T pour lui faire plusieurs propositions de produits (complémentaires ou indépendants) à la fois afin de répondre au mieux à son besoin. Cette démarche est gagnante pour le client et l'entreprise car l'un est conseillé et équipé pour répondre à son besoin de manière efficace et l'autre permet à l'entreprise d'augmenter son chiffre d'affaires.

Le cas d'étude qui nous concerne est le *cross-sell* Auto MRH qui correspond au fait de vendre un contrat d'assurance MRH à un client détenant déjà un contrat d'assurance Auto.

Le but de ce mémoire est d'aider les agents à déterminer les profils de clients qui seront susceptibles de faire du *cross-sell* Auto MRH.

Pour mener à bien cette étude, la base de données étant construite sur un périmètre d'étude bien défini, nous cherchons à comprendre le fonctionnement du *cross-sell* Auto MRH à travers différentes analyses.

Pour ce faire, l'ordre des différentes actions permettant d'arriver au *cross-sell* est observé. En effet, toutes les actions peuvent être menées en même temps. Autrement dit, l'individu peut réaliser le même jour les 4 actions suivantes : le devis Auto, l'affaire nouvelle Mon Auto, le devis Ma Maison, l'affaire nouvelle. On l'appellera "*cross-sell* de type1". Ainsi, plusieurs possibilités peuvent ressortir mais les plus majoritaires seront mentionnées dans le tableau ci-dessous :

| Devis Mon Auto | Affaire Nouvelle Mon Auto | Devis Ma Maison | Affaire Nouvelle Ma Maison | Répartition |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 41% |
| 1 | 1 | 1 | 2 | 3% |
| 1 | 1 | 2 | 2 | 36% |
| 1 | 1 | 2 | 3 | 7% |

TABLE 1 – Ordre de réalisation des étapes du *cross-sell*

Il apparait selon le tableau que dans 41% des cas toutes les étapes ont lieu le même jour. Il existe néanmoins des cas où les étapes sont réalisées à différents moments. Pour ces cas, l'écart de temps entre ces différentes actions est étudié. Généralement, l'écart de temps entre deux actions est de 4 mois.

A partir des analyses réalisées, il est constaté que la plupart des clients réalisant le *cross-sell* le font dans les 3 mois. Le schéma suivant permet de mieux appréhender le phénomène du *cross-sell*.

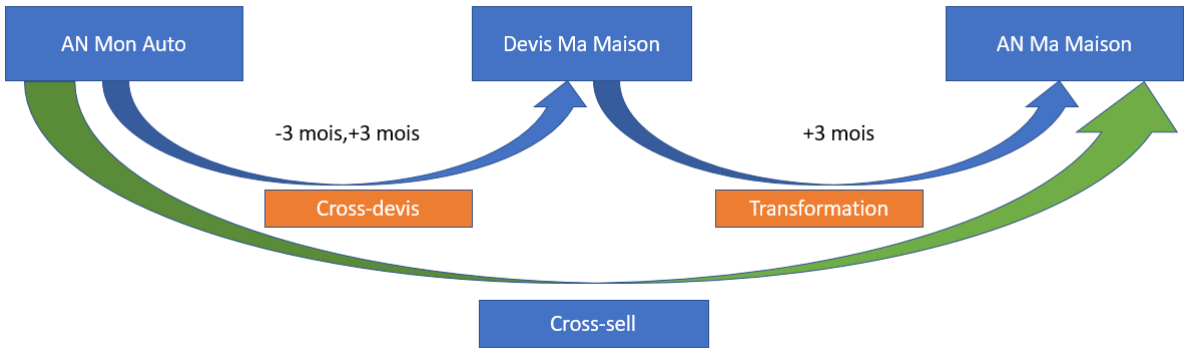



FIGURE 1 – Schéma récapitulant le *cross-sell*, le cross-devis et le taux de transformation

Une première analyse univariée est ensuite réalisée afin d'avoir une idée des profils qui réaliseraient du *cross-sell*. Disposant de peu de variables quantitatives, elles sont toutes discrétisées grâce aux arbres de décision. Il en ressort quelques variables discriminantes comme l'ancienneté du permis de conduire, la multi-détention, l'ancienneté du client dans le portefeuille, la segmentation client, la région, le type de formule Auto... Les données distributeurs ne semblent pas être discriminantes selon les statistiques descriptives.

La méthode *stepwise* est ensuite utilisée pour sélectionner les variables qui auront plus de significativité dans notre modèle. Seulement la variable donnant la catégorie socio-professionnelle est supprimée. Pour réduire davantage nos variables, nous nous référons

à l'ordre d'importance donné par le modèle XGBOOST. Les variables les plus représentantes sont ainsi sélectionnées. Une étude de corrélation est réalisée sur ces variables en fixant le seuil à 60%.

Pour déterminer les profils des individus, 3 modèles ont été utilisés :la régression logistique, les arbres de décisions, le modèle XGBOOST. La performance de ces trois modèles est jugée selon l'AUC sur la base de test, l'interprétabilité et la facilité opérationnelle.

| Modèles | AUC(test) | Interprétabilité | Facilité opérationnelle |
|---|---|---|---|
| GLM | 0.7324 | + | + |
| CART | 0.7128 | + | + |
| XGBOOST | 0.749 | - | - |

TABLE 2 – Comparaison des trois modèles selon l'AUC, l'interprétabilité, la facilité opérationnelle

Parmi ces trois modèles, le modèle logistique est ainsi choisi. Ses prédictions sont acceptables. A partir des différents coefficients déterminés par le modèle, il est possible d'interpréter chacune des modalités des variables. Des profils peuvent être construits en croisant plusieurs modalités de différentes variables et il est possible d'obtenir des probabilités de *cross-sell* associées. La formule utilisée pour la déterminer est l'inverse de la fonction logit. L'interprétation du modèle GLM est relativement simple.

Les résultats du modèle sont les suivants :

| Variables | Modalités | Coefficient | Odd-ratio |
|---|---|---|---|
| Ancienneté permis | [10,41[ | 0 | 1 |
| | <10 | -0,25 | 0,78 |
| | >=41 | 0,37 | 1,45 |
| Multi-détention | Aucun produit | 0 | 1 |
| | mono_auto | -0,73 | 0,48 |
| | mono_autres | 0,16 | 1,18 |
| | multi_auto | -1,02 | 0,36 |
| | multi_autres | -0,72 | 0,49 |
| Ancienneté client | [0,1[ | 0,00 | 1,00 |
| | [1,5[ | -0,55 | 0,58 |
| | >=11 | -0,48 | 0,62 |
| | [5,11[ | -0,49 | 0,61 |
| Mode logement | Locataires et autres | 0 | 1 |
| | Propriétaires | 0,17 | 1,18 |

| | | | |
|---|---|---|---|
| Segment client | B50, Sinistrés, Multi-sinistrés | 0 | 1 |
| | 18-20 | -0,60 | 0,55 |
| | 21-30 | -0,30 | 0,74 |
| | Autres | -0,20 | 0,82 |
| | Creusement de Bonus | -0,20 | 0,82 |
| | Pro | -0,47 | 0,63 |
| | Sans antécédents | -0,78 | 0,46 |
| Formules | F4 | 0 | 1 |
| | F1 | -0,61 | 0,54 |
| | F2 | -0,33 | 0,72 |
| | F3 | -0,10 | 0,90 |
| | F5 | 0,09 | 1,10 |
| | F6 | -0,26 | 0,77 |
| Fractionnement Prime Auto | Mensuel | 0 | 1 |
| | Annuel | -0,39 | 0,68 |
| Situation maritale | Célibataire | 0 | 1 |
| | Autres | 0,43 | 1,54 |
| | COHABITEE | 0,16 | 1,18 |
| | Marié | 0,44 | 1,56 |
| Prime annuelle Auto | [49,199[ | 0 | 1 |
| | [0,49[ | 0,43 | 1,53 |
| | [199,774[ | 0,21 | 1,24 |
| | >=774 | -0,30 | 0,74 |
| Mode d'habitation | Maison | 0 | 1 |
| | Appartement | 0,17 | 1,19 |
| | Autres | -0,75 | 0,47 |
| Type de garage | Autres | 0 | 1 |
| | STREET | -0,12 | 0,89 |
| Segment Epargne | Non spécialisé | 0 | 1 |
| | Expert | 0,08 | 1,08 |
| Région | Région 67 | 0 | 1 |
| | Region64 | -0,19 | 0,82 |
| | Region65 | 0,26 | 1,29 |
| | Region66 | 0,12 | 1,12 |
| | Region68 | 0,09 | 1,10 |
| Agence santé | Active | 0 | 1 |
| | Non Active | -0,22 | 0,80 |

TABLE 3 – Résultats de la régression logistique

Par la suite, un exemple d'application de nos travaux consistera à augmenter le taux de *cross-sell* sur une sous-population. Le modèle qui nous permettra de le faire est l'algorithme CART. Un taux de *cross-sell* de 45% est déterminé sur une population de 6%.

Les profils de cette population sont repérés par les premiers nœuds de l'arbre calibré.

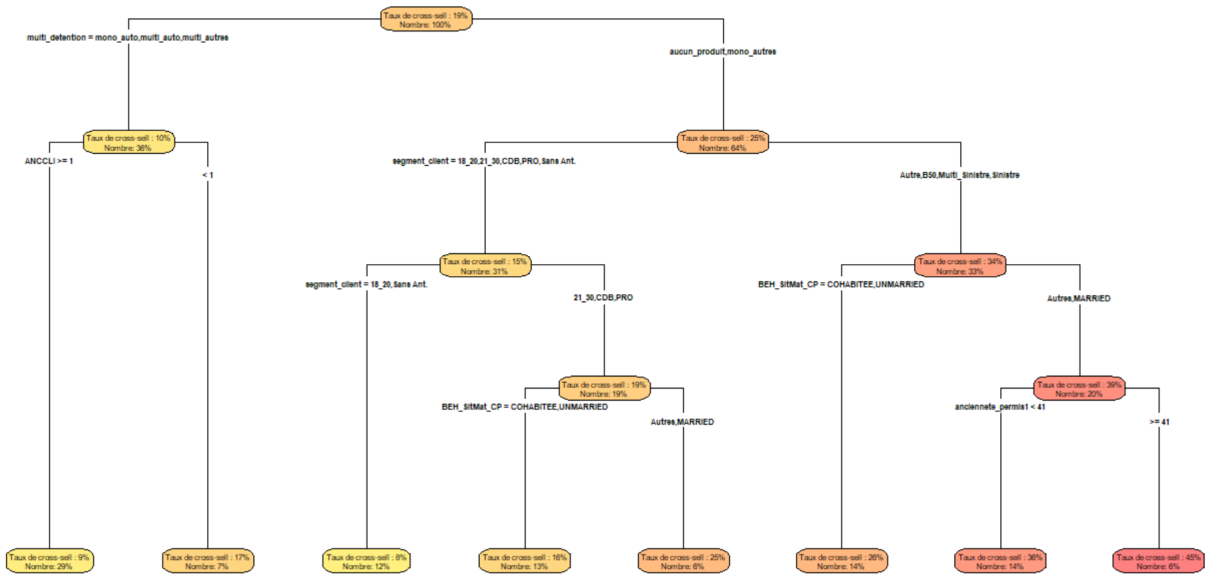



FIGURE 2 – Arbre de décision

— Selon la détention du produit, sont sélectionnées les personnes ne détenant aucun produit donc les prospects et les personnes étant déjà clientes chez AXA mais détenant des contrats uniques autres que l'auto et l'habitation.
— Pour ce qui est de la segmentation client, les sinistrés, les multi-sinistrés, les Bonus 50 et les autres(les personnes n'appartenant à aucun des segments cités dans le premier chapitre) apparaissent.
— Les mariés et les autres(concernent les personnes n'appartenant ni à la catégorie marié et célibataire) sont sélectionnés.

Une autre approche est d'utiliser les scores donnés par le GLM. Avec le modèle GLM, il est possible d'obtenir la probabilité de multi-équipement de chaque individu soit en faisant le croisement de plusieurs variables. Une maquette sous Excel est construite avec un code couleur donnant le taux de *cross-sell* résultant d'un croisement de variables discriminantes. Des croisements donnant des probabilités élevées seraient intéressants pour l'agent.

| Variables | Valeurs | Contribution |
|---|---|---|
| Ancienneté permis | 41 | |
| Multi-détention | Mono autres | |
| Formule Auto | F5 | |
| Situation matrimoniale | marrié | |
| Région | 65 | |
| Total | | 75% |

FIGURE 3 – Exemple d'alerte sur quelques caractéristiques

A partir de cette application, les agents pourront se souvenir par exemple qu'une personne ayant une détention "Mono Autres" aura tendance à augmenter sa possibilité à réussir un *cross-sell*.

Une dernière approche consistera à cibler les agents qui auraient dû réaliser plus de *cross-sell* mais qui ne l'ont pas réalisé. Pour ce faire, nous utiliserons les segments de l'arbre défini plus haut mais en se limitant à l'avant-dernier noeud de la branche de droite.

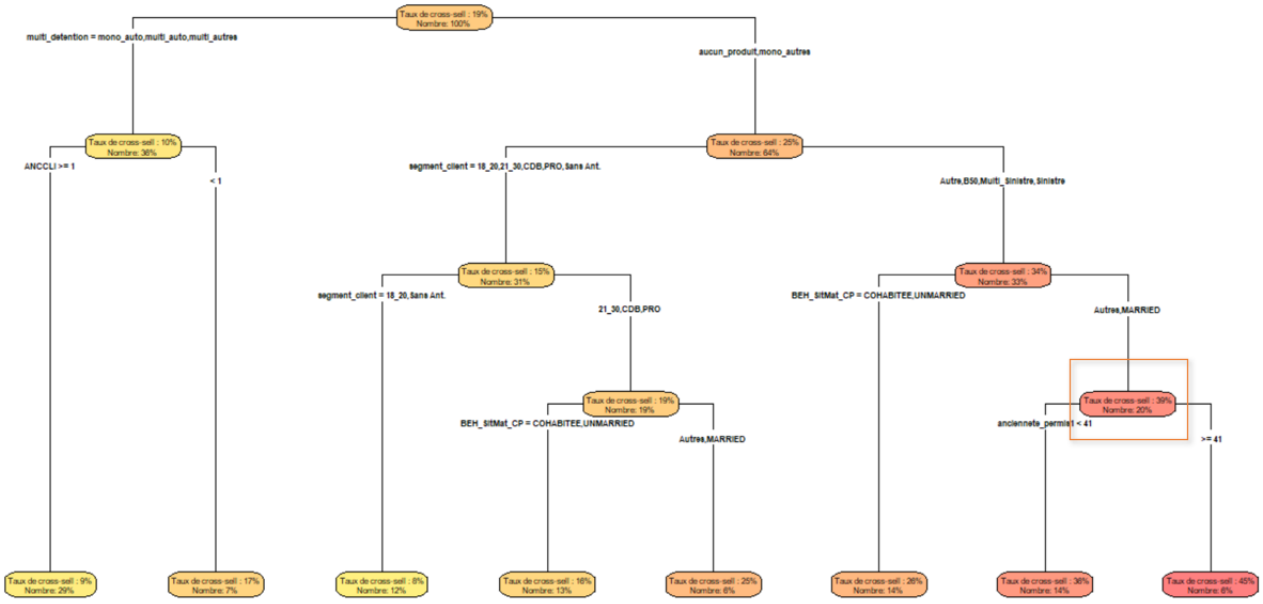



FIGURE 4 – Noeud à utiliser

Le modèle GLM sera utilisé pour la prédiction en filtrant notre base sur les segments de l'arbre. Les agents sont reconnus dans les bases AXA grâce au code gestionnaire. Nous comparons ainsi les taux de *cross-sell* réels et prédits par code gestionnaire.

Pour les agents dont le taux de *cross-sell* prédit est supérieur au taux de *cross-sell* réel, nous cherchons à déterminer combien de contrats nous aurions pu gagner si l'attention avait été portée envers les profils du segment choisi.

Après calcul, nous obtenons un nombre de contrats d'environ 3000 qui aurait pu être gagnés.

Il sera alors possible de contacter ces agents pour montrer cette alerte.

# Summary

In a competitive environment and seeking to maximize its turnover, AXA France decides to help its agents to multi-equip its customers. The multi-equipment or *cross-sell* in English is a marketing technique consisting in taking advantage of the opportunity to be with the customer at a given time to make several product proposals (including complementary or independent) at the same time in order to best meet their needs. This approach is a winner for the client and the company because one is advised and equipped to meet its need effectively and the other allows the company to increase its turnover.

The case study that concerns us is the *cross-sell* Auto MRH which corresponds to the fact to sell an MRH insurance contract to a client who already holds an insurance contract Auto.

The purpose of this dissertation is to help agents determine the profiles of customers who will be likely to *cross-sell* Auto MRH.

To carry out this study, the database being built on a perimeter well-defined study, we seek to understand the operation of the Auto MRH *cross-sell* through different analyses.

To do this, the order of the different actions leading to the *cross-sell* is observed. Indeed, all actions can be carried out at the same time. Other said, the individual can perform the following 4 actions on the same day : the Auto quote, the business new My Auto, the My House quote, the new deal. It will be called "*cross-sell* of type1". Thus, several possibilities can emerge but the most majority will be mentioned in the table below :

| Quotation Mon Auto | Contract Mon Auto | Quotation Ma Maison | Contract Ma Maison | Répartition |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 41% |
| 1 | 1 | 1 | 2 | 3% |
| 1 | 1 | 2 | 2 | 36% |
| 1 | 1 | 2 | 3 | 7% |

TABLE 4 – Order of performing the *cross-sell* steps

It appears from the table that in 41% of cases all the steps take place the same day. However, there are cases where the steps are performed at different times. For In these cases, the time lapse between these different actions is studied. Generally, the difference of time between two actions is 4 months.

From the analyses carried out, it is noted that most of the customers carrying out the *cross-sell* do it within 3 months. The following diagram provides a better understanding of the *cross-sell* phenomenon.

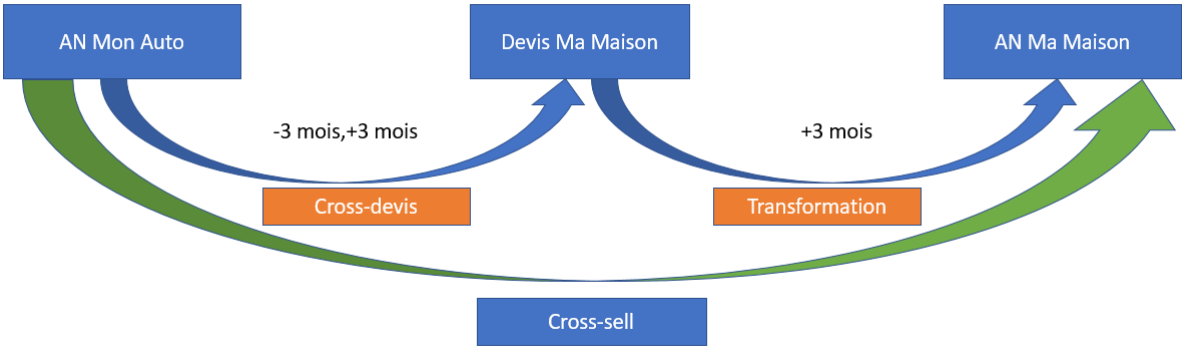



FIGURE 5 – Diagram summarizing the *cross-sell*, the cross-quote and the conversion rate

A first univariate analysis is then carried out in order to have an idea of the profiles that would *cross-sell*. With few quantitative variables, they are all discretized using decision trees. From this analysis emerges a few variables criteria such as the seniority of the driving licence, multiple holdings, seniority of the customer in the portfolio, customer segmentation, region, type of Auto plan. . . The distributor data does not seem to be discriminating according to the descriptive statistics.

With this first analysis, the stepwise method is used to select the variables that will have more significance in our model. Only the variable With this first analysis, the stepwise method is used to select the variables that will have more significance in our model. Only the variable giving the socio-professional category is deleted. In order to further

reduce our variables, we refer to the order of importance given by the XGBOOST model. The most representative variables are thus selected. A correlation study is performed on these variables by setting the threshold at 60%.

To determine the profiles of individuals, 3 models were used : regression logistics, decision trees, the XGBOOST model. The performance of these three models is studied through the AUC obtained during the application on the test base.

| Modèles | AUC(test) | Interprétabilité | Facilité opérationnelle |
|---|---|---|---|
| GLM | 0.7324 | + | + |
| CART | 0.7128 | + | + |
| XGBOOST | 0.749 | - | - |

TABLE 5 – Comparison of the three models according to AUC, interpretability, operational ease

Among these three models, the logistic model is thus chosen. His predictions are acceptable. From the different coefficients determined by the model, it is possible to interpret each of the modalities of the variables. Profiles can be built in crossing several modalities of different variables and it is possible to obtain associated *cross-sell* probabilities. The formula used to determine it is the inverse of the housing function. The interpretation of the GLM model is relatively simple.

The results of the model are as follows :

| Variables | Modalités | Coefficient | Odd-ratio |
|---|---|---|---|
| Ancienneté permis | [10,41[ | 0 | 1 |
| | <10 | -0,25 | 0,78 |
| | >=41 | 0,37 | 1,45 |
| Multi-détention | Aucun produit | 0 | 1 |
| | mono_auto | -0,73 | 0,48 |
| | mono_autres | 0,16 | 1,18 |
| | multi_auto | -1,02 | 0,36 |
| | multi_autres | -0,72 | 0,49 |
| Ancienneté client | [0,1[ | 0,00 | 1,00 |
| | [1,5[ | -0,55 | 0,58 |
| | >=11 | -0,48 | 0,62 |
| | [5,11[ | -0,49 | 0,61 |
| Mode logement | Locataires et autres | 0 | 1 |
| | Propriétaires | 0,17 | 1,18 |

| | | | |
|---|---|---|---|
| Segment client | B50, Sinistrés, Multi-sinistrés | 0 | 1 |
| | 18-20 | -0,60 | 0,55 |
| | 21-30 | -0,30 | 0,74 |
| | Autres | -0,20 | 0,82 |
| | Creusement de Bonus | -0,20 | 0,82 |
| | Pro | -0,47 | 0,63 |
| | Sans antécédents | -0,78 | 0,46 |
| Formules | F4 | 0 | 1 |
| | F1 | -0,61 | 0,54 |
| | F2 | -0,33 | 0,72 |
| | F3 | -0,10 | 0,90 |
| | F5 | 0,09 | 1,10 |
| | F6 | -0,26 | 0,77 |
| Fractionnement Prime Auto | Mensuel | 0 | 1 |
| | Annuel | -0,39 | 0,68 |
| Situation maritale | Célibataire | 0 | 1 |
| | Autres | 0,43 | 1,54 |
| | COHABITEE | 0,16 | 1,18 |
| | Marié | 0,44 | 1,56 |
| Prime annuelle Auto | [49,199[ | 0 | 1 |
| | [0,49[ | 0,43 | 1,53 |
| | [199,774[ | 0,21 | 1,24 |
| | >=774 | -0,30 | 0,74 |
| Mode d'habitation | Maison | 0 | 1 |
| | Appartement | 0,17 | 1,19 |
| | Autres | -0,75 | 0,47 |
| Type de garage | Autres | 0 | 1 |
| | STREET | -0,12 | 0,89 |
| Segment Epargne | Non spécialisé | 0 | 1 |
| | Expert | 0,08 | 1,08 |
| Région | Région 67 | 0 | 1 |
| | Region64 | -0,19 | 0,82 |
| | Region65 | 0,26 | 1,29 |
| | Region66 | 0,12 | 1,12 |
| | Region68 | 0,09 | 1,10 |
| Agence santé | Active | 0 | 1 |
| | Non Active | -0,22 | 0,80 |

Thereafter, an example of application of our work will consist in increasing the rate of *cross-sell* on a sub-population. The model that will allow us to do this is the algorithm CART rhythm. A *cross-sell* rate of 45% is determined on a population of 6%. The profiles of this population are identified by the first nodes of the calibrated tree.

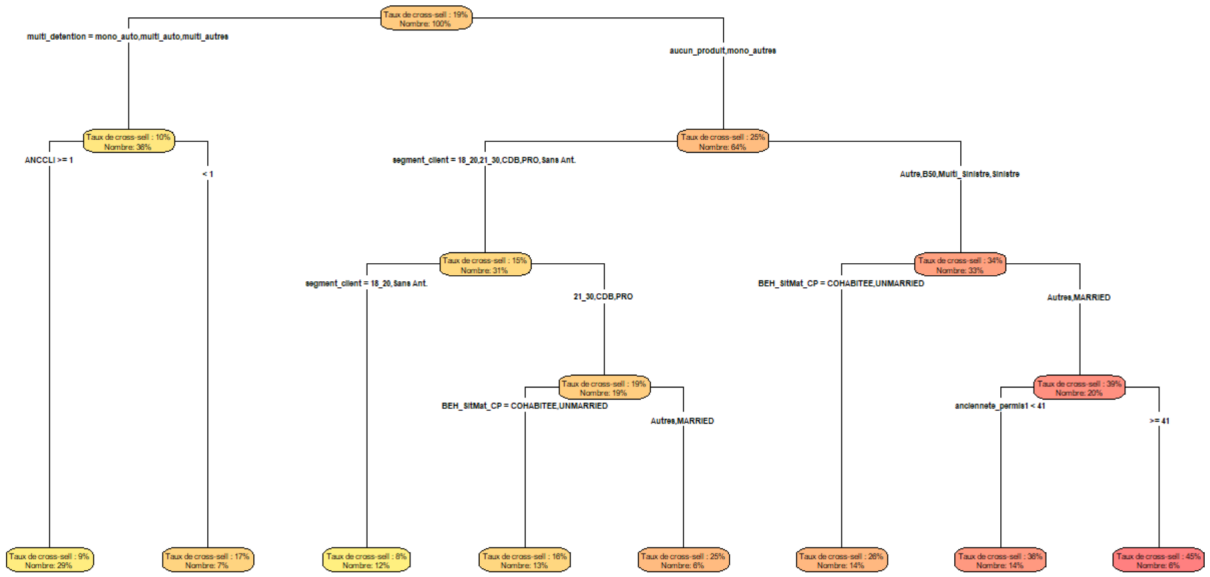



FIGURE 6 – Arbre de décision

— Depending on the holding of the product, people who do not hold any therefore produces prospects and people who are already customers of AXA but holding single policies other than auto and home.
— With regard to customer segmentation, victims, multiple victims, Bonuses 50 and the others (people who do not belong to any of the segments mentioned in the first chapter) appear.
— Married people and others (concerns people who do not belong to the category married and single) are selected.

Another approach is to use the scores given by the GLM. With the GLM model, it is possible to obtain the multi-equipment probability of each individual either by crossing several variables. A model in Excel is built with a color code giving the *cross-sell* rate resulting from a crossing of a discriminant variable. Crosses giving high probabilities would be interesting for the agent.

| Variables | Valeurs | Contribution |
|---|---|---|
| Ancienneté permis | 41 | |
| Multi-détention | Mono autres | |
| Formule Auto | F5 | |
| Situation matrimoniale | marrié | |
| Région | 65 | |
| Total | | 75% |

FIGURE 7 – Example of alert on some characteristics

From this application, agents will be able to remember, for example, that a person with a "Mono autres" holding will tend to increase their possibility to achieve a *cross-sell*.

A third approach will be to target agents who should have achieved more *cross-sell* but who have not realized it. To do this, we will use the segments of tree defined above but limited to the penultimate node of the branch of right.

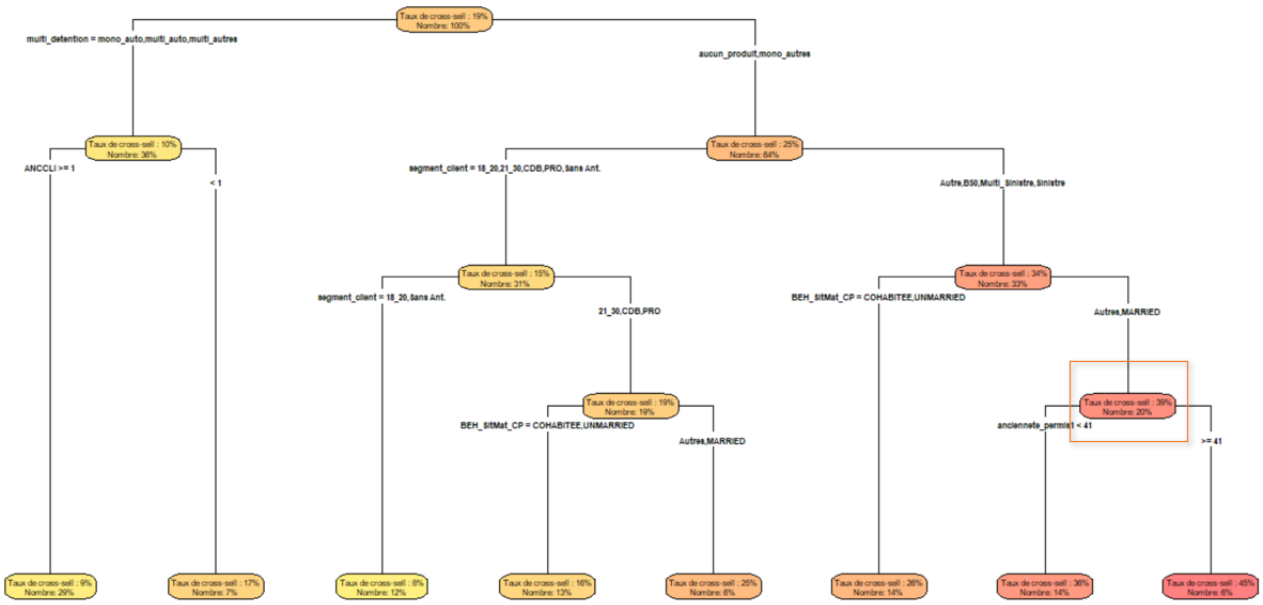



FIGURE 8 – Noeud à utiliser

The GLM model will be used for prediction by filtering our base on segments of the tree. The agents are recognized in the AXA databases thanks to the manager code. We let's compare the actual and predicted *cross-sell* rates by manager code. For agents whose predicted *cross-sell* rate is higher than the actual *cross-sell* rate, we seek to determine how many contracts we could have won if the attention had been brought to the profiles of the selected segment.

After calculation, we get a number of contracts of about 3000 that could have been won.

It will then be possible to contact these agents to show this alert.

# Table des matières

# Introduction

Dans un contexte concurrentiel, augmenter le chiffre d'affaires et fidéliser ses clients font partie des premiers objectifs d'un assureur. De ce fait, le multi-équipement est un moyen qui serait nécessaire pour remédier à cela. En effet, le multi-équipement est une technique marketing visant à proposer plusieurs produits complémentaires ou indépendants à un client afin de mieux répondre à ses besoins. Réussir à multi-équiper un client conduira donc à augmenter le chiffre d'affaires de l'entreprise et à maintenir le client dans son portefeuille.

Sur le portefeuille AXA, les produits auto et habitation sont les produits phares vendus souvent en produit d'entrée à la souscription. Ils représentent ainsi une grande majorité des contrats chez axa. En considérant que tout le monde ait une maison donc devrait avoir une assurance habitation, une couverture MRH est alors proposée. L'étude du *cross-sell* portera donc sur ces deux produits. Il s'agira d'encourager les agents à proposer une couverture MRH à un client possédant déjà une couverture auto.

L'objectif de ce mémoire est de déterminer les différents profils susceptibles de réaliser du *cross-sell* auto MRH. Dans un premier temps, nous présenterons le contexte de l'étude et la base de données construite pour la réalisation de notre étude. Ensuite, nous sortirons quelques analyses afin de comprendre le fonctionnement du *cross-sell* auto MRH. Pour finir, la régression logistique et les modèles de machine learning tels que les arbres de décision et l'*extreme gradient boosting* seront utilisés pour déterminer les profils des assurés. Cela nécessitera une comparaison entre ces différents modèles qui conduira à n'en choisir qu'un seul.

# Chapitre 1

# Contexte général

## 1.1 Le marché de l'assurance IARD en France

L'assurance IARD (Incendie, Accidents et Risques Divers) couramment connue sous le nom d' assurance de biens et de responsabilités est principalement destinée aux particuliers et aux entreprises. Elle prend en charge les sinistres ne relevant pas de la vie humaine.

Le marché de l'assurance IARD se décompose en plusieurs catégories selon le bien à protéger. Cependant, les proportions de cotisations diffèrent d'une catégorie à une autre.

La répartition en termes de cotisation est la suivante :

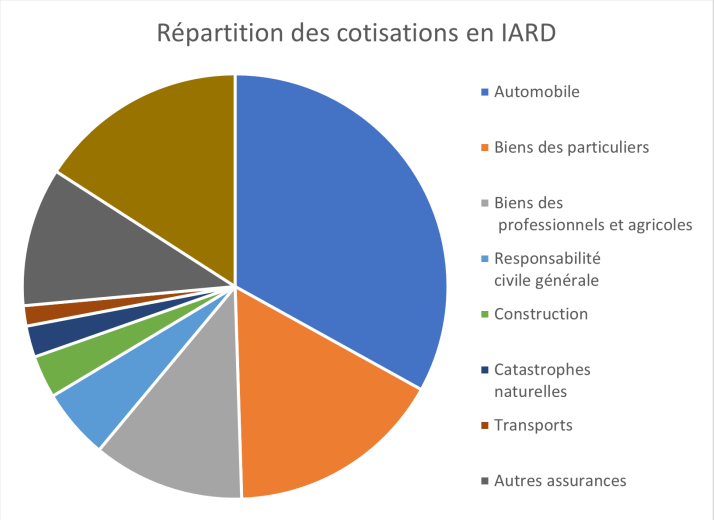



FIGURE 1.1 – Répartition des cotisations sur le marché IARD en France en 2020

Les assurances automobile et habitation occupent les deux premières places sur ce marché. C'est sur ces deux branches que se focalisera ce mémoire.

## 1.2 L'assurance Automobile en France

L'assurance automobile possède plusieurs garanties obligatoires comme facultatives :

— La responsabilité civile : cette garantie est obligatoire et indemnise les dégâts causés à un tiers. Ces dégâts peuvent être corporels comme matériels. Cependant, elle ne s'applique pas dans certains cas notamment en cas de permis de conduire non valide ou en cas d'une action volontaire.

— Les garanties facultatives ou complémentaires : dommages tout accident, dommages collision, Incendie, vol, bris de glace, castatrophes naturelles...

Il est aussi important de souligner que les cotisations perçues par le marché automobile restent en constante évolution depuis 2016.

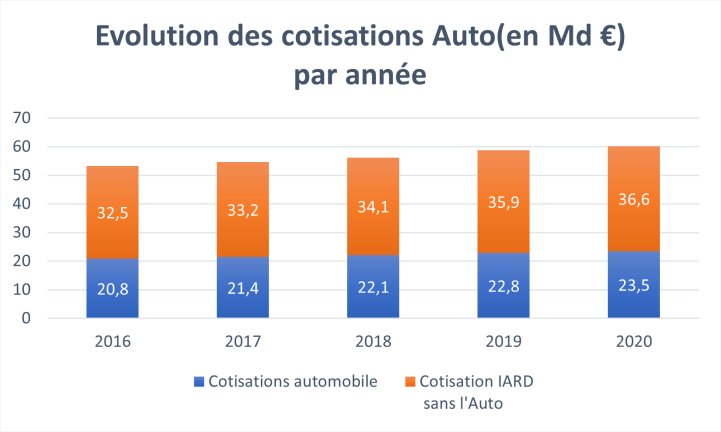



FIGURE 1.2 – Cotisations perçues sur le marché Auto et autres catégories en 2020

## 1.3 L'assurance multirisque habitation

L'assurance habitation est obligatoire pour les locataires, les copropriétaires... Elle présente plusieurs garanties que sont :

— Incendie : Elle indemnise les incidents liés à la fumée, au feu...

— Vol et vandalime : Lors de la signature des contrats, des montants à ne pas dépasser sont fixés par l'assureur en cas de perte, de vol, ...

— Dégâts des eaux : Des prestations sont offertes afin de réparer les dégradations subies par l'assuré et son voisin si celui-ci en est affecté.

— Bris de glace

— Castastrophes naturelles

— Responsabilité civile

Concernant la répartition des cotisations perçues par année, nous nous intéressons au graphique suivant :

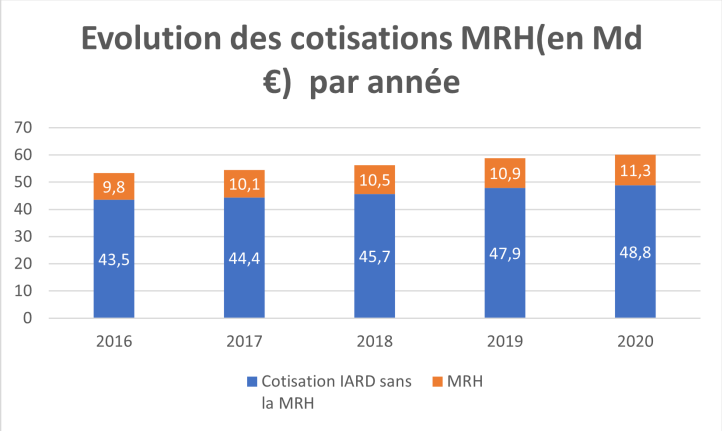



FIGURE 1.3 – Cotisations perçues sur le marché Habitation et autres catégories IARD en 2020

Il est important de noter que les cotisations du marché habitation augmentent au fil des années. Nous passons de 9,3 milliards d'euros en 2016 à 11,3 milliards en 2020.

## 1.4 Présentation d'AXA France

AXA est la résultante de plusieurs sociétés d'assurance dont la plus ancienne fut fondée en 1817. C'est un groupe français qui s'étend à l'international et est spécialisé dans le domaine de l'assurance et de la gestion d'actifs. AXA est présent dans 64 pays et compte dans son portefeuille 107 millions de clients. Il propose différents services toujours dans le but de répondre aux besoins présentés par les particuliers et les entreprises. Ses trois domaines centraux sont l'assurance dommage, l'assurance vie, la gestion d'actifs.

AXA France est une société du groupe AXA et est le leader de l'assurance en France. Sur le territoire français, environ 33 000 personnes employés travaillent pour satisfaire le besoin de leurs clients. Elles sont au service d'environ 7,6 millions de clients. Elle réalise un chiffre d'affaires de 96,7 milliards d'Euros en 2020. Afin de mieux répondre aux besoins de la clientèle, AXA France possède plusieurs entités : les entités opérationnelles, les entités pilotage et supports, les entités sœurs.


FIGURE 1.4 – Entité d'AXA France

## 1.5 Présentation de la direction et de l'équipe

AXA IARD Particuliers a pour vocation de regrouper les activités IARD (Incendies – Accidents et Risques Divers), la Distribution, la vente au travers des différents réseaux du groupe. La Direction de l'Offre IARD définit et met en oeuvre les politiques techniques en matière de produits d'assurances Dommages (Non-Vie), tant pour le lancement des nouveaux produits que pour la maintenance du portefeuille (gestion des affaires nouvelles, des résiliations, etc.) . La direction a pour but de suivre les tendances, d'évaluer les marchés, de proposer des services adaptés au besoin d'assurance des clients. Elle oeuvre quotidiennement à répondre aux besoins de clients, à travers différentes actions :

— Gérer la tarification commerciale des produits d'assurance

— Evaluer les risques de chaque contrat de toutes les branches

— Suivre la rentabilité des contrats, la sinistralité, le risque d'anti-sélection et la performance commerciale.

— Construire la stratégie des budgets commerciaux, et coordonner les réseaux de distribution et l'Animation commerciale.

La Direction de l'Offre est repartie en plusieurs départements qui peuvent être soit spécialisés sur une branche (Automobile, ou Multi-Risques Habitation (MRH),MultiRisque Professionnel (MRP)), soit transverse ou multi-contributeurs, c'est-à-dire qui font intervenir toutes les branches. Au sein de la direction de l'offre IARD, les travaux ont été effectués principalement dans l'équipe P&C Business Analytics rattachée au service développement de l'offre et pilotage transverse. Ce service est notamment chargé du suivi du business, de la production et la gestion des projets d'analyses transverses sur la ré-

tention et la satisfaction client et de la mise en place de campagne marketing. L'équipe P&C Business Analytics a pour rôle principal :

— La production, l'analyse et la communication des Key Performance Indicators (KPI) transverses afin de suivre le business AUTO, MRH et MRP et d'alerter des évolutions atypiques et/ou inattendues. Cela se fait à partir des reportings à fréquence hebdomadaire, mensuelle et trimestrielle.

— La production d'éventuels projets de modélisation et d' études statistiques autour des indicateurs. Des études peuvent être nécessaires pour pouvoir expliquer et justifier les évolutions du business qui ressortent lors de la communication des KPIs. Aussi, des études peuvent permettre d'identifier des actions pouvant être mises en oeuvre pour optimiser les process commerciaux et tarifaires.

Ce pôle étant un pôle transverse, nous étudierons deux types de produits d'assurance : Mon Auto et Ma Maison.

## 1.6 Contexte et Objectif de l'étude

Dans cette étude, nous appellerons *cross-sell* le fait de fournir à un client à la fois une couverture d'assurance automobile et une couverture d'assurance habitation. Le taux de *cross-sell* fait l'objet d'un suivi regulier à travers des reportings.

L'analyse et le suivi de ces reporting permet à l'assureur de suivre son activité. L'un des indicateurs importants de ces reportings est le taux *cross-sell* Auto vers MRH(Multi Risques Habitation) qui correspond au fait de vendre un contrat d'assurance MRH à un client détenant déjà un contrat d'assurance Auto. Il existe également le le *cross-sell* MRH vers Auto. Le cas d'étude qui nous concerne est le *cross-sell* Auto MRH. En effet, intuitivement il est plus facile pour un individu détenant un contrat MRH de réaliser un contrat Auto. La période de réalisation du *cross-sell* Auto MRH considéré dans les reportings est de 3 à 4 mois après la date d'affaires nouvelles Auto. Le produit MRH que nous étudierons est le produit Ma maison et en ce qui concerne le produit Auto, nous travaillerons avec le produit Mon Auto. Avant d'arriver à réaliser une affaire MRH, plusieurs possibilités et étapes sont mises en avant. Le *cross-sell* passe en premier lieu par l'étape du cross-devis. Le cross-devis est la réalisation de deux devis de produits différents. Ensuite, après le cross-devis, viendra l'étape de la concrétisation. Des détails seront donnés dans un chapitre dédié.

Selon les statistiques ressorties, nous constatons que nous avons un taux de cross-devis qui est de 27% avec un taux de transformation de l'ordre de 70%. Cela signifie qu'un client après avoir fait son devis Auto réalise son devis MRH dans 27% des cas et ce devis MRH est transformé en contrat dans 75% des cas. Le taux de transformation est la proportion de devis convertis parmi tous les devis (convertis et non-convertis). Cependant parmi les affaires nouvelles Auto et non-détenteurs MRH, seulement 19%

réalisent du *cross-sell*. Nous observons donc un faible taux de *cross-sell* pour un taux de transformation élevé lorsqu'un devis MRH est réalisé. En d'autres termes, peu de personnes réalisent du *cross-sell* pour un taux de transformation élevé.

L'objectif de ce mémoire sera donc d'aider les agents à cibler les profils de personnes susceptibles de réaliser du *cross-sell*. Pour cela, dans un premier temps, nous chercherons à comprendre le *cross-sell* et son fonctionnement et ensuite déterminer les profils étant susceptibles de faire du *cross-sell* à travers des modèles. Le modèle choisi sera utilisé par les agents afin d'être alertés sur les types de personnes pouvant se multiéquiper.

## 1.7 Présentation produits

### 1.7.1 Produit Mon Auto

Mon Auto est une nouvelle offre Automobile de la branche Auto qui s'adresse aux particuliers et aux professionnels. Elle remplace les anciens produits Auto "Référence" et "C&G" qui est un produit web. Depuis son arrivée en Mars 2019, Mon Auto est l'offre auto qui est proposée en priorité pour tous nouveaux contrats. Cependant, il est souvent possible de retrouver des Demandes de Tarifs qui se font encore sur l'ancien produit Référence. En effet, il existe un périmètre encore éligible à Reference tel que les personnes morales, les voiturettes, les Véhicules très hauts de gamme ... Mon Auto comprend un socle de garanties indispensables, auxquelles des garanties et/ou des franchises peuvent être ajoutés afin de garantir une couverture optimale. Mon Auto comprend 6 formules.

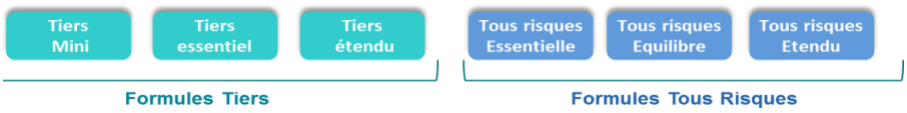



FIGURE 1.5 – Formules du produit Mon Auto

### 1.7.2 Produit Ma Maison

Tout comme le produit Mon Auto, le service MRH d'AXA France a lancé en 2017 un nouveau produit "Ma Maison" qui vient après le produit "Confort". Le produit Confort a pour vocation de disparaître. Le produit Ma Maison propose des garanties et options intéressantes qui répond davantage aux besoins du client. Ils proposent par exemple des options à la carte ce qui permet au client d'avoir le contrat le plus adapté à ses besoins. Nous avons par exemple les options piscine, une options casse des appareils nomades... Très attractifs, il permet à AXA France d'avoir une place importante sur le marché d'habitation qui est de plus en plus concurrentiel.

## 1.8 Constitution de la base de données

### 1.8.1 Description des bases

La base de données finale qui sera utilisée pour étudier le *cross-sell* est issue de la fusion de plusieurs bases. En effet, étant donné qu'il s'agit de deux produits différents, nous nous sommes appuyés sur les bases regroupant les informations devis et affaires nouvelles Mon Auto et Ma Maison.

Les informations que nous utiliserons le plus sont les informations qui se trouvent dans la base des devis "Mon Auto". Dans celle-ci se trouvent les informations sur :

— le devis :

-Le numéro de devis
-La date d'émission du devis

— le client :

-L'âge
-L'adresse
-Ancienneté du permis
-La situation maritale : marié, célibataire...
-Le type de logement : appartement, maison...
-La catégorie socio-professionnelle

— le véhicule

Nous utilisons les bases contrats afin de récupérer la date d'émission de l'affaire nouvelle qui est la date à laquelle le devis se transforme en contrat. Nous n'avons besoin de récupérer que les numéros de contrats et la date d'émission.

Avant de réaliser un contrat, le client passe par plusieurs étapes. Pour d'abord réaliser un devis, le client peut soit se rendre soit en agence, soit sur le site internet d'Axa France. L'origine d'un devis peut-être "naturelle" c'est-à-dire le premier contact et la souscription se font en agence ou "hybride" si le premier contact s'effectue dans un premier temps en ligne et la souscription en agence.

Un devis est une proposition commerciale sous forme de document contenant les informations liées à la tarification dans lequel AXA s'engage à conserver un tarif pour une période précise. L'ensemble des différentes étapes réalisées sans que le document ne soit édité représente les demandes de tarifs.

Lorsqu'un client arrive en agence pour la réalisation d'un contrat, un identifiant IDAXAPAC lui est crée ainsi qu'un projet. Par ce projet, plusieurs devis peuvent être réalisés mais évidemment un seul se concrétisera. Cependant, lorsque le projet est crée, les étapes de demandes de tarifs sont les suivantes :

— **Le brouillon ou draft** : A cette première étape, le client entre ses informations qui permettront d'établir un tarif et effectue le choix de ses garanties.

— **Tarif vu** : Le client a la possibilité de voir le tarif que lui propose l'agent.

— **Prix validé** : Le choix doit être validé par le client.

— **Summary seen** : A cette étape, les informations rentrées sont résumées avant d'accepter finalement les choix.

— **Contractualised unsigned** : C'est l'étape finale qui correspond à la signature avant que le devis ne contractualise.

Dans les bases devis Mon Auto comme Ma Maison, toutes ces étapes sont mentionnées. La variable mentionnant ces étapes se nomme "Pol_Statutdevis". Pour réaliser notre étude de *cross-sell*, nous nous intéresserons à la dernière étape du devis portant la modalité "QUOTATION_CONTRACTUALIZED_UNSIGNED" de la variable donnant le statut du devis. Généralement, c'est le dernier devis qui contractualise.

Un client peut avoir plusieurs numéros de contrats notés "Numcnt" dans les bases. Il est important de noter qu'un client chez AXA représente un foyer c'est-à-dire un ensemble de personnes vivant à la même adresse ou sous un même toit. Un individu peut avoir un statut souscripteur, assuré... La variable utilisée pour désigner le numéro client est nommé "Nmcli". Dans notre étude, nous nous intéresserons à la maille client. Ainsi, dans la section suivante, nous décrirons étape par étape la construction de la base finale.

### 1.8.2 Etapes de construction de la base finale

La construction de notre base finale a nécessité plusieurs étapes. Un schéma simplifié expliquant la méthode de construction est présenté ci-dessous :

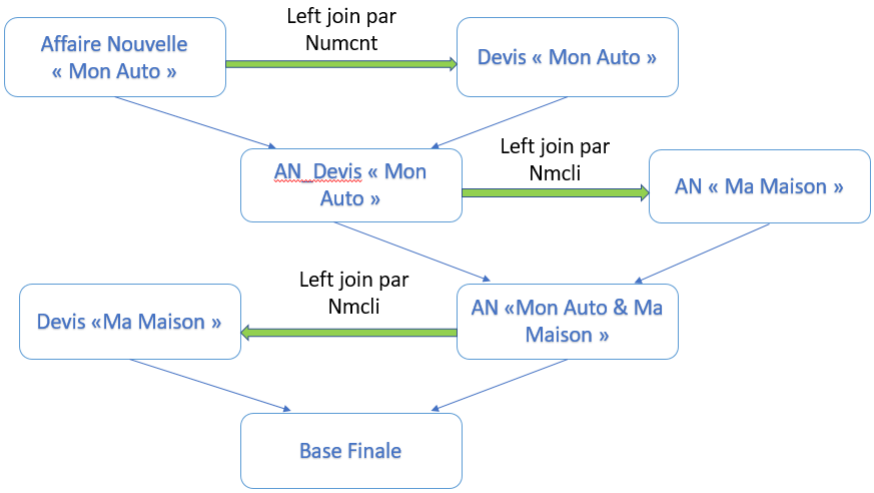



FIGURE 1.6 – Schéma simplifié expliquant la méthode de construction de la base de données

Premièrement, nous partons des affaires nouvelles(AN) Auto car nous ne considérons que le *cross-sell* des affaires nouvelles "Mon Auto" vers les affaires nouvelles "Ma Maison". Dans cette table, les numéros de contrats sont déjà dédoublonnés c'est-à-dire que nous avons une ligne pour un numéro de contrat. Par contre, il est possible de trouver sur plusieurs lignes le même numéro de client. Cela s'explique par le fait que le client (représentant un foyer comme décrit dans la sous-section précédente) ait plusieurs personnes dans son foyer et à chaque personne correspond un numéro de contrat. Afin de récupérer la date de devis et les différentes informations concernant le devis et le client, nous effectuons une jointure par numéro de contrat. Il existe des cas où des affaires nouvelles se retrouvent sans devis. Cela peut s'expliquer par le fait que l'agent va au bout du processus sans générer de document. Nous en avons peu dans ce cas soit 0,03%.

Une fois les affaires nouvelles et les devis Auto réunis, nous cherchons ensuite à récupérer la date d'émission des affaires nouvelles MRH. Etant à la maille client, nous arrivons donc à faire cette jointure entre notre première base contenant les AN et les devis Auto et la base contrat MRH. Cependant, un problème se pose car un client peut également avoir plusieurs contrats habitation. Pour résoudre ce problème, nous choisissons de ne garder qu'un contrat MRH par numéro client en privilégiant le contrat MRH qui vient après l'AN Mon Auto (car nous sommes dans le cas Auto vers MRH) et le plus récent par rapport à la date de contrat Auto. Cette date nous permet à cet instant de déterminer si un client était détenteur MRH ou pas au moment de la date de devis Auto.

Enfin, nous cherchons à avoir la date d'émission du devis MRH en procédant également par jointure avec pour clef le numéro client. Si un client réalise plusieurs devis MRH alors nous privilégerons celui qui vient après le devis Mon Auto et le plus proche.

Afin d'avoir assez d'arguments pour expliquer le phénomène du *cross-sell*, nous cherchons à rajouter différentes informations à notre base de données pour l'enrichir.

### 1.8.3 Enrichissement d'information

En plus de la base obtenue, nous collectons d'autres informations supplémentaires qui permettront de mieux expliquer le *cross-sell*. De ce fait, une variable représentant une segmentation client faite par AXA est ajoutée. Elle distingue les clients selon les caractéristiques suivantes :

-Les personnes de 18 à 20 ans
-Les personnes de 21 à 30 ans non sinistrées
-Les Bonus 50
-Les sinistrés : les personnes ayant eu un unique sinistre dans les 36 derniers mois
-Les multi-sinistrés : les personnes ayant eu plus d'un sinsitre dans les 36 mois
-Les professionnels
-Les creusements de bonus : parmi les clients, certains n'ont pas eu le temps de devenir Bonus 50, sans qu'ils soient mauvais pour autant. Cette classe permet de capter les bons conducteurs âgés de plus de 30 ans et n'ayant pas suffisamment d'ancienneté de permis pour devenir Bonus50.
-Les Autres : Toutes les personnes ne se trouvant pas dans les classes précédentes.

Pour mieux synthétiser ces caractéristiques et les différencier, nous ressortons le schéma suivant :

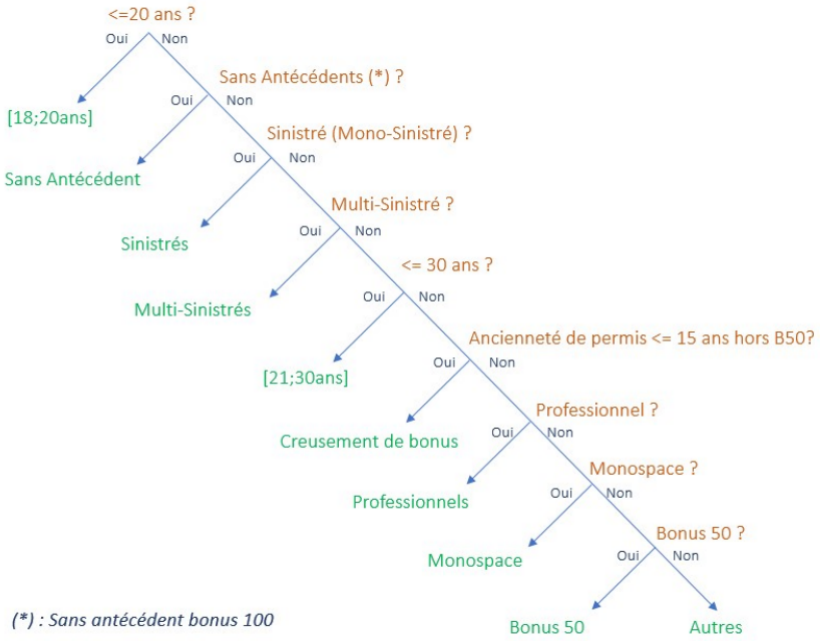



FIGURE 1.7 – Segmentation client

En vue d'obtenir plus d'informations expliquant le *cross-sell*, nous enrichissons notre base avec des données concurrence donnant le nombre de concurrents dans un rayon de 500 mètres.

Nous rajoutons également des données distributeurs. Celles-ci concernent les agences actives ou non en santé, prévoyance, MRP. Aussi, les commissions des agents sont rajoutées. Elles sont réparties en déciles. Le premier décile est le niveau qui sépare d'un côté les 10% d'agences les moins bien commissionnés et de l'autre, les 90% les mieux commissionnés. Le deuxième décile est le niveau de commission pour lequel 20% touchent moins et 80% touchent plus.

Une variable donnant la détention du client est aussi construite. Il est question de savoir si le client détient un contrat autre que la MRH, l'auto ... Pour ce faire, les catégories suivantes ont été construites :

— Aucun produit : il s'agit des prospects
— Mono auto : ce sont les clients détenant un seul contrat Auto
— Mono autres : Ces clients détiennent un seul contrat autre que l'auto et la MRH
— Multi auto : Il s'agit des personnes détenant plusieurs contrats auto
— Multi autres : Ces personnes détiennent plusieurs contrats autre que l'auto et la MRH.

### 1.8.4 Périmètre d'étude

Dans ce mémoire, nous nous sommes focalisés sur les nouveaux produits Mon Auto et Ma Maison de l'auto et la MRH. En ce qui concerne la base des affaires nouvelles Mon Auto, nous nous intéressons à la période du 01 Janvier 2021 au 30 novembre 2021. Pour choisir la période de la base devis Mon Auto, nous nous basons sur le fait que lorsqu'un devis est réalisé, la plupart souscrit en moins de 4 mois. Nous illustrons cela à travers le graphe suivant :

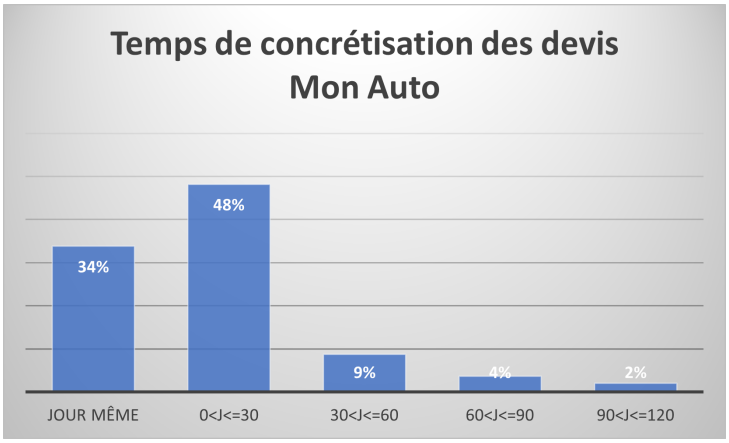



FIGURE 1.8 – Répartition des affaires nouvelles Mon Auto après leur devis

Nous constatons sur ce graphique que plus de 90% des clients concrétisent leur devis Mon Auto dans les 4 mois. Concernant donc la période des devis Mon Auto, nous reculons 4 mois en arrière par rapport à la date d'affaires nouvelles choisie. Afin de limiter le nombre d'affaires nouvelles sans devis, nous prenons la période du 01 Juin 2020 à fin novembre 2021.

Ensuite, la période d'affaires nouvelles MRH concernera la période de Janvier 2020 à fin novembre 2021 car nous chercherons à savoir si un individu disposait un contrat MRH avant sa souscription Auto.

Nous utiliserons la règle des 4 mois pour choisir la date des devis MRH. Nous optons pour la période de Juin 2019 à Novembre 2021.

# Chapitre 2

# Compréhension du *cross-sell* Auto MRH

## 2.1 Généralités sur le *cross-sell*

Le *cross-sell* ou multi-équipement est une technique de vente consistant à proposer à un client un produit différent en plus du produit qu'il détient. Le multi-équipement permettra aux entreprises d'augmenter le chiffre d'affaires mais aussi de maintenir le client dans son portefeuille.

Tout d'abord, du côté de l'assuré, le *cross-sell* permettra de se couvrir d'un risque supplémentaire non couvert et peut lui permettre sans doute de bénéficier une réduction. En cas d'excès de tarification, le client détenant déjà d'autres produits sera avantagé dans la mesure où ce dernier pourrait bénéficier d'options intéressantes.

Du côté de l'assureur, la multi-détention est un facteur clé pour augmenter son chiffre d'affaires. En multi équipant l'assuré, le total des primes versées par l'assuré augmente. Cela est moins onéreux pour un assureur de prendre en charge le même assuré avec de nouveaux produits que d'acquérir de nouveaux assurés.

Par ailleurs, le multi-équipement est un facteur clé pour fidéliser ses clients. C'est cette fidélité du client dans le portefeuille qui permettra à l'assureur d'augmenter son chiffre d'affaires. Par cette technique marketing, l'assureur aura la possibilité de mieux connaître ses assurés.

Pour un assureur, le multi-équipement étant important, il est important de chercher à développer cette pratique. Nous chercherons donc à aider l'agent à multi-équiper en déterminant les caractéristiques des clients qui seraient susceptible de souscrire plusieurs contrats.

Dans ce mémoire, nous traiterons le *cross-sell* sur deux produits spécifiques que sont les produits Auto et MRH. Nous parlerons ainsi de *cross-sell* Auto MRH. Dans cette expression, nous prenons compte du sens c'est-à-dire qu'on parlera de *cross-sell* auto

MRH lorsqu'un client souscrira à un contrat Auto dans un premier temps puis à un contrat MRH.

## 2.2 Période de réalisation du *cross-sell* et du cross-devis

Comme signifié plus haut, le cas de *cross-sell* nous concernant est le *cross-sell* Auto vers MRH. Dans cette partie, il s'agira de montrer à quel moment une personne après avoir réalisé son affaire nouvelle Mon Auto réalisera une affaire nouvelle Ma Maison. Nous nous intéresserons particulièrement aux personnes ne détenant pas d'affaires nouvelles MRH au moment du devis Mon Auto. On parlera donc de "Non-détenteurs MRH".

Dans notre base finale, nous avons la répartition suivante :

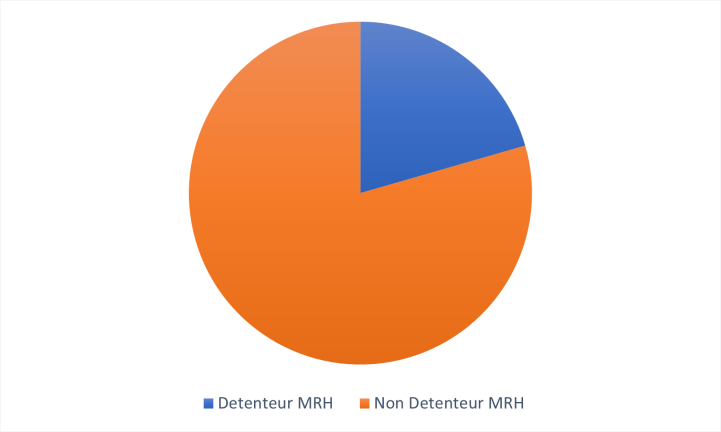



FIGURE 2.1 – Répartition selon la détention MRH

79% de non-détenteurs MRH contre 21% de détenteurs sont observés. Un regard particulier est porté aux non-détenteurs car nous supposons que si l'individu possède déjà un contrat Ma Maison, il sera moins enclin à réaliser une autre affaire nouvelle Ma Maison.

Dès lors, avec cette population, nous observons les différentes actions faites autour de l'affaire nouvelle Mon Auto, avant et après. Les analyses qui seront réalisées aideront à comprendre le *cross-sell* auto vers MRH. Nous parlons de cross-devis lorsqu'un devis Ma Maison est réalisé après un devis Mon Auto ou après une affaire nouvelle Mon Auto.

Nous observons dans un premier temps le comportement des devis Ma Maison autour des devis Mon Auto.

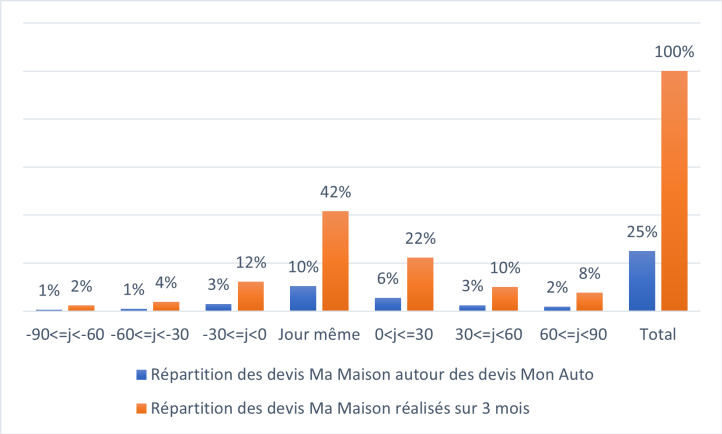



FIGURE 2.2 – Répartition des devis Ma Maison autour des devis Mon Auto selon le moment de réalisation

25% des clients non-détenteurs MRH détenant un contrat Mon Auto non-détenteurs MRH réalisent leur devis Ma Maison 3 mois autour de leur devis Mon Auto.

La répartition des devis Ma Maison autour de l'affaire nouvelle Mon Auto est aussi observée :

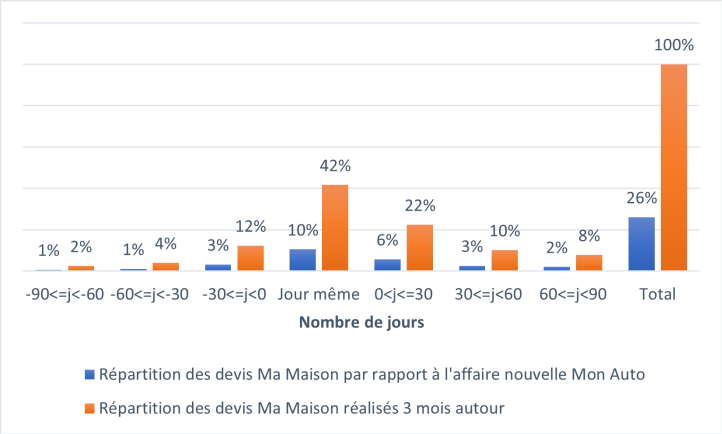



FIGURE 2.3 – Répartition du devis Ma Maison autour de l'affaire nouvelle Mon Auto selon le moment de réalisation

De ce graphique ressort que 26% des clients non-détenteurs MRH détenant un contrat Mon Auto réalisent leur devis Ma Maison 3 mois autour (c'est-à-dire avant et après) de leur AN Mon Auto. Parmi ces 26%, 42% le réalisent le jour même de l'affaire nouvelle Mon Auto.

La période du cross-devis est alors restreinte sur 3 mois. Aisni, tous les devis intervenant après 4 mois ne seront plus considérés comme des cross-devis.

En ce qui concerne le délai entre une affaire nouvelle Ma Maison et une affaire nouvelle Mon Auto, la répartition suivante est observée :

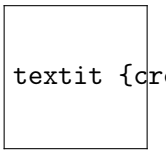



FIGURE 2.4 – Répartition de l'affaire nouvelle Ma Maison autour de l'affaire nouvelle Mon Auto selon le moment de réalisation

De ce graphique, il ressort que 19% des clients non détenteurs MRH détenant un contrat Mon Auto réalisent leur affaire nouvelle Ma Maison 3 mois après et parmi eux 54% le font le jour même de l'affaire nouvelle Mon Auto.

En ayant remarqué que les affaires nouvelles Ma Maison se réalisent majoritairement dans les 3 mois, nous disons qu'il y a *cross-sell* Auto MRH lorsqu'une personne souscrit un contrat Ma Maison dans les 3 mois après son contrat Mon Auto.

Il est aussi important d'observer le temps de réalisation de l'affaire nouvelle Ma Maison lorsqu'un client réalise son cross-devis après l'affaire nouvelle Mon Auto.

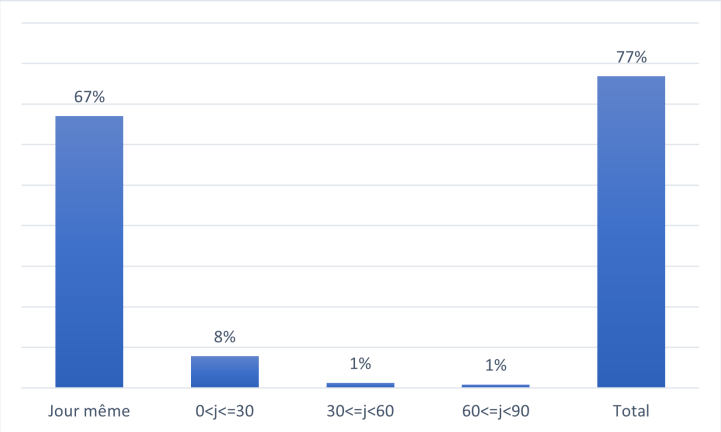



FIGURE 2.5 – Répartition du taux de transformation Ma Maison 3 mois après la réalisation d'un devis Ma Maison et d'une affaire nouvelle Mon Auto

Après que l'affaire nouvelle Ma Maison soit réalisée après le devis Ma Maison, nous avons un taux de transformation de 77% dans les 3 mois dont 67% le même jour.

Un schéma récapitulant le cross-devis, le *cross-sell*, le taux de transformation est présenté ci-dessous :

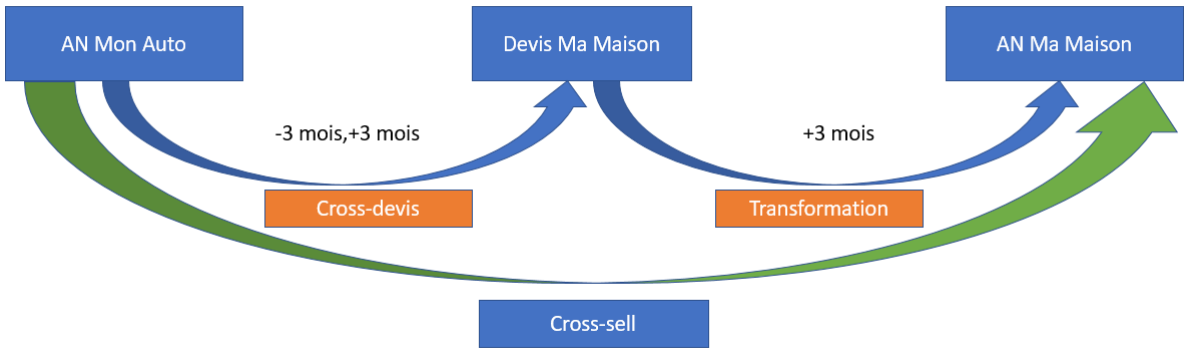



FIGURE 2.6 – Schéma récapitulant le *cross-sell*, le cross-devis et le taux de transformation

## 2.3 Les différents types de *cross-sell*

Toujours dans l'optique de comprendre le *cross-sell*, nous cherchons à comprendre l'ordre des différentes étapes pour arriver au *cross-sell*.

En effet, toutes les actions peuvent être menées en même temps. Autrement dit, l'individu peut réaliser le même jour les 4 actions suivantes : le devis Auto, l'affaire nouvelle Mon Auto, le devis Ma Maison, l'affaire nouvelle. On l'appellera "*cross-sell* de type 1".

Ainsi, plusieurs possibilités peuvent ressortir mais nous mentionnerons les plus majoritaires dans le tableau ci-dessous :

| Devis Mon Auto | Affaire Nouvelle Mon Auto | Devis Ma Maison | Affaire Nouvelle Ma Maison | Répartition |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 41% |
| 1 | 1 | 1 | 2 | 3% |
| 1 | 1 | 2 | 2 | 36% |
| 1 | 1 | 2 | 3 | 7% |

TABLE 2.1 – Ordre de réalisation des étapes du *cross-sell*

Les chiffres représentés dans ce tableau représente l'ordre de réalisation des actions permettant d'arriver au *cross-sell*.

*cross-sell* de type 1 :

Nous dirons que nous aurons affaire à un *cross-sell* de type 1 lorsque toutes les actions se feront le même jour.

Dans notre base, 41% réalisent le *cross-sell* en un temps.

*cross-sell* de type 2 :

Le *cross-sell* de type 2 se définit simplement par deux étapes. Ainsi, 47% des affaires nouvelles Mon Auto non-détenteurs MRH réalisent leur *cross-sell* en deux temps. Ils se répartissent comme suit :

36% réalisent leur devis Auto et leur affaire nouvelle Auto le même jour et ensuite le devis MRH et l'affaire nouvelle MRH le même jour. Il serait intéressant de regarder le temps de réalisation entre ces deux phases.

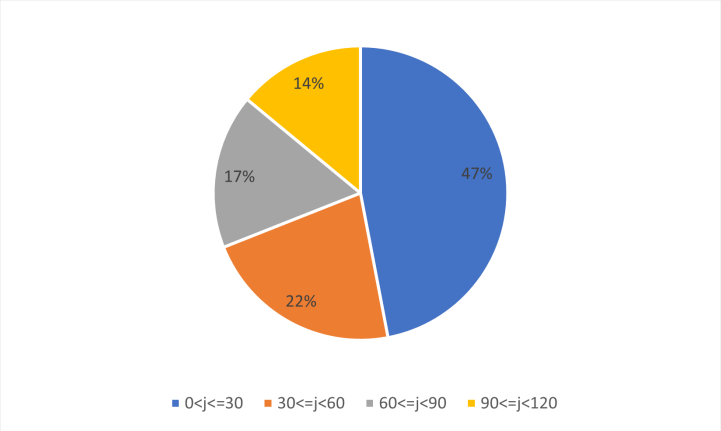



FIGURE 2.7 – Temps de réalisation

Plus de la moitié du *cross-sell* de type 2 se réalisent majoritairement dans les deux mois.

3% de notre population réalisent le devis Auto, l'affaire nouvelle Auto, le devis Ma Maison le même jour et ensuite l'affaire nouvelle Ma Maison.

Il peut arriver que le devis Ma Maison se fasse en premier et ensuite les autres actions le même jour. Nous avons 7% dans ces cas. Nous cherchons ensuite le temps de réalisation entre ces deux phases :

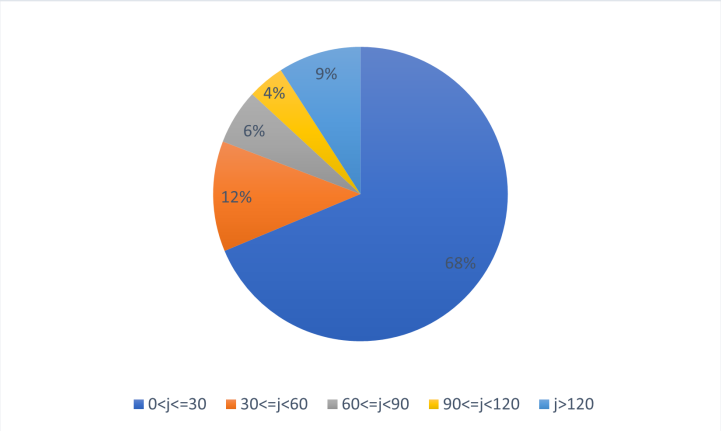


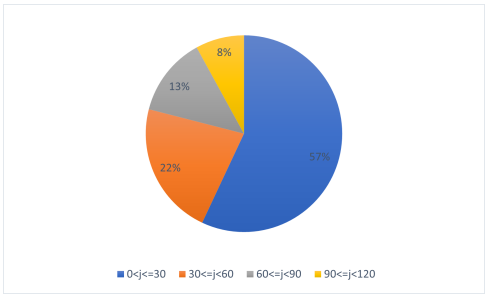
FIGURE 2.8 – Temps de réalisation du *cross-sell* après un devis Ma Maison

Le délai entre ces deux phases se réalisent principalement dans les deux mois.

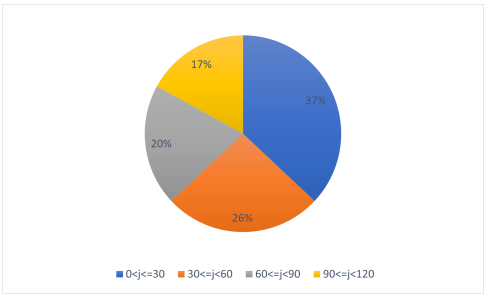
*cross-sell* de type 3

Ce type de *cross-sell* est ainsi défini comme étant un *cross-sell* qui se définirait en 3 étapes. Ces cas sont un peu rares. Nous n'en avons que 7% dans notre base.

Nous en avons certains qui réalisent leur devis Auto et affaire nouvelle le même jour, ensuite leur devis MRH et en dernière position leur affaire nouvelle. 7% de la population se trouvent dans ce cas. Nous ressortons le délai entre les différentes actions.




(a) Délai entre la première et la deuxième action

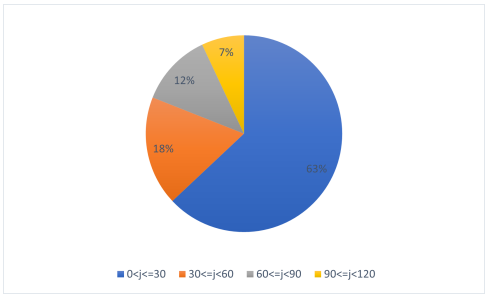



(b) Délai entre la deuxième et la troisième action

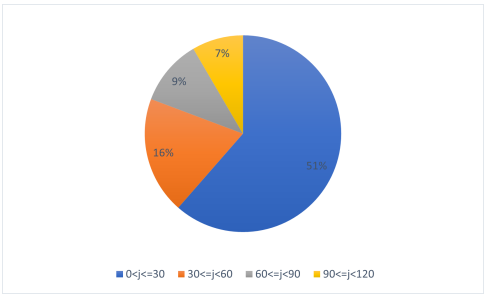
FIGURE 2.9 – Temps de réalisation des différentes actions

Pour ce qui concerne ce cas de *cross-sell*, le devis Ma Maison se réalise beaucoup plus rapidement dans les 30 jours mais prend assez de temps avant de concrétiser.

Observons le cas où nous avons le devis en première action, ensuite le devis et l'affaire nouvelle le même jour puis enfin l'affaire nouvelle Ma Maison.




(a) Délai entre la première et la deuxième action




(b) Délai entre la deuxième et la troisième action

FIGURE 2.10 – Temps de réalisation des différentes actions

Lorsque le devis Ma Maison est réalisé en première action, dans les deux mois qui

suivent 81% de la population réalisent leur devis et affaire nouvelle Mon Auto et se concrétisent un peu plus rapidement en affaire nouvelle MRH.

Au vue de ces différentes analyses, nous pouvons constater que le *cross-sell* en général ne prend pas assez de temps à se réaliser. Cette information sera sûrement utile à l'agent. Afin d'avoir une première détection des profils, nous réaliserons des statistiques descriptives.

# Chapitre 3

# Profil et tendances favorisant le cross sell

Avant de réaliser nos modèles, l'analyse descriptive est une étape importante pour avoir une première idée des résultats que pourront produire les modèles. Il est important également de vérifier les liens entre les variables en étudiant leur corrélation.

## 3.1 Analyse des corrélations

Il s'agira dans cette partie de réaliser une première sélection de variables afin d'obtenir des variables de faible corrélation dans notre modèle. Les corrélations seront calculées en fonction de la nature des variables dont nous disposons. Pour les variables qualitatives, nous utiliserons le V de cramer et la formule de Pearson pour les variables quantitatives.

Le V de Cramer

Le V de Cramer est défini de la façon suivante :

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}}$$

Avec :

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Cette métrique est appelée la distance de Khi2. Nous définissons également la valeur maximale du khi2.

$$\chi^2_max = nmin(I-1, J-1)$$

Avec :

— n le nombre total d'observations
— I le nombre de lignes
— J le nombre de colonnes

Le V de cramer prend ses valeurs entre 0 et 1. Plus il se rapproche de 1, plus les variables sont corrélées.

La méthode Pearson

Cette méthode consiste à analyser la corrélation entre deux variables quantitatives et prend ses valeurs entre -1 et 1. L'expression de son coefficient est la suivante :

$$r(X,Y) = \frac{cov(X,Y)}{\delta_X \delta_Y}$$

Avec :

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

Avec $\delta_X$, $\delta_Y$ les écarts-types respectifs des variables X et Y ; $\bar{X}$, $\bar{Y}$ les moyennes respectives des variables X et Y ; N le nombre d'observations ; $X_i$ et $Y_i$ les observations respectives i des variables X et Y.

## 3.2 Analyses univariées

A partir de la base précédemment construite, nous ressortirons une analyse univariée du taux de *cross-sell*. Il s'agira de donner la répartition du taux de *cross-sell* en fonction des caractéristiques des clients Auto. Cela nous permettra dans un premier temps d'avoir une tendance de profils favorisant le *cross-sell*.

Tout d'abord, les variables quantitatives ont été transformées en variables catégorielles au travers des arbres de décision. La théorie de ces arbres est expliquée un plus bas dans ce mémoire.

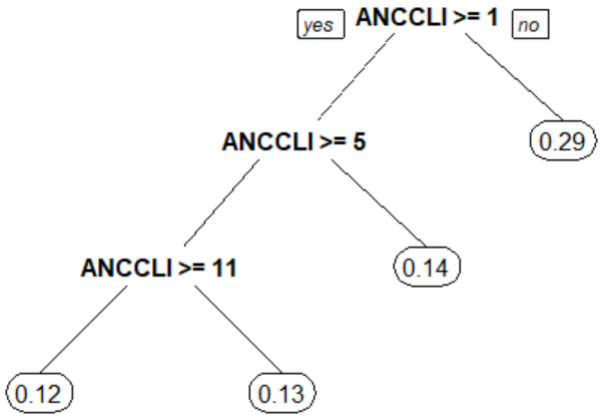



FIGURE 3.1 – Discrétisation de la variable Ancienneté

Cet arbre nous a permis, pour la variable ancienneté permis d'avoir les classes suivantes : [0,1[,[1,5[,[5,11[,[11,+∞[. Nous procédons de la même manière en discrétisant les variables durée de circulation, la prime Auto, le nombre de véhicule.

Ancienneté client

Nous commençons par regarder la répartition de l'ancienneté du client.

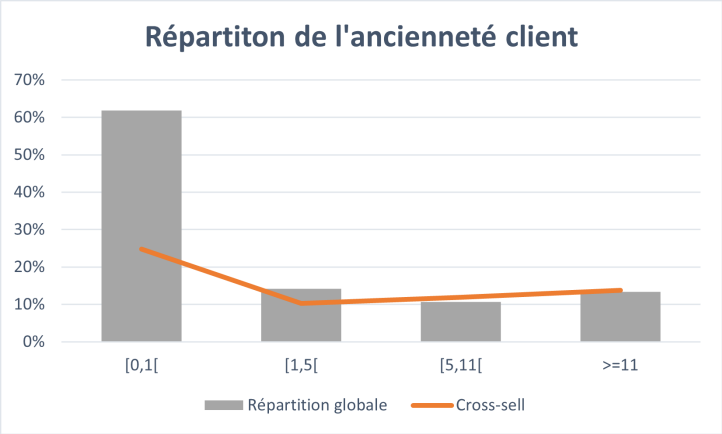



FIGURE 3.2 – Ancienneté client dans le portefeuille

De ce graphique ressort que les clients possédant une faible ancienneté dans le portefeuille sont beaucoup plus susceptibles de faire du *cross-sell*.

Ancienneté permis

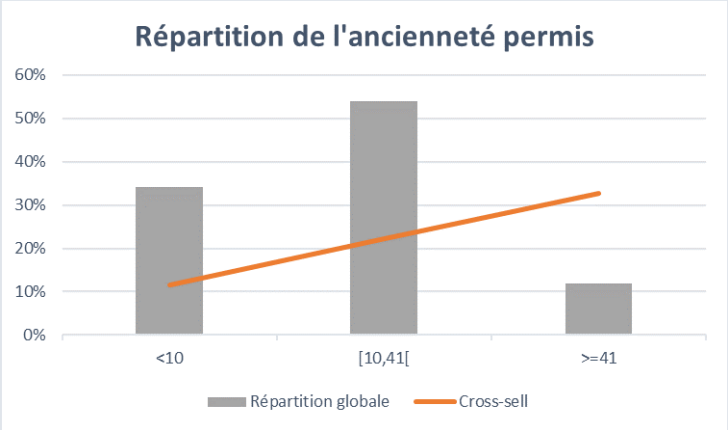



FIGURE 3.3 – Répartition de l'ancienneté permis

Les anciens conducteurs semblent plus enclins à réaliser du *cross-sell*.

Type client

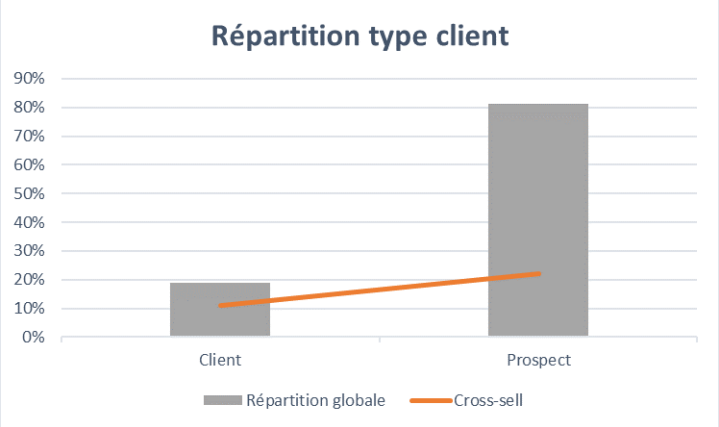



FIGURE 3.4 – Taux de *cross-sell* par typologie client

Les prospects ont plus tendance à réaliser du *cross-sell* que les clients.Nous observons une pente un peu élevée ce qui signifierait que la variable type client pourrait être une variable impactante pour notre étude.

Région

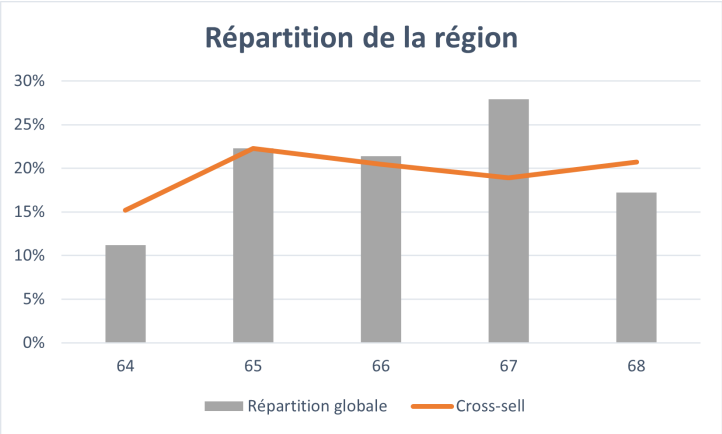



FIGURE 3.5 – Taux de *cross-sell* par Région

Nous constatons que les personnes vivant en régions 65 et 68 représentant respectivement le Nord-Est et le Sud-Ouest réalisent plus de *cross-sell* par rapport aux autres régions.

La segmentation client

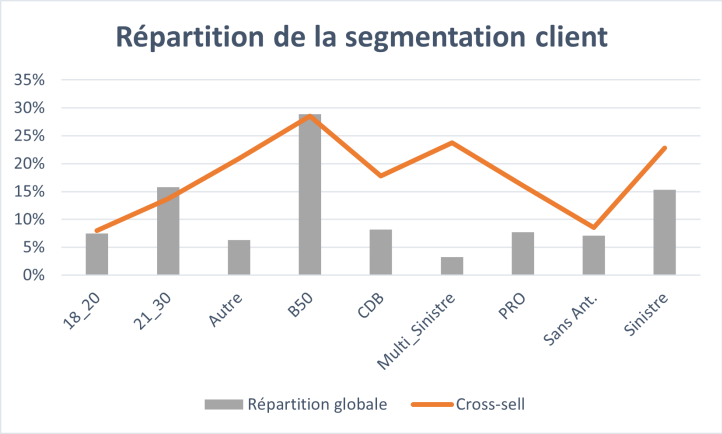



FIGURE 3.6 – Taux de *cross-sell* par la segmentation client

En ce qui concerne la segmentation client, nous observons une forte présence de Bonus 50, ensuite les sinistrés qui présentent un taux de *cross-sell* par rapport aux autres. Cette variable aura sûrement de la significativité dans notre modèle.

La multi-détention

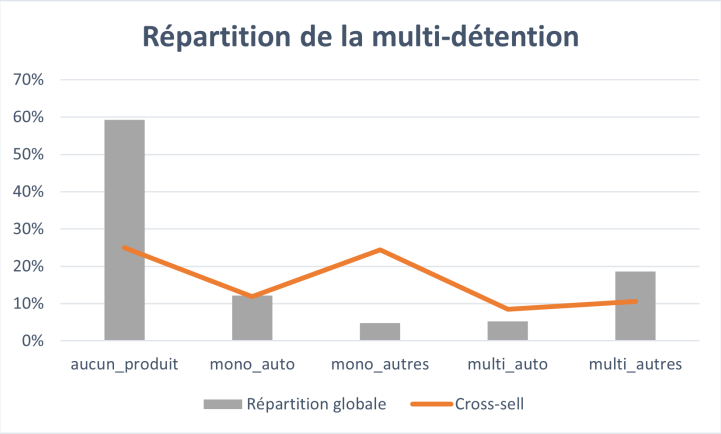



FIGURE 3.7 – Taux de *cross-sell* par multi-détention

La multi-détention est une variable créée donnant le type de contrat detenu par chaque personne. Les autres sont les types de contrats hors Auto et MRH.

Sur ce graphique, il en ressort que les personnes ne détenant aucun type de contrat et ceux détenant d'autres produits que l'auto et la MRH réalisent beaucoup plus de *cross-sell*.

Les formules Auto

La gamme Mon Auto contient 6 formules avec des notations particulières :
— F1 : Tiers Mini
— F2 : Tiers essentielle
— F3 : Tiers étendue
— F4 : Tous risques essentielles
— F5 : Tous risques équilibre
— F6 : Tous risques étendue

Les formules augmentent avec leur niveau de garantie en partant de 1 à 6. Nous observons la répartition suivante :

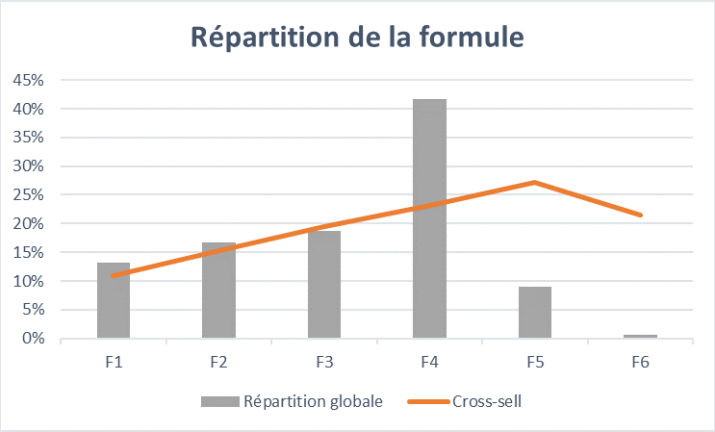



FIGURE 3.8 – Taux de *cross-sell* par Formule Auto

La formule auto pourrait être significative pour le *cross-sell*. Les personnes souscrivant à la formule F5 sont plus enclins à faire du *cross-sell*.

Commissions

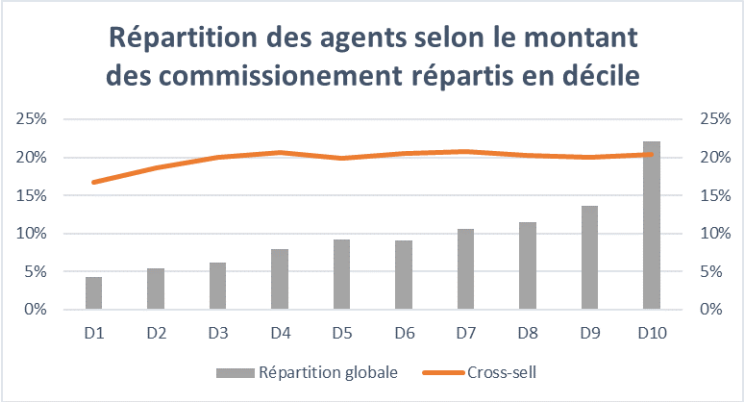



FIGURE 3.9 – Taux de *cross-sell* par Niveau de commission

Les commissions des agents regroupées en quantiles ont une tendance presque linéaire. Cette variable semble ne pas avoir un impact significatif sur le taux de *cross-sell*.

# Chapitre 4

# Aspect théorique de la Modélisation

Les méthodes de modélisation que nous utiliserons dans ce mémoire sont des méthodes d'apprentissage. Il en existe deux types que sont l'apprentissage supervisé et non-supervisé. La différence entre ces deux méthodes d'apprentissage se trouve au niveau de la connaissance de la variable à prédire. Pour ce qui est de l'apprentissage supervisé, dans le cas d'une classification, connaissant la classe des individus, le modèle sera plus habilité à déterminer les caractéristiques associées afin de les réutiliser sur d'autres données. Nous pouvons citer en exemple les glm, les algorithmes de machine learning...

En revanche, en ce qui concerne l'apprentissage non-supervisé, les classes d'individus ne sont pas connus. L'algorithme essaiera donc de regrouper ces classes dans des groupes homogènes ce qui permettra de mieux appréhender la base de données. Comme exemples de méthode d'apprentissage non supervisé, nous avons la méthode des K-means, la classification ascendante hiérarchique, l'analyse en composante principale (ACP)...

## 4.1 Méthode *boosting*

Le *boosting*

Le but du *boosting* est de créer des algorithmes puissants et efficaces en partant d'algorithmes présentant de moindres performances. Les algorithmes présentant les performances faibles sont appelés *weak learners* et ceux présentant des performances élevées les *strong learner*. Les *boostings* permettent de passer des *weak learners* aux *strong learners*. Sur la base d'apprentissage, la récurrence des arbres est définie. A chaque itération, une importance est donnée aux individus mal prédits. De ce fait à l'itération k, une grande attention est apportée aux individus mal ajustés de l'étape précédente. Une augmentation de leur poids est alors observée.

Le modèle final est reconnu par une pondération d'arbres dont les classes auront un poids plus élevées.

Le gradient boosting

Pour expliquer l'algorithme du gradient boosting, nous utiliserons les notations suivantes : y la variable réponse et x le vecteur aléatoire. Considérons que l'échantillon d'apprentissage comprend N individus. L'objectif est de pouvoir déterminer une approximation $\hat{F}$ de la fonction $F^*$ reliant ces deux types de variables. Il est pourtant possible de définir $F^*$ comme étant la fonction qui minimise l'espérance de la fonction de perte $L(.,.)$ :

$$F^* = E_{yx}[L(y, F(x)] = E_x[E_y[L(y, F(x)|y]|x] \tag{4.1}$$

Il est possible d'avoir différentes formes de fonction de perte selon le problème qui se pose :

— $L(y, F) = (y - F)^2$ pour la régression
— $L(y, F) = |y - F|$, pour la régression
— $L(y, F) = log(1 + e^{-2yF})$, pour la classification

Pour résoudre l'équation précédente, nous considérons l'expression de $F^*$ comme suit :

$$F^*(x) = \sum_{m=0}^{M} f_m(x)$$

Avec :

— $f_0(x)$ une hypothèse initiale
— $f_m(x)$ incréments obtenus par itération par la descente du gradient

Ces derniers sont définis par l'expression ci-dessous :

$$f_m(x) = -\rho_m g_m(x)$$

— $g_m(x) = -[\frac{\partial E_y[L(y,F(x))|x]}{\partial F(x)}]_{F(x)=F_{m-1}(x)}$
— $F_{m-1}(x) = \sum_{i=0}^{m-1} f_i(x)$
— $\rho_m = \text{argmin}_\rho \, E_{yx}[L(y, F_{m-1}(x) - \rho g_m)]$

$-g_m$ est la direction du gradient et $\rho_m$ la règle de recherche. Il existe une expression commune décrivant F comme modèle additif :

$$F(x; (\alpha_m, b_m)_1^M) = \sum_{m=1}^{M} \alpha_m h(x, b_m)$$

Avec h une fonction paramétrique avec entrée x et des paramètres b=$b_1, ..., b_M$ qui représente un classificateur faible.

Afin de minimiser cette fonction, il en ressort l'équation suivante :

$$(\alpha_m, b_m) = \text{argmin}_\alpha, b \sum_{i=1}^{N} L(y_i, \sum_{m=1}^{M} \alpha'_m h(x_i, b'_m)$$

La fonction qui relie x et y peut avoir l'estimation suivante :

$$F_m(x) = F_{m-1}(x) + \rho_m h(x, b_m)$$

Si la variable réponse y est dichotomique alors la fonction de perte a la forme suivante :

$$L(y, F) = log(1 + e^{-2yF})$$

Où

$$F(X) = \frac{1}{2} log[\frac{P(y = 0|x)}{P(y = 1|x)}]$$

On a ainsi :

$$\tilde{y_i} = -(\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)})_{F(x)=F_{m-1}(x)} = \frac{2y_i}{(1 + exp(2y_i F_{m-1(x_i)})}$$

$$\rho_m = \text{argmin}_\rho \sum_{i=1}^{N} log(1 + exp(-2y_i(F_{m-1}(x_i) + \rho h(x_i, b_m))))$$

XGBOOST

Le terme "XGBOOST" est un sigle du groupe de mot "Extreme Gradient Boosting". C'est une méthode qui part au-delà de la descente du gradient. En plus de la dérivée du gradient, cet algorithme propose une dérivée supplémentaire pour en obtenir le minimum ce qui permet encore plus de diminuer le taux d'erreur.

## 4.2 Les modèles linéaires généralisés

### 4.2.1 Les modèles linéaires

Les modèles linéaires sont généralement utilisés pour expliquer une variable Y en fonction d'une combinaison linéaire de variables explicatives $X_1, X_2, ..., X_{n-1}, X_n$ et d'un terme aléatoire $\epsilon$ suivant une loi normale $\mathcal{N}(0, \sigma^2)$. Il est souvent mis sous forme matricielle.

Ce modèle reposant sur la normalité de la variable Y, il est possible de le généraliser avec les modèles linéaires généralisés.

### 4.2.2 Les modèles linéaires généralisés

Concernant les modèles linéaires généralisés, la variable réponse doit être de la famille exponentielle ayant pour fonction de densité :

$$f_{\theta,\phi}(y) = exp(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)) \tag{4.2}$$

Avec :

— $\theta$ : le paramètre
— $\phi$ : le paramètre de dispersion
— Les fonctions a,b et c sont fixées

L'une des composantes importantes d'un modèle linéaire généralisé est la fonction lien. Elle permet de modéliser l'espérance de la variable Y. Selon la loi étudiée, il existe une fonction lien particulière qui se doit d'être inversible. Nous donnons quelques exemples dans le tableau suivant :

| Loi | Fonction lien |
|---|---|
| Binomiale | $ln(\frac{\mu}{1-\mu})$ |
| Normale | $\mu$ |
| Poisson | $ln(\mu)$ |
| Gamma | $\frac{1}{\mu}$ |

TABLE 4.1 – Fonction lien

### 4.2.3 La régression logistique

La régression logistique est un cas particulier du modèle linéaire généralisé. Dans notre cas, notre variable réponse est la variable "*cross-sell*" qui prend deux modalités soit 0 ou 1.

Théorie du modèle

Le modèle de régression logistique modélise la loi de $Y|X = x$ par une loi de Bernouilli de paramètre $P(Y = 1|X = x)$. La loi de Bernouilli étant définie comme suit :

$$P(Y = 1) = p; P(Y = 0) = 1 - p$$

Si nous considérons les $Y_i$ comme la variable affaire nouvelle Mon Auto suivant la loi de Bernoulli qui prend 1 s'il fait du *cross-sell* ou 0 sinon, nous chercherons donc à modéliser la somme des $Y_i$ qui représentera donc une loi binomiale.

La densité de la loi binomiale se présente comme suit :

$$f(k) = C_n^k p^k (1-p)^{n-k} \tag{4.3}$$

Avec :

-k=0,1,...,n
-n le nombre d'observations
-p la probabilité de réaliser du *cross-sell* donc $P(Y_i) = 1$
L'expression sous la forme de la densité d'une famille exponetielle donne la formule suivante :

$$f(k) = exp(klog(\frac{p}{1-p}) + nlog(1-p) + log(C_n^k) \tag{4.4}$$

De cette expression, nous identifions les paramètres de la densité d'une famille exponentielle :
— $\theta$=g(p)=log($\frac{1}{1-p}$) avec g la fonction lien
— $\phi$=1
— b($\theta$)=1
— $c(k, \phi) = log(C_n^k) + nlog(1-p)$

En ce qui concerne le modèle logistique, nous ressortons alors l'expression de la fonction lien logit de la loi binomiale que nous avons noté dans le tableau donné à la section précédente :

$$logit[P(Y = 1|X = x)] = ln[\frac{P(Y = 1|X = x)}{P(Y = 0|X = x)}] = X'\beta$$

La fonction logit a la forme suivante :

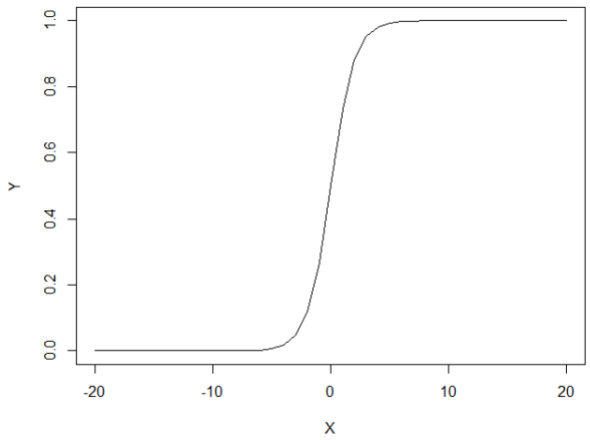



FIGURE 4.1 – Fonction Logit

Par conséquent :

$$P(Y = 1|X = x) = 1 - P(Y = 0|X = x) = \frac{1}{1 + exp(-X\beta)}$$

X étant le vecteur représentant les variables et $\beta$ les coefficients associés. A partir des

propriétés de l'espérance, on a :

$$E[Y|X = x] = P(Y = 1|X = x).1 + P(Y = 0|X = x).0 = P(Y = 1|X = x)$$

Estimation des coefficients du modèle

Les coefficients du modèle par maximum de vraisemblance. Pour cela, nous utilserons la méthode de dérivation de la log-vraisemblance. Nous définissons dans un premier temps la vraisemblance :

$$L(y_1, ..., yn, \beta) = \prod_{i=1}^{n} (f(y_i), \beta) \tag{4.5}$$

Nous notons $L(\beta)$ l'expression de la log-vraisemblance :

$$L(\beta) = log(\prod_{i=1}^{n} (f(y_i), \beta)) = log(\prod_{i=1}^{n} C_n^k p^{y_i} (1-p)^{n-y_i} \tag{4.6}$$

En intégrant l'expression de p dans l'équation, on a :

$$L(\beta) = \sum_{i=1}^{n} log(C_n^k (\frac{exp(X\beta_i)}{1 + exp(X\beta_i}})^{y_i})(1 - \frac{exp(X\beta_i)}{1 + exp(X\beta_i}})^{n-y_i}))$$

En simplifiant notre équation, nous obtenons l'expression suivante :

$$L(\beta) = C + \sum_{i=1}^{n} y_i X\beta_i - nlog(1 + exp(X\beta_i))$$

Afin de déterminer les $\beta$ optimaux, nous résolvons l'équation suivante :

$$\frac{\partial L(\beta)}{\partial \beta_i} = 0 \tag{4.7}$$

D'où :

$$\sum_{i=1}^{n} x_i y_i - \frac{nx_i}{1 + exp(-X\beta)} = 0 \tag{4.8}$$

Il n'existe pas de formule explicite permettant de résoudre cette équation pour trouver les $\beta_i$. Il est possible d'utiliser une approximation en nous référant à la méthode de descente de gradient.

Les odds ratios

Afin de pouvoir mieux interpréter les coefficients $\beta_i$ déterminés, nous définissons l'odds qui est le ratio de chances. Dans notre étude par exemple, elle définit la probabilité qu'un

individu fasse du *cross-sell* par rapport un individu n'en faisant pas. L'odds est défini de la façon suivante :

$$ODDS = \frac{P(Y = 1|X = x)}{1 - P(Y = 1|X = x)} \tag{4.9}$$

Dans le cas où nous avons pour une même variable deux modalités différentes, l'expression de l'odd ratio est la suivante :

$$OR = \frac{\frac{P(Y=1|X=x_j)}{1-P(Y=1|X=x_j)}}{\frac{P(Y=1|X=x_{j'})}{1-P(Y=1|X=x_{j'})}} \tag{4.10}$$

En ce qui concerne la régression logistique, les ODDS RATIO valent $e^{\beta_j}$

Avantages et Inconvénients de la régression logistique

— L'un des premiers avantages de la régression logistique est qu'elle nous permet de connaître les coefficients des variables explicatives qui expliquent le modèles. Aussi, des comparaisons peuvent être faites d'une modalité à une autre.

— La sélection des variables est faite sur l'ensemble du modèle. Nous avons donc un ajustement global.

— Parmi les inconvénients de ce modèle, il est à noter qu'une étude est réalisée avant de l'appliquer notamment l'étude de corrélation qui évite l'insertion de biais.

### 4.2.4 Processus de sélection de variables

Après l'étude de l'importance de variables, nous avons gardé 19 variables. En partant du modèle complet, nous éliminons les variables ne présentant pas une pvalue significative en utilisant le test de type III.

Méthode de sélection par type III

Lorsque nous voulons vérifier qu'une variable est pertinente ou pas dans notre modèle de régression logistique, nous nous référons à cette méthode. Si nous supposons k variables possédant m modalités, le test de type III nous donne deux hypothèses :

$H_0$ : $\beta_{k1} = \beta_{k2} = ... = \beta_{km}$ soit la nullité de tous les coefficients.

$H_1$ : il existe j tel que $\beta_{kj} \neq 0$ au moins un coefficient n'étant pas nul

Afin de valider l'hypothèse, la mesure *Likelihood Ratio Test* est utilisée. Nous pouvons ainsi mesurer la différence entre les deux modèles obtenus (avec et sans la variable testée). La statistique pour ce test est la suivante :

$$S = 2ln(L(\hat{\beta}) - L(\hat{\beta_{H_0}})$$

Ce test suit une loi asymptotique de $\chi^2$ à m dégrés de liberté avec m représentant le nombre de modalités. Le test pour conserver la modalité est le suivant :

Si $P(S > \chi^2_m) \leq \alpha$ alors la variable n'est pas supprimée,sinon nous supprimons la variable.

Dans les travaux de ce mémoire, $\alpha$ représentera le risque de première espèce pour rejeter $H_0$. Nous lui donnerons la valeur de 5%.

Le test de Wald

Ce test nous a permis de sélectionner les modalités de nos variables. Nous définissons l'hypothèse $H_0$ comme suit :

$$H_0 : \beta_j = 0$$

En définissant la statistique de test $T = \frac{\beta_j^2}{Var(\hat{\beta}_j^2)}$, elle est comparée à la statistique du test de $\chi^2_1$. La règle est la suivante :

— Si $P(T > \chi^2_1) \leq \alpha$ alors la modalité de la variable est significative
— Si $P(T > \chi^2_1) > \alpha$ alors la modalité n'est pas significative

## 4.3 Arbre de décision

L'arbre de décision est un modèle de prédiction qui est représenté graphiquement sous forme d'une procédure de classification. Il engendre des résultats sous la forme de plusieurs critères logiques. Si tous les critères sont remplis alors l'individu appartient à une classe donnée de la variable cible. L'efficacité des arbres se trouve dans le fait de pouvoir regrouper les individus grâce aux conditions sur les variables explicatives. Nous avons deux catégories différentes d'arbre de décisions :

— Arbres de classification dont la variable cible est catégorielle
— Arbre de régression dont la variable cible est continue.

Un arbre de décision est un arbre où :

— Chaque Noeud interne est un attribut.
— Chaque Branche d'un noeud est un test sur un attribut (ou une variable).
— Les feuilles correspondent à une classe donnée.

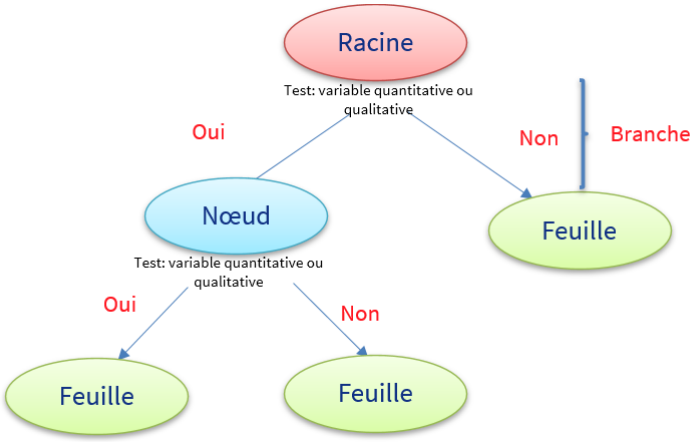



FIGURE 4.2 – Représentation d'un arbre de décision

Le principe de l'arbre est de construire un arbre optimal avec la plupart des variables. Pour construire un arbre, nous nous référons à deux étapes :

-Etape 1 : La construction

Dans cette partie, il s'agira de distribuer les informations dans les nœuds à partir de critères bien définis. Le critère de séparation peut être basé sur l'entropie, le gain d'information, l'indice de gini qui est utilisé par les algorithmes CART pour les variables explicatives discrètes. Nous définissons ci-dessous un algorithme d'arbre de décision.

Algorithme pour les arbres de décision

Initialisation : Racine= Noeud Initial

A chaque noeud :

1. Trouver la meilleure répartition pour chaque variable
2. Déterminer la variable explicative qui segmente le mieux la population
3. Séparation des individus dans les modalités de la variable choisie

Dans chacun des groupes crées, nous utilisons les mêmes étapes jusqu'à atteindre un critère d'arrêt.

Arrêt de l'algorithme : Critère d'arrêt atteint, les feuilles de l'arbre sont les noeuds terminaux.

Les noeuds sont terminaux lorsque le critère d'homogénéité apparaît c'est-à-dire les individus de ce noeud sont de la même classe. Le classement des observations est bien établi.

-Etape 2 : L'élagage de l'arbre

Nous obtenons après l'étape de la construction un grand arbre composé d'un nombre important de feuilles. Les informations les plus importantes se trouvent dans les premières divisions. Afin d'éviter le sur-appprentissage, cette phase permettra d'extraire les feuilles contenant le moins d'information. Pour résoudre ce problème, une fonction de complexité a été défini afin de la minimiser. L'expression de cette dernière est la suivante :

$$c_\beta(A) = E(A) + \alpha|A| \tag{4.11}$$

où |A| est le nombre de feuilles de l'arbre et $\beta$ un réel strictement positif. E(.) est l'erreur de prédiction de l'arbre définit comme suit :

$$E(A) = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}_{g_A(X_i)\neq Y_i} \tag{4.12}$$

où N est le nombre d'observations dans la base de données, $g_A(x)$, une fonction d'application permettant d'associer une sortie y binaire à un vecteur de i variables $(X_1, ..., X_i)$

L'idée pour minimiser la fonction de complexité est de construire un arbre plus grand sans toutefois construire l'arbre complet.

Plusieurs solutions existent pour élaguer l'arbre.

L'une des méthodes consiste à répartir la base en base d'apprentissage et de validation. Des erreurs d'apprentissage et de validation seront calculées. Les erreurs auront tendance à décroître jusqu'à arriver à un niveau de sur-apprentissage.

Un schéma représentatif des erreurs en fonction de la complexité du modèle est défini ci-dessous :

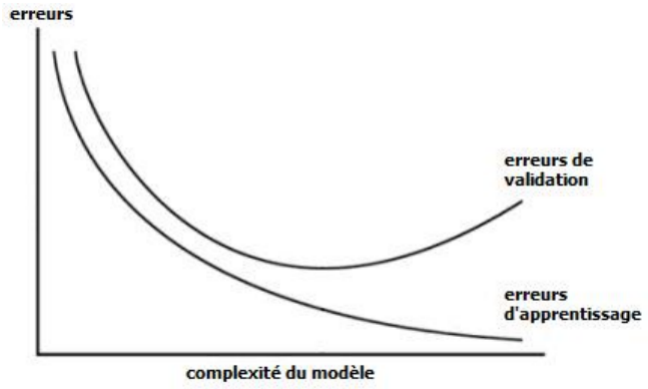



FIGURE 4.3 – Erreur en fonction de la complexité de l'arbre

Une autre méthode est d'utiliser la méthode de validation croisée. Cette méthode consiste à diviser notre base en K sous-ensembles. Dans une première itération, la base de test est représentée par le premier sous-ensemble et les K-1 sous-ensembles constitueront

la base d'apprentissage. A la seconde itération, la base test représentera le second sous-ensemble et ainsi de suite.

Avantages et inconvénients des arbres de décision

Avantages

Les arbres de décisions présentent quelques avantages :

— Les arbres de décision sont faciles à lire, à mettre en place et à être interprétés.
— Il n'est pas obligatoire de spécifier la nature des variables. Elles peuvent être soit qualitatives soit quantitatives

Inconvénients

— Les arbres de décision présentent un manque de robustesse dans la mesure où ceux-ci dépendent de l'échantillon que nous étudions.
— Etant non-paramétriques, il est impossible d'avoir un coefficient propre pour chacune des modalités comme les modèles GLM.
— Les optimums détectés par l'arbre sont locaux. En d'autres termes, c'est à chaque division de noeud que se fait le choix de la variable. L'effet de la variable est vu au moment de cette division pas sur la totalité de l'arbre.

## 4.4 Performance du modèle

La matrice de confusion

La matrice de confusion est un indicateur utilisé afin de comparer les variables prédites $\hat{X}$ des variables observées X par rapport à notre variable cible. Nous déduisons les bonnes et mauvaises estimations à partir d'un seuil de probabilité choisi. Elle nous permet de ressortir certains indicateurs nous permettant de juger la performance de notre modèle. Elle est représentée de la façon suivante :

| | Classe 0 | Classe 1 |
|---|---|---|
| Classe 0 | VN | FP |
| Classe 1 | FN | VP |

Table 4.2 – La matrice de confusion

Avec :
— VN : les vrais négatifs qui représentent la prédiction correcte des non *cross-sell* et qui le sont réellement
— FP : les faux positifs qui représentent les individus prédits comme *cross-sell* mais qui ne le sont pas en réalité
— FN : les faux négatifs représentant les individus ayant réalisés du *cross-sell* mais prédits comme non *cross-sell*

— VP : les vrais positifs qui représentent le classement des individus réalisant du *cross-sell* et qui le sont en réalité

Nous pouvons déduire de cette matrice plusieurs indicateurs que nous définissons :

— Le taux d'erreur représentant le taux de mal classés

$$Erreur = \frac{FP + FN}{VN + FP + FN + VP}$$

— La spécificité représentant les vrais négatifs parmi tous les réels négatifs

$$Specificite = \frac{VN}{VN + FP}$$

— La sensibilité représentant le taux de vrais positifs retrouvés par le modèle parmi les positifs

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

La courbe ROC et l'AUC

La courbe ROC est un instrument la plupart du temps utilisé pour évaluer la qualité d'un modèle. Elle représente la sensibilité en fonction de 1-spécificité avec des seuils variant entre 0 et 1. La sensibilité représente la proportion de vrais positifs et la spécificité la proportion de vrais négatifs .

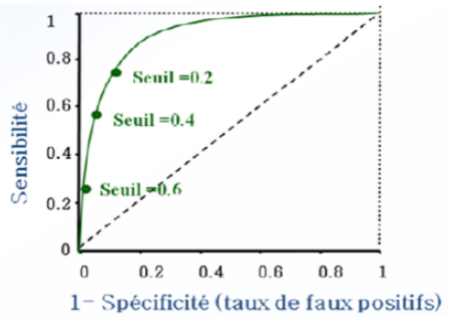



FIGURE 4.4 – Représentation d'une courbe ROC

L'aire sous la courbe est connue sous le nom de AUC(*Area Under the ROC Curve*). Elle permet de juger la performance de notre modèle.

Pour un AUC compris entre 0.5 et 0.7 exclus, la discrimination du modèle est jugée faible. Un modèle ayant une discrimination acceptable a son AUC compris entre 0.7 et 0.8. Plus l'AUC se rapproche de 1 plus notre modèle est performant.

# Chapitre 5

# Aspect pratique de la Modélisation

## 5.1 La régression logistique

### 5.1.1 Importance des variables

Dans un premier temps, il était important de connaître quelles variables allaient intervenir dans le modèle. La méthode utilisée est la méthode stepwise.

La méthode *stepwise* est la combinaison des méthodes backward et forward.

La méthode *backward* consiste à partir d'un modèle complet, à retirer une à une les variables ne jouant un rôle significatif dans notre modèle.

La méthode *forward* consiste quant à elle à partir d'un modèle vide et à y ajouter les variables une à une afin de voir quel ajout aurait une importance significative dans notre modèle.

La méthode *stepwise* consistera à regarder les variables communes entre les deux méthodes. Les ajouts et retraits des variables ont été réalisés selon le critère de l'AIC.

L'AIC correspond au critère d'information d'Akaike. Cet indicateur est défini de la manière suivante :

$$AIC = -2Log(\tilde{L}) + 2(\tilde{k})$$

Avec $\tilde{L}$ la vraisemblance maximisée et $\tilde{k}$ le nombre de paramètres libres (à estimer) dans le modèle. Nous dirons que plus l'AIC est faible, plus nous avons un bon modèle.

Ainsi, cette méthode nous conduit à ne supprimer que la variable donnant la catégorie socioprofessionnelle.

Pour sélectionner nos variables, nous utiliserons la méthode XGBOOST citée plus haut. En effet, le modèle XGBOOST permet d'obtenir une hiérarchie des variables selon la mesure du gain. Elle les classe de la plus significative à la moins significative. Dans notre base, nous avons 23 variables composées de variables qualitatives comme quantitatives. Concernant les variables qualitatives, les modalités sont transformées en variables binaires. Dans la hiérarchie des variables que propose l'algorithme, Ces indicatrices sont aussi présentées par ordre d'importances. L'ordre suivant est ainsi obtenu :

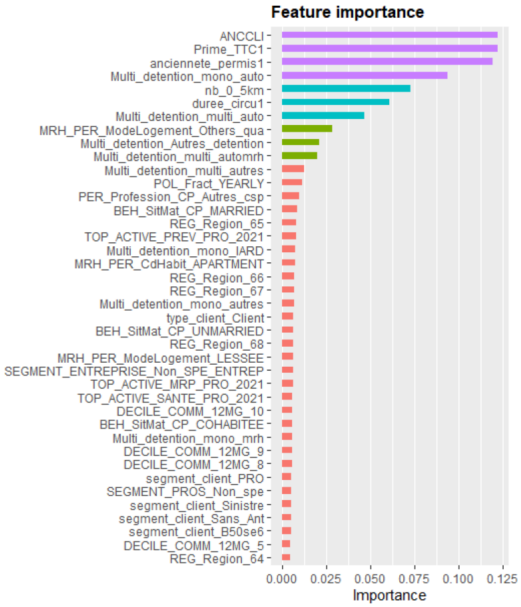



FIGURE 5.1 – Ordre d'importance des variables

Nous pouvons constater que l'ancienneté du client, la prime auto, l'ancienneté du permis, la multi-détention... semble être des variables importantes pour le XGBOOST. Dans le cas des variables présentant plusieurs modalités, nous sélectionnons la variable dès que l'une de ses modalités apparaît selon l'ordre d'importance. Certaines variables sont significativement corrélées dans ce modèle. Afin de conserver un certain nombre de variables, nous étudions la corrélation entre les variables et nous supprimons une variable si elle est corrélée à une variable plus importante qu'elle dans la hiérarchisation. Les variables quantitatives étant peu nombreuses, elles sont discrétisées. Pour cela, un seuil de corrélation à 60% est fixé. Pour des raisons de visibilité, les différentes variables seront mises en annexe.

Nous obtenons ainsi les résultats suivants :

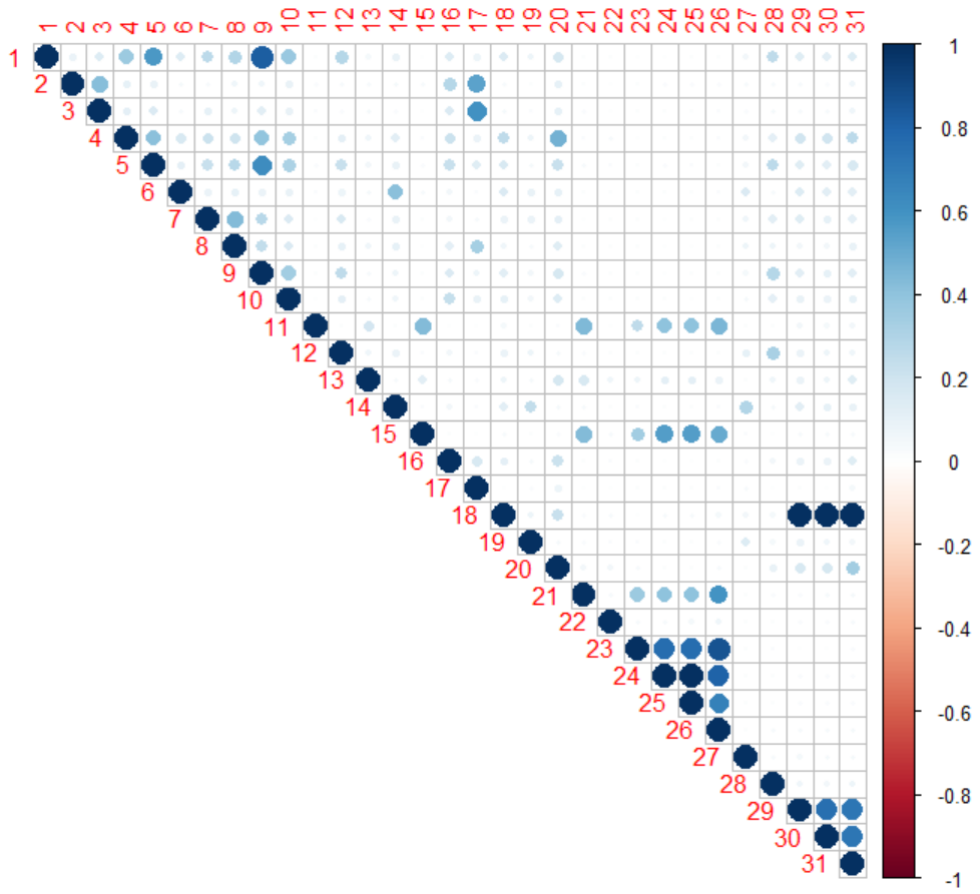



FIGURE 5.2 – Corrélation entre les variables qualitatives

Les variables corrélées sont celles qui présentent des cases fortement bleutées. Les variables portant les informations distributeurs sont corrélées entre elles. La variable donnant la catégorie âge est corrélée à l'ancienneté du permis. Une sélection est ainsi réalisée.

### 5.1.2 Conservation de variables

Dans un premier temps, nous partons du modèle avec les variables retenues par le modèle XGBOOST et l'étude des corrélations. Après avoir appliqué notre modèle, toutes les modalités ne sont pas significatives. Nous cherchons à regrouper certaines modalités des variables afin de les rendre significatives et de donner plus de significativité à notre modèle.

Pour savoir si une modalité est importante ou pas, nous nous référons au test de Wald que nous avons décrit plus haut. Ainsi, lorsqu'une modalité présente une pvalue supérieure à 5%, elle est considérée comme non significative et doit être regroupée avec une autre variable.

En procédant ainsi, il est possible d'avoir plusieurs cas :

— Le cas voulu où la modalité considérée non pertinente devient significative une fois associée à une autre modalité

— La variable peut perdre de la significativité lorsque les modalités sont regroupées. Dans ce cas, la variable est supprimée.

— Il est possible qu'aucun regroupement de modalités ne donne de significativité alors la variable est aussi supprimée.

Nous donnons ainsi un exemple de quelques modalités d'une variable qui ont dû être regroupées avec leur pvalue après avoir appliqué la régression logistique.

| | Coefficients | P-value |
|---|---|---|
| concurr_classconcur_35_63 | -0.19727 | 7.93e-07 |
| concurr_classconcur_63_74 | -1.66186 | 6.36e-10 |
| concurr_classconcur_74 | -0.12830 | 0.56097 |

TABLE 5.1 – Variables donnant le nombre de concurrents

Cette variable donnée dans le tableau représente le nombre de concurrents dans un rayon de 500 m par rapport à la commune. La pvalue donnée par la régression logistique indique que la modalité "nombre de concurrents supérieur à 74" n'est pas significative. Nous décidons donc de la regrouper avec la modalité donnant le nombre de concurrents compris entre 63 et 74 ce qui nous permet d'obtenir la modalité "nombre de concurrents supérieur à 63".

Nous utilisons cette méthode pour toutes les modalités des autres variables.

Ainsi grâce au test de type III, nous obtenons des variables significatives c'est-à-dire avec une pvalue inférieure à 5%.

Nous obtenons donc un modèle de 16 variables pour notre modèle finale.

### 5.1.3 Analyse des résultats

Après avoir sélectionné et regroupé nos variables, nous cherchons à expliquer les différentes modalités des variables en utilisant les odds ratios et un score construit qui prend ses valeurs entre 0 et 100. Plus ce score est élevé, plus la modalité correspondante discrimine le taux de *cross-sell*.

| Variables | Modalités | Coefficient | Odd-ratio |
|---|---|---|---|
| Ancienneté permis | [10,41[ | 0 | 1 |
| | <10 | -0,25 | 0,78 |
| | >=41 | 0,37 | 1,45 |
| Multi-détention | Aucun produit | 0 | 1 |
| | mono_auto | -0,73 | 0,48 |
| | mono_autres | 0,16 | 1,18 |
| | multi_auto | -1,02 | 0,36 |
| | multi_autres | -0,72 | 0,49 |
| Ancienneté client | [0,1[ | 0,00 | 1,00 |
| | [1,5[ | -0,55 | 0,58 |
| | >=11 | -0,48 | 0,62 |
| | [5,11[ | -0,49 | 0,61 |
| Mode logement | Locataires et autres | 0 | 1 |
| | Propriétaires | 0,17 | 1,18 |
| Segment client | B50, Sinistrés, Multi-sinistrés | 0 | 1 |
| | 18-20 | -0,60 | 0,55 |
| | 21-30 | -0,30 | 0,74 |
| | Autres | -0,20 | 0,82 |
| | Creusement de Bonus | -0,20 | 0,82 |
| | Pro | -0,47 | 0,63 |
| | Sans antécédents | -0,78 | 0,46 |
| Formules | F4 | 0 | 1 |
| | F1 | -0,61 | 0,54 |
| | F2 | -0,33 | 0,72 |
| | F3 | -0,10 | 0,90 |
| | F5 | 0,09 | 1,10 |
| | F6 | -0,26 | 0,77 |
| Fractionnement Prime Auto | Mensuel | 0 | 1 |
| | Annuel | -0,39 | 0,68 |
| Situation maritale | Célibataire | 0 | 1 |
| | Autres | 0,43 | 1,54 |
| | COHABITEE | 0,16 | 1,18 |
| | Marié | 0,44 | 1,56 |
| Prime annuelle Auto | [49,199[ | 0 | 1 |
| | [0,49[ | 0,43 | 1,53 |
| | [199,774[ | 0,21 | 1,24 |
| | >=774 | -0,30 | 0,74 |

| | | | |
|---|---|---|---|
| Mode d'habitation | Maison | 0 | 1 |
| | Appartement | 0,17 | 1,19 |
| | Autres | -0,75 | 0,47 |
| Type de garage | Autres | 0 | 1 |
| | STREET | -0,12 | 0,89 |
| | Non spécialisé | 0 | 1 |
| Segment Epargne | Expert | 0,08 | 1,08 |
| Région | Région 67 | 0 | 1 |
| | Region64 | -0,19 | 0,82 |
| | Region65 | 0,26 | 1,29 |
| | Region66 | 0,12 | 1,12 |
| | Region68 | 0,09 | 1,10 |
| Agence santé | Active | 0 | 1 |
| | Non Active | -0,22 | 0,80 |

TABLE 5.2 – Résultats de la régression logistique

D'après ces résultats de score, les affirmations suivantes sont déduites :

— Les propriétaires ont plus tendance à réaliser du *cross-sell* contrairement aux locataires.
— La variable multi-détention discrimine fortement notre modèle. Les personnes détenant un unique contrat n'étant ni auto ni MRH ont une forte probabilité de de faire du *cross-sell* Auto MRH. Ensuite, viens ceux ne détenant aucun produit chez AXA c'est-à-dire les prospects. Pour les clients ne détenant pas de produits, ils seront attirés par le fait d'avoir plusieurs produits en vue d'une réduction tarifaire.
— Au niveau du fractionnement de la prime Auto, les personnes payant leur prime annuellement auront tendance à faire moins de *cross-sell*.
— Concernant la région, les personnes habitant la région 65 réalisent plus de *cross-sell* par rapport aux autres régions.
— Comme observé dans les statistiques descriptives, nous observons que les prospects ont plus tendance à réaliser du *cross-sell* donc les personnes ayant une ancienneté de 0. En effet, les personnes ayant une ancienneté importante ont sûrement eu le temps de se multiéquiper durant leur année de contrat. S'ils souscrivent donc à un contrat Auto, ils n'auront peut-être pas besoin d'avoir un contrat MRH puisqu'ils l'auront déjà.
— Les bonus 50, les sinistrés, les multi-sinistrés sont les segments qui réalisent le plus de *cross-sell* comme observé dans les statistiques descriptives.
— Nous observons également que les anciens conducteurs ont tendance à faire plus de *cross-sell*.
— Les individus payant le moins de prime annuelle ont une appétence à faire du *cross-sell*. En payant moins de prime, ceux-ci auront tendance à s'octroyer de nouveaux contrat sur lesquels ils pourront certainement avoir des réductions.
— En ce qui concerne le type de garage, les personnes garant leur voiture dans la

rue se multiéquipent moins rapidement.
— Comme vu dans les statistiques descriptives, les segments épargnes et les agences n'ont pas de significativité pertinente sur le *cross-sell*.

Pour ce qui est de la performance de notre modèle, nous obtenons la courbe ROC suivante :

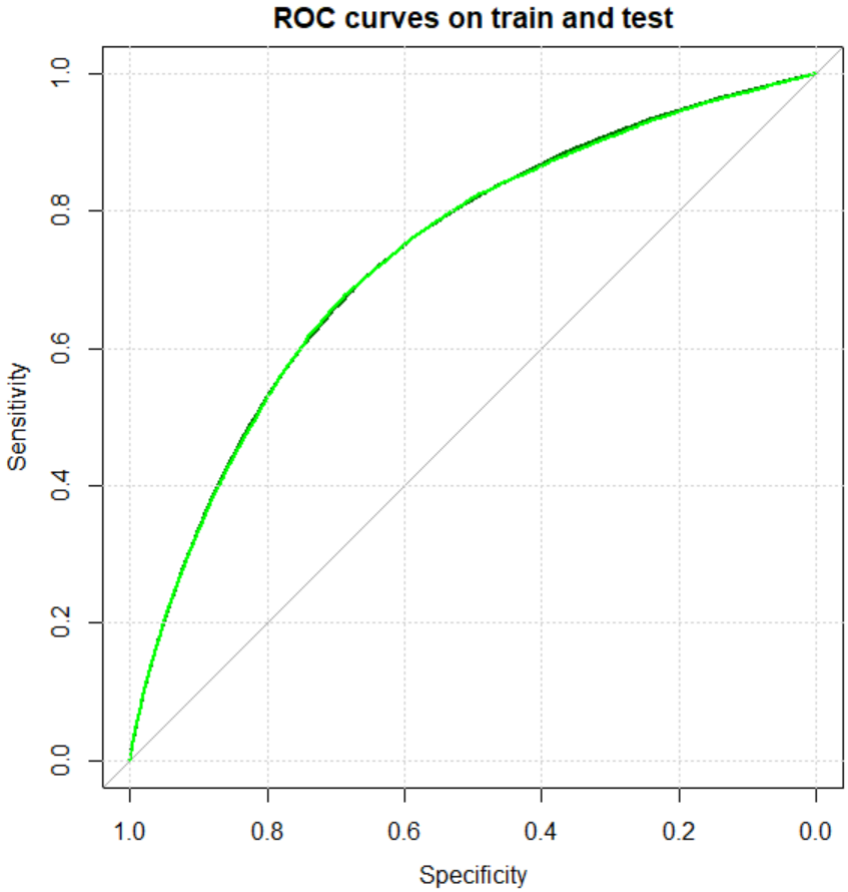



FIGURE 5.3 – Courbe ROC du GLM

Nous avons calculé l'aire sous la courbe ROC sur la base d'apprentissage ainsi que sur la base de test. Nous obtenons ainsi un AUC de 0.7333 sur la base d'apprentissage et de 0.7324 sur la base de test.

Une autre méthode pour vérifier la stabilité du modèle est de calculer le taux d'erreur à travers la matrice de confusion sur la base d'apprentissage et sur le test.

Pour choisir le seuil qui donnerait de bonnes prédictions des vrais positifs qui représentent la modalité 1 et des vrais négatifs représentant la modalité 0, les courbes de sensibilités et de spécificités sont construites en fonction des seuils de probabilité. L'intersection entre les deux courbes permet de donner le meilleur compromis.

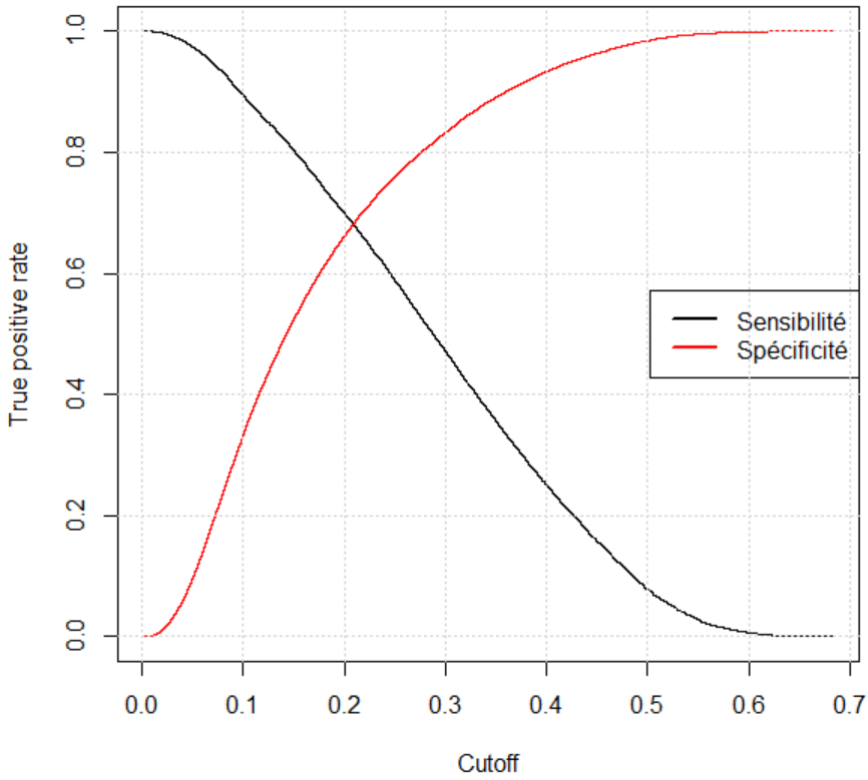



FIGURE 5.4 – La spécificité et la sensibilité en fonction des différents seuils de probabilité selon le modèle logistique

Le seuil donnant le meilleur compromis est de l'ordre de 0.21. A partir de ce seuil, les matrices de confusions suivantes ressortent :

| | | Classes prédites | |
|---|---|---|---|
| | | 0 | 1 |
| Classes réelles | 0 | 69.23% | 30.77% |
| | 1 | 33.13% | 66.87% |

TABLE 5.3 – Matrice de confusion sur la base d'apprentissage

| | | Classes prédites | |
|---|---|---|---|
| | | 0 | 1 |
| Classes réelles | 0 | 69.21% | 30.79% |
| | 1 | 32.72% | 67.28% |

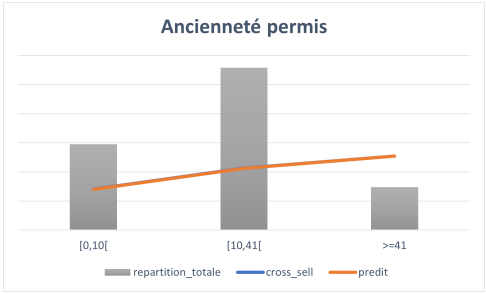
TABLE 5.4 – Matrice de confusion sur la base de test

Sur la base d'apprentissage, nous avons un taux d'erreur de 0.3123668 contre un taux de 0.3117221 sur la base de test. Ces deux taux sont relativement proches. Nous pouvons conclure sur la stabilité de notre modèle.

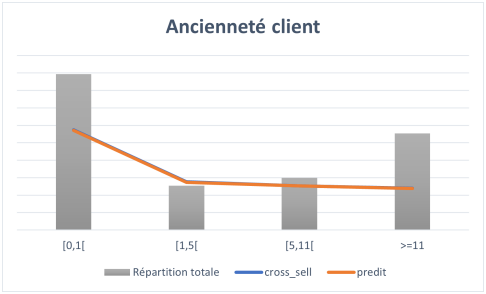
### 5.1.4 Validation du modèle

Afin de valider nos modèles, il est possible de passer par les prédictions des analyses univariées qui seront comparées au réel. Nous décidons de sélectionner les variables

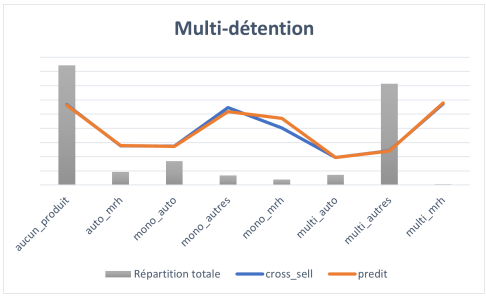
discriminant notre taux de *cross-sell*.



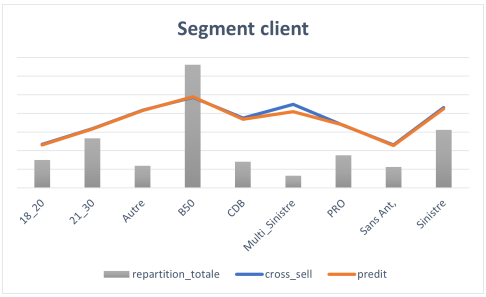

(a) Ancienneté permis




(b) Ancienneté client




(c) Multi-détention




(d) Segment Client

FIGURE 5.5 – Prédictions univariées

Le modèle de régression logistique semble bien prédire les différentes variables. Nous retrouvons quelques différences au niveau de certaines variables comme la multi-détention, la segmentation client qui sont d'environ de 1%.

## 5.2 Les arbres de décision

La mise en place de l'arbre de décision a été réalisée sur le package R. Les librairies qui ont été utilisées sont les librairies rpart pour la réalisation du modèle et rpart.plot pour schématiser l'arbre en question. Nous notons également que les variables corrélées ont été rétirées pour réaliser l'arbre. Les paramètres qui ont été utilisés sont les suivants :

— cp le paramètre de complexité. Il permet de minimiser l'erreur de validation croisée en vue d'éviter du surapprentissage. L'arbre de taille maximale est obtenu lorsque sa valeur prend 0. L'éloigner de 0 sans pour autant qu'il soit très grand permettra d'avoir un arbre plus lisible et plus petit. Un schéma représentant le cp choisi avec la taille de l'arbre est présenté ci-dessous.

— Minbucket utilisé pour définir le nombre minimum d'individus par feuille

Afin de déterminer la valeur du cp minimisant la fonction d'erreur, nous utilisons

la fonction *plotcp* de la librairie rpart. Le graphique obtenu nous permettra de choisir un cp adapté en fonction du nombre de feuilles afin d'éviter du surapprentissage. En choisissant un arbre moins complexe, le risque de surapprentissage est moins élevé. Ainsi, nous choisissons un cp de 0.001 nous donnant le graphique suivant :

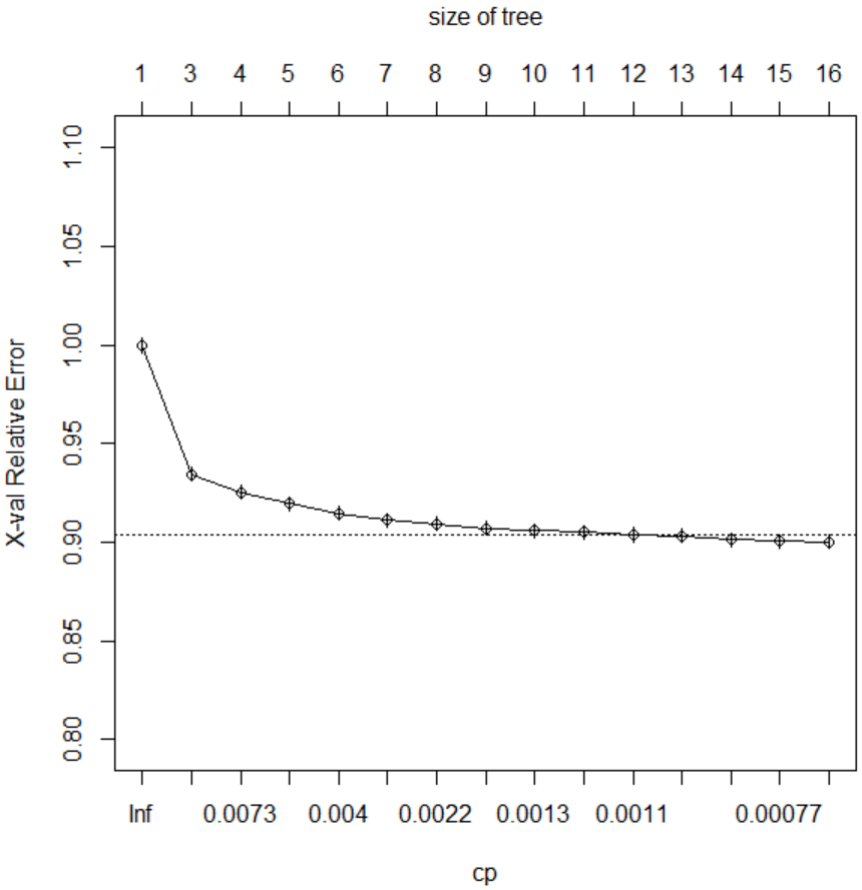



FIGURE 5.6 – L'Erreur de validation croisée en fonction de la valeur du paramètre de complexité

En ayant constaté la stabilité sur la base d'apprentissage pour différentes valeurs de cp avec l'indicateur donnant la différence entre les moyennes observées et prédites, nous cherchons à déterminer le cp minimisant cet indicateur sur la base de test. Ainsi, nous sélectionnons la moyenne prédite la plus proche de la moyenne observée sur le test. Nous obtenons donc un cp de l'ordre de 0.008 qui nous donne un arbre à 4 feuilles.

Par ailleurs, la hiérarchie des variables ressortie par l'arbre est la suivante :

L'ancienneté client, la multi-détention, l'ancienneté permis, la segmentation client sont les variables les plus discriminantes du modèle.

Ainsi, nous obtenons l'arbre suivant :

| Variables | Importance |
|---|---|
| Ancienneté client | 1504 |
| Multi-détention | 1376 |
| Ancienneté Permis | 1342 |
| Segment client | 1183 |
| Âge | 963 |
| Type client | 629 |
| Mode logement | 442 |
| Situation Maritale | 360 |
| Nombre de véhicule | 257 |
| Prime | 189 |
| Usage du véhicule | 31 |
| Formule | 27 |
| Type logement | 21 |
| Durée de circulationdu véhicule | 5 |

TABLE 5.5 – Importance des variables selon l'arbre

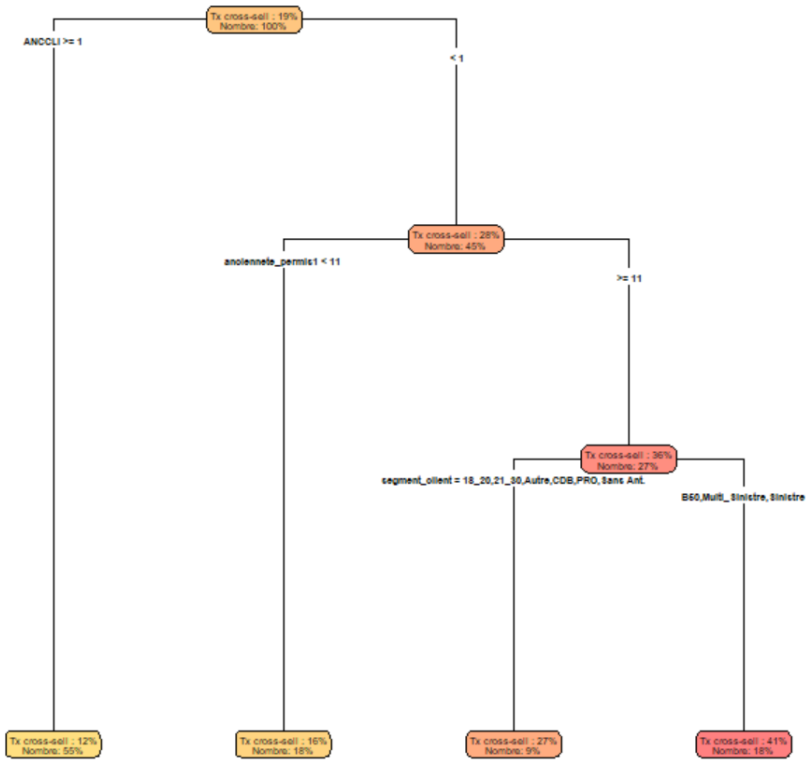



FIGURE 5.7 – Arbre expliquant le taux de *cross-sell*

A partir de cet arbre, nous remarquons qu'au premier noeud, les personnes n'ayant pas d'ancienneté dans le portefeuille sont susceptibles de faire du *cross-sell* ce qui est conforme à la régression logistique faite plus haut.

Ensuite, les anciens conducteurs et les segments clients Bonus 50, sinistrés, multi-sinistrés permettent d'obtenir un taux un peu plus élevé soit 41% avec un pourcentage de 18%.

La courbe ROC obtenue de ce modèle est la suivante :

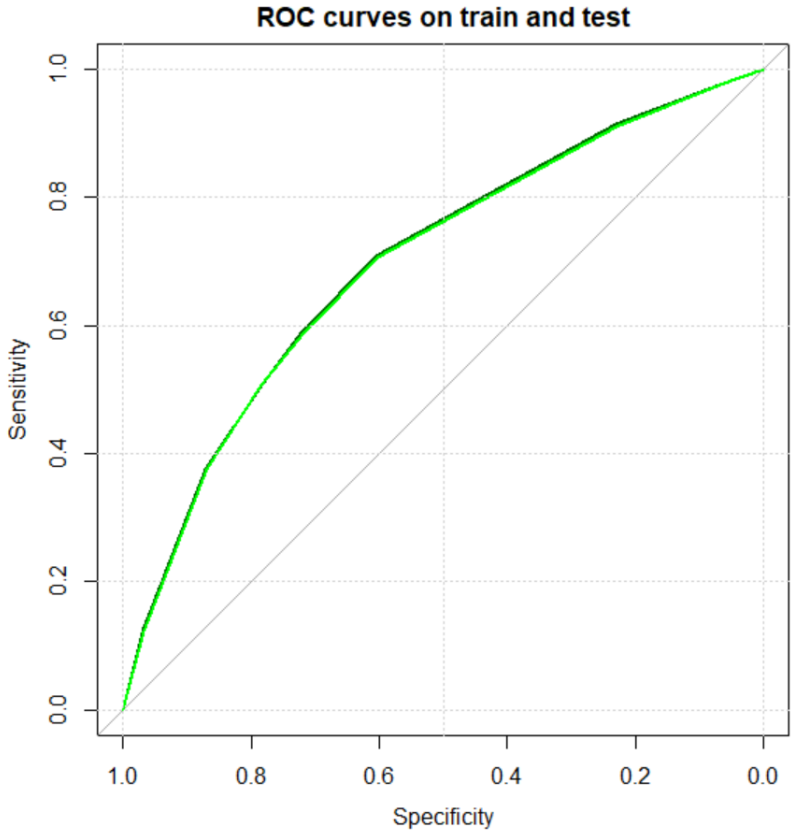



FIGURE 5.8 – Courbe ROC - Arbre de décision

Nous obtenons un AUC de 0.7128 sur l'apprentissage et de 0.7135 sur la base de test.

L'évolution de la sensibilité et de la spécificité en fonction des seuils de probabilité est déduite :

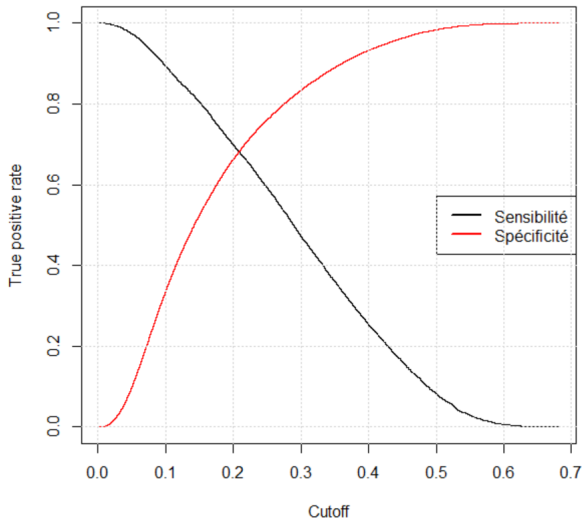



FIGURE 5.9 – La spécificité et la sensibilité en fonction des différents seuils de probabilité selon l'algorithme CART

Les matrices de confusion sur l'apprentissage et le test sont observées pour un seuil de 0.21 représentant l'intersection des deux courbes.

| | | Classes prédites | |
|---|---|---|---|
| | | 0 | 1 |
| Classes réelles | 0 | 66.81% | 33.19% |
| | 1 | 32.98% | 67.02% |

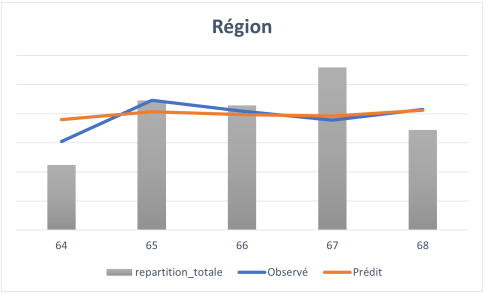
TABLE 5.6 – Matrice de confusion sur la base d'apprentissage

| | | Classes prédites | |
|---|---|---|---|
| | | 0 | 1 |
| Classes réelles | 0 | 66.79% | 33.21% |
| | 1 | 32.76% | 67.24% |

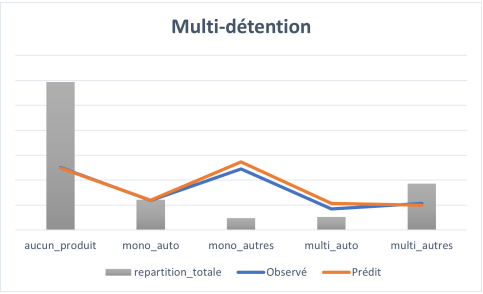
TABLE 5.7 – Matrice de confusion sur la base de test

Les taux d'erreur sur la base d'apprentissage et la base de test sont respectivement de 0.1960882 et de 0.1976414.

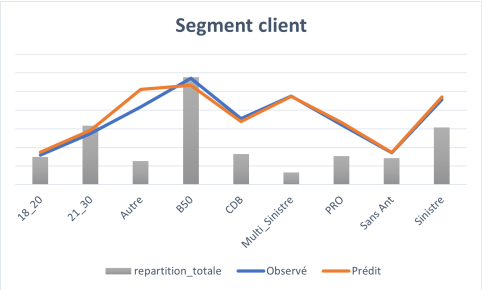
Comme pour la régression logistique, nous cherchons à comparer le réel et le prédit à travers quelques statistiques descriptives.




(a) Région




(b) Multi-détention




(c) La segmentation client

FIGURE 5.10 – Prédictions univariées

Contrairement à la régression logistique, les arbres de décision ne prédisent pas correctement sur certaines modalités des variables. En effet, l'un des inconvénients est qu'il ne permet pas d'obtenir des coefficients précis pour les modalités des variables.

## 5.3 Le gradient boosting

Cette méthode d'apprentissage a été réalisé sous R avec la library XGBOOST. Le package caret a été utilisé dans le but de tester plusieurs paramètres. Les paramètres sont les suivants :

— Objective : binary :logistic
— nrounds : le nombre d'itérations
— maxdepth : la profondeur maximale de l'arbre
— min_child_weight : le poids minimal des feuilles
— subsample : le nombre d'individus échantillonnés aléatoirement
— colsample_bytree : la proportion de colonnes considérée à chaque échantillon
— eta : le taux d'apprentissage permettant d'éviter le surapprentissage. Il prend ses valeurs entre 0 et 1.
— gamma : Ce paramètre prend ses valeurs entre 0 et $+\infty$. Il minimise la fonction de perte afin de favoriser la fraction d'un nœud.

Nous testons les paramètres suivants :

| Paramètres | Valeurs |
|---|---|
| nrounds | 100/150/200 |
| maxdepth | 8 / 10 / 16 |
| min_child_weight | 1 / 5 |
| subsample | 0.5 / 1 |
| colsample_bytree | 0.5 / 1 |
| Gamma | 0.3 / 0.5 / 1 |
| eta | 0.1 / 0.2/ 0.5 |

TABLE 5.8 – Paramètres testés du XGBOOST

Après avoir testé plusieurs paramètres, la grille de recherche nous permet d'obtenir la combinaison suivante :

— nrounds : 150
— maxdepth : 4
— min_child_weight : 1
— subsample : 0.5
— colsample_bytree : 1
— eta : 0.1
— gamma : 0.5

Nous représentons les 30 premières variables ressortant de ce modèle.

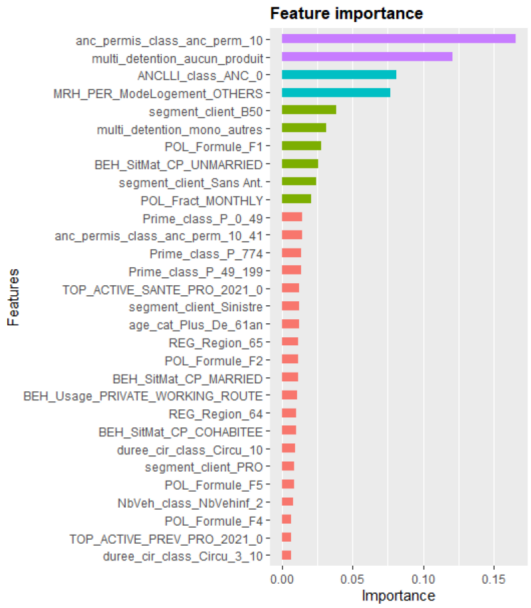



FIGURE 5.11 – Importance des variables

Les variables telles que l'ancienneté permis, la multi-détention, l'ancienneté du por-

tefeuille client ressortent comme importantes, ce qui est similaire par rapport aux autres modèles.

La courbe ROC suivante est obtenue :

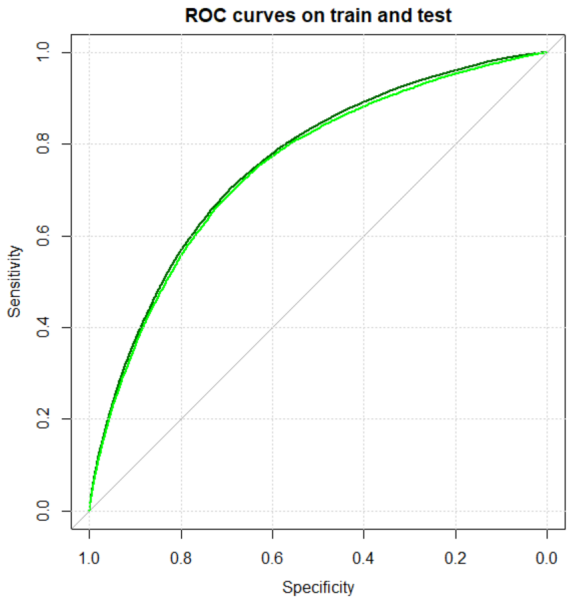



FIGURE 5.12 – Courbe ROC - XGBOOST

De cette courbe ROC, nous obtenons un AUC de 0.7577 sur la base d'apprentissage et de 0.749 sur la base de test.

Comme les deux précédents modèles, la sensibilité et la spécificité sont représentées :

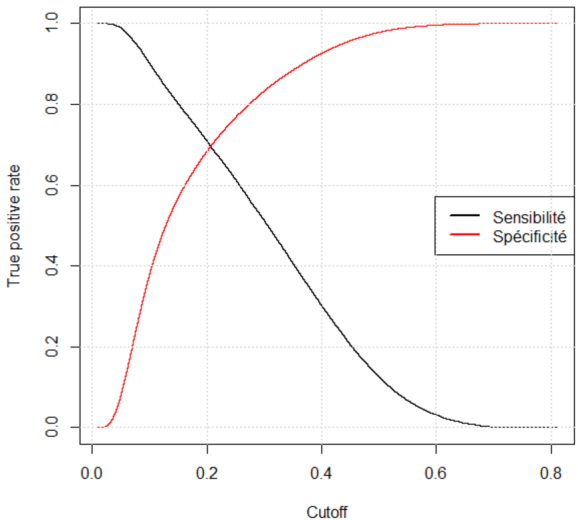


FIGURE 5.13 – La spécificité et la sensibilité en fonction des différents seuils de probabilité selon le modèle XGBOOST

Pour un seuil discriminant de 0.20, les matrices de confusion suivantes sont obtenues :

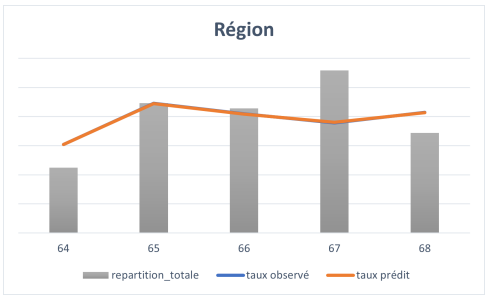
| | | Classes prédites | |
|---|---|---|---|
| | | 0 | 1 |
| Classes réelles | 0 | 68,38% | 31,62% |
| | 1 | 28,97% | 71,03% |

TABLE 5.9 – Matrice de confusion sur la base d'apprentissage

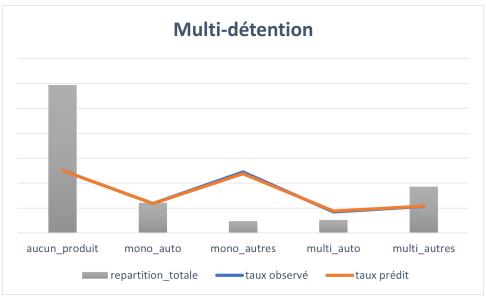
| | | Classes prédites | |
|---|---|---|---|
| | | 0 | 1 |
| Classes réelles | 0 | 68,05% | 31,95% |
| | 1 | 29,70% | 70,30% |

TABLE 5.10 – Matrice de confusion sur la base de test

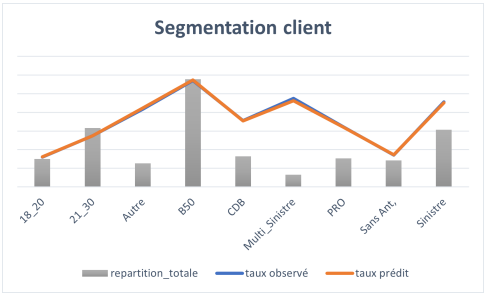
Afin de vérifier la stabilité de notre modèle, nous utilisons les statistiques descriptives de quelques variables :




(a) Région




(b) Multi-détention




(c) La segmentation client

FIGURE 5.14 – Prédictions univariées

Les prédictions suivent bien les tendances observées pour la plupart des variables. Nous pouvons ainsi juger d'une bonne stabilité de notre modèle.

## 5.4 Comparaison des modèles

Afin de comparer nos modèles, les variables les plus pertinentes qui se retrouvent dans chacun des modèles seront évoquées. Certaines variables jugées pertinentes par un modèle ressortent comme pertinentes dans les deux autres. Ce sont les variables telles que :

— La multi-détention
— L'ancienneté client dans le portefeuille
— L'ancienneté permis
— La segmentation client

L'AUC sur la base de test, l'interprétabilité et la facilité opérationnelle nous permettront de comparer nos modèles :

| Modèles | AUC(test) | Interprétabilité | Facilité opérationnelle |
|---|---|---|---|
| GLM | 0.7324 | + | + |
| CART | 0.7128 | + | + |
| XGBOOST | 0.749 | - | - |

TABLE 5.11 – Comparaison des trois modèles selon l'AUC, l'interprétabilité, la facilité opérationnelle

Nous remarquons que selon l'AUC, le modèle donnant une meilleure performance est le modèle XGBOOST. Ces différents modèles possèdent des AUC acceptables de plus de 70% mais chacun possédant des avantages et des inconvénients, nous utiliserons celui qui correspondrait au mieux à nos attentes.

En ce qui concerne l'algorithme CART, la lecture étant beaucoup plus simple que les autres présentent de faible performance de prédiction. Aussi, en fonction du paramètre de complexité choisi pour éviter le surapprentissage, une grande dimension de l'arbre entraîne une lecture de plus en plus difficile.

Les arbres boostés présentent certe des performances de prédictions intéressantes mais l'interprétation n'est pas aussi aisée.

Le modèle GLM correspond le mieux à notre enjeu. Ses prédictions sont acceptables. A partir des différents coefficients déterminés par le modèle, il est possible d'interpréter chacune des modalités des variables. Des profils peuvent être construits en croisant plusieurs modalités de différentes variables et d'en obtenir des probabilités. La formule utilisée pour la déterminer est l'inverse de la fonction logit. L'interprétation du modèle GLM est relativement simple.

## 5.5 Limites et améliorations possibles du modèle

Plusieurs points peuvent être relevés dans cette étude. L'ajout d'informations supplémentaires concernant les distributeurs et de données de concurrence selon la commune auraient pu apporter plus de significativité à notre modèle. Par exemple, il pourrait être intéressant de rechercher les lieux de localisation des agences à savoir si ces agences se trouvent dans une zone rurale, urbaine... Pour ce qui est de la concurrence, rechercher le taux de concurrence et le taux de pénétration AXA pour les codes communes existantes dans les bases AXA pourrait d'avantage expliquer le taux de *cross-sell*.

D'autres informations concernant le client pourraient être rajoutées telles que des informations donnant le revenu ou le niveau de vie du client.

# Chapitre 6

# Exemple d'application du modèle

Dans cette partie du mémoire, nous chercherons à appliquer notre modèle validé en vue d'aider les agents à cibler les personnes qui lui permettront d'augementer son taux de *cross-sell*. Trois approches pourront être utilisées.

Première approche

L'idée sera d'utiliser l'arbre de décision calibré plus haut pour trouver les profils qui feront augmenter le taux de *cross-sell*. Au départ, les agents réalisent 19% de *cross-sell* en divisant la base pour l'apprentissage et la validation. Pour faire gagner du temps à l'agent à appeler les personnes ayant une forte probabilité de se multi-équiper, nous aurons une sous-population ayant des caractéristiques plus restreintes qui feront augmenter le taux de *cross-sell*. Les indicateurs donnés dans l'arbre seront alors le pourcentage de la population concernée par rapport à la population totale et le taux de *cross-sell* réalisé par celle-ci en fonction des caractéristiques données dans les branches.

Le noeud qui nous sera intéressant est le noeud permettant d'avoir le plus fort taux de *cross-sell* et une population pas très restreinte afin d'avoir des informations fiables. L'arbre suivant est ainsi obtenu :

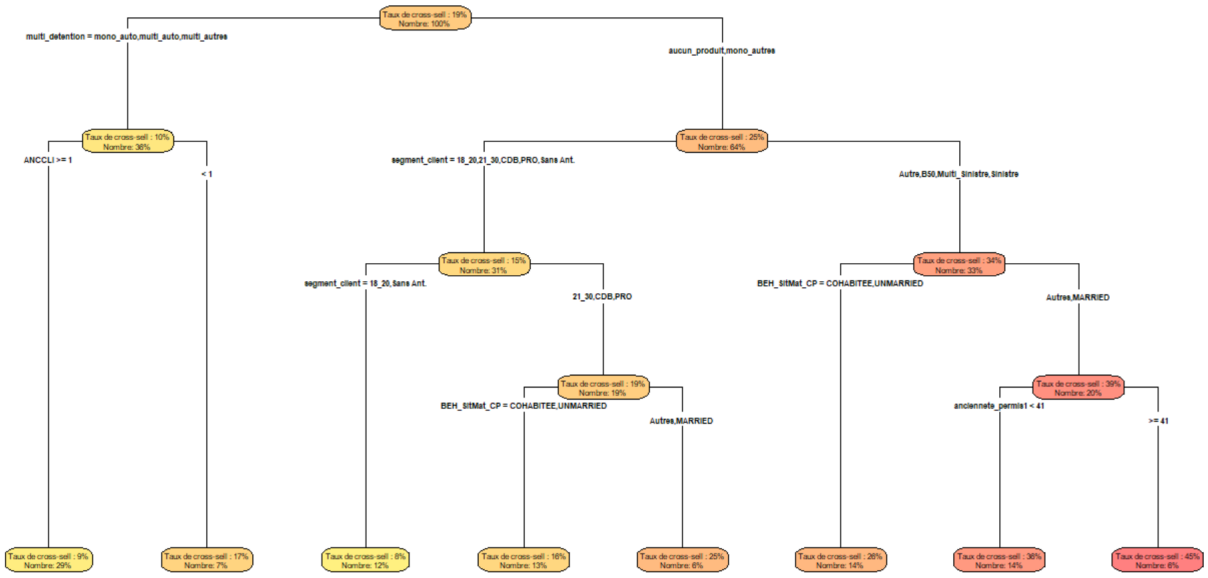



FIGURE 6.1 – Arbre de décision

En nous basant sur le noeud de droite, nous avons la plus grande probabilité de *cross-sell* qui est de 45% avec une sous-population de 6% de la base totale. Ces pourcentages correspondent à une population ayant les caractéristiques suivantes :

— Selon la détention du produit, nous avons les personnes ne détenant aucun produit donc les prospects. Aussi, sont sélectionnées les personnes étant déjà clientes chez AXA mais détenant des contrats uniques autres que l'auto et l'habitation.

— Pour ce qui est de la segmentation client, les sinistrés, les multi-sinistrés, les Bonus 50 et les autres(concernant les personnes n'appartenant à aucun des segments cités dans le premier chapitre)

— Les mariés et les autres(concernent les personnes n'appartenant pas à la catégorie marié et célibataire)
— Il y a enfin les anciens conducteurs ayant une ancienneté de permis d'au moins 41 mois.

Ces profils nous permettront de donner un score de probabilité de *cross-sell* qui sera intégrer dans un outil utilisé par les agents appelé *Salesforce*. Ce score de probabilité sera déterminé en croisant les profils de l'arbre. Nous justifions cela par le fait que le GLM donne plus de précision en prédiction par rapport à l'arbre.

Salesforce est une interface dont dispose les agents afin de suivre de manière efficace et personnalisé la relation client. Lorsqu'un individu fait un devis, les différentes informations concernant ce dernier sont entrées à partir de cette interface. Une fiche client est

alors créée lors de la souscription et à l'intérieur de cette fiche se trouve des scores. Un atout de cette plateforme est qu'elle permet à l'agent de piloter son activité, de recevoir des alertes, des rappels, des prises de rendez-vous etc.

Ainsi, lorsque le client rentrera ses informations, selon les profils définis dans l'arbre, une alerte apparaîtra en indiquant à l'agent que ce client aurait plus d'appétence à faire du *cross-sell* à partir du score calculé par le GLM. Le score pourra être mis à jour en fonction des informations qui seront modifiées.

Deuxième approche

Une autre approche est d'utiliser les scores donnés par le GLM. Avec le modèle GLM, il est possible d'obtenir la probabilité de multi-équipement de chaque individu en faisant le croisement de plusieurs variables. Une maquette sous Excel est construite avec un code couleur donnant le taux de *cross-sell* résultant d'un croisement de variables discriminantes. Des croisements donnant des probabilités élevées seraient intéressants pour l'agent.

| Variables | Valeurs | Contribution |
|---|---|---|
| Ancienneté permis | 41 | |
| Multi-détention | Mono autres | |
| Formule Auto | F5 | |
| Situation matrimoniale | marrié | |
| Région | 65 | |
| Total | | 75% |

FIGURE 6.2 – Exemple d'alerte sur quelques caractéristiques

A partir de cette application, les agents pourront se souvenir par exemple qu'une personne ayant une détention "Mono Autres" aura tendance à augmenter sa probabilité à réussir un *cross-sell*.

Troisième approche

Cette troisième approche consistera à cibler les agents qui auraient dû réaliser plus de *cross-sell* mais qui ne l'ont pas fait.

Pour ce faire, nous utiliserons les segments de l'arbre défini plus haut mais en se limitant à l'avant-dernier noeud de la branche de droite.

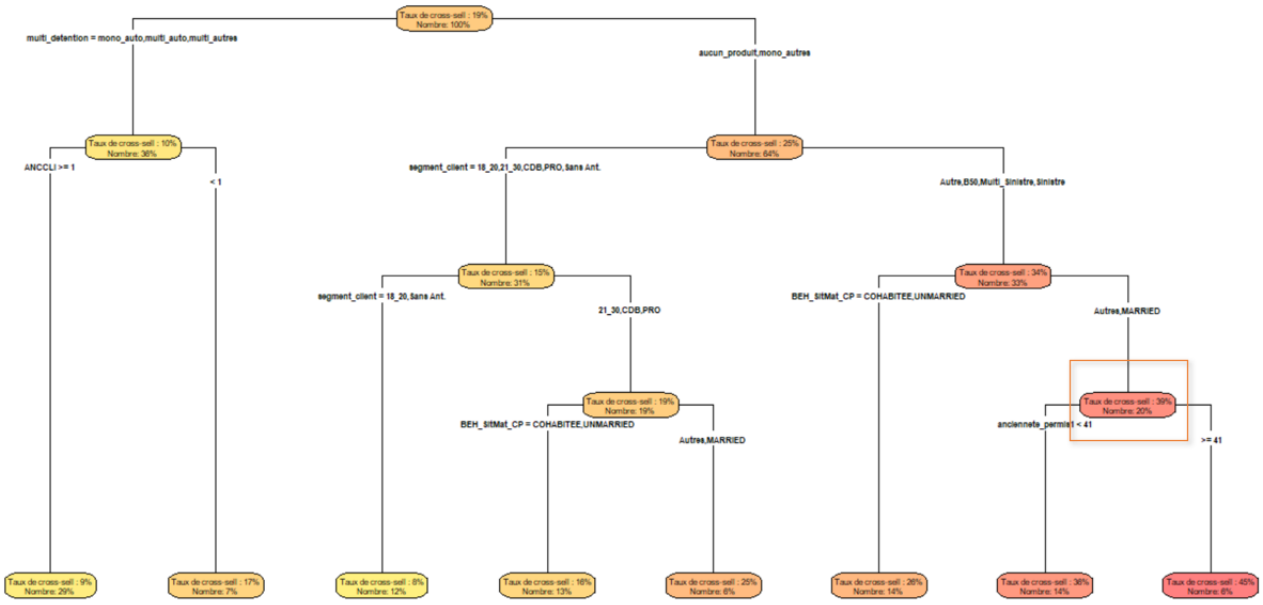



FIGURE 6.3 – Noeud à utiliser

Le modèle GLM sera utilisé pour la prédiction en filtrant notre base sur les segments de l'arbre. Les agents sont reconnus dans les bases AXA grâce au code gestionnaire. Nous comparons ainsi les taux de *cross-sell* réels et prédits par code gestionnaire.

Pour les agents dont le taux de *cross-sell* prédit est supérieur au taux de *cross-sell* réel, nous cherchons à déterminer combien de contrats nous aurions pu gagner si l'attention avait été portée envers les profils du segment choisi.

Après calcul, nous obtenons un nombre de contrats d'environ 3000 qui auraient pu être gagnés.

Il sera alors possible de contacter ces agents pour montrer cette alerte.

# Conclusion

L'objectif du mémoire est de déterminer les types de profils qui seraient plus enclins à réaliser du *cross-sell* Auto MRH. Pour ce faire, les probabilités estimées ont été calculées à partir de la régression logistique, des arbres de décisions et du modèle XGBOOST. Les variables les plus impactantes telles que l'ancienneté permis, l'ancienneté client, la multi-détention, la segmentation client ressortent dans les trois modèles. Aussi, les performances de ces modèles sont relativement acceptables avec des AUC supérieures à 70%.

Tout d'abord, nous optons pour la régression logistique car elle présente une interprétabilité plus explicite en donnant un coefficient à chacune des modalités des variables. Ainsi, l'impact de chaque modalité sur la variable cible peut être interprété. Il est aussi possible de réaliser des profils en procédant à des croisements de modalités des variables qui conduira à déterminer la probabilité de chaque profil.

Ensuite, les arbres de décision présentent certes une lecture simple mais les performances de prédictions ne sont pas aussi bonnes que celles de la régression logistique et du XGBOOST. Le modèle XGBOOST quant à lui ne possède pas une interprétabilité plus aisée comme le GLM. Nous décidons ainsi de garder la régression logistique.

Une fois le modèle sélectionné, nous recherchons à déterminer le type de profil qui augmenterait le taux de *cross-sell*. Le taux de *cross-sell* augmente sur une partie restreinte de la population. L'arbre de décision nous permet ainsi de les cibler. Ainsi, 10% de la population sont ciblés donnant un taux de *cross-sell* de 49% contre 19% au départ. Dans un premier temps, les agents pourront être attentifs à ce type de profils s'ils veulent augmenter leur taux. Aussi, la régression logistique est ainsi utilisée pour prédire le taux de *cross-sell* par agent sur ces individus. En ayant déterminé le nombre de devis réalisés par chacun d'eux, à partir du taux de *cross-sell* réel et prédit, les agents qui auraient dû faire plus de *cross-sell* sont ciblés. Ces agents ciblés auraient réalisé ainsi plus de 3000 contrats en plus. Ces analyses montrent ainsi l'utilité de notre modèle.

Cependant, l'ajout de certaines informations portant sur les distributeurs, les données de concurrence et marché auraient pu apporter plus de compréhension à notre étude. Des informations portant sur la situation de vie du client pourraient être intéressantes.

# Table des figures

# Liste des tableaux

# Annexe A

# Numérotation et Variables

| | |
|---|---|
| 1 | anc_permis_class |
| 2 | multi_detention |
| 3 | ANCLLI_class |
| 4 | MRH_PER_ModeLogement |
| 5 | segment_client |
| 6 | POL_Formule |
| 7 | POL_Fract |
| 8 | Prime_class |
| 9 | age_cat |
| 10 | BEH_SitMat_CP |
| 11 | TOP_ACTIVE_SANTE_PRO_2021 |
| 12 | BEH_Usage |
| 13 | REG_Region |
| 14 | duree_cir_class |
| 15 | TOP_ACTIVE_PREV_PRO_2021 |
| 16 | NbVeh_class |
| 17 | type_client |
| 18 | BEH_TypeGarage |
| 19 | VEH_Energie |
| 20 | MRH_PER_CdHabit |

| | |
|---|---|
| 21 | TOP_ACTIVE_MRP_PRO_2021 |
| 22 | concurr_class |
| 23 | SEGMENT_ENTREPRISE |
| 24 | SEGMENT_PROS |
| 25 | SEGMENT_EPARGNE |
| 26 | DECILE_COMM_12MG |
| 27 | PER_ModeAchatVeh |
| 28 | PER_Profession_CP |
| 29 | BEH_TypeGarageIndoor |
| 30 | BEH_TypeGarageClosed |
| 31 | BEH_TypeGarageCollective |

# Bibliographie


[ARMENTERAS, 2022] ARMENTERAS, M. (2022). *Définition d'une valeur contrat à la souscription pour les contrats d'assurance habitation.* Mémoire d'actuariat.

[CHALENDARD, 2014] CHALENDARD, R. (2014). *Optimisation du multi-équipement dans le cadre d'une vente additionnelle.* Mémoire d'actuariat.

[Charguéraud, 2016] CHARGUÉRAUD, A. (2016). *Le taux de transformation en automobile : comparaison de différentes méthodes d'apprentissage.* Mémoire d'actuariat.

[Coulibaly, 2021] COULIBALY, A. (2021). *Modélisation des sinistres de la garantie incendie en Multirisque Professionnelle : L'apport de machine learning.* Mémoire d'actuariat.

[DENNIEL, 2021] DENNIEL, C. (2021). *Lissage des résidus par Krigeage dans la création d'un zonier : Application sur un portefeuille MRH.* Mémoire d'actuariat.

[DOUMBIA, 2021] DOUMBIA, M. (2021). *Les critères qui discriminent le taux de transformation des devis Auto.* Rapport étude interne.

[France Assureurs, 2020] FRANCE ASSUREURS (2020). L'assurance franÇaise données clés 2020. `https://www.franceassureurs.fr/wp-content/uploads/VF-Donnees-cles-2020.pdf`.

[Herling, 2019] HERLING, A.-S. (2019). *Evaluation de Ia qualité de la souscription en assurance automobile grâce à la construction d'un indicateur de valeur.* Mémoire d'actuariat.

[KAREEM, 2021] KAREEM, F. (2021). *Modélisation et Analyse du taux de résiliation Assuré de la branche Automobile.* Mémoire d'études statistiques.

[MARKAOUI, 2016] MARKAOUI, H. (2016). *Analyse de la probabilité de résiliation en assurance automobile : comparatif de deux méthodologies d'estimation et conséquences sur la tarification.* Mémoire d'actuariat.

[PEYRILLER, 2019] PEYRILLER, A. (2019). *Prédiction des résiliations en santé individuelle.* Mémoire d'actuariat.

[RAITI, 2016] RAITI, M. (2016). *Analyse technique du multi-équipement dans le cadre d'une stratégie Post-ANI.* Mémoire d'actuariat.

[ROBERT, 2019] ROBERT, T. (2019). *Modélisation du taux de transformation et élasticité au prix.* Mémoire d'actuariat.

[VICAIRE, 2017] VICAIRE, E. (2017). *L'open data et les réseaux neuronaux : vers une amélioration de la prédictibilité des sinistres.* Mémoire d'actuariat.