

Mémoire présenté devant l'ENSAE Paris  
pour l'obtention du diplôme de la filière Actuariat  
et l'admission à l'Institut des Actuaire  
le 16/10/2022

Par : **Eugénie Rinelle TOUKAM**

Titre : **Clients multi-équipés : étude du comportement  
de résiliation et amélioration de la politique  
de renouvellement.**

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

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

*Membres présents du jury de la filière*

*Entreprise : AXA France*

*Nom : Pierre PICARD*

*Signature :*

**AXA France I.A.R.D.**  
Société Anonyme au Capital de 214 799 030 €  
Entreprise régie par le Code des Assurances  
Siège social : 313, Terrasses de l'Arche  
92727 NANTERRE CEDEX  
722 057 460 RCS Nanterre

*Membres présents du jury de l'Institut  
des Actuaire*

*Directeur du mémoire en entreprise :*

*Nom : Gérald LUCAS*

*Signature :*

*glucas*

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

Secrétariat :

Signature du responsable entreprise

*glucas*

Bibliothèque :

Signature du candidat

*Emp*

Mémoire présenté devant l'ENSAE Paris  
pour l'obtention du diplôme de la filière Actuariat  
et l'admission à l'Institut des Actuaire  
le 16/10/2022

Par : **Eugénie Rinelle TOUKAM**

Titre : **Clients multi-équipés : étude du comportement  
de résiliation et amélioration de la politique  
de renouvellement.**

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

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

*Membres présents du jury de la filière*

*Entreprise : AXA France*

*Nom : Pierre PICARD*

*Signature :*

*Membres présents du jury de l'Institut  
des Actuaire*

*Directeur du mémoire en entreprise :*

*Nom : Gérald LUCAS*

*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)**

Secrétariat :

Signature du responsable entreprise

Bibliothèque :

Signature du candidat

*A mon Papa parti le 30/01/2022.*

# Résumé.

La fidélisation de la clientèle est une problématique clé des assureurs. Elle est encore plus cruciale pour les clients multi-équipés qui ont la particularité de détenir plusieurs contrats d'assurance auprès d'un même assureur. Au sein d'AXA France les politiques de renouvellement des contrats se font par ligne de business (branche par branche), ainsi les spécificités des clients multi-équipés ne sont pas toujours prises en compte. Or ces derniers détiennent une grande part du portefeuille et la non-satisfaction sur une branche mineure est susceptible d'entraîner une résiliation complète de l'ensemble du portefeuille d'un client multi-équipé. L'une des branches les plus multi-équipées du périmètre entreprises d'AXA est la responsabilité civile avec un taux de multi-équipement d'environ 56% en 2021. Elle est également l'une des branches les plus rentables d'où le besoin de fidélisation.

Dans ce travail, nous nous sommes appesantis sur l'étude des comportements de résiliation des clients multi-équipés et l'amélioration des politiques de renouvellement dans le but de prendre en compte efficacement le multi-équipement. A cet effet, les modèles logistique, CART et Random Forest ont été utilisés pour modéliser la résiliation. Le taux majoration client a été identifié comme l'un des principaux facteurs de résiliation. Partant de ce constat la deuxième partie du travail s'est consacrée à la compréhension de la majoration des clients multi-équipés sujets à la résiliation. Cette problématique a été résolue en utilisant la distance de Levenshtein pour retrouver les jumeaux de chaque potentiel résilieur. L'algorithme utilisé pour rechercher les jumeaux consistait à créer une variable concaténant les caractéristiques de chaque client, puis à calculer la distance de Levenshtein entre les clients en utilisant cette nouvelle variable et enfin, à déduire les jumeaux qui correspondaient aux clients ayant la plus faible distance. L'analyse minutieuse des résiliés et de leurs jumeaux non résiliés a permis de mettre en place un ensemble d'alertes devant être utilisé par AXA pour les renouvellements futurs afin de fidéliser davantage la clientèle multi-équipée.

**Mots clés :** *Fidélisation, Majoration, Multi-équipement, Renouvellement, Responsabilité civile.*

# Abstract.

Customer loyalty is a key issue for insurers, and it is even more crucial for multi-equipment customers, who have the particularity of holding several insurance contracts with the same insurer. At AXA France, contract renewal policies are based on lines of business (branch by branch), so the specific characteristics of multi-equity clients are not always taken into account. This means that the specific characteristics of multi-equity clients are not always taken into account. However, multi-equity clients account for the bulk of the portfolio and, moreover, failure to satisfy a minor line of business is likely to result in the termination of the entire portfolio of a multi-equity client. One of the most multi-equipped lines of business in AXA's corporate perimeter is civil liability, with a multi-equipment rate of approximately 56% in 2021, it is also one of the most profitable lines of business. Thus, it is essential to increase their loyalty.

Therefore, we have focused on the study of the cancellation behavior of multi-equipped customers and the improvement of renewal policies in order to effectively take into account the multi-equipment. To this end, logistic, CART and Random Forest models were used to model termination of contracts. It revealed that the customer mark-up rate is one of the main factors of termination. Based on this observation, the second part of the work was devoted to understanding the markup of multi-equipped customers subject to terminate. This problem was solved by using the Levenshtein distance to find the twins of each potential chaser. The algorithm used to find the twins consisted of creating a variable concatenating the characteristics of each customer, then calculating the Levenshtein distance between the customers using this new variable and finally, deducing the twins that corresponded to the customers with the lowest distance. The careful analysis of the terminated customers and their unterminated twins allowed us to set up a set of alerts to be used by AXA for future renewals in order to increase the loyalty of the multi-equipped customers.

**Key words** : *Loyalty, Surcharge, Multi-equipment, Renewal, Civil liability.*

# Remerciements.

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

- Tout d’abord, je tiens à remercier profondément Gérard LUCAS mon directeur de mémoire pour sa disponibilité, son encadrement, ses conseils qui m’ont permis de donner le meilleur de moi pour l’accomplissement de ce travail.
- Merci à Véronique MARPILLAT, responsable de l’équipe Actuariat Produit IARD Entreprises d’AXA France de m’avoir accueilli au sein de son équipe, et d’avoir témoigné une grande confiance dans la réalisation de mes travaux.
- Merci à chacun des membres de l’équipe Actuariat Produit IARD Entreprises d’AXA France pour votre bienveillance et votre joie de vivre qui m’a accompagné au long de mon alternance.
- Je remercie Pierre PICARD mon tuteur académique ENSAE pour sa disponibilité et ses recommandations qui ont grandement contribué à l’amélioration de ce travail.
- Je remercie Christine AGBOTSE et Marie Thérèse MBIA NDI pour avoir apporté un second regard sur ce mémoire.
- Je finirais par remercier ma famille, principalement mon mari Jacques MANGUELE pour leur soutien sans faille pendant les périodes difficiles traversées au cours de ce mémoire. Leur amour pour moi n’est plus à démontrer!!!

# Table des matières

|                                                                                                                                                                |            |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>Dédicace</b>                                                                                                                                                | <b>i</b>   |
| <b>Résumé.</b>                                                                                                                                                 | <b>i</b>   |
| <b>Abstract.</b>                                                                                                                                               | <b>ii</b>  |
| <b>Remerciements.</b>                                                                                                                                          | <b>iii</b> |
| <b>INTRODUCTION</b>                                                                                                                                            | <b>2</b>   |
| <b>1 PRÉSENTATION DU CADRE DE L'ÉTUDE.</b>                                                                                                                     | <b>4</b>   |
| 1.1 AXA : un leader du secteur de l'assurance. . . . .                                                                                                         | 4          |
| 1.2 Politique de renouvellement des contrats d'assurance au sein d'AXA France<br>IARD Entreprises. . . . .                                                     | 5          |
| 1.3 Qu'est ce que le multi-équipement ? . . . . .                                                                                                              | 6          |
| 1.4 Qu'est ce que la responsabilité civile Entreprise ? Et quels sont les produits d'as-<br>surance d'AXA IARD Entreprises multi-équipés avec elle ? . . . . . | 7          |
| 1.4.1 La responsabilité civile Entreprise (RC). . . . .                                                                                                        | 7          |
| 1.4.2 Présentation des branches des clients RC multi-équipés. . . . .                                                                                          | 9          |
| 1.5 Conclusion partielle. . . . .                                                                                                                              | 10         |
| <b>2 CONSTRUCTION ET DESCRIPTION DE LA BASE DE DONNÉES.</b>                                                                                                    | <b>11</b>  |
| 2.1 Construction de la base de données. . . . .                                                                                                                | 11         |
| 2.2 Présentation des variables de la base finale. . . . .                                                                                                      | 13         |
| 2.2.1 Description des variables de la base finale. . . . .                                                                                                     | 13         |
| 2.2.2 Analyse des variables communes aux clients mono-équipés et multi-équipés.                                                                                | 13         |
| 2.2.3 Analyse des variables spécifiques aux clients multi-équipés. . . . .                                                                                     | 20         |
| 2.3 Analyse croisée des variables. . . . .                                                                                                                     | 25         |
| 2.3.1 Corrélation entre les variables catégorielles. . . . .                                                                                                   | 25         |
| 2.3.2 Corrélation entre les variables continues. . . . .                                                                                                       | 27         |
| 2.4 Traitement des données. . . . .                                                                                                                            | 29         |
| 2.4.1 Traitement des valeurs manquantes. . . . .                                                                                                               | 29         |
| 2.4.2 Création de nouvelles variables. . . . .                                                                                                                 | 30         |
| 2.4.3 Choix des variables de modélisation. . . . .                                                                                                             | 31         |

|          |                                                                                     |           |
|----------|-------------------------------------------------------------------------------------|-----------|
| 2.5      | Conclusion partielle. . . . .                                                       | 32        |
| <b>3</b> | <b>MODÉLISATION DE LA RÉSILIATION DES CLIENTS EN RESPON-</b>                        |           |
|          | <b>SABILITÉ CIVILE ENTREPRISE.</b>                                                  | <b>33</b> |
| 3.1      | Approche méthodologique de la modélisation de la résiliation. . . . .               | 33        |
| 3.1.1    | Modèle de régression logistique binaire. . . . .                                    | 33        |
| 3.1.2    | Les arbres de décision. . . . .                                                     | 35        |
| 3.1.3    | Random forest : une méthode d'agrégation des arbres de décision. . . . .            | 38        |
| 3.1.4    | Appréciation et interprétation des modèles. . . . .                                 | 39        |
| 3.2      | Présentation des résultats. . . . .                                                 | 46        |
| 3.2.1    | Modélisation de la résiliation par un modèle logistique. . . . .                    | 46        |
| 3.2.2    | Présentation des modèles de résiliation implémentés part CART et Ran-               |           |
|          | dom forest. . . . .                                                                 | 61        |
| 3.3      | Conclusion partielle. . . . .                                                       | 63        |
| <b>4</b> | <b>OPTIMISATION DES POLITIQUES DE RENOUVELLEMENT DES CLIENTS</b>                    |           |
|          | <b>RC MULTI-ÉQUIPÉS.</b>                                                            | <b>65</b> |
| 4.1      | Approche méthodologique de la distance de Levenshtein. . . . .                      | 65        |
| 4.1.1    | Définition de la Distance de Levenshtein. . . . .                                   | 66        |
| 4.1.2    | Algorithme de recherche des jumeaux des potentiels résiliers . . . . .              | 67        |
| 4.1.3    | Principe d'évaluation de la majoration optimale entre les potentiels rési-          |           |
|          | liers et leurs jumeaux. . . . .                                                     | 68        |
| 4.2      | Distance de Levenshtein : de la recherche des jumeaux à l'optimisation de la        |           |
|          | majoration. . . . .                                                                 | 68        |
| 4.2.1    | Première approche : utilisation de la distance de Levenshtein non pondérée. . . . . | 69        |
| 4.2.2    | Deuxième approche : prise en compte des corrélations entre les variables            |           |
|          | et le taux de majoration. . . . .                                                   | 71        |
| 4.3      | Interprétation des résultats et importance opérationnelle de l'approche de Le-      |           |
|          | venshtein. . . . .                                                                  | 73        |
| 4.3.1    | Interprétation des résultats . . . . .                                              | 73        |
| 4.3.2    | Importance opérationnelle de l'approche de Levenshtein. . . . .                     | 77        |
|          | <b>CONCLUSION.</b>                                                                  | <b>79</b> |
|          | <b>Abréviations.</b>                                                                | <b>81</b> |
|          | <b>Bibliographie</b>                                                                | <b>82</b> |
|          | <b>Annexe A Base de données et modélisation.</b>                                    | <b>84</b> |
| A.1      | Base de données. . . . .                                                            | 84        |
| A.2      | Analyse des variables et modélisation de la résiliation. . . . .                    | 87        |



|                                                                       |           |
|-----------------------------------------------------------------------|-----------|
| A.3 Caractérisation des prédictions du modèle de résiliation. . . . . | 88        |
| <b>Annexe B Modèles linéaires généralisés.</b>                        | <b>89</b> |
| <b>Note de synthèse.</b>                                              | <b>92</b> |
| <b>Executive summary.</b>                                             | <b>98</b> |

# INTRODUCTION

Le multi-équipement renvoie à deux phénomènes distincts en marketing selon notamment qu'on évoque le domaine des biens d'équipement ou celui des services. Dans le domaine des biens d'équipement, le multi-équipement désigne le phénomène par lequel un foyer est équipé de plusieurs exemplaires d'un même bien qui à l'origine était "normalement" détenu de manière unique au sein du foyer. On peut ainsi parler de multi-équipement pour un téléviseur, un réfrigérateur, un ordinateur portable, etc. Par ailleurs, dans le domaine des services et notamment dans celui de l'assurance, le multi-équipement désigne le phénomène par lequel un client possède plusieurs produits ou plusieurs contrats auprès d'un même assureur. Un client d'AXA France possédant à titre d'exemple un contrat d'assurance automobile et un contrat d'assurance responsabilité civile sera dit multi-équipé. Au sens large, le multi-équipement désigne le phénomène par lequel un client possède plusieurs produits ou services auprès d'une même entreprise ou d'un même réseau. Souvent utilisé comme indicateur clé de pilotage et d'évaluation des performances en assurance<sup>1</sup>, le multi-équipement représente un enjeu crucial pour les entreprises car il permet de développer le potentiel des clients actuels sans en acquérir de nouveaux à un coût élevé.

Au vu de sa pertinence pour les analyses marketing, le multi-équipement a déjà fait l'objet de plusieurs études qui aboutissent à la conclusion selon laquelle le multi-équipement est un moyen efficace de fidélisation de la clientèle. En effet, en étant multi-équipés, les clients sont plus disposés à maintenir leurs relations commerciales avec l'entreprise (Reinartz et Kumar, 2003). Il contribue donc à la rentabilisation de la relation avec le client par la diminution des coûts d'acquisition<sup>2</sup>. Ceci devient d'autant plus important pour les assureurs car les coûts d'acquisition client atteignent entre 200 € et 400 €<sup>3</sup>. Ajouté à ces coûts d'acquisition très élevés, la forte concurrence dans le secteur de l'assurance entraîne une croissance des attentes clients de plus en plus forte ; ainsi la fidélisation client devient un axe pertinent à explorer et en particulier la fidélisation des multi-équipés. Ngobo (2004) montre que les variables affectant le

---

1. <https://www.definitions-marketing.com/definition/multi-equipement/>

2. Le coût d'acquisition d'un client est le prix moyen payé ou investi par une entreprise ou une organisation pour convertir une cible potentielle en un client.

3. [http://archives.lesechos.fr/archives/cercle/2014/10/31/cercle\\_117024.htm](http://archives.lesechos.fr/archives/cercle/2014/10/31/cercle_117024.htm)

comportement de multi-équipement sont d'ordre marketing, attitudinales ou encore sociodémographiques. Toutefois, on est sans ignorer que le prix reste un critère fondamental d'acquisition et de fidélisation de la clientèle, ainsi les décisions de retarification en assurance doivent se faire de façon plus minutieuse pour les clients multi-équipés. En effet, une non-satisfaction d'un client sur une branche mineure peut l'inciter à résilier l'ensemble de ses contrats.

Au sein d'AXA France, l'une des branches les plus multi-équipées du périmètre entreprise est la responsabilité civile avec un taux d'environ 56% en 2021. Fidéliser la clientèle de la branche responsabilité civile revient à bien connaître ses clients ainsi que les produits d'assurance qu'ils possèdent, ceci afin d'ajuster les décisions de retarification et optimiser la conservation des bons clients. En effet, à ce jour, la tarification des affaires nouvelles et les renouvellements au sein d'Axa France n'intègrent pas la composante du multi-équipement, ainsi on se demande si cela n'affecte pas la rentabilité du portefeuille ? en d'autres termes, est-ce que la prise en compte du multi-équipement dans les politiques de renouvellement conduit à une fidélisation de la clientèle ? L'objectif de notre mémoire est d'analyser la rentabilité IARD entreprises via le prisme complet du client pour proposer une politique tarifaire et de renouvellement optimisée qui valorise davantage le multi-équipement, pour ainsi améliorer la rentabilité et la fidélisation des clients du portefeuille. Pour atteindre nos objectifs, nous allons suivre le canevas ci-dessous :

1. Présentation de cadre de l'étude ;
2. Construction de la base de données et aperçu général sur les données ;
3. Modélisation de la résiliation de notre portefeuille et identification des profils de résiliation ;
4. Optimisation de la majoration des clients résiliés ;
5. Conclusion et axes d'améliorations de l'étude.

# Chapitre 1

## PRÉSENTATION DU CADRE DE L'ÉTUDE.

Dans ce premier chapitre, il est question de situer le cadre du travail partant de la présentation d'AXA à la description de la problématique, et ceci en passant par un aperçu général sur l'ensemble des produits d'assurance utilisés dans cette étude.

### 1.1 AXA : un leader du secteur de l'assurance.

Créé en 1994 suite à la fusion de plusieurs sociétés d'assurance, AXA est un groupe international français spécialisé dans l'assurance et la gestion d'actifs. Proposant aux clients particuliers et entreprises des services en matière d'assurance, de prévoyance, d'épargne, de transmission de patrimoine et bancaire. Classé selon Forbes comme le quatrième plus grand assureur au rang mondiale<sup>1</sup> en 2022, AXA Groupe détenait en 2017 un chiffre d'affaires évalué à 100,193 milliards d'euros avec 107 millions de clients dans le monde dont 6,3 millions en France.

Comme tout grand groupe international, la structure d'AXA est complexe et est constituée de plusieurs filiales qui se caractérisent par leur position géographique et le métier exercé. En France, AXA compte environ 17 filiales dont AXA France IARD qui est la filiale au sein de laquelle mon alternance s'est déroulée et a ainsi donné lieu à la rédaction du présent travail.

#### **Zoom sur AXA France IARD.**

Avec un chiffre d'affaires de 6,5 milliards d'euros en 2021, AXA France IARD est la filiale d'AXA en France spécialisée dans les assurances dommages. Pour rappel, les assurances dommages sont des assurances qui protègent financièrement l'assuré contre des événements dommageables définis. Elles regroupent les assurances des biens et les assurances de responsabilité. Au sein d'AXA France IARD, une segmentation est faite en fonction du type du client : particulier ou entreprise. Ainsi, AXA France IARD Particuliers et AXA France IARD En-

---

1. <https://www.atlas-mag.net/article/forbes-2000-top-20-des-assureurs-les-plus-importants-en-2018>

treprises sont les deux grandes entités de la filiale qui s'occupent respectivement des clients "particuliers" et des clients "entreprises". Cette segmentation est justifiée à la fois par le fait que le volume de chiffre d'affaires géré pour les clients entreprises est très élevé ce qui entraîne une grande charge de sinistre en cas de sinistralité, mais aussi les clients entreprises sont très peu homogènes. Ainsi, la tarification des contrats d'assurance pour les entreprises se fait de façon plus minutieuse pour tenir compte des spécificités de chaque client. Par ailleurs, les lois qui régissent l'assurance des particuliers ne sont pas les mêmes que celles qui s'appliquent à l'assurance des clients entreprises. Par exemple la loi Hamon qui s'applique uniquement aux clients particuliers en leur permettant de résilier leur contrat d'assurance après 1 an d'adhésion, sans être exposé à des pénalités.

AXA France IARD Entreprises dispose d'un chiffre d'affaires qui s'estime à 3 milliards d'euros. Il fournit à l'ensemble de ses clients une large gamme de produits d'assurance partant de l'assurance des biens à l'assurance responsabilité (les contrats d'assurance responsabilité civile, les contrats d'assurance automobile, les contrats d'assurance Risque Industriel, etc). Par ailleurs, le produit responsabilité civile est l'un des produits les plus rentables du portefeuille. Par ailleurs, avec un taux de multi-équipement d'environ 56% en 2021, il est l'un des produits le plus multi-équipé du portefeuille IARD Entreprises. Au vu de ceci, ce travail va s'appesantir sur les clients de la branche responsabilité civile du portefeuille. Nous allons principalement nous intéresser aux clients multi-équipés en analysant leurs comportements de résiliation et en proposant des approches à utiliser lors des politiques de renouvellement pour fidéliser davantage ce groupe de client. Pour mieux comprendre les différents termes techniques utilisés par la suite, nous allons dans les parties suivantes définir le multi-équipement, présenter le processus de renouvellement des contrats entreprises et enfin présenter les différents produits d'AXA France IARD Entreprises en faisant un zoom sur la responsabilité civile qui est le cœur de notre étude.

## 1.2 Politique de renouvellement des contrats d'assurance au sein d'AXA France IARD Entreprises.

Les contrats d'assurance sont souscrits pour une durée limitée qui peut être mensuelle, semestrielle, annuelle, ... Au sein d'AXA France IARD Entreprises, la majorité des contrats ont une échéance annuelle. Ainsi, à l'anniversaire du contrat, le client décide s'il résilie ou non son contrat. De son côté, en fonction des objectifs fixés et de la rentabilité du contrat, AXA revalorise la prime d'assurance de l'année précédente et propose au client une nouvelle prime. Cette nouvelle prime est égale à *la prime initiale \* taux de majoration*. Avec *la prime initiale* la prime du client à l'année précédente et *taux de majoration* le pourcentage d'augmentation ou de diminution de cette prime. Le taux de majoration tient compte de la sinistralité du client, du taux d'inflation, de l'exposition au risque et des objectifs internes d'AXA.

Une fois la nouvelle prime d'assurance évaluée, elle est proposée au client et celui-ci peut

donc décider si oui ou non, il accepte la proposition : c'est le **renouvellement du contrat**. Si le client accepte le taux de majoration proposé, son contrat est directement renouvelé. S'il refuse, des négociations peuvent se faire avec le souscripteur afin de trouver un nouveau taux de majoration qui satisfait les deux parties.

### **Comment le taux de majoration est évalué ?**

Comme indiqué précédemment, le taux de majoration appliqué à un contrat dépend principalement de la sinistralité du client, du taux d'inflation, de l'exposition au risque et des objectifs internes fixés par AXA. Concernant l'exposition au risque, un processus de surveillance de la qualité du portefeuille est mis sur pied par AXA. Ce processus consiste à surveiller mensuellement la rentabilité du portefeuille afin de détecter très rapidement les grandes expositions et d'anticiper la sur-majoration à appliquer à ces contrats. Implicitement la surveillance du portefeuille, c'est donc l'identification automatique des contrats non-rentables sur la base de critères nationaux de la direction technique. Une fois un contrat détecté, s'en suit un traitement qui consiste à analyser l'état du contrat et la situation du client, puis à saisir des actions correctives ou préventives en proposant à ce client un taux de majoration proportionnel à son risque. L'objectif de ce processus de surveillance de portefeuille est d'anticiper sur les risques à venir et également de retrouver une rentabilité technique pour l'ensemble des clients.

## **1.3 Qu'est ce que le multi-équipement ?**

Comme présenté en introduction , le multi-équipement désigne le phénomène par lequel un client possède plusieurs produits ou services auprès d'une même entreprise ou d'un même réseau. Il est souvent utilisé comme indicateur clé de pilotage et d'évaluation des performances en assurance et représente un enjeu crucial pour les entreprises, car il permet de développer le potentiel des clients actuels sans en acquérir des nouveaux à un coût élevé. En assurance, les coûts d'acquisition varient entre 200 € et 400 €<sup>2</sup>, d'où l'importance de mieux maîtriser le comportement des clients multi-équipés afin de les fidéliser davantage. Ceci dit, de quoi est constitué le multi-équipement de la branche responsabilité civile entreprises d'AXA France IARD Entreprises ? En d'autres termes, quelles sont les branches détenues par les clients multi-équipés ?

---

2. [http://archives.lesechos.fr/archives/cercle/2014/10/31/cercle\\_117024.htm](http://archives.lesechos.fr/archives/cercle/2014/10/31/cercle_117024.htm)

## 1.4 Qu'est ce que la responsabilité civile Entreprise ? Et quels sont les produits d'assurance d'AXA IARD Entreprises multi-équipés avec elle ?

Le périmètre IARD Entreprises d'AXA France propose à l'ensemble de ses clients de multiples contrats en fonction de leurs besoins. On peut citer ici les contrats d'assurance responsabilité civile, les contrats d'assurance automobile, les contrats d'assurance Risque Industriel, etc. Notre travail consiste en l'analyse du multi-équipement des clients de la branche responsabilité civile. Cette étude pourra par la suite s'étendre aux autres branches IARD en adoptant la même démarche méthodologique. Dans cette partie, sera présentée premièrement la responsabilité civile, ensuite l'ensemble des branches du périmètre IARD Entreprises multi-équipées avec la responsabilité civile.

### 1.4.1 La responsabilité civile Entreprise (RC).

L'assurance responsabilité civile des entreprises vise à réparer les dommages corporels, matériels ou immatériels causés à un des tiers ou des clients. La responsabilité civile d'une entreprise concerne également les préjudices qui pourraient intervenir après la livraison d'un bien ou d'un service. Il existe donc de nombreuses situations dans lesquelles la responsabilité de l'entreprise peut être engagée. Au regard de sa complexité et des différents périmètres qu'elle couvre, le pilotage de la branche RC Entreprises au sein d'AXA France est organisé autour de plusieurs segments d'activités. Le tableau 1.1 donne la segmentation de la branche RC.

TABLEAU 1.1 – Segmentation de la branche RC.

| CODE SEGMENT | DESCRIPTION SEGMENT                                                   |
|--------------|-----------------------------------------------------------------------|
| S01          | Industrie, commerce, Eaux et déchets, BTP (Bâtiment Travaux Publics). |
| S02          | Prestataires de services, Enseignement et Formation.                  |
| S03          | Collectivité locales.                                                 |
| S04          | Responsabilité Du Dirigeant (RDD).                                    |
| S06.1        | Sports et loisirs hors chasse.                                        |
| S06.2        | Chasse.                                                               |
| S07          | Transports et services.                                               |
| S08          | Professions réglementaires.                                           |
| S09          | RC médicale.                                                          |
| S10          | RC Autres.                                                            |

**Source : Interne AXA.**

Notons que la branche RC Entreprises est considérée comme une branche d'intensité : ce n'est pas la fréquence de sinistralité qui la caractérise, mais plutôt l'intensité des sinistres. Ainsi, elle fait partie de la famille des sinistres à développement long, même si en pratique une grande

partie des sinistres sont à déroulement court (un an). Pour les sinistres les plus graves, il faut parfois attendre plusieurs années avant la liquidation (actions en justice, sinistres corporels).

AXA France IARD Entreprises gère principalement les clients entreprises et donc pour des besoins de logistique et de comptabilité, ces clients logent le plus souvent l'ensemble de leurs contrats d'assurance chez un même assureur ainsi le taux de multi-équipement est souvent très élevé pour ces types de clients. À titre d'exemple, en 2021 le taux de multi-équipement pour la branche responsabilité civile était d'environ 56% ce qui signifie que près de 3 clients sur 5 de la branche responsabilité civile ont au moins un deuxième contrat au sein du portefeuille.

Au sein du portefeuille d'AXA France IARD Entreprises, la part de la RC est de 450 millions d'euros, soit environ 15% du chiffre d'affaires global. Les principales branches des clients RC multi-équipés sont l'automobile, le risque industriel, le risque technique, l'immeuble, la construction et le transport terrestre. La figure 1.1 nous donne la part de chacune de ses branches dans la prime d'assurance de l'ensemble des clients RC. On constate qu'avec une proportion de 28.9%, la branche automobile est la branche majoritaire des clients RC multi-équipés.

FIGURE. 1.1 – Multi-équipement RC 2020 : Répartition de la prime AXA.

|                             |                                   |                                 |                           |
|-----------------------------|-----------------------------------|---------------------------------|---------------------------|
| AUTOMOBILE<br>28,9% ; 234M€ | RC Multi-équipé<br>22,22% ; 180M€ | RI<br>20,75% ; 168M€            | CONST<br>11,28% ;<br>91M€ |
|                             |                                   | RC Mono-équipé<br>11,12% ; 90M€ | RT<br>3,7% ;<br>30M€      |
|                             |                                   |                                 | TT 2,06% ;<br>17M€        |

**Remarque :** RI : Risque industriel ; CONST : Construction ;

RT : Risque Technique ; TT : Transport Terrestre.

Prime globale du portefeuille : 806M €.

N.B. : Entre le 450M€ (chiffre d'affaires global de la RC) et la somme des primes RC mono-équipé + multi-équipé, la différence vient des contrats en coassurance.

**Source : Interne AXA.**



## 1.4.2 Présentation des branches des clients RC multi-équipés.

### a) L'automobile.

L'assurance automobile a pour principale fonction de prévoir l'indemnisation des dommages pouvant être causés à un tiers par un conducteur et/ou l'un de ses passagers. Elle permet de couvrir les frais engendrés lors de blessures causées à un individu (frais de prise en charge hospitalière, etc.) et en cas de dégâts matériels (sur d'autres véhicules, mais aussi sur des biens immobiliers, etc.).  L'automobile fait partir de l'une des plus grandes branches de l'assurance non-vie du marché français et concentre l'essentiel de l'activité avec 23,1 milliards d'euros des encaissements et 39% de parts de marché en 2020<sup>3</sup>. Comme le montre la figure 1.1, l'assurance automobile est aussi la principale branche des clients RC Entreprises multi-équipés ceci s'explique en partie par le fait que, le code des assurances oblige tout détenteur d'un contrat d'assurance automobile d'avoir la garantie RC.

### b) Les risques industriels (RI).

Le risque industriel est défini comme un évènement accidentel se produisant sur un site industriel mettant en jeu des produits et/ou des procédés dangereux et entraînant des conséquences immédiates graves pour le personnel, les riverains, les biens et l'environnement<sup>4</sup>. L'assurance risque industriel est ainsi un contrat d'assurance multirisque destiné à l'assurance de tous locaux professionnels utilisés par des ateliers, des unités de fabrication ou de transformation. Elle couvre l'assuré notamment contre les incendies, dégâts des eaux, vols ou tentatives de vol, brise de glace, actes de vandalisme, catastrophes naturelles, mais aussi et surtout contre les dommages liés à la production. L'objectif principal de cette assurance est de couvrir l'entreprise en cas de sinistre, tant sur le plan matériel que financier. Au sein du portefeuille d'AXA France IARD Entreprises, l'assurance risques industriels représente 21% du multi-équipement des clients de la branche RC entreprises.

### c) Les risques techniques (RT).

L'assurance risque technique cible principalement les maîtres d'ouvrage, les propriétaires, locataires, détenteurs ou loueurs de tout type d'équipement (les engins de chantier, les machines de production, les matériels informatiques, matériels et appareils électroniques, ...). Avec 3,7% de primes, il est le quatrième produits le plus détenu par les clients RC entreprises multi-équipés.

### d) La construction.

L'assurance construction est un contrat obligatoire auquel le constructeur doit souscrire pour couvrir les risques liés à la construction. Celle-ci comprend l'assurance "dommages ou-

---

3. <https://www.atlas-mag.net/article/le-marche-d-assurance-dommages-en-france>

4. <https://assurance-risques-industriels.com/>

vrages" s'appliquant après la réception de l'ouvrage et se terminant avec la garantie décennale, l'assurance "responsabilité civile professionnelle" permettant de couvrir les conséquences des fautes d'une activité professionnelle. Cette assurance a une logique de protection du propriétaire, elle détermine les garanties liées au contrat souscrit et sécurise le consommateur contre les dommages importants jusqu'à 10 ans après la livraison des travaux. Tous les constructeurs sont concernés par l'assurance construction (entrepreneur, architecte, etc.) du moment qu'ils sont liés au maître de l'ouvrage, ils bénéficient d'une couverture lors de graves dommages (solidité du bien) et offrent aux consommateurs la sécurité de financer des réparations nécessaires. 11% des clients RC multi-équipés possèdent au moins un contrat de cette branche.

#### **e) Le transport terrestre (TT).**

Le contrat d'assurance transport terrestre (TT) est un produit conçu pour procurer aux acteurs du transport une protection efficace contre le risque de perte et de dommages. Avec une part de 2%, c'est l'une des branches les moins détenues par les clients RC multi-équipés.

## **1.5 Conclusion partielle.**

Ce premier chapitre avait pour but de présenter le cadre général du travail. Ceci a été fait en décrivant de façon sommaire AXA et en particulier AXA France IARD Entreprises. Ensuite, nous avons défini le multi-équipement, le processus de renouvellement et enfin la branche RC Entreprises et ses branches multi-équipées. Dans la suite, nous allons dans un premier temps présenter la base de données et ensuite analyser celle-ci afin de ressortir les premières conclusions sur notre objectif numéro un qui est celui d'étudier le comportement de résiliation des clients multi-équipés de la branche RC Entreprises.

# Chapitre 2

## CONSTRUCTION ET DESCRIPTION DE LA BASE DE DONNÉES.

Ce chapitre est consacré à la présentation des données utilisées dans le cadre de l'étude. Il commencera par une présentation du processus de construction de la base de données suivi d'une analyse uni et bi dimensionnelle des variables de la base, et enfin du traitement des données et du choix des variables pertinentes pour la suite de l'étude.

*N.B. : Par soucis de confidentialité, les résultats de ce travail ont été ajustés mais l'ensemble des conclusions ont été conservées.*

### 2.1 Construction de la base de données.

Dans cette étude, nous utilisons les données clients de la branche responsabilité civile entreprises sur la période 2019 à 2021. La base de données finale a été construite à partir des bases utilisées en interne pour les suivis d'activités, les modèles de tarification et les analyses de rentabilité. Ces bases sont les suivantes :

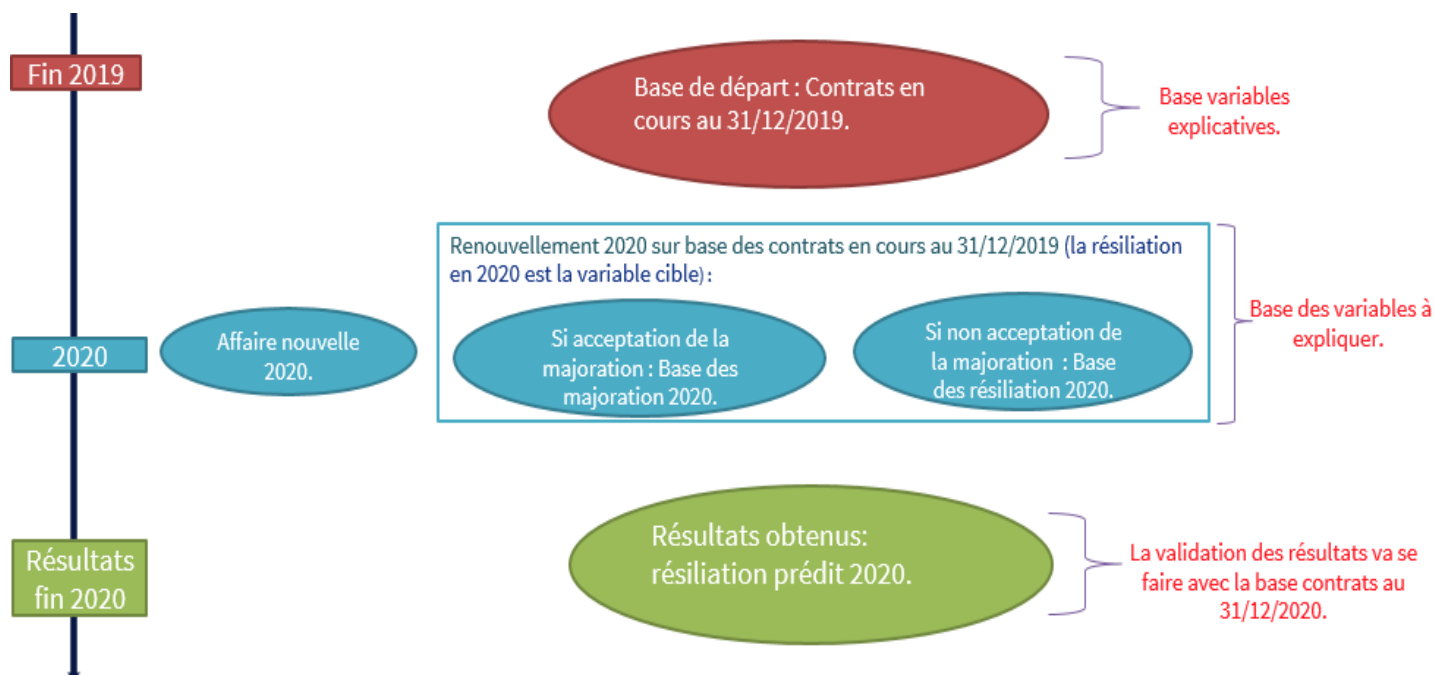
- **Base\_ACEA** : il s'agit de la base renseignant sur l'ensemble des contrats RC pour une année donnée. En 2020, la base ACEA comptait 163 484 contrats dont 15 633 résiliés au cours de cette année (soit un taux de résiliations d'environ 10%). Cette base a été utilisée comme base centrale et nous y avons ajouté les informations sur le multi-équipement, la majoration et la sinistralité en utilisant d'autres sources de données.
- **Base\_OVC** : encore appelée base outil vision client, c'est la base qui renseigne sur l'ensemble des clients entreprises. Dans cette base, a été principalement recueilli les informations personnelles de chaque client (le numéro Siret du client, l'indice,...). Ces informations ont permis de joindre la base des contrats RC et les bases des contrats des autres branches ceci afin d'avoir la composante multi-équipement.
- **Base\_Vision** : il s'agit de la base renseignant sur l'ensemble des contrats du portefeuille entreprise pour une année donnée. Un filtre a été fait sur les contrats des principales

branches des clients RC multi-équipés (automobile, risques industriels, risques techniques, construction et transport terrestre). Cette base a permis d’avoir les informations sur les autres contrats des clients RC multi-équipés, notamment leurs taux de résiliations, leurs sinistralités et leurs caractéristiques.

- **Base\_bilmaj** : c’est la base du bilan des majorations, elle donne pour chaque année les majorations appliquées à chaque contrat du portefeuille IARD Entreprise lors de la période de renouvellement.
- **Base\_RC\_QP** : c’est la base donnant la majoration définitivement appliquée aux clients qui n’ont pas accepté les majorations automatiques appliquées lors du renouvellement.
- **Base\_excl** : elle renseigne sur la majoration des contrats qui ont été exclus de la base du bilan des majorations d’une année à l’autre.
- **Base\_sin** : c’est la base sinistre, elle donne pour chaque année la sinistralité et coût du sinistre de chaque contrat du portefeuille IARD Entreprise.

Le processus de construction de la base finale d’analyse a consisté à fusionner les différentes bases ci-dessus pour obtenir une base renseignant sur la résiliation, la majoration, la sinistralité et le multi-équipement des clients de la branche RC entreprises. La figure 2.1 donne le schéma de construction de la base de données pour l’année 2020 en renseignant sur le processus de fusion des bases intermédiaires.

FIGURE. 2.1 – Schéma de construction de la base de données 2020.



Source : Interne AXA.

Au final obtenu la base de données obtenue comportait environ 353 000 contrats (chaque

contrat étant décrit par 37 variables) sur les années 2019 à 2021 présentait les caractéristiques suivantes :

TABLEAU 2.1 – Aperçu général de la base de données.

| Année        | Nombre de contrats RC | Primes RC en M€ | Taux de résiliation | Taux de majoration RC | Taux de multi-équipement |
|--------------|-----------------------|-----------------|---------------------|-----------------------|--------------------------|
| 2019         | 113 397               | 266             | 10,63%              | 2,25%                 | 42,88%                   |
| 2020         | 119 258               | 282             | 10,29%              | 2,40%                 | 55,45%                   |
| 2021         | 120 317               | 288             | 9,97%               | 3,35%                 | 56,20%                   |
| <b>Total</b> | <b>352 972</b>        | <b>836</b>      | <b>10,29%</b>       | <b>2,68%</b>          | <b>51,67%</b>            |

Source : Interne AXA.

Sur les trois années d'études, près d'un client sur deux de la branche RC est multi-équipé (tableau 2.1) ; toutefois, le phénomène du multi-équipement augmente au fil des années. En effet il est passé d'un taux de 43% en 2020 à un taux de 56% en 2021 soit une augmentation d'environ 13 . De plus, le taux de majoration des contrats RC a largement augmenté avec le temps et à l'inverse le taux de résiliation a diminué passant de 10,63% en 2019 à 9,97% en 2021. Pour mieux comprendre l'ensemble du portefeuille, nous allons par la suite analyser l'ensemble des variables de la base ainsi que les corrélations existant entre elles afin d'avoir une vision plus approfondie de notre portefeuille.

## 2.2 Présentation des variables de la base finale.

### 2.2.1 Description des variables de la base finale.

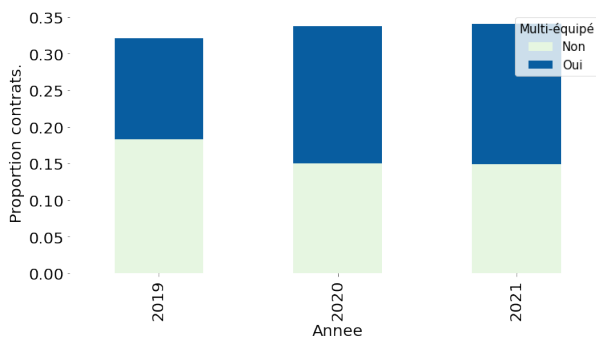
Le tableau A.1 (annexe A) présente la liste des variables de la base finale. Les variables en **couleur bleue** sont les variables qui ont été construites à l'aide des variables initiales. Nous expliquerons par la suite le processus de construction de chacune d'elles et nous présenterons également la répartition des variables de la base de données suivant le volume de primes RC et l'année d'étude. Ceci permettra dans un premier temps de voir si la composition du portefeuille a considérablement changé entre 2019 et 2021, puis de détecter les clients qui ont la plus grande importance en terme de primes dans le portefeuille.

### 2.2.2 Analyse des variables communes aux clients mono-équipés et multi-équipés.

Pour commencer, est présenté pour chaque année d'étude le nombre de contrats RC du portefeuille, le volume de primes et la résiliation. Le volume de primes RC et le nombre de contrats RC ont légèrement augmenté entre 2019 et 2021 (figures 2.2 et 2.3). Quant au multi-équipement ( $\frac{\text{Nombre de contrats multiequipés}}{\text{Nombre de contrats total}}$ ), il a augmenté d'environ 13 points partant d'environ

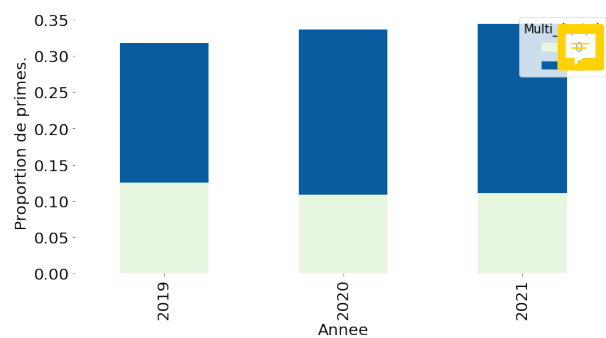
43% en 2019 à 56% en 2021. Par ailleurs, lorsqu'on s'intéresse au multi-équipement vision prime ( $\frac{\text{Prime des contrats multiequipés}}{\text{Prime total}}$ ), on a une augmentation plutôt de 7 points (de 61% en 2019 à 68% en 2021). Ces résultats montrent d'une part que le taux de multi-équipement est en croissance progressive au sein du portefeuille et d'autre part que ce sont ces derniers qui détiennent la grande partie des primes RC (près de 70%). Avec ce constat, la présente étude, qui vise à proposer des politiques tarifaires pour la fidélisation de ces clients multi-équipés, devient très intéressante.

FIGURE. 2.2 – Répartition du nombre de contrats et multi-équipement par année.



Source : Interne AXA.

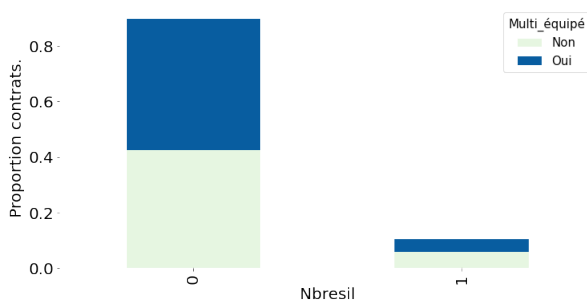
FIGURE. 2.3 – Répartition de la prime RC et du multi-équipement par année.



Source : Interne AXA.

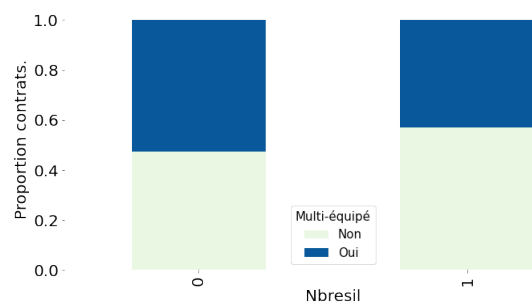
L'indicateur de la résiliation est noté dans notre base **nbresil** il vaut 0 si le contrat n'a pas été résilié au cours de l'année d'étude ; et 1 si le contrat a été résilié au cours de l'année d'étude. Au sein de notre portefeuille RC, on a un taux de résiliation d'environ 10% (figure 2.4). Par ailleurs, les clients mono-équipés ont un taux de résiliation plus élevé que les multi-équipés (figure 2.5).

FIGURE. 2.4 – Répartition des clients mono et multi-équipé suivant la résiliation.



Source : Interne AXA.

FIGURE. 2.5 – Répartition des clients mono et multi-équipé suivant la résiliation (base 100).



Source : Interne AXA.

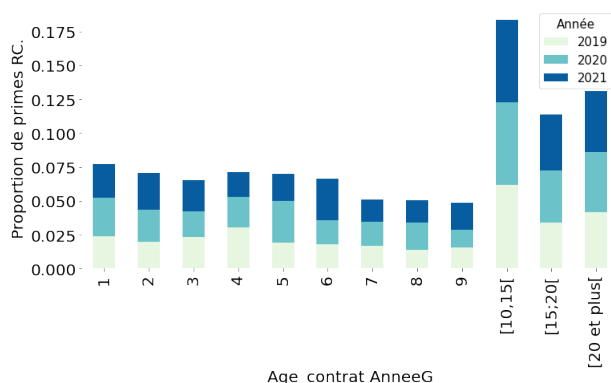
Dans la suite des analyses, sera passé en revue l'ensemble des variables. Nous allons principalement nous attarder sur la répartition des primes RC au sein de chaque variable suivant les années d'étude (2019, 2020 et 2021) et du comportement de résiliation. Ceci permettra de comprendre si la composition du portefeuille est restée homogène au cours du temps et également

d'identifier les variables qui expliquent le mieux la résiliation des clients RC .

#### a) Âge contrat en Année.

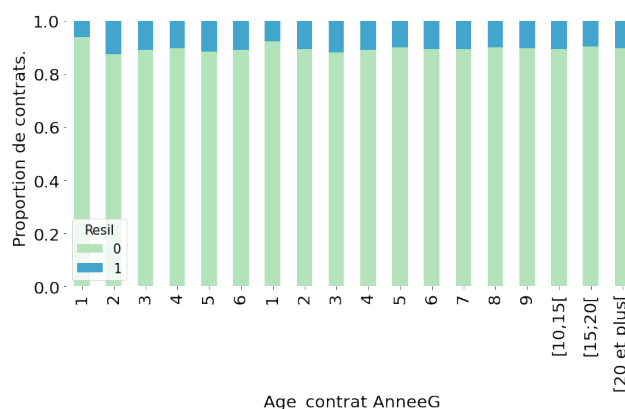
L'âge contrat en année renseigne sur l'ancienneté du client, c'est-à-dire la différence entre l'année d'étude et l'année de souscription du contrat. La figure 2.6 donne la répartition de cette variable suivant les années d'étude . On constate que de 2019 à 2021, la répartition de l'âge des contrats en portefeuille est restée pratiquement la même. On note également que le portefeuille est relativement âgé, car environ 50% des contrats ont plus de 5 ans d'âge. Par ailleurs, l'analyse de la répartition de la résiliation en fonction de l'ancienneté du client (voir figure 2.7) montre qu'à la première année du contrat, le taux de résiliation est plus faible.

FIGURE. 2.6 – Répartition de l'âge du contrat suivant l'année d'étude et la prime.



Source : Interne AXA.

FIGURE. 2.7 – Âge contrat vs Résiliation



Source : Interne AXA.

#### b) Indice.

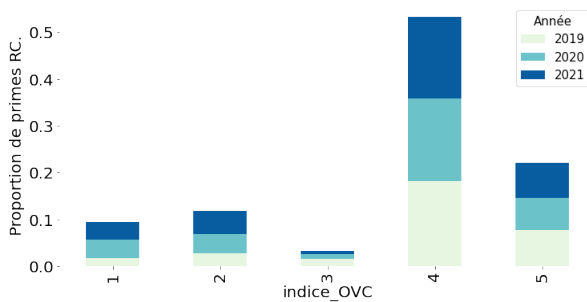
L'indice est un indicateur qui positionne sur une échelle de 1 à 5 la valeur économique consolidée des contrats IARD Entreprise d'un client. Cet indice est calculé sur la base des résultats historiques et projetés des dits contrats, en appliquant si nécessaire une mutualisation d'expérience au sein du segment de risque de chaque contrat. La construction de cet indice prend en compte la valeur du client en euro et en unité de prime, et la valeur du distributeur. Plus spécifiquement :

- Indice = 1 : représente 5% des moins bons clients ;
- Indice = 5 : correspond à 5% des meilleurs clients ;
- Indice = 3 : centré en € autour de 0, sur la valeur client (entre ] - 200; 200[) ;
- Indice = 2 : correspond aux reste des clients de valeur inférieure à -200 euros ;
- Indice = 4 : correspond aux reste des clients de valeur supérieure à 200 euros.

Cet indice aide principalement à prendre en compte de l'importance du client pour AXA et à déterminer le niveau de souplesse par rapport à la majoration préconisée. La figure 2.8 donne la

répartition de l'indice dans notre portefeuille. On remarque que la proportion de clients d'indice 1 et 2 a augmenté entre 2019 et 2021. De plus, environ 25% de prime est détenue par les clients d'indice 1 et 2 ce qui est assez élevé, car ces clients sont en général ceux susceptibles d'avoir des pertes énormes donc on devrait minimiser leur présence dans le portefeuille pour minimiser les risques. Pour ce qui est de la résiliation au sein du portefeuille, la figure 2.9 donnant la proportion des clients résiliés suivant l'indice montre que le taux de résiliation est sensiblement le même pour les 5 catégories de clients.

FIGURE. 2.8 – Répartition de l'indice suivant l'année d'étude et la prime.



Source : Interne AXA.

FIGURE. 2.9 – Indice vs Résiliation



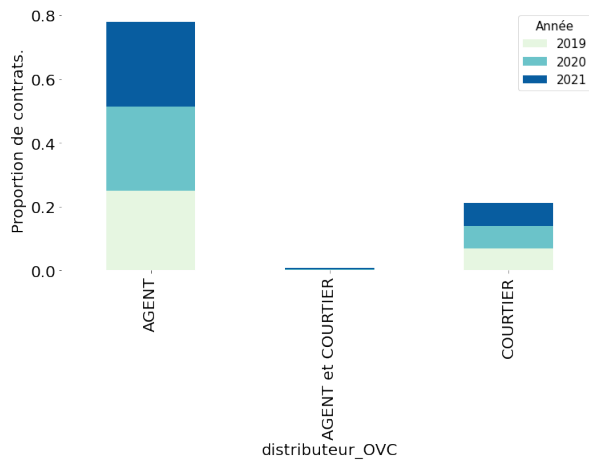
Source : Interne AXA.

### c) Distributeur.

La distribution des produits d'assurance d'AXA est réalisée par les Agents ou les Courtiers. Au sein de notre portefeuille RC, on constate comme l'indique la figure 2.10 qu'en termes de nombre de contrats, les agents ont plus de 3 fois plus de clients que les courtiers. Par contre, lorsqu'on regarde la prime (figure 2.11), environ 45% est détenue par les courtiers. Ce qui signifie que les courtiers assurent principalement les gros clients du portefeuille. Notons toutefois qu'entre 2019 et 2021, la part agent courtiers est restée la même. Concernant la résiliation, la figure 2.12 montre que les clients gérés uniquement par les agents résilient moins que les autres.

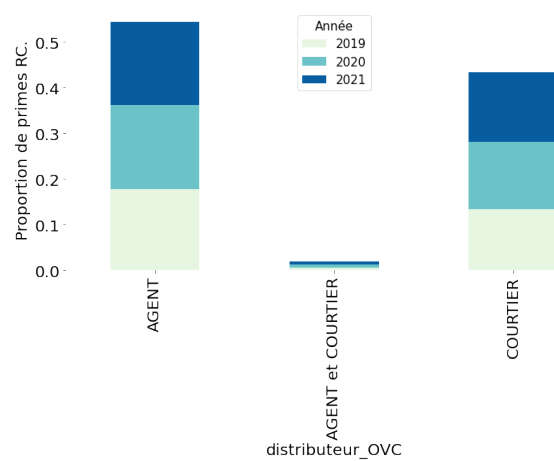


FIGURE. 2.10 – Répartition du nombre de contrats selon le distributeur et par année.



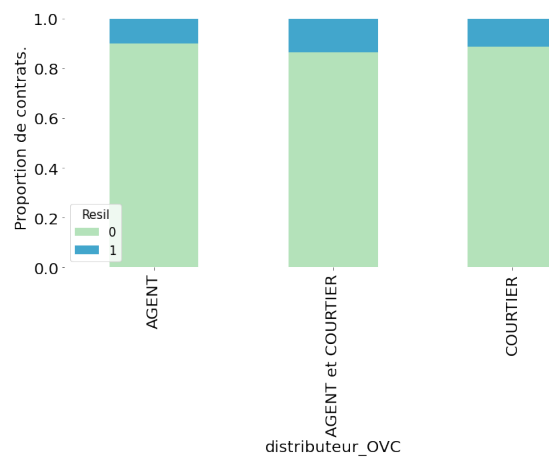
Source : Interne AXA.

FIGURE. 2.11 – Répartition de la prime selon le distributeur et par année.



Source : Interne AXA.

FIGURE. 2.12 – Distributeur vs Résiliation



Source : Interne AXA.

#### d) Région.

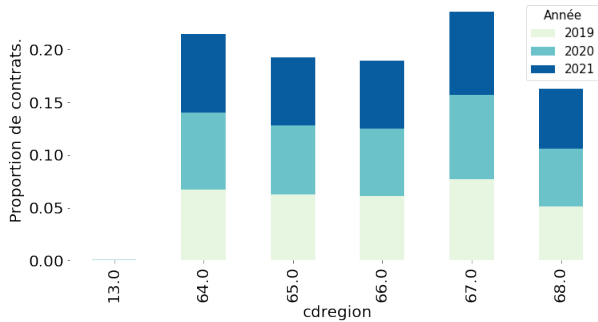
Dans la base de données, la variable **Cdregion** indique le code de la région du client, c'est-à-dire sa zone géographique. Elle est codifiée comme suit :

- 13 : DOM-TOM
- 64 : Île de France
- 65 : Nord Est
- 66 : Ouest
- 67 : Sud Est
- 68 : Sud Ouest

Au sein de notre portefeuille RC, on constate comme l'indique la figure 2.13 qu'en termes de nombre de contrats, on a à peu près le même volume de contrats dans toutes les régions hors

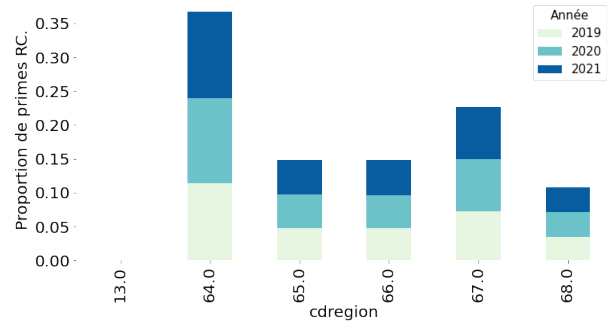
DOM-TOM. Par contre, lorsqu'on regarde les primes RC (figure 2.14), environ 37% est localisé en Île-de-France. Notons toutefois qu'entre 2019 et 2021, la répartition du portefeuille dans ces régions est restée la même.

FIGURE. 2.13 – Répartition du nombre de contrats selon la région et par année.



Source : Interne AXA.

FIGURE. 2.14 – Répartition de la prime selon la région et par année.



Source : Interne AXA.

#### e) Effet.

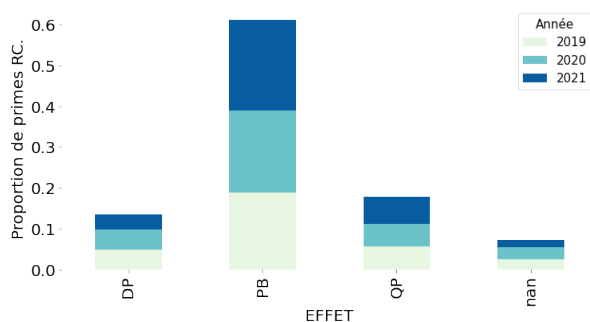
En fonction de la majoration appliquée lors du renouvellement, on peut avoir 3 cas de figure :

- PB : Presse Bouton, c'est la majoration automatique à tous les contrats.
- QP : Qualité du Portefeuille, c'est la majoration supplémentaire pour les contrats ayant une forte sinistralité.
- DP : Défense de Portefeuille, c'est la majoration qui a été appliquée en dessous du seuil initial (PB).

Exemple : Un client bénéficie d'un contrat RC, on majore les contrats en moyenne de 5%, c'est le **PB**. Sauf que ce client a un indice faible (indice = 1), donc on lui ajoute une majoration supplémentaire de 4%, c'est la **QP**, puis finalement, il refuse la majoration totale de 9%, mais accepte à 8%. Dans ce cas, la **DP** est de 1% puisque nous majorons de 8% au lieu de 9%. Le total des majorations s'appelle l'**effet prix (EP)** et s'écrit de la manière suivante :  $EP = PB + QP - DP$ .

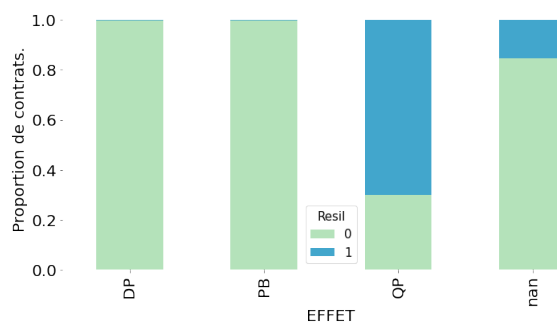
Au sein de notre portefeuille RC, on constate que les renouvellements des contrats se font majoritairement en utilisant l'effet PB (voir figure 2.15). De plus, entre 2019 et 2021, les parts PB, QP, DP sont restées les mêmes dans notre portefeuille. Par ailleurs, l'analyse de la figure 2.16 donnant la proportion des clients résiliés suivant la majoration appliquée pour le renouvellement montre que près de 80% des clients passés en QP résilient leurs contrats. Ceci signifie que juste 20% des clients du portefeuille acceptent une forte majoration lors de la période du renouvellement. Par contre, les clients qui passent par la PB ou la DP ne résilient presque pas leurs contrats RC. Ce résultat indique déjà que les souscripteurs doivent prêter une attention particulière aux contrats qui passent en QP pendant le renouvellement, car ce sont ces derniers qui sont les plus susceptibles de résilier.

FIGURE. 2.15 – Répartition de l'effet suivant l'année d'étude et la prime.



Source : Interne AXA.

FIGURE. 2.16 – Effet vs Résiliation



Source : Interne AXA.

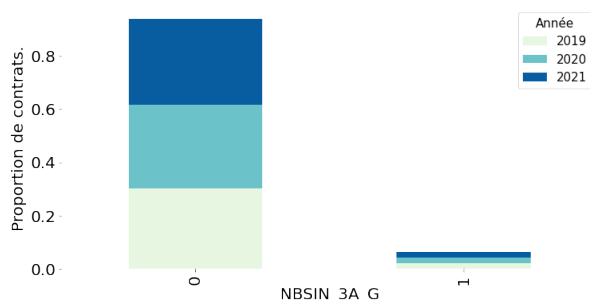
#### f) Sinistralité des trois dernières années.

Dans la base de données, la variable **Nbsin\_3A** indique la sinistralité du client sur les trois dernières années. Pour les analyses, nous l'avons groupé en deux catégories :

- 0 : aucun sinistre enregistré sur les trois dernières années ;
- 1 : au moins un sinistre enregistré sur les trois dernières années.

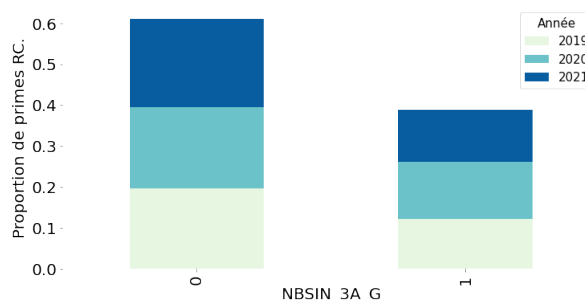
Au sein de notre portefeuille RC on constate comme l'indique la figure 2.17 qu'en termes de nombre de contrats, on a moins de 10% de sinistralité sur les 3 dernières années. Par contre, lorsqu'on regarde la prime (figure 2.18), ces clients sinistrés ont environ 40% des primes RC. Ceci indique que ce sont principalement les gros clients du portefeuille qui enregistrent des sinistres. Par ailleurs, l'analyse de la figure 2.19 montre que les clients non-sinistrés enregistrent dans l'ensemble un taux de résiliation plus élevé que les clients sinistrés.

FIGURE. 2.17 – Répartition du volume de sinistralité suivant l'année d'étude.



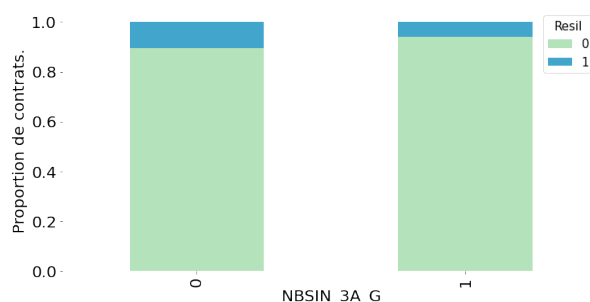
Source : Interne AXA.

FIGURE. 2.18 – Répartition de la sinistralité suivant l'année d'étude et la prime.



Source : Interne AXA.

FIGURE. 2.19 – Sinistralité vs Résiliation

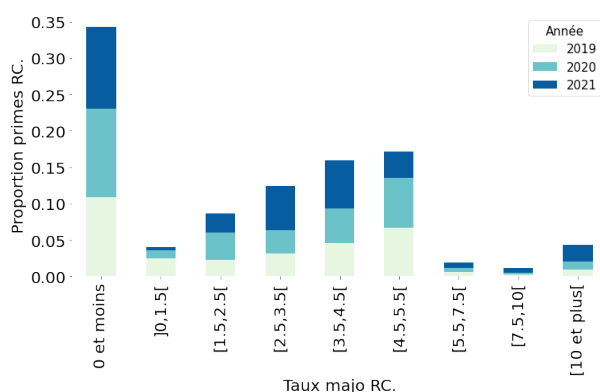


Source : Interne AXA.

### g) Taux de majoration RC.

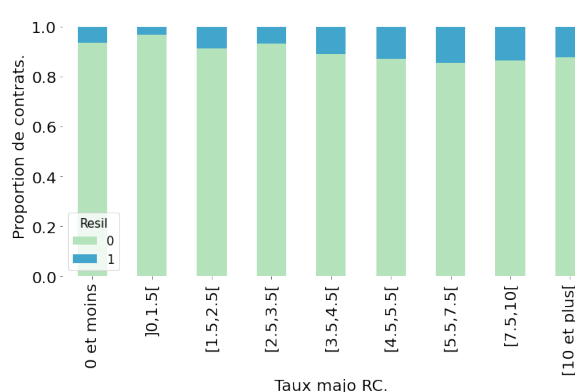
Le taux de majoration du contrat RC noté  $Tx\_majo\_RC = \frac{Majoration\ du\ contrats\ RC}{Prime\ du\ contrat\ RC}$ . Il indique de combien va varier la prime d'assurance lors du renouvellement. Pour faciliter les analyses, nous l'avons groupé en catégories. La figure 2.20 montre que sur les trois années 2019 à 2021, la répartition de majoration RC était pratiquement la même et ces majorations sont majoritairement inférieures à 5.5%. De plus on constate que près de 35% des contrats RC ont eu des majorations inférieures ou égales à 0 (les primes ont été revues à la baisse ou alors n'ont pas été modifiées). Lorsqu'on observe la variation du taux de résiliation en fonction de la majoration RC (2.21) on constate que plus le taux de majoration est élevé, plus la proportion des résiliés devient importante.

FIGURE. 2.20 – Répartition de la prime selon le taux de majo RC et par année.



Source : Interne AXA.

FIGURE. 2.21 – Taux majoration RC vs Résiliation



Source : Interne AXA.

### 2.2.3 Analyse des variables spécifiques aux clients multi-équipés.

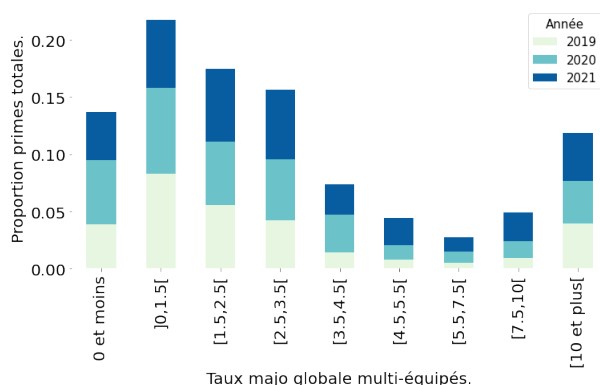
Dans cette partie, nous analysons les variables spécifiques aux clients multi-équipés telles que : le taux de majoration global du client multi-équipé ; la différence entre le taux de majoration RC et le taux de majoration multi-équipé ; l'évolution du multi-équipement ; le nombre de contrats du client ; le nombre de branches du client ; la liste des branches du client ; et la résilia-

tion des autres branches du client. Notons que pour ces analyses nous nous sommes restreints à la base des clients multi-équipés comportant 183 mille clients RC.

#### a) Taux de majoration global du client multi-équipé.

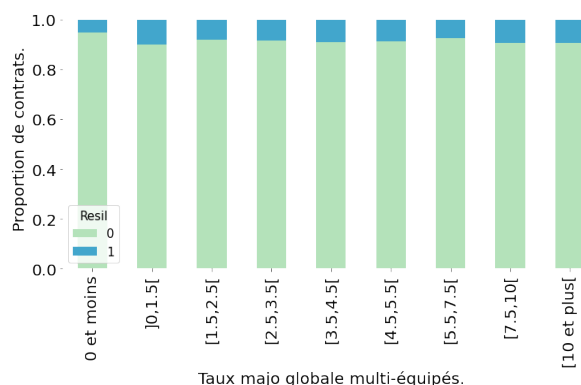
Le taux de majoration du client multi-équipé noté  $Tx\_majo\_multi = \frac{\text{Majoration globale des contrats du client}}{\text{Prime globale des contrats du client}}$ , indique le taux de majoration global qu'un client multi-équipé a eu sur l'ensemble de ses contrats. Pour faciliter les analyses, nous l'avons groupé en catégories. La figure 2.22 montre que contrairement aux majoration RC (voir figure 2.21) le taux de majoration global des multi-équipés est beaucoup plus élevé et près de 15% des clients multi-équipés ont reçu des majorations globale au-dessus de 10%.

FIGURE. 2.22 – Répartition de la prime selon le taux de majo global et par année.



Source : Interne AXA.

FIGURE. 2.23 – Taux majoration multi-équipé vs Résiliation



Source : Interne AXA.

#### b) Différence entre le taux de majoration RC et le taux de majoration multi-équipé

**Diff\_majo** est la variable de notre base qui compare<sup>1</sup> le taux de majoration de la branche RC d'un client au taux de majoration sur l'ensemble de ses branches. Elle est codée comme suit :

- Diff\_majo= -1 si taux de majoration global < taux de majoration branche RC ;
- Diff\_majo= 0 si taux de majoration global = taux de majoration branche RC ;
- Diff\_majo= 1 si taux de majoration global > taux de majoration branche RC.

Cette variable a été introduite suite au constat fait dans le tableau 2.24 ci-dessous. En effet, à plus de 3.5% de majoration RC, le taux de majoration globale devient dans l'ensemble inférieure à la majoration RC du client ce qui signifie que la majoration globale du portefeuille est beaucoup plus tirée par la majoration de la branche RC. Par exemple, 25.71% des contrats RC ont eu des majorations entre [4.5,5.5], et juste 10% de ces derniers ont eu des majoration

1. Notons que nous avons au préalable groupé les variables taux de majoration RC et taux de majoration globale pour faire cette comparaison.

globale supérieures à 5.5%. Donc pour environ 90% des clients de majoration RC entre [4.5,5.5[ ont eu des majorations globales inférieures à 5.5%.

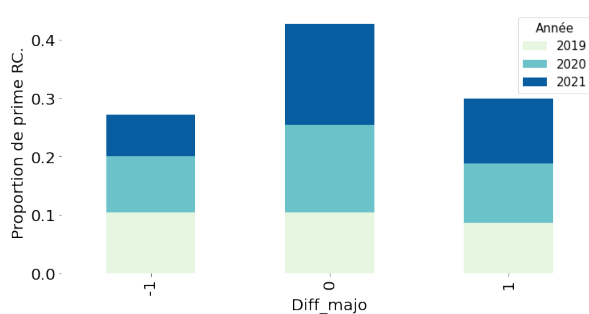
TABLEAU 2.2 – Taux majoration RC vs Taux de majoration globale.

| Tx majo global | Tx_majo RC |            |         |           |           |           |           |           |          |              | Total   |
|----------------|------------|------------|---------|-----------|-----------|-----------|-----------|-----------|----------|--------------|---------|
|                | NA         | 0 et moins | ]0,1.5[ | ]1.5,2.5[ | ]2.5,3.5[ | ]3.5,4.5[ | ]4.5,5.5[ | ]5.5,7.5[ | ]7.5,10[ | ]10 et plus[ |         |
| 0 et moins     |            | 8,84%      | 0,01%   | 3,08%     |           |           |           |           |          |              | 16,77%  |
| ]0,1.5[        |            |            | 0,25%   |           | 2,82%     |           |           |           |          |              | 5,87%   |
| ]1.5,2.5[      |            |            |         | 3,96%     |           |           |           |           |          |              | 8,86%   |
| ]2.5,3.5[      |            |            |         |           | 4,61%     |           |           |           |          |              | 10,90%  |
| ]3.5,4.5[      | 0,34%      | 9,47%      |         |           |           | 8,34%     |           |           |          |              | 17,39%  |
| ]4.5,5.5[      |            |            | 0,30%   | 9,35%     |           |           | 8,62%     |           |          |              | 21,18%  |
| ]5.5,7.5[      |            |            |         |           | 4,40%     | 2,37%     |           | 1,91%     |          |              | 6,36%   |
| ]7.5,10[       |            |            |         |           |           |           | 2,55%     | 0,26%     | 0,96%    |              | 3,49%   |
| ]10 et plus[   |            |            |         |           |           |           |           |           | 0,09%    | 2,79%        | 9,17%   |
| Total          | 0,34%      | 18,31%     | 0,56%   | 16,39%    | 11,83%    | 16,30%    | 25,71%    | 4,00%     | 1,91%    | 4,64%        | 100,00% |
|                |            |            |         |           |           |           |           |           |          |              | 30,58%  |
|                |            |            |         |           |           |           |           |           |          |              | 37,49%  |
|                |            |            |         |           |           |           |           |           |          |              | 29,13%  |

Source : Interne AXA.

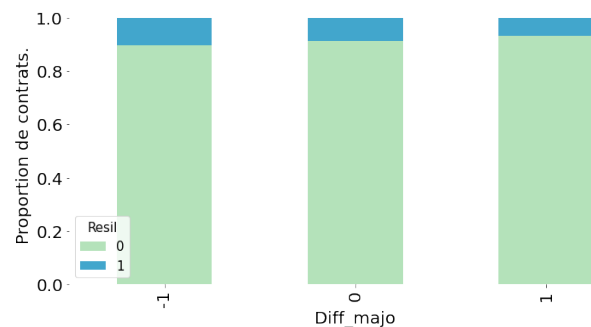
La figure 2.24 donne une répartition de la différence de majoration entre la majoration globale des clients multi-équipés et leurs majorations RC. Elle montre qu'environ 30% des clients RC multi-équipés ont reçu des majorations RC supérieures à leurs majorations globales. De plus en analysant la figure 2.25 on constate que les clients ayant reçu des majorations RC au-dessus de leurs majorations globales ont eu plus de résiliation.

FIGURE. 2.24 – Taux majoration RC vs Taux de majoration global.



Source : Interne AXA.

FIGURE. 2.25 – Diff majoration vs Résiliation.



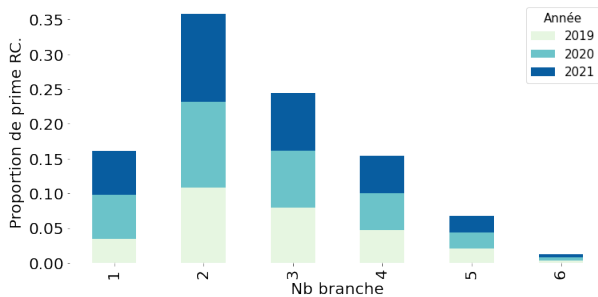
Source : Interne AXA.

### c) Nombre de branches.

Nb **branche** est la variable de notre base qui donne le nombre de branches du client multi-équipé. Notons que le nombre de branches = 1 signifie que le client a plusieurs contrats RC et pas de contrats sur les autres branches. On constate que les clients multi-équipés ont

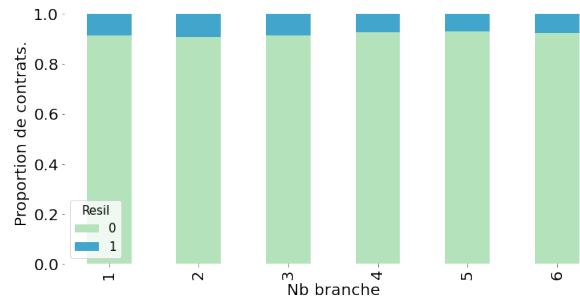
majoritairement souscrit à 2 (voir figure 2.26). Pour ce qui est de la résiliation (figure 2.27), la proportion des résiliés est plus élevée chez les clients ayant moins de 3 branches.

FIGURE. 2.26 – Répartition du nombre de branches suivant l’année d’étude et la prime.



Source : Interne AXA.

FIGURE. 2.27 – Nombre de branches vs résiliation.



Source : Interne AXA.

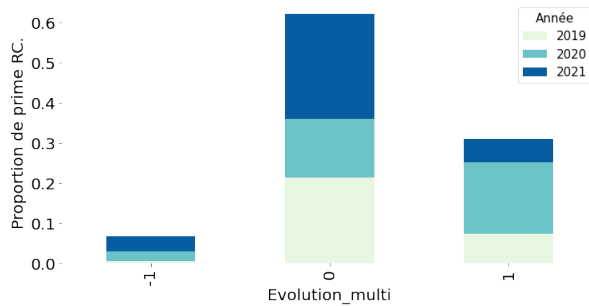
#### d) Évolution du multi-équipement.

**Évolution\_multi** est la variable qui compare le nombre de contrats du client entre l’année N-1 et l’année N. Elle est codée comme suit :

- Évolution\_multi = -1 : nombre contrats année N-1 > Nombre contrats année N (le client a résilié au moins l’un de ses contrats) ;
- Évolution\_multi = 0 : nombre contrats année N-1 = Nombre contrats année N (le client n’a résilié ou a souscrit à autant d’affaires nouvelles que de résiliation) ;
- Évolution\_multi = 1 : nombre contrats année N-1 < Nombre contrats année N (le client a souscrit à au moins une affaire nouvelle).

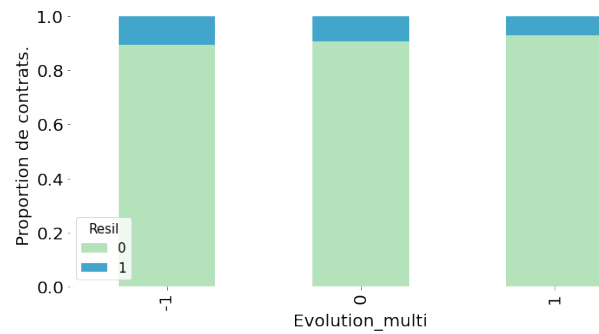
Sur la figure 2.28 présentant la répartition de l’évolution du multi-équipement au sein du portefeuille, on constate que 2020 a été une année spécifique avec près de 50% des clients multi-équipés qui ont agrandi leur portefeuille. Par ailleurs, lorsqu’on regarde les années 2019 et 2021, on remarque qu’environ 5% des clients multi-équipés ont résilié au moins un des contrats leur portefeuille. Concernant la résiliation (figure 2.29, les clients multi-équipés en décroissance de portefeuille ( évolution multi = -1) ont une plus grande proportion de résiliés.

FIGURE. 2.28 – Répartition de l'évolution du multi-équipement suivant l'année d'étude et la prime.



Source : Interne AXA.

FIGURE. 2.29 – Évolution multi-équipement vs résiliation.



Source : Interne AXA.

#### e) Liste des branches du client multi-équipé.

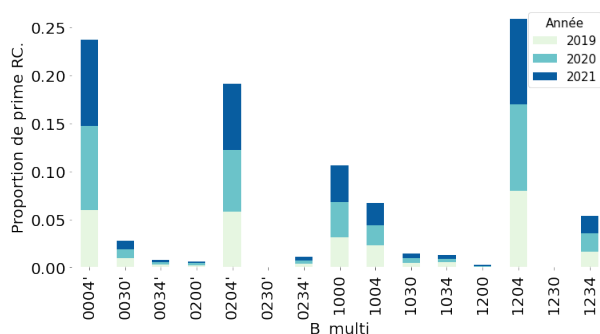
**B\_multi** est une variable qui a été construite à l'aide des variables indiquant le nombre de contrats du client RC sur les autres branches. Elle a été codée *XYZT* avec  $X = automobile$  ;  $Y = Risqueindustriel$  ;  $Z = Construction$  ;  $T = Autre\ branche$ . A titre d'exemple :

- 0000= mono-équipé ;
- 1000= multi-équipé à la branche automobile uniquement ;
- 1200= multi-équipé aux branches automobile et Risque industriel ;
- 1230= multi-équipé aux branches automobile , Risque industriel et Construction ;
- 1234= multi-équipé aux branches automobile , Risque industriel, Construction et une autre branche.

La figure 2.30 donnant l'évolution de la répartition des branches du multi-équipent RC indique qu'une grande partie des clients multi-équipés ont les branches automobile et risque industriel. Quand au taux de résiliation (voir figure 2.31, il est difficile de se prononcer car les branches qui enregistrent les plus grandes proportions de résiliés (0200= 1230=multi-équipé à la branche risque industriel 1200=1230=multi-équipé aux branches automobile et Risque industriel ; 1230=multi-équipé aux branches automobile , Risque industriel et Construction) ont très peu de clients.

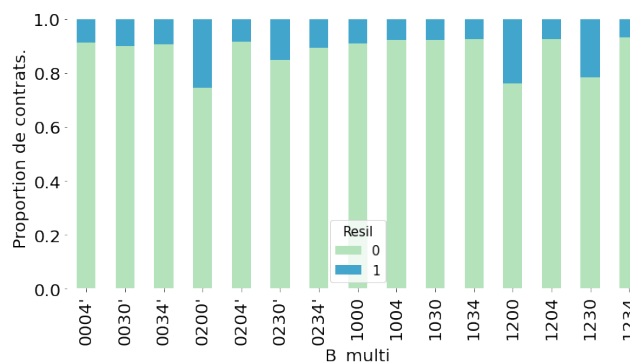


FIGURE. 2.30 – Répartition des branches du multi-équipement suivant l'année d'étude et la prime.



Source : Interne AXA.

FIGURE. 2.31 – Branches multi-équipement vs résiliation.



Source : Interne AXA.

Les analyses précédentes ont permis d'avoir un aperçu global sur chacune des variables de notre base caractérisant les clients RC Entreprises ainsi que la répartition de la résiliation au sein de chacune d'elles. Grâce à cette analyse, nous avons remarqué que le portefeuille est resté homogène au cours des trois années 2019, 2020 et 2021. Ceci signifie que, toutes choses égales par ailleurs, les conclusions tirées dans cette analyse seront valables pour les années à venir et notamment pour 2022.

## 2.3 Analyse croisée des variables.

Bien que nous ayons grâce à l'analyse réalisée précédemment une idée sur la composition de notre portefeuille, il est important de savoir comment ces variables sont corrélées entre elles. En effet, dans les modèles statistiques, il est primordial de découvrir et de quantifier à quel point les variables sont liées, car certaines de ces dépendances affaiblissent les performances des modèles et peuvent conduire à des conclusions fallacieuses. Pour analyser la corrélation entre les variables, plusieurs méthodes statistiques peuvent être utilisées en fonction du type des variables (qualitative, quantitative). On peut noter ici le test du khi-deux, la corrélation de Pearson, l'analyse de la variance (ANOVA), ... Pour analyser la corrélation entre les variables, nous allons utiliser le V de Cramer pour les variables qualitatives et la corrélation de Pearson pour les variables quantitatives.

### 2.3.1 Corrélation entre les variables catégorielles.

Pour savoir si deux variables catégorielles sont liées, on utilise le test statistique<sup>2</sup> du khi-deux. Dans ce test, l'hypothèse nulle est simplement « les deux variables testées sont indépendantes ». Enfin, le test s'accompagne d'une statistique de test qui participe à la décision de

<sup>2</sup>. Un test statistique est une procédure de décision entre deux hypothèses. Il s'agit d'une démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique, appelée hypothèse nulle  $H_0$ , en fonction d'un jeu de données.

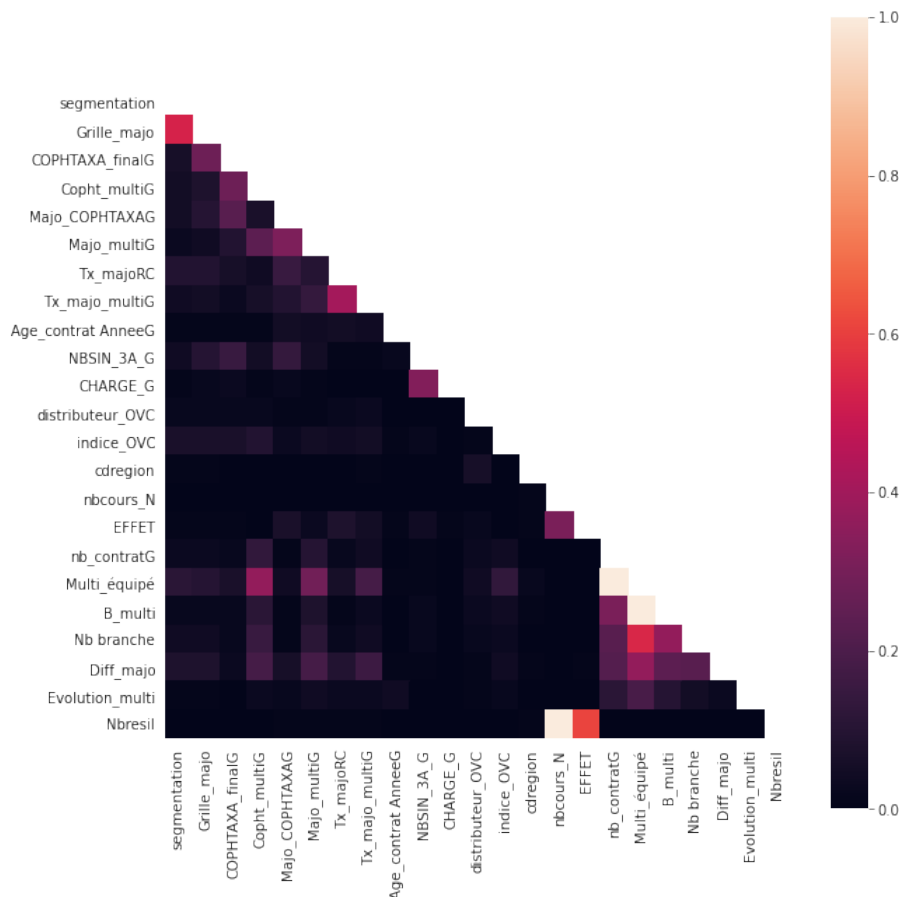
rejeter ou non l'hypothèse nulle. Cette statistique par la manière dont le test est construit suit la loi de khi-deux avec un certain degré de liberté.

### Mais comment décider de rejeter ou non l'hypothèse nulle ?

Chaque test statistique dispose de ce qu'on appelle la **p-valeur**. On peut la voir comme une valeur de référence pour décider du rejet ou non de l'hypothèse nulle. Si cette dernière est en dessous d'un seuil  $\alpha\%$  alors on rejette l'hypothèse nulle (le seuil de  $\alpha\%$  est d'usage de praticien et il peut varier suivant la problématique).

Après avoir évalué la dépendance des variables, on mesure le niveau de corrélation entre les deux variables en utilisant le **V de Cramer**. Il se calcule à l'aide de la statistique de test, du degré de liberté et des dimensions de la table de contingence du test de khi-deux. Il renvoie une valeur entre 0 et 1. Si la valeur est renvoyée, est supérieure à 0.9, on peut qualifier la relation de très forte. Si la valeur est inférieure à 0.01, on peut qualifier la relation de faible.

FIGURE. 2.32 – Matrice de corrélation entre les variables catégorielles : V de Cramer.



Source : Interne AXA.

La figure 2.32 donne la matrice de corrélation du V de Cramer entre les variables catégorielles. On constate que :

- Les variables segmentation et grille\_majo sont fortement corrélées avec une corrélation d'environ 0,6. Également les variables prime-RC groupées en classe et la grille\_majo sont fortement corrélées avec un taux autour de 0,4. Ces fortes corrélations s'expliquent par le

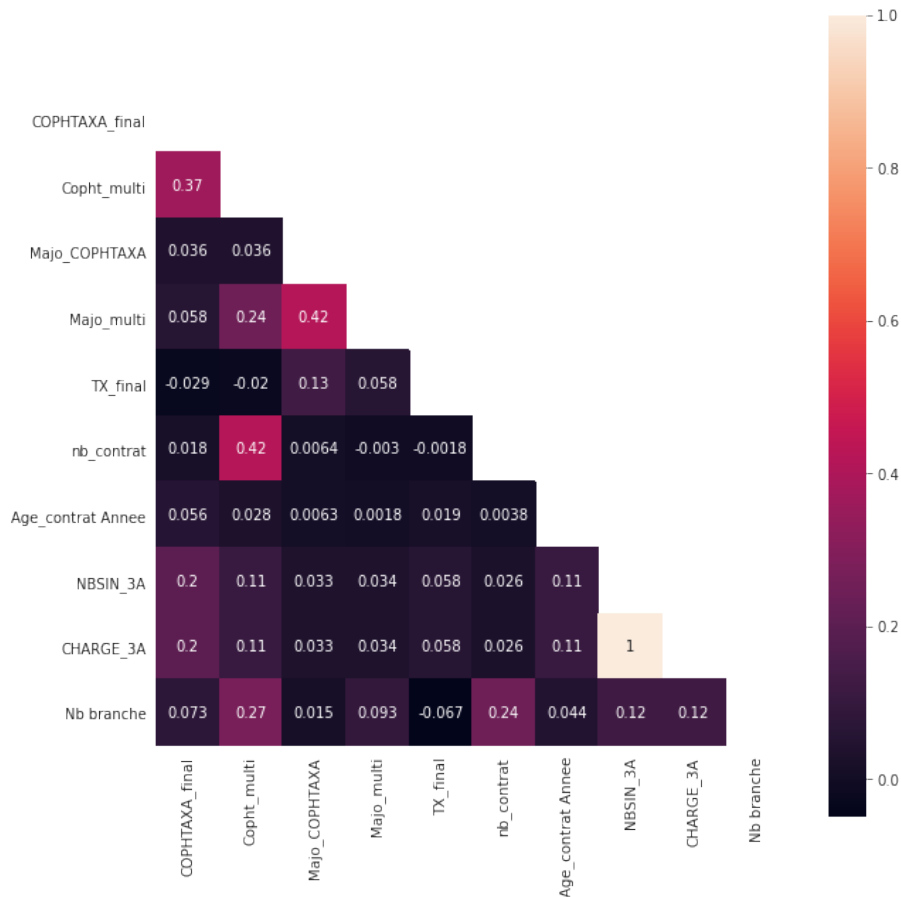
fait que la grille\_majo est une variable construite en utilisant les segments des clients et leurs niveaux de primes. Vu cette forte liaison entre les 3 variables, elles ne pourront être utilisées simultanément dans les modèles.

- Les variables prime-RC et majo-RC, puis prime-RC et prime-Multi sont très corrélées avec un taux autour de 0,4. La corrélation entre prime-RC et majo RC s'explique par le fait que la majoration du client est proportionnelle à sa prime. D'autre part, la corrélation entre prime-RC et prime-Multi s'explique par le fait que la prime-Multi est égale à la prime-RC plus la prime des autres branches du multi-équipement.
- Les variables nbresil et nbcours sont parfaitement corrélées avec un taux de 1. En effet nbresil est l'indicatrice de la résiliation et nbcours indique si le contrat est en cours. Donc si nbresil=1 alors nbcours=0 et vice-versa d'où la parfaite corrélation.
- Les variables nbresil et effet sont fortement corrélées avec un taux autour de 0.7. Ceci s'explique par le fait que plus de 80% des contrats qui passent en QP finissent en résiliation.
- Les variables du multi-équipement sont également très corrélées. Notamment, l'indicateur du multi-équipement est fortement corrélé à B\_multi (liste des branches du multi-équipement), nombre de branches et nombre de contrats. Ceci s'explique par le fait que les clients mono-équipés ont tous les mêmes caractéristiques en terme de nombre de branche et de contrat.

### 2.3.2 Corrélation entre les variables continues.

Le coefficient de Pearson est un indice reflétant une relation linéaire entre deux variables continues. Le coefficient de corrélation varie entre  $-1$  et  $+1$ ,  $0$  reflétant une relation nulle entre les deux variables, une valeur négative (corrélation négative) signifiant que lorsqu'une des variables augmente, l'autre diminue ; tandis qu'une valeur positive (corrélation positive) indique que les deux variables varient ensemble dans le même sens. Mathématiquement, le coefficient de corrélation de Pearson entre deux variables aléatoires réelles  $X$  et  $Y$  ayant chacune une variance finie, noté  $Cor(X, Y)$  est défini par :  $r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = Cor(X, Y)$ , où  $Cov(X, Y)$  désigne la covariance des variables  $X$  et  $Y$ ,  $\sigma_X$  et  $\sigma_Y$  désignent leurs écart-types.

FIGURE. 2.33 – Matrice de corrélation entre les variables continues : Pearson.



**Source : Interne AXA.**

La figure 2.33 présentant la matrice de corrélation de Pearson entre les variables montre que :

- Les variables prime-RC et prime-Multi sont très corrélées avec un taux de 0.37. Cette forte corrélation s'explique par le fait que la prime-Multi est égale à la prime-RC plus la prime des autres branches du multi-équipement.
- On a également une forte corrélation positive entre les majorations et les primes qui s'explique par le fait que les majorations sont proportionnelles aux primes.
- Les variables prime-multi, nombre de branches et nombre de contrats sont également fortement corrélées positivement. Cette corrélation est due au fait que les clients ayant plusieurs contrats ont pour la plupart souscrit à des produits de branches différentes ; et aussi, chaque contrat a une prime qui lui est propre. Ainsi la prime-Multi qui est la somme des primes de tous les contrats du client multi-équipé augmente avec le nombre de contrats. D'où la corrélation positive entre la prime-multi et le nombre de contrats.
- Les variables charges sinistres et nombre sinistres sont parfaitement corrélées positivement. En effet les clients n'ayant pas eu de sinistre n'ont pas de charges sinistres et également, plus le nombre de sinistres est important plus la charge sera élevée. Ainsi il est évident que ces deux variables sont très corrélées.

## 2.4 Traitement des données.

Nous avons présenté ci-dessus la base de données partant de sa construction à l'analyse des différentes variables. Comme nous l'avons constaté lors des analyses, notre base de données contient très peu d'anomalies. Ceci est particulièrement dû au fait que les bases intermédiaires utilisées pour sa construction sont des bases couramment utilisées en interne et donc sont constamment mises à jour. Toutefois notre base contient quand même quelques coquilles qu'il serait judicieux de ~~les~~ traiter avant de passer à l'étape de modélisation.

### 2.4.1 Traitement des valeurs manquantes.

En statistiques, les données manquantes ou les valeurs manquantes se produisent lorsqu'aucune valeur de données n'est représentée pour une variable pour une observation donnée. Les données manquantes sont courantes et peuvent avoir un effet significatif sur l'inférence, les performances de prédiction ou toute autre utilisation faite avec les données. Des données manquantes peuvent exister dans les données en raison d'une « omission de réponse » pour l'observation donnée. Dans cette situation, aucune information n'est fournie pour un élément, pour plusieurs éléments ou pour une unité entière, c'est-à-dire un sujet. Les données peuvent également être manquantes de façon aléatoire ou prévisible. En fonction du type de données manquantes, des traitements appropriés doivent être réalisés. Bien que des méthodes sophistiquées existent de plus en plus pour palier au problème de valeurs manquantes, il est très conseillé de faire des études sur des données minimisant le taux de valeurs manquantes.

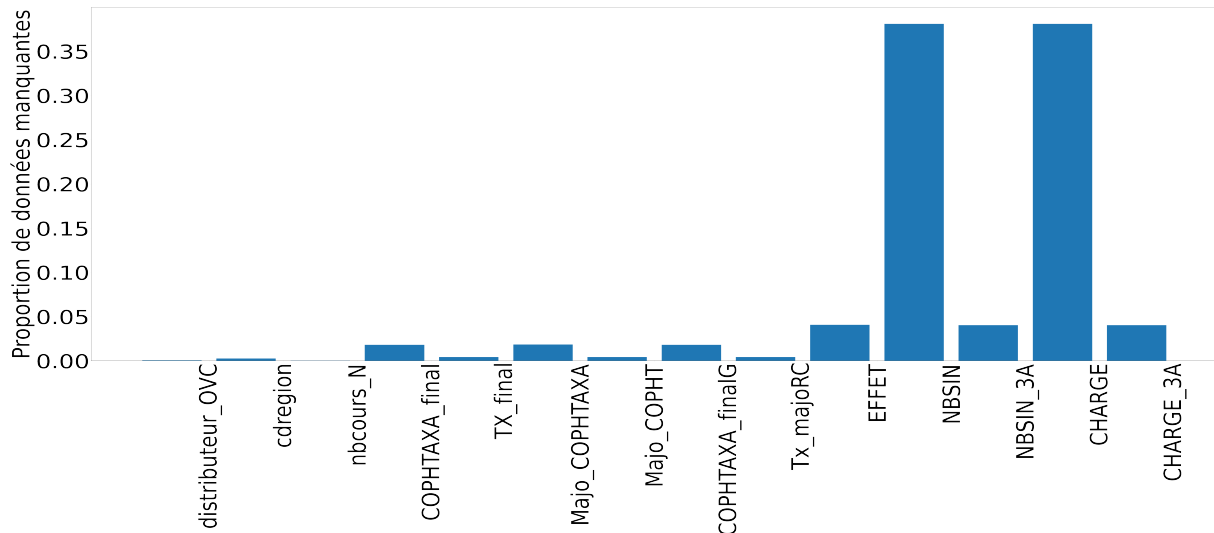
Au sein de notre base de données, la proportion de données manquantes dans la base s'élève à 1.73%, ceci rassure sur la qualité des données que nous utiliserons pour les modèles. Pour traiter ces valeurs manquantes, nous avons premièrement ciblé les variables qui regorgent le plus de données manquantes, ensuite étudié la nature de ces dernières et enfin proposé des approches de correction. Parmi les 37 variables explicatives, 14 contiennent au moins une valeur manquante, la figure 2.34 ci-dessous donnant le taux de valeurs manquantes par variable présente le taux de valeurs manquantes dans chacune de ces variables. On constate que :

- Deux variables détiennent environ 70% des valeurs manquantes de la base, il s'agit du nombre de sinistres enregistré l'année précédente et de sa charge. Pour ces deux variables, la valeur manquante renvoie l'information selon laquelle le client n'a pas eu de sinistre au cours de l'année précédente. Ainsi, pour corriger ces valeurs manquantes, nous avons remplacé par 0 les différentes observations manquantes observées au niveau de la sinistralité.
- Les variables catégorielles possédant des valeurs manquantes sont le distributeur, l'indice, la région et l'effet ; on remarque que le taux de données manquantes pour chacune d'elles est très en dessous de 5%. Pour ces variables, les valeurs manquantes vont être corrigées en remplaçant la donnée manquante par le mode. A titre d'exemple, les valeurs manquantes de la variable distributeur seront remplacées par la modalité "Agent" qui est le canal de

distributeur de près de 70% des contrats RC de notre base de données (voir figure 2.10).

- Les autres variables continues hors sinistralité de la base de données possédant des données manquantes sont la prime, la majoration et le taux de majoration. Vu les faibles taux de données manquantes pour ces variables, nous avons fait des imputations par la médiane en remplaçant les observations manquantes de la prime et de la majoration par leurs valeurs médianes.

FIGURE. 2.34 – Répartition des valeurs manquantes au sein des variables.



Source : Interne AXA.

## 2.4.2 Création de nouvelles variables.

En analyse statistique, la performance des modèles est étroitement liée à la représentativité des données et taille de l'échantillon utilisé (plus l'échantillon est grand, plus la mesure est précise). Pour éviter d'aboutir à des conclusions fallacieuses, il est important de vérifier si également au sein des différentes variables les modalités sont assez représentatives. En théorie, une modalité est représentative si elle contient au moins 5% de l'échantillon. Dans le cadre des analyses descriptives, nous avons remarqué que certaines modalités étaient très marginales. L'objectif ici est de créer de nouvelles variables en regroupant les modalités peu représentées pour obtenir une base avec des modalités plus significatives.

**Remarque :** Dans les analyses descriptives, nous avons anticipé cette partie créant des variables catégorielles à base des variables continues (exemple prime\_G qui est la variable prime regroupée en classe). Ces regroupements ont été fait en s'inspirant de la dichotomisation appliquée en interne sur les variables continues pour les modèles de tarification et dans les politiques de renouvellement. Nous présentons dans le tableau A.2 (annexe A) les différentes variables créées d'une part en transformant les variables continues et d'autre part en groupant les modalités des variables continues.

### 2.4.3 Choix des variables de modélisation.

Grâce aux analyses descriptives précédemment réalisées, nous avons constaté que parmi les variables, plusieurs étaient fortement corrélées entre elles de par leurs constructions. Pour éviter le phénomène de multi-collinéarité (forte corrélation entre les variables explicatives) qui entraînerait une instabilité des modèles, et les rendrait in-interprétables, nous n'allons pas introduire dans les modèles les variables fortement corrélées. À titre d'exemple, on n'inclura pas simultanément le segment et la grille de majoration (grille-majo) car ces deux variables sont fortement corrélées comme l'indique la figure 2.32. Ou encore une variable continue et la variable catégorielle obtenue par transformation de cette dernière. Le tableau 2.3 ci-dessous donne la liste des variables retenues au final pour les modèles.

TABLEAU 2.3 – Liste des variables choisies pour la modélisation.

| Variables       | Modalités      | Description                                                                                                                                                         |
|-----------------|----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Nbresil         | 0,1            | Indicateur de la résiliation du contrat l'année N.                                                                                                                  |
| Annee           | 2019,2021,2022 | Année d'étude.                                                                                                                                                      |
| Cdregion_G      | 64,67,...      | Région du client.                                                                                                                                                   |
| Distributeur_G  | Agent,Courtier | Distributeur (agent courtier).                                                                                                                                      |
| Grille_majoG    | S01<5K,...     | Grille utilisée comme base de majoration pour le renouvellement.                                                                                                    |
| Indice_OVCG     | 1_2, 3,4_5     | Indice du client.                                                                                                                                                   |
| Nbbranche_G     | 1,2, ...       | Nombre de branches du client.                                                                                                                                       |
| Nbsinistre3A_G  | 0,1,           | Nombre de sinistres enregistré sur les 3 dernières années.                                                                                                          |
| TX_majo_RCG     | [0,0.25[,...   | Taux de majoration du contrat RC.                                                                                                                                   |
| Tx_majo_multiG  | [0,0.25[,...   | Taux de majoration du client sur l'ensemble de ses contrats.                                                                                                        |
| Age_contratG    | 1,2, ...       | Âge du contrat en année.                                                                                                                                            |
| B_Multi         | 1234, 1034,... | Les différentes branches du client multi-équipé.                                                                                                                    |
| Multi_équipé    | 0,1            | Indicateur du multi-équipement.                                                                                                                                     |
| Diff_majo       | -1,0,1         | Variable indiquant si le taux de majoration RC est supérieur ou inférieur au taux de majoration global du client.                                                   |
| Evolution_multi | -1,0,1         | Évolution du multi-équipement (il indique si par rapport à l'année précédente le client à résilier l'un de ces contrat ou alors souscrit à des affaires nouvelles). |

## 2.5 Conclusion partielle.

Dans cette deuxième partie du travail, nous avons présenté la base de données partant de la construction à la description des variables. Ensuite, les différents traitements qui ont été effectués pour corriger les anomalies de la base. Enfin, sur la base des analyses réalisées, les variables intéressantes pour la modélisation ont été choisies. Les parties suivantes seront consacrées à la présentation de la méthodologie pour l'atteinte des objectifs ainsi que des résultats obtenus.



## Chapitre 3

# MODÉLISATION DE LA RÉSILIATION DES CLIENTS EN RESPONSABILITÉ CIVILE ENTREPRISE.

Les modèles de tarification actuels consistent en la segmentation des assurés en plusieurs classes qui diffèrent par leurs caractéristiques. Ces dernières sont des critères essentiels qui permettent d'obtenir un tarif viable. Cependant, même si certains assurés appartiennent à une même classe, il n'empêche que leur résiliation peut varier entre eux. En effet, chaque assuré a un comportement qui lui est propre, et le modèle tarifaire, qui tente d'estimer au mieux ce risque, n'a peut-être pas tous les éléments nécessaires pour les discriminer (informations protégées par la législation, informations inaccessibles, ...). Un paramètre à ce jour non exploité dans les modèles tarifaires est le multi-équipement, or cet indicateur est un paramètre fort de fidélisation client. De manière indirecte, le multi-équipement permet de renforcer la fidélisation des « bons » clients et de diversifier les risques au sein du portefeuille.

Cette partie de notre travail sera consacrée à la modélisation de la résiliation des clients de notre portefeuille. Il s'attardera en particulier sur la résiliation des clients multi-équipés, ceci afin de voir si le comportement de ces derniers en termes de résiliation diffère de celui des clients mono-équipés. Nous allons présenter dans un premier temps l'approche méthodologique utilisée et ensuite les résultats obtenus en l'appliquant à notre base de données.

### 3.1 Approche méthodologique de la modélisation de la résiliation.

#### 3.1.1 Modèle de régression logistique binaire.

La régression logistique ou régression binomiale est une méthode statistique couramment utilisée pour aborder la modélisation d'une variable binaire binomiale ou de Bernoulli (Exemple :

absence ou présence d'une pathologie, possession ou non d'un produit, résiliation ou non du contrat,...). Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien (la présentation du modèle linéaire généralisé se est faite en annexe B). Dans le cadre de notre étude, nous cherchons à expliquer premièrement la résiliation des clients de notre portefeuille. Nous présenterons ici la formulation mathématique du modèle de régression logistique, ses avantages et ses limites. Nous nous inspirons dans cette partie de **Xavier D'Haultfoeuille, 2018**.

#### a) Principe mathématique.

Posons  $Y$  l'indicateur de la résiliation ( $Y \in \{0, 1\}$  avec 1 = résilié) et  $X = (X_1; \dots; X_k) \in \mathbb{R}_k$  le vecteur des variables explicatives. En pratique, les modèles linéaires sont mal adaptés pour étudier  $Y$ . En effet, comme  $Y \in \{0, 1\}$ , on a :

$$E(Y|X) = P(Y = 1|X) \in [0; 1]. \quad (3.1)$$

Dans un modèle linéaire (sous l'hypothèse d'exogénéité  $E(\varepsilon|X) = 0$ ), on a  $E(Y|X) = X' \beta_0$ . Mais rien n'assure que  $X' \beta_0 \in [0; 1]$ . Pour que (3.1) soit satisfaite, on va supposer que :

$$E(Y|X) = F(X' \beta_0), \quad (3.2)$$

où  $F$  est une fonction (connue) strictement croissante bijective de  $\mathbb{R}$  dans  $]0; 1[$ , donc une fonction de répartition. Le modèle (3.2) peut s'interpréter en termes de *variables latentes*. En effet, supposons qu'il existe une variable continue  $Y^* \in \mathbb{R}$  telle que  $Y = \mathbf{1}\{Y^* \geq 0\}$ . Supposons par ailleurs que  $Y^*$  suive un modèle linéaire :

$$Y^* = X' \beta_0 + \varepsilon, \quad (3.3)$$

où  $\varepsilon$  est indépendante de  $X$  et a pour fonction de répartition  $F$ . Alors

$$\begin{aligned} P(Y = 1|X) &= P(X' \beta_0 + \varepsilon \geq 0|X) \\ &= P(\varepsilon \leq X' \beta_0|X) \\ &= F(X' \beta_0). \end{aligned}$$

On retrouve donc l'équation (3.2). L'interprétation en termes de variables latentes sera donc naturelle. A priori n'importe quel choix est possible pour  $F$ . En particulier pour la loi logistique,

$$F(x) = \frac{1}{1 + e^{-x}}$$

on obtient le **modèle logistique**.

## b) Odds-ratios.

L'odd-ratio encore appelé rapport des chances, rapport des cotes ou risque relatif rapproché, est une mesure statistique exprimant le degré de dépendance entre des variables aléatoires qualitatives. L'un des avantages du modèle logistique est qu'il permet de calculer très facilement ce ratio par la formule suivante :

$$\frac{P(Y = 1|X)}{P(Y = 0|X)}.$$

## c) Identification et estimation des paramètres.

### c-1) Identification des paramètres.

Revenons à notre équation du modèle logistique :  $Y = \mathbb{1}\{X'\beta_0 + \varepsilon \geq 0\}$ . Alors deux questions se posent : (i) pourquoi fixer le seuil à 0 ? (ii) Pourquoi  $Var(\varepsilon) = \frac{\pi^2}{3}$  ? La raison est que le modèle n'est pas identifiable. En effet en notant  $s$  le seuil on a :

$$Y = \mathbb{1}\{\beta_{0-1} + X'_{-1}\beta_{0-1} + \varepsilon \geq s\} \Leftrightarrow Y = \mathbb{1}\{\beta_{0-1} - s + X'_{-1}\beta_{0-1} + \varepsilon \geq 0\}$$

on ne peut pas identifier séparément la constante  $\beta_{0-1}$  et le seuil  $s$ . On fixe donc (arbitrairement)  $s$  à 0. De même, on ne peut pas identifier de façon jointe  $\beta_0$  et la variance  $\sigma^2$  du résidu  $\varepsilon$  ; on fixe donc arbitrairement  $\sigma$ . En effet

$$Y = \mathbb{1}\{\beta_0 + \sigma\varepsilon \geq 0\} \Leftrightarrow Y = \mathbb{1}\{X'\frac{\beta_0}{\sigma} + \varepsilon \geq 0\}$$

En conclusion, si  $s$  et  $\sigma$  sont fixés et  $E(XX')$  est inversible, **le modèle est identifié**.

### c-2) Estimation des paramètres.

On s'intéresse à l'estimation de  $\beta_0$  à partir d'un échantillon indépendant et identiquement distribué  $((Y_1, X_1) \dots (Y_n, X_n))$ . Étant donné que le modèle (3.3) est paramétrique, on peut l'estimer par **la méthode du maximum de vraisemblance**.

## 3.1.2 Les arbres de décision.

Les arbres de décision sont des méthodes non-paramétriques (ne nécessitent pas d'hypothèse sur la distribution des données), cette approche non-paramétrique est un gros avantage par rapport à l'approche paramétrique du modèle logistique présenté en section à cause de :

- Ils modélisent spontanément les non-linéarités et les interactions, ce qui n'est possible dans les modèles logistiques qu'après amélioration des prédicteurs linéaires.
- Ils sont susceptibles d'isoler des groupes de comportement de résilience significative à partir d'un nombre restreint de règles de décision.

- Ils ne nécessitent pas la spécification de la distribution des données. En effet préciser une distribution pour les données est souvent difficile et très approximatif, ce qui peut conduire à des résultats erronés.
- Ils sont facilement interprétables.

Toutefois, ces avantages vont de pair avec certains inconvénients bien connus, tels que la faible stabilité du modèle et un biais vers la sélection des variables catégorielles lorsqu'elles présentent un nombre de modalités élevées.

### **a) Généralités.**

Les arbres de décision sont une famille majeure de méthodes d'exploration de données et se classent parmi les techniques d'apprentissage supervisé. Leur objectif est, à partir des données, de répartir un ensemble d'individus en des groupes les plus homogènes possibles du point de vue d'une variable cible. Différentes techniques d'induction d'arbre existent, parmi lesquelles la méthode CHAID (Chi-Squared Automatic Interaction Detector) développée par Gordon Kass en 1980, CART (Classification and Regression Trees) développée par Leo Breiman en 1984, la méthode C4.5 par Quinlan, en 1993, et plus récemment la construction d'arbres de décision basée sur une approche Bayésienne (Arbre de décision MODL) de Nicolas et al. en 2017.

Les premiers algorithmes d'apprentissage automatique d'arbre de décision sont basés sur un pré-élagage. Le principe de construction consiste, à partir de la racine de l'arbre, c'est-à-dire la totalité de l'ensemble d'apprentissage, à choisir parmi toutes les variables explicatives celle qui donne la meilleure partition selon un critère de segmentation. Puis de façon récursive, on applique l'algorithme de segmentation sur les feuilles. Le processus s'arrête quand pour chaque feuille on ne peut plus améliorer le critère de segmentation. Toute la problématique de la construction des arbres de décision consiste à faire le bon choix de la variable de coupure et des points de coupure. Les méthodes C4.5, CART et CHAID utilisent plusieurs paramètres pour apprendre leur arbre "optimum" : paramètres de choix des variables, de coupure de variables numériques ou de groupage des données catégorielles, et paramètre de post-élagage de l'arbre. Toutefois, aucune de ces méthodes ne propose un critère global et homogène prenant en compte la structure de l'arbre, le choix des variables de coupures et les performances de l'arbre. Dans le cadre de l'étude du comportement de résiliation dans notre base, nous allons utiliser la méthode CART. Nous présentons dans la suite le principe mathématique de ce modèle et son algorithme.

### **b) Classification And Regression Tree (CART).**

Pour commencer, rappelons que parmi les arbres CART (Classification And Regression Tree), nous avons les arbres de classification et de régression ; et ces deux types de modèles se distinguent par la nature de la variable d'intérêt.

- Si cette variable est qualitative et qu'on veut prévoir la classe ou modalité à laquelle va appartenir la réponse, alors on utilisera un arbre de classification.

- Si la variable d'intérêt est quantitative, on utilisera un arbre de régression. Mais le principe reste le même : déterminer des sous-groupes dans la population selon leurs variables explicatives et dont les valeurs de la variable d'intérêt sont aussi similaires que possible.

Un arbre CART débute à partir d'une racine ou nœud initial puis se divise en deux branches conduisant à deux nouveaux nœuds qui se divisent (ou non) en de nouvelles branches et donc à des nouveaux nœuds, etc. Les nœuds terminaux, aussi appelés feuilles, sont situés en bas de l'arbre et regroupent des ensembles homogènes d'observations. Ces observations partageant des combinaisons de modalités de variables explicatives ayant un effet commun sur la variable d'intérêt et ainsi ayant des valeurs pour la variable d'intérêt aussi homogène que possible. Ces feuilles donnent une estimation de la variable d'intérêt sachant certaines valeurs des variables explicatives.

La construction d'un arbre binaire consiste à déterminer une séquence de nœuds. En effet, on commence à la racine ou nœud initial contenant l'ensemble de l'échantillon que l'on divise selon une partition des modalités d'une variable explicative en deux classes pour obtenir deux sous-ensembles de l'échantillon les plus homogènes possibles. Ces deux sous-ensembles sont des nouveaux nœuds que l'on va chercher à diviser en deux nouveaux sous-ensembles. La division se fait selon une partition des modalités d'une variable explicative autre que celle qui a déjà servi pour la première division. Ou alors selon une partition du sous-groupe de modalités créées lors de la première division. Cette répartition donne lieu à quatre nœuds terminaux que l'on peut encore diviser en deux sous-groupes, etc ; et la division se fait à l'aide d'une fonction appelée **fonctions d'hétérogénéité**.

Les fonctions d'hétérogénéité sont des fonctions positives qui croissent en fonction de l'hétérogénéité des valeurs prises par la variable d'intérêt au sein d'un même nœud. Ces fonctions diffèrent selon la nature de la variable d'intérêt.

- **Pour les arbres de classification (Variable d'intérêt discrète à  $k$  modalité) :** une fonction d'hétérogénéité est une fonction  $i : I_k \rightarrow \mathbb{R}^+$ ,  $(p_1, \dots, p_k) \mapsto i(p_1, \dots, p_k)$  avec  $I_k = \{p_1, \dots, p_k\} \in [0, 1]^k : \sum_{i=1}^k p_i = 1$ . En pratique, la probabilité  $p_{j,N}$  est estimée par  $\frac{n_{j,N}}{n_N}$  où  $n_{j,N}$  est l'effectif dans le nœud de la modalité et  $n_N$  l'effectif total du nœud  $N$ . Ainsi pour mesurer la réduction d'hétérogénéité lors de la division d'un nœud  $N$  en deux nœuds fils gauche  $N_G$  et droit  $N_D$  on applique la formule suivante :

$$\Delta = i(N) - (\mathbb{P}[N_G]i(N_G) + \mathbb{P}[N_D]i(N_D)). \quad (3.4)$$

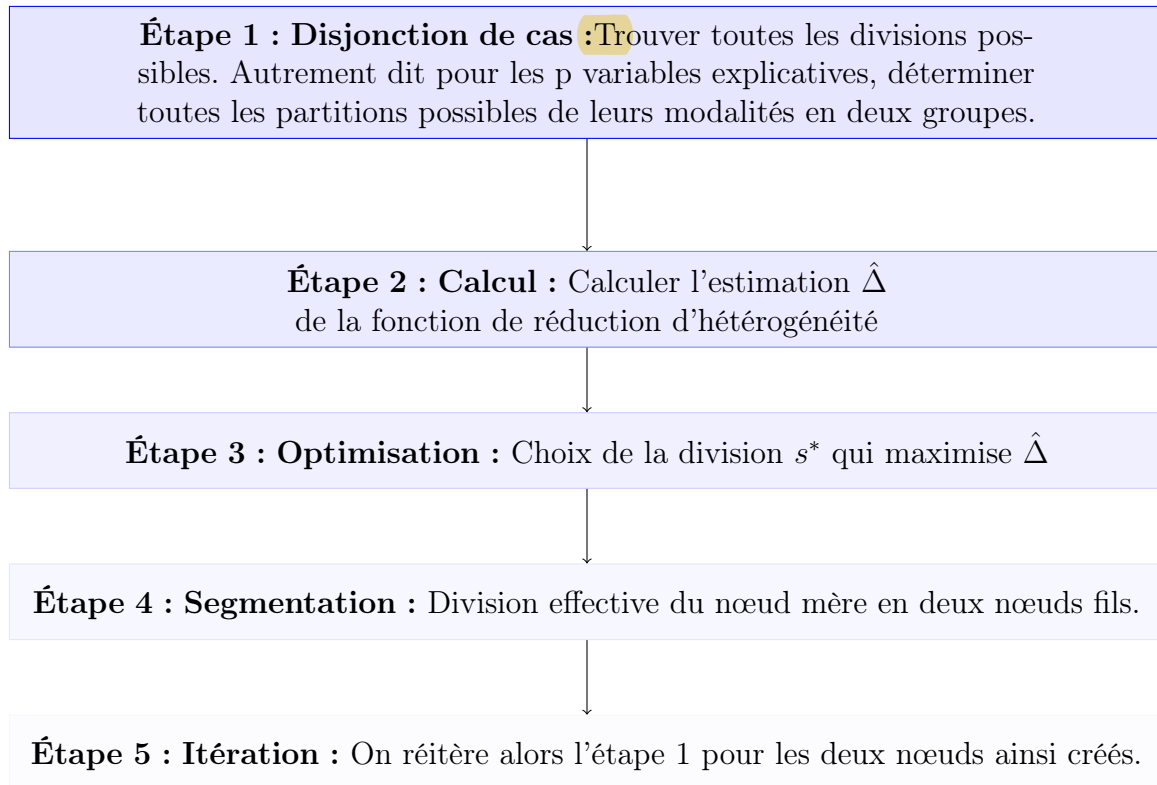
- **Pour les arbres de régression (Variable d'intérêt continue)** la fonction d'hétérogénéité est la variance intra-nœud défini par :

$$\Delta = n_N(\mathbb{V}[Y|N] - (\mathbb{P}[N_G]\mathbb{V}[Y|N_G] + \mathbb{P}[N_D]\mathbb{V}[Y|N_D])). \quad (3.5)$$

Dans l'algorithme de création d'un nœud dans un modèle CART, la division d'un nœud

revient à maximiser la réduction d'hétérogénéité, on cherche donc à tester tous les cas envisageables pour  $N_G$  et  $N_D$  afin de trouver la séparation  $s^*$  qui maximise l'estimateur  $\hat{\Delta}$  de  $\Delta$ . L'algorithme simplifié peut se présenter comme suit :

FIGURE. 3.1 – Algorithme CART.



La construction d'un arbre CART consiste donc à répéter ce schéma pour les deux nouveaux nœuds créés en utilisant seulement le sous-ensemble, de variables explicatives et modalités, générées par la division. La division optimale, qui correspond au choix d'une variable explicative et d'une partition de ses modalités, forme alors le premier nœud et conduit à de nouveaux nœuds appelés nœuds fils. On répète cette opération jusqu'à obtenir un profil par feuille (nœud terminal), et on obtient ainsi un arbre appelé **arbre saturé**.

### 3.1.3 Random forest : une méthode d'agrégation des arbres de décision.

Les arbres CART construits sont en général très instables et donc les estimations, ou prévisions produites, varient beaucoup. En effet plus l'arbre est étendu, plus les nœuds finaux (et donc les estimations) dépendent de certains profils et peuvent donner des estimations très bonnes comme très mauvaises lorsqu'il s'agit de prévoir des observations autres que celles qui ont été utilisées pour la construction de l'arbre. L'idée essentielle d'une méthode ensembliste est d'offrir une meilleure stabilité et fiabilité aux résultats, en palliant le sur-apprentissage, principal défaut des arbres de décision.

Dans une méthode d'agrégation de modèles, l'idée est de construire un grand nombre d'arbres puis de réunir les résultats pour obtenir une unique estimation. Cette procédure permet de réduire considérablement la variance de l'estimation et d'obtenir des estimateurs plus consistants. Nous utiliserons dans le cadre de notre étude le random forest (foret aléatoire) qui est un modèle d'agrégation dits parallèle c'est-à-dire qui rassemblent des arbres construits indépendamment des uns des autres.

L'algorithme des forêts aléatoires commence par la construction d'une multitude d'arbres par bootstrap (si l'on dispose d'un échantillon de taille  $n$ , on tire aléatoirement avec remise des observations pour constituer un nouvel échantillon de taille  $n_i$ , on répète cette opération un grand nombre de fois pour ainsi obtenir des échantillons à partir desquels on va construire  $B$  arbres de taille maximale). Après construction des échantillons, on tire aléatoirement un certain nombre  $m$  de variables explicatives, et on considère les divisions possibles basées sur ce sous-ensemble pour la construction des arbres saturés.

Ajouter cet aléa dans la construction des arbres permet de rendre les arbres construits plus « indépendants » et de réduire donc la variance de l'estimation. De plus, dans le cas où il y a de nombreuses variables explicatives, et particulièrement des variables explicatives avec un grand nombre de modalités, la sélection aléatoire d'un sous-groupe de variables explicatives permet d'utiliser des variables qui n'auraient peut-être pas été sélectionnées si elles avaient été confrontées à toutes les variables. Si on note  $p$  le nombre total de variables explicatives, le nombre de variables explicatives tirées aléatoirement est par défaut :

- $m = \sqrt{p}$  pour un arbre de classification ;
- $m = \frac{p}{3}$  pour un arbre de régression.

L'algorithme d'un modèle de random forest est donné dans la figure 3.2.

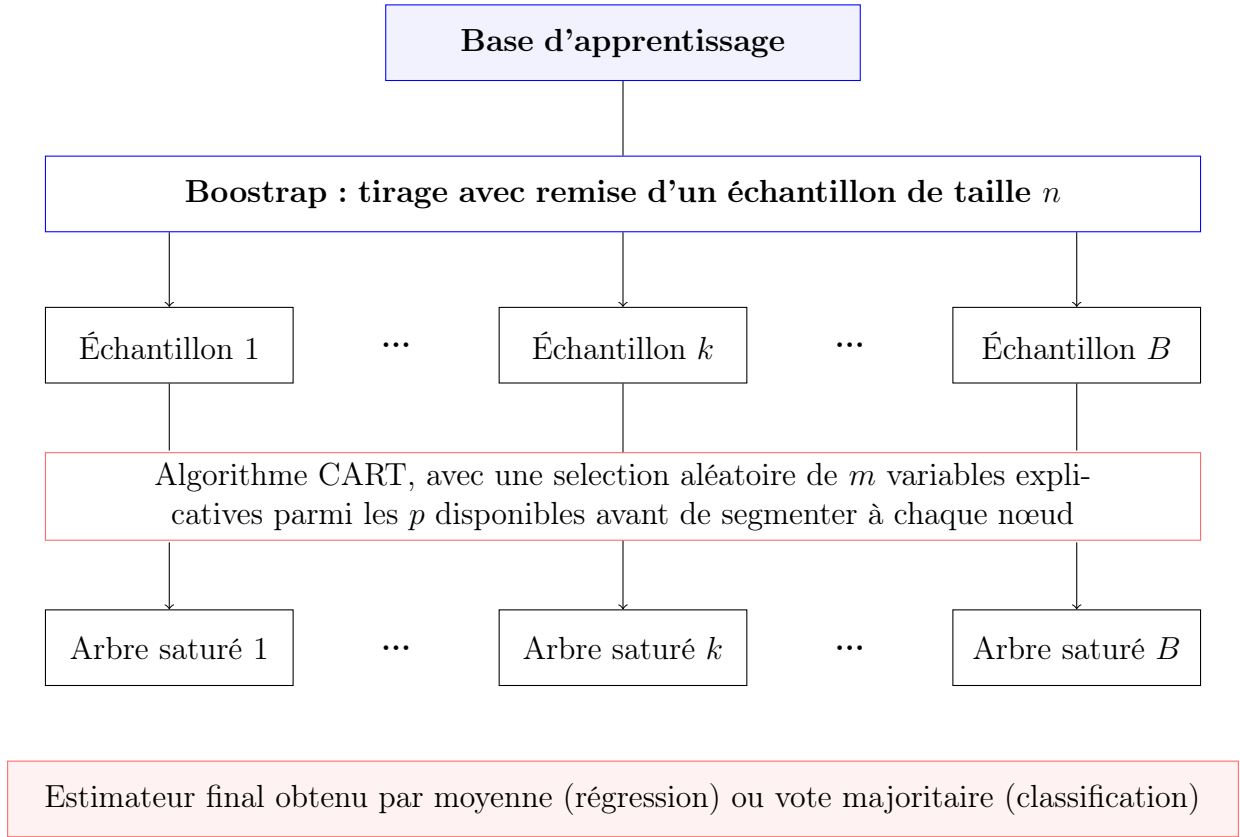
### 3.1.4 Appréciation et interprétation des modèles.

Afin d'apprécier les différents modèles implémentés, et d'évaluer celui qui nous permet d'obtenir des meilleures prédictions pour notre problématique, nous allons utiliser quelques techniques aussi bien algorithmiques, statistiques, que graphiques. Cette partie décrit l'ensemble des techniques qui seront utilisées pour la validation des modèles.

#### a) SHAP Values : une approche d'interprétabilité des modèles de machine learning.

Les algorithmes des modèles de machine learning sont parfois tellement complexes qu'ils se révèlent être de véritables boîtes noires (CART, Random Forest, les réseaux de neurones, ...). Ce qui les fait perdre en interprétabilité. Pour des problématiques métiers, cette perte d'interprétabilité est une très grande limite, car les résultats obtenus doivent pouvoir être expliqué simplement. Par exemple en assurance, il n'est pas judicieux d'utiliser un modèle boîte noire pour définir une prime d'assurance, car les souscripteurs doivent connaître comment

FIGURE. 3.2 – Algorithme Random Forest.



celle-ci est évaluée afin de pouvoir mieux expliquer au client les paramètres qui rentrent en jeu dans le calcul de sa prime. Pour modéliser la résiliation des contrats RC, nous allons utiliser les modèles logistique, les arbres de décision et les random forest. Afin de mieux comprendre la constitution des modèles, nous avons utilisé les valeurs de Shapley.

Les SHAP Values (SHapley Additive exPlanations) sont une méthode basée sur la théorie des jeux coopératifs et utilisée pour augmenter la transparence et l'interprétabilité des modèles d'apprentissage automatique. En effet, on souhaite expliquer la prédiction  $f(x)$  associée à une observation  $x$ . À chaque couple observation-prédiction  $(x, f(x))$ , est associée un jeu :

- Les joueurs sont les valeurs  $x_i$  prises par  $x$  sur chaque variable d'entrée.
- La quantité à répartir équitablement entre tous ces joueurs est la différence entre la prédiction  $f(x)$  et la moyenne des prédictions  $\mathbb{E}[f(X)]$ .
- La fonction caractéristique du jeu est :  $c(u) = \mathbb{E}[f(X)|X_u = x_u]$ , définie pour toutes variables d'entrée.

Si on a  $n$  variables d'entrée, la valeur de Shapley (encore appelée SHAP Values) associée à une modalité  $x_i$  est donc définie par :

$$\phi_i(f, x) = \sum_{u \subseteq \{1, \dots, n\} \setminus i} \frac{(n - |u| - 1)! |u|!}{n!} [\mathbb{E}[f(X)|X_{u \cup \{i\}} = x_{u \cup \{i\}}] - \mathbb{E}[f(X)|X_u = x_u]] \quad (3.6)$$

La somme des valeurs de Shapley d'une observation  $x$  est égale à l'écart entre la prévision



$f(x)$  et la moyenne des prévisions  $\mathbb{E}[f(X)]$  :

$$f(x) - \mathbb{E}[f(X)] = \sum_{i=1}^n \phi_i \quad (3.7)$$

Cette formule 3.7 montre que, le coefficient  $\phi_i$  explique comment les valeurs  $x_i$  contribuent à décaler la prévision  $f(x)$  de la moyenne  $\mathbb{E}[f(X)]$  des prévisions. Deux modalités  $x_i$  et  $x_j$  prises isolément peuvent n'avoir aucun pouvoir de prédiction, alors qu'elles peuvent être très informatives une fois couplées ensemble. L'intérêt majeur de l'approche basée sur le théorème de Shapley est donc de tenir compte des effets d'interactions, car la contribution de chaque valeur  $x_i$  est moyennée dans chacune des coalitions de valeurs  $u$  auxquelles elle appartient.

### **Estimation numériquement des SHAP Values.**

Bien qu'il soit simple de définir mathématiquement les valeurs de Shapley, leur estimation, elle, n'est pas aisée. Le calcul d'un coefficient de Shapley, en reprenant la formule 3.6 pose deux difficultés :

- Estimer les espérances conditionnelles ;
- Faire face à l'explosion combinatoire du nombre de coalitions à parcourir, lorsque le nombre  $n$  de joueurs augmente. Le nombre de coalitions à parcourir est exponentiel, en  $2^n$ .

Pour l'estimation numérique des valeurs de Shapley, la bibliothèque Python SHAP propose différents algorithmes d'estimation spécifiques à un type de modèle. Elle contient à ce jour cinq algorithmes. Nous ne rentrerons pas ici dans les détails de chaque algorithme.

- Linear Explainer : pour des modèles linéaires, souvent déjà interprétable ;
- Kernel Explainer : fonctionne pour tout type de modèle ;
- Tree Explainer : pour les modèles basés sur les arbres de décision ;
- Deep Explainer : pour les réseaux de neurones ;
- Gradient Explainer : pour les réseaux de neurones.

Dans le cadre de notre application, nous allons principalement utiliser les algorithmes Linear Explainer, Kernel Explainer et Tree Explainer de la bibliothèque SHAP.

### **b) Validation croisée.**

La cross validation ou validation croisée est une méthode statistique qui permet d'évaluer les performances des modèles d'apprentissage automatique. Lorsqu'on entraîne un modèle sur des données étiquetées, on émet l'hypothèse qu'il doit également fonctionner sur de nouvelles données. Une confirmation supplémentaire sera tout de même nécessaire pour s'assurer de l'exactitude ou non de ces prédictions. La validation croisée permet justement de vérifier si cette hypothèse est valide, ou non. On pourra ensuite choisir l'algorithme de Machine Learning approprié pour effectuer une tâche précise. Il existe plusieurs méthodes pour effectuer une

validation croisée. Les deux principales techniques sont : le **Train-Test split** et la **méthode K-Folds**.

La méthode du Train-Test Split consiste à décomposer l'ensemble de données de façon aléatoire. Une partie permettant d'entraîner le modèle de Machine Learning (environ 80%), tandis que l'autre servant pour le test de validation. C'est une technique qui se montre très efficace, exceptée si les données disponibles sont limitées. Dans ce cas, on peut se retrouver avec des informations manquantes, qui n'ont pas été utilisées pour l'entraînement. Les résultats obtenus peuvent ainsi être biaisés. Par contre, lorsque l'ensemble de données est vaste et qu'on arrive à faire une distribution égale entre les deux échantillons, l'approche Train-Test Split est tout à fait adaptée.

La limite du Train-Test Split est qu'on utilise qu'une partie du jeu de données pour entraîner, et qu'une partie pour tester ce qui peut entraîner un sur-apprentissage du modèle. La méthode **K-Folds** vient palier à cette limite ; elle va permettre d'utiliser l'intégralité du jeu de données pour l'entraînement et pour la validation. Son principe est le suivant :

- On découpe le jeu de données en  $k$  parties (folds en anglais) à peu près égales. Tour à tour, chacune des  $k$  parties est utilisée comme jeu de test. Le reste (autrement dit, l'union des  $k - 1$  autres parties) est utilisé pour l'entraînement. Dans le cas d'un problème de classification, on s'efforce à stratifier la validation croisée pour éviter d'introduire des biais.
- À la fin, chaque point (ou observation) a servi 1 fois dans un jeu de test,  $(k - 1)$  fois dans un jeu d'entraînement.
- L'évaluation des performances du modèle peut ainsi se faire soit en évaluant les prédictions faites sur l'ensemble des données, soit en moyennant les performances obtenues sur les  $k$  folds.

### c) Matrice de confusion.

La matrice de confusion est en quelque sorte un résumé des résultats de prédiction pour un problème particulier de classification. Elle compare les données réelles pour une variable cible à celles prédites par un modèle. Les prédictions justes et fausses sont révélées et réparties par classe, ce qui permet de les comparer avec des valeurs définies. La matrice de confusion permet ainsi de connaître d'une part les différentes erreurs commises par un algorithme de prédiction, mais plus important encore, de connaître les différents types d'erreurs commises. En les analysant, il est possible de déterminer les résultats qui indiquent comment ces erreurs ont eu lieu.

Les résultats d'une matrice de confusion sont classés en quatre grandes catégories : **les vrais positifs (VP)**, **les vrais négatifs (VN)**, **les faux positifs (FP)** et **les faux négatifs (FN)**. Considérons notre modèle de résiliation, la matrice de confusion va se présenter comme suit :

TABLEAU 3.1 – Matrice de confusion.

| <div> <div>Observations</div> <div>Prédictions</div> </div> | Non Résilié | Résilié |
|-------------------------------------------------------------|-------------|---------|
|                                                             | VN          | FN      |
| Non Résilié                                                 |             |         |
| Résilié                                                     | FP          | VP      |

- Les vrais positifs (VP) indiquent les cas où les prédictions et les valeurs réelles sont effectivement positives. Dans l'exemple du modèle de résiliation, cela signifie que le modèle reconnaît comme client résilié un client qui a réellement résilié.
- Les vrais négatifs (VN) indiquent par contre les cas où les prédictions et les valeurs réelles sont toutes les deux négatives. Avec le même exemple, le modèle indique alors qu'un client n'a pas résilié ce qui est bien le cas.
- Les faux positifs (FP) indiquent quant à eux une prédiction positive contraire à la valeur réelle qui est négative. Ils sont également considérés comme des **erreurs de type 1**. En poursuivant avec l'exemple, cela implique les cas où des non résiliés sont considérés comme des résiliés.
- Les faux négatifs (FN) font référence aux cas où les prédictions seraient négatives alors que les valeurs réelles sont positives. Ils sont également considérés comme des **erreurs de type 2**. Dans le cas de notre modèle de résiliation, client résilié réellement comme n'ayant pas résilié.

On remarque que la matrice de confusion nous permet de déceler 2 types d'erreur : **l'erreur de type 1** et **l'erreur de type 2**. Dans le cadre de notre modèle de résiliation, la première erreur semble moins risquée, puisqu'elle va permettre de mettre des politiques en place pour maintenir des clients qui à la base ne comptaient pas résilier. Par contre, le coût financier peut être important pour AXA s'il s'évertue à développer des politiques de retentions pour des résiliations fictives. La seconde catégorie d'erreur est particulièrement risquée. Déclarer client résilié réellement comme n'ayant pas résilié peut avoir de graves conséquences non seulement sur la rentabilité mais aussi sur la politique de fidélisation des clients. On privilégie donc dans les modèles l'erreur de type 1 par rapport à l'erreur de type 2.

Toujours dans le cadre de la matrice de confusion, différentes métriques peuvent être calculées afin de faciliter les interprétations . C'est par exemple le cas **du taux d'erreur, de la précision, du rappel et du F1 score**. Ces indicateurs permettent de mieux apprécier la qualité de précision du modèle.

- **Le taux d'erreur ou ERR** (error rate en anglais) =  $\frac{FP+FN}{VN+VP+FP+FN}$ . Plus il est bas, mieux c'est. Le meilleur taux d'erreur possible est de 0, mais il est rarement atteint par un modèle dans la pratique.
- **L'Accuracy** =  $\frac{VN+VP}{VN+VP+FP+FN} = 1 - ERR$ . Il permet d'apporter une réponse à la question suivante : de toutes les classes positives et négatives, combien parmi elles ont été prédites

correctement ? Des valeurs élevées de ce paramètre sont souvent souhaitables.

- **La précision**  $= \frac{VP}{VP+FP}$ . Ce paramètre répond donc à la question suivante : sur tous les résiliés prédits, combien sont réellement résiliés ?
- **La Sensibilité (ou recall)**  $= \frac{VP}{VP+FN}$ . Il permet de répondre à la question suivante : sur tous les résiliés, combien ont été correctement prédits ?
- **La Spécificité ou taux de vrai négatif (TVN)**  $= \frac{VN}{VN+FN}$  : mesure le ratio entre le nombre de prédictions négatives correctes et le nombre total de cas réels négatifs.
- **Le Taux de faux positif (TFP)**  $= \frac{FP}{VN+FN}$ .
- **Le score F1 (ou F-measure en anglais)**  $=$

$$\frac{2 * Précision * Sensibilité}{Précision + Sensibilité} :$$

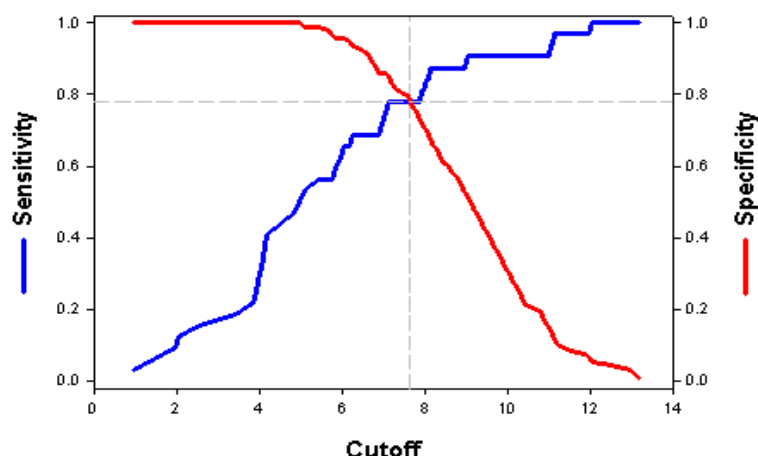
c'est une moyenne harmonique de la précision et du rappel. Sa valeur est maximale lorsque le rappel et la précision sont équivalents.

#### d) Détermination du seuil optimal : cutt-off.

Dans la plupart des logiciels, les modèles de classification prédisent des variables latentes. Par exemple dans un modèle logistique, on obtient en sortie :  $Y^* = X' \beta_0 + \varepsilon$ . L'idée est donc de déterminer le seuil  $s$  telle que  $Y = \mathbb{1}\{Y^* \geq s\}$  qui permet d'avoir une bonne répartition entre les  $FP, FN, VP$  et  $VN$ . En effet si  $s$  est trop petit, le modèle va classer tous les individus en positif, si  $s$  est trop grand, tous les individus sont classés en négatif. En pratique pour déterminer  $s$  optimal, on utilise le procédé suivant :

- On fait varier  $s$  entre  $Y_{min}^*$  et  $Y_{max}^*$  ;
- Pour chaque valeur de  $s$  on calcule *la spécificité* et *la sensibilité* ;
- La valeur optimale de  $s$  s'obtient quand *sensibilité* = *spécificité*.

FIGURE. 3.3 – Détermination du cutt-off optimal.

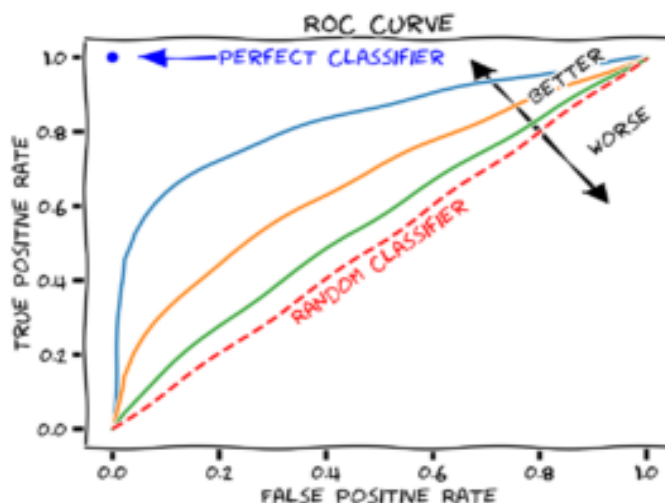


Source : Stack Overflow.

#### e) Courbes ROC et AUC.

La courbe ROC (Receiver Operating Characteristic) représente la *sensibilité* en fonction de  $1 - \text{spécificité}$  pour toutes les valeurs seuils possibles. Cette courbe sert à comparer différents classifieurs. Plus une courbe a des valeurs élevées, plus l'aire sous la courbe est grande, moins le classifieur fait d'erreur.

FIGURE. 3.4 – Courbe ROC.



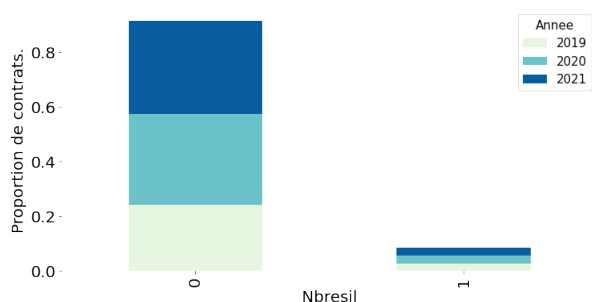
Source : IDBC.

L'aire sous la courbe ROC (ou Area Under the Curve, AUC) peut être interprétée comme la probabilité que, parmi deux contrats choisis au hasard, un résilié et un non-résilié, la valeur du marqueur soit plus élevée pour le résilié que pour le non-résilié. Par conséquent, une AUC de 0.5(50%) indique que le marqueur est non-informatif. Une augmentation de l'AUC indique une amélioration des capacités discriminatoires, avec un maximum de 1(100%).

## 3.2 Présentation des résultats.

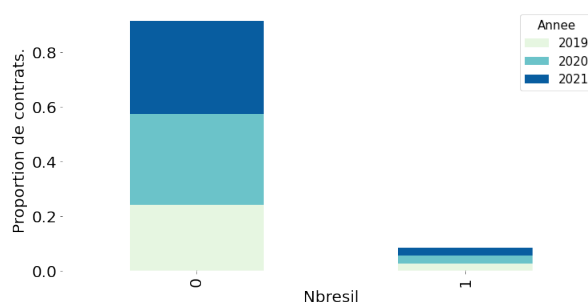
En utilisant la base de données définitive constituée des variables présentées dans le tableau 2.3, nous avons dans un premier temps segmenté notre base en 90/10 (90% de la base pour l'entraînement et 10% pour le test). Ci-dessous les figures 3.5 et 3.6 donnent la répartition de la résiliation dans les échantillons train et test. On a bien une même proportion de résiliés dans les deux échantillons train et test ce qui est nous garantit sur la représentativité de ces échantillons.

FIGURE. 3.5 – Distribution résiliation dans l'échantillon train.



Source : Interne AXA.

FIGURE. 3.6 – Distribution résiliation dans l'échantillon test.



Source : Interne AXA.

### 3.2.1 Modélisation de la résiliation par un modèle logistique.

La modélisation de la résiliation des clients RC par les modèles logistique a été implémenté avec le logiciel AKUR8<sup>1</sup>. Deux grandes classes de modèles ont été implémentées, une première portant sur l'ensemble des clients du portefeuille et une seconde portant uniquement sur les clients mutli-équipés du portefeuille. Nous présentons dans le tableau 3.2 les résultats des deux meilleurs modèles.

TABLEAU 3.2 – Résultats du modèle de résiliation des clients RC.

| Modèles           | Nomdre de variables | AUC    | Remarques                                                      |
|-------------------|---------------------|--------|----------------------------------------------------------------|
| Modèle mono+multi | 9                   | 61.80% | Modèle sur ma totalité des clients RC de la base de données.   |
| Modèle multi      | 10                  | 61.51% | Modèle sur les clients RC multi-équipés de la base de données. |

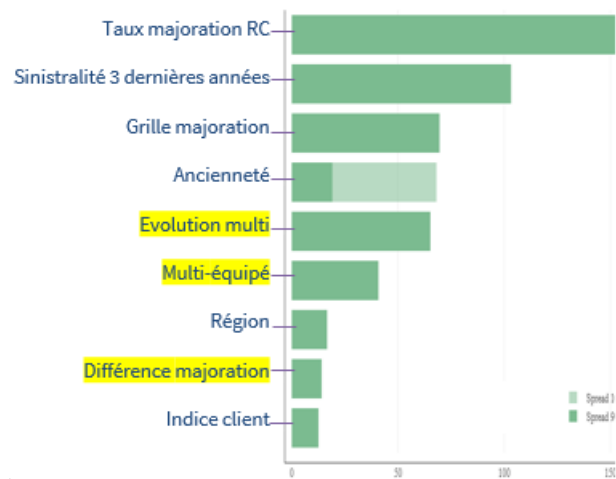
Source : Interne AXA.

Le tableau 3.2 ci-dessus nous indique que sur l'ensemble des variables explicatives, 9 se sont avérées significatives pour l'explication de la résiliation du modèle global avec un pouvoir explicatif d'environ 62% (ce qui signifie que le modèle permet de discriminer avec une probabilité

1. AKUR8 est un outil utilisé en interne pour la tarification.

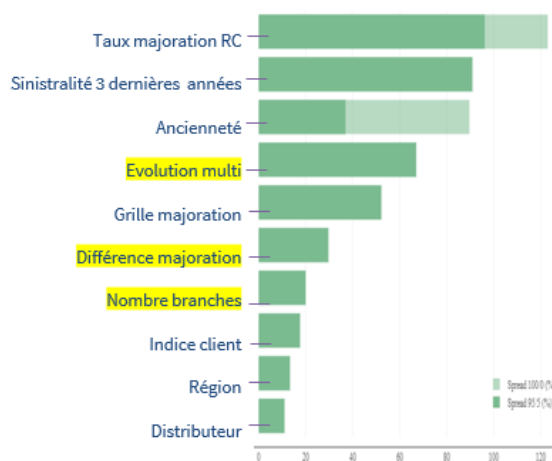
de 0.62 les résiliés des non résiliés). Quant au modèle de résiliation des clients multi-équipés, on a 10 variables qui se sont avérées significatives et toujours avec un pouvoir explicatif d'environ 62%. Les figures 3.7 et 3.8 ci-dessous présentent l'ensemble de ces variables classées par ordre d'importance dans le modèle.

FIGURE. 3.7 – Modèle de résiliation mono+multi : variables explicatives.



Source : Interne AXA.

FIGURE. 3.8 – Modèle de résiliation multi : variables explicatives.



Source : Interne AXA.

Ces figures 3.7 et 3.8 montrent que le taux de majoration RC et la sinistralité des trois dernières années sont les variables qui expliquent le plus la résiliation tant dans la population globale que dans le groupe des multi-équipés. D'autre part, les variables du multi-équipement sont pertinentes pour l'analyse de la résiliation des contrats RC (multi-équipé, Différence de majoration, Nombre de branches, évolution multi-équipement). Pour mieux comprendre ces deux modèles, nous allons par la suite analyser chacune de ces variables et procéder à l'analyse des prédictions effectuées sur la base de test.

#### a) Consistance dans le temps : time consistency.

Notre base de données porte sur les clients RC de 2019 à 2021, ainsi pour vérifier la stabilité et la robustesse des résultats dans le temps, il est important de vérifier la consistance de chaque variable explicative dans le temps. Le tableau 3.3 nous renseigne sur ce sujet. On constate qu'en dehors de l'indice, toutes les autres variables de notre modèle sont consistantes dans le temps. Ceci nous montre que les conclusions qui seront tirées des analyses pourront rester vraies dans le temps si la composition du portefeuille ne subit pas de changement majeur.

TABLEAU 3.3 – Résultats du modèle de résiliation des clients RC.

| Variables                       | Modèle mono+multi                                       | Modèle multi          |
|---------------------------------|---------------------------------------------------------|-----------------------|
| Taux majoration RC              | OK                                                      | OK                    |
| Sinistralité 3 dernières années | OK                                                      | OK                    |
| Grille majoration               | OK                                                      | OK                    |
| Région                          | OK                                                      | OK                    |
| Âge contrat                     | OK                                                      | OK                    |
| Indice                          | Non pour indice=3 (Même tendance mais spread important) | Non pour indice=3 ... |
| Distributeur                    | OK                                                      | OK                    |
| Évolution multi-équipement      | OK                                                      | OK                    |
| Multi-équipé                    | OK                                                      | -                     |
| Différence de majoration        | -                                                       | OK                    |
| Nombre de branches              | -                                                       | OK                    |

**Source** : Interne AXA.

## b) Analyse des variables explicatives des modèles de résiliation.

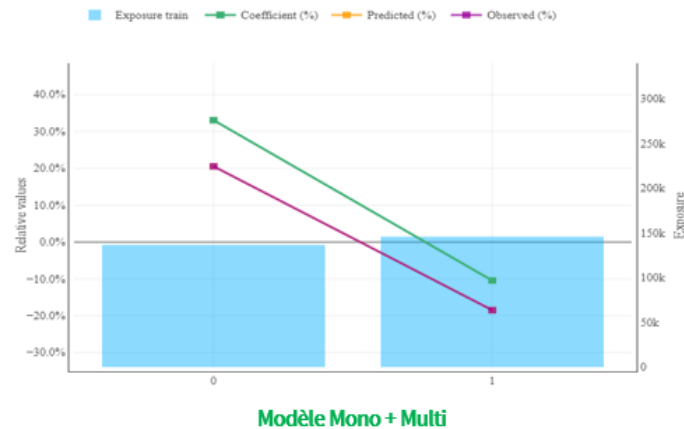
Dans cette partie nous allons analyser les modèles de résiliation en étudiant l'évolution de la résiliation au sein de chaque variable. A titre de rappelle la base globale est constituée de 352 mille contrats RC avec un chiffre d'affaires de 836 millions d'euros. Et la base restreinte sur les clients multi-équipés compte 182 mille contrats RC avec un chiffre d'affaires de 548 millions d'euros. Quand on inclut les contrats des autres branches de ces clients multi-équipés, on est à 953 mille contrats et un chiffre d'affaire de 3.165 milliards d'euros.

Sur les figures 3.10 à 3.17 on a les informations suivantes :

- Les histogrammes en bleu indiquent l'effectif des contrats RC de chaque modalité.
- Les courbes en vert donnent les coefficients des modalités dans le modèle de résiliation. Pour les clients non résiliés (voir figure 3.9), ce coefficient est autour de 32% ce qui signifie que toutes choses égales par ailleurs, les clients mono-équipés ont 32% de plus de chance de résiliés que les multi-équipés.
- Les courbes en jaune donnent les prédictions du taux de résiliation dans les différentes modalités.
- Les courbes en violet donnent les valeurs observées du taux de résiliation dans les différentes modalités. On remarque que ces courbes sont superposées sur les courbes en jaune ce qui renseigne sur la bonne qualité des prédictions.
- Pour toutes les variables présentes dans le modèle de résiliation de l'ensemble du portefeuille et dans le modèle du groupe des multi-équipés, nous avons représenté en :
  - (A) La répartition du taux de majoration RC dans le modèle mono+multi.
  - (B) La répartition du taux de majoration RC dans le modèle multi.



FIGURE. 3.9 – Modèle résiliation : Multi-équipement.



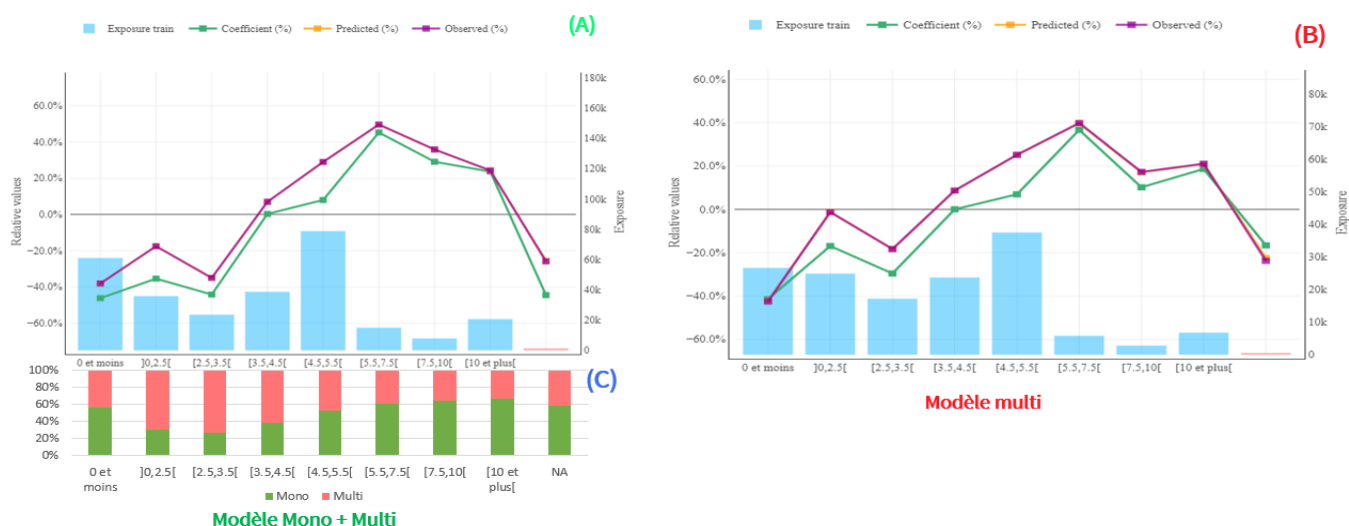
Source : Interne AXA.

- (C) La proportion mono vs multi dans la base de données.

#### b-1) Taux de majoration RC.

Dans les figures 3.10 A et B donnant l'évolution des modèles de résiliation en fonction du taux de majoration RC, on constate que le taux de résiliation augmente avec le taux de majoration RC. Par ailleurs lorsque la majoration dépasse 4.5%, la probabilité de résilier devient très grande aussi bien sur l'ensemble du portefeuille (3.10 A) que chez les multi-équipés (3.10 B). Cette conclusion s'explique par le fait que plus le taux de majoration est élevé, plus la prime d'assurance du client devient très forte, et donc logiquement il aura tendance à aller vers la concurrence si celle-ci propose des prix plus attractifs. Toutefois on remarque que pour les taux de majoration très élevés (supérieure ou égale à 7.5%) la probabilité de résiliation des clients du portefeuille chute, et cette chute est encore plus prononcée chez les clients multi-équipés (3.10 B).

FIGURE. 3.10 – Modèle résiliation : Taux de majoration RC.

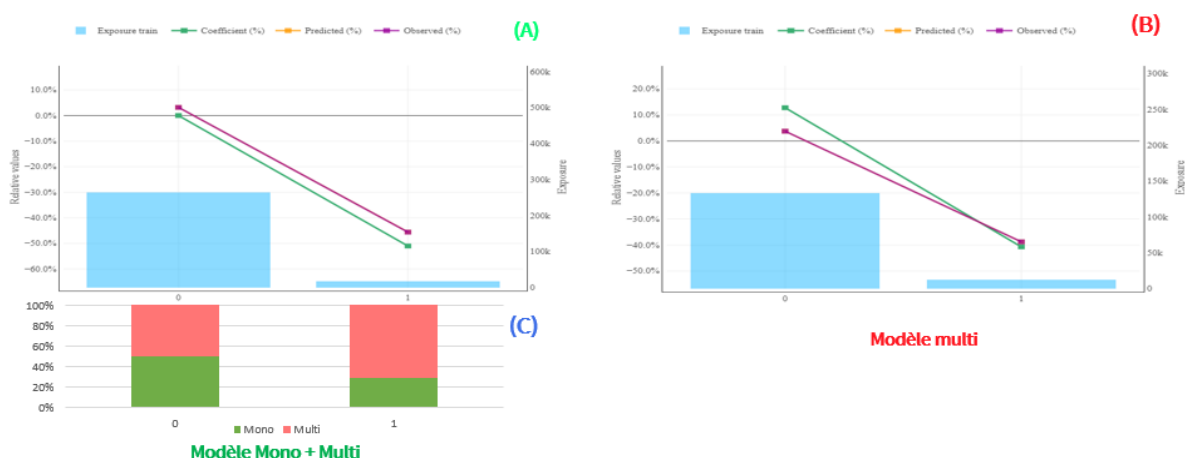


Source : Interne AXA.

### b-2) Sinistralité des trois dernières années.

La figure 3.11 B donnant l'évolution des modèles de résiliation en fonction de la sinistralité montre que les multi-équipés non sinistrés ont plus tendance à résilier. Toutes choses égales par ailleurs, leurs taux de résiliation dépassent de 10% celui des clients sinistrés. Contrairement à ce qu'on aurait imaginé, on s'aperçoit que ce sont les clients sinistrés qui sont les plus fidèles de notre portefeuille!!! Ce résultat pourrait surprendre mais non, en effet chaque client entreprise possède un relevé d'information qui retrace l'historique de sa sinistralité. Ce document permet donc aux assureurs de connaître les comportements des clients dès la souscription et de pouvoir donc proposer des tarifs adaptés aux degrés d'exposition aux risques. Ainsi, les clients sinistrés savent d'avance qu'ils seront pénalisés s'ils décident de résilier leurs contrats pour la concurrence donc préfèrent rester chez leurs anciens assureurs.

FIGURE. 3.11 – Modèle résiliation : Sinistralité 3A.

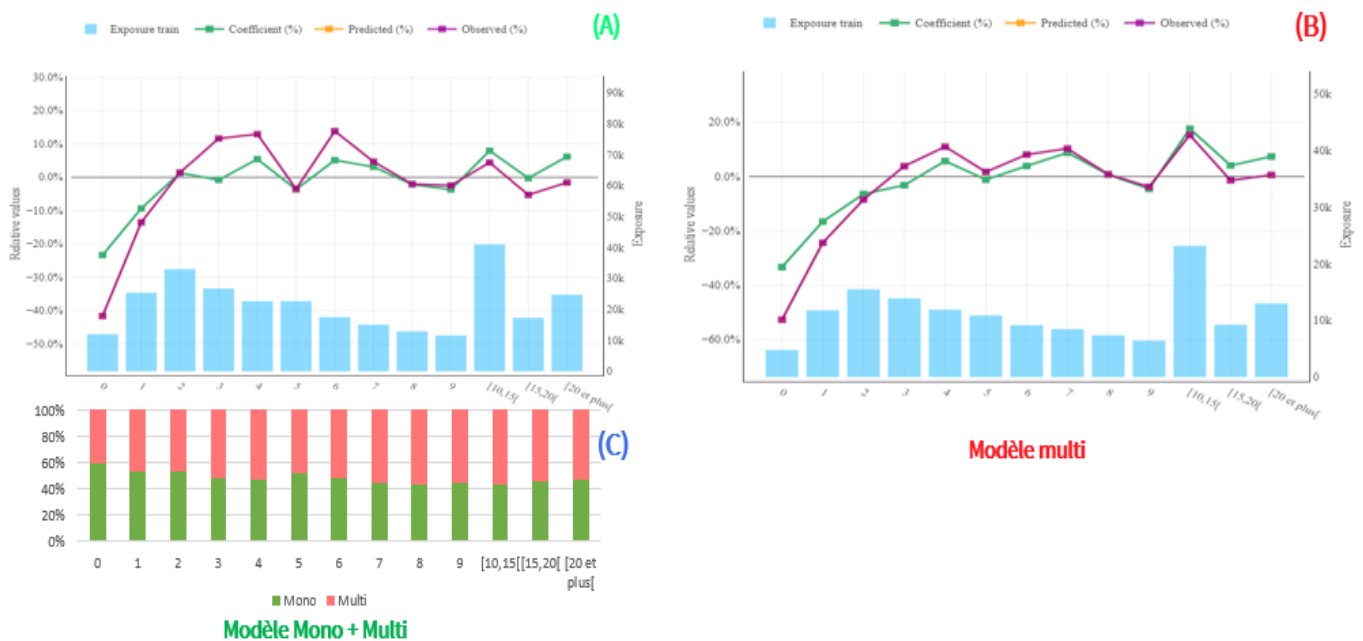


Source : Interne AXA.

### b-3) Âge contrat.

Au sein de notre modèle de résiliation, on s'aperçoit que le taux de résiliation augmente avec l'âge du contrat (figure 3.12 A et B) et ceci tant dans la population globale (3.12 A) que dans le groupe des multi-équipés (3.12 B). Par ailleurs on constate que les clients multi-équipés sont plus fidèles. En effet, toutes choses égales par ailleurs, les clients RC (3.12 A) résilient à partir de deux ans d'ancienneté au sein du portefeuille. Or dans le groupe des multi-équipés (3.12 B), la résiliation n'est fréquente qu'à partir de la quatrième année du contrat. Ce résultat nous indique qu'une bonne politique de fidélisation des clients RC multi-équipés devrait s'attarder davantage sur les clients ayant plus de 3 ans d'ancienneté dans le portefeuille d'AXA.

FIGURE. 3.12 – Modèle résiliation : Âge contrat.



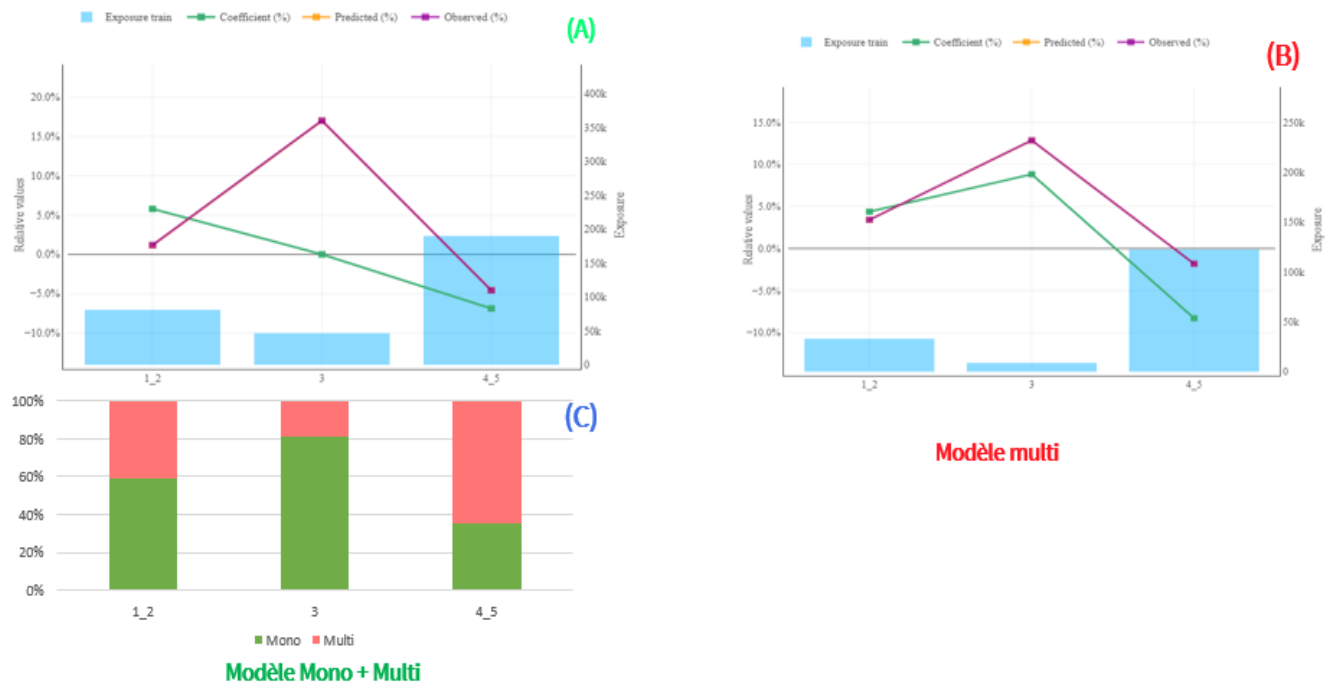
Source : Interne AXA.

#### b-4) Indice.

Pour ce qui est de l'indice (figures 3.13 A et B), notre modèle de résiliation nous indique que les clients d'indice 3 sont ceux qui résilient le plus et ceci que ce soit dans la population globale (3.13 A) que dans le groupe des multi-équipés (3.13 B). D'autre part, on constate que la propension à résilier des clients d'indice 1 et 2 est plus forte que celle des clients d'indice 4 et 5. En effet au sein des multi-équipés (3.13 B), toutes choses égales par ailleurs les clients d'indice 1 et 2 ont 2.5% de plus de chance de résilier. Par ailleurs, lorsqu'on regarde dans le groupe d'indice 4 et 5, la probabilité de ne pas résilier est plus faible chez les multi-équipés. On a en effet dans le groupe des clients d'indice 4 et 5 multi-équipés (3.13 A) une probabilité de ne pas résilier autour de 2.5% contre une probabilité de 5% sur l'ensemble de l'échantillon. Or avec une proportion de près de 70%, la part des multi-équipés est plus forte chez ces clients (voir figure C de la figure 3.13). Ceci nous montre que la résiliation des clients d'indice 4 et 5 au sein des multi-équipés est assez particulière et devrait être traitée avec plus de précaution car en effet en

plus d'être des clients très rentables pour AXA, ils détiennent le gros du chiffre d'affaires (plus de 70%, voir figure 2.8) donc les politiques de fidélisation de la clientèle devraient s'accroître autour d'eux afin de diminuer au mieux le taux de résiliation.

FIGURE. 3.13 – Modèle résiliation : Indice.

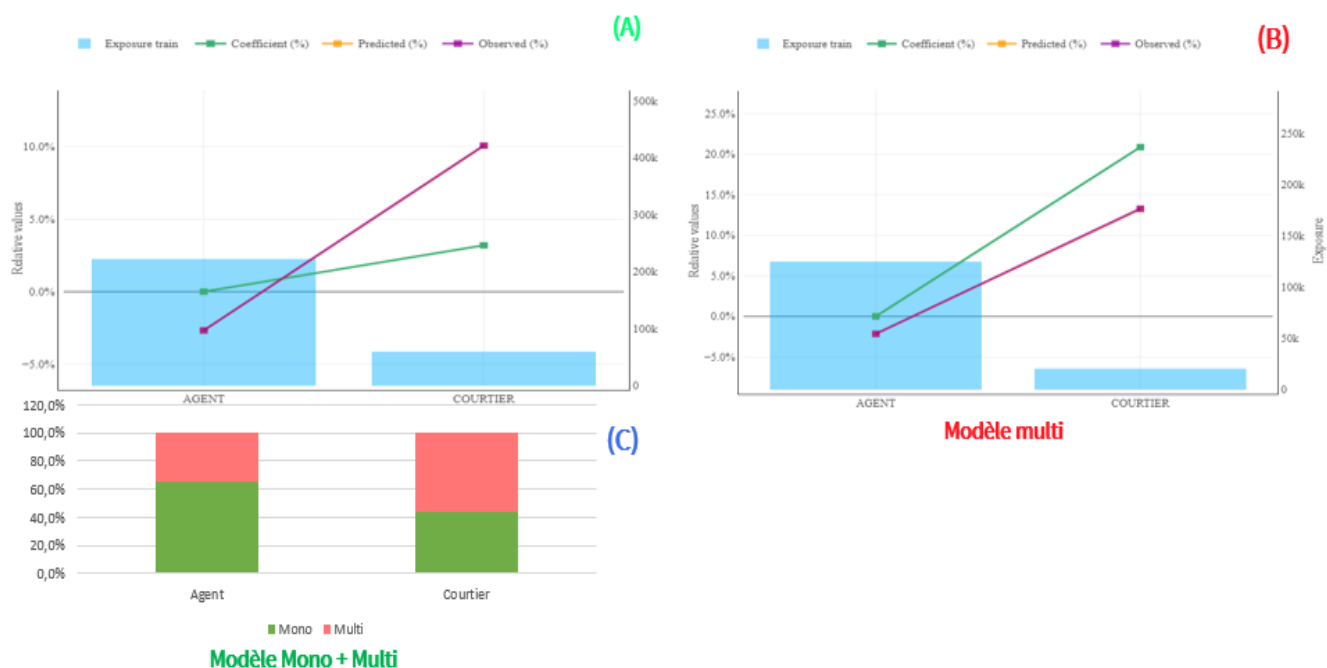


Source : Interne AXA.

### b-5) Distributeur.

Dans le modèle de résiliation, le distributeur apparaît comme un facteur de résiliation que ce soit dans la population globale que dans le groupe des multi-équipés (figure 3.14 B). Plus précisément, toutes choses égales par ailleurs, le taux de résiliation chez les courtiers dépasse d'environ 10% celui des agents. L'une des raisons qui explique cette différence est que les courtiers distribuent les produits d'assurance de plusieurs assureurs. Lorsqu'AXA propose un prix élevé pour un client donné, le courtier peut dans le but de conserver sa clientèle, comparer avec les offres de la concurrence et ainsi offrir au client l'offre la moins onéreuse. Cette latitude qu'a le courtier d'avoir une vision globale du marché peut conduire ses clients de facilement passer d'AXA pour un autre assureur d'où la hausse du taux de résiliation des clients détenus dans le portefeuille du courtier. Les analyses descriptives nous ont montré que les courtiers détenaient les plus gros clients du portefeuille (plus 45% de chiffre d'affaires pour moins de 30% des clients voir figures 2.10 et 2.11). Donc la fidélisation des clients RC Entreprise du portefeuille des courtiers devient un enjeu important.

FIGURE. 3.14 – Modèle résiliation : Distributeur.

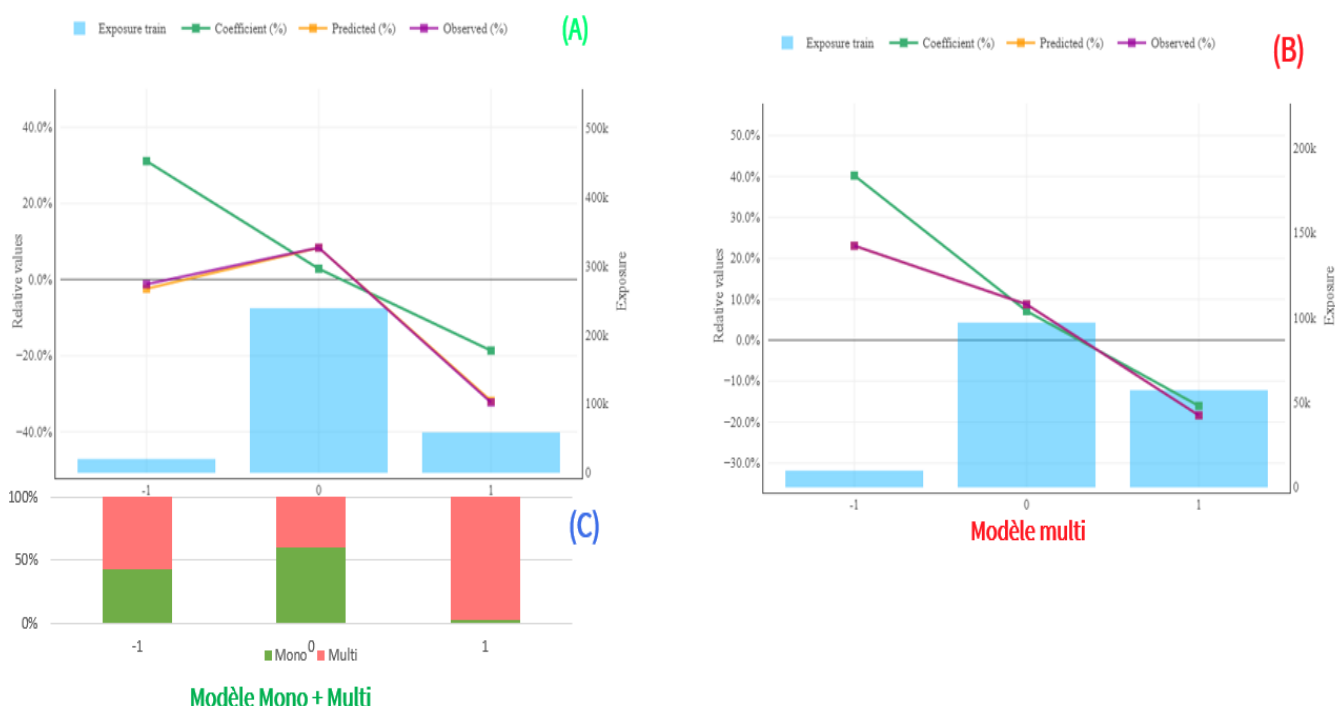


Source : Interne AXA.

#### b-6) Évolution multi-équipement.

Le modèle de résiliation nous montre que concernant l'évolution du multi-équipement (figures 3.15 A et B), les clients en décroissance de portefeuille (les clients ayant résilié au moins un de leur contrat de l'année N à l'année N-1) sont ceux pour lesquelles les risques de résiliation sont très élevés. En effet on constate que la probabilité de résilier chez ces derniers est environ 25% plus élevée que la résiliation chez les autres clients. A l'inverse les client en croissance de portefeuille (les clients ayant souscrit à au moins un nouveau contrat de l'année N à l'année N-1) ont 20% de chance de plus de ne pas résilier. Ces résultats nous montrent qu'une bonne politique de fidélisation des clients multi-équipés devrait davantage s'attarder sur les clients en décroissance de portefeuille.

FIGURE. 3.15 – Modèle résiliation : Évolution multi-équipement.

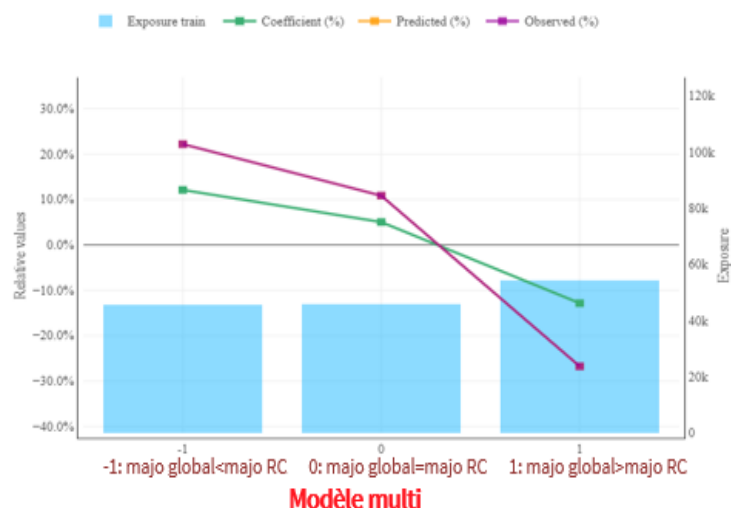


Source : Interne AXA.

#### b-7) Différence de majoration.

Le modèle de résiliation nous indique également que la diff-majo (position du taux de majoration globale par rapport à la majoration RC des clients multi-équipés) est un facteur important de résiliation chez les clients multi-équipés (voir figure 3.16). Plus spécifiquement, le modèle nous enseigne qu'à caractéristique égale, les clients multi-équipés ayant reçu une majoration RC supérieure à la majoration globale de l'ensemble de leurs contrats (diff\_majo = -1) sont avec une probabilité de résiliation d'environ 20% ceux qui sont les plus susceptibles à résilier (3.16). A l'inverse, les clients multi-équipés de majoration RC inférieure à la majoration globale (diff\_majo=1) sont les clients les plus fidèles du portefeuille avec une probabilité de ne pas résilier d'environ 30%. Nous avons également remarquer en analysant le taux de majoration RC dans le modèle (voir figure 3.10) que le taux de résiliation augmentait fortement avec le taux majoration. Alors on peut se demander si les clients de majoration RC supérieure à la majoration globale ne résilient pas du fait de leur majoration RC très élevée à la base? Pour répondre à cette interrogation nous allons introduire dans notre modèle de résiliation les variables croisées.

FIGURE. 3.16 – Modèle résiliation : Différence de majoration.

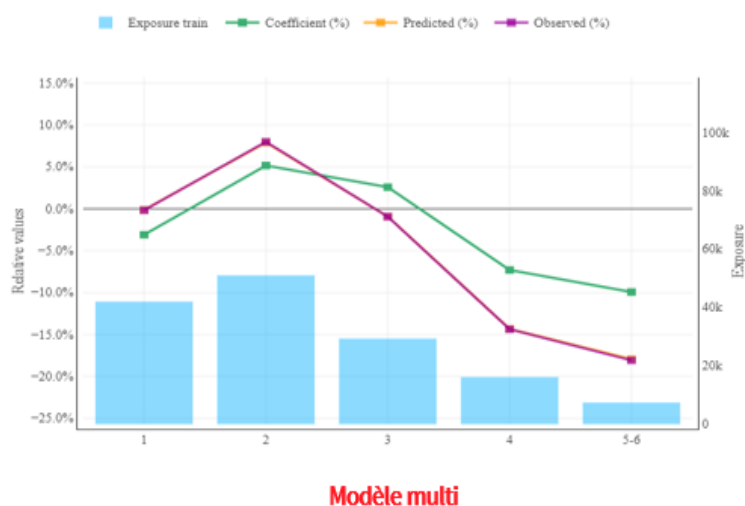


Source : Interne AXA.

#### b-8) Nombre de branches.

Dans le modèle de résiliation des clients multi-équipés, le nombre de branches du client apparaît également comme un critère principale de résiliation (figure 3.17). On constate que plus le poids du multi-équipement est important plus le taux de résiliation diminue. En d'autre termes, les clients multi-équipés ayant des contrats sur plusieurs branches sont ceux qui résilient le moins leur contrat RC. Ce résultat nous montre qu'encourager les clients multi-équipés à diversifier leurs portefeuilles en souscrivant à plusieurs branches pourrait être un bon moyen de fidélisation de la clientèle.

FIGURE. 3.17 – Modèle résiliation : Nombre de branches.



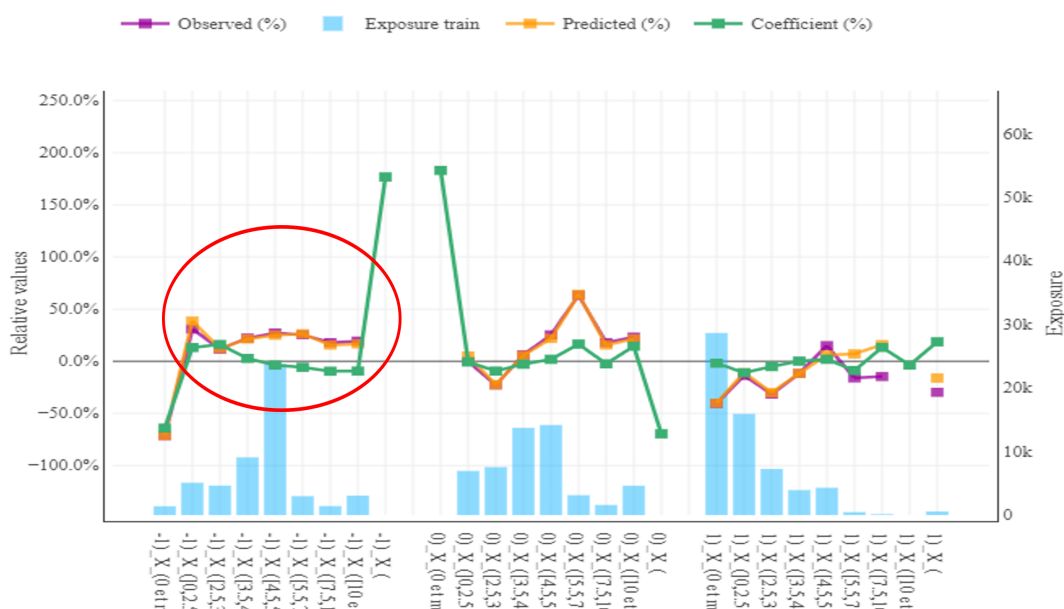
Source : Interne AXA.

### c) Modèle de résiliation avec interaction entre les variables.

En analysant les variables explicatives des modèles de résiliation présentés en 3.7, nous avons soupçonné l'existence de quelques liaisons dans les variables. En particulier, en étudiant le comportement de résiliation en fonction de la majoration RC (voir figure 3.10), on s'est demandé si les clients qui recevaient de fortes majorations étaient en particulier les clients d'indice 1 et 2 du portefeuille ? Puis ensuite la forte résiliation des clients de  $\text{diff\_majo} = -1$  (majoration RC supérieure à la majoration globale) nous a fait nous poser la question de savoir si ces clients résiliaient du fait de leurs majorations RC initialement très fortes ? Pour répondre à ces interrogations nous avons introduit dans le modèle de résiliation des multi-équipés les interactions entre les variables (taux majoration vs Indice ; taux majoration RC vs Diff\_majo ; ...).

De l'ensemble des interactions introduites, une seule s'est avérée significative à savoir l'interaction entre **le taux de majoration RC et la variable Différence majoration**. la figure 3.18 présentant les résultats de cette interaction nous montre que, comme nous l'avons soupçonné au départ, les clients ayant reçu une majoration RC supérieure ou égale à leur majoration globale ( $\text{Diff\_majo} = -1$  ou  $\text{Diff\_majo} = 0$ ) ont dans l'ensemble reçu des fortes majorations RC (l'histogramme en bleu montre que la grande majorité de ces clients ont des majorations supérieures à 3.5%). Donc on comprend qu'une grande partie des clients de  $\text{Diff\_majo} = -1$  résilient fortement du fait des taux de majoration RC très élevés qu'ils ont reçu à la base.

FIGURE. 3.18 – Modèle résiliation : Diff\_majo vs Taux majoration RC.



Source : Interne AXA.

### d) Analyse des prédictions du modèle de résiliation.

Afin d'évaluer la pertinence de notre modèle de résiliation, nous avons réalisé des prédictions sur notre échantillon de test. Pour simplifier les analyses, nous nous sommes concentrés sur les

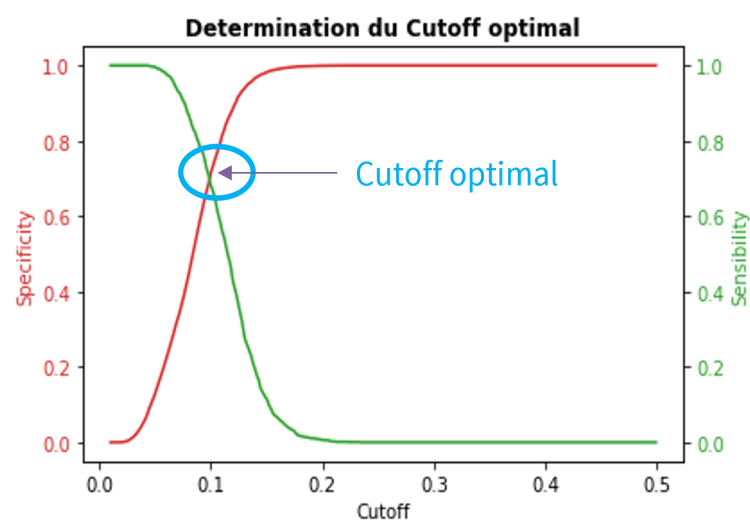


clients multi-équipés. Cette partie sera consacrée à la présentation des résultats obtenus.

**d-1) Choix du seuil optimal.**

En utilisant la formule (3.3), les coefficients du modèle de résiliation nous ont permis de prédire la probabilité de résiliation des clients de la base test. Une fois ces probabilités calculées, nous avons utilisé la méthode du **Cutoff** décrite à la section 3.1.4 pour déterminer le seuil optimal de résiliation. la figure 3.19 nous donne l'évolution de la sensibilité en fonction de la spécificité de notre modèle. L'analyse de la figure nous montre que la probabilité minimale de résiliation se trouve autour de 1.3%, et la maximale à 22.45%. Le croisement des courbes de sensibilité et spécificité se fait au seuil de 9.7% qui correspond au seuil optimal de résiliation dans notre modèle. En d'autre termes, un client ayant une probabilité de résiliation supérieure à 9.7% sera considéré comme résilié.

FIGURE. 3.19 – Seuil optimal de résiliation.



Source : Interne AXA.

**d-2) Matrice de confusion.**

A l'aide du seuil optimal de résiliation précédemment déterminé, nous avons construit la matrice de confusion des prédictions (base test) en pourcentage du nombre de contrats (voir tableau 3.4) puis en pourcentage du chiffre d'affaires (voir tableau 3.5).

TABLEAU 3.4 – Matrice de confusion en pourcentage du nombre de contrats.

| <b>Prédictions \ Observations</b> | <b>Observations</b> |             | <b>Total</b> |
|-----------------------------------|---------------------|-------------|--------------|
|                                   | Non Résilié         | Résilié     |              |
| Non Résilié                       | 58.7%               | 2.1%        | 60.9%        |
| Résilié                           | 32.5%               | 6.6%        | 39.1%        |
| <b>Total</b>                      | <b>91.3%</b>        | <b>8.7%</b> | <b>100%</b>  |

Source : Interne AXA.

TABLEAU 3.5 – Matrice de confusion en pourcentage de la prime.

| <b>Observations</b><br><b>Prédictions</b> | Non Résilié  | Résilié      | <b>Total</b> |
|-------------------------------------------|--------------|--------------|--------------|
| Non Résilié                               | 68.7%        | 4.0%         | 72.7%        |
| Résilié                                   | 21.1%        | 6.2%         | 27.3%        |
| <b>Total</b>                              | <b>89.8%</b> | <b>10.2%</b> | <b>100%</b>  |

**Source : Interne AXA.**

Ces résultats en tableau 3.4 et 3.5 nous montrent qu'en terme de nombre de contrats on a une accuracy de 65.3% et de 75% en terme de chiffre d'affaires. Ceci signifie qu'en effet notre modèle prédit à 65.3% le comportement de résiliation des clients de notre portefeuille, et ces 65.3% de "bonne prédiction" possèdent 75% du chiffre d'affaires de notre portefeuille. Une autre statistique importante à mentionner est que l'erreur du type 2 est très faible dans notre modèle. En effet on commet cette erreur dans notre modèle juste dans 2% des cas ( 2% clients sont déclarés comme non résiliés alors qu'ils ont résilié). En plus dans notre modèle de résiliation, nous avons une sensibilité d'environ 75%, ce qui signifie que 3 clients résiliés sur 4 ont été correctement prédits.

### **d-3) Caractérisation des différents groupe.**

A présent que nous avons construit notre modèle de résiliation et évaluer les différentes performances, il est important d'explorer davantage la matrice de confusion afin de savoir quelles sont les caractéristiques des clients qui se situent dans chaque classe. En d'autre termes on aimerait savoir sur quel groupe de client on se trompe le moins.

Pour une première approche de caractérisation, nous avons faits les analyses variable par variable afin de voir quel était la répartition des VP, FN, FP et VN au sein de chacune des modalités (voir figures A.1 à A.6 en annexe A.2). Cette première approche nous a permis de repérer les variables pour lesquelles chacun de ces groupes se démarquait le plus des autres :

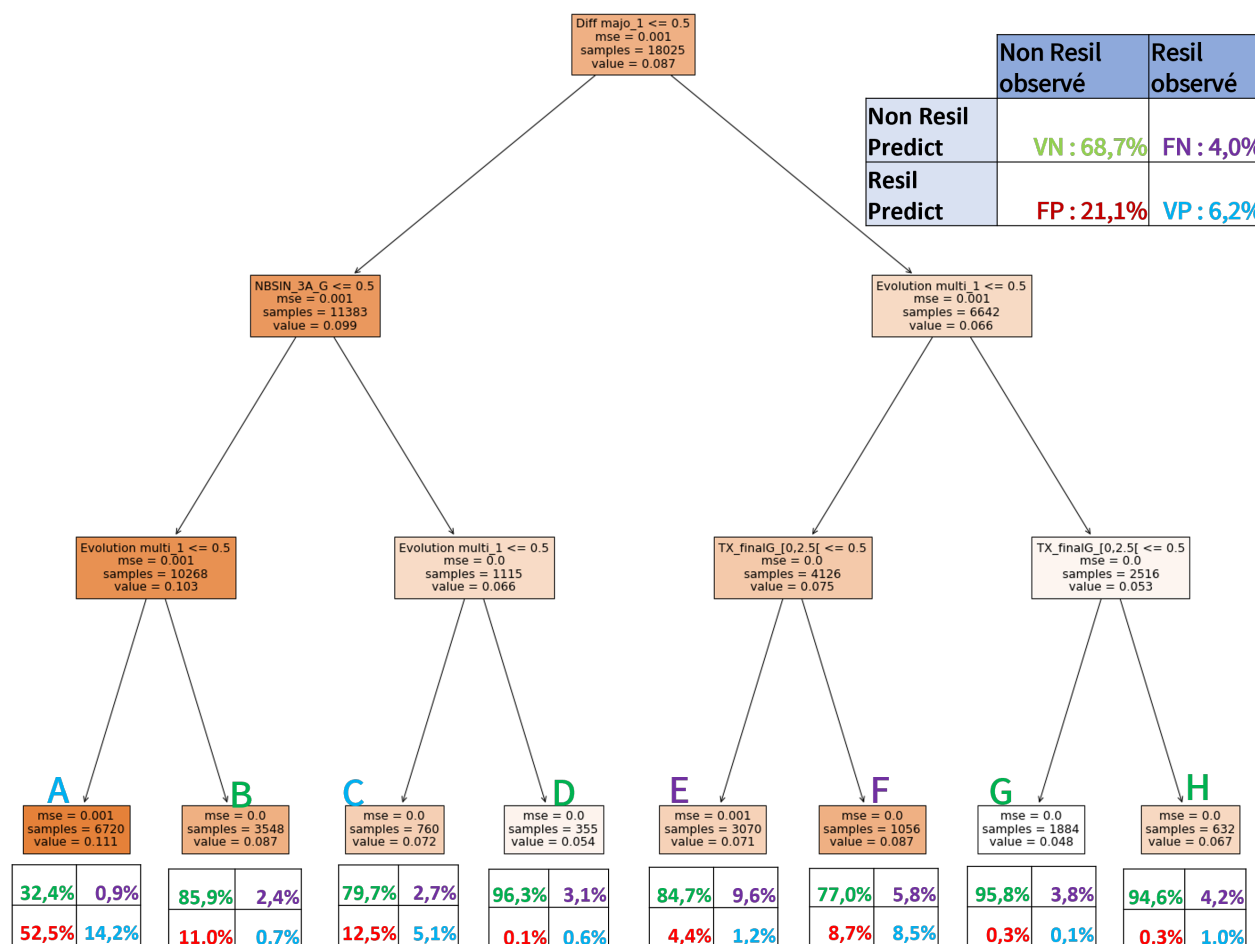
- La prédiction de la résiliation des clients en croissance de portefeuille est satisfaisante et le taux de résiliation de ces derniers sont faibles. Par ailleurs, les résiliés sont majoritairement dans le groupe des clients en décroissance de portefeuille et le modèle se trompe grandement pour ces derniers.
- Le modèle se trompe grandement sur la prédiction de la résiliation des clients sinistrés.
- On a moins de chances de prédire le correctement la résiliation des clients ayant plus de 4 branches.
- La prédiction semble meilleure pour les clients d'indice 3 et 1\_2.

Cette première approche reste très uni-variée or on aimerait avoir une caractérisation beaucoup plus large qui permettrait d'identifier de façon spécifique les clients. Pour corriger la limite

de cette première approche nous avons réalisé des caractérisations multi-variées en implémentant un arbre CART de profondeur 4 sur les prédictions. la figure 3.20 nous montre que la différence de majoration et l'évolution du multi-équipement sont les 2 grands facteurs de segmentation de la résiliation en RC et de façon générale, on peut classer les clients en 3 grands groupes :

- Les clients les plus susceptibles de résilier et pour lesquels le modèle prédit bien la résiliation (A, C) :
  - (A) : ce sont les clients ayant une majoration globale inférieure ou égale à la majoration RC, n'ayant enregistré aucun sinistre les trois dernières années et en plus étant en décroissance de portefeuille ;
  - (C) : les clients ayant une majoration globale inférieure ou égale à la majoration RC, ayant eu au moins 1 sinistre les 3 dernières années et en décroissance de portefeuille.
- Les clients les plus fidèles et dont le modèle prédit bien la non résiliation ((B, D, G, H)) :
  - (B) : ces clients ont une majoration globale inférieure ou égale à la majoration RC, n'ont pas eu de sinistres les 3 dernières années et sont en croissance de portefeuille ;
  - (D) : ces clients ont une majoration globale inférieure ou égale à la majoration RC, ayant eu au moins 1 sinistre les 3 dernières années et sont en croissance de portefeuille.
  - (G et H) : ces clients ont une majoration globale supérieur à la majoration RC et sont en croissance de portefeuille.
- Les clients pour lesquels le modèle commet le plus l'erreur de type II (prédit à tort la non résiliation) E et F : ce sont les clients dont la majoration globale est supérieure à la majoration RC et dont le portefeuille est en décroissance. Ceci est encore plus accentué pour les clients de majoration RC est supérieure 2.5

FIGURE. 3.20 – CART sur la prédiction des résiliations.



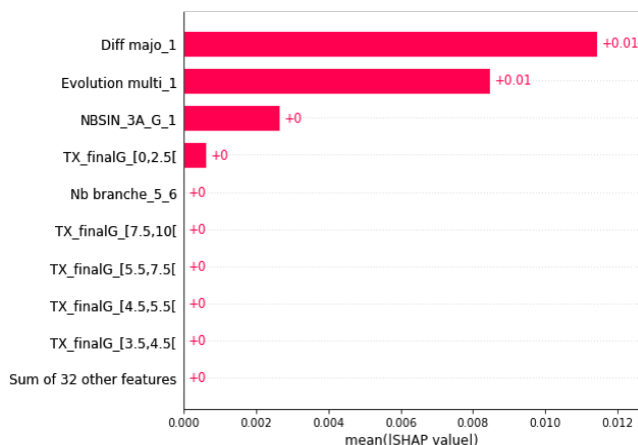
Les couleurs verte, rouge, violet et bleue renvoient aux taux de VN, FP, FN et VP

Source : Interne AXA.

### d-3) Analyse des SHAP Values.

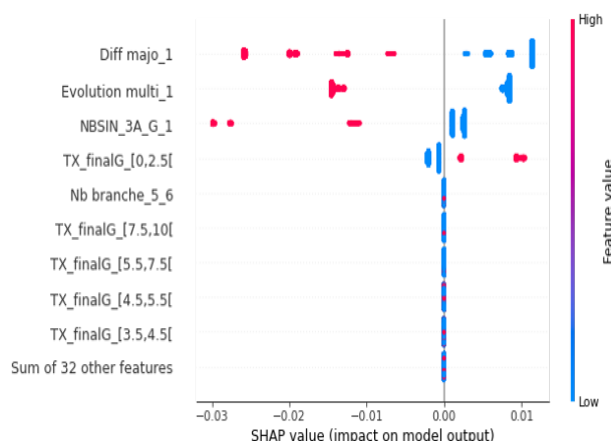
Pour aller plus loin, les prédictions ont également été caractérisées par les SHAP values sur la figure 3.21, l'importance des variables est calculée en moyennant la valeur absolue des valeurs de SHAP. On remarque que tout comme l'arbre CART, quatre variables se démarquent des autres : la différence de majoration ; l'évolution du multi-équipement ; la sinistralité des trois dernières années et le taux de majoration RC. Sur la figure 3.22, les SHAP values sont représentées pour chaque variable dans leur ordre d'importance, chaque point représente une valeur de Shap (pour un exemple), les points rouges représentent des valeurs élevées de la variable et les points bleus des valeurs basses de la variable. Il en ressort qu'une majoration RC en dessous de la majoration globale diminue la probabilité de résilié du client. Par ailleurs, la décroissance du portefeuille augmente le risque de résiliation. Ces résultats rejoignent les conclusions tirées précédemment avec la caractérisation par l'arbre de décision.

FIGURE. 3.21 – Importance globale des variables en utilisant les valeurs de Shap (moyenne valeurs absolues).



Source : Interne AXA.

FIGURE. 3.22 – Importance globale des variables en utilisant les valeurs de Shap (impact sur l'output).



Source : Interne AXA.

### 3.2.2 Présentation des modèles de résiliation implémentés part CART et Random forest.

#### a) Analyse des résultats des modèles.

Une autre approche pour la modélisation de la résiliation a été d'utiliser les arbres de décisions que nous avons par la suite sophistiqué par les méthodes ensembliste du type des Random forest. Pour l'implémentation des modèles CART et Random forest nous avons joué sur les différents hyperparamètres tels que le critère de sélection, la métrique du score, la profondeur des arbres, taille minimale de l'échantillon par feuille (min\_sample). Les meilleurs modèles obtenus sont consignés dans le tableau 3.6 ci-dessous. Notons que nous avons implémenté ces modèles sur la base de données des 182 mille contrats RC multi-équipés et ceci en utilisant les variables de notre base de modélisation consignées au tableau 2.3. Rappelons également que les échantillons de train et test sont les mêmes que nous avons utilisés pour la construction du modèle logit. Nous avons conservé ces mêmes échantillons pour pouvoir comparer avec exactitude les résultats des modèles de résiliation obtenus par les différentes approches.

L'analyse des résultats des modèles CART et Random forest (tableau 3.6) nous montre qu'avec un score de 53,2% (le score ici représente la probabilité de bien prédire la résiliation d'un client), le modèle de Random forest est moins bon en terme d'AUC que le modèle CART. Par ailleurs en comparant avec l'AUC du modèle logistique (voir tableau 3.3), on constate que le modèle logistique a la plus grande AUC. Toutefois, l'AUC n'est pas le seul indicateur pour évaluer les performances intrinsèques des modèles. Pour pouvoir faire des comparaisons plus approfondies, nous allons dans la suite analyser les performances prédictives des modèles CART et Random forest.

TABLEAU 3.6 – Résultats du modèle de résiliation : CART et Random forest.

| Modèles     | Critère de sélection | Métrique du score | Score | Autres hyperparamètres         |
|-------------|----------------------|-------------------|-------|--------------------------------|
| Modèle CART | Gini                 | ROC_AUC           | 54,9% | Profondeur = 2, min sample =50 |
| Modèle RF   | Gini                 | ROC_AUC           | 53,2% | Profondeur = 3, min sample =10 |

**Source : Interne AXA.**

## b) Analyse des performances prédictives des modèles CART et Random forest.

L'analyse des matrices de confusion 3.7 et 3.8 présentant respectivement les performances prédictives des modèles CART et Random forest sur la base test nous montre que les deux modèles ont des performances prédictives très médiocres. En effet on a une sensibilité de 0% ce qui signifie que tous les résiliés ont été prédits comme n'ayant pas résiliés. Or dans le modèle logistique présenté précédemment nous avons obtenu une sensibilité d'environ 75% (voir tableau 3.4). Ceci nous permet de conclure que le modèle logistique est plus adapté pour la prédiction de la résiliation dans notre base de données.

TABLEAU 3.7 – Matrice de confusion modèle CART.

| <b>Observations</b><br><b>Prédictions</b> | Non Résilié   | Résilié     | Total       |
|-------------------------------------------|---------------|-------------|-------------|
| Non Résilié                               | 91.18%        | 8.64%       | 99.83%      |
| Résilié                                   | 0.12%         | 0.06%       | 0.17%       |
| <b>Total</b>                              | <b>91.30%</b> | <b>8.7%</b> | <b>100%</b> |

**Source : Interne AXA.**

TABLEAU 3.8 – Matrice de confusion modèle Random Forest.

| <b>Observations</b><br><b>Prédictions</b> | Non Résilié   | Résilié     | Total       |
|-------------------------------------------|---------------|-------------|-------------|
| Non Résilié                               | 91.30%        | 8.70%       | 100%        |
| Résilié                                   | 0.00%         | 0.00%       | 0.00%       |
| <b>Total</b>                              | <b>91.30%</b> | <b>8.7%</b> | <b>100%</b> |

**Source : Interne AXA.**

En réalité, les modèles CART et Random Forest peinent à bien prédire la résiliation du fait du jeu de données déséquilibré. Notre base de données est constituée à 91% des clients n'ayant pas résiliés, or les algorithmes de machine learning standards ont la faiblesse d'être mis en difficulté face à de telle données<sup>2</sup>. Pour pallier cette limite, plusieurs solutions sont utilisées en théorie par exemple le ré-équilibrage de la base de données qui peut se faire suivant différentes approches en fonction du jeu de données de base :

2. <https://www.aquiladata.fr/insights/comment-gerer-le-desequilibre-des-classes-dans-un-jeu-de-donnees/>

- **L’undersampling** lorsque l’on dispose d’un très grand nombre d’observations (à minima  $> 10K$ ). Il s’agit ici simplement de retirer aléatoirement des instances de la classe majoritaire afin de ré-équilibrer les proportions. On perd toutefois de l’information, et il y a donc un risque d’underfitting.
- **L’oversampling** lorsque l’on dispose d’un nombre limité d’observations ( $< 10K$ ), ou bien si le temps de calcul n’est pas un problème. Il s’agit ici de dupliquer aléatoirement certaines instances des classes minoritaires, rendant ainsi leur signal plus puissant. Il y a toutefois ici un risque d’overfitting.
- **La méthode des class weights** permet de prendre en compte le caractère biaisé de la distribution du dataset et de créer un modèle pénalisé. Il s’agit ici de simplement attribuer des poids différents aux différentes classes de notre dataset, en donnant un poids plus important aux classes minoritaires, afin d’influencer le modèle lors de son entraînement. Nous pénalisons ainsi plus fortement une erreur de classification d’une classe minoritaire par rapport à une erreur de classification d’une classe majoritaire.

Étant donné que dans le cadre de ce travail, nous souhaitons garder la structure de départ de notre base de données, nous n’allons pas adopter cette solution.

### 3.3 Conclusion partielle.

Dans cette partie, nous avons répondu au premier objectif qui consistait à analyser la résiliation des clients multi-équipés. Pour y arriver, nous avons utilisé les modèles logistique, CART et Random forest pour la prédiction de la résiliation. En analysant les différents modèles, nous avons retenu le modèle logistique comme meilleur modèle ceci en raison de ses bonnes performances prédictives (voir 3.4) et de son interprétabilité. Sur la base de ce modèle nous avons déjà tiré quelques conclusions<sup>3</sup> :

- Le poids de la majoration RC dans la majoration globale d’un client multi-équipé influence sur son comportement de résiliation :
  - Si le taux de majoration de la branche RC du client est inférieur au taux de majoration global sur l’ensemble du portefeuille, alors le client a de faibles chances de résilier ;
  - Si le taux de majoration RC est supérieur à son taux de majoration global, et si de plus le client n’a pas eu de sinistralité sur les trois dernières années alors attention le client a de fortes chances de résilier.
- L’évolution du multi-équipement est un indicateur pertinent de fidélisation des clients multi-équipés :

---

3. N.B : Ces conclusions sont beaucoup plus d’ordre qualitatif au regard de la taille de notre base de données utilisée pour les analyses.

- Si un client est en croissance de portefeuille (le client a augmenté son nombre de contrats de l'année N-1) alors ce dernier a peu de chances de résilier ;
  - Si un client est en décroissance de portefeuille et de plus n'a pas eu de sinistre sur les 3 dernières années alors attention il a de fortes chances de résilier.
- Si le nombre de branches d'un client multi-équipé est supérieur à 3 alors sa probabilité de résilier est faible ;
- Si le distributeur est agent, la probabilité de résilier du client diminue ;
- Si un client a résilié au moins un de ses contrats l'année N-1, alors attention il a de fortes chances de résilier son contrat. Et ceci d'autant plus si son taux de majoration RC est supérieur à son taux de majoration global.

Après cette première partie de notre travail qui répond à notre premier objectif à savoir identifier le profil réel des clients résiliants de la branche RC, nous constatons que l'un des grands motifs de résiliation tourne autour des primes d'assurance des clients, notamment la majoration appliquée lors du renouvellement. Ainsi, quelle majoration optimale doit être proposée aux clients pour limiter la résiliation des clients tout en restant rentable ? Le chapitre suivant sera consacré à cette problématique.



## Chapitre 4

# OPTIMISATION DES POLITIQUES DE RENOUVELLEMENT DES CLIENTS RC MULTI-ÉQUIPÉS.

L'objectif principal de cette étude est de proposer aux souscripteurs des approches qui pourront permettre de réduire la résiliation lors du renouvellement. En effet, si un client est prédit comme résilié, il serait judicieux de vérifier sa rentabilité. De plus, pour les clients rentables présentant des grands risque de résiliation, faire des gestes commerciaux pour réduire le risque et ainsi fidéliser les bons clients. Le modèle de résiliation a permis de caractériser de façon spécifique les clients résiliés de notre base de données. Ayant identifié les clients les plus sujets à la résiliation, cette partie de notre travail consiste à présenter les approches pouvant être utilisé pour réduire les risques de résiliation de ces derniers.

Le modèle de résiliation a révélé que l'un des facteurs clés de résiliation des clients était le taux de majoration. Étant donné que notre objectif de base est de réduire la résiliation au sein de notre portefeuille, il est nécessaire d'analyser les majorations reçues par les clients lors du renouvellement. Pour cela, nous proposons une approche pouvant aider à mieux comprendre la majoration des clients RC lors des politiques de renouvellement. Elle consiste à trouver pour chaque client prédit comme résilié la majoration moyenne de ses "jumeaux" non résiliés et ensuite comparer ces différentes majorations. La détermination des "jumeaux" ici se fera en utilisant la distance de Levenshtein.

### 4.1 Approche méthodologique de la distance de Levenshtein.

Lorsqu'on modélise la résiliation, on se rend compte comme nous l'avons dit précédemment qu'un des facteurs clés de résiliation est le taux de majoration. Alors une première approche pour diminuer la résiliation serait de baisser le taux de majoration des potentiels résiliés.

Toutefois, des questions essentielles se posent : de combien diminuer la majoration des clients prédits comme résilié tout en restant rentable ? Tous les potentiels résiliés méritent-ils d'être fidélisés ?

Pour répondre à ces interrogations, notre approche consiste à rechercher pour les potentiels résiliés leurs "jumeaux" non résiliés puis comparer ces taux de majoration et ensuite affecter à notre potentiel résilié une majoration semblable à celle de ces "jumeaux". Cette approche à l'avantage d'être non seulement facile à implémenter, mais aussi d'assurer une rentabilité globale du portefeuille et individuelle des clients. En effet, les taux de majorations appliqués aux clients pendant le renouvellement sont calculés sous contrainte de rentabilité globale du portefeuille et individuelle des clients. Ainsi, étant donné deux clients  $X$  et  $Y$  ayant les mêmes caractéristiques (même indice, même segment, même distributeur, ...), si on affecte le taux de majoration  $t_y$  calculé pour le client  $Y$  au client  $X$  les deux clients restent rentables.

Dans cette étude, le terme "jumeaux" renvoie aux clients ayant les caractéristiques à peu près similaires. Pour identifier les paires de jumeaux, la distance de Levenshtein sera utilisée.

#### 4.1.1 Définition de la Distance de Levenshtein.

La distance de Levenshtein est une distance, au sens mathématique du terme, donnant une mesure de la différence entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre. La distance de Levenshtein entre deux chaînes  $a$  et  $b$  est donc le coût minimal pour transformer  $a$  en  $b$  en effectuant les seules opérations élémentaires (au niveau d'un caractère) suivantes (on agit exclusivement sur  $a$  et successivement sur ses transformés) : **substitution**, **insertion (ou ajout)**, **suppression (ou effacement)**. Formellement, on définit cette distance entre deux chaînes  $a$  et  $b$  par :

$$\text{lev}(a, b) = \begin{cases} \max(|a|, |b|) & \text{si } \min(|a|, |b|) = 0, \\ \text{lev}(a-1, b-1) & \text{si } a[0] = b[0], \\ 1 + \min \begin{cases} \text{lev}(a-1, b) \\ \text{lev}(a, b-1) \\ \text{lev}(a-1, b-1) \end{cases} & \text{sinon.} \end{cases} \quad (4.1)$$

Avec  $|a|$  le cardinal de  $a$  (ou son nombre de lettres), et  $a-1$  la chaîne  $a$  tronquée de sa 1ère lettre  $a[0]$ .

A titre d'illustration, si  $A = \text{« résiliation »}$  et  $B = \text{« résiliation »}$ , alors  $\text{lev}(A, B) = 0$  parce qu'aucune opération n'a été réalisée. Si  $A = \text{« résiliation »}$  et  $B = \text{« résilyation »}$ , alors  $\text{lev}(A, B) = 1$ , parce qu'il y a eu une substitution (changement du  $i$  en  $y$ ), et que l'on ne peut pas en faire moins.

### 4.1.2 Algorithme de recherche des jumeaux des potentiels résiliers .

L'algorithme à suivre pour rechercher les jumeaux consiste à créer une variable concaténant les caractéristiques de chaque client, puis calculer la distance de Levenshtein entre les clients en utilisant cette nouvelle variable et enfin, déduire les jumeaux qui correspondent aux clients ayant la plus faible distance. Plus spécifiquement elle consiste à :

- **Créer l'indicateur de ressemblance** : dans la base de données, on crée une variable pouvant être utilisée comme indicateur de ressemblance entre les clients. Ce dernier pour nous est simplement la concaténation de tous les caractéristiques de chaque client. Par exemple prenons deux clients ( $c1$  et  $c2$ ) tels que,  $c1$  : Évolution du multi-équipement= $-1$ , distributeur=agent, âge contrat = 1 ; et  $c2$  : Évolution du multi-équipement= $1$ , distributeur=courtier, âge contrat = 3. Alors pour ces deux clients l'indicateur de ressemblance noté  $id\_ressemblance$  est  $-1agent1$  et  $1courtier3$  pour  $c1$  et  $c2$  respectivement.
- **Ré-codage des modalités** : on pourrait évaluer les distances entre les clients  $c1$  et  $c2$  en faisant  $distance(c1, c2) = lev(-1agent1, 1courtier3)$ . Mais on remarque que les chaînes de caractères n'ont pas la même taille ( $|-1agent1| = 7$  et  $|1courtier3| = 10$  ce qui pourrait conduire à une source d'erreurs dans le calcul des distances. Alors pour pallier à cela, une solution consiste à ré-coder les modalités de telle sorte que chacune tienne sur un seul caractère. Exemple : Pour l'âge du contrat :  $1 = 1, 2 = 2, 3 = 3$ . Pour l'évolution du multi-équipement :  $-1 = 4, 0 = 5, 1 = 6$ . Pour le distributeur  $agent = a, courtier = c$ . Remarquons que pour ré-coder l'évolution du multi-équipement, on a été obligé de commencer par 4 car l'âge du contrat occupe déjà les valeurs 1, 2, 3. Ainsi on aura  $c1 : id\_ressemblance = 4a1$  ;  $c2 : id\_ressemblance = 6c3$ .
- **Concaténation des variables** : le nouvel indicateur de ressemblance est calculé en concaténant les variables ré-codées en fonction de leur importance dans le modèle de résiliation (la variable la plus importante occupe la première position dans la chaîne et ainsi de suite).
- **Segmentation de la base** : on divise la base de données en deux pour avoir une base des résiliés et une base des non résiliés. Ensuite on retire de ces deux bases les clients sinistrés. En effet suite à une sinistralité, la majoration appliquée au client lors du renouvellement peut être particulière. Pour éviter les aléas liés aux antécédents de sinistralité, on se restreint sur les clients non sinistrés.
- **Calcul de la distance de Levenshtein** : pour chaque client de la base des résiliés on évalue la distance entre lui et tous les clients non résiliés. Pour tenir compte de l'importance de chaque variable dans la distance, on pondère par la position de cette dernière

dans la chaîne de caractères. Ainsi la nouvelle forme de la distance de Levenshtein est :

$$\text{lev}(a, b) = \begin{cases} \max(|a|, |b|) & \text{si } \min(|a|, |b|) = 0, \\ \frac{|a| - |a-1|}{|a|} + \text{lev}(a-1, b-1) & \text{si } a[0] = b[0], \\ 1 + \frac{|a| - |a-1|}{|a|} + \min \begin{cases} \text{lev}(a-1, b) \\ \text{lev}(a, b-1) \\ \text{lev}(a-1, b-1) \end{cases} & \text{sinon.} \end{cases} \quad (4.2)$$

- **Identification des jumeaux** : après le calcul des distances, les jumeaux du client résilié correspondent à ceux ayant la distance de Levenshtein minimal.

#### 4.1.3 Principe d'évaluation de la majoration optimale entre les potentiels résiliers et leurs jumeaux.

Après avoir trouvé les jumeaux des potentiels résiliers, la deuxième étape consiste à déterminer la majoration optimale à les proposer pour réduire le risque de résiliation. L'approche est donc la suivante :

- **Calculer la majoration moyenne des jumeaux** : pour chaque client prédit comme potentiel résilieur, on calcule la moyenne des majorations obtenue par ses jumeaux.
- **Nuage de points de la majoration des potentiels résiliers vs majoration des jumeaux non résiliés et évaluation de la position du résilié par rapport à celle de ses jumeaux** : on construit le nuage de points portant en abscisse la majoration des potentiels résiliers et en ordonnée celle de leurs jumeaux. Puis on étudie le comportement du nuage, plus précisément la position des clients par rapport à la première bissectrice. Les clients situés en dessous de la première bissectrice sont ceux qui ont eu des majorations supérieures à celles de leurs jumeaux.
- **Conclure** : ayant donc pour l'ensemble des potentiels résiliers ceux qui ont été sur-majorés comparativement à leurs jumeaux, proposer aux potentiels résiliers des majorations semblables à celle de leurs jumeaux non résiliés.

## 4.2 Distance de Levenshtein : de la recherche des jumeaux à l'optimisation de la majoration.

Le modèle de résiliation (modèle logistique) présenté en section 3.2.1 a permis d'effectuer des prédictions sur la base test constituée de 18 mille clients multi-équipés. En utilisant ces prédictions de résiliation et la résiliation observée au sein de notre portefeuille, nous avons construit quatre groupes de clients : les vrais positifs (VP : on les prédit comme résilié et ils

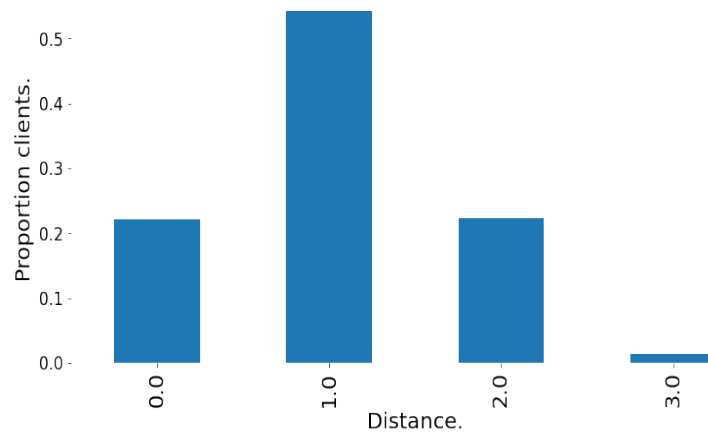
ont effectivement résilié), les vrais négatifs (VN : on les prédit comme non résilié et ils n'ont effectivement pas résilié), les faux négatifs (FN : on les prédit comme résilié alors qu'ils n'ont pas résilié), et enfin les faux positifs (FP : on les prédit comme non résilié alors qu'ils ont résilié). Dans le cadre de notre étude, nous nous concentrons uniquement sur le groupe des VP et des VN, l'idée étant comme présentée en section 4.1 de rechercher les jumeaux des VP dans la base des VN et ensuite optimiser leurs majorations. Notons que nous nous sommes aussi restreints sur la base des clients non sinistrés et pour la recherche des jumeaux, nous sommes partis d'une approche simple que nous avons enrichi pas à pas. Dans la suite, nous présentons les différentes approches utilisées ; rappelons également que la base de données est la même dans chaque cas.

#### 4.2.1 Première approche : utilisation de la distance de Levenshtein non pondérée.

La première approche consiste à concaténer les caractéristiques des variables cryptées en fonction de leur importance dans le modèle de résiliation (exceptés le taux de majoration RC et la sinistralité) et calculer la distance entre les clients en utilisant la formule de la distance de Levenshtein non pondérée (formule 4.1).

La concaténation des variables s'est faite suivant cet ordre : *Ancienneté du contrat + Évolution du multi-équipement + Grille majoration + Différence majoration + Nombre de branches + Indice + Région + Distributeur + Année*. En suivant la suite de l'algorithme de recherche des jumeaux présentés en section 4.1 (mais avec la formule de Levenshtein non pondérée), nous avons identifié les jumeaux des clients résiliés (base VP). La figure 4.1 ci-dessous nous donne la distance moyenne séparant chaque client résilié de ses jumeaux non résiliés. On constate que les jumeaux sont à plus de 50% à une distance inférieure ou égale à 1 donc on a plutôt des «vrais jumeaux». On rappelle que si la distance = 0 alors on a exactement les mêmes caractéristiques chez le client résilié et ses jumeaux ; distance = 1 une seule caractéristique diffère entre le client résilié de ses jumeaux ;...

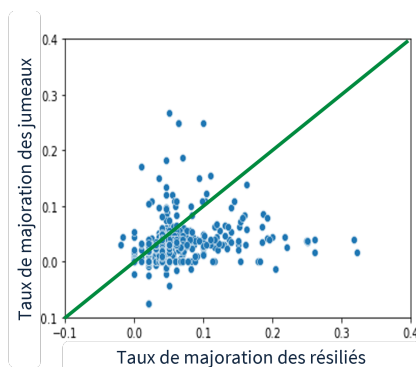
FIGURE. 4.1 – 1ère Approche : Distance de Levenshtein entre les résiliés et leurs jumeaux non résiliés.



Source : Interne AXA.

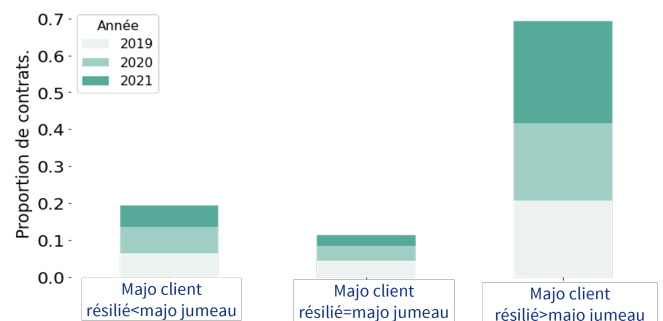
Pour comparer la majoration RC des clients résiliés par rapport à celle de leurs jumeaux non résiliés, nous présentons premièrement dans la figure 4.2 ci-dessous le nuage de points entre la majoration RC des clients résiliés et celle de leurs jumeaux non résiliés. En plus du nuage de points, la figure 4.3 donne la position des clients par rapport à celle de leurs jumeaux non résiliés (position par rapport à la première bissectrice). On constate que le nuage de point est beaucoup plus concentré en dessous de la première bissectrice. Ce qui signifie qu'à caractéristiques «similaires», les résiliés ont dans l'ensemble reçu une forte majoration. En analysant de façon détaillée avec la figure 4.3, on observe qu'environ 70% des résiliés ont reçu des majorations supérieures aux majorations reçues par leurs jumeaux et qu'en 2021 cette proportion était encore plus forte et avoisinait les 80%.

FIGURE. 4.2 – 1ère Approche : Nuage de points entre majoration des résiliés et jumeaux non résiliés.



Source : Interne AXA.

FIGURE. 4.3 – 1ère Approche : Comparaison majoration des résiliés et jumeaux non résiliés.



Source : Interne AXA.

Ce premier résultat nous montre qu'à caractéristique similaire aux clients non résiliés, plus 75% des clients de la base des résiliés ont reçu des majorations plus élevées. Toutefois rappelons que la distance de Levenshtein a été calculée ici en utilisant la formule non pondérée donc ce

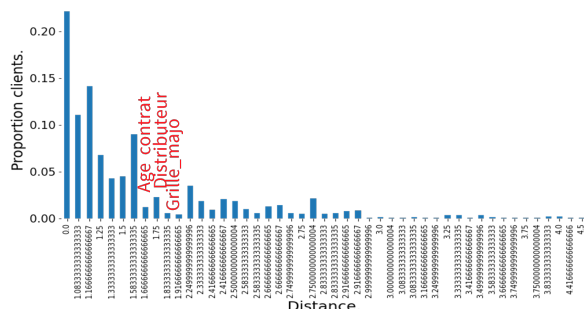
résultat nécessite des améliorations. En effet, supposons que les clients sont juste caractérisés par «distributeur», «différence de majoration» et «Indice». Soient les clients suivant : A (distributeur=agent ; différence de majoration=1 ; indice=5) ; B (distributeur=agent ; différence de majoration=1 ; indice=4) ; C (distributeur= agent ; différence de majoration=0 ; indice=5), alors en utilisant la formule non pondérée de la distance de Levenshtein,  $Lev(A,B)=1$  et  $Lev(A,C)=1$ . Or dans notre modèle de résiliation, «différence de majoration» explique mieux la résiliation que «indice». Donc on aimerait que B soit plus proche de A que C ie  $Lev(A,B) < Lev(A,C)$ . Pour corriger cette erreur, on va utiliser la formule pondérée de la distance de Levenshtein.

#### 4.2.2 Deuxième approche : prise en compte des corrélations entre les variables et le taux de majoration.

Dans la première approche, nous avons concaténé les variables en fonction de leurs importances dans le modèle de résiliation. Or sur le plan métier, certaines variables ont une très forte influence sur la majoration par rapport à d'autres et nous l'avons par ailleurs montré en faisant l'analyse croisée de la majoration et des variables. En effet, les courtiers par exemple appliquent des majorations beaucoup plus faibles que les agents lors du renouvellement, ce qui signifie qu'à caractéristiques égales, un client qui a pour distributeur un agent sera plus majoré que s'il était chez un courtier. Pour donc atténuer les biais liés aux trois variables les plus déterminantes dans la majoration client à savoir l'ancienneté, le distributeur et la grille de majoration, nous allons les placer en dernière position dans la construction de l'indicateur de ressemblance et ensuite calculer la distance en utilisant la formule pondérée de la distance de Levenshtein.

L'indicateur de ressemblance pour cette approche vaut donc : *Évolution multi-équipement + Différence de majoration + Nombre de branches + Indice + Région + Ancienneté contrat + Distributeur + Grille majoration + Année*. En suivant la suite de l'algorithme de recherche des jumeaux présenté en section 4.1, nous avons identifié les jumeaux des clients résiliés (base VP). La figure 4.4 nous donne la distance moyenne séparant chaque client résilié de ses jumeaux non résiliés. On constate que les jumeaux sont en moyenne à une distance de 1 donc on a plutôt des «vrais jumeaux». On rappelle que si la distance = 0 alors on a exactement les mêmes caractéristiques chez le client résilié et ses jumeaux ; distance  $\in [1, 2[$  une seule caractéristique diffère entre le client résilié de ses jumeaux. En évaluant la majoration moyenne chez les résiliés et leurs jumeaux non résiliés, nous avons constaté que le taux de majoration moyen des résiliés était d'environ 4% pour tous niveaux de distances et celui de leurs jumeaux de 3% (figure 4.5). Donc à caractéristiques similaires, les clients résiliés avaient reçu plus de majoration que leurs jumeaux non résiliés.

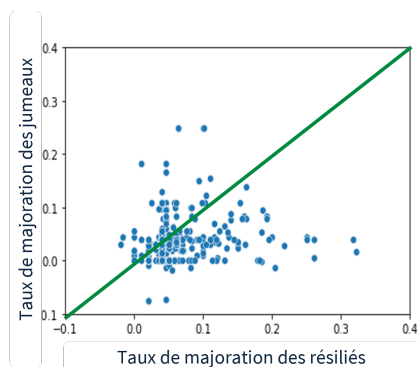
FIGURE. 4.4 – 2ème Approche : Distance de Levenshtein entre les résiliés et leurs jumeaux non résiliés.



Source : Interne AXA.

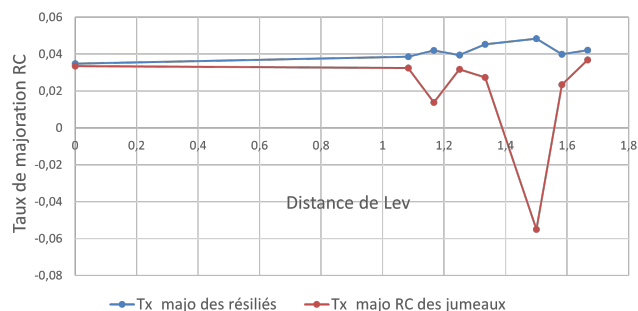
En analysant les positions des résiliés par rapport à leurs jumeaux non résiliés (voir figure 4.6 et 4.7), On constate que près de 60% des résiliés ont reçu des majorations supérieures aux majorations reçues par leurs «jumeaux» et en 2021, cette proportion s’est encore plus accentuée passant à environ 75%. Ceci signifie qu’à caractéristiques «similaires», les résiliés ont dans l’ensemble reçu une forte majoration.

FIGURE. 4.6 – 2ème Approche : Nuage de points entre majoration des résiliés et jumeaux non résiliés.



Source : Interne AXA.

FIGURE. 4.5 – Taux de majoration des résiliés vs jumeaux non résiliés.



Source : Interne AXA.

FIGURE. 4.7 – 2ème Approche : Comparaison majoration des résiliés et jumeaux non résiliés.

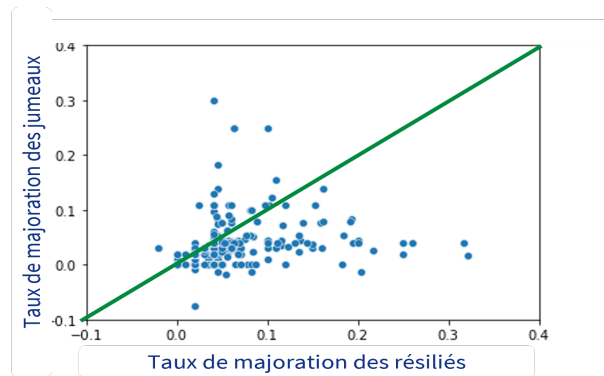


Source : Interne AXA.

Pour éliminer le biais lié à la corrélation des variables (l’ancienneté, le distributeur et la grille de majoration) avec la majoration, on va faire un zoom sur les clients dont la distance de Levenshtein avec leur jumeau a été inférieure à 1.67. Ce sont en effet les clients résiliés ayant au maximum un caractère qui les différencie de leurs jumeaux non résiliés, et en plus le caractère qui différencie n’est ni l’ancienneté, ni le distributeur, ni la grille de majoration. En analysant les figures 4.8 et 4.9, on constate que près de 50% des résiliés ont reçu des majorations supérieures aux majorations reçues par leurs «jumeaux» et qu’en 2021 cette proportion était d’environ 60%.

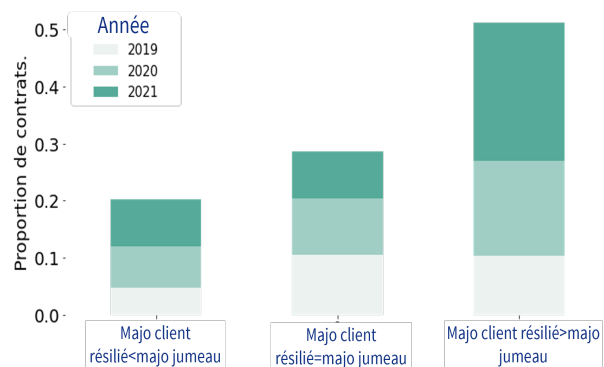


FIGURE. 4.8 – 2ème Approche : Nuage de points entre majoration des résiliés et jumeaux non résiliés (Zoom sur les clients de distance  $<1.67$  ie le caractère qui différencie n'est pas l'âge du contrat, le distributeur ou la grille majo).



Source : Interne AXA.

FIGURE. 4.9 – 2ème Approche : Comparaison majoration des résiliés et jumeaux non résiliés (Zoom sur les clients de distance  $<1.67$  ie le caractère qui différencie n'est pas l'âge du contrat, le distributeur ou la grille majo).



Source : Interne AXA.

Grâce à cette approche, nous constatons qu'à caractéristiques similaires, environ 50% des résiliés ont reçu des majorations supérieures à celles de leurs jumeaux non résiliés. Pour ces clients, on pourrait penser que s'ils avaient reçu des majorations similaires à la majoration moyenne de leurs jumeaux non résiliés, leurs risques de résiliation auraient diminué.

## 4.3 Interprétation des résultats et importance opérationnelle de l'approche de Levenshtein.

### 4.3.1 Interprétation des résultats

Nous avons précédemment fait des analyses comparatives sur les majorations reçues par les clients résiliés et leurs jumeaux n'ayant pas résilié. Cette analyse a montré que les clients qui résilient ont dans plus de 50% des cas reçu des fortes majorations comparativement à

leurs jumeaux non résiliés. Ce résultat vient confirmer les conclusions tirées du modèle de résiliation : **le taux de majoration RC est le principal facteur de résiliation des clients multi-équipés de la branche RC.**

Dans cette partie, nous aimerions savoir quelles sont les caractéristiques des clients résiliés ayant reçu des majorations différentes de celles de leurs jumeaux. À cet effet, la variable "position\_jumeau" définit comme suit a été créée :

- "position\_jumeau" = 1 : le client résilié a reçu une majoration supérieure à la majoration moyenne de ses jumeaux non résiliés ;
- "position\_jumeau" = 0 : le client résilié a reçu une majoration égale à la majoration moyenne de ses jumeaux non résiliés ;
- "position\_jumeau" = -1 : le client résilié a reçu une majoration inférieure à la majoration moyenne de ses jumeaux non résiliés.

De façon générale en analysant le tableau 4.1, on constate qu'avec une majoration moyenne de **300€** les résiliés ont des majorations plus élevées que leurs jumeaux non résiliés. Par ailleurs, en observant les primes d'assurance, on constate qu'avec un taux de majoration de **10.56%** les résiliés dont la majoration est supérieure à celle des jumeaux ont en moyenne des primes plus faibles que leurs jumeaux non résiliés. On peut donc conclure que ces derniers ont résilié du fait des taux de majoration très élevés. Par contre, les résiliés ayant des majorations plus faible que leurs jumeaux (taux de majoration de **9.53%**) ont des primes d'assurance plus fortes. Ils ont alors probablement résilié du fait des primes RC très élevées.

TABLEAU 4.1 – Comparaison résiliés vs jumeaux non résiliés (Taux de majoration | primes en € | montant de majoration en €).

| Position jumeaux | Majo client résilié < majo jumeau | Majo client résilié = majo jumeau | Majo client résilié > majo jumeau | Moyenne            |
|------------------|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|
| Résiliés         | 9,53%   2659€   253€              | 8,88%   2564€   227€              | 10,56%   3337€   352€             | 10%   3000€   300€ |
| Jumeaux          | 14,12%   1763€   249€             | 9,47%   2341€   221€              | 5,48%   4664€   255€              | 7%   3500€   245€  |

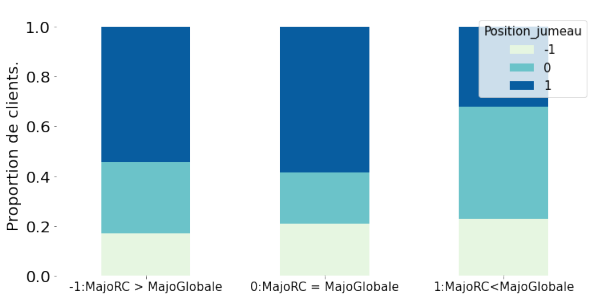
Ayant remarqué que les niveaux de primes et les montants de majoration en euros variaient significativement selon la position du jumeau, nous allons faire des caractérisations suivant les variables d'étude afin de voir si des conclusions similaires sont observées.

#### a) Différence de majoration.

Tout d'abord, en analysant la position du jumeau suivant les différents niveaux de majoration (figures 4.10 et 4.11), on constate que majoritairement, ce sont les clients de majoration supérieure ou égale à la majoration globale qui ont des fortes majorations comparativement à leurs jumeaux non résiliés. De plus, les clients de majoration RC inférieure à la majoration globale résilient pour un motif autre que la majoration. Par ce résultat, on comprend que les po-

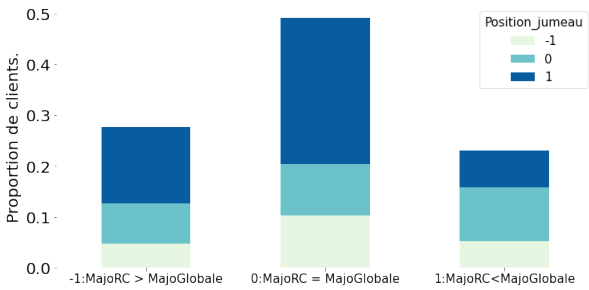
litiques de fidélisation des clients doivent se faire en prenant en compte le niveau de majoration global du portefeuille.

FIGURE. 4.10 – Diff-majoration du client résilié vs majoration des jumeaux (base 100).



Source : Interne AXA.

FIGURE. 4.11 – Diff-majoration du client résilié vs majoration des jumeaux.



Source : Interne AXA.

TABLEAU 4.2 – Différence de majoration : majoration des clients résiliés (Taux de majoration | primes en € | montant de majoration en €).

| Position jumeau | MajoRC > MajoGlobale  | MajoRC = MajoGlobale  | MajoRC < MajoGlobale | Moyenne               |
|-----------------|-----------------------|-----------------------|----------------------|-----------------------|
| -1              | 10,79%   2798€   302€ | 10,08%   2709€   273€ | 6,24%   2362€   147€ | 9,53%   2659€   253€  |
| 0               | 9,65%   2871€   277€  | 9,42%   3015€   284€  | 5,17%   1497€   77€  | 8,88%   2564€   227€  |
| 1               | 11,38%   4065€   462€ | 11,29%   2473€   279€ | 5,51%   3838€   211€ | 10,56%   3337€   352€ |
| Moyenne         | 10,93%   3547€   387€ | 10,57%   2644€   279€ | 5,59%   2527€   141€ | 10%   3000€   300€    |

TABLEAU 4.3 – Différence de majoration : majoration des jumeaux non résiliés (Taux de majoration | primes en € | montant de majoration en €).

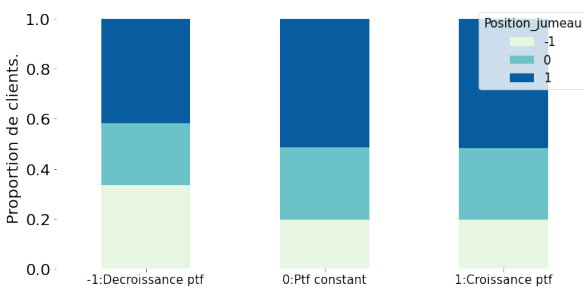
| Position jumeau | MajoRC > MajoGlobale  | MajoRC = MajoGlobale  | MajoRC < MajoGlobale   | Moyenne               |
|-----------------|-----------------------|-----------------------|------------------------|-----------------------|
| -1              | 21,3%   1524€   324€  | 11,33%   2321€   263€ | 8,71%   1253€   109€   | 14,12%   1763€   249€ |
| 0               | 10,49%   2436€   255€ | 10,91%   2463€   268€ | 5,28%   2036€   107€   | 9,47%   2341€   221€  |
| 1               | 6,12%   5157€   315€  | 6,32%   4749€   300€  | -5,66%   2467€   -139€ | 5,48%   4664€   255€  |
| Moyenne         | 7,8%   3860€   301€   | 7,55%   3791€   286€  | 1,03%   2004€   20€    | 7%   3500€   245€     |

L'analyse en fonction de la différence de majoration des résiliés (tableau 4.2) et des jumeaux non résiliés (tableau 4.3) montre que les écarts de majorations entre les résiliés et leurs jumeaux sont beaucoup plus perceptibles chez les clients de position jumeau = 1 (majoration client résilié > majoration moyenne de ses « jumeaux »). Par ailleurs, les résiliés de majoration RC inférieure majoration jumeaux ont généralement des primes plus hautes que leurs jumeaux. De plus les résiliés de majoration RC supérieure à la majoration globale ont dans l'ensemble de montants de majoration en euros qui dépassent largement la majoration moyenne des résiliés (387€ contre 300€ en moyenne).

## b) Évolution multi-équipement.

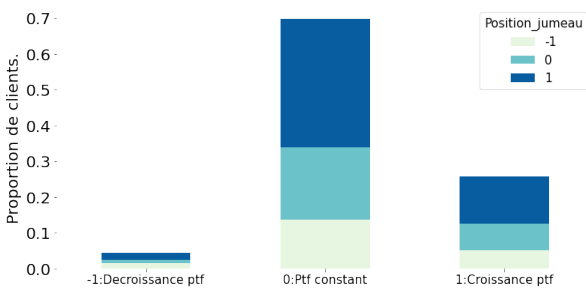
Pour l'évolution du multi-équipement, ce sont majoritairement les clients en croissance de portefeuille qui ont des fortes majorations comparativement à leurs jumeaux non résiliés (figures 4.12 et 4.13). Le modèle de résiliation (section 3.2.1) a montré que l'évolution du multi-équipement conduit à une baisse de la résiliation. Mais avec les résultats de la distance de Levenshtein, on constate que si les clients en croissance de portefeuille ont des fortes majorations, on annule le gain lié à l'évolution du multi-équipement.

FIGURE. 4.12 – Évolution multi-équipement du client résilié vs majoration des jumeaux (base 100).



Source : Interne AXA.

FIGURE. 4.13 – Évolution multi-équipement du client résilié vs majoration des jumeaux.



Source : Interne AXA.

TABLEAU 4.4 – Évolution multi-équipement : majoration des clients résiliés (Taux de majoration | primes en € | montant de majoration en €).

| Position jumeau | Décroissance portefeuille | Portefeuille constant | Croissance portefeuille | Moyenne               |
|-----------------|---------------------------|-----------------------|-------------------------|-----------------------|
| -1              | 11,02%   708€   78€       | 9,18%   2726€   250€  | 10,64%   2728€   290€   | 9,53%   2659€   253€  |
| 0               | 9,48%   4225€   400€      | 8,23%   2365€   194€  | 10,16%   2633€   267€   | 8,88%   2564€   227€  |
| 1               | 11,6%   4291€   498€      | 10,01%   3283€   328€ | 12,08%   3088€   373€   | 10,56%   3337€   352€ |
| Moyenne         | 11%   3993€   439€        | 9,48%   2934€   278€  | 11,24%   2869€   322€   | 10%   3000€   300€    |

TABLEAU 4.5 – Évolution multi-équipement : majoration des jumeaux non résiliés (Taux de majoration | primes en € | montant de majoration en €).

| Position jumeau | Décroissance portefeuille | Portefeuille constant | Croissance portefeuille | Moyenne               |
|-----------------|---------------------------|-----------------------|-------------------------|-----------------------|
| -1              | 8,48%   745€   63€        | 15,18%   1599€   242€ | 12,11%   2444€   296€   | 14,12%   1763€   249€ |
| 0               | 11,54%   2420€   279€     | 9,05%   2246€   203€  | 9,9%   2577€   255€     | 9,47%   2341€   221€  |
| 1               | 6,27%   3797€   238€      | 5,36%   4528€   242€  | 5,59%   5591€   312€    | 5,48%   4664€   255€  |
| Moy             | 7,37%   3193€   235€      | 6,86%   3381€   232€  | 7,26%   4001€   290€    | 7%   3500€   245€     |

L'analyse de la position\_jumeau suivant l'évolution du multi-équipement des résiliés (tableau 4.2) et des jumeaux non résiliés (tableau 4.3) montre que les écarts de majorations

entre les résiliés et leurs jumeaux sont beaucoup plus perceptibles chez les clients de position jumeau = 1 (majoration client résilié > majoration moyenne de ses « jumeaux »). Par ailleurs, les résiliés de majoration RC inférieure majoration jumeaux ont généralement des primes plus hautes que leurs jumeaux. De plus les résiliés en croissance de portefeuille ont dans l'ensemble de montants de majoration en euros qui dépassent la majoration moyenne des résiliés (322€ contre 300€ en moyenne). Notons toutefois qu'au regard du faible effectif des clients en décroissance de portefeuille les résultats observés dans ces tableau sont pas significatifs pour tirer des conclusions.

#### 4.3.2 Importance opérationnelle de l'approche de Levenshtein.

L'approche de Levenshtein pour l'analyse des majorations a conduit à des conclusions d'une part sur la position de la majoration RC par rapport à la majoration globale du client multi-équipé et d'autre part sur l'évolution du multi-équipement. Nous rappelons ici ces grands résultats et les conclusions opérationnelles qui en découlent.

- Pour les clients multi-équipés, lorsque la majoration RC est au-dessus de la majoration globale, les risques de résiliation sont très élevés. Au vu de ce résultat, nous suggérons de proposer aux clients multi-équipés une majoration flat pour tous ses contrats. En effet, lorsqu'un client multi-équipé s'aperçoit qu'il est très majoré sur une branche particulière, il est fort probable qu'il résilie son contrat sur cette branche. Cette première résiliation va conduire à une baisse de son portefeuille (nombre de contrats année N inférieure au nombre de contrats année N-1). De plus, comme il a été montré dans le modèle de résiliation, la décroissance du portefeuille augmente de 40% les risques de résiliation. Ces risques cumulés de résiliation sont donc susceptibles d'entraîner la résiliation de tous les contrats du client initialement multi-équipé, ce qui pourrait être une perte pour AXA ceci d'autant plus si le client avait un portefeuille rentable au global. Une majoration unique pour tous les contrats du client sera donc un bon paramètre de fidélisation en ce sens qu'il évitera les risques de résiliation pour la branche la plus majorée.
- Une deuxième conclusion tirée de l'analyse des majorations des résiliés par la distance de Levenshtein concerne l'évolution du multi-équipement. Les clients en croissance de portefeuille ne doivent pas être considérés comme des "acquis" pour AXA, en effet dans le modèle de résiliation nous avons conclu que si un client multi-équipé accroît son portefeuille, il a peu de chances de résilier. Mais avec l'analyse de la distance de Levenshtein on remarque que si ces mêmes clients reçoivent des majorations supérieures à la majoration moyenne, leurs probabilités de résilier augmente considérablement. En d'autres termes, les souscripteurs ne doivent pas majorer fortement les clients en croissance de portefeuille. Les majorations doivent être standardisées pour favoriser davantage la fidélisation des clients multi-équipés.
- La troisième conclusion à tirer de cette analyse est que les résiliés ont deux grands facteurs

qui les différencie de leurs jumeaux non résiliés (la prime et le montant de la majoration en euro). En général, les résiliés qui ont reçu des fortes majorations comparativement à leurs jumeaux non résiliés ont également eu des montants de majorations en euros très élevés (par rapport à leurs jumeaux et à la moyenne des majorations des résiliés). Dans le même sens, les clients qui résilient malgré qu'ils aient eu des taux de majoration plus faibles que leurs jumeaux ont des primes d'assurances très élevées. Ce résultat nous montre que les souscripteurs doivent prêter attention au niveau de primes des clients lors du renouvellement. Une majoration de 5% pour un client ayant une prime de 10 000 euros (soit 500 euros de majoration) est très différente d'une majoration de 5% pour un client ayant une prime de 100 euros (soit 5 euros de majoration). Étant donné que la politique de renouvellement actuelle d'AXA propose des majorations suivant les niveaux de primes, les souscripteurs doivent davantage prêter attention aux clients ayant des grosses primes d'assurance car ces derniers sont beaucoup plus sensibles à une augmentation de leurs primes que les autres clients du portefeuille.

# CONCLUSION.

Dans le cadre de ce travail, il était question d'étudier le comportement de résiliation des clients multi-équipés de la branche responsabilité civile, ceci dans le but de proposer des approches pouvant permettre d'améliorer la politique de renouvellement. Pour répondre à cette problématique, la modélisation de la résiliation a été faite en utilisant les modèles logistique, CART et Random Forest. Après l'analyse et la comparaison des différents modèles, le modèle logistique a été retenu comme meilleur modèle, ceci en raison de ses bonnes performances prédictives (sensibilité de 75%) et de son interprétabilité. Ce modèle a révélé que la majoration client est l'un des principaux facteurs de résiliation, et le pic de résiliation chez les clients multi-équipés s'observe à un taux de majoration de 4.5%. Ajouté à cela, le poids de la majoration RC dans la majoration globale d'un client multi-équipé et l'évolution du multi-équipement ont été identifiés comme des grands facteurs de résiliation.

Partant de ces constats, la deuxième partie du travail s'est consacrée à la compréhension de la majoration des clients multi-équipés sujets à la résiliation. Cette problématique a été abordée en utilisant la distance de Levenshtein pour retrouver les jumeaux de chaque potentiel résilieur. L'algorithme utilisé pour rechercher les jumeaux consistait à créer une variable concaténant les caractéristiques de chaque client, puis à calculer la distance de Levenshtein entre les clients en utilisant cette nouvelle variable et enfin, à déduire les jumeaux qui correspondaient aux clients ayant la plus faible distance. L'implémentation de cet algorithme sur une base de données de 183 000 clients multi-équipés montre qu'environ 70% des clients résiliés ont des distances de Levenshtein inférieures à 2 de leurs jumeaux. Ceci indique qu'on pouvait trouver la majoration optimale des résiliés à partir de celle de leurs jumeaux non résiliés. En évaluant la majoration moyenne, on obtient un taux de majoration moyen des résiliés d'environ 4% et celui de leurs jumeaux de 3%. Ainsi, à caractéristiques similaires, les clients résiliés ont fait l'objet de plus de majoration que leurs jumeaux non résiliés.

Une caractérisation des jumeaux résiliés a ensuite permis de confirmer les résultats obtenus dans le modèle de résiliation et d'identifier la prime et le montant de la majoration en euros comme d'autres facteurs de résiliation. Sur ces bases, les alertes ont été transmises aux sous-

cripteurs dans le but de fidéliser davantage la clientèle multi-équipée qui fait partir des clients les plus rentables. De plus, deux grandes recommandations ont été proposées à AXA :

- Proposer à un client multi-équipé une majoration flat pour tous ses contrats. En effet, lorsqu'un client multi-équipé s'aperçoit qu'il est très majoré sur une branche particulière, il est fort probable qu'il résilie son contrat sur cette branche. Cette première résiliation va conduire à une baisse de son portefeuille (nombre de contrats année N inférieure au nombre de contrats année N-1). De plus, comme il a été montré dans le modèle de résiliation, la décroissance du portefeuille augmente de 40% les risques de résiliation. Ces risques cumulés de résiliation sont susceptibles d'entraîner la résiliation de tous les contrats du client initialement multi-équipé, ce qui pourrait être une perte pour AXA ceci d'autant plus si le client avait un portefeuille rentable au global. Une majoration unique pour tous les contrats du client semblerait donc être un bon paramètre de fidélisation en ce sens qu'il évitera les risques de résiliation pour la branche la plus majorée.
- Sachant l'importance de la croissance de portefeuille mais aussi le fait d'éviter la décroissance, il serait intéressant d'anticiper sur les demandes des clients en leurs proposant de nouveaux contrats, par exemple un client du BTP (Bâtiment des Travaux Publics) avec une flotte automobile, pourrait se voir proposer un contrat d'assurance construction ou d'assurance responsabilité civile. Et en cas de résiliation d'un contrat une année, le souscripteur pourrait lui reproposer ce contrat l'année suivante.

Ce travail, bien que pertinent pour la compréhension de la résiliation des clients multi-équipés, a employé des techniques beaucoup plus descriptives pour la résolution de la problématique. Le choix de ces techniques d'analyse a été orienté par la spécificité des clients du portefeuille (trop hétérogènes) et le besoin d'avoir des résultats très opérationnels pour AXA. De plus, ce travail s'est limité à l'analyse de la résiliation de la branche responsabilité civile des clients multi-équipés. Pour l'approfondir davantage, nous avons suggéré d'étudier la résiliation des autres branches des clients multi-équipés afin de tirer des conclusions plus générales sur la position de leurs majorations par rapport à la majoration globale, sur l'évolution du multi-équipement et sur l'analyse de la rentabilité du client.



# Abréviations.

BTP : Bâtiments des Travaux Publics.

CART : Classification and Regression Tree.

DP : Défense de Portefeuille.

FN : Faux Négatif.

FP : Faux Positif.

IARD : Incendies, Accidents et Risques Divers.

OVC : Outil Vision Client.

PB : Presse Bouton.

QP : Qualité de portefeuille.

RC : Responsabilité civile.

RI : Risque Industriel.

RT : Risque Technique.

TT : Transport Terrestre.

VN : Vrai Négatif.

VP : Vrai Positif.

# Bibliographie

- [1] Reinartz, W. J., Kumar, V. (2003), *The impact of customer relationship characteristics on profitable lifetime duration*, Journal of marketing, 67(1), 77-99.
- [2] Ngobo, P. V. (2004), *Drivers of customers' cross-buying intentions.*, European Journal of Marketing, 38(9/10), 1129-1157.
- [3] CHARPENTIER A. (2013), *Cours Actuariat IARD -ACT2040 - Partie 4 : Modèles linéaires généralisés*, <https://freakonometrics.hypotheses.org/>, Université du Québec à Montreal.
- [4] BREIMAN L. (2001) *Random Forests*. Statistics Department, Berkeley
- [5] BERK R.A. (2004) *An introduction to Ensemble Methods for Data Analysis*. Department of Statistics
- [6] Xavier D'Haultfoeuille (2018), *Cours Econométrie 2 ENSAE, Chapitre 3 : modèles binaires*.

# Annexes

# Annexe A

## Base de données et modélisation.

### A.1 Base de données.

#### Liste des variables de la base initiale.

TABLEAU A.1 – Liste des variables de la base finale construite.

| Variables                       | Modalités       | Descriptions                                                               |
|---------------------------------|-----------------|----------------------------------------------------------------------------|
| Annee                           | 2019,2021,2022  | Année d'étude.                                                             |
| Cdregion                        | 64,67,...       | Région du client.                                                          |
| Charge total                    | 1000,...        | Charge total de la sinistralité de l'année N-1.                            |
| Charge total_3A                 | 1000,...        | Charge total de la sinistralité sur les 3 dernières années.                |
| CophtAxa_multi<br>(Prime_multi) | 1000,...        | Cotisation potentielle hors taxe du client sur l'ensemble de ses contrats. |
| CophtAXA_RC<br>(Prime Rc)       | 1000,...        | Cotisation potentielle hors taxe du contrat RC .                           |
| Distributeur_OVC                | Agent,Courtier  | Distributeur (agent courtier).                                             |
| DTFAN_ACEA                      | 12-Jan-2009,... | Date d'effet.                                                              |
| Effet                           | QP,PB,DP        | Type de majoration appliqué au contrat lors du renouvellement.             |
| Grille_majo                     | S01<5K,...      | Grille utilisée comme base de majoration pour le renouvellement.           |
| Indice_OVC                      | 1,2,3,4,5       | Indice client.                                                             |
| Majo_RC                         | 1000,...        | Montant de la majoration du contrat RC.                                    |
| Majo_multi                      | 1000,...        | Montant de la majoration du client sur l'ensemble de ses contrats.         |
| Nb branche                      | 1,2,...         | Nombre de branches du client.                                              |
| Continuer à la page suivante    |                 |                                                                            |

**Tableau A.1 – Liste des variables de la base finale construite.**

| <b>Variables</b>    | <b>Modalités</b> | <b>Descriptions</b>                                                                                                                                                |
|---------------------|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Nb_RC               | 1,2,...          | Nombre de contrats RC du client.                                                                                                                                   |
| Nb_Auto             | 0,1,2,...        | Nombre de contrats Automobile du client.                                                                                                                           |
| Nb_Const            | 0,1,2,...        | Nombre de contrat Construction du client.                                                                                                                          |
| Nb_RI               | 0,1,2,...        | Nombre de contrats RI du client.                                                                                                                                   |
| Nb_RT               | 0,1,2,...        | Nombre de contrats RT du client.                                                                                                                                   |
| Nb_TT               | 0,1,2,...        | Nombre de contrats TT du client.                                                                                                                                   |
| Nbcours             | 0,1              | Indicateur du contrat en cours l'année N.                                                                                                                          |
| Nbresil             | 0,1              | Indicateur du contrat en cours l'année N.                                                                                                                          |
| Nbsinistre          | 0,1,2,...        | Nombre de sinistres enregistré l'année N-1.                                                                                                                        |
| Nbsinistre3A        | 0,1,2,...        | Nombre de sinistres enregistré sur les 3 dernières années.                                                                                                         |
| Numcnt              |                  | Numéro du client.                                                                                                                                                  |
| Resil_autre branche | 0,1              | Indicateur de la résiliation des autres branches du client.                                                                                                        |
| Segmentation        | S01,...          | Segment du contrat.                                                                                                                                                |
| TX_majo_RC          | 0,0.1,0.25,...   | Taux de majoration du contrat RC.                                                                                                                                  |
| Tx_majo_multi       | 0,0.1,0.25,...   | Taux de majoration du client sur l'ensemble de ses contrats.                                                                                                       |
| Age_contrat Annee   | 1,2, ...         | Âge du contrat en année.                                                                                                                                           |
| Age_contrat mois    | 12,15, ...       | Âge du contrat en mois.                                                                                                                                            |
| B_Multi             | 1234, 1034,...   | Les différentes branches du client multi-équipé.                                                                                                                   |
| Multi_équipé        | 0,1              | Indicateur du multi-équipement.                                                                                                                                    |
| Nb_contrat          | 1,2,...          | Nombre de contrats du client.                                                                                                                                      |
| Diff_majo           | -1,0,1           | Variable indiquant si le taux de majoration RC est supérieur ou inférieur au taux de majoration global du client.                                                  |
| Evolution_multi     | -1,0,1           | Évolution du multi-équipement (il indique si par rapport à l'année précédente le client a résilié l'un de ces contrat ou alors souscrit à des affaires nouvelles). |

## Liste des variables créées.

TABLEAU A.2 – Liste des variables créées.

| <b>Variables</b>             | <b>Variables initiales</b> | <b>Description</b>                                |
|------------------------------|----------------------------|---------------------------------------------------|
| Prime_multiG                 | Prime_multi                | Variable prime_multi transformée en catégorielle. |
| Prime_RCG                    | Prime_RC                   | Variable Prime_RC transformée en catégorielle.    |
| Continuer à la page suivante |                            |                                                   |

**Tableau A.2 – Liste des variables créées.**

| <b>Variables</b> | <b>Variables initiales</b> | <b>Description</b>                                                                                                                               |
|------------------|----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| Majo_RCG         | Majo_RC                    | Variable Majo_RC transformée en catégorielle.                                                                                                    |
| Majo_multiG      | Majo_multi                 | Variable Majo_multi transformée en catégorielle.                                                                                                 |
| TX_majo_RCG      | TX_majo_RC                 | Variable TX_majo_RC transformée en catégorielle.                                                                                                 |
| Tx_majo_multiG   | Tx_majo_multi              | Variable Tx_majo_multi transformée en catégorielle.                                                                                              |
| Nbsinistre3A_G   | Nbsinistre3A               | Variable Nbsinistre3A transformée en catégorielle.                                                                                               |
| Charge3A_G       | Charge3A                   | Variable Charge3A transformée en catégorielle.                                                                                                   |
| Nbbranche_G      | Nbbranche                  | On a groupé les branches 5 et 6.                                                                                                                 |
| Cdregion_G       | Cdregion                   | On a regroupé les modalités 13=DOM TOM et 67=Sud Est car en interne, les DOM-TOM sont gérés par la région Sud Est.                               |
| Distributeur_G   | Distributeur_OVC           | On a regroupé les modalités "agent et courtier" et "agent".                                                                                      |
| Grille_majoG     | Grille_majo                | Les S01 et S04 de moins de 5% ont été regroupées entre eux et les autres modalités de moins de 5% ont été regroupées avec la modalité "Autres" . |
| Indice_OVCG      | Indice_OVC                 | On a regroupé les modalités 1 et 2 puis 4 et 5.                                                                                                  |
| Age_contratG     | Age_contrat                | A plus de 10 ans, on a regroupé en classe ([10,15[ ; [15,20[ ; [20 et plus]).                                                                    |
| Nbcontrat_G      | Nbcontrat                  | A plus de 5 contrats, on a regroupé les modalités ([5,10[ ; [10 et plus]).                                                                       |

## A.2 Analyse des variables et modélisation de la résiliation.

### Analyse des corrélations entre les variables.

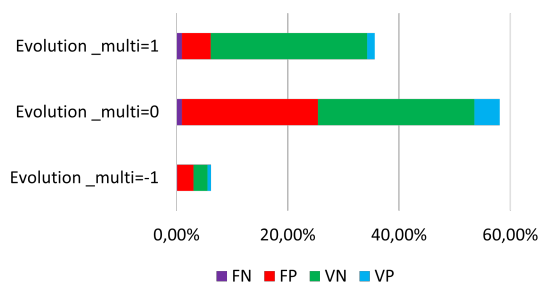
TABLEAU A.3 – Test du khi-deux d'indépendance entre la résiliation et les autres variables du modèle

| Variables                  | Statistique du test | Pvalue        | Degré de liberté |
|----------------------------|---------------------|---------------|------------------|
| Segmentation               | 713.415052          | 1.225293e-140 | 17               |
| Grille majo majoration     | 681.480866          | 3.007713e-132 | 19               |
| Cophtaxa RC (Gp)           | 131.696009          | 5.647018e-26  | 6                |
| Cophtaxa Global (Gp)       | 1026.432414         | 2.340586e-217 | 7                |
| Majoration RC (Gp)         | 2316.926428         | 0.000000e+00  | 7                |
| Majoration global (Gp)     | 1968.950886         | 0.000000e+00  | 7                |
| Taux majoration RC (Gp)    | 2999.935768         | 0.000000e+00  | 8                |
| Taux majoration multi (Gp) | 2107.204378         | 0.000000e+00  | 8                |
| Age contrat (Gp)           | 775.592099          | 7.212836e-154 | 17               |
| Sinitralité (Gp)           | 498.679561          | 1.841925e-110 | 1                |
| Charge sinistre (Gp)       | 295.018429          | 1.286220e-62  | 4                |
| Distributeur               | 131.635009          | 2.402753e-28  | 3                |
| Indice client              | 184.786757          | 6.988462e-39  | 4                |
| Code région                | 113.660726          | 6.894979e-23  | 5                |
| Nb cours                   | 352872.144529       | 0.000000e+00  | 1                |
| Effet                      | 215273.513401       | 0.000000e+00  | 3                |
| Nombre contrat (Gp)        | 1315.821396         | 2.385714e-282 | 5                |
| Multi-équipé               | 1230.220155         | 1.650760e-269 | 1                |
| Branche multiéquipement    | 1696.584592         | 0.000000e+00  | 15               |
| Nombre branche (Gp)        | 770.364481          | 2.978734e-164 | 5                |
| Diff-majo                  | 461.680249          | 5.589948e-101 | 2                |
| Évolution multiéquipement  | 863.256891          | 3.516802e-188 | 2                |

**Source : Interne AXA.**

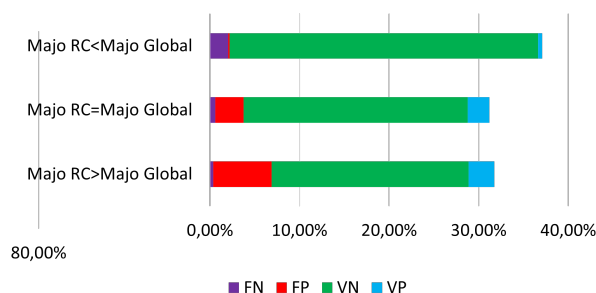
## A.3 Caractérisation des prédictions du modèle de résilience.

FIGURE. A.1 – Caractérisation des prédictions : Évolution du multi-équipement.



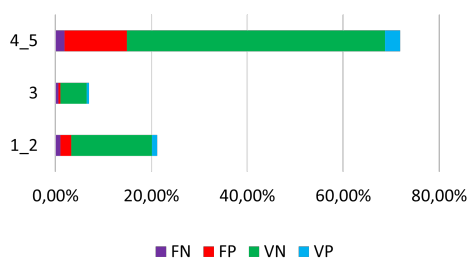
Source : Interne AXA.

FIGURE. A.2 – Caractérisation des prédictions : Différence de majoration.



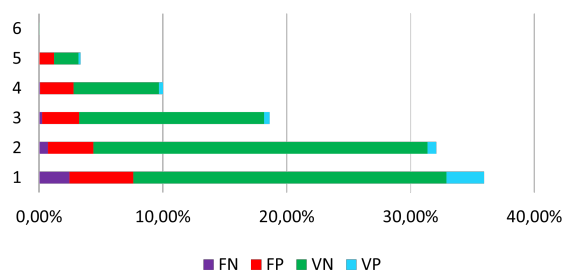
Source : Interne AXA.

FIGURE. A.3 – Caractérisation des prédictions : Indice client.



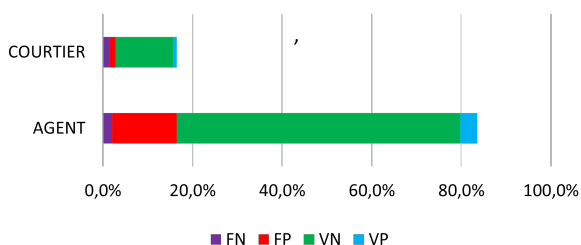
Source : Interne AXA.

FIGURE. A.4 – Caractérisation des prédictions : nombre de branches.



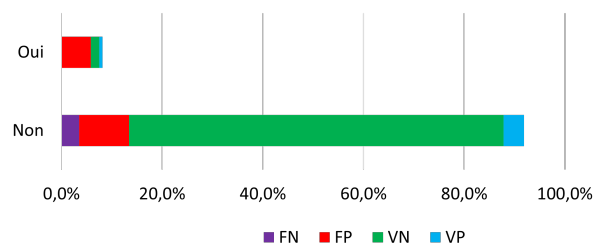
Source : Interne AXA.

FIGURE. A.5 – Caractérisation des prédictions : distributeur.



Source : Interne AXA.

FIGURE. A.6 – Caractérisation des prédictions : Sinistralité des 3 dernières années.



Source : Interne AXA.



## Annexe B

### Modèles linéaires généralisés.

En statistique, l'un des modèles le plus plébiscité est le modèle linéaire gaussien (le modèle de régression linéaire classique) ceci en raison de sa simplicité. Toutefois, il est peu adapté pour répondre aux problématiques assurantielles et plus précisément en tarification :

- La loi normale n'est souvent pas adaptée à la modélisation des variables d'intérêt. En effet il s'agit d'une loi continue et à valeur dans  $\mathbb{R}$ . Or en assurance on s'intéresse la plupart du temps au nombre de sinistres à valeurs dans  $\mathbb{N}$ , au coût d'un sinistre à valeurs dans  $\mathbb{R}^+$  ou à la probabilité d'avoir un sinistre à valeurs dans  $[0, 1]$ .
- La relation linéaire entre la variable réponse et les variables explicatives n'est pas nécessairement adaptée à toutes les modélisations et impose d'importantes limitations.
- L'homoscédasticité supposée dans le modèle linéaire gaussien impose aussi certaines limites et ne traduit pas nécessairement la réalité des variables dont on souhaite étudier le "comportement".

Ainsi pour palier à ses limites tout en conservant ses bonnes propriétés (simplicité, explicabilité) les modèles linéaires généralisés sont couramment utilisés par les actuaires pour les questions de tarification. Dans la suite, en s'appuyant sur le chapitre 9 de Denuitn Charpentier(2005), nous allons présenter le modèle GLM partant de sa formulation à son estimation.

#### Définition des GLM.

Pour commencer, nous définissons d'abord la *famille exponentielle*. La famille exponentielle regroupe les lois de probabilité dont la densité est donnée par :

$$\ln f_X(x; \theta) = \frac{\theta x - b(\theta)}{a(\phi)} + c(x, \phi), x \in \mathbb{X}, \quad (\text{B.1})$$

avec  $\mathbb{X}$  le support de la variable aléatoire ;  $a, b, c$  trois fonctions différentiables propres à la loi,  $\theta$  le paramètre d'intérêt et  $\phi$  le paramètre de dispersion.

Ainsi définit la famille exponentielle, un modèle linéaire généralisé (GLM) est caractérisé par trois hypothèses :

- **Une loi de probabilité** : les  $(Y_i)_i$  sont indépendantes et  $Y_i$  suit une loi  $f_{exp}(\theta_i, \Phi_i, a, b, c)$  appartenant à la famille exponentielle.
- **Une fonction déterministe** : le vecteur de variables explicatives  $X_i$  donne le prédicteur linéaire  $\eta_i = X_i^T \beta$ .
- **Une fonction lien**  $g : \mathbb{R} \rightarrow \mathbb{X}$  monotone, différentiable et inversible telle que  $\mathbb{E}[Y_i] = g^{-1}(\eta_i)$ , pour  $i \in 1, \dots, n$ , ou  $\theta_i$  est le paramètre d'échelle,  $\phi$  le paramètre de dispersion et  $a, b, c$  trois fonctions.

Notons que les paramètres  $\theta_i$  sont liés au prédicteur linéaire par la relation :

$$\mu_i = \mathbb{E}(Y_i) = b'(\beta_i) = g^{-1}(\eta_i). \quad (\text{B.2})$$

Si  $\beta_i = \eta_i$ , alors la fonction lien est dite canonique. C'est-à-dire  $\beta = g(b'(\beta_i))$  ie  $g(x) = ((b')^{-1}(x))$ . Le paramètre de dispersion  $\phi_i$  est en général commun à toutes les observations à une transformation connue près. On suppose généralement que  $\phi_i = \frac{\phi}{\omega_i}$  pour  $\omega_i$  un poids connu, typiquement 1 pour les lois positives ou  $n_i$  pour la loi Binomiale.

## Estimation des paramètres.

L'estimateur du vecteur de paramètres  $\beta$  et du paramètre  $\Phi$  se fait par la méthode de maximum de vraisemblance. La log-vraisemblance est :

$$\ln \mathbb{L}(\beta, \Phi) = \sum_{i=1}^n \frac{\theta_i(\beta) y_i - b(\theta_i(\beta))}{a(\phi_i)} + \sum_{i=1}^n c(y_i, \phi_i),$$

avec  $\theta_i = (b')^{-1}(g^{-1}(\eta_i))$ . Les équations du score dans le cas général sont :

$$\forall j = 1, \dots, d, \sum_{i=1}^n \frac{y_i - \mu_i}{a(\phi_i) V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} = 0. \quad (\text{B.3})$$

Dans le cas du lien canonique ( $g(.) = \ln(.)$ ), on a :

$$\forall j = 1, \dots, d, \sum_{i=1}^n \frac{y_i - \mu_i}{a(\phi_i)^2} x_{ij} = 0. \quad (\text{B.4})$$

La solution explicite de  $\beta$  peut se déterminer par la méthode de Newton pour les équations du score.

## Mesure d'adéquation.

Le modèle le plus simple appelé **modèle nul** correspond au modèle avec un seul paramètre commun  $\mu$  toutes les observations  $y = (y_1; \dots; y_n)$ . Toute la variation des observations  $y_1; \dots; y_n$  est expliquée par la composante aléatoire. A l'opposé, le plus complexe appelé **modèle complet** correspond avec autant de paramètre que d'observations. Toute la variation des observations  $y_1; \dots; y_n$  est expliquée par la composante déterministe. Notons  $\log L(\mu; \phi; y)$  la log-vraisemblance en fonction du paramètre de moyenne  $\mu$  plutôt que du paramètre canonique  $\theta$ . La déviance pour un modèle ayant  $p$  variables explicatives est par le ratio de log-vraisemblance suivant :

$$D(\hat{\mu}; \phi; y) = 2\ln L(y; \phi; y) - 2\ln L(\hat{\mu}; \phi; y).$$

Pour la famille exponentielle, on obtient :

$$D(\hat{\mu}; \phi; y) = \sum_{i=1}^n \frac{2y_i(\theta(\hat{\mu}_i) - \theta(y_i)) - b(\theta(\hat{\mu}_i)) + b(\theta(y_i))}{a(\phi)}, \quad (\text{B.5})$$

ou  $\mu_i$  le paramètre calibré ayant les  $p$  variables explicatives,

$\theta(\hat{\mu}_i)$  le paramètre canonique pour  $\hat{\mu}_i$ ,

$\theta(y)$  le paramètre canonique pour le modèle complet.

Plus le modèle est explicatif plus la déviance est petite. Cependant, la déviance est asymptotiquement de loi du chi-deux  $(n - k)$  : on ne dispose plus d'une loi exacte comme pour les modèles linéaires. Ainsi, l'analyse de déviance permet de sélectionner les variables explicatives les plus importantes mais on ne peut plus interpréter ces résultats en terme d'orthogonalité.

# Note de synthèse.

Le multi-équipement désigne le phénomène par lequel un client possède plusieurs produits ou services auprès d'une même entreprise ou d'un même réseau. Il est un enjeu crucial pour les entreprises car il permet de développer le potentiel des clients du portefeuille sans en acquérir de nouveaux à un coût élevé. Au sein d'AXA France, une entité est spécialisée dans l'assurance des clients "entreprises" (AXA France IARD Entreprises). Pour des besoins de logistique et de comptabilité, ces clients logent le plus souvent l'ensemble de leurs contrats d'assurance chez un même assureur ce qui contribue grandement au développement du multi-équipement.

La responsabilité civile Entreprise (RC), produit d'assurance visant à réparer les dommages corporels, matériels ou immatériels causés à un des tiers ou des clients, est avec 15% du chiffre d'affaires l'un des principaux produits d'AXA France IARD Entreprises. Par ailleurs, avec un taux de multi-équipement d'environ 56% en nombre de clients et de 68% en chiffre d'affaires, il est le produit le plus multi-équipé. De ce fait, la fidélisation des clients de la branche RC devient un enjeu crucial pour AXA, car une non-satisfaction d'un client sur cette branche peut l'inciter à résilier l'ensemble de ses contrats.

À ce jour, les renouvellements des contrats d'assurance au sein d'Axa France n'intègrent pas la composante du multi-équipement, on se demande alors si la prise en compte du multi-équipement dans les politiques de renouvellement pourrait conduire à une plus forte fidélisation de la clientèle. Pour répondre à cette question, nous avons dans un premier temps construit des indicateurs servant à caractériser le multi-équipement des clients. Ensuite, nous avons analysé la résiliation des clients du portefeuille RC en considérant aussi bien les clients mono-équipés que multi-équipés. Afin de percevoir l'impact du multi-équipement sur la résiliation des clients RC, une seconde analyse utilisant les indicateurs de multi-équipement préalablement construits a été faite uniquement sur les clients multi-équipés.

L'étude du comportement de résiliation s'est faite avec les modèles logistique, CART et Random Forest. Après analyse et comparaison des différents modèles, le modèle logistique a été retenu comme meilleur modèle, ceci en raison de ses bonnes performances prédictives donnant

une sensibilité d'environ 75% (contre une sensibilité de moins de 1% pour les CART et Random Forest).

TABLEAU 4.1 – Matrice de confusion modèle logistique.

| <b>Prédictions \ Observations</b> | Non Résilié  | Résilié     | <b>Total</b> |
|-----------------------------------|--------------|-------------|--------------|
| Non Résilié                       | 58.7%        | 2.1%        | 60.9%        |
| Résilié                           | 32.5%        | 6.6%        | 39.1%        |
| <b>Total</b>                      | <b>91.3%</b> | <b>8.7%</b> | <b>100%</b>  |

**Source : Interne AXA.**

L'analyse de ce modèle logistique a révélé que le taux majoration client (pourcentage d'augmentation de la prime d'assurance lors du renouvellement) est l'un des principaux facteurs de résiliation. Plus spécifiquement, le pic de résiliation chez les clients multi-équipés s'observait à un taux de majoration de 4.5% contre 3.5% dans le portefeuille global (mono-équipé et multi-équipés). Ce résultat nous a conforté sur la nécessité d'appliquer des taux de majorations spécifiques aux clients multi-équipés lors du renouvellement. En analysant le modèle de résiliation, les alertes suivantes ont été transmises aux souscripteurs dans le but de fidéliser davantage la clientèle multi-équipée qui fait partir des clients les plus rentables d'AXA :

- Le poids de la majoration RC dans la majoration globale d'un client multi-équipé influe sur son comportement de résiliation :
  - Si le taux de majoration de la branche RC du client est inférieur au taux de majoration global sur l'ensemble du portefeuille, alors le client a de faibles chances de résilier ;
  - Si le taux de majoration RC est supérieur à son taux de majoration global, et si de plus le client n'a pas eu de sinistralité sur les trois dernières années alors attention le client a de fortes chances de résilier.
- L'évolution du multi-équipement est un indicateur pertinent de fidélisation des clients multi-équipés :
  - Si un client est en croissance de portefeuille (le client a augmenté son nombre de contrats de l'année N-1) alors ce dernier a peu de chances de résilier ;
  - Si un client est en décroissance de portefeuille et de plus n'a pas eu de sinistre sur les 3 dernières années alors attention, il a de fortes chances de résilier.
- Si le nombre de branches d'un client multi-équipé est supérieur à 3 alors sa probabilité de résilier est faible ;
- Si le distributeur est agent, la probabilité de résilier du client diminue.

Ayant constaté que les clients de majoration RC supérieure à la majoration globale (moyenne des majorations de tous les contrats du client) présentaient plus de chance de résilier leur contrat RC, la deuxième partie du travail s'est consacrée à la compréhension de la majoration des clients multi-équipés sujets à la résiliation. Cette problématique a été résolue en utilisant la distance de Levenshtein pour retrouver les jumeaux de chaque potentiel résilieur. L'algorithme utilisé pour rechercher les jumeaux consistait à créer une variable concaténant les caractéristiques de chaque client, puis à calculer la distance de Levenshtein entre les clients en utilisant cette nouvelle variable et enfin, à déduire les jumeaux qui correspondaient aux clients ayant la plus faible distance. Plus spécifiquement elle s'est faite en suivant ces différentes étapes :

- **Création l'indicateur de ressemblance** : dans la base de données, créer une variable pouvant être utilisée comme indicateur de ressemblance entre les clients. Ce dernier pour nous était simplement la concaténation de toutes les caractéristiques de chaque client. Par exemple prenons deux clients (c1 et c2) tels que, c1 : Évolution du multi-équipement=-1, distributeur=agent, ancienneté contrat = 1 ; et c2 : Évolution du multi-équipement=1, distributeur=courtier, ancienneté contrat = 3. Alors pour ces deux clients, l'indicateur de ressemblance noté *id\_ressemblance* est *-1agent1* et *1courtier3* pour c1 et c2 respectivement.
- **Ré-codage des modalités** : on aurait pu évaluer les distances entre les clients c1 et c2 en faisant  $distance(c1, c2) = lev(-1agent1, 1courtier3)$ . Mais on a remarqué que les chaînes de caractères n'ont pas la même taille ( $| -1agent1 | = 7$  et  $| 1courtier3 | = 10$ ) ce qui pouvait conduire à une source d'erreurs dans le calcul des distances. Alors pour pallier à cela, la solution consistait à ré-coder les modalités de telle sorte que chacune tienne sur un seul caractère. Exemple : Pour l'ancienneté du contrat : 1 = 1, 2 = 2, 3 = 3. Pour l'évolution du multi-équipement : -1 = 4, 0 = 5, 1 = 6. Pour le distributeur *agent* = *a*, *courtier* = *c*. Remarquons que pour ré-coder l'évolution du multi-équipement, on a été obligé de commencer par 4 car l'ancienneté du contrat occupait déjà les valeurs 1, 2, 3. Au final on a obtenu c1 : *id\_ressemblance* = *4a1* ; c2 : *id\_ressemblance* = *6c3*.
- **Concaténation des variables** : le nouvel indicateur de ressemblance a été calculé en concaténant les variables ré-codées en fonction de leur importance dans le modèle de résiliation (la variable la plus importante occupant la première position dans la chaîne et ainsi de suite).
- **Segmentation de la base** : on a divisé la base de données en deux pour avoir une base des résiliés et une base des non résiliés. Ensuite on a retiré de ces deux bases les clients sinistrés. En effet suite à une sinistralité, la majoration appliquée au client lors du renouvellement peut être particulière. Pour éviter les aléas liés aux antécédents de sinistralité, on s'est restreint sur les clients non sinistrés.
- **Calcul de la distance de Levenshtein** : pour chaque client de la base des résiliés on a évalué la distance entre lui et tous les clients non résiliés. Pour tenir compte de

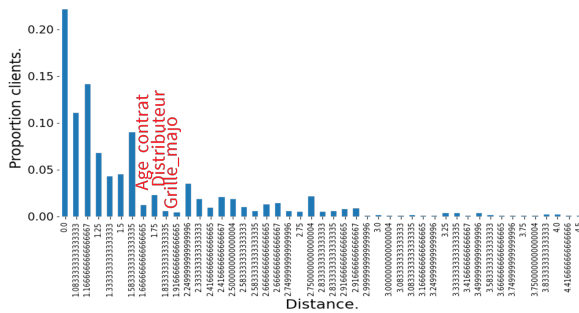
l'importance de chaque variable dans la distance, on a pondéré par la position de cette dernière dans la chaîne de caractères. La formule de la distance de Levenshtein utilisée à cet effet est :

$$\text{lev}(a, b) = \begin{cases} \max(|a|, |b|) & \text{si } \min(|a|, |b|) = 0, \\ \frac{|a|-|a-1|}{|a|} + \text{lev}(a-1, b-1) & \text{si } a[0] = b[0], \\ 1 + \frac{|a|-|a-1|}{|a|} + \min \begin{cases} \text{lev}(a-1, b) \\ \text{lev}(a, b-1) \\ \text{lev}(a-1, b-1) \end{cases} & \text{sinon.} \end{cases} \quad (4.6)$$

- **Identification des jumeaux** : après le calcul des distances, les jumeaux du client résilié correspondaient à ceux ayant la distance de Levenshtein minimal.

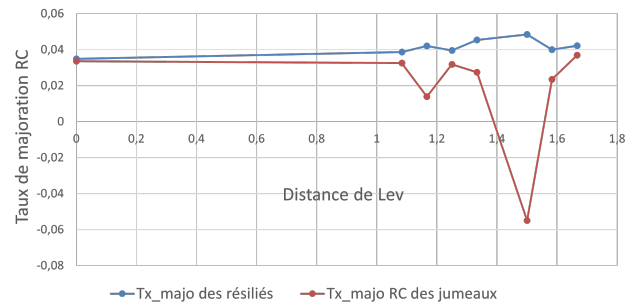
En implémentant sur une base de données de 183 000 clients RC multi-équipés, nous avons remarqué qu'environ 70% de nos clients résiliés avaient des distances inférieures à 2 de leurs jumeaux (figure 4.1). Ceci indiquant qu'on pouvait trouver la majoration optimale des résiliés à partir de celle de leurs jumeaux non résiliés. Puis en évaluant la majoration moyenne chez les résiliés et leurs jumeaux non résiliés, il en est ressorti que le taux de majoration moyen des résiliés était d'environ 4% pour tous niveaux de distances et celui de leurs jumeaux de 3% (figure 4.2). Donc à caractéristiques similaires, les clients résiliés avaient reçu plus de majoration que leurs jumeaux non résiliés.

FIGURE. 4.1 – Distance de Levenshtein entre les résiliés et leurs jumeaux non résiliés.



Source : Interne AXA.

FIGURE. 4.2 – Taux de majoration des résiliés vs jumeaux non résiliés.



Source : Interne AXA.

Une analyse plus fine suivant les montants de majoration en euros et les primes nous a révélé que :

- Chez les clients multi-équipés, lorsque la majoration RC est au-dessus de la majoration globale, les risques de résiliation sont énormes. Au vu de ce résultat, nous avons suggéré de proposer à un client multi-équipé une majoration flat pour tous ses contrats. En effet,

lorsqu'un client multi-équipé s'aperçoit qu'il est très majoré sur une branche particulière, il est fort probable qu'il résilie son contrat sur cette branche. Cette première résiliation va conduire à une baisse de son portefeuille (nombre de contrats année N inférieure au nombre de contrats année N-1). De plus, comme il a été montré dans le modèle de résiliation, la décroissance du portefeuille augmente de 40% les risques de résiliation. Ces risques cumulés de résiliation sont donc susceptibles d'entraîner la résiliation de tous les contrats du client initialement multi-équipé, ce qui pourrait être une perte pour AXA ceci d'autant plus si le client avait un portefeuille rentable au global. Une majoration unique pour tous les contrats du client semblerait donc être un bon paramètre de fidélisation en ce sens qu'il évitera les risques de résiliation pour la branche la plus majorée.

- Les clients multi-équipés en croissance de portefeuille ne doivent pas être considérés comme des "acquis" pour AXA, en effet dans le modèle de résiliation nous avons conclu que si un client multi-équipé accroît son portefeuille, il a peu de chances de résilier. Mais avec l'analyse de la distance de Levenshtein on a remarqué que si ces mêmes clients recevaient des majorations supérieures à la majoration moyenne, leurs probabilités de résilier augmentaient considérablement. En d'autres termes, les souscripteurs ne doivent pas majorer fortement les clients en croissance de portefeuille. Les majorations doivent être standardisées pour favoriser davantage la fidélisation des clients multi-équipés.
- Les résiliés ont deux grands facteurs qui les différencient de leurs jumeaux non résiliés (la prime et le montant de la majoration en euro). En général, les résiliés qui ont reçu des fortes majorations comparativement à leurs jumeaux non résiliés ont également eu des montants de majorations en euros très élevés (par rapport à leurs jumeaux et à la moyenne des majorations des résiliés). Dans le même sens, les clients qui résilient malgré qu'ils aient eu des taux de majoration plus faibles que leurs jumeaux ont des primes d'assurances très élevées. Ce résultat montre que les souscripteurs doivent faire attention au niveau de primes des clients lors du renouvellement. Une majoration de 5% pour un client ayant une prime de 10 000 euros (soit 500 euros de majoration) est très différente d'une majoration de 5% pour un client ayant une prime de 100 euros (soit 5 euros de majoration). Étant donné que la politique de renouvellement actuelle d'AXA propose des majorations suivant les niveaux de primes, les souscripteurs doivent davantage prêter attention aux clients ayant des grosses primes d'assurance car ces derniers sont beaucoup plus sensibles à une augmentation de leurs primes que les autres clients du portefeuille.

Au final, ce travail nous a permis d'analyser les clients RC multi-équipés dans leur globalité. Nous nous sommes appesantis sur les comportements de résiliation des clients. Sur la base des résultats obtenus, un ensemble d'alertes a été mis sur pieds et devront servir lors des renouvellements futurs, afin d'améliorer la politique actuelle et mieux fidéliser la clientèle multi-équipée. Ce travail s'est toutefois limité à l'analyse de la résiliation de la branche RC des



clients multi-équipés. Pour l'approfondir davantage, nous avons suggéré d'étudier la résiliation des autres branches des clients multi-équipés afin de tirer des conclusions plus générales sur la position de leurs majorations par rapport à la majoration globale et sur l'évolution du multi-équipement.

# Executive summary.

Multi-equipment is the phenomenon whereby a customer owns several products or services from the same company or network. A crucial issue for businesses, multi-equipment makes it possible to develop the potential of existing customers without acquiring new ones at a high cost. Within AXA France, one entity specializes in commercial insurance (AXA France IARD Entreprises), and therefore mainly manages "business" clients. For logistical and accounting reasons, these clients generally place all their insurance policies with the same insurer, which contributes greatly to the development of multi-equipment.

Corporate liability (CR), an insurance product designed to cover bodily injury, property damage or consequential loss caused to third parties or clients, is one of AXA France IARD Entreprises' main products, accounting for 15% of revenues. In addition, with a multi-equipment rate of around 56% in terms of the number of clients and 68% in terms of revenues, it is the most multi-equipped product. As a result, customer loyalty in the liability line of business is becoming a crucial issue for AXA, since customer dissatisfaction with this line of business can lead them to cancel all of their policies.

To date, Axa France's insurance policy renewals do not include the multi-equipment component, so we wonder if taking multi-equipment into account in renewal policies could lead to greater customer loyalty? To answer this question, we first constructed indicators to characterize customer multi-equipment. Then, we analyzed contrat termination in the CR portfolio, considering single and multi-equipment customers. In order to perceive the impact of multi-equipment on the termination of CR customers, a second analysis using the previously constructed indicators of multi-equipment was performed only on multi-equipped customers.

The study of termination behavior was conducted by using logistic, CART and Random Forest models. After analysis and comparison of the different models, the logistic model was selected as the best model because of its good predictive performance, giving a sensitivity of about 75% (compared to a sensitivity of less than 1% for CART).

TABLEAU 4.2 – Confusion matrix logistic model.

| <b>Observations</b><br><b>Predictions</b> | Not Ter-<br>minated | Terminated  | <b>Total</b> |
|-------------------------------------------|---------------------|-------------|--------------|
| Not Terminated                            | 58.7%               | 2.1%        | 60.9%        |
| Terminated                                | 32.5%               | 6.6%        | 39.1%        |
| <b>Total</b>                              | <b>91.3%</b>        | <b>8.7%</b> | <b>100%</b>  |

**Source : Interne AXA.**

The analysis of this logistic model revealed that the customer surcharge rate (percentage increase in insurance premium at renewal) is one of the main drivers of termination. Specifically, the peak of termination for multi-equipment customers was observed at a markup of 4.5% versus 3.5% for the overall portfolio (single and multi-equipment). This result reinforced the need to apply specific markup rates to multi-equipment customers at renewal. In analyzing the termination model, the following alerts were sent to policyholders in an effort to increase the loyalty of the multi-equipment customer base, which is one of AXA's most profitable customers :

- The weight of the liability surcharge in the overall surcharge of a multi-equipment client influences his or her cancellation behavior :
  - the customer's liability line markup is lower than the overall markup on the entire portfolio, then the customer is unlikely to terminate ;
  - the client's liability markup rate is higher than its overall markup rate, and if the client has had no claims experience over the past three years, then the client is likely to terminate.
- The evolution of multi-equipment is a relevant indicator of the loyalty of multi-equipped clients :
  - a customer is growing his portfolio (the customer has increased his number of contracts from the year N-1) then he is unlikely to cancel ;
  - a client is in portfolio decrease and moreover has not had any claim over the last 3 years, then beware, he has a high chance of cancelling.
- the number of branches of a multi-equipped customer is higher than 3 then his probability to cancel is low ;
- the distributor is an agent, the customer's probability of terminating decreases.

Having found that customers who received CR markups greater than their overall markups (average of the markups of all the customer's contracts) were more likely to terminate their CR contract, the second part of the work focused on understanding the markup of multi-equity customers prone to termination. This problem was solved by using Levenshtein distance to

find the twins of each potential terminator. The algorithm used to find the twins consisted of creating a variable concatenating the characteristics of each customer, then calculating the Levenshtein distance between the customers using this new variable and finally, deducing the twins that corresponded to the customers with the lowest distance. More specifically, it was done by following these different steps :

- **Creation of the similarity indicator** : in the database, create a variable that can be used as an indicator of similarity between customers. The latter for us was simply the concatenation of all characteristics of each customer. For example, let's take two customers (c1 and c2) such that, c1 : Multi-equipment evolution=-1, distributor=agent, contract age = 1 ; and c2 : Multi-equipment evolution=1, distributor=broker, contract age = 3. Then for these two customers the similarity indicator noted *id\_ressemblance* is -1agent1 and 1broker3 for c1 and c2 respectively.
- **Re-coding of the modalities** : we could have evaluated the distances between the customers c1 and c2 by doing  $distance(c1, c2) = lev(-1agent1, 1broker3)$ . But we noticed that the strings do not have the same size ( $| -1agent1 | = 7$  and  $| 1broker3 | = 10$ ) which could lead to a source of errors in the calculation of the distances. So to overcome this, the solution was to re-code the terms so that each one fits on a single character. Example : For the age of the contract : 1 = 1, 2 = 2, 3 = 3. For the evolution of the multi-equipment : -1 = 4, 0 = 5, 1 = 6. For the distributor *agent* = *a*, *broker* = *c*. Note that to re-code the evolution of the multi-equipment, we were obliged to start with 4 because the age of the contract already occupies the values 1, 2, 3. Thus we have c1 : *id\_resemblance* = 4a1 ; c2 : *id\_resemblance* = 6c3.
- **Concatenation of variables** : the new similarity indicator was calculated by concatenating the re-coded variables according to their importance in the termination model (the most important variable occupies the first position in the chain and so on).
- **Segmentation of the base** : we divided the database in two to have a base of the terminated and a base of the not terminated. Then we withdrew from these two bases the damaged customers. Indeed, following a claim, the increase applied to the customer at the time of renewal can be particular. In order to avoid the uncertainties linked to the history of claims, we have restricted ourselves to customers who have not made a claim. of the Levenshtein distance : for each client in the terminated base, we evaluated the distance between him and all the nonterminated clients. To take into account the importance of each variable in the distance, we weighted it by its position in the string. The Levenshtein

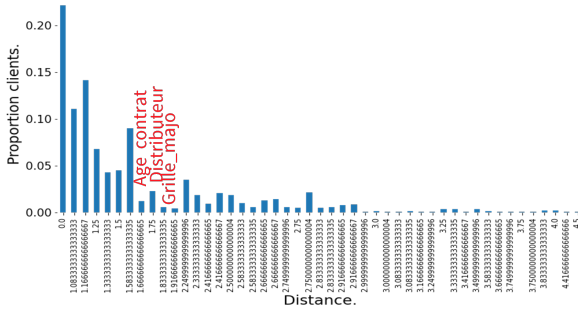
distance formula used for this purpose is :

$$\text{lev}(a, b) = \begin{cases} \max(|a|, |b|) & \text{si } \min(|a|, |b|) = 0, \\ \frac{|a| - |a-1|}{|a|} + \text{lev}(a-1, b-1) & \text{si } a[0] = b[0], \\ 1 + \frac{|a| - |a-1|}{|a|} + \min \begin{cases} \text{lev}(a-1, b) \\ \text{lev}(a, b-1) \\ \text{lev}(a-1, b-1) \end{cases} & \text{sinon.} \end{cases} \quad (4.7)$$

- **Twin identification** : after computing the distances, the terminated client's twins matched those with the minimum Levenshtein distance.

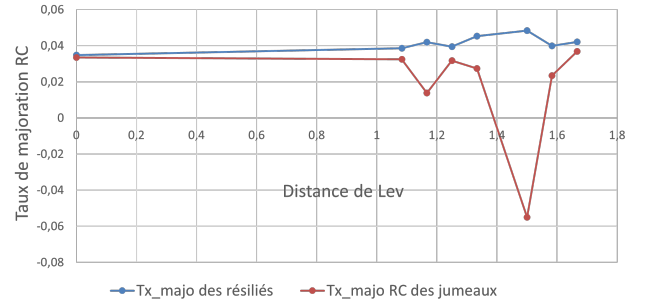
After implementing on a database of 183,000 multi-equipped RC customers, we found that about 70% of our terminated customers had distances less than 2 of their twins (4.3). Then, when we evaluated the average markup for terminated customers and their non-cancelled twins, we found that the average markup for terminated customers was about 4% for all distance levels and for their twins was about 3% (Figure 4.4). Thus, for similar characteristics, terminated clients received more markup than their non-terminated twins.

FIGURE. 4.3 – Levenshtein distance between the terminated and their untermi-  
nated twins.



Source : Interne AXA.

FIGURE. 4.4 – Terminated vs. untermi-  
nated twin markup rate.



Source : Interne AXA.

A more detailed analysis according to the amounts of markup in euros and premiums revealed that :

- multi-equipment clients, when the RC mark-up is higher than the overall mark-up, the risk of cancellation is enormous. In view of this result, we suggested that multi-equipment customers be offered a flat surcharge for all their contracts. Indeed, when a multi-equipped customer realizes that he has a high mark-up on a particular branch, he is likely to cancel his contract on this branch. This first cancellation will lead to a decrease in his portfolio (number of contracts year N lower than the number of contracts year N-1). Moreover, as shown in the termination model, the decrease in the portfolio increases the termination

risks by 40%. These cumulative risks of termination are therefore likely to lead to the termination of all the contracts of the initially multi-equipped client, which could be a loss for AXA, especially if the client had a profitable portfolio overall. A single surcharge for all of the client's contracts would therefore seem to be a good parameter for building loyalty in that it would avoid the risk of termination for the branch with the highest surcharge.

- Multi-equity customers with portfolio growth should not be taken for granted by AXA, as in the termination model we concluded that if a multi-equity customer increases his portfolio, he is unlikely to terminate. But with the Levenshtein distance analysis we noticed that if these same customers received higher than average markups, their probability of terminating increased considerably. In other words, underwriters should not heavily markup growing customers. Markups should be standardized to further encourage retention of multi-equity customers.
- The terminated have two main factors that differentiate them from their non-terminated twins (the premium and the amount of the euro markup). In general, terminated customers who received high markups compared to their non-cancelled twins also had very high Euro markup amounts (compared to their twins and to the average terminated customer's markup). Similarly, clients who cancel despite having had lower surcharge rates than their twins have very high insurance premiums. This result shows that underwriters need to be careful about the level of customer premiums at renewal. A 5 percent surcharge for a client with a premium of €10,000 (i.e., €500 surcharge) is very different from a 5 percent surcharge for a client with a premium of €100 (i.e., €5 surcharge). Given that AXA's current renewal policy offers surcharges based on premium levels, underwriters need to pay more attention to customers with large insurance premiums because they are much more sensitive to a premium increase than other customers in the portfolio.

In the end, this work allowed us to analyze the multi-equipped RC customers in their entirety. We focused on the termination behavior of the clients. Based on the results obtained, a set of alerts was set up and should be used during future renewals, in order to improve the current policy and better retain the multi-equipped customers. However, this work was limited to the analysis of the termination of the liability branch of multi-equipped clients. In order to further investigate the issue, we suggested studying the termination of other lines of business of multi-equipped clients in order to draw more general conclusions on the position of their mark-ups in relation to the overall mark-up and the evolution of multi-equipment.

# Table des figures

|      |                                                                                     |    |
|------|-------------------------------------------------------------------------------------|----|
| 1.1  | Multi-équipement RC 2020 : Répartition de la prime AXA. . . . .                     | 8  |
| 2.1  | Schéma de construction de la base de données 2020. . . . .                          | 12 |
| 2.2  | Répartition du nombre de contrats et multi-équipement par année. . . . .            | 14 |
| 2.3  | Répartition de la prime RC et du multi-équipement par année. . . . .                | 14 |
| 2.4  | Répartition des clients mono et multi-équipé suivant la résiliation. . . . .        | 14 |
| 2.5  | Répartition des clients mono et multi-équipé suivant la résiliation (base 100). . . | 14 |
| 2.6  | Répartition de l'âge du contrat suivant l'année d'étude et la prime. . . . .        | 15 |
| 2.7  | Âge contrat vs Résiliation . . . . .                                                | 15 |
| 2.8  | Répartition de l'indice suivant l'année d'étude et la prime. . . . .                | 16 |
| 2.9  | Indice vs Résiliation . . . . .                                                     | 16 |
| 2.10 | Répartition du nombre de contrats selon le distributeur et par année. . . . .       | 17 |
| 2.11 | Répartition de la prime selon le distributeur et par année. . . . .                 | 17 |
| 2.12 | Distributeur vs Résiliation . . . . .                                               | 17 |
| 2.13 | Répartition du nombre de contrats selon la région et par année. . . . .             | 18 |
| 2.14 | Répartition de la prime selon la région et par année. . . . .                       | 18 |
| 2.15 | Répartition de l'effet suivant l'année d'étude et la prime. . . . .                 | 19 |
| 2.16 | Effet vs Résiliation . . . . .                                                      | 19 |
| 2.17 | Répartition du volume de sinistralité suivant l'année d'étude. . . . .              | 19 |
| 2.18 | Répartition de la sinistralité suivant l'année d'étude et la prime. . . . .         | 19 |
| 2.19 | Sinistralité vs Résiliation . . . . .                                               | 20 |
| 2.20 | Répartition de la prime selon le taux de majoration RC et par année. . . . .        | 20 |
| 2.21 | Taux majoration RC vs Résiliation . . . . .                                         | 20 |
| 2.22 | Répartition de la prime selon le taux de majoration global et par année. . . . .    | 21 |
| 2.23 | Taux majoration multi-équipé vs Résiliation . . . . .                               | 21 |
| 2.24 | Taux majoration RC vs Taux de majoration global. . . . .                            | 22 |
| 2.25 | Diff majoration vs Résiliation. . . . .                                             | 22 |
| 2.26 | Répartition du nombre de branches suivant l'année d'étude et la prime. . . . .      | 23 |
| 2.27 | Nombre de branches vs résiliation. . . . .                                          | 23 |
| 2.28 | Répartition de l'évolution du multi-équipement suivant l'année d'étude et la prime. | 24 |
| 2.29 | Évolution multi-équipement vs résiliation. . . . .                                  | 24 |
| 2.30 | Répartition des branches du multi-équipement suivant l'année d'étude et la prime.   | 25 |

|      |                                                                                                            |    |
|------|------------------------------------------------------------------------------------------------------------|----|
| 2.31 | Branches multi-équipement vs résiliation. . . . .                                                          | 25 |
| 2.32 | Matrice de corrélation entre les variables catégorielles : V de Cramer. . . . .                            | 26 |
| 2.33 | Matrice de corrélation entre les variables continues : Pearson. . . . .                                    | 28 |
| 2.34 | Répartition des valeurs manquantes au sein des variables. . . . .                                          | 30 |
| 3.1  | Algorithme CART. . . . .                                                                                   | 38 |
| 3.2  | Algorithme Random Forest. . . . .                                                                          | 40 |
| 3.3  | Détermination du cutt-off optimal. . . . .                                                                 | 45 |
| 3.4  | Courbe ROC. . . . .                                                                                        | 45 |
| 3.5  | Distribution résiliation dans l'échantillon train. . . . .                                                 | 46 |
| 3.6  | Distribution résiliation dans l'échantillon test. . . . .                                                  | 46 |
| 3.7  | Modèle de résiliation mono+multi : variables explicatives. . . . .                                         | 47 |
| 3.8  | Modèle de résiliation multi : variables explicatives. . . . .                                              | 47 |
| 3.9  | Modèle résiliation : Multi-équipement. . . . .                                                             | 49 |
| 3.10 | Modèle résiliation : Taux de majoration RC. . . . .                                                        | 50 |
| 3.11 | Modèle résiliation : Sinistralité 3A. . . . .                                                              | 50 |
| 3.12 | Modèle résiliation : Âge contrat. . . . .                                                                  | 51 |
| 3.13 | Modèle résiliation : Indice. . . . .                                                                       | 52 |
| 3.14 | Modèle résiliation : Distributeur. . . . .                                                                 | 53 |
| 3.15 | Modèle résiliation : Évolution multi-équipement. . . . .                                                   | 54 |
| 3.16 | Modèle résiliation : Différence de majoration. . . . .                                                     | 55 |
| 3.17 | Modèle résiliation : Nombre de branches. . . . .                                                           | 55 |
| 3.18 | Modèle résiliation : Diff_majo vs Taux majoration RC. . . . .                                              | 56 |
| 3.19 | Seuil optimal de résiliation. . . . .                                                                      | 57 |
| 3.20 | CART sur la prédiction des résiliations. . . . .                                                           | 60 |
| 3.21 | Importance globale des variables en utilisant les valeurs de Shap (moyenne va-<br>leurs absolues). . . . . | 61 |
| 3.22 | Importance globale des variables en utilisant les valeurs de Shap (impact sur<br>l'output). . . . .        | 61 |
| 4.1  | 1ère Approche : Distance de Levenshtein entre les résiliés et leurs jumeaux non<br>résiliés. . . . .       | 70 |
| 4.2  | 1ère Approche : Nuage de points entre majoration des résiliés et jumeaux non<br>résiliés. . . . .          | 70 |
| 4.3  | 1ère Approche : Comparaison majoration des résiliés et jumeaux non résiliés. . . . .                       | 70 |
| 4.4  | 2ème Approche : Distance de Levenshtein entre les résiliés et leurs jumeaux non<br>résiliés. . . . .       | 72 |
| 4.5  | Taux de majoration des résiliés vs jumeaux non résiliés. . . . .                                           | 72 |
| 4.6  | 2ème Approche : Nuage de points entre majoration des résiliés et jumeaux non<br>résiliés. . . . .          | 72 |



|      |                                                                                                                                                                                                                                         |     |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.7  | 2ème Approche : Comparaison majoration des résiliés et jumeaux non résiliés. .                                                                                                                                                          | 72  |
| 4.8  | 2ème Approche : Nuage de points entre majoration des résiliés et jumeaux non résiliés (Zoom sur les clients de distance $<1.67$ ie le caractère qui différencie n'est pas l'âge du contrat, le distributeur ou la grille majo). . . . . | 73  |
| 4.9  | 2ème Approche : Comparaison majoration des résiliés et jumeaux non résiliés (Zoom sur les clients de distance $<1.67$ ie le caractère qui différencie n'est pas l'âge du contrat, le distributeur ou la grille majo). . . . .           | 73  |
| 4.10 | Diff-majoration du client résilié vs majoration des jumeaux (base 100). . . . .                                                                                                                                                         | 75  |
| 4.11 | Diff-majoration du client résilié vs majoration des jumeaux. . . . .                                                                                                                                                                    | 75  |
| 4.12 | Évolution multi-équipement du client résilié vs majoration des jumeaux (base 100). . . . .                                                                                                                                              | 76  |
| 4.13 | Évolution multi-équipement n du client résilié vs majoration des jumeaux. . . .                                                                                                                                                         | 76  |
| A.1  | Caractérisation des prédictions : Évolution du multi-équipement. . . . .                                                                                                                                                                | 88  |
| A.2  | Caractérisation des prédictions : Différence de majoration. . . . .                                                                                                                                                                     | 88  |
| A.3  | Caractérisation des prédictions : Indice client. . . . .                                                                                                                                                                                | 88  |
| A.4  | Caractérisation des prédictions : nombre de branches. . . . .                                                                                                                                                                           | 88  |
| A.5  | Caractérisation des prédictions : distributeur. . . . .                                                                                                                                                                                 | 88  |
| A.6  | Caractérisation des prédictions : Sinistralité des 3 dernières années. . . . .                                                                                                                                                          | 88  |
| 4.1  | Distance de Levenshtein entre les résiliés et leurs jumeaux non résiliés. . . . .                                                                                                                                                       | 95  |
| 4.2  | Taux de majoration des résiliés vs jumeaux non résiliés. . . . .                                                                                                                                                                        | 95  |
| 4.3  | Levenshtein distance between the terminated and their unterminated twins. . .                                                                                                                                                           | 101 |
| 4.4  | Terminated vs. unterminated twin markup rate. . . . .                                                                                                                                                                                   | 101 |

# Liste des tableaux

|     |                                                                                                                                                |    |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | Segmentation de la branche RC. . . . .                                                                                                         | 7  |
| 2.1 | Aperçu général de la base de données. . . . .                                                                                                  | 13 |
| 2.2 | Taux majoration RC vs Taux de majoration globale. . . . .                                                                                      | 22 |
| 2.3 | Liste des variables choisies pour la modélisation. . . . .                                                                                     | 31 |
| 3.1 | Matrice de confusion. . . . .                                                                                                                  | 43 |
| 3.2 | Résultats du modèle de résiliation des clients RC. . . . .                                                                                     | 46 |
| 3.3 | Résultats du modèle de résiliation des clients RC. . . . .                                                                                     | 48 |
| 3.4 | Matrice de confusion en pourcentage du nombre de contrats. . . . .                                                                             | 57 |
| 3.5 | Matrice de confusion en pourcentage de la prime. . . . .                                                                                       | 58 |
| 3.6 | Résultats du modèle de résiliation : CART et Random forest. . . . .                                                                            | 62 |
| 3.7 | Matrice de confusion modèle CART. . . . .                                                                                                      | 62 |
| 3.8 | Matrice de confusion modèle Random Forest. . . . .                                                                                             | 62 |
| 4.1 | Comparaison résiliés vs jumeaux non résiliés (Taux de majoration   primes en €<br>  montant de majoration en €). . . . .                       | 74 |
| 4.2 | Différence de majoration : majoration des clients résiliés (Taux de majoration  <br>primes en €   montant de majoration en €). . . . .         | 75 |
| 4.3 | Différence de majoration : majoration des jumeaux non résiliés (Taux de ma-<br>joration   primes en €   montant de majoration en €). . . . .   | 75 |
| 4.4 | Évolution multi-équipement : majoration des clients résiliés (Taux de majoration<br>  primes en €   montant de majoration en €). . . . .       | 76 |
| 4.5 | Évolution multi-équipement : majoration des jumeaux non résiliés (Taux de ma-<br>joration   primes en €   montant de majoration en €). . . . . | 76 |
| A.1 | Liste des variables de la base finale construite. . . . .                                                                                      | 84 |
| A.2 | Liste des variables créées. . . . .                                                                                                            | 85 |
| A.3 | Test du khi-deux d'indépendance entre la résiliation et les autres<br>variables du modèle . . . . .                                            | 87 |
| 4.1 | Matrice de confusion modèle logistique. . . . .                                                                                                | 93 |
| 4.2 | Confusion matrix logistic model. . . . .                                                                                                       | 99 |