Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires

le 27/06/2022

Par : Aicha KOROVAEV
Titre : Modélisation de la durée de vie des contrats d'assurance habitation - application à l'optimisation de la rentabilité

Confidentialité : Non ☐ ☑ Oui (Durée : ☐ 1 an ☐ ☑ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : Allianz IARD
Signature :

Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :

Quentin Guibert

Directeur de Mémoire en entreprise :
Nom : Yann QUELEN
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

Résumé

Le modèle de distribution des compagnies d'assurance reposait traditionnellement sur les agents et les courtiers. Mais depuis quelques années, un nouveau moyen de distribution s'est développé : le canal direct. Celui-ci attire de plus en plus de clients puisqu'il permet de trouver et souscrire une offre d'assurance n'importe où, à n'importe quel moment et de bénéficier de tarifs plus attractifs. Cependant, plusieurs assurés préfèrent toujours un interlocuteur et sont donc plus rassurés par le modèle de distribution classique. Ce mode de distribution a néanmoins une part de marché limitée par celle des mutuelles et des bancassureurs. Certaines compagnies ont donc décidé d'adopter une stratégie «multi-accès» offrant aux assurés ou prospects plusieurs canaux d'interaction basés non seulement sur des réseaux de distribution physiques mais aussi sur le canal direct c'est-à-dire sur Internet et les plateformes téléphoniques.

Pour optimiser la rentabilité de la stratégie, il est donc important de pouvoir identifier les contrats les plus rentables à long terme suivant chaque moyen de distribution et d'encourager la souscription sur ces contrats. En partant du fait qu'il est plus bénéfique de garder un client plutôt que d'attirer un nouveau, les profils les plus rentables peuvent être ceux qui durent le plus. Garder un contrat permet d'économiser en coût d'acquisition. De même, plus la durée de vie du contrat dans le portefeuille augmente, plus le ratio sinistres sur primes (S/C) baisse globalement grâce à la revalorisation de la prime. Cependant, à durée de vie du contrat égale, les profils les plus rentables sont ceux ayant un meilleur ratio sinistre sur prime. De plus, même si nous voulons encourager la souscription sur les profils les plus loyaux, il y a des contraintes comprenant le coût d'acquision client qui pourraient nous empêcher de choisir certains profils. Ainsi pour optimiser la rentabilité, il faut aller plus loin que déterminer simplement les profils qui durent le plus. Nous allons alors essayer de trouver, dans chaque parcours, les profils ayant le meilleur S/C sur leur durée de vie qu'il faudrait idéalement avoir en affaires nouvelles afin de minimiser cet indicateur de rentabilité sur plusieurs années suivant plusieurs contraintes prenant en compte les coûts.

Le sujet de ce mémoire consiste donc à déterminer la durée de vie des contrats d'assurance habitation en modélisant les taux de résiliation sur plusieurs périodes afin d'optimiser la rentabilité de la stratégie multi-accès. Et pour modéliser ces taux de résiliation pendant plusieurs périodes de durée égale, seront utilisés 3 modèles de prédiction *machine learning* sur un portefeuille de contrats d'assurance habitation : un GLM (*Generalized Linear Model*), un *Random Forest* et un XGBoost afin de choisir le meilleur. Ces modèles aideront à prédire, pour chaque période donnée, si un assuré va résilier son contrat ou non. Nous pourrons donc à la fin de chaque période, voir quelles sont les variables qui impactent le plus les taux de résiliation et notamment déterminer l'impact des parcours. En combinant les modèles, il sera possible d'estimer les durées de vie. Ces durées de vie vont non seulement permettre de pouvoir identifier les profils les plus loyaux mais aussi d'établir une approche d'optimisation de la rentabilité sur 4 ans basée sur la sélection des profils à l'affaire nouvelle dans chaque parcours.

Mots-clés : *Assurance habitation*, *Durée de vie*, *Rétention*, *Machine Learning*, *Optimisation sous contraintes*.

Abstract

The distribution model of insurance companies has traditionally relied on agents and brokers. But in recent years, a new means of distribution has developed : the direct channel. It is attracting more and more customers since it makes it possible to find and subscribe to an insurance offer anywhere, anytime and benefit from more attractive prices. But many insureds still prefer an interlocutor and are therefore more reassured by the classic distribution model. However, this mode of distribution has a market share limited by that of mutuals and bancassurance companies. Some companies have therefore decided to adopt a "multi-access" strategy offering clients several interaction channels based not only on classical distribution mode but also on the direct channel, that is to say on the Internet and telephone platforms.

To optimize the profitability of the strategy, it is therefore important to be able to identify the most profitable customers in the long term according to each means of distribution and to encourage subscription on these contracts. Starting from the fact that it is more beneficial to keep a customer rather than attracting a new one, the most profitable profiles can be the ones that last the longest. Keeping a contract helps in saving on acquisition costs. Also, the longer the duration of the contract in the portfolio increases, the more the claims-to-premium ratio (S/C) decreases overall thanks to the revaluation of the premium. However, with similar duration, the most profitable profiles are those with a better claim-to-premium ratio. Also, even if we want to encourage the subscription on the profiles that last the longest, there are constraints including the cost of customer acquisition and others that could prevent us from choosing certain profiles. So to optimize profitability, we must go further than simply determining loyal profiles. We will then try to find, in each course, the profiles with the best S/C over their duration in order to minimize this indicator of profitability over several years following several constraints taking costs into account.

The subject of this thesis is therefore to determine the duration of home insurance contracts by modeling churn rates over several periods in order to optimize the profitability of the multi-access strategy. And to model these churn rates for several periods of equal duration, we will use 3 machine learning prediction models on a portfolio of home insurance contracts : a GLM (Generalized Linear Model), a Random Forest and an XGBoost in order to choose the best. These models will help us predict, for each given period, whether an insured will terminate their contract or not. At the end of each period, we will therefore be able to see which variables have the most impact on churn rates and in particular determine the impact of the courses. By combining the churn rate models, we will be able to predict the duration. These lifetimes will not only allow us to be able to identify the profiles that last the longest in our portfolio but also to establish an approach to optimize profitability over 4 years based on the profiles selection.

Keywords : *Home Insurance*, *Duration*, *Churn rate*, *Machine Learning*, *Constrained Optimization*.

Note de Synthèse

Notre objectif, tout au long de ce mémoire, est de modéliser la durée de vie des contrats d'assurance habitation afin d'établir une stratégie d'optimisation de la rentabilité (sur plusieurs années) basée sur la sélection des profils à l'affaire nouvelle de chaque parcours client de la stratégie multi-accès.

Avec l'émergence de l'utilisation d'internet comme mode de distribution des produits d'assurance, certaines compagnies d'assurance ont décidé d'adopter une stratégie multi-accès afin de profiter des moyens de distribution traditionnelle mais aussi du canal direct. Un assuré du canal direct utilise internet et les plateformes téléphoniques comme moyen de contact avec l'assureur.
La stratégie multi-accès d'Allianz se décompose en 3 étapes :
— Une étape *Sourcing* : elle permet de différencier les contrats dont le devis a été fait sur internet (D) des contrats dont le devis a été fait sur le canal traditionnel (T) c'est-à-dire en se rendant en agence.
— Une étape Souscription : les prospects ou clients ont la possibilité de souscrire en agence (A), en contactant la plateforme téléphonique (M) d'Allianz ou en souscrivant sur le site Web (W) d'Allianz.
— Une étape Gestion de contrat : les clients peuvent faire gérer leur contrat en agence (A) ou par la plateforme téléphonique d'Allianz (M).
La stratégie multi-accès est ainsi constituée de 8 parcours obtenus en croisant ces étapes : 2 parcours agence (contrats souscrits en agence), 4 parcours plateforme et 2 parcours Web.

Les parcours client de la stratégie multi-accès d'Allianz ont un coût d'acquisition client différent. Ce coût peut être plus élevé que la prime moyenne pour certains parcours. Ainsi, pour l'amortir, il est nécessaire de garder le client le plus longtemps possible en portefeuille. De même, le ratio Sinistres à Primes (S/C) qui est un important indicateur de rentabilité baisse globalement lorsque la durée de vie des contrats augmente grâce à la revalorisation de la prime. Tout cela montre qu'il est plus rentable de garder un client le plus longtemps possible dans le portefeuille.
Pour parvenir à optimiser la rentabilité, nous souhaitons donc identifier les profils qui durent le plus longtemps dans le portefeuille afin d'encourager la souscription sur ces profils. Dans le cas où le parcours aurait un impact sur la durée de vie, les clients pourraient être orientés vers le parcours où ils dureraient le plus afin d'optimiser la rentabilité.
Par ailleurs, nous savons aussi que lorsque les durées de vie sont égales, les assurés ayant un meilleur S/C sont plus rentables. Pour établir une approche plus solide d'optimisation de la rentabilité à long terme de la stratégie multi-accès, il faudrait identifier dans chaque parcours les profils les plus rentables (ayant le meilleur S/C sur leur durée de vie) à avoir en affaire nouvelle suivant un coût d'acquisition total donné et plusieurs autres contraintes. Cette optimisation va de ce fait permettre de connaître les parcours qu'il faudrait développer et lesquels pourraient être éliminés.

Données

Plusieurs bases de données d'Allianz pour les contrats d'assurance habitation sont utilisées afin de créer la base de modélisation. Ces bases ont permis d'obtenir 1 244 324 affaires nouvelles en assurance habitation souscrites entre le 1er Janvier 2015 et le 31 Décembre 2020 et d'observer leur éventuelle

résiliation du 1er Janvier 2015 au 31 Décembre 2020. Les valeurs manquantes et affaires nouvelles non retrouvées dans certaines bases d'Allianz utilisées ont entrainé la suppression de 4% du nombre total d'affaires nouvelles initialement récupérées. Le tableau 1 suivant présente les principales variables explicatives constituant la base, les variables zonier se terminant par `FREQ` désignent le zonier technique et `COM` le zonier commercial.

Variables contrat	**Variables client**	**Variables zonier**
— `parcours` — `PrEnt` (prime) — `qualJur` (qualité juridique) — `typeHab` (type d'habitation) — `natRes` (nature de la résidence) — `covid` (résiliations enregistrées pendant la Covid-19)	— `csp` (catégorie socio-professionnelle) — `AgeCat` (tranche d'âge) — `situationFam` (situation familiale) — `nb_contrats` (nombre de contrats)	— `BDG_FREQ` (zonier bris de glace) — `DDE_FREQ` (zonier dégâts des eaux) — `VOL_FREQ` (zonier vol) — `INC_com` (zonier incendie)

TABLE 1 – Principales variables explicatives constituant la base de donnée créée

Faire une analyse descriptive de la base créée a permis de voir que les parcours ont des profils différents. Par exemple, dans les parcours Web se trouvent peu de propriétaires (moins de 7%) alors que dans les parcours agence il y en a plus de 27%. L'analyse descriptive montre aussi que les taux de résiliation par parcours sont différents ce qui pourrait faire croire que les parcours ont un impact sur les résiliations et donc la durée de vie du contrat.

Un problème pour la modélisation de la durée de vie est tout de suite relevé : les parcours Web n'ont qu'une année entière d'ancienneté alors que certains parcours en ont 5 ans. C'est l'une des raisons qui nous a poussés à opter pour la modélisation des durées de vie en passant par les taux de résiliation par période. Ainsi pour une période donnée, il serait possible de déterminer l'impact des parcours sur l'acte de résiliation et de le projeter afin de modéliser les durées de vie de tous les parcours sur le même horizon. Les modèles de durées de vie comme le modèle de Cox ou même le *Random Survival Forest* auraient pu être utilisés. Cependant, la limite du modèle de Cox est l'hypothèse des hazards proportionnels non vérifiée par certaines variables. Le *Random Survival Forest* pallie cet inconvénient puisqu'il n'a pas besoin que cette hypothèse soit vérifiée mais il est plus difficile à interpréter.
Notre **plan d'action** est décrit dans la figure 1 suivante.

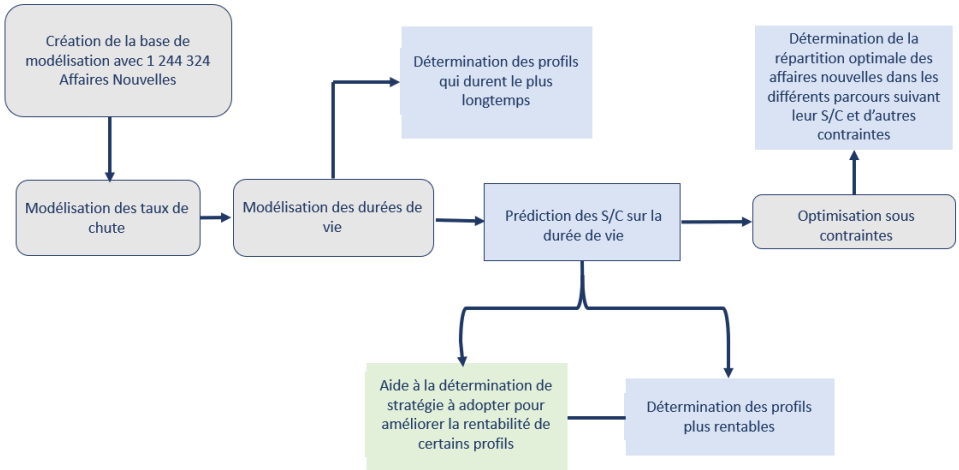


FIGURE 1 – Plan d'action

Il se décompose principalement en trois étapes après la création de la base de modélisation. Elles sont résumées dans la figure 1 et expliquées plus en détail par la suite. Les principales étapes sont coloriées en gris tandis que les résultats sont identifiés en bleu (les résultats possibles en vert).

Modélisation des taux de résiliation pendant plusieurs périodes

Nous disposons de 1 244 324 affaires nouvelles souscrites entre 2015 et 2020, soit 5 ans d'ancienneté mais la modélisation a été arrêtée à 4 ans, nous verrons un peu plus tard pourquoi. Nous souhaitons donc savoir sur 4 ans, **quels contrats vont survivre et au bout de combien de temps certains contrats vont être résiliés**. N'oublions pas non plus le fait que certains parcours n'ont qu'une année d'ancienneté nous amenant alors à trouver des pistes pour modéliser leur durée sur plus d'une année. Il est donc nécessaire de modéliser les taux de résiliation **par période**.

La variable à expliquer est Y=0 (survie) et Y=1 (résiliation). Une période de modélisation de 1 an a été choisie. Une période plus petite (6 mois) a été testée mais les performances de la période de modélisation de 1 an sont plus satisfaisantes. Les taux de résiliation par 6 mois semblent être trop faibles à modéliser puisque le déséquilibre des classes est trop grand. Le déséquilibre pourrait entraîner que seule l'appartenance à la classe majoritaire va être bien prédite.
Pour prédire les durées de vie, il faut donc modéliser les taux de résiliation pendant la première année puis les taux de résiliation pendant la période N sachant que les contrats ont survécus jusqu'à la période N-1, $N > 2$**.**
Puisque nous souhaitons prédire la durée de vie du contrat à l'affaire nouvelle, les variables vues à l'affaire nouvelle vont être utilisées quelle que soit la période de modélisation même si plusieurs variables pourraient changer à la suite d'un avenant (changement du niveau du capital par exemple, majoration de la prime).

Avant de commencer la modélisation, plusieurs problèmes ont été repérés en plus de celui lié à l'ancienneté des parcours. Ils incluent **un volume des affaires nouvelles faibles pour certains parcours**. Cela pourrait entrainer que les résiliations de ces parcours ne vont pas être identifiées par le modèle et donc être mal prédites vu que ces parcours sont très peu représentés. Pour éviter ce problème, les parcours ayant des caractéristiques communes et des taux de résiliation proches ont été regroupés.
Il y a aussi **le déséquilibre des classes** qui peut poser problème. En effet, les taux de résiliation ne dépassent pas 18%, la classe minoritaire est donc la classe résiliation. Pour équilibrer la base, des méthodes de ré-échantillonnage comme le SMOTE (*Synthetic Minority Over-Sampling Technique*) et le *Random Under Sampling* sont testés après avoir utilisé la base d'entrainement sans ré-échantillonnage.

Pour chaque période, trois modèles sont testés : un Modèle Linéaire Généralisé (GLM), un *Random Forest* (RF) et un *eXtreme Gradient Boosting* (XGBoost). Pour choisir le modèle à utiliser pour le calcul des durées de vie, une règle de décision a été définie. Celle-ci consiste à utiliser des métriques pertinentes comme l'AUC, l'AUCPr, le rappel, la précision, le logloss et de faire une :
— comparaison des valeurs des métriques des différents modèles,
— comparaison des valeurs des métriques des différents modèles suivant les parcours pour vérifier que les résiliations des parcours avec peu de données sont prédits comme celles des autres parcours,
— comparaison des taux de résiliation prédits par les différents modèles avec ceux qui sont réellement observés dans la base test suivant la variable discriminante pour tous les modèles (la qualité juridique).
Vu que nous voulons calculer les durées de vie moyennes, il faut donc que les durées de vie prédites par notre modèle soit cohérentes avec les durées réelles. Pour cela, il faut que les taux de résiliation de chaque année soient les plus proches possible des taux de résiliation réels.

C'est ainsi que **nous avons globalement choisi le XGBoost vu qu'il répond le plus à notre règle de décision** même si les performances des modèles ne sont pas très différentes. Les performances obtenues sont juste acceptables. L'AUC tourne autour de 70% pour chaque période. Le rappel tourne autour de 60%. La précision et l'AUCPr se dégradent de période en période. La précision passe de 34% pendant la première année à 16% à la quatrième. Ce qui nous a poussés à arrêter la modélisation à 4 ans.

Impact des parcours et autres variables
L'analyse de l'importance des variables a permis de voir que les parcours ont très peu d'impact sur la résiliation mais aussi que les variables les plus discriminantes sont entre autres la qualité juridique, le nombre de contrats du client, l'âge du client et les zoniers. L'importance des variables n'est pas la même pour tous les modèles. Globalement, les mêmes variables sont retrouvées parmi les variables les plus importantes au fil des périodes mais leur importance diminue. Quel que soit le modèle testé, le parcours est très peu important.
Après avoir choisi le XGBoost pour la modélisation des durées de vie, des outils d'interprétation de modèles boites noires comme le *Partial Dependence Plot* (PDP), l'*Individual Conditional Expectation Plot* (ICE) et les SHAP *values* ont été utilisés pour renforcer l'observation faite quant à l'importance des variables. Cela a permis d'affirmer que les parcours ont très peu d'influence sur les durées de vie. Ces outils ont aussi aidé à confirmer que l'année de souscription n'a pas d'impact sur la résiliation à moins qu'il n'y ait eu une perturbation comme la Covid-19 au cours de l'année. Les durées de vie sont stables dans le temps (hors période de Covid-19). La Covid-19 a fait baisser les résiliations à cause des confinements et des couvre-feux. Par ailleurs, les propriétaires ont moins de chance de résilier que les locataires tandis que les assurés ayant plusieurs contrats ont moins de chance de résilier que ceux qui n'en ont qu'un. Une durée de vie plus élevée pour un parcours par rapport à un autre pourrait par exemple s'expliquer par le fait qu'il y ait plus de de propriétaires dans ce parcours.

Modélisation de la durée de vie

Une fois les taux de résiliations modélisés, il est possible de prédire à la fin de chaque période quel contrat va être résilié et quel contrat va survivre.

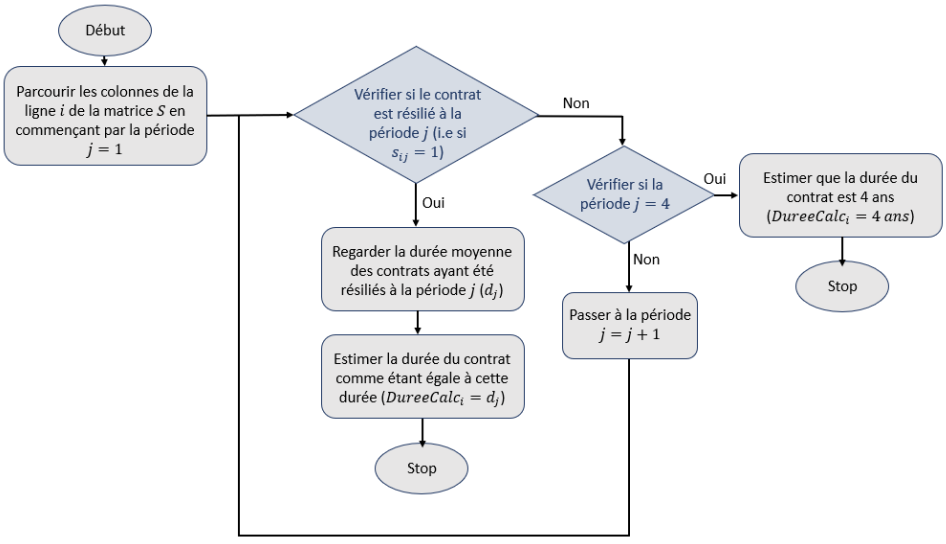


FIGURE 2 – Schéma résumant l'algorithme d'estimation de la durée de vie du contrat i notée $DureeCalc_i$ à partir de la matrice de survie S et du vecteur de durées de vie moyennes D

Si n est le nombre total de contrats, la matrice $\boldsymbol{S}$ (de taille $n \times 4$) est une matrice dont chaque ligne i représente un contrat et chaque colonne j contient la prédiction de la résiliation ou non du contrat pour la $j-ème$ période (prédiction obtenue avec le modèle de taux de résiliation de la période),

$$S = (s_{ij}) \text{ avec } 1 \leq i \leq n \text{ et } 1 \leq j \leq 4, \text{ et } s_{ij} \in \{1(\text{résiliation}), 0(\text{survie})\}. \quad (1)$$

Le vecteur de durées de vie moyennes de taille 4 et noté $\boldsymbol{D}= (d_1, d_2, d_3, d_4)$ présente la durée moyenne réelle des contrats résiliés pendant la j-ème période, j allant de 1 à 4. Pour estimer les durées de vie, un algorithme prenant en entrée la matrice S et le vecteur D a été utilisé. Pour le contrat i, l'algorithme est décrit dans la figure 2 ci-dessus.

Ainsi, en l'utilisant sur les contrats de la base dont la durée réelle sur 4 ans n'est pas connue, le processus va aboutir sur une base avec des contrats souscrits entre 2015 et 2020 avec leur durée de vie moyenne sur 4 ans. Les contrats ont donc tous été observés sur la même durée.
Le tableau 2 suivant présente les taux de résiliation prédits pendant la première, la deuxième, la troisième et la quatrième année ainsi que les durées de vie moyennes prédites suivant le parcours.

	Taux de résiliation par année				
Parcours	1ère	2ème	3ème	4ème	Durée de vie moyenne
agence digitale (ADA)	20%	19%	19%	7%	3,08 ans
agence traditionnelle (ATA)	19%	19%	17%	7%	3,12 ans
plateforme traditionnelle (MTM)	21%	20%	19%	8%	3,00 ans
full Web (WDM)	30%	26%	25%	14%	2,65 ans

TABLE 2 – Taux de résiliation prédits pendant la première, la deuxième, la troisième et la quatrième année ainsi que les durées de vie moyennes prédites suivant le parcours.

Les parcours agence ont les taux les plus faibles et les durées de vie les plus élevées puisque s'y trouvent plus de profils susceptibles de durer longtemps (les parcours agence ont plus de 27% de propriétaires tandis que les parcours Web en ont moins de 7%). Sur 4 ans, la durée de vie des propriétaires est d'environ 3,5 ans et est supérieure à celle des locataires qui est égale à 2,6 ans soit une différence de presqu'une année. Il est possible de lire cette information dans le tableau 3 suivant, où sont présentées les durées de vie suivant ses variables les plus discriminantes. Pour chaque variable, les durées sont décroissantes par modalité.

Variable	**Modalités**	**Durée de vie moyenne**
Qualité juridique	Propriétaire	3,5 ans
	Locataire	2,6 ans
Age du client	Seniors	3,3 ans
	Adultes	3 ans
	Jeunes	2,6 ans
Nombre de contrats	Plusieurs contrats	3,2 ans
	1 contrat	2,8 ans

TABLE 3 – Durées de vie moyennes (sur 4 ans) suivant 3 variables discriminantes

Les durées de vie estimées vont finalement permettre d'optimiser la rentabilité.

Optimisation de la rentabilité

Profils les plus loyaux

Pour déterminer les caractéristiques des profils les plus loyaux, les contrats sont regroupés suivant les variables discriminantes pour la durée, ce qui résulte sur 354 285 profils de durée de vie. Ainsi, afin d'identifier ce qui définit un profil qui dure longtemps dans le portefeuille, un arbre de décision va être utilisé. L'arbre de décision permet de déterminer à quelle classe de durée de vie appartient un contrat suivant plusieurs critères.
En interprétant l'arbre obtenu, nous remarquons que les propriétaires, les locataires seniors ou locataires adultes multi-détenteurs sont généralement les profils les plus loyaux (durent plus de 3 ans). Les propriétaires ont 87% de chance de durer plus de 3 ans. Nous avons déjà vu que le parcours n'a pas d'influence sur la résiliation. Il n'y a plus besoin d'identifier les profils qui auraient duré plus longtemps s'ils étaient dans un autre parcours.

Profils les plus rentables

Un indicateur de rentabilité a ensuite été pris en compte. Il s'agit du ratio Sinistres à Primes (S/C). Cela va permettre d'identifier les profils les plus rentables ainsi que les profils moins rentables suivant leur S/C et leur durée de vie. De ce fait, il sera possible de déterminer les profils qui pourraient continuer à être des affaires nouvelles, les profils à éviter dans chaque parcours ainsi que les profils dont nous pourrions essayer d'améliorer la rentabilité.
Les profils les plus rentables sont dans tous les parcours des locataires d'appartement loyaux. Les profils à éviter dans les parcours Web sont les propriétaires. Pour les profils ayant un S/C élevé mais une durée de vie faible comme les locataires d'appartement pas loyaux, il pourrait être envisagé de leur offrir des mois gratuits vu que leur S/C laisse une marge suffisamment importante pour baisser leur prix. Il faudrait faire une étude beaucoup plus poussée pour savoir exactement quelle stratégie adopter pour améliorer la rentabilité.

Optimisation sous contraintes

Finalement, l'optimisation sous contraintes va permettre d'identifier les profils qui pourraient être choisis dans chaque parcours afin de maximiser la rentabilité tout en satisfaisant plusieurs contraintes. Les 4 contraintes comprennent les coûts (le coût d'acquisition notamment), le nombre total de contrats et la somme totale des primes émises. Les coûts sont supposées fixes par parcours. Chaque parcours est divisé en 2 ou 3 *clusters* (c'est-à-dire groupes avec des S/C homogènes) soit un *cluster* avec des profils plus rentables et un *cluster* avec des profils moins rentables.
Les inconnus du problème sont donc les nombres de contrats à avoir dans chacun des 16 *clusters* afin de minimiser le S/C sur 4 ans suivant les contraintes. La fonction objective est alors le rapport sinistres à primes des profils sur 4 ans. Elle n'est pas convexe, ce qui provoque qu'il est possible de tomber sur un optimum local au lieu de l'optimum global recherché. Mais en utilisant des algorithmes d'optimisation globale comme l'**évolution différentielle** et le ***basin-hopping***, l'optimum global semble avoir été trouvé.
Nous avons donc pu déterminer les profils qu'il faudrait avoir en affaires nouvelles dans chaque parcours afin d'obtenir un S/C minimal sur 4 ans égal à 53,67% (10% de moins que le S/C sur 4 ans actuel) tout en respectant les contraintes. Par la même occasion, sont identifiés les parcours qui devraient être plus développés (le parcours Web WDM notamment) et les parcours qui pourraient être éliminés (généralement des parcours plateforme).

Bilan et perspectives

Ce mémoire montre qu'il est possible de prédire les durées de vie moyennes en modélisant les taux de résiliation par période. Les modèles ont aussi permis de conclure que les parcours ont très peu d'impact sur la durée de vie, il n'y a donc plus besoin d'identifier les profils qui auraient durer plus longtemps s'ils étaient dans un autre parcours. Prédire les durées de vie a permis d'identifier que les profils les plus loyaux sont généralement des propriétaires, des locataires seniors ou des locataires adultes ayant plusieurs contrats.
L'optimisation sous contraintes a permis de voir quels profils il faudrait avoir en affaire nouvelle dans chaque parcours et qu'il est possible de déterminer les parcours de la stratégie multi-accès qui pourrait être éliminé. Il pourrait être intéressant de prendre des probabilités en compte afin de ne pas négliger le comportement du client. En effet, nous orientons les clients sans se préoccuper vraiment de leurs préférences.

En essayant d'améliorer la prédiction des durées de vie en passant par l'amélioration des performances des modèles de taux de résiliation (par l'introduction de certaines variables comme l'indice de compétivité, la fréquence des sinistres prédite) et en corrigeant les limites de la stratégie d'optimisation sous contraintes (comme les coûts fixes par parcours), nous pourrions tirer avantage de la stratégie d'optimisation avec beaucoup plus de certitudes. L'optimisation a aussi permis d'avoir une idée sur les profils les plus rentables suivant leur durée de vie et le ratio sinistres à prime ainsi qu'une idée sur comment améliorer les rentabilités. Ceci pourrait constituer une aide permettant d'élaborer une stratégie pour améliorer la rentabilité des profils cibles peu rentables.

Synthesis Note

Our objective, throughout this thesis, is to model the duration of home insurance contracts in order to establish a strategy for optimizing profitability (over several years) based on the selection of the new clients' profiles of each multi-access strategy's customer course.

With the emergence of the use of the Internet as a method of distributing insurance products, some insurance companies have decided to adopt a multi-access strategy in order to take advantage of traditional means of distribution but also of the direct channel. A direct channel insured uses the Internet and telephone platforms as a means of contact with the insurer.
Allianz's multi-access strategy can be broken down into 3 steps :
— A Sourcing step : it makes it possible to differentiate the contracts for which the estimate was made on the internet (D) from the contracts for which the estimate was made on the traditional way (T).
— A Subscription step : prospects or customers have the option of subscribing at an agency (A), by contacting the Allianz telephone platform (M) or by subscribing on the Allianz website (W).
— A contract management step : customers can have their contract managed in an agency (A) or through the Allianz telephone platform (M).
The multi-access strategy is therefore made up of 8 routes obtained by crossing these steps :
2 agency routes (contract taken out in agency), 4 platform routes and 2 web routes.

The customer courses of Allianz's multi-access strategy have a different customer acquisition cost. This cost may be higher than the average premium for some routes. Thus, to amortize it, it is necessary to keep the client as long as possible in the portfolio. Also, the Claims to Premiums (S / C) ratio, which is an important indicator of profitability, drops overall when the duration of the contracts increases thanks to the revaluation of the premium. All of this shows that it is more profitable to keep a client as long as possible in the portfolio.
To achieve optimal profitability, we therefore want to identify the profiles that last the longest in the portfolio in order to encourage subscription to these profiles. In case the route has an impact on the duration, customers could be directed to the route where they last the longest in order to optimize profitability.
On the other hand, we also know that when the duration is equal, policyholders with better S/C are more profitable. To establish a more solid approach to optimizing the long-term profitability of the multi-access strategy, it would be necessary to identify in each course the profiles (with the best S/C over their lifetime) to have in new business according to a total cost acquisition given and several other constraints. This optimization will therefore make it possible to know which courses must be developed and which could be eliminated.

Data

Several Allianz databases for home insurance contracts are used to create the modeling base. These bases made it possible to obtain 1 244 324 new business in home insurance taken out between January

1, 2015 and December 31, 2020 and to observe their possible churn from January 1, 2015 to December 31, 2020. Missing values and new business not found in some Allianz databases used resulted in the deletion of 4% of the total number of new business recovered. The following table 4 presents the main explanatory variables constituting the base, the zoning variables ending with `FREQ` designate the technical zoning and `COM` the commercial zoning.

Contract variables	**Customer variables**	**Zoning variables**
— `parcours` (path) — `PrEnt` (premium) — `qualJur` (legal quality) — `typeHab` (type of dwelling) — `natRes` (nature of residence) — `covid` (churns recorded during Covid-19)	— `csp` (socio-professional category) — `AgeCat` (age range) — `situationFam` (family status) — `nb_contracts` (number of contracts)	— `BDG_FREQ` (glass breakage zone) — `DDE_FREQ` (water damage zone) — `VOL_FREQ` (zone vol) — `INC_com` (fire zone)

TABLE 4 – Main explanatory variables constituting the created database

Carrying out a descriptive analysis of this database made it possible to see that the routes have different profiles. For example, web courses have few owners (less than 7%) while agency courses have over 27%. The descriptive analysis also shows that the churn rates per course are different, which could lead to believe that the courses have an impact on churn.

A problem for the duration modeling is immediately noted : Web courses are only a full year old while some courses are 5 years old. This is one of the reasons that leads us to opt for the duration modeling by passing through the churn rates per period. Thus for a given period, it would be possible to determine the impact of the courses and to project them. Survival models like the Cox model or even the Random Survival Forest could be used. However, the limit of the Cox model is the assumption of proportional hazards not verified by certain variables. The Random Survival Forest overcomes this drawback since it does not need this assumption to be verified but it is more difficult to interpret.
Our **action plan** is described in the following figure 3.

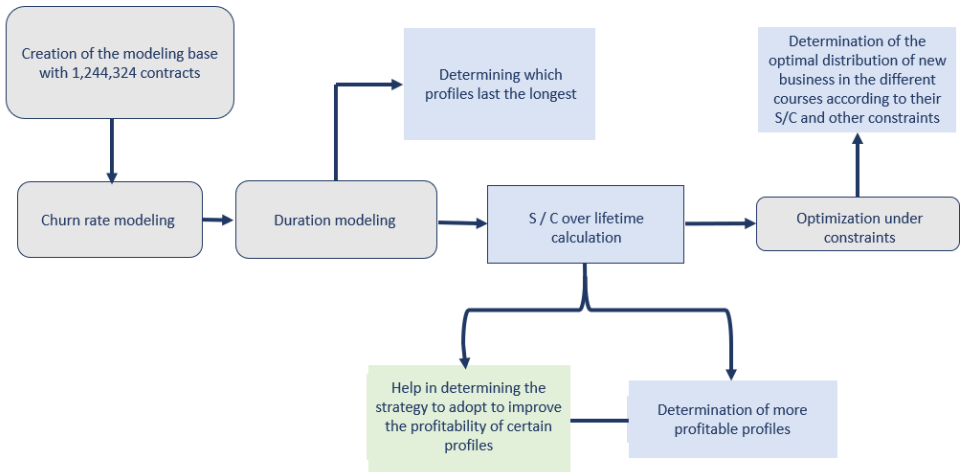


FIGURE 3 – Action plan

It is mainly broken down into four stages which are summarized in the figure above and explained in more detail below. The main steps are colored in gray while the results are identified in blue (possible results in green).

Modelling churn rates over multiple periods

We have 1,244,324 new business subscribed between 2015 and 2020, i.e. 5 years of seniority but the modeling was stopped at 4 years, we will see a little later why. We therefore want to know over 4 years, **which contracts will survive and after how long certain contracts will be terminated**. Let's not forget the fact that some courses are only one year old, leading us to find ways to model their duration over more than a year. It is therefore necessary to model the fall rates **per period**.

The variable to be explained is Y = 0 (no churn) and Y = 1 (churn). A modeling period of 1 year was chosen. A smaller period (6 months) has been tested but the performance of the 1 year modeling period is more satisfying. Churn rates per 6 months seem to be too low to model since the class imbalance is too large. The imbalance could lead to only belonging to the majority class will be to be well predicted.
To predict duration, it is therefore necessary to model the churn rates during the first year then the churn rates during period N knowing that the contracts have survived until period N-1, $N > 2$**.**
Since we want to predict the lifetime of the contract for the new business, the variables seen at the the new business will be used regardless of the modeling period, even if several variables could change following an amendment (change in capital level or premium increase for example).

Before starting the modeling, several problems were identified in addition to the one related to the age of the routes. They include **low new business volume for certain routes**. This could lead to the churns of certain routes not being identified by the model and therefore being poorly predicted since these routes are very poorly represented. To avoid this problem, routes with common characteristics and similar churn rates have been grouped together.
There is also **class imbalance** which can be a problem. Indeed, churn rates do not exceed 18%, the minority class is therefore the termination class. To balance the database, resampling methods such as SMOTE (*Synthetic Minority Over-Sampling Technique*) and *Random Under Sampling* are tested after using the training base without resampling.

For each period, we have tested three models : a Generalized Linear Model (GLM), a Random Forest (RF) and an eXtreme Gradient Boosting (XGBoost).

To choose the model to use for the calculation of duration, a decision rule has been defined. This consists of using relevant metrics like AUC, AUCPr, recall, precision, logloss and doing a :
— comparison of the values of the different models metrics,
— comparison of the values of the different models metrics according to the routes to check if the churns of the ones with little data are predicted as the churns of the other courses,
— comparison of the churn rates predicted by the different models with those actually observed in the test base according to the discriminant variable for all the models. Since we want to calculate average duration, it is therefore necessary that the lifetimes predicted by our model are consistent with the actual lifetimes. This requires that the churn rates for each year be as close as possible to the actual churn rates.
This is how we chose the XGBoost overall since it best meets our decision rule even if the performances of the models are not very different. The performances obtained are just acceptable. The AUC is around 70% for each period. The recall is around 60%. Precision and AUCPr degrade from period to period. Precision drops from 34% in Year 1 to 16% in Year 4, what pushed us to stop modeling at 4 years old.

Impact of the routes and other variables on the duration
The analysis of the importance of the variables made it possible to see that the paths have very

little impact on the termination but also that the most discriminating variables are among others the legal quality, the number of contracts of the customer, the age of the customer and the zoning. The importance of the variables is not the same for all the models. Overall, the same variables are found among the most important variables over time, but their importance decreases. Regardless of the model tested, the course is very unimportant.
After choosing XGBoost for duuration modeling, tools for interpreting black box models such as Partial Dependence Plot (PDP) and Individual Conditional Expectation Plot (ICE) and SHAP values to reinforce the observation made as to the importance of the variables. This made it possible to assert that routes have very little influence on lifespans. These tools also helped confirm that the year of subscription does not impact termination unless there was a disruption like Covid-19 during the year. The duration is stable over time (excluding the Covid-19 period). Covid-19 has reduced churn due to lock-downs and curfews. Furthermore, owners are less likely to cancel than tenants, while policyholders with several contracts are less likely to cancel than those with only one. A longer duration for one route compared to another could for example be explained by the fact that there are more owners in that route.

Duration Modeling

Once churn rates are modeled, it is possible to predict at the end of each period which contract will be churned and which contract will survive.

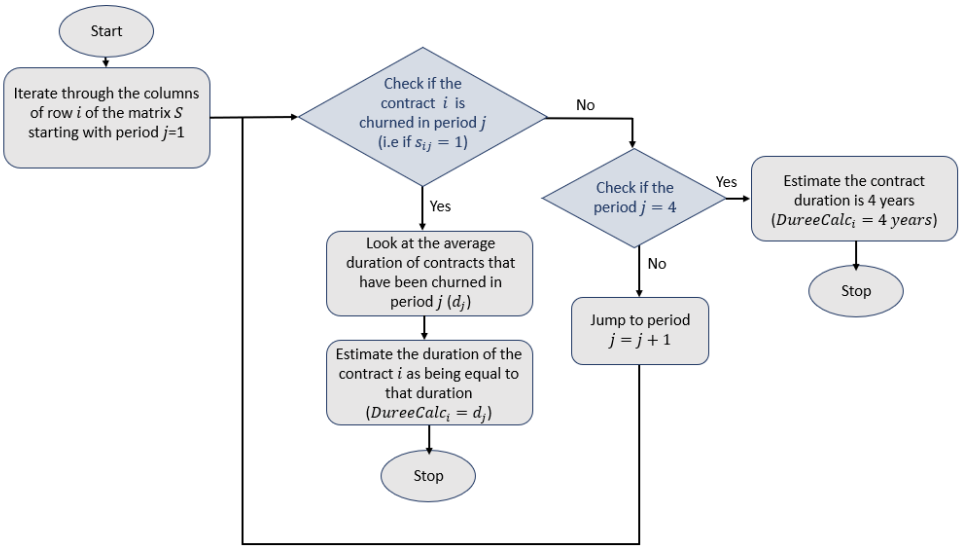


FIGURE 4 – Diagram summarizing the algorithm that estimates the duration of the contract i denoted $DureeCalc_i$ from the survival matrix S and the vector of average duration D

If n is the total number of contracts, the matrix $\boldsymbol{S}$ (of size $n \times 4$) is a matrix where each row i represents a contract and each column j contains the prediction of the churn or survival of the contract for the $j-th$ period (prediction obtained with the churn rate model of the period).

$$S = (s_{ij}) \text{ where } 1 \leq i \leq n \text{ and } 1 \leq j \leq 4, \text{ and } s_{ij} \in \{1(\text{churn}), 0(\text{no churn})\}. \quad (2)$$

The vector of average duration of size 4 and denoted $\boldsymbol{D}= (d_1, d_2, d_3, d_4)$ presents the real average duration of contracts terminated during the j-th period, j ranging from 1 to 4. To estimate the duration, an algorithm taking as input the matrix S and the vector D is used. It is described in the diagram 4 above (for the contract i).

Thus, by using it on the database contracts whose real duration over 4 years is not known, the process will end up on a database with contracts taken out between 2015 and 2020 with their average duration over 4 years. The contracts were therefore all observed over the same period.
The following table 5 presents the predicted churn rates during the first, second, third and fourth year as well as the predicted average lifetimes according to the course.

	Churn rate per year				
Course	1st	2nd	3rd	4th	Average duration
digital agency (ADA)	20%	19%	19%	7%	3.08 years
traditional agency (ATA)	19%	19%	17%	7%	3.12 years
traditional platform (MTM)	21%	20%	19%	8%	3.00 years
full Web (WDM)	30%	26%	25%	14%	2.65 years

TABLE 5 – Predicted churn rates during the first, second, third and fourth year as well as the predicted average duration according to the course.

Agency routes have the lowest rates and the highest duration since there are more profiles likely to last a long time (agency journeys have more than 27% of owners while web journeys have fewer than 7%). Over 4 years, the duration of owners is approximately 3.5 years and is greater than that of tenants, which is equal to 2.6 years, a difference of almost a year. It is possible to read this information in the following table 6, where the duration is presented according to its most discriminating variables. For each variable, the lifespans are decreasing by modality.

Variable	**Terms**	**Average duration**
Legal quality	Owner	3.5 years
	Tenant	2.6 years
Client's age	Seniors	3.3 years
	Adults	3 years
	Youth	2.6 years
Number of contracts	Several contracts	3.2 years
	1 contract	2.8 years

TABLE 6 – Average duration (over 4 years) according to 3 discriminating variables

The estimated duration will ultimately help optimize profitability.

Profitability optimization

Most Loyal Profiles

To determine the characteristics of the most loyal profiles, the contracts are grouped according to the most discriminating variables for the duration, which results in 354,285 duration profiles. So, in order to identify what defines a profile that lasts long in the portfolio, a decision tree is going to be used. The decision tree makes it possible to determine to which duration class the contract belongs according to several criteria.
By interpreting the classification decision tree obtained, we notice that owners, senior tenants or adult multi-holders tenants are generally the most loyal profiles (they last more than 3 years). Owners have an 87% chance of lasting more than 3 years. We have already seen that the course has no influence on churn. Therefore, there is no need to identify the profiles that would have lasted longer if they were in another course.

Most profitable profiles

We then take into account a profitability indicator, the Claims to Premiums (S/C) ratio. This will make it possible to identify the most profitable profiles as well as the less profitable profiles according to their S/C and their duration. As a result, it will be possible to determine not only the profiles that could continue to be new business but also the profiles to avoid in each route and the profiles whose profitability we could try to improve.
The most profitable profiles are in all the routes loyal apartment tenants. The profiles to avoid in web courses are the owners. For profiles with a high S/C but a low duration, it could be considered to offer them free months since their S/C leaves a sufficiently large margin to lower their price. Much more study would be needed to know exactly what strategy to adopt to improve profitability.

Optimization under constraints

Finally, the optimization under constraints will make it possible to identify the profiles that could be chosen in each route in order to maximize profitability while satisfying several constraints including costs. The 4 constraints include the costs (the cost of acquisition in particular), the total number of contracts and the total amount of premiums issued. Costs are assumed to be fixed per route. Each route is divided into 2 or 3 *clusters* (i.e groups with homogeneous S/Cs) i.e a *cluster* with more profitable profiles and a *cluster* with less profitable profiles.
The unknown variables of the problem are therefore the number of contracts to have in each of the 16 clusters in order to minimize the S/C over 4 years according to the constraints. The objective function is then the claims-to-premiums ratio of the profiles over 4 years. It is not convex, which means that it is possible to find a local optimum instead of the desired global optimum. But by using global optimization algorithms like **differential evolution** and **basin-hopping**, the global optimum seems to have been found.
We were therefore able to determine which profiles we should have in new business in each course in order to obtain a minimum S/C over 4 years equal to 53.67% (10% less than the current 4-year S/C) while respecting the constraints. At the same time, the routes that should be more developed (the Web path WDM) and the courses that could be eliminated (generally platform routes) are identified.

Conclusion

This thesis shows that it is possible to predict average duration by modeling the churn rates per period. The models also made it possible to conclude that the routes have very little impact on duration, so there is no longer any need to identify the profiles that would have lasted longer if they were in another route. Predicting duration has identified that the profiles that last the longest are generally owners, senior tenants or adult tenants with multiple contracts.
Optimization under constraints has made it possible to see which profiles should be in new business in each route and that it is possible to determine which path of the multi-access strategy could be eliminated. It could be interesting to take probabilities into account in order not to neglect the behavior of the customer. Indeed, we guide customers without worrying about their preferences.

By trying to improve the prediction of duration which is possible by improving the performance of churn rate models (thanks to the introduction of some variables such as the frequency of predicted claims) and by correcting the limits of the optimization strategy (fixed costs per route for example), we could take advantage of the constrained optimization strategy with much more certainty. The optimization also made it possible to have an idea of the most profitable profiles according to their duration and the S/C ratio as well as an idea on how to improve profitability. This could constitute an aid allowing to elaborate a strategy to improve the profitability of low profitable target profiles.

Remerciements

J'aimerais adresser mes remerciements les plus sincères à mon tuteur en entreprise David JAOUEN non seulement pour son implication, son suivi, sa disponibilité, ses bons conseils et toutes les connaissances que j'ai acquises grâce à lui qui m'ont permis de pouvoir écrire ce mémoire mais aussi pour sa grande gentillesse et sa positivité. Je remercie également Yann QUELEN, manager de l'équipe Pilotage Technique IARD chez Allianz, pour son suivi et ses conseils.

Merci aussi à tous les membres de la squad Pilotage Multi Accès, Christine DE GANDT, Mamihasina RAKOTOBE pour leur accueil chaleureux et leurs conseils tout au long du stage. Merci aussi à Ilias KAMAL pour nos échanges et pour sa gentillesse ainsi qu'aux membres de l'écosystème Ma vie quotidienne pour leur accueil.

Mes remerciements s'adressent vivement à Anne-Charlotte BONGARD pour son suivi et ses conseils ainsi que Quentin GUIBERT et plus généralement à toute l'université Paris Dauphine qui m'a permis d'avoir suffisamment de connaissance pour mener à bien ce stage. Je remercie également Nicolas FORCADEL ainsi que les professeurs du département Génie Mathématique de l'INSA Rouen pour l'enseignement et pour m'avoir donné la chance de faire cette double formation.

J'aimerais finalement remercier toute ma famille et particulièrement ma mère, qui est *la meilleure personne du monde* et ma favorite aussi sans oublier ma petite soeur, *la deuxième meilleure personne du monde* mais pas ma favorite. Merci aussi à mes très chers amis.

Table des matières

Introduction

A l'ère du digital, les moyens de distribution des produits d'assurance se sont vu bouleverser par Internet. Un nouveau canal de distribution s'est développé, s'ajoutant ainsi aux canaux traditionnels qui reposent sur les agences et les sociétés de courtage. Il s'agit du canal direct. Un assuré du canal direct utilise Internet mais aussi les plateformes téléphoniques comme moyen de contact avec l'assureur. Depuis quelques années, ce canal attire de plus en plus de clients puisqu'il permet non seulement de comparer les différentes offres du marché et de trouver celle qui répond le plus aux attentes, mais aussi de souscrire une offre quand le prospect le souhaite, où il le souhaite et de bénéficier de tarifs avantageux. Cependant, certains prospects ou clients restent plus rassurés par les moyens de distribution traditionnelle car ils permettent, entre autres, d'avoir un contact direct avec un interocuteur. Cela pourrait expliquer pourquoi le canal direct a du mal à croître en part de marché comme nous le pensions c'est-à-dire comme il l'a fait au Royaume-Uni. La part de marché du mode de distribution traditionnelle reste quant-à-elle limitée par celles des bancassureurs et des mutuelles d'assurances. De ce fait, certaines compagnies d'assurance ont décidé d'adopter une stratégie multi-accès. Cette dernière repose sur les moyens de distribution traditionnelle mais aussi sur les plateformes téléphoniques et Internet afin de permettre à l'assuré ou au prospect d'avoir le parcours client qui lui correspond le plus. Un prospect peut, par exemple, faire un devis sur Internet puis contacter la plateforme pour souscrire et faire gérer son contrat par une agence. Nous avons, de ce fait, 8 parcours client avec la stratégie multi-accès d'Allianz que nous expliquerons plus en détail dans la première partie du mémoire : 2 parcours agence (contrats souscrits en agence), 4 parcours plateforme et 2 parcours Web.

Les profils rencontrés dans les différents parcours client de la stratégie multi-accès ne sont pas forcément les mêmes et chacun de ces parcours a un coût d'acquisition client différent. Le coût d'acquisition est le coût qui a permis d'attirer un prospect vers les produits d'Allianz. Pour les parcours Web, il comprend les commissions des comparateurs internet, pour les parcours plateforme, il comprend les coûts des appels téléphoniques tandis que pour les agents, une partie des couts d'acquisition est passée dans les commissions. C'est un coût qui peut être plus élevé que la prime moyenne notamment pour les parcours plateforme. Ainsi, pour l'amortir, il est nécessaire de garder le client le plus longtemps possible en portefeuille. De même, le ratio Sinistres sur Primes (S/C) qui est un important indicateur de rentabilité baisse globalement lorsque la durée de vie des contrats dans le portefeuille augmente. Cette baisse est obtenue grâce à la revalorisation de la prime.

Notre objectif principal est d'optimiser la rentabilité sur plusieurs années de la stratégie multi-accès suivant plusieurs contraintes comprenant le coût d'acquisition. Cette optimisation passe donc par l'identification des profils les plus rentables suivant plusieurs contraintes afin d'encourager la souscription sur ces contrats. Optimiser la rentabilité à long terme de la stratégie multi-accès reviendrait donc à choisir les clients qui vont le plus durer en portefeuille car comme nous l'avions dit plus tôt, le ratio S/C baisse lorsque la durée de vie augmente et le coût d'acquisition peut être amorti.
Cela nous pousse à nous poser plusieurs questions : est-ce que le parcours client choisi a un impact sur la durée ? Par exemple, est-ce qu'un client avec un profil donné a plus de chance de résilier lorsqu'il est

dans un parcours Web que lorsqu'il est dans un parcours agence ? Est-ce qu'en orientant un profil vers un parcours plutôt qu'un autre, il va durer plus longtemps ? Si c'est le cas, nous pourrions orienter les clients vers le(s) parcours où ils dureraient le plus afin d'optimiser la rentabilité.
Lorsque les durées de vie sont égales, les assurés ayant un meilleur S/C sont plus rentables. Pour établir une approche plus solide d'optimisation de la rentabilité de la stratégie multi-accès sur plusieurs année, il faudrait cibler les contrats ayant un meilleur S/C (sur leur durée de vie) suivant un coût d'acquisition total donné et plusieurs autres contraintes.

Le sujet du mémoire est donc de modéliser les durées de vie des contrats d'assurance habitation afin d'optimiser la rentabilité et donc de contribuer au pilotage de la stratégie multi-accès de l'entreprise. Pour cela, nous allons modéliser les taux de résiliation des affaires nouvelles en assurance habitation en utilisant plusieurs méthodes de prédiction *machine learning*.
Dans la première partie, nous allons expliquer plus en détail le contexte ainsi que la création de la base de données qui va être utilisée pour la modélisation. Dans la seconde partie, nous présenterons les différents modèles *machine learning* utilisés pour modéliser les taux avant de modéliser les durées de vie en utilisant les modèles de taux de résiliation dans la troisième partie. Nous allons finalement utiliser les durées de vie obtenues afin d'essayer d'optimiser la rentabilité dans la quatrième partie.

Chapitre 1

Contexte général et base d'étude

1.1 Contexte général

1.1.1 Le contrat d'assurance

Un contrat d'assurance [1] se définit comme ≪un contrat par lequel l'assureur promet au souscripteur (pour son compte ou celui d'un tiers) après le paiement d'une prime ou cotisation, une prestation généralement pécuniaire suite à la réalisation d'un risque.≫.

Il **engage deux ou plusieurs autres parties** qui sont :

— **L'assureur**
L'assureur s'engage à indemniser le bénéficiaire du contrat d'assurance en cas de sinistre. Dans le droit des assurances, il peut s'agir d'un intermédiaire d'assurances (agent général d'assurances ou courtiers), d'une société civile (SAM), d'une société commerciale (SA), d'une société européenne.

— **Le souscripteur**
Le souscripteur, nommé ≪ preneur d'assurance ≫ (*policy holder* en anglais) en droit communautaire, est le signataire des documents contractuels et celui qui prend l'engagement de verser les primes. L'assurance pouvant être contractée par un tiers pour le compte de l'assuré, le souscripteur n'est pas forcément l'assuré.

— **L'assuré**
Le risque assuré dépend des intérêts de l'assuré en assurance dommage. L'assuré bénéficie des prestations de l'assureur aprés la réalisation dudit sinistre.

— **Les tiers bénéficiaires**
Les tiers bénéficiaires représentent les personnes (différentes de l'assuré) qui bénéficient des prestations de l'assureur après la réalisation dudit sinistre.

Il y a aussi deux grands types de tiers bénéficiaires : les créanciers privilégiés [2] et les victimes en assurance de responsabilité.

L'article L 124-3 du Code des Assurances et la jurisprudence donnent à la victime d'un dommage un droit d'action directe à l'encontre de l'assureur et du responsable assuré. Cette victime peut donc bénéficier des prestations de l'assureur à cause du dommage causé par l'assuré.

Le contrat couvre des risques qui constituent les **garanties**.

1. Définition du contrat d'assurance
2. Ils concernent les créanciers qui bénéficient d'un privilège. Il y a l'exemple du voisin ou du propriétaire d'un immeuble loué ayant un privilège sur les meubles du bien assuré avec le risque locatif ou le recours du voisin.

L'assuré peut avoir le choix entre plusieurs formules, de moyenne à haut de gamme, les garanties et leurs niveaux de couverture variant suivant la formule.
Les garanties généralement incluses dans un contrat d'assurance habitation comprennent une garantie couvrant le bien immobilier, une garantie incendie, une garantie catastrophe naturelle, une garantie dégât des eaux, une garantie bris de glace, une garantie cambriolage, une garantie mobilier [3], une garantie couvrant les équipements [4], une garantie couvrant les objets de valeur [5],etc.
Certaines garanties ont aussi une franchise.
Chez Allianz, il y a un socle de garanties en base indissociables et obligatoires : Incendie y compris Tempête Grêle Neige (TGN), Attentats, Catastrophes Naturelles et Technologiques, Responsabilité Civile (RC), Dégâts des Eaux (DDE). Il y a aussi 3 garanties avec des profondeurs de couverture gérées sous formes de réglettes : vol-vandalisme (il a 2 niveaux de garantie), bris de glace (il a 2 niveaux) et assistance (elle a 3 niveaux).

Lorsque l'assuré trouve qu'il n'est plus suffisamment protégé avec les garanties proposées par son contrat, il peut le résilier. Avant 2015, les assurés devaient attendre la date d'échéance pour pouvoir résilier leur assurance, à moins qu'un évènement validé par le Code des assurances ne justifie une résiliation anticipée. Puis **le 17 mars 2014, la loi Hamon a été introduite**.

1.1.2 La loi Hamon et la loi chatel

La loi Hamon

Du nom de Benoît Hamon, alors ministre délégué à l'économie sociale et solidaire à la consommation, la loi Hamon (LOI n° 2014-344 du 17 mars 2014) fait partie des lois sur l'assurance qui ont apporté de nombreux changements au sein du Code de la consommation, pour favoriser la concurrence et ainsi protéger davantage le consommateur en facilitant la résiliation des contrats. Elle est entrée en vigueur en 2015.
Elle donne le droit aux assurés de résilier leur contrat après la **première échéance**, sans avoir à donner de motif de résiliation à son assureur et sans pénalité financière. Ainsi, que les garanties du contrat ne correspondent plus à l'assuré ou qu'il trouve les cotisations trop coûteuses, il peut résilier sans donner de raison à sa compagnie d'assurance !

Cependant, les contrats d'assurance sont toujours reconduits tacitement mais cette reconduction tacite est réglementée par la loi Chatel.

La loi Chatel

Entrée en rigueur le 1er août 2005, la loi Chatel (LOI n° 2005-67 du 28 janvier 2005 tendant à conforter la confiance et la protection du consommateur) a été proposée par le député Luc Châtel afin de permettre aux assurés de changer plus facilement d'assurance. La loi Chatel oblige l'assureur à envoyer un avis d'échéance annuelle entre 3 mois et deux semaines (en assurance habitation) avant la date limite de résiliation.
La date limite de résiliation doit être clairement lisible, dans un encadré. Dans le cas contraire, l'assuré a la possibilité de résilier son contrat dans les 20 jours (en assurance habitation) à partir de la date d'envoi de l'avis. Si l'assureur n'a pas adressé d'avis d'échéance, l'assuré pourra alors résilier son

3. prend en charge les meubles ou objets présents dans le logement de l'assuré
4. high-tech, électroménager, hifi, vidéo...
5. la définition d'objet de valeur varie d'une compagnie à une autre et d'un contrat à un autre. Certains couvrent les bijoux, les oeuvres d'art ou les deux. Cette garantie est toujours assortie d'un plafond d'indemnisation.

contrat à tout moment. Cette résiliation prendra effet dès le lendemain de l'envoi de la lettre recommandée de résiliation. Elle ne se limite pas à la non reconduction d'un contrat à la date d'échéance. Elle englobe un grand nombre de décisions concernant le commerce en général. Tous les types de contrats à tacite reconduction sont soumis à la loi Chatel. Cela concerne donc les contrats d'assurance, les abonnements, contrats d'entretien, etc. Il y a toutefois quelques exceptions dans chaque catégorie.

Les loi Hamon et Chatel permettent donc aux assurés de résilier plus facilement leurs contrats. Nous allons à présent voir quels sont les différents motifs de résiliation.

1.1.3 Les motifs de résiliation

L'assuré et l'assureur peuvent résilier le contrat d'assurance pour diverses raisons.

Les motifs de résiliation par l'assuré

Les motifs qui permettent à l'assuré de résilier son contrat avant la première échéance ne sont pas totalement les mêmes qu'après la première échéance. Rappelons qu'après la première échéance, l'assuré n'est plus obligé de fournir un motif de résiliation grâce à la loi Hamon. Les différents motifs sont résumés dans le tableau 1.1 suivant.

Avant la première échéance	**Après la première échéance**
— Le changement de la situation de l'assuré, comme un déménagement impliquant une modification du risque couvert (Article L.113-16 du code des assurances). — La perte du bien assuré : par exemple, la vente de l'habitation assurée (Article L.121-10 du Code des assurances). — Le décès de l'assuré : dans ce cas, la couverture est maintenue au profit de l'héritier de la personne décédée qui peut conserver ou non le contrat d'après le code des assurances (Article L.121-10 du code des assurances).	— L'augmentation de la prime ou alors l'obtention d'un tarif plus avantageux chez un autre assureur. — L'insatisfaction de la qualité du service de l'assureur (suite à un délai de règlement du montant des sinistres trop élevé par exemple). — Motifs de résiliation avant la première échéance.

TABLE 1.1 – Différents motifs de résiliations par l'assuré avant et après la première échéance

Les motifs de résiliation par l'assureur

Le contrat d'assurance peut aussi être résilié par l'assureur [6]. Il doit motiver les raisons de la résiliation. Ces différents motifs sont résumés dans le tableau 1.2 suivant.

6. Article L.113-12-1 du code des assurances

Avant la première échéance	Après la première échéance
— Changement de situation (entrainant une modification du risque) non déclaré : dès que l'assuré a connaissance du changement, il doit le déclarer à l'assureur par lettre recommandée dans un délai de 15 jours au maximum. Si cette modification entraîne une aggravation du risque, l'assureur peut, soit résilier le contrat, soit proposer une augmentation de la prime. — Fausse déclaration, par exemple donner une superficie du logement assuré différent de la superficie réelle lors de la souscription du contrat d'assurance : lorsque l'assureur remarque que l'assuré a fait une fausse déclaration ou n'a pas déclaré précisemment le risque, il peut résilier le contrat d'assurance. Il doit notifier l'assuré par lettre recommandée et la résiliation prendra effet 10 jours après. La partie de cotisation correspondant à la période comprise entre la résiliation effective du contrat et l'échéance initialement prévue doit être remboursée à l'assuré.	— Sinistralité importante : l'assureur doit, dans ce cas, envoyer une lettre recommandée à l'assuré au moins deux mois avant la date d'échéance. — Non-paiement des primes : l'assureur peut résilier le contrat d'assurance lorsqu'il y a non paiement des primes. Il envoie d'abord une lettre recommandée à l'assuré s'il ne paie pas dans les dix jours qui suivent la réception de l'avis d'échéance. Puis, si trente jours après la réception de la lettre, l'assuré ne paie toujours pas la prime, le contrat d'assurance fait d'abord l'objet d'une suspension de garantie. Si la situation n'est toujours pas réglée dans les dix jours suivants, l'assureur résilie alors le contrat et la prime annuelle reste intégralement due à l'assureur. — Motifs de résiliation avant la première échéance.

TABLE 1.2 – Différents motifs de résiliations par l'assureur avant et après la première échéance

Nous venons donc de voir qu'il y a plusieurs motifs qui peuvent expliquer la résiliation d'un contrat. L'assuré peut alors utiliser un des différents moyens de distribution des produits d'assurance pour souscrire un nouveau contrat d'assurance.

1.1.4 Les moyens de distribution des produits d'assurance

Les moyens de distribution traditionnelle

Contrairement à d'autres produits, la distribution des produits d'assurance est soumise à une réglementation afin de protéger le consommateur (LAMBERT, 1998). Les entreprises relevant du code des assurances ont la possibilité de distribuer leurs produits en faisant recours à des moyens de distribution allant des intermédiaires (comme les agents généraux ou les courtiers qui constituent les distributeurs traditionnels) à leurs propres réseaux de salariés.

En vertu de l'article R.511-2 du code des assurances, les produits des entreprises relevant de ce code ne peuvent être présentés que par quatre catégories de personnes qui sont :
— les courtiers ;
— les agents généraux ;
— les salariés des courtiers, agents généraux ou entreprises d'assurance et
— les mandataires non salariés des courtiers, agents généraux ou entreprises d'assurance.
Des intermédiaires occasionnels énumérés aux articles R.512-3 à R.512-5 du code des assurances peuvent aussi distribuer des produits d'assurance pour certaines opérations particulières ou accessoires à d'autres contrats. Par exemple, les vendeurs d'objets mobiliers peuvent vendre des assurances garantissant contre le vol ou la perte des objets vendus par leurs soins.

Les agents généraux

Les agents généraux d'assurance sont des professionnels indépendants qui représentent, en vertu d'un mandat dit traité de nomination, une ou plusieurs entreprises d'assurance avec la ou lesquelles ils travaillent exclusivement (Lambert, 1998). Cette exclusivité entraîne que les agents généraux connaissent très bien les produits d'assurance de(s) compagnie(s) d'assurance dont ils sont mandataires. Ils peuvent néanmoins collaborer avec plusieurs compagnies lorsque cette collaboration ne porte alors que sur des produits qui ne sont pas concurrents. Par exemple, ils vont travailler avec la compagnie A pour l'assurance santé et avec la compagnie B pour l'assurance habitation.

Les agents généraux sont rémunérés par des commissions comprenant une commission de gestion (rémunère les travaux de gestion effectués par l'agent général) et une commission d'apport (rémunère l'acquisition d'un contrat d'assurance et est fixée en pourcentage des primes). Une double fonction est exercée par la plupart d'entre eux : une fonction administrative (de par l'encaissement de primes ainsi que l'indemnisation des sinistres) et une fonction commerciale (de prospection).

L'agent général n'est pas propriétaire du portefeuille de contrats qu'il apporte à sa compagnie mandante. Cependant, il détient des droits de créance sur les commissions afférentes au portefeuille de l'agence. Il est possible que l'agent général récupère ces droits de créance en cessant ses fonctions, par la perception d'une indemnité compensatrice à la charge de sa compagnie, en pourcentage des commissions, ou bien par la vente à titre onéreux de son agence à un successeur agréé par sa compagnie.

Les courtiers

Les courtiers exercent à titre individuel, ou représentent des sociétés de taille variable, filiales de sociétés étrangères et de banques qui peuvent leur confier des pouvoirs de souscription, de gestion et de règlement des sinistres (Lambert, 1998). Ces sociétés les rémunèrent généralement par des commissions. Le courtier n'est pas lié exclusivement à une compagnie d'assurance et est mandataire de ses clients. Chez un courtier, l'assuré peut choisir par exemple, une assurance habitation de la compagnie X, Y ou Z. Cela entraîne que les courtiers peuvent ne pas connaître aussi bien les produits qu'ils proposent que les agents généraux.

D'après l'article 109 du code de commerce, les courtiers ont le statut de commerçant. Ils peuvent exercer des activités connexes de courtage de réassurance, de gestion de risques ou d'audit en assurance.

Les concurrents des distributeurs traditionnels

La réglementation des moyens de distribution des produits d'assurance n'a pas empêché l'infiltration du marché par de nouveaux intervenants comme les banques et les sociétés sans intermédiaires. Les sociétés sans intermédiaires peuvent représenter la vente directe par les compagnies d'assurance ou la distribution effectuée pour leur propre compte par les mutuelles sans intermédiaires, qui ne peuvent statutairement avoir recours à des intermédiaires rémunérés.

La bancassurance

La bancassurance se définit comme la distribution de produits d'assurance aux guichets des banques et des établissements financiers. L'Espagne et la France ont été les premiers pays à l'initier. C'est face à un marché mur et très concurrentiel en matière bancaire dans les années 70 en France que les banques françaises se sont lancées dans la bancassurance [7]. Celle-ci représentait pour les banques, une nouvelle source de profit, un moyen de diversifier leur activité bancaire et de fidéliser leurs clients.

La distribution directe

Avec ce canal, le prospect ou client a la possibilité d'être directement en contact avec la compagnie

7. Caillol et Giraud, 2017

d'assurances par plateforme téléphonique ou en ligne. L'implantation des assureurs directs en France s'est réalisée par vagues successives, depuis les années 70 (Lambert, 1998). Il faut attendre l'émergence des comparateurs d'assurance en 2010 pour voir l'émergence de ce canal.

1.1.5 L'émergence du canal direct

Le développement du canal direct est principalement lié à celui des comparateurs d'assurance.

Comparateurs d'assurance

Les comparateurs d'assurances sont des programmes informatiques permettant de faire simultanément des devis par plusieurs compagnies d'assurance afin de comparer les garanties et les primes. Nous assistons à leur émergence en 2010 grâce à *Assurland*[8] avec l'arrivée du comparateur *Lelynx.fr*, filiale du leader britannique des comparateurs en ligne *Confused.com*. Le marché de l'assurance directe en France reste toujours difficile à conquérir alors que certains espéraient qu'il se développe autant qu'au Royaume-Uni où plus de la moitié des contrats sont désormais souscrits par un comparateur. C'est *Confused.com* créé en 2002 par l'assureur direct Admiral qui a a transformé la distribution d'assurance auto au Royaume-Uni. La combinaison de prix d'assurance en hausse et de l'application de la loi Chatel, qui limite les effets de la tacite reconduction, a aidé à gagner en part de marché en France mais celle-ci reste toujours faible (0.2% en 2019 (FFA, 2019)).

Si les comparateurs ont permis de développer le marché du canal direct c'est parcequ'ils sont, sur le Web, la source majeure d'information pour les personnes voulant souscrire un contrat d'assurance dommage. D'après un article de l'argus de l'assurance[9] :
— 39% des individus qui ont souscrit un contrat ou formulé une demande pour une assurance auto en ligne ont utilisé un comparateur dans leur parcours. Cela s'explique par le fait qu'un comparateur permet d'avoir une vue d'ensemble sur les différentes offres du marché de l'assurance.
— 61% des personnes qui ont eu recours aux comparateurs le font parce qu'ils sont faciles à utiliser. Les comparateurs banissent les termes techniques ce qui permet de comprendre plus facilement l'assurance et de faciliter la recherche des tarifs les plus compétitifs.

Cependant, les comparateurs n'affichent pas les offres de tous les assureurs mais seulement ceux des assureurs avec qui ils sont en partenariat. Les offres affichées sur les comparateurs d'assurance suite à un devis sont proposées par des sociétés indépendantes des compagnies d'assurances qui rémunèrent le plus souvent les comparateurs par un intéressement sur les contrats souscrits par l'intermédiaire des comparateurs. Nous retrouvons principalement dans les comparateurs, les compagnies d'assurance directe et les sociétés de courtage. Les mutuelles et les bancassureurs n'en font pas partie.

Avantages et limites du canal direct

Si le canal direct se développe c'est parce qu'il présente plusieurs avantages vu qu'il :

— permet d'avoir une large vision du marché grâce aux comparateurs d'assurance,
— facilite la souscription en donnant au prospect la possibilité de la faire à tout moment et n'importe où,
— permet de trouver des tarifs avantageux ainsi qu'un produit répondant aux besoins du prospect ou client.

8. Les Echos, 2010
9. Chabrier, 2020

Malgré les avantages qu'il présente, le canal direct ne se développe pas en France comme il l'a si considérablement fait au Royaume-Uni. Sa part de marché reste faible en France (0,2% en 2019 d'après la FFA (2019)) et a même baissé de 0.8% entre 2018 et 2019 (en comparant avec les *données clés 2018* de la FFA). Cela s'explique par le fait que des prospects présentent une certaine réticence à ce canal. Celle-ci repose sur :
(1) l'appréhension de l'acte d'achat sur internet,
(2) l'idée qu'une assurance souscrite sur Internet offrirait une moins bonne gestion de sinistres,
(3) la sécurité.
La première raison expliquant la réticence repose sur le fait que certains assurés considèrent la qualité du service et les conseils de confiance d'un agent comme une partie importante de leur décision d'achat. Ils n'envisagent donc pas d'acheter un contrat d'assurance sur internet. Pour ce qui est de la troisième raison, elle n'est pas spécifique à l'assurance. Sur le digital, se pose actuellement la question de la sécurité et de la confidentialité des données. Cette volonté de garder la maitrise sur ses informations se comprend au vu des enjeux de cybersécurité, de piratages informatiques fréquents et de revente de données personnelles sur le Web. Une étude [10] sur la perception des données personnelles après l'adoption du règlement général sur la protection des données (RGPD) a été menée et publiée le 31 octobre 2019 conjointement par Médiamétrie et la Chaire Valeurs et politiques des informations personnelles de l'Institut Mines-Télécom [11]. D'après cette étude, les français sont de plus en plus vigilants sur internet (globalement, 67% des internautes français sont plus vigilants sur internet par rapport aux années précédentes (contre 54% en 2017)). De même, 70% des personnes interrogées utilisent des pseudonymes ou une fausse identité sur internet. Ces individus pourraient donc être réticents au fait de souscrire un contrat d'assurance en ligne puisqu'il faut fournir des informations sur son identité.

En outre, ce canal n'est pas forcément rentable. A leur début, les assureurs directs attiraient à eux tous les « mauvais risques », comme les jeunes ou les résiliés. Ainsi, Direct Assurance a dû attendre 10 ans après son lancement avant de réaliser des bénéfices. Elle a dû d'abord segmenter les risques. Aussi les lois Hamon et Chatel, décrites plus tôt, combinées avec la facilité qu'offre le canal direct pour souscrire une offre d'assurance, pourraient favoriser la volatilité des clients. Ceux-ci peuvent résilier après la première année d'échéance et souscrire facilement chez un concurrent pour bénéficier des tarifs plus attractifs en terme de prix par exemple.

Nous remarquons donc que les forces du canal direct représentent également ses faiblesses :
— La compagnie d'assurance peut perdre des clients aussi facilement qu'elle en gagne.
— La compagnie d'assurance peut attirer des clients qui vont lui faire perdre de l'argent.
Il est donc nécessaire de bien choisir les profils à accepter dans ce canal pour optimiser la rentabilité. Même si sa part de marché peine à croître considérablement, il reste quand même un canal qui attire des clients et qui mérite donc d'être exploité.

1.1.6 Comparaison des parts de marché des moyens de distribution

Le graphique ci-dessous (figure 1.1), ayant pour source FFA - *Données clés 2019* (FFA, 2019), présente les parts de marché des compagnies d'assurances non-vie en 2019. Sont donc prises en compte, les compagnie d'assurances de dommages aux biens mais également celles de responsabilités et de dommages aux personnes (santé et accidents corporels).

10. L'étude a été menée auprès d'un échantillon de 2017 internautes âgés de 15 ans et plus, représentatifs de la population internaute. La représentativité de l'échantillon a été assurée par la méthode des quotas (sexe, tranche d'âge en 5 classes, CSP en 5 classes et région Paris-Province) sur la base de l'enquête de référence de la population d'internautes en France, l'Observatoire des Usages Internet.
11. Waelbroeck et al., 2019

Nous remarquons que les sociétés avec intermédiaires (moyens de distribution traditionnelle) sont leaders du marché devant les mutuelles d'assurance sans intermédiaires (MAAF, MAIF, MACIF...). Les bancassureurs sont en 3ème position et en comparant avec les autres années, nous avons constaté qu'ils gagnent chaque année un peu plus de part de marché depuis quelques temps et que ce gain atteint parfois 1%. L'assurance directe a toujours la part de marché la plus faible (0.2%) et l'écart avec ses autres concurrents est très élevé. En comparant avec les données de 2018, nous remarquons que sa part de marché a baissé. C'est peut-être trop optimiste de penser que ce canal va continuer à conquérir encore plus de monde pour finalement atteindre le même niveau qu'au Royaume-Uni.

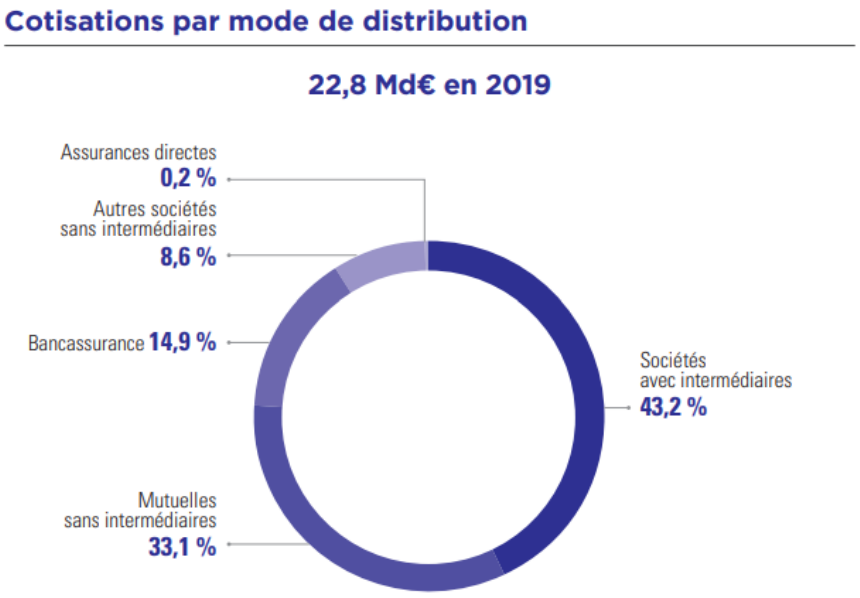


FIGURE 1.1 – Parts de marché par moyen de distribution en 2019 - Source : FFA - *Données clés 2019*

La part de marché des modes de distribution traditionnelle est limitée par celle des bancassureurs et des mutuelles d'assurances tandis que le canal direct attire de plus en plus de clients mais sa croissance reste aussi limitée. Afin de tirer profit de tous ces canaux, Allianz a donc décidé d'adopter une stratégie multi-accès.

1.1.7 La Stratégie multi-accès

La stratégie multi-accès repose sur les moyens de distribution traditionnelle mais aussi sur les plateformes téléphoniques et internet afin de permettre à l'assuré ou au prospect d'avoir le parcours client qui lui correspond le plus. Un prospect peut, par exemple, faire un devis sur internet puis contacter la plateforme pour souscrire et faire gérer son contrat par une agence.

La stratégie multi-accès se résume donc en 3 étapes :
- Une étape ***Sourcing*** : elle permet d'identifier la source du contrat, c'est-à-dire de différencier les contrats dont le devis a été fait sur internet des contrats dont le devis a été fait de manière traditionnelle en agence. Les contrats qui viennent du canal digital sont identifiés par la lettre (D) et ceux du canal traditionnel par la lettre (T).
- Une étape **Souscription** : les prospects ou clients ont la possibilité de souscrire en agence (A), en

contactant la plateforme téléphonique (M) d'Allianz ou en souscrivant sur le site Web (W) d'Allianz.
- Une étape **Gestion de contrat (intermédiaire)** : les clients peuvent faire gérer leur contrat en agence (A) ou par la plateforme téléphonique d'Allianz (M).

Le croisement de ces 3 étapes résulte sur 8 parcours clients possibles. Ils sont codés de telle sorte que la première lettre représente le canal de souscription, la deuxième lettre représente le sourcing et la troisième lettre représente l'intermédiaire (gérant du contrat).

La figure 1.2 suivante résume les différentes étapes du parcours client ainsi que les différents parcours obtenus en croisant les étapes.

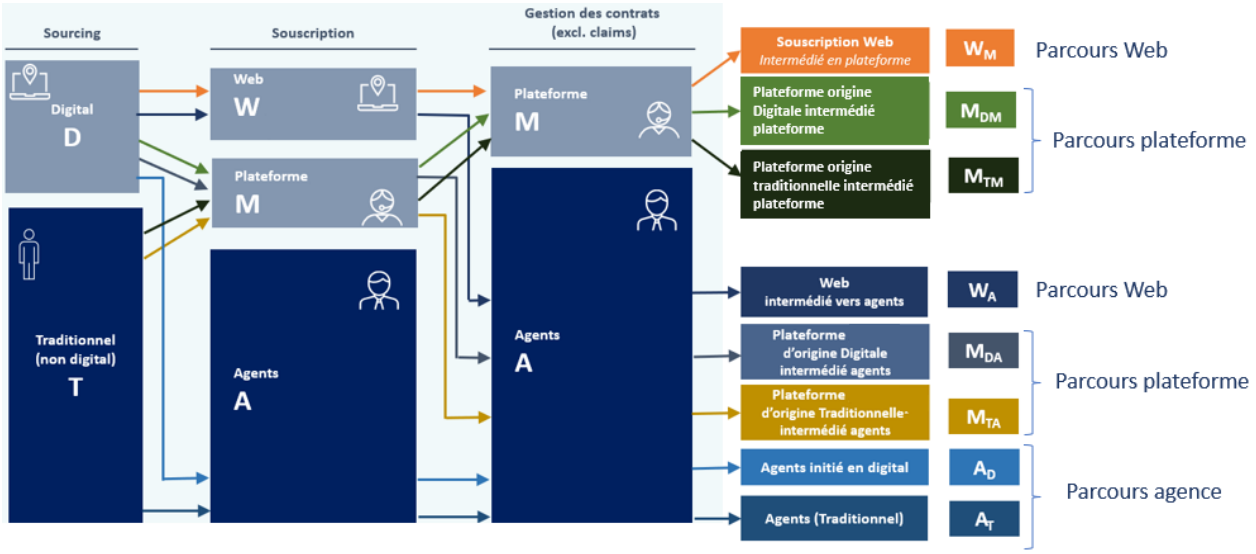


FIGURE 1.2 – Différentes étapes des parcours client de la stratégie multi-accès

Ainsi, le parcours ATA (qui constitue le parcours traditionnel) représente les contrats qui viennent d'une agence, souscrits et gérés en agence. Il représente le pourcentage le plus élevé du portefeuille. Il y a aussi le parcours ADA qui représente les contrats souscrits et gérés en agence mais d'origine digitale. Ces deux parcours constituent les parcours agence (contrats souscrits en agence).
Il existe aussi les parcours Web : WDA et WDM ainsi que les parcours plateformes : MTA [12], MTM, MDA et MDM.
Si un client fait son devis en ligne puis contacte la plateforme téléphonique d'Allianz pour souscrire et fait finalement gérer son contrat en agence alors il a le parcours MDA.

Ces parcours pourraient attirer des profils différents. Les clients les plus sensibles au prix pourraient être retrouvés dans les parcours digitaux vu que ce sont les parcours dans lesquels les prix les plus faibles sont généralement offerts. Avec la loi Hamon, ils peuvent changer d'assurance avec un préavis de 2 mois à n'importe quel mois après la première échéance de leur contrat. Cela, combiné avec la facilité à trouver une meilleure offre et à souscrire sur internet, peut les encourager à résilier et à souscrire chez un concurrent.

Les agences ont plus intérêt à fidéliser les clients puisqu'ils ont une commission sur la prime annuelle du client. Parmi les qualités de l'agent, il y a : avoir de l'empathie, un bon relationnel, être à l'écoute du client, savoir s'adapter à chaque client et situation, etc. Ces qualités peuvent aider à créer une relation

12. regroupe les contrats souscrits avec la plateforme téléphonique, dont le devis a été fait en agence et qui sont gérés en agence

de confiance avec l'assuré et le rendre plus loyal. **Cela pourrait faire croire que les assurés sont plus loyaux lorsqu'ils se trouvent en agence**. Dans ce cas, un assuré aurait plus de chance de résilier s'il se trouve dans un parcours Web que dans un parcours agence.

Les agents ont aussi des taux de protocole qui permettent de prendre en compte leur rentabilité et les poussent à bien sélectionner les assurés. Dans le canal direct, afin d'écarter les prospects potentiellement pas rentables, Allianz n'apparait dans le comparateur et ne propose un tarif suite à un devis sur son site que pour les profils les moins risqués. Par exemple lorsqu'un sinistré fait un devis sur Internet, Allianz n'apparaît pas dans le comparateur.

Grâce à la stratégie multi-accès, il y a 8 parcours client. Nous n'allons évidemment pas «ouvrir la porte» à tous les profils dans n'importe quel parcours au risque de ne plus être rentable. Ainsi, afin d'optimiser la rentabilité, il faut déterminer quels profils peuvent toujours être reçus comme affaire nouvelle dans chaque parcours, quels profils doivent être évités dans un parcours et donc dans quel(s) parcours il faut orienter un profil donné. En partant du constat qu'un profil qui dure plus longtemps est plus rentable, cela reviendrait à identifier les profils qui durent le plus longtemps dans chaque parcours et à orienter la souscription sur ces profils. Cependant, lorsque deux assurés ont la même durée de vie, le plus rentable est celui ayant le meilleur S/C (ratio sinistres à primes), ce qui pousse alors à aller plus loin qu'identifier simplement les profils qui durent le plus longtemps en prenant en compte leur ratio sinistres à primes (sur leur durée de vie) dans une approche d'optimisation de la rentabilité sous contraintes.

Mais qu'est-ce qui justifie que retenir un client le plus longtemps possible est plus bénéfique ? Pour répondre à cette question, nous allons donc évaluer l'avantage de la rétention client.

1.1.8 Avantages de la rétention client

Coût d'acquisition

Une prime d'assurance comprend la couverture du risque, les taxes et contributions, les frais d'acquisition et les frais d'administration.

Le coût d'acquisition client désigne le montant moyen dépensé pour générer un client. Retenir un client au lieu d'attirer un nouveau permet donc d'économiser en coût d'acquisition. Les parcours de la stratégie multi-accès ont des coûts d'acquisition différents voire très différents pour certains. Ainsi, pour encore plus économiser en coût d'acquisition, il faut savoir quel est le parcours idéal où il faudrait orienter une affaire nouvelle.

Chez Allianz, le coût d'acquisition se décompose en coût d'acquisition *sourcing* et coût d'acquisition hors *sourcing*. Le *sourcing* est l'origine du contrat, le premier contact avec le futur client.

— **Le coût d'acquisition** *sourcing* est constitué des investissements média et des commissions comparateurs. Il est net de la refacturation de *leads* et de celle de l'intermédiation aux agents. Les *leads* désignent des contacts que la compagnie espère voir se transformer en client. Le coût d'acquisition *sourcing* est donc égal à 0€ pour les moyens de distribution traditionnelle.

— **Le coût d'acquisition hors** *sourcing* se décompose en 2 parties : la première est commune à tous les parcours (il s'agit des frais fixes d'Allianz France attribuable à la vente de contrats) et la deuxième est spécifique aux parcours plateforme (c'est le coût des appels téléphoniques).

Une partie des coûts d'acquisition des agents est passée en commission. Les coûts annuels d'administration des agents sont donc plus élevés. Le coût d'administration est un ensemble de coûts récurrents. Il se décompose en trois parties :
— La première partie est commune à tous les parcours : elle représente les frais fixes d'Allianz France attribuable à la gestion des contrats.

— La deuxième partie est spécifique aux parcours gérés en plateforme : elle représente le coût des appels pour la gestion.
— La troisième partie est spécifique aux parcours gérés en agence : elle représente les commissions des agents.

Le coût d'acquisition peut être plus élevé que la prime moyenne pour certains parcours. La figure 1.3 suivante permet de comparer les coûts d'acquisition des parcours ainsi que leur prime moyenne mais aussi le coût annuel d'administration (fin 2020).

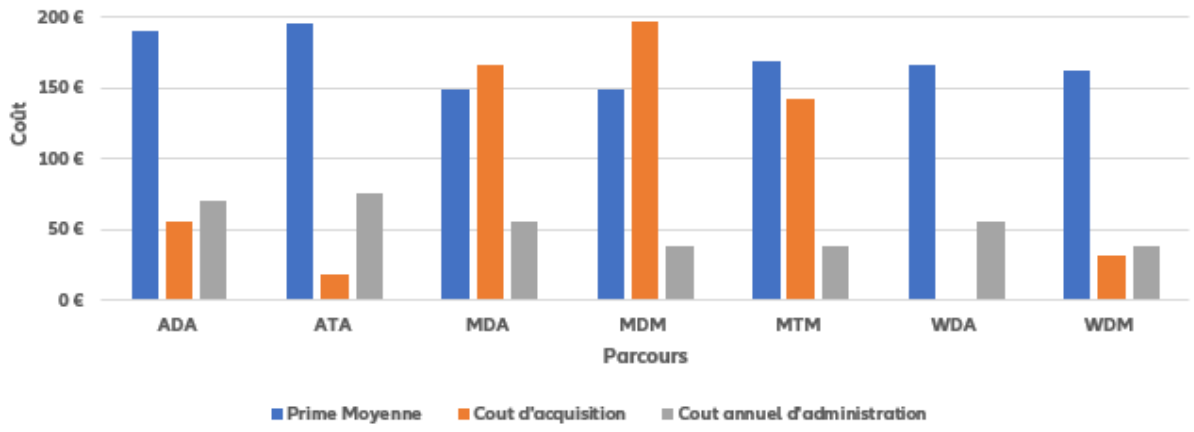


FIGURE 1.3 – Prime moyenne, coût d'acquisition client et coût annuel d'administration par parcours en fin 2020

Nous remarquons que les parcours Web ont des coûts peu élévés mais pour la plupart des parcours plateforme, le coût d'acquisition est plus élevé que la prime moyenne. Afin de pouvoir amortir le coût d'acquisition, il faudrait que le client reste le plus longtemps possible dans le portefeuille. Le coût d'acquisition pourrait influencer le choix du parcours où il serait préférable d'orienter le client. Par exemple, si le client va avoir la même durée quel que soit le parcours, il faudrait l'orienter vers le parcours ayant les coûts les plus bas afin d'économiser au mieux.
En essayant de déterminer les profils idéaux à avoir en affaire nouvelle dans chaque parcours, il est donc important de prendre en compte les différents coûts.

Ratio Sinistres à Primes (S/C)

Faire durer le client le plus longtemps possible dans le portefeuille est aussi bénéfique pour le ratio S/C qui est un important indicateur de rentabilité. Il est égal à la charge totale des sinistres sur la somme totale des cotisations encaissées. En effet, plus le contrat reste dans le portefeuille, plus le ratio sinistres à primes baisse en général. Cela est possible grâce à la revalorisation. En effet, les contrats subissent généralement une majoration à chaque échéance de leur contrat d'assurance habitation afin de mieux couvrir les sinistres. Cependant, il peut parfois arriver qu'il y ait des évenements naturels et des graves qui se produisent et qui font que le ratio sinistres à primes ne baisse pas quand la durée augmente.

La courbe suivante (figure 1.4) représente le ratio sinistres sur primes en 2019 des contrats habitation en portefeuille suivant leur durée de vie (en année) dans le portefeuille. Nous choisissons l'année 2019 qui est une année normale contrairement à l'année 2020 qui a été perturbée par la Covid-19. Par soucis de confidentialité, nous ne montrerons pas les chiffres. Les sinistres ont été observés du 1er Janvier 2019 au 31 Décembre 2019. La charge totale des sinistres comprend les charges attritionnelles, les graves et les événements naturels.

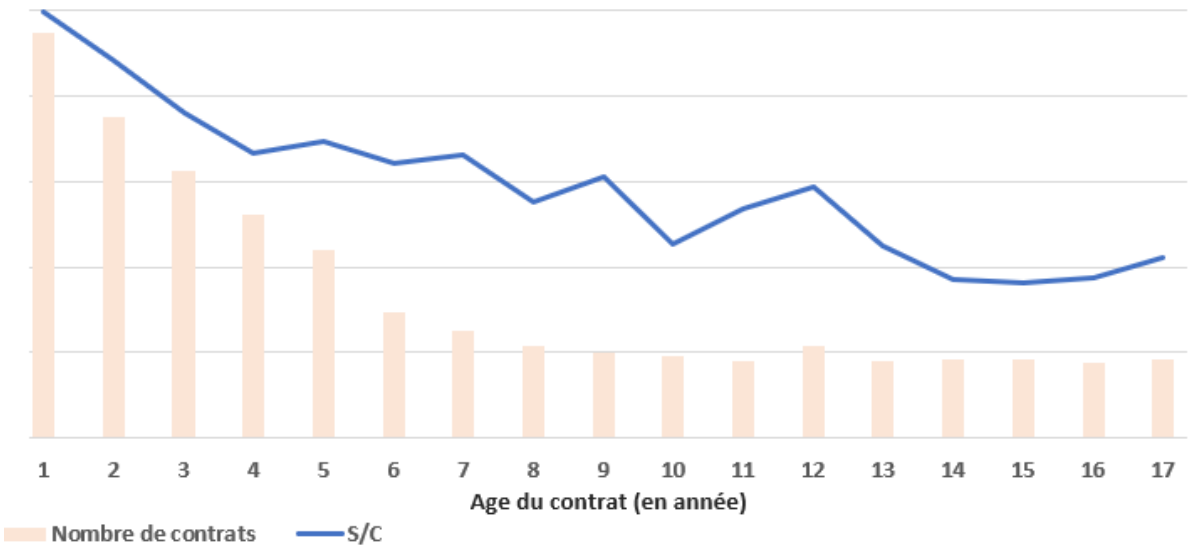


FIGURE 1.4 – Ratio sinistres à primes (S/C) en 2019 des contrats habitation en portefeuille suivant leur durée de vie (en année) dans le portefeuille

Nous remarquons que l'allure de la courbe est descendante. Quand le nombre d'années passées dans le portefeuille augmente, le ratio sinistres à primes baisse. La courbe n'est pas lisse car comme nous l'avions dit plus tôt, il a pu y avoir pendant l'année, un événement naturel ou un grave qui entraîne que le S/C ne baisse pas toujours lorsque la durée de vie augmente.

Nous venons donc de voir que garder un client le plus longtemps possible dans le portefeuille permet d'amortir le coût d'acquisition et d'avoir un S/C global plus faible. Pour optimitiser la rentabilité de la stratégie multi-accès, il est nécessaire d'identifier les profils qui durent le plus longtemps dans chaque parcours afin d'orienter la souscription sur ces contrats. A durée de vie égale, les clients ayant un S/C et un coût d'acquisition plus faibles sont plus rentables. Ainsi, en essayant de déterminer sur quels profils il faudrait orienter la souscription dans chaque parcours, il est nécessaire de prendre en compte le S/C sur la durée de vie de contrat mais aussi le coût d'acquisition, le coût d'administration et d'autres contraintes. Cela nous pousse alors à modéliser les durées de vie puis à faire une optimisation de la rentabilité et donc une minimisation du S/C sous contraintes (comprenant les coûts) dans la suite du mémoire.

Maintenant que nous comprenons bien le contexte de notre étude, nous pouvons passer à la création de la base d'étude pour la modélisation.

1.2 Création de la base d'étude

Pour créer notre base de modélisation, nous allons utiliser plusieurs bases de données d'Allianz.

1.2.1 Bases mouvements annuelles des contrats habitation

Ce sont des bases annuelles qui contiennent les entrées et les sorties des produits habitation du 1er Janvier au 31er Décembre de l'année donnée. Elles vont donc permettre de récupérer les affaires nouvelles (AFN) ainsi que certaines de leurs caractéristiques et d'observer leur éventuelle résiliation. Nous allons utiliser les bases de 2015 à 2020. La stratégie multi-accès a commencé en 2015 et nous estimons un recul avec 5 ans d'ancienneté suffisant.
Ces bases contiennent les différents mouvements des contrats habitation pendant l'année donnée, ce qui entraine qu'il peut y avoir plusieurs lignes pour un contrat. Il faut donc bien identifier les affaires nouvelles pour récupérer leurs caractéristiques et les ajouter à notre base d'étude. Il faut éviter de récupérer des caractéristiques qui ne correspondent pas à celles de l'affaire nouvelle mais plutôt à celles de la résiliation. Il faut de même identifier les résiliations correspondantes à ces AFN et ajouter leurs caractéristiques à notre base d'étude. Les résiliations sont reconnues grâce à la variable `topres` et les affaires nouvelles (AFN) grâce à la variable `topent`.
Sachant qu'une résiliation ou une AFN peut être annulée, plusieurs cas peuvent être retrouvés :

- `topent`= 1 : le contrat est une AFN.
- `topent`=-1 : le contrat est une AFN annulée.
- `topent`= 0 : le contrat n'est plus une AFN.
- `topres`= 1 : le contrat a été résilié.
- `topres`=-1 : la résiliation du contrat a été annulée.
- `topres`= 0 : le contrat n'est pas résilié.

Par exemple, si l'assuré résilie puis annule la résiliation, le contrat a, dans la base, une ligne avec `topres` $= 1$ puis une ligne avec `topres` $= -1$ ce qui entraîne que la somme de la variable `topres` du contrat est égale à 0. Le contrat n'a pas été résilié. C'est pourquoi, afin de vérifier qu'il y a bien eu résiliation ou non, il faut regarder la somme des valeurs de la variable `topres` du contrat. S'il est égal à 1 alors il y a bien résiliation. De même, une AFN est enregistrée si la somme de `topent` du contrat est égale à 1.

$$\sum \texttt{topent}_{contrat} = 0 : \text{Le contrat n'est pas une AFN.}$$
$$\sum \texttt{topent}_{contrat} = 1 : \text{Le contrat est une AFN.}$$
$$\sum \texttt{topres}_{contrat} = 0 : \text{Le contrat n'a pas été résilié.}$$
$$\sum \texttt{topres}_{contrat} = 1 : \text{Le contrat a été resilié.}$$

Il peut y avoir, en effet, plusieurs lignes correspondantes à des ≪résiliations≫ d'un même contrat. S'il y a bien résiliation, il n'y a qu'une seule qui doit être utilisée pour récupérer la date de résiliation et l'âge du contrat. Il en est de même pour l'AFN.
La date de résiliation permet de repérer la ≪ vraie ≫ résiliation. Ainsi la ≪ vraie ≫ résiliation correspond à la ligne où la date de résiliation est la plus récente par rapport aux autres résiliations (vu que les autres résiliations ont été annulées). De même, il peut y avoir un mouvement d'entrée (identifié avec `topent` $= 1$) qui sera ensuite annulé (`topent` $= -1$ dans ce cas) et un mouvement d'entrée correspondant à la ≪ vraie ≫ affaire nouvelle (identifié avec `topent` $= 1$). La ≪ vraie ≫ affaire nouvelle correspond à celle dont la date de souscription est la plus récente (vu que les autres AFN ont été annulées). Une illustration est faite dans le tableau 1.3 suivant. Il présente deux contrats (5434567 et

9065789) dans les bases mouvements annuelles et ces deux contrats ont plusieurs lignes respectivement en affaire nouvelle et résiliation, la ligne correspondante à la vraie AFN ou résiliation étant coloriée en bleu.

Numéro de contrat	`topent`	`topres`	date de souscription	date de résiliation
5434567	1	0	13 Janvier 2015	-
5434567	-1	0	13 Janvier 2015	-
5434567	1	0	18 Janvier 2015	-
9065789	0	1	-	17 Mars 2015
9065789	0	-1	-	17 Mars 2015
9065789	0	1	-	23 Mars 2015

TABLE 1.3 – Tableau présentant les caractéristiques de deux contrats dans les bases mouvements annuelles, le premier étant d'abord une affaire nouvelle annulée puis une vraie affaire nouvelle (5434567) et le deuxième ayant subi une résiliation annulée avant d'être réellement résilié (9065789)

Nous allons alors créer 3 tables à partir des bases mouvements :

- Une table affaires nouvelles : ne contenant que les affaires nouvelles et leurs caractéristiques
- Une table résiliation : ne contenant que les résiliations avec l'âge du contrat et la date de résiliation correspondants.
- Une table qui compte pour chaque contrat, la somme de `topres` et la somme de `topent`. Cette table permet donc d'identifier les affaires nouvelles et voir si elles ont été résiliées ou non grâce à la variable `Resil`.

`Resil` = 1 s'il y a eu résiliation,
`Resil` = 0 sinon.

Joindre ces 3 tables va permettre d'avoir une base d'AFN (affaires nouvelles) dont le numéro de contrat est la clé. Chaque ligne correspond a une AFN, avec les caractéristiques de l'AFN (comme la qualité juridique, le type d'habitation...), l'âge du contrat et la date de résiliation s'il y a bien eu résiliation. Cependant, la base de modélisation est, à ce stade, incomplète.

Les bases mouvements permettent de récupérer des variables comme la qualité juridique de l'AFN, le type d'habitation, le niveau du capital, le nombre de pièces... Cependant, il manque certaines variables contrat pour compléter la base d'étude comme le nombre de sinistres antérieurs à l'affaire nouvelle, le nombre d'enfants... Nous allons donc utiliser les bases périodiques des contrats habitation pour récupérer ces variables à l'affaire nouvelle.

1.2.2 Bases périodiques des contrats habitation

Elles contiennent les caractéristiques des contrats habitation et les différents avenants qu'il y a eu sur ces contrats pendant une période donnée. Les bases périodiques du 1er Janvier au 31 Décembre des années 2015 à 2020 vont être utilisées. Les périodes de modification sont repérées grâce aux dates de début et de fin de modification.
Comme tous les avenants sur le contrat habitation se trouvent dans ces bases, plusieurs lignes peuvent correspondre à un contrat. La ligne correspondante à l'affaire nouvelle est repérée grâce à la variable `topent` qui est la même que celle des bases mouvements. Malheureusement, cette variable présente près de 10% de valeurs manquantes ce qui fait qu'elle ne permet pas d'identifier toutes les AFN. Quand

elle est manquante, les contrats sont triés par date de début de modification du contrat (de la plus ancienne à la plus récente) et les doublons sont supprimés. Grâce à cela, nous sommes sûrs de garder la première vision du contrat dans les bases périodiques, cette vision va contenir les caractéristiques du contrat à l'affaire nouvelle. **Il y a cependant 3% d'affaires nouvelles non trouvées dans les bases périodiques, nous les supprimons alors car les contrats doivent forcément se trouver dans les bases périodiques**. Ces affaires nouvelles pourraient être des affaires nouvelles qui ont été annulées sans que ça le soit spécifier dans les bases mouvements.

Après avoir complété la base d'AFN avec les variables contrats de la base périodique, il manque toujours certaines variables client comme la catégorie socio-professionnelle, le nombre de contrats du client... Les bases clients d'Allianz vont donc être utilisées pour les obtenir.

1.2.3 Bases annuelles client

Ces bases sont des bases annuelles qui contiennent tous les changements sur les caractéristiques du client pendant l'année. Les changements sont enregistrés par période. La période est repérée grâce aux date de début et de fin de modification. Ainsi pour trouver les caractéristiques du client lors de la souscription du contrat habitation de la base AFN créée, nous utilisons la ligne de la base client dont la période de modification contient la date de traitement de l'AFN. Si pour une AFN, cette ligne n'est pas trouvée, la ligne correspondante à la première modification après la date de traitement de l'AFN est utilisée.

Il faut désormais récupérer la variable donnant le parcours client de chaque contrat en utilisant la base parcours.

1.2.4 Bases parcours

Ces bases contiennent le parcours des contrats du 1er Janvier 2015 au 31 Décembre 2020. Le parcours a été déterminé en identifiant les 3 étapes composant la stratégie multi-accès (décrites plus en détail dans la section 1.1.7) :

— *Sourcing* : déterminée en vérifiant si le devis a été fait en ligne ou par le moyen traditionnel.

— Souscription : déterminée en identifiant si le contrat a été souscrit en Web, par plateforme téléphonique ou en agence.

— Gestion : déterminée en vérifiant si le contrat est géré en agence ou en plateforme.

Il y a une ligne pour chaque contrat, il suffit donc de faire la jointure avec le numéro du contrat pour récupérer son parcours.

A cette étape, il ne manque plus que les zoniers pour compléter la base d'AFN.

1.2.5 Bases des zoniers

Sont utilisés le zonier technique et le zonier commercial. Ces zoniers donnent la fréquence estimée des risques couverts suivant l'iris ou le code insee de l'assuré : incendie, dommage électrique, bris de glace, vol, etc.

Une jointure est faite entre la base d'AFN et les 4 dernières bases mentionnées pour obtenir la base de modélisation. Cette base contient 1 244 324 affaires nouvelles. La suppression des affaires nouvelles non trouvées dans les bases périodiques a entrainé une réduction de 3% de la base d'AFN initiale. La base contient les principales variables suivantes.

Variables client

Le tableau 1.4 suivant présente les variables client ainsi que la répartition de leurs modalités et le pourcentage de valeurs manquantes de ces variables. Les valeurs manquantes de la `csp` et de la situation familiale forment la modalité `Inconnu` puisqu'elle dépasse 20%. Celles de la tranche d'âge sont supprimées puisqu'elles sont inférieures à 1%.

Variable	**Description**	**Modalités**	**Répartition des modalités**	**Valeurs manquantes**
`csp`	catégorie socio-professionnelle de l'assuré	- Artisans, commerçants, chefs d'entreprise, Professions intermédiaires - Cadres et professions intellectuelles supérieures - Employés - Ouvriers - Retraités - Autres personnes sans activité professionnelle - Inconnu	10% 6% 19% 10% 9% 13% 33%	0%
`AgeCat`	tranche d'âge du client	Moins de 25 ans (jeunes) Adultes [25 ans, 65 ans] Plus de 65 ans (seniors)	18% 68% 13%	0.66%
`situationFam`	situation familiale de l'assuré	marié célibataire Inconnu	25% 51% 24%	0%
`nb_contrats`	nombre de contrats détenus par le client	1 2 ou plus	42% 58%	0%

TABLE 1.4 – Tableau présentant les variables client

Variables zonier, prime et nombre de pièces

Variable	**Description**	**Valeurs manquantes**
`PrEnt`	prime à l'affaire nouvelle	0%
`nb_Piece`	nombre de pièces assuré	0.01%
`BDG_FREQ`	zonier technique bris de glace	0.6%
`DEL_FREQ`	zonier technique dommage électrique	0.6%
`RC_FREQ`	zonier technique responsabilité civile	0.6%
`VOL_FREQ`	zonier technique vol	0.6%
`DDE_FREQ`	zonier technique dégât des eaux	0.6%
`INC_FREQ`	zonier technique incendie	0.6%
`INC_com`	zonier commercial incendie	0.6%

TABLE 1.5 – Tableau présentant les variables quantitatives (la prime, le nombre de pièces et les variables zonier)

Le tableau 1.5 ci-dessus présente les variables quantitatives (variables zonier, prime et niveau du capital) ainsi que le pourcentage de valeurs manquantes de ces variables. Certaines variables zonier sont présentées dans le tableau B.2 en Annexe B.2. Les variables zonier ont moins d'1% de valeurs manquantes qui sont donc supprimées. Les valeurs manquantes correspondent à des zones Outre-mer ou la Corse. La variable nombre de pièces va être croisée avec la variable niveau du capital pour former la variable `nivCap_nbPiece` puisque le niveau du capital dépend du nombre de pièces.

Variables contrat

Variable	**Description**	**Modalités**	**Répartition des modalités**	**Valeurs manquantes**
`parcours`	parcours client	voir 1.5	voir 1.5	0%
`origine`	origine du contrat	Digitale Non Digitale	8% 92%	0%
`intermediaire`	gérant du contrat	Agence Plateforme	95% 5%	0%
`qualJur`	qualité juridique	Locataire Proprietaire	69% 31%	0%
`typeHab`	type d'habitation	Appartement Maison	62% 38%	0%
`natRes`	nature de la résidence	Principale Secondaire	90% 10%	0%
`franchise`	franchise	Moins de 225 225 et plus	19% 81%	0%
`nivCap`	niveau du capital	1 (moins prestigieux) 2 3 4 (plus prestigieux)	64% 20% 8% 8%	0%
`generation`	année de souscription du contrat	1 (2015) 2 (2016) 3 (2017) 4 (2018) 5 (2019) 6 (2020)	15,33% 17,30% 17,26% 16,95% 16,96% 16,20%	0%
`covid`	résiliations faites pendant la covid	Hors Covid-19 Standard Confinement Couvre-feu	84% 14% 0.5% 1.5%	0%
`objet_val`	pourcentage d'objets de valeur	0% 1% et plus	86% 14%	0%
`code_firme`	réduction sur la prime	1354 Autres Sans firme	53% 27% 20%	0%

TABLE 1.6 – Tableau présentant des variables contrat

Le tableau 1.6 présente les variables contrat qui n'ont pas de valeurs manquantes ainsi que la répartition

de leurs modalités. Précisons que la variable `nivCap` présente des valeurs manquantes mais puisqu'il n'y en a que 5, elles sont très négligeables (et sont supprimées). D'autres variables contrat sont présentés dans le tableau B.1 en Annexe B.2.

Dates et âge du contrat
Il y a aussi dans la base, la date d'effet de l'affaire nouvelle, la date de résiliation du contrat (s'il a bien été résilié), la date de traitement du mouvement et l'âge du contrat.

La suppression des valeurs manquantes a réduit de 0.7% la base de données. Ainsi près de 4% des données initiales ont été supprimées (en prenant en compte les affaires nouvelles supprimées car ne se trouvant pas dans les bases périodiques).
Une fois que la base d'étude complète obtenue, nous pouvons passer à l'analyse descriptive afin de découvrir les caractéristiques de notre portefeuille et l'impact que les variables ont sur le taux de résiliation.

1.3 Analyse descriptive

Rappelons qu'il y a 8 parcours clients de la stratégie multi-accès codés comme suit : la première lettre représente le canal de **souscription**, la deuxième lettre représente l' **origine** digitale ou traditionnelle du devis et la troisième lettre représente l'intermédiaire (le **gérant du contrat**). Ainsi, il y a :
les parcours agence (souscrits en agence)

- ATA (parcours traditionnel) représente les contrats qui viennent d'une agence, souscrits en agence et gérés en agence.
- ADA : qui représente les contrats souscrits et gérés en agence mais d'origine digitale.

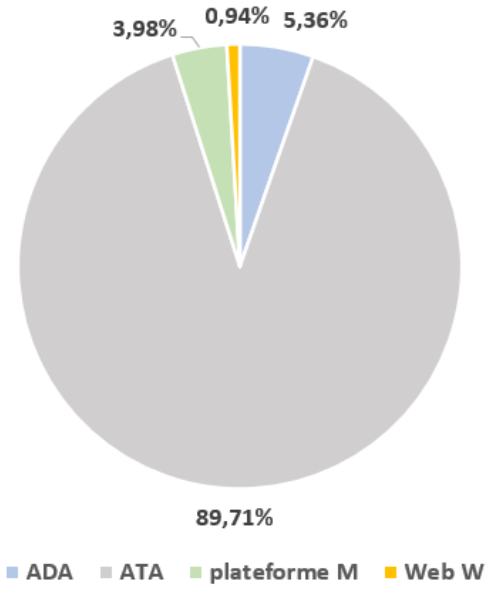
les parcours plateforme (souscrits avec la plateforme téléphonique)

- MDA : représente les contrats souscrits en plateforme, d'origine digitale et gérés en agence.
- MDM : représente les contrats souscrits et gérés en plateforme, d'origine digitale.
- MTA : regroupe les contrats souscrits avec la plateforme téléphonique, d'origine traditionnelle et gérés en agence.
- MTM : regroupe les contrats souscrits et gérés par la plateforme mais venant des agences traditionnelles.

les parcours Web (souscrits sur le Web)

- WDA : regroupe les contrats venant du digital et souscrits en Web mais gérés en agence.
- WDM : regroupe les contrats venant du digital et souscrits en Web mais gérés en plateforme.

Rappelons aussi que notre portefeuille est constitué de 1 244 324 affaires nouvelles (AFN) obtenues du 1er Janvier 2015 au 31 Décembre 2020. La majorité de ces affaires nouvelles se trouvent dans le parcours ATA. La figure 1.5 suivante présente la répartition des affaires nouvelles de notre base de données suivant le parcours. La domination du parcours ATA (parcours traditionnel) s'explique par le fait que c'est le parcours qui existe depuis la création des produits d'assurance habitation. La stratégie multi-accès avec l'assurance habitation a commencé en 2015. Mais tous les parcours n'ont pas démarré cette année.



Parcours	Pourcentage du portefeuille
ADA	5,36%
ATA	89,71%
MDA	0,13%
MDM	1,37%
MTA	0,30%
MTM	2,17%
WDA	0,19%
WDM	0,76%
Total général	**100,00%**

FIGURE 1.5 – Répartition des AFN de notre base de donnée (souscrits du 1er Janvier 2016 au 31 Décembre 2020) suivant le parcours

Nous nous intéressons donc désormais à la date où a eu lieu la première souscription d'affaire nouvelle dans notre périmètre d'étude (du 1er Janvier 2015 au 31 Décembre 2020) suivant le parcours. Elle est présentée dans le tableau 1.7 suivant.

Parcours	Date de souscription de la première affaire nouvelle
ADA	02/01/2016
ATA	01/01/2015
MDA	21/11/2018
MDM	01/01/2015
MTA	26/12/2017
MTM	01/01/2015
WDA	14/04/2019
WDM	12/04/2019

TABLE 1.7 – Tableau présentant la date de la première affaire nouvelle de chaque parcours à partir du 1er Janvier 2015

Ce tableau montre bien que tous les parcours n'ont pas la même ancienneté. Pour les parcours Web WDA et WDM dont la première affaire nouvelle a été souscrite en Avril 2019, l'ancienneté ne représente qu'une année entière, la deuxième année n'étant pas complète. Vu que nous souhaitons modéliser la durée de vie sur une période totale définie, il faudrait normalement avoir la même ancienneté pour tous les parcours. Mais nous ne pouvons pas modéliser la durée de vie sur un an. Nous savons cependant qu'il est possible d'utiliser les autres parcours pour modéliser les durées de vie des parcours qui n'ont qu'une année d'ancienneté. Nous verrons donc plus en détail ultérieurement comment nous comptons résoudre ce problème.

Nous allons maintenant comparer les taux de résiliation par parcours afin d'avoir une idée de leur

valeur mais aussi afin de vérifier si les parcours présentent des taux similaires. La figure 1.6 suivante présentent les taux de résiliation des affaires nouvelles pendant leur première année ainsi que le nombre de contrats dans chaque parcours.

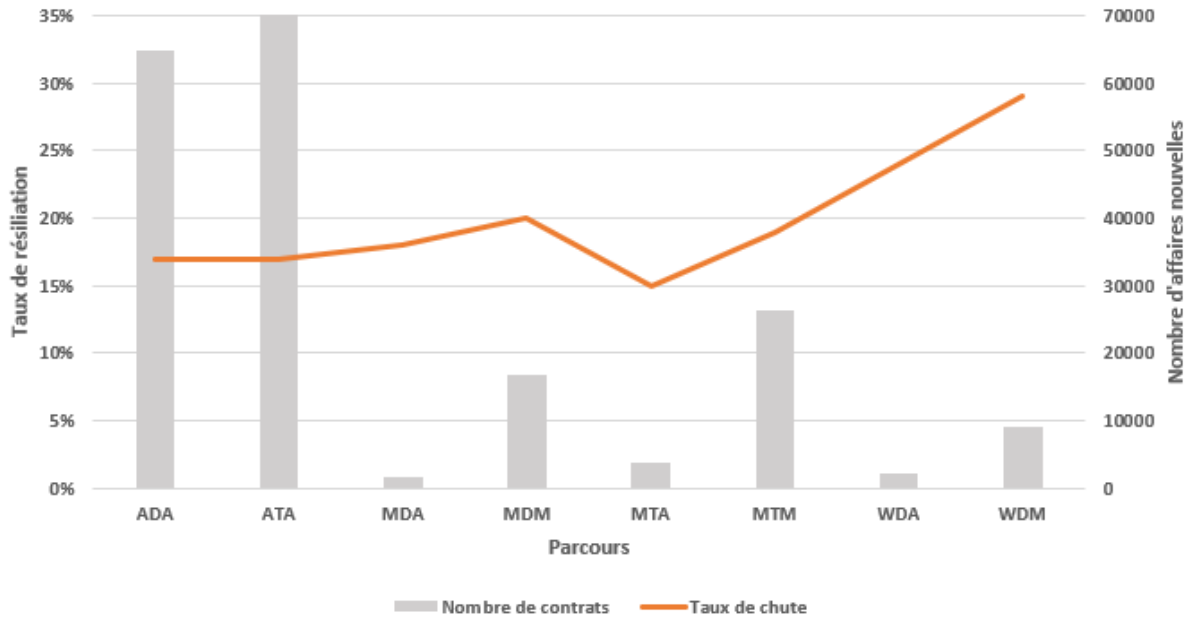


FIGURE 1.6 – Taux de résiliation (et nombre d'affaires nouvelles) pendant la première année suivant le parcours

Nous remarquons que les taux de résiliation par parcours sont différents. Les parcours Web ainsi que le parcours MDM [13] ont les taux de résiliation les plus élevés. Les contrats du parcours MTA [14] ainsi que les contrats souscrits et gérés en agence (ADA, ATA) ont les taux de résiliation les plus faibles.
Qu'est-ce qui pourrait expliquer ces différences ? Un assuré aurait-il plus de chance de résilier en étant dans un parcours plutôt que dans un autre ? Nous avons tendance à penser que les assurés se trouvant en agence ont moins de chance de résilier puisque les agents sont plus aptes à fidéliser le client (grâce à leurs conseils à la souscription, à leurs offres pour encourager le multi-équipement...). Ces différences de taux seraient-ils aussi ou entièrement dues aux profils qui se trouvent dans chaque parcours ?
Afin d'essayer d'avoir une première idée sur la réponse à ces questions, il est nécessaire de déterminer les profils qui se trouvent dans chaque parcours mais aussi d'identifier par la même occasion les profils des clients qui résilient le plus (et donc les variables qui semblent discriminantes pour la résiliation). Une analyse suivant les variables qui différentient le plus les profils va donc être faite. Les taux de résiliation calculés sont tous des taux de résiliation pendant la première année.

Nous commençons notre analyse par la variable qualité juridique : locataire ou propriétaire.

1.3.1 Qualité Juridique

La figure 1.7 suivante présente la répartition des propriétaires/locataires des affaires nouvelles de notre base suivant le parcours. Nous remarquons que les parcours ADA, ATA et MTM sont les parcours

13. contrats d'origine digitale souscrits et gérés en agence
14. souscrits en plateforme et gérés en agence

qui ont plus de 20% de propriétaires. Cela pourrait s'expliquer par le fait que les propriétaires sont généralement orientés en agence tandis que les propriétaires vivant loin d'une agence sont orientés vers le parcours MTM.

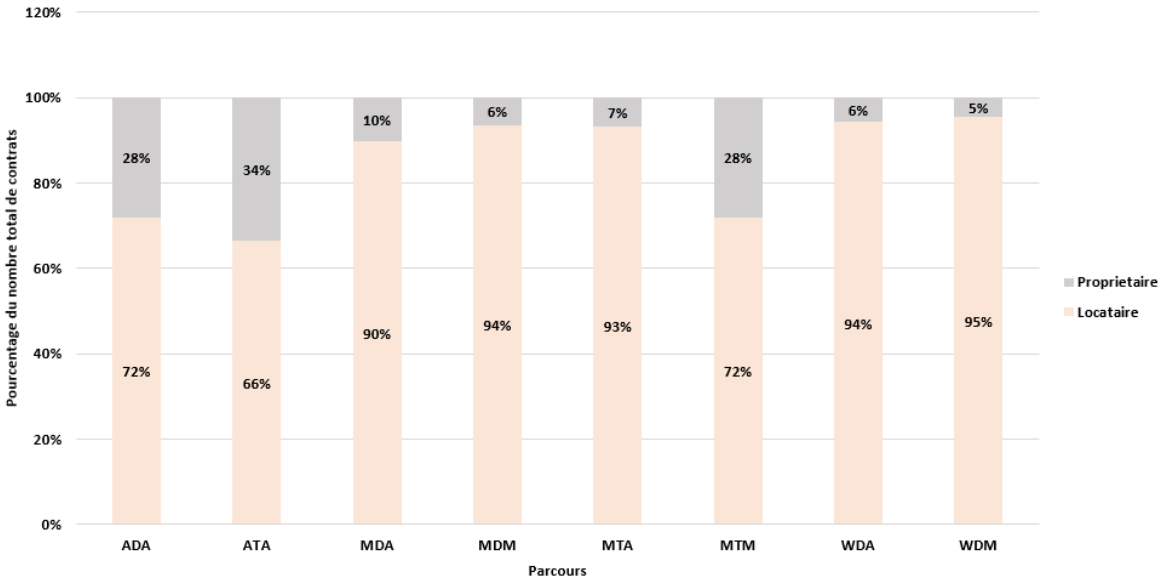


FIGURE 1.7 – Répartition des AFN suivant la qualité juridique et le parcours

La figure 1.8 suivante présente les taux de résiliation pendant la première année des affaires nouvelles de la base suivant le parcours et la qualité juridique.

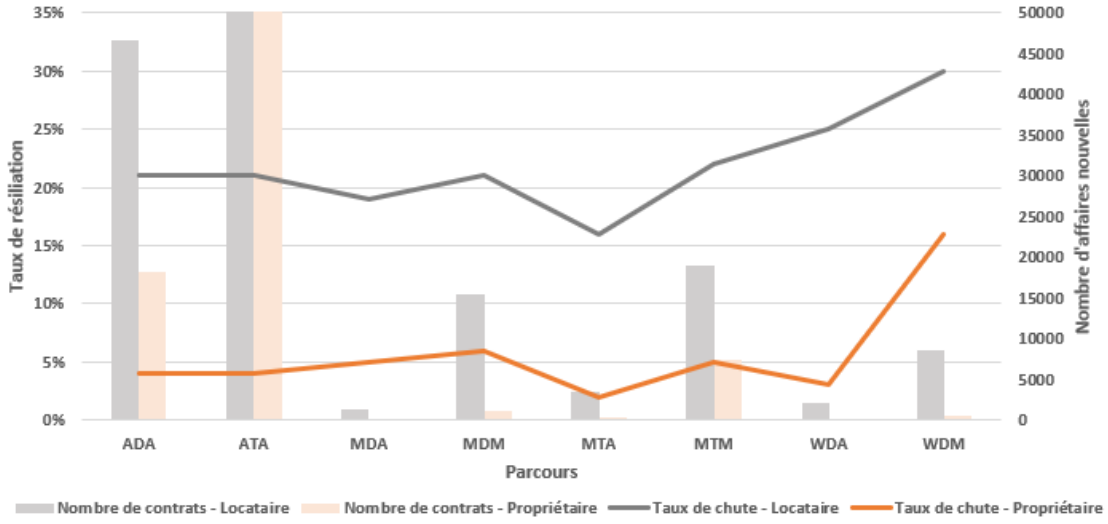


FIGURE 1.8 – Nombre d'AFN et taux de résiliation pendant la première année suivant la qualité juridique et le parcours

Afin de faciliter la visualisation et l'interprétation, la totalité du nombre de contrats du parcours ATA n'est pas affiché puisque le nombre de contrats dans le parcours ATA est élevé par rapport à celui des autres parcours.

Lorsque nous regardons les taux de résiliation pendant la première année, nous remarquons que les taux de résiliation des locataires sont considérablement plus élevés que ceux des propriétaires (différence d'environ 15% quel que soit le parcours). Ainsi, pour les locataires :

— Les contrats intermédiés en agence ont toujours un meilleur taux que ceux non intermédiés.

— Les parcours Web ont des taux de résiliation plus élevés que ceux des autres parcours.

— Le parcours MTA représentant les contrats souscrits en plateforme et gérés en agence semble avoir le taux de résiliation le moins élevé mais il a peu encore de données. Il en est de même pour le parcours MDA. Les parcours agence ATA et ADA ont les taux les plus faibles après ces parcours.

Pour les propriétaires, les parcours Web (WDA et WDM) et les parcours plateformes d'origine digitale (MDA et MDM) et celui d'origine traditionnelle intermédié en agence (MTA) ont des volumes trop faibles pour être interprétés.

Une autre variable permettant de différencier les profils est l'âge du client.

1.3.2 Age du client

Puisque les individus appartenant à la même tranche d'âge ont tendance à avoir les mêmes comportements, l'âge du client est segmenté en 3 catégories :

— Moins de 25 ans : qui correspond aux jeunes.

— [25 ans, 65 ans] : qui correspond aux adultes.

— Plus de 65 ans : qui correspond aux seniors.

Pour faciliter l'interprétation, nous regardons la répartition des catégories socio-professionnelles selon l'origine (digitale ou traditionnelle) du contrat. La figure 1.9 suivante présente la répartition des différentes catégories d'âge du client suivant l'origine (digitale ou non).
En observant la figure 1.9, nous remarquons que ce sont les adultes qui constituent la majorité du portefeuille quel que soit le parcours. Les plus de 65 ans sont les moins présents surtout dans les parcours digitaux où ils ne représentent que 7%.

La figure 1.10 suivante présente le taux de résiliation (et le nombre d'affaires nouvelles) pendant la première année suivant l'âge du client et l'origine.
Nous constatons que les tendances sont les mêmes quelle que soit l'origine du contrat. Ce sont les moins de 25 ans qui résilient le plus. Cette tranche d'âge pourrait être constituée d'étudiants, de jeunes actifs ou de personnes sans activité professionnelle, qui n'ont pas encore une vie stable (déménagement après changement d'université, pour stage, pour premier emploi, pour un autre emploi...) et qui pourraient être sensibles au prix. Ce qui explique pourquoi ils résilient plus. Les seniors sont ceux qui résilient le moins (9% contre 31% pour les jeunes dans le non digital).
Nous avons aussi remarqué que les jeunes sont les seuls qui résilient plus quand ils viennent du non digital. Cela pourrait s'expliquer par le fait qu'en déménageant les jeunes du canal direct ont plus tendance à garder leur assurance habitation (vu que ceux en agence pourraient déménager loin de l'agence).

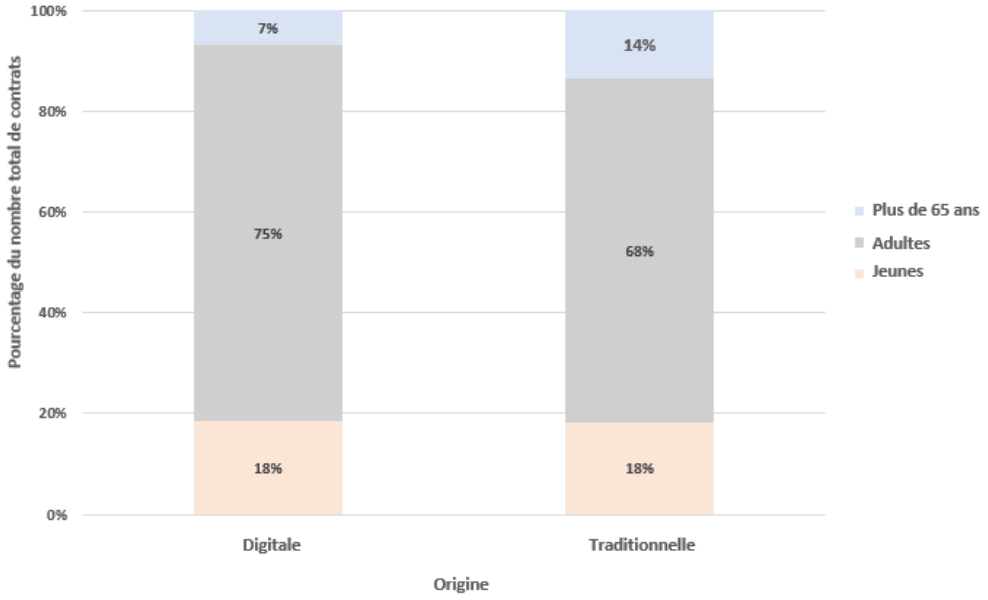


FIGURE 1.9 – Répartition des différentes catégories d'âge du client suivant l'origine du parcours

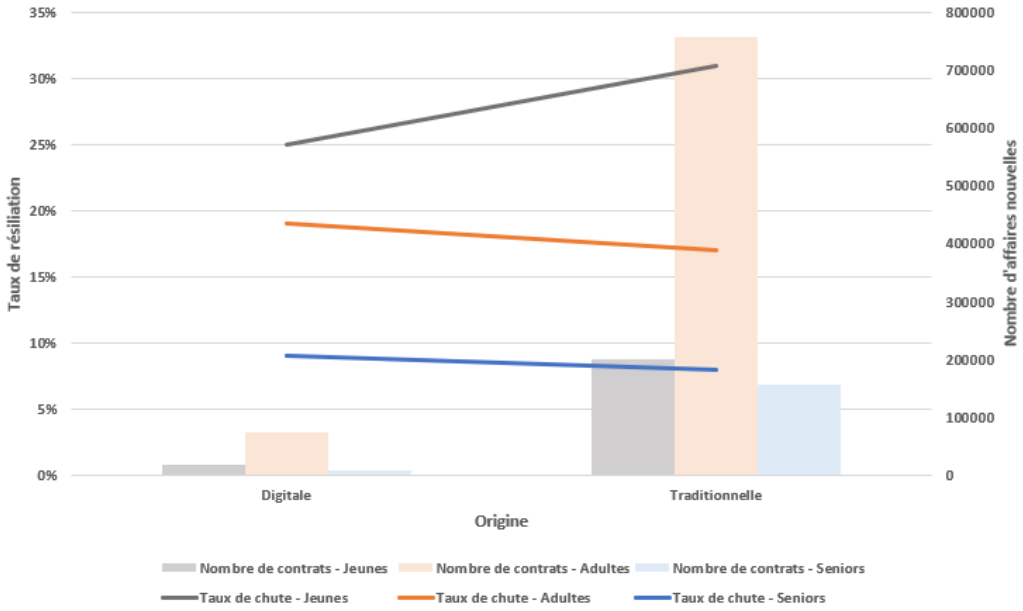


FIGURE 1.10 – Nombre de contrats et taux de résiliation pendant la première année suivant l'origine et l'âge du client

Une autre variable permettant de différencier les profils est le nombre de contrats détenus par le client dans la compagnie d'assurance.

1.3.3 Nombre de contrats du client

L'affaire nouvelle est incluse dans le nombre de contrats. Pour faciliter l'interprétation et pour que chaque modalité soit bien représentée, le nombre de contrats des clients a été réparti en groupes :

— 1 : représente les clients n'ayant que l'affaire nouvelle habitation étudiée comme contrat (mono-détenteur).

— 2 : représente les clients ayant 2 contrats ou plus chez Allianz, le contrat habitation compris (multi-détenteur).

La figure 1.11 présente la répartition du nombre de contrats du client suivant le parcours du client.
Nous remarquons que c'est dans le parcours traditionnel (ATA) et le parcours des contrats souscrits et gérés en agence mais d'origine digitale (ADA) qu'il y a le plus grand pourcentage de clients possédant plus de 2 contrats (respectivement 63% et 59%). La plupart des AFN en assurance habitation en agence se font avec des personnes qui étaient déjà des clients de l'agence, ce qui explique que nous retrouvons plus de clients avec plus d'un contrat dans les parcours agence ADA et ATA. Cela fait partie du processus de fidélisation des agents. Les agents peuvent, par exemple, proposer au client une réduction s'ils prennent leur contrat d'assurance habitation en plus de leur contrat d'assurance automobile chez Allianz.
Dans certains parcours comme les parcours Web (WDA et WDM), il y a peu de clients avec plusieurs contrats (moins de 15%). Le pourcentage de clients avec plus de 2 contrats se trouvant dans le parcours WDM n'atteint même pas 5%. Cela pourrait s'expliquer par le fait que ces parcours sont encore récents mais aussi que les clients se trouvant dans ces parcours sont plus difficiles à multi-équiper.

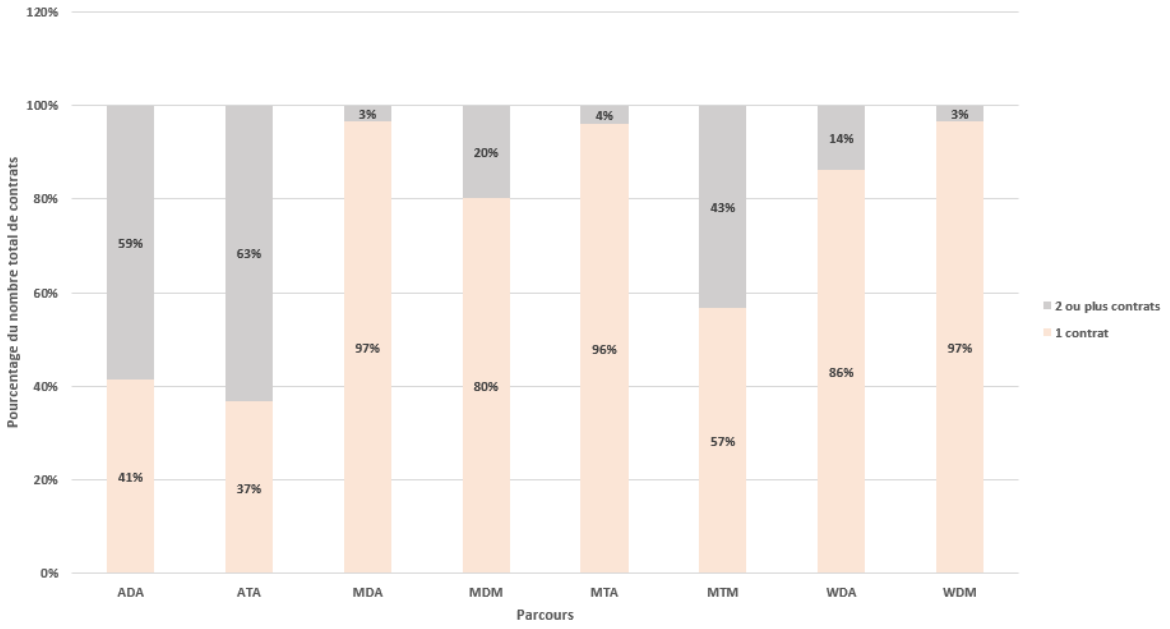


FIGURE 1.11 – Répartition des deux catégories de nombre de contrats du client suivant le parcours

La figure 1.12 ci-dessous présente les taux de résiliation pendant la première année des parcours ADA, ATA et MTM où se trouvent le plus de clients multi-équipés.
En observant les taux de résiliation, nous remarquons que la tendance est la même quel que soit le parcours. Les clients qui ont plusieurs contrats chez Allianz résilient moins que ceux qui n'en ont qu'un (différence d'environ 6%). Cela pourrait s'expliquer par le fait que :

— le client pourrait avoir des réductions liés à son nombre de contrats. Ainsi, résilier son contrat habitation pourrait lui enlever une ou des réduction(s), ce qui l'en dissuaderait.
— le client est peut-être suffisamment satisfait des services d'Allianz, ce qui l'a poussé à prendre un autre contrat. Cette satisfaction pourrait le pousser à ne pas résilier.

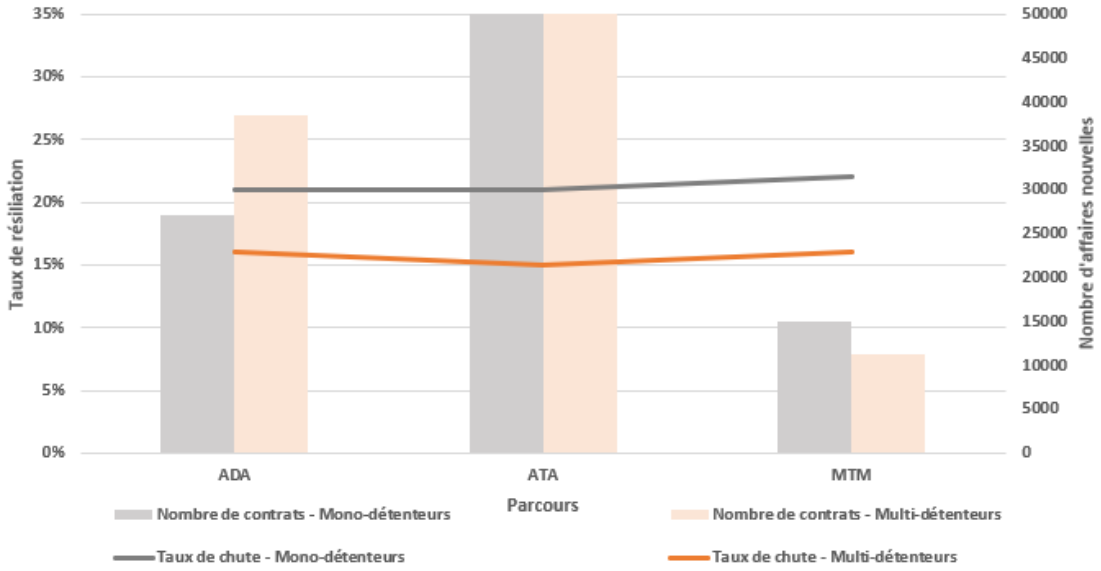


FIGURE 1.12 – Taux de résiliation (et nombre d'affaires nouvelles) suivant le nombre de contrats du client pendant la première année et suivant le parcours

Deux autres variables qui permettent aussi de différencier les profils sont le type d'habitation et la catégorie socio-professionnelle du client. L'analyse suivant ces variables a été mise en Annexe B.3. Les différentes analyses ont permis de voir que les parcours présentent des profils différents. C'est dans les parcours agence (ATA et ADA) que se trouvent le plus de profils moins susceptibles de résilier, ce qui pourrait expliquer les taux de résiliations par parcours.
Nous voulons finalement vérifier si les taux de résiliation sont constants à partir d'une période car si c'est le cas, il ne faut modéliser les taux de résiliation que pendant cette période pour prédire les durées de vie.

1.3.4 Age du contrat

La figure 1.13 suivante présente les taux de résiliation suivant l'âge du contrat (en année).
En comparant les taux de résiliation suivant l'âge du contrat, nous remarquons qu'ils diminuent un peu plus chaque année. Il passe de 18% pour la première année à 10% pour la quatrième année dans le portefeuille. La baisse des taux pourrait s'expliquer par le fait qu'il y ait de moins en moins de profils susceptibles de résilier au fil des années. Par exemple, les locataires ont plus tendance à résilier que les propriétaires, ce qui entraîne qu'il y aura moins de locataires l'année suivante. Cela pourrait entraîner une baisse des taux de résiliation puisque le pourcentage de propriétaires va augmenter, les propriétaires ayant moins tendance à résilier.

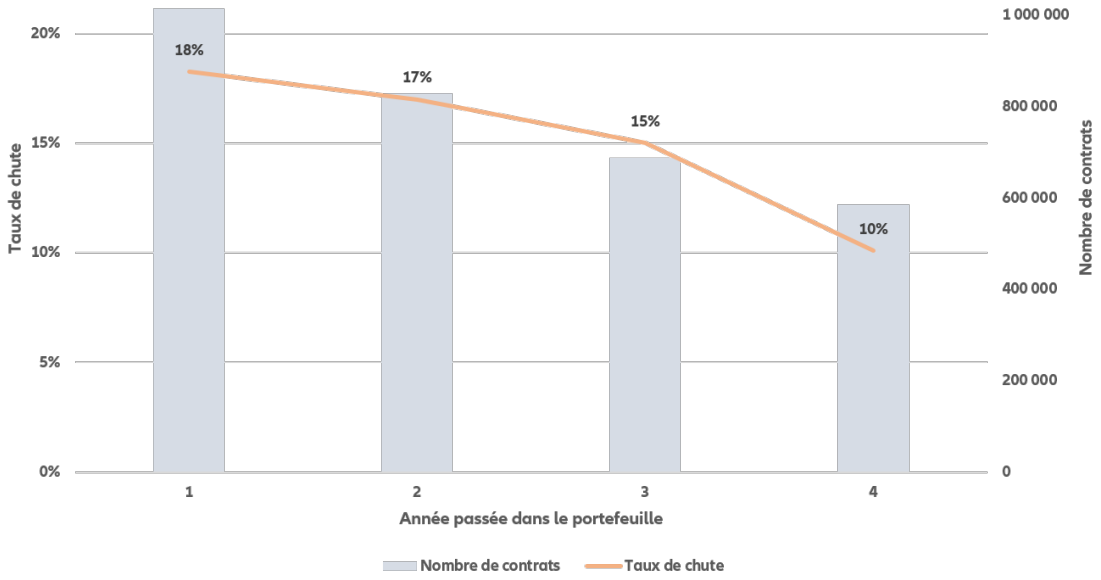


FIGURE 1.13 – Taux de résiliation suivant l'âge du contrat en année et nombre de contrats au début de chaque année passée dans le portefeuille

En faisant l'analyse descriptive, nous avons aussi remarqué que les taux de résiliation obtenus suivant une variable donnée pourraient s'expliquer par une corrélation entre cette variable et une ou d'autres variables. Par exemple, les taux de résiliation pour certaines catégories socio-profesionnelles pourraient s'expliquer par le fait qu'il y n'ait que des locataires dans ces catégories. Lorsque des variables sont corrélées, une des variables est donc redondante, il faudrait l'éliminer du modèle pour respecter la parcimonie.
Il est donc nécessaire de faire une étude de corrélation afin de pouvoir supprimer les variables redondantes pour notre modélisation.

1.4 Etude de corrélation

Afin d'étudier la corrélation entre les variables de la base de modélisation, nous allons utiliser la matrice de corrélation pour déterminer la corrélation entre les variables quantitatives ainsi que le V de Cramer (voir Annexe A.1) pour déterminer la corrélation entre les variables catégorielles.

1.4.1 Règle de décision

La corrélation est supposée forte à partir de 60% [15]. Les parcours sont corrélés à plus de 60% à l'origine et à l'intermédiaire. Ces deux variables vont donc être supprimées afin de garder les parcours. Lorsqu'une variable est corrélée à plusieurs autres variables, cette variable est gardée et les autres éliminées. Elle est la variable qui contient le plus d'informations par rapport aux autres.
Lorsque deux variables sont corrélées, l'une des deux va être supprimée. Nous choisissons celle qui nous semble la plus pertinente à garder en utilisant nos connaissances de l'assurance habitation. Par exemple, entre la variable `ext_dommages` [16] et la variable `typeHab` [17], nous allons choisir la variable `typeHab` puisqu'elle est plus facile à interpréter.

15. voir Annexe A.1
16. permettant de voir s'il y a des extensions de garantie dans le contrat
17. type d'habitation

1.4.2 Visualisation des corrélations

La visualisation des corrélations est obtenue grâce au package *seaborn* de python.

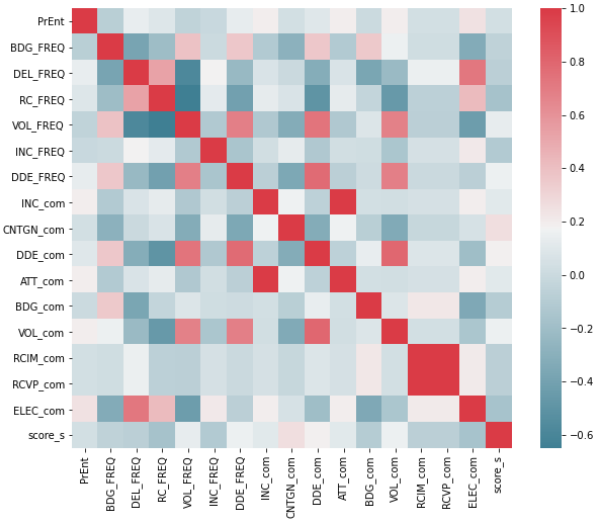


FIGURE 1.14 – Figure représentant la matrice de corrélation des variables quantitatives : prime (`PrEnt`) et variables zonier

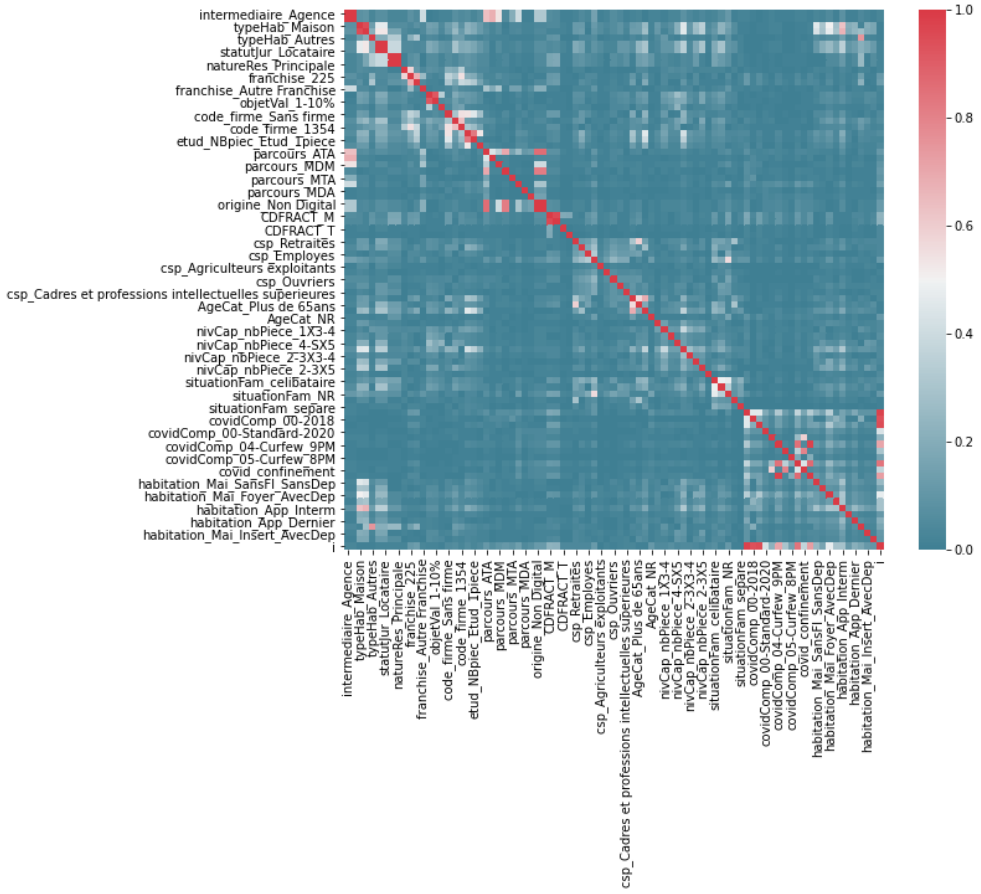


FIGURE 1.15 – Figure représentant le V de Cramer des variables qualitatives

La figure 1.14 ci-dessus présente la visualisation de la matrice de corrélation des variables quantitatives : prime (`PrEnt`) et les différents zoniers. En observant la visualisation de la matrice de corrélation des variables quantitatives, nous remarquons qu'elles sont bien corrélées entre elles. Ce n'est pas surprenant vu que ce sont presque toutes des variables zonier. Nous allons donc en éliminer plusieurs en suivant notre règle de décision.

La figure 1.15 ci-dessus présente la visualisation du V de Cramer des variables qualitatives. En observant la visualisation du V de Cramer des variables qualitatives, nous remarquons qu'elles sont indépendantes dans l'ensemble (vu la dominance du bleu) mais il y a quelques corrélations. Nous allons donc éliminer certaines variables en suivant notre règle de décision.

1.4.3 Variables à éliminer

En suivant notre règle de décision, nous allons donc éliminer les variables suivantes, entre autres :
- La variable `RAN` (option remboursement à neuf) corrélée à plus de 60% à la variable `dEL` (dommages éléctriques). Cela pourrait s'expliquer par le fait que les assurés ayant la garantie dommages éléctriques ont aussi la garantie remboursement à neuf.
- La variable `ext_dommages` corrélée à plus de 60% à la variable `locataire`[18].
- La variable `origine` et `intermediaire` corrélée à 80% à la variable `parcours`. Puisque nous souhaitons voir l'impact des parcours, nous allons garder la variable `parcours`.
- plusieurs variables des zoniers (presque 40%). Nous avons donc garder les zoniers bris de glace, dommages électriques, responsabilité civile, responsabilité civile vie privée, incendie, catastrophe naturelle, dégât des eaux et le score de sécheresse.

Ainsi près de 20% des variables ont été éliminées suite à l'étude des corrélations.

1.4.4 Variables à croiser

Les modalités des variables qualité juridique `qualJur` (`locataire/proprietaire`) et type d'habitation `typeHab` (`maison/appartement`) sont corrélées à plus de 50%. Puisque nous souhaitons les garder toutes les deux, nous les regroupons donc en une seule variable avec 4 modalités :
`Appartement_locataire`, `Appartement_proprietaire`, `Maison_locataire`, `Maison_proprietaire`.
Croiser les variables permet de prendre en compte l'interaction entre les variables concernées. Il peut donc permettre d'améliorer les performances du modèle surtout quand c'est un GLM (Modèle Linéaire Généralisé).

Notre objectif est de modéliser la durée de vie des contrats d'assurance habitation afin d'optimiser la rentabilité de la stratégie multi-accès. Après avoir vu plus en détail le contexte, la création de la base et avoir fait l'analyse descriptive de la base, nous allons entamer la modélisation dans les parties suivantes.

18. Suite à la transformation des variables qualitatives en variables numériques, chaque modalité de ces variables devient une variable à part entière. La variable `typeHab` donne alors les variables `proprietaire` et `locataire`.

Chapitre 2

Modélisation des taux de résiliation

Afin de pouvoir déterminer les profils qui durent le plus longtemps dans notre portefeuille tout en observant l'impact des variables sur la résiliation à chaque échéance, nous voulons modéliser les durées de vie des contrats dans le portefeuille. Nous disposons des affaires nouvelles de 2015 à 2020, soit 5 ans d'ancienneté. Nous verrons plus tard pourquoi nous avons arrêté notre modélisation à 4 ans.
Nous souhaitons donc modéliser la résiliation ou la survie d'un contrat d'assurance pendant plusieurs périodes données. La durée totale des périodes doit former 4 ans, ce qui va permettre de savoir, au bout de 4 ans, quels contrats ont été résiliés et quels contrats ont survécu. Une explication détaillée du processus de calcul des durées de vie en utilisant les modélisations des taux de résiliation pendant plusieurs périodes sera donnée dans la troisième partie.

Nous allons d'abord choisir comme période de modélisation **1 an**. Nous souhaitons également voir l'impact de certaines variables, comme le parcours, pour chaque année où le contrat a survécu dans le portefeuille. Nous allons ensuite essayer de trouver une période de modélisation plus petite afin d'obtenir des durées beaucoup plus précises.

La variable à expliquer pour chaque période est donc

$Y = 1$ le contrat a été résilié à la fin de la période de modélisation,
$Y = 0$ sinon.

Puisque nous souhaitons prédire la durée de vie du contrat à l'affaire nouvelle, nous allons évidemment utiliser les variables qui peuvent expliquer la résiliation et ces variables doivent être disponibles à l'affaire nouvelle. Cela entraîne que pour chaque période de modélisation, nous gardons les mêmes variables explicatives même si nous savons que plusieurs variables peuvent changer suite à un avenant (changement du niveau du capital par exemple) ou autre (majoration de la prime).
Les variables explicatives sont donc celles qui constituent la base de modélisation construite. Les variables inutiles sont suprimées de la base. Celles-ci représentent principalement les dates d'effet de l'affaire nouvelle, de résiliation, de traitement du mouvement... (et ne sont donc pas utiles pour la modélisation).

Nous allons utiliser le package $h2o$ [1] (H2O, 2021) sur python (Van Rossum et Drake, 2009) pour construire les modèles.

Avant de procéder à la modélisation, il faut se rappeler que pendant l'analyse descriptive, nous avions remarqué certains éléments qui pourraient poser problème lors de la modélisation. Il est donc

1. défini en Annexe A.6

important de revoir en détail ces éléments et comment nous pourrions éviter ou résoudre les problèmes qu'ils pourraient engendrer.

2.1 Problèmes identifiés et solutions envisagées

Le premier problème décelé concerne l'ancienneté de certains parcours. C'est ce qui nous a principalement poussé à faire l'estimation des durées de vie par la modélisation des taux de résiliation par période.

2.1.1 Ancienneté de certains parcours inférieure à 4 ans

Le tableau 2.1 suivant montre la date de souscription de la première affaire nouvelle sur le périmètre d'étude (1er Janvier 2015 au 31 Décembre 2020) et l'ancienneté (en année entière) du parcours au 31 Décembre 2020.

Parcours	Date de souscription de la première affaire nouvelle	Ancienneté
ADA	02/01/2016	4 ans
ATA	01/01/2015	5 ans
MDA	21/11/2018	2 ans
MDM	01/01/2015	5 ans
MTA	26/12/2017	3 ans
MTM	01/01/2015	5 ans
WDA	14/04/2019	1 an
WDM	12/04/2019	1 an

TABLE 2.1 – Tableau présentant la date de souscription de la première affaire nouvelle dans chaque parcours ainsi que l'ancienneté en année complète de chaque parcours

Nous remarquons que les parcours Web n'ont qu'un an d'ancienneté alors que nous souhaitons modéliser les durées de vie sur le plus grand intervalle de temps possible. D'autres parcours ont aussi moins de 5 ans d'ancienneté alors que nous souhaitons modéliser tous les parcours sur la même durée. Cela poserait problème dans le cas où les parcours auraient un impact sur la résiliation. Dans ce cas, il faudrait donc essayer de trouver des pistes afin de modéliser les durées de vie de ces parcours pendant le même horizon que les autres parcours. Pour cela, deux pistes auraient pu être exploitées, elles sont décrites en Annexe A.8.

Le deuxième problème relevé lors de l'analyse descriptive est la sous-représentation de certains parcours et la dominance du parcours traditionnel ATA.

2.1.2 Déséquilibre de la représentation des parcours

Comme déjà remarqué pendant l'analyse descriptive (section 1.3), il y a un déséquilibre entre les volumes de données des différents parcours. Le parcours traditionnel ATA représente à lui seul 90% du portefeuille tandis que certains parcours ont un volume de données faible. Nous allons donc regrouper les parcours ayant des caractéristiques communes et des taux de résiliation proches, c'est-à-dire les :
— parcours Web WDA et WDM en W,
— parcours plateforme intermédié en agence MTA et MDA en MA,
— parcours plateforme sans intermédiaire MTM et MDM en MM.

Nous n'allons aussi pas manquer de vérifier les performances du modèle suivant le parcours afin de nous assurer que le modèle n'est pas performant qu'avec le parcours le plus représenté (ATA).

Un autre problème soulevé dans la première partie est le déséquilibre entre les deux classes du modèle de classification.

2.1.3 Déséquilibre des classes

Lors de l'analyse descriptive (sous-section 1.3.4), nous avions remarqué avec la figure 1.13 que les taux de résiliation n'étaient pas élevés et ne dépassaient pas 18% ce qui entraine que les deux classes de modélisation sont déséquilibrées. La classe 1 : résiliation est la classe minoritaire.
La figure 1.13 montre bien que les classes sont déséquilibrées et que le déséquilibre s'accroît un peu plus chaque année passée par le contrat dans le portefeuille puisque les taux passent de 18% à 10%. Ce déséquilibre de classe entre les classes majoritaires et minoritaires pourrait biaiser les performances prédictives des algorithmes de machine learning vers la classe majoritaire. En d'autres termes, cela pourrait résulter sur des modèles qui ne prédisent bien que l'appartenance à la classe majoritaire et qui vont ignorer la classe minoritaire.

Il existe plusieurs moyens pour essayer de résoudre ce problème.

Méthodes de ré-échantillonnage

Il consiste à modifier l'ensemble de données utilisé afin d'avoir des données plus équilibrées avant de procéder à la modélisation. Il existe deux catégories de méthodes :

— **Le sous-échantillonnage (*Under Sampling*)** permet de diminuer le nombre d'observations de la (des) classe(s) majoritaire(s) dans le but d'obtenir un ratio classe minoritaire/classe majoritaire satisfaisant (Branco et al., 2016). L'inconvénient du sous-échantillonnage est qu'il pourrait supprimer les données potentiellement utiles. Mais il peut être très utile quand la taille de l'ensemble de données est suffisamment grande. Vu que notre base de données compte plus d'un million de lignes et que le parcours ATA représente à lui tout seul 90% des données, il est possible de sous-échantilloner les données en supprimant aléatoirement les observations de la classe majoritaire se trouvant dans le parcours ATA. Cela permettrait à la fois, de réduire le déséquilibre des classes mais aussi la prépondérance du parcours ATA qui constitue le premier problème évoqué dans la sous-section 2.1.2.

  Au lieu de diminuer le nombre d'éléments dans la classe majoritaire, celui de la classe minoritaire pourrait être augmenté grâce à un sur-échantillonnage.

— **Le sur-échantillonnage (*over Sampling*)** permet d'augmenter le nombre d'observations de la (des) classe(s) minoritaire(s) dans le but d'obtenir un ratio classe minoritaire/ classe majoritaire satisfaisant. Il consiste soit à échantillonner chaque membre de la classe minoritaire avec remise (*random over sampling*), soit à créer des observations synthétiques par échantillonnage aléatoire à partir de l'ensemble des variables explicatives (SMOTE - *Synthetic Minority Over-sampling Technique*).
  Ces deux méthodes de sur-échantillonnage sont présentées dans le tableau 2.2 suivant.

Random Over Sampling	**SMOTE (*Synthetic Minority Over Sampling Technique*)**
Il consiste à échantillonner les données avec remise afin d'augmenter le nombre d'éléments dans la classe minoritaire. Le principal inconvénient de cette méthode est qu'il ne fait que dupliquer des exemples existants ce qui augmente le risque de sur-ajustement (*overfitting*) comme vu dans l'étude de Branco et al. (2016). En sur-échantillonnant l'ensemble de données, il est assez courant que le modèle génère une règle de classification pour couvrir un seul exemple répliqué. Un deuxième inconvénient est qu'il augmente le nombre d'observations de la classe minoritaire, augmentant ainsi le temps d'apprentissage. Cet inconvénient se retrouve dans toutes les méthodes de sur-échantillonnage.	SMOTE, introduit dans l'article de Chawla et al. (2002), permet de créer des observations synthétiques par échantillonnage aléatoire à partir de l'ensemble de données. Il vise à équilibrer la distribution des classes en augmentant les individus de classes minoritaires. SMOTE génére de nouveaux échantillons par interpolation linéaire en utilisant les données de la classe minoritaire et celles de leurs proches voisins. Ces observations synthétiques sont obtenus en sélectionnant au hasard un ou plusieurs des k plus proches voisins pour chaque exemple dans la classe minoritaire. Voici le détail du processus en 3 étapes : **Étape 1 :** Définition de l'ensemble de classes minoritaires A. Pour chaque $x \in A$, les k plus proches voisins de x sont obtenus en calculant la distance euclidienne entre x et tous les autres échantillons de l'ensemble A. **Étape 2 :** La fréquence d'échantillonnage N est fixée en fonction de la proportion déséquilibrée. Elle peut aussi être donnée. Pour chaque $x \in A$, N exemples (c'est-à-dire $x_1, x_2, \ldots x_N$) sont choisis au hasard parmi ses k plus proches voisins, ils définissent l'ensemble A_1 . **Étape 3 :** Pour chaque point $x_k \in A_1$, $(k = 1, 2, 3 \ldots N)$, la formule suivante est utilisée pour générer un nouveau point $$x' = x + rand(0,1) * \mid x - x_k \mid, \qquad (2.1)$$ $rand(0,1)$ représente le nombre aléatoire compris entre 0 et 1.

Table 2.2 – Tableau présentant deux techniques de sur-échantillonnage

Dans leur article, Weiss et al. (2007) ont fait une série d'expériences afin de déterminer la meilleure des méthodes notamment entre le sur-échantillonnage et le sous-échantillonnage. Sur la base des résultats de tous les ensembles de données, il n'y a pas un gagnant unanime entre ces 2 méthodes. Il faut choisir la technique la plus adaptée suivant les caractéristiques de l'ensemble de données. Ils ont cependant remarqué que le SMOTE pourrait apporter des améliorations par rapport au *Random Over Sampling*.
Ainsi, parmi les méthodes que nous venons d'évoquer afin d'essayer de résoudre le problème engendré par le déséquilibre des classes, vont être testées :
— une méthode de sur-échantillonnage : SMOTE,
— une méthode de sous-échantillonnage : *Random Under Sampling* avec uniquement les contrats de la classe majoritaire appartenant au parcours sur-représenté ATA. Cela va aussi permettre de mieux équilibrer la représentation des parcours.

Modification du seuil de prédiction

Les algorithmes de classification de machine learning prédisent une probabilité ou un score d'appartenance à une classe afin de pouvoir déterminer dans quelle classe l'individu va appartenir. Ils utilisent donc un seuil qui pourrait être par défaut 0,5 lorsqu'il y a 2 classes. Ainsi, tous les individus dont les probabilités ou scores d'appartenance sont supérieur(e)s ou égal(e)s au seuil sont considérés comme appartenant à une classe et tous les autres individus sont considérés comme appartenant à l'autre classe. Pour les problèmes de classification qui présentent un déséquilibre de classes, le seuil par défaut pourrait donner de mauvaises performances. Une approche simple et efficace pour améliorer les performances d'un classificateur sur un problème de classification déséquilibrée consiste à régler le seuil utilisé pour définir à quelle classe va appartenir l'individu.
Lors de la modélisation, nous allons utiliser le package *h2o* (défini en Annexe A.6) sur python. Pour les problèmes de classification binaire, le seuil de prédiction utilisé avec *h2o* est celui qui va permettre de maximiser le score F1 pour l'ensemble de données. Le score F1 [2] fournit une mesure de la capacité d'un classificateur binaire à bien classer les cas positifs étant donné une valeur seuil. Un score F1 égal à 1 signifie que le modèle a correctement identifié toutes les résiliations et n'a pas marqué une non-résiliation comme une résiliation. Si le modèle peine à à identifier toutes les résiliations en plus de classer plusieurs non-résiliations comme étant des résiliations, cela se traduira par un score F1 plus proche de 0. Sa formule de calcul est mis en Annexe A.6.2.

Une étape importante lors d'une modélisation consiste à évaluer différents modèles les uns par rapport aux autres. Choisir la mauvaise mesure d'évaluation ou ne pas comprendre ce que signifie réellement les métriques pourrait amener à ne pas garder le modèle le plus performant. Nous allons donc voir les différentes métriques qui seront utilisées pour évaluer la qualité des modèles.

2.2 Métriques permettant d'évaluer les modèles

Il existe plusieurs métriques pour évaluer les performances de modèles classificateurs. Il ne faut pas oublier de prendre en compte le fait que les deux classes sont déséquilibrées et qu'il faut utiliser des métriques pertinentes par rapport à ce déséquilibre (He et Garcia, 2009).

2.2.1 Précision et Rappel

En travaillant avec un ensemble de données déséquilibré, il est toujours utile d'utiliser une matrice de confusion afin d'évaluer le modèle d'apprentissage. Elle permet de compter le nombre de vrais positifs [3] (TP), vrais négatifs [4] (TN), faux positifs [5] (FP) et faux négatifs [6] (FN). Le vrai positif (TP) désigne une prédiction que le classificateur a correctement prédite comme étant dans la classe positive. Le vrai négatif (TN) désigne une prédiction que le classificateur prédit correctement comme étant dans la classe négative. Le faux positif (FP) désigne une prédiction que le classificateur prédit à tort comme étant dans la classe positive alors qu'elle est dans la classe négative. Le faux négatif (FN) désigne une prédiction où le classificateur prédit à tort la classe positive comme étant négative.
Ces éléments permettent de déterminer différentes mesures de performance comme :

2. Performance and prediction, h2o
3. en anglais *True Positive*
4. en anglais *True Negative*
5. en anglais *False Positive*
6. en anglais *False Negative*

— la **précision** : elle indique quelle portion des prédictions en tant que classe positive sont réellement positives. Elle se calcule comme suit

$$\text{Précision} = \frac{TP}{TP + FP}. \tag{2.2}$$

— le **rappel** : il indique quelle fraction de tous les échantillons positifs est correctement prédite comme positive par le classificateur. Il est également connu sous le nom de taux de vrais positifs (TPR), sensibilité, probabilité de détection. Le rappel se calcule comme suit

$$\text{Rappel} = \frac{TP}{TP + FN}. \tag{2.3}$$

— l'***accuracy*** : elle indique la fraction de prédictions correctes faites par le modèle.

$$Accuracy = \frac{TP + TN}{TN + FP + FN + TN}. \tag{2.4}$$

Une des métriques les plus utilisées pour évaluer un modèle de classification est l'*accuracy* parfois appelée précision en français. Cependant, elle n'est pas très pertinente pour évaluer les performances avec des classes déséquilibrées. Prenons le cas d'un ensemble de données déséquilibré avec un déséquilibre de classe de ratio 1 :100. Dans ce problème, pour chaque exemple de la classe minoritaire (classe 1) correspond 100 exemples pour la classe majoritaire (classe 0). Ainsi, un modèle qui prédit la classe majoritaire (classe 0) pour tous les exemples de l'ensemble de test aura une *accuracy* de classification de 99%, ce qui parait excellent alors que le modèle n'a prédit aucun élément de la classe minoritaire correctement. Dans notre cas et dans plusieurs cas comme la détection de fraude, de spam par exemple, il est très important que le modèle puisse bien classer les contrats appartenant à la classe minoritaire (pour nous les résiliations). Utiliser l'*accuracy* comme métrique n'est pas vraiment appropriée car l'impact des individus les moins représentés, mais les plus importants, est réduit par rapport à celui de la classe majoritaire.
Il est donc plus pertinent d'utiliser la précision et le rappel afin de déterminer quelle fraction des résiliations réellement observées ont été prédites par le modèle (rappel) et quelle fraction des résiliations prédites par le modèle sont observées réellement (précision). Ainsi, dans un monde parfait, nous voudrions un modèle qui a une précision égale à 1 et un rappel égal à 1. Ceci n'est pas fréquemment le cas pour un modèle de *machine learning* et il y a souvent un compromis entre ces deux métriques. En effet, il existe une relation inverse générale entre elles. Cela signifie que si l'une augmente, l'autre diminue le plus souvent.

En plus de ces éléments, il y a d'autres outils qui aident à l'évaluation des performances des modèles prédictifs de classification binaire.

2.2.2 AUC

L'aire sous la courbe AUC correspond à l'intégrale de la fonction ROC (voir Annexe A.2). Ses valeurs sont comprises entre 0 et 1. Ainsi, une AUC de 0,5 signifie que la probabilité qu'une instance positive se classe plus haut qu'une instance négative est de 0,5 et donc aléatoire. Un classificateur parfait classerait toujours une instance positive plus haut qu'une instance négative et aurait une AUC égale à 1 d'après He et Garcia (2009).

Suivant la valeur de l'AUC obtenue, la qualité du modèle peut être interprétée comme suit :

— $AUC = 0.5$: mauvaise.

— $AUC \in]50; 60]$: médiocre.

— $AUC \in [70; 80[$: acceptable.
— $AUC \in [80; 90[$: excellente.
— $AUC \geq 90$: exceptionnelle.

2.2.3 AUCPr (Aire sous la courbe de Précision-Rappel)

Lorsqu'il y a un déséquilibre dans les observations entre les deux classes, analyser à la fois la précision et le rappel (appelé *recall* en anglais) est utile. Une courbe de précision rappel (*Precision-Recall curve* en anglais) est un graphe ayant la précision comme axe des ordonnées et le rappel comme axe des abscisses pour différents seuils. Il présente donc des similarités avec la courbe ROC [7]. Il peut être utilisé afin de déterminer un seuil donnant le compromis souhaité entre le rappel et la précision. La figure 2.1 suivante présente une courbe de précision rappel (*Precison-Recall curve*) où il est possible de voir la relation entre le rappel et la précision suivant les seuils k et m.

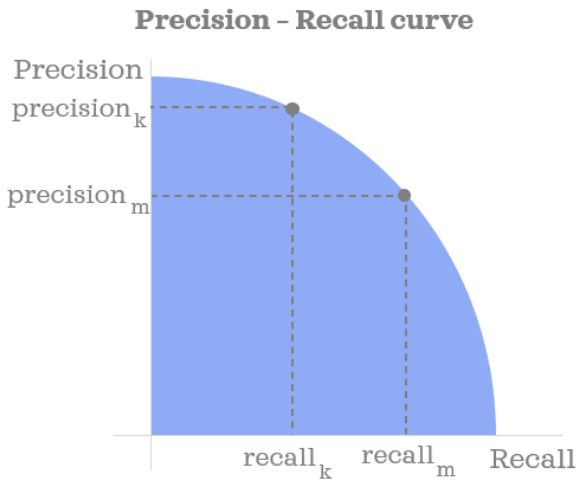


FIGURE 2.1 – Exemple de courbe Précision-Rappel (*Precision-Recall Curve* en anglais)

Nous remarquons que le seuil m donne un meilleur rappel mais une précision plus faible que le seuil k qui donne donc une meilleure précision mais un rappel plus faible.
L'AUCPr représente donc l'aire sous la courbe de précision-rappel. Nous la voyons en bleu sur la figure 2.1. Elle n'est cependant pas facile à interpréter. La meilleure AUCPr est celle qui est la plus proche de 100%.

La dernière métrique est un peu différente des métriques déjà définies puisqu'elle a une approche probabiliste. Elle reste cependant tout aussi efficace pour évaluer les performances des modèles de classification.

2.2.4 LogLoss

Le logloss n'est pas un concept intuitif comme la précision par exemple. Il s'obtient grâce à la formule suivante

7. voir Annexe A.2

$$H_p(q) = \frac{-1}{N}\sum_{i=1}^{N} y_i \log(p(y_i)) + (1-y_i)\log(p(1-y_i)), \tag{2.5}$$

où y est la variable à expliquer et $p(y)$ est la probabilité prédite que y soit égal à 1 (c'est-à-dire qu'il y ait résiliation dans notre cas) pour tous les N points de la base.
Cette formule indique que, pour chaque résiliation (y=1), $log(p(y))$ est ajouté, c'est-à-dire le log de probabilité que le contrat soit résilié. Inversement, $log(1-p(y))$ est ajouté, c'est-à-dire la probabilité log qu'il ne soit pas résilié, pour chaque contrat non résilié (y=0). Ainsi, **plus le logloss est petit, meilleur est le modèle**.

Nous pouvons désormais amorcer la modélisation.

2.3 Modélisation des taux de résiliation par période d'une année

2.3.1 GLM

Nous utilisons tout d'abord le Modèle Linéaire Généralisé qui est connu pour la facilité d'interprétation des prédictions grâce aux coefficients. C'est pourquoi c'est l'un des modèles les plus utilisés en Actuariat. Le GLM va être construit avec la fonction `H2OGeneralizedLinearEstimator` du package $h2o$ [8] sur python.

Eléments Théoriques

Le modèle linéaire généralisé
Soit $(y_1, y_2, \ldots, y_n)$ le vecteur d'observation obtenu avec la réalisation de la variable aléatoire $Y = (Y_1, Y_2, \ldots, Y_n)^t$ et $X_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^t$ le i-ème vecteur ligne des variables explicatives associées à l'observation i. X est donc une matrice de taille $n \times (p+1)$ dont les lignes sont les vecteurs lignes X_i^t et les variables correspondantes peuvent être quantitatives ou qualitatives.
Popularisé par McCullagh et Nelder (1989), le modèle linéaire généralisé permet de modéliser une relation non-linéaire entre la variable aléatoire $Y \in \mathbb{R}^n$ et les p variables explicatives X en utilisant une fonction lien [9] g comme suit, β étant le vecteur des $p+1$ paramètres

$$g(E[Y_i|X_i]) = X_i\beta. \tag{2.6}$$

Pour la modélisation des taux de résiliation, la variable à expliquer (l'acte de résiliation) est binaire (elle prend les valeurs 0 ou 1). Elle suit donc une loi de Bernoulli [10].
La loi de Bernoulli de paramètre μ_i appartient bien à la famille exponentielle avec les paramètres suivants

$$\theta_i = \log(\frac{\mu_i}{1-\mu_i}),\ \phi = 1 \text{ et } v(\theta_i) = \log(1-\exp(\theta_i)),$$

où $\theta_i \in \mathbb{R}$ = paramètre canonique ou de la moyenne, $\phi \in \mathbb{R}$ = paramètre de dispersion et v fonction définie sur $\mathbb{R}$ deux fois dérivable.

8. défini en Annexe A.6
9. voir Annexe A.3
10. elle peut être trouvée dans le tableau A.2 mis en Annexe A.3 et présentant les valeurs des paramètres des lois exponentielles usuelles

Sa fonction de lien va donc être le *logit*. Il s'agit alors d'une régression logistique avec

$$g(\mu_i) = \theta_i = \text{logit}(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right) \implies \mu_i(X) = \frac{e^{X_i\beta}}{1-e^{X_i\beta}}. \tag{2.7}$$

En supposant les (Y_i, X_i) indépendants et identiquement distribués, pour toute observation $i \in 1, \ldots, n$, la log-vraisemblance permettant de déterminer les coefficients β de la régression logistique peut avoir pour formule

$$\begin{aligned} \text{Log}(L) = \mathcal{L}_n(Y \mid X, \beta) &= \ln \prod_{i=1}^{n} \left(g^{-1}(X_i\beta)^{Y_i} \left(1 - g^{-1}(X_i\beta)\right)^{1-Y_i} \right), \\ &= \sum_{i=1}^{n} Y_i \ln\left(g^{-1}(X_i\beta)\right) + (1-Y_i)\ln\left(1 - g^{-1}(X_i\beta)\right). \end{aligned} \tag{2.8}$$

Pour trouver β_j maximisant cette log-vraisemblance, il faut donc résoudre les équations de vraisemblance suivantes

$$\frac{\partial \text{Log}(L)}{\partial \beta_j} = 0, \ \forall j = 0, ..., p. \tag{2.9}$$

Nous sommes aussi tentés d'utiliser un modèle de régression pénalisée.

Les modèles de régression pénalisée

Pour faire un GLM, les variables explicatives sont supposées indépendantes. Dans la réalité, les variables d'entrée pour la modélisation ne le sont pas toutes. Il pourrait alors y avoir un problème de conditionnement de la matrice X^tX à cause de la colinéarité entre certaines variables. Cette matrice devient donc difficile à inverser. L'ajout de la pénalisation résout ce problème vu qu'elle permet d'avoir une solution explicite au problème d'inversion. En effet, ajouter de la pénalisation revient à ajouter λ à toutes les valeurs propres de X^tX.

Le GLM peut également sur-apprendre (*overfit* en anglais) c'est-à-dire être très performant sur la base d'entraînement sans pour autant parvenir à généraliser cela sur la base de test. Ce sur-ajustement peut être évité en ajoutant une pénalité au modèle à forte variance, réduisant ainsi les coefficients β.

La pénalisation est un paramètre supplémentaire du modèle linéaire généralisé qui peut être choisi en essayant de maximiser une métrique permettant d'évaluer les performances du GLM comme l'AUC. Les mêmes notations que précedemment sont conservées.

Pour un GLM pénalisé, les coefficients β sont estimés comme suit

$$\hat{\beta}(\lambda) = arg \min_{\beta \in \mathbb{R}^{p+1}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ell(y_i \mid x_i, \beta) + \lambda ||\beta||_l^l \right\}, \tag{2.10}$$

où $\ell(y_i \mid x_i, \beta)$ est le terme de log-vraisemblance associé à l'observation i.

— Si $l = 2$ alors il s'agit d'une régression Ridge.

— Si $l = 1$ alors il s'agit d'une régression LASSO (*Least Absolute Shrinkage and Selection Operator*).

La manière dont la régression LASSO contraint les coefficients est différente de celle de la régression Ridge. Contrairement à la régression Ridge, la pénalisation LASSO permet de sélectionner les variables explicatives en annulant certains coefficients β_j et les variables correspondantes (variables j telles que $\beta_j = 0$) peuvent être retirées des variables explicatives. L'idée de la pénalisation LASSO est de

contraindre les coefficients petits à être nuls afin de rendre les autres coefficients plus significatifs. Néanmois, lorsque des variables sont corrélées, LASSO en choisit une, celle qui est la plus liée à la cible souvent, masquant ainsi l'influence des autres variables (corrélées à la variable choisie). Cela se retrouve dans plusieurs techniques utilisant un mécanisme de sélection de variables comme les arbres de décision [11].

Une troisième pénalisation combinant ces deux pénalisations a été introduite par Zou et Hastie (2005) et est notée *Elastic Net*. L'idée est d'utiliser une pondération avec une partie Ridge et une partie LASSO, soit pour $\lambda, \alpha > 0$

$$\beta_{\lambda,\alpha^{EN}} = arg \min_{\beta \in \mathbb{R}^{p+1}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i \mid x_i, \beta \right) + \lambda(\alpha \|\beta\|_1 + (1-\alpha)\|\beta\|_2^2) \right\}. \tag{2.11}$$

La méthode *Elastic Net* combine ainsi les atouts des méthodes Ridge et LASSO c'est-à-dire :

— partage des poids comme la régression Ridge. Ainsi lorsque les variables sont corrélées, ce n'est pas une seule variable qui va avoir un coefficient réduit à 0 par sélection arbiraire,

— capacité de sélection des variables de la régression LASSO conservée (coefficients nuls).

Pour déterminer les paramètres λ et $\alpha > 0$, nous pouvons toujours utiliser une grille permettant de trouver les valeurs de ces paramètres qui vont maximiser une métrique donnée.

Pour lutter contre l'*overfitting*, la validation croisée [12] va aussi être utilisée.

Sélection des variables

Selon la loi de parcimonie de «*'Occam's Razor'*», la meilleure explication à un problème est celle qui implique le moins d'hypothèses possibles. Ainsi, la sélection des variables devient un élément indispensable de la construction de modèles d'apprentissage automatique afin d'éliminer toutes les variables redondantes et dépendantes d'autres variables. N'oublions pas que le GLM suppose que les variables explicatives sont indépendantes. Elle consiste à réduire le nombre de variables d'entrée lors du développement d'un modèle prédictif. Réduire le nombre de variables d'entrée permet non seulement de réduire le temps de calcul de la modélisation mais aussi, dans certains cas, d'améliorer les performances du modèle.

Différentes méthodes de sélection des variables

Embedded Methods

Les méthodes intégrées ou *Embedded Methods* en anglais, sont itératives dans le sens où elles prennent en charge chaque itération du processus de formation du modèle et extraient soigneusement les caractéristiques qui contribuent le plus à la formation pour une itération particulière. Parmi ces techniques, se trouve la régularisation LASSO.

Wrapper-based

Ces méthodes considèrent la sélection d'un ensemble de variables comme un problème de recherche. Ils sont utilisés avec un critère d'évaluation [13] (AIC, BIC, P-value par exemple). Dans le tableau 2.3 ci-dessous sont présentés 3 méthodes *Wrapper-based* :

11. voir Annexe A.5
12. définie en Annexe A.7.1
13. voir Annexe A.4

Nom	Description
Forward Feature Elimination	Il s'agit d'une méthode itérative qui commence par la variable la plus performante par rapport à la cible. Ensuite, elle sélectionne une autre variable qui donne les meilleures performances en combinaison avec la première variable. Ce processus se poursuit jusqu'à ce que le critère prédéfini soit atteint. Cette méthode pourrait permettre d'éliminer les variables colinéaires puisqu'une fois l'une des variables colinéaires sélectionnées, les autres pourraient ne plus être significatives. Nous pouvons aussi utiliser une liste initiale dans laquelle nous allons mettre les variables que nous souhaitons conserver pour la modélisation.
Backward Feature Elimination	Cette méthode fonctionne exactement à l'opposé de la méthode *Forward Feature Selection*. Elle commence avec toutes les variables pour construire un modèle. Ensuite, elle élimine une à une les variables qui donnent les pires performances selon le critère d'évaluation. Ce processus se poursuit jusqu'à ce que le critère prédéfini soit atteint.
Exhaustive Feature Selection	C'est la méthode de sélection de variables la plus robuste parmi les 3 *Wrapper-based* cités. Il s'agit d'une évaluation par force brute de chaque sous-ensemble de variables. Cela signifie qu'elle essaie toutes les combinaisons possibles des variables et renvoie le sous-ensemble le plus performant. Son inconvénient est donc qu'elle a un temps de calcul trop élevé comparé aux deux autres.

TABLE 2.3 – Tableau présentant des méthodes de sélection des variables *Wrapper-based*

Méthodes de sélection des variables choisies

Nous allons donc faire une sélection des variables en utilisant la méthode ascendante (*Forward Feature Elimination*) afin de sélectionner les variables à utiliser pour construire le GLM sans régularisation. L'algorithme implémentant la selection ascendante des variables nous permet de définir une liste initiale qui va contenir toutes les variables que nous voudrons conserver dans le modèle. Cette liste contient donc les différents parcours puisque nous souhaitons voir l'impact des parcours sur la résiliation (il faut donc forcément les utiliser lors de la modélisation). L'algorithme a été codé sur python. Le critère de sélection des variables choisi est la p-value [14] puisque nous souhaitons garder les variables significatives afin de pouvoir bien interpréter leur coefficient.

Nous souhaitons aussi tester le GLM pénalisé afin de voir si les performances obtenues vont être meilleures que celles d'un GLM non pénalisé. Pour ce GLM, nous allons aussi faire une sélection des variables en utilisant la pénalisation LASSO. Celle-ci a la propriété de réduire certaines variables à zéro de telle sorte que ces variables peuvent être supprimées du modèle. Pour faire un GLM LASSO au lieu d'un GLM sans pénalisation, il faut définir la valeur de λ en paramètre du modèle GLM avec le package

14. voir Annexe A.4.1

h2o. Il faut donc d'abord déterminer la valeur de λ en utilisant une grille de recherche (*gridsearch* sur *h2o*). C'est une grille qui va permettre de créer plusieurs modèles LASSO avec différentes valeurs de λ afin de déterminer quelle valeur de lambda permet de maximiser la performance du LASSO suivant une métrique choisie. Il faut aussi tenir compte du nombre de variables éliminées car plus la pénalisation est grande, plus il peut y avoir de variables jugées non explicatives.

En utilisant le *gridsearch* de *h2o*, nous avons essayé de déterminer le paramètre λ de la pénalisation LASSO qui va maximiser l'AUC puisque nous avions remarqué que l'AUC du modèle avec un λ pris par défaut était médiocre. Nous avons donc résumé dans le tableau 2.4 suivant, les 4 valeurs de lambda donnant les 4 meilleurs AUC ainsi que le nombre de variables dont les coefficients ont été réduits à 0 par la régularisation LASSO. La valeur de λ donnant la meilleure AUC ne permet de supprimer que 3 variables, la valeur de λ permettant de supprimer 6 variables a une AUC proche de la meilleure AUC. Avec $\lambda = 0.2$, 11 variables sont éliminées et l'AUC baisse de 4%. Nous avons donc choisi $\lambda = 0.1$ permettant ainsi de supprimer 6 variables.

Valeur de λ	AUC	Nombre de variables avec $\beta = 0$
0.001	72.93%	3
0.0001	72.85%	2
0.1	72.7%	6
0.2	68%	11

TABLE 2.4 – Tableau présentant l'AUC et nombre de variables avec $\beta = 0$ obtenu après régression LASSO suivant λ

Variables non retenues lors de la sélection

Précisons d'abord que pour la modélisation, les variables qualitatives sont transformées en variables numériques, ce qui entraîne que leurs modalités deviennent des variables à part entière.
Ainsi, les 8 variables qui n'ont pas été retenues après la sélection des variables par la méthode ascendante sont :

— 2 modalités de la situation familiale : `marie` et `celibataire`,

— 2 modalités de catégories socio-professionnelles : `professions intermediaires` et `inconnu`,

— 2 modalités de `nivCap_nbPiece` [15] : `nivCap_nbPiece_2-3X5` [16] et `nivCap_nbPiece_2-3X3-4` [17],

— 2 variables zonier : `BDG_com` (bris de glace commercial) et `score_s` (score sécheresse).

Les 6 variables jugées inutiles par la sélection de variables par la pénalisation LASSO sont :

— 2 variables zoniers comme celle du bris de glace (zonier commercial) ainsi que le zonier responsabilité civile vie privée RCVP (zonier commercial),

— 2 modalités de de la situation familiale : `marie` et `celibataire`,

— 2 modalités de catégories socio-professionnelles : `professions intermediaires` et `inconnu`.

Nous remarquons que parmi les 6 variables non retenues par la pénalisation LASSO, 5 n'ont pas aussi été retenues par la méthode ascendante.

Une fois que nous connaissons les variables à éliminer grâce à la sélection des variables, nous allons donc construire le modèle GLM avec les variables retenues. Après avoir obtenu les modèles, leurs

15. niveau du capital croisé avec le nombre de pièces.
16. niveau 2 ou 3 avec un nombre de pièces supérieur ou égal à 5.
17. niveau 2 ou 3 avec 3 ou 4 pièces.

performances seront évaluées en utilisant les métriques déjà définies plus tôt.

Performances

Le tableau 2.5 suivant résume les valeurs des différentes métriques obtenues avec un modèle GLM et un modèle GLM pénalisé afin de déterminer le plus performant. Rappelons que le λ choisi pour le GLM LASSO est le λ qui permet de maximiser l'AUC tout en éliminant 6 variables comme vu dans la partie sélection des variables.
Nous remarquons que les valeurs des métriques des deux modèles sont proches. Nous choisissons donc de garder le modèle GLM sans pénalisation. N'oublions pas aussi qu'avec le GLM pénalisé, nous ne pouvons pas calculer la p-value puisque l'hypothèse nulle avec la régularisation LASSO est différente de l'hypothèse nulle d'un modèle linéaire généralisé sans régularisation.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision	AIC
GLM	0.423	72.8%	35.17%	62.84%	31.51%	152 739
GLM pénalisé $\lambda = 0.1$	0.420	72.7%	34.98%	62.80%	32%	152 924

TABLE 2.5 – Tableau comparant les valeurs des métriques obtenues avec la base test pour un GLM normal et un GLM pénalisé avec $\lambda = 0.1$ pour la première période d'une année

Rappelons que pour essayer de lutter contre le problème engendré par le déséquilibre des classes, nous avons testé :
— une méthode de sur-échantillonnage : SMOTE,
— une méthode de sous-échantillonnage : *Random Under Sampling* sur les contrats de la classe majoritaire appartenant au parcours sur-représenté ATA (afin d'essayer aussi de mieux équilibrer la représentation des parcours).
Le tableau 2.6 suivant permet de comparer les valeurs des métriques obtenues avec la base test pour un GLM normal et les GLM construits avec des données ré-échantillonnées pour la première période d'une année.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision	AIC
GLM	0.423	72.8%	35.17%	62.84%	31.51%	152 739
GLM équilibré SMOTE	0.435	71.6%	34.2%	64.25%	29.98%	152 782
GLM équilibré *Under Sampling*	0.447	71%	32.6%	63%	29.83%	153 284

TABLE 2.6 – Tableau comparant les valeurs des métriques obtenues avec la base test pour un GLM normal et les GLM construits avec des données ré-échantillonnées pour la première période d'une année

Nous remarquons alors qu'en sur-échantillonnant les données afin d'équilibrer les deux classes, il n'y a pas de grande différence entre les performances des différents modèles. Les modèles avec les classes équilibrées ont une AUC, une AUCPr et une précision un peu plus faible, cependant leur rappel est plus élevé.

Vu qu'en équilibrant les données, la représentation des parcours change, nous allons aussi regarder les performances par parcours avec les différents GLM construits. Les performances sont proches d'un modèle à l'autre ce qui fait que ce n'est pas un élément qui va influencer le choix du modèle GLM que

nous allons garder.
Le tableau 2.7 suivant présente les valeurs des métriques sur la base test pour chaque parcours obtenues avec le GLM sans ré-échantillonnage pour la première année. Il permet donc de voir que les parcours les moins représentés n'ont pas été «ignorés» par le modèle vu que les performances sur ces parcours sont proches voire supérieures aux performances globales.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
ADA	0.43	70%	27%	77%	23%
ATA	0.42	73%	35%	62%	31%
MM	0.393	72.20%	34%	62%	27.6%
MA	0.41	85%	55%	65%	78%
W	0.39	93%	78%	82%	76%

TABLE 2.7 – Tableau présentant les valeurs des métriques sur la base test pour chaque parcours obtenues avec le GLM normal pour la première année

Equilibrer les classes n'a pas permis d'améliorer les performances comme nous le souhaitions, nous allons donc utiliser le GLM avec les classes non équilibrées.

Après avoir déterminé le meilleur GLM construit, nous allons interpréter les coefficients obtenus afin de voir comment les variables influencent les résiliations. Ceci permettra aussi de mieux comprendre le modèle et de vérifier si les informations fournies sont cohérentes.

Interprétation des coefficients

Le coefficient d'une variable β d'un **modèle linéaire** permet de voir l'effet d'un changement d'une unité de la variable explicative si celle-ci est quantitative ou de comparer les différences entre les différentes modalités d'une variable catégorielle lorsque la variable explicative est qualitative. Par exemple, si un modèle linéaire associe un coefficient égal à 0.03 à la modalité `résidence principale` de la variable `type de résidence`, nous pouvons déduire que le fait que la résidence de l'habitation assurée soit une résidence principale au lieu d'une résidence secondaire fait augmenter le taux de résiliation de 0,03. 0 représente le coefficient de la deuxième modalité de la variable `type de résidence` qui est égale à `secondaire`. En effet, pour les variables catégorielles, une catégorie a 0 comme coefficient, c'est la catégorie de référence. Si le logement n'est pas une résidence principale alors c'est forcément une résidence secondaire. L'interprétation est légèrement moins simple pour les variables quantitatives. Lorsque le coefficient de la prime est de -0.077 par exemple, et que la prime a pour valeur 100€, l'effet de la prime sur la prédiction est de $100 \times -0.077 = -0.77$. La probabilité de résiliation diminue pour les primes plus élevées.
Dans un modèle linéaire simple, le coefficient de la variable explicative correspond à la pente de la droite. Si le coefficient est positif, la direction de la droite est «vers le haut» et s'il est négatif, «vers le bas». Le signe du coefficient détermine donc si l'effet de la variable explicative sur la variable à expliquer est positif ou négatif. De même pour le modèle linéaire généralisé, un coefficient négatif signifie que la variable correspondante a une relation négative avec la variable à expliquer. Lorsque les autres variables explicatives gardent une valeur constante, une augmentation de la variable explicative ayant un coefficient négatif entraîne une diminution de la variable à expliquer lorsque les variables sont indépendantes.

Les coefficients β d'un modèle linéaire généralisé peuvent donc être interprétrés de la même manière

que ceux d'un modèle linéaire. Il faut cependant garder en tête qu'ils donnent la variation du *logit* et non la moyenne. En effet, nous avons déjà vu que le modèle linéaire généralisé avec lien *logit* est défini comme suit (en reprenant les mêmes notations que la partie Eléments théoriques 2.3.1)

$$\frac{\mu_i}{1-\mu_i} = \exp\left(\mathbf{X}_i\beta\right). \tag{2.12}$$

Ainsi, l'interprétation des coefficients pour déterminer l'impact des variables sur la résiliation avec le modèle linéaire généralisé suit la notion d'***odds ratio***.
Pour comprendre ce qu'est un *odds ratio*, il faut d'abord connaître la définition des cotes (*odds* en anglais). Une cote est le rapport des probabilités de deux résultats qui s'excluent mutuellement. Par exemple, si le modèle prédit une probabilité de résiliation de 10%, la probabilité de ne pas résilier est de $100\% - 10\% = 90\%$. Les chances sont donc de 10% contre 90%. Diviser les deux côtés par 90% donne alors 0,11 contre 1. Ainsi, l'*odds* de 0,11 n'est qu'une façon différente de dire que la probabilité de résiliation est de 10%. Cela veut aussi signifier que l'assuré a 0.11 fois plus de chance de résilier que de ne pas résilier.

Un rapport de cotes ou *Odds Ratio* en anglais (OR), bien défini dans l'article de Szumilas et Magdalena (2010), est donc une mesure d'association entre une certaine propriété A et une deuxième propriété B dans une population. Plus précisément, il indique comment la présence ou l'absence de la propriété A a un effet sur la présence ou l'absence de la propriété B.
Le rapport de cotes peut donc être utilisé pour déterminer si une exposition particulière est un facteur de risque pour la résiliation et pour comparer l'ampleur de divers facteurs de risque pour la résiliation. Pour comparer les probabilités d'obtenir $Y = 1$ (dans notre cas que le contrat soit résilié) entre 2 individus x et $x\prime$

$$OR(x, x\prime) = \frac{\mu(x)/[1-(\mu(x))]}{\mu(x\prime)/[1-(\mu(x\prime))]} = \frac{odds(x)}{odds(x\prime)}, \tag{2.13}$$

avec

$$odds(x) = \frac{\mu(x)}{1-\mu(x)}.$$

Pour déterminer l'influence d'une variable X_j, il faut donc comparer les probabilités de succès de deux observations x et $x\prime$ dont seule la j-ème variable est différente. L'*odds ratio* est alors donné par

$$OR(x, x\prime) = \frac{odds(x)}{odds(x\prime)} = \exp(\beta_j(x_j - x_j\prime)), \text{ avec } j \in [1, p]. \tag{2.14}$$

Ainsi, pour les variables continues, lorsque X_j passe de x à $x+1$, l'odds ratio est

$$OR(x+1, x) = \exp(\beta_j((x+1) - x)) = \exp(\beta_j). \tag{2.15}$$

Et pour les variables binaires, la variable X_j n'a que deux valeurs possibles 0 et 1, l'*odds ratio* est donc

$$OR(1, 0) = \exp(\beta_j((1-0)) = \exp(\beta_j). \tag{2.16}$$

En d'autres termes, la fonction exponentielle du coefficient ($\exp(\beta_j)$) est le rapport de cotes associé à une augmentation d'une unité de l'exposition pour une variable continue ou le rapport de cotes associé aux deux valeurs d'une variable binaire [18]. C'est pourquoi l'*odds ratio* (OR) permet d'interpréter les coefficients associés aux variables explicatives du modèle GLM. Suivant la valeur de l'*odds ratio* (OR), il est possible de voir comment la variable influence la résiliation :

18. Les variables qualitatives de la base ont été transformées en variables binaires lors de la modélisation.

— $OR = 1$ alors $\beta = 0$, la variable associée n'affecte pas la variable à expliquer.
— $OR > 1$ alors $\beta > 0$, la variable associée a une influence positive sur la variable à expliquer.
— $OR < 1$ alors $\beta < 0$, la variable associée a une influence négative sur la variable à expliquer.

Le tableau 2.8 suivant montre les valeurs des coefficients obtenues avec le GLM gardé. La p-value, montrant si la variable est significative [19], peut aussi être vue. Il y a 11 variables qui ne sont pas significatives soit 18% des variables.

	variables	p_value	coefficients	exp(coefficients)
1	**habxJur_Appartement_Locataire** vs habxJur_Maison_Proprietaire	0	1,31	3,71
2	**habxJur_Maison_Locataire** vs habxJur_Maison_Proprietaire	0	1,21	3,35
3	**covid_covid_stand** vs couvre_feu et confinement	0	1,09	2,97
4	**nb_contrats** (un seul vs plusieurs)	0	0,43	1,53
5	**AgeCat_Plus de 65ans** vs jeunes	0	-0,32	0,72
6	**generation**	0	-0,31	0,73
7	**Code_firme_1354** vs code_firme_Autres	0	0,30	1,35
8	**Csp_Employés** vs csp_Autres personnes sans activité professionnelle	0	-0,28	0,75
9	**nivCap_nbPiece_1X1-2** vs nivCap_nbPiece le plus prestigieux	2,196E-13	0,26	1,30
10	**code_firme_Sans firme** vs code_firme_Autres	0	0,25	1,28

TABLE 2.8 – P-value, coefficient, et exponentielle des coefficients des 10 variables les plus importantes du modèle GLM de la première année

Les 10 variables présentes dans le tableau sont toutes significatives. Cela permet de conclure que :
— Les locataires d'appartement ont 3,71 fois plus de chance de résilier que les propriétaires de maison. Les locataires de maison ont 3,35 fois plus de chance de résilier que les propriétaires de maison. Ils ont presque autant de chance de résilier que les locataires d'appartement.
— Pendant le confinement ou le couvre-feu, les assurés avaient presque 3 fois moins de chance de résilier que pendant la période de Covid-19 sans confinement ni couvre feu (période de Covid-19 standart). Après le confinement ou le couvre-feu, les assurés en ont profité pour résilier leur contrat.

19. La p-value est inférieure ou égale à 5%.

— Les assurés ayant un seul contrat (l'affaire nouvelle) ont 53% plus de chance de résilier que les assurés ayant plusieurs contrats.
— Les seniors ont 28% moins de chance de résilier que les jeunes. Ils ont 18% moins de chance de résilier que les adultes.
— Les assurés ayant le code firme 1354 (avantage de -25% ou -23% sur la prime affaire nouvelle) ont 35% plus de chance de résilier que les assurés ayant un autre code firme (qui est généralement plus avantageux). Cependant, ce code firme a un coefficient différent de 7% de celle de la variable `sans code firme`. Les assurés ayant le code firme 1354 et les assurés qui n'ont pas de code firme ont donc presque autant de chance de résilier pendant la première année du contrat dans le portefeuille. Cela peut s'expliquer par le fait que les clients ayant l'avantage code firme 1354 ont tendance à résilier après la première échéance pour un tarif plus avantageux chez un concurrent vu qu'ils perdent leur réduction de prime.
— Les employés ont 25% moins de chance de résilier que les personnes sans activité professionnelle.
— Les assurés qui ont le niveau de capital le plus bas avec 1 ou 2 pièces ont 30% plus de chance de résilier que les assurés ayant un niveau de capital plus prestigieux. Les assurés ayant ce niveau de capital sont plus suceptibles de déménager vu que les logements assurés sont en général des studios.

Zoom sur les parcours

Rappelons tout d'abord qu'en observant les performances obtenues suivant le parcours grâce au tableau 2.7 vu plus tôt, nous avons remarqué que les performances par parcours sont aussi acceptables que les performances globales du modèle. Les parcours sous-représentés (plateforme MM, MA et Web W) n'ont pas été «ignorés» par le modèle. Nous pouvons donc regarder les coefficients et l'importance des parcours obtenus avec le modèle GLM.
Le tableau 2.9 suivant présente les différents coefficients des parcours ainsi que leur p-value. Le parcours Web W a été choisi comme référence [20]. Nous remarquons que leur coefficient est proche et les différences sont moins de 1%. Cela voudrait dire que les chances de résiliation sont presque les mêmes suivant le parcours. Cependant, seul les parcours plateforme MM et MA sont significatifs.
Nous constatons aussi que les coefficients des parcours significatifs sont faibles (autour de 0,01 en valeur absolu). Précisons que le coefficient le plus élevé du modèle est égal à 1,31 et le coefficient médian 0,12 (correspond à la variable zonier commercial incendie).

Parcours	Coefficient	Odds Ratio	P-value
ADA vs W	-0,01368	0.986	0.10
ATA vs W	-0,01369	0.986	0.3
MM vs W	-0,01367	0.986	9e-12
MA vs W	-0,01369	0.986	3e-8

TABLE 2.9 – Tableau présentant le coefficient, l'*odds ratio* et la p-value de chaque parcours obtenus avec le GLM pour la première année

Les mêmes remarques sont faites avec les GLMs construits après ré-échantillonnage de la base (SMOTE et *Random Under Sampling*), le ré-échantillonnage ayant permis d'augmenter la représentation des parcours avec peu de volume. Cela pousse à déduire que les parcours sont très peu importants.

Ainsi nous serions tenter de dire que les parcours ont très peu d'influence sur la résiliation pendant la première année d'après le GLM.

20. Vu qu'il y a 5 modalités possibles, il faut supprimer une modalité parce qu'elle peut être retrouvée lorsqu'aucune des autres modalités n'est vérifiée. Elle représente la référence.

2.3.2 Random Forest

Afin de trouver le modèle ayant les meilleures performances, nous allons à présent modéliser la résiliation en utilisant le *Random Forest*. Nous utilisons les mêmes variables explicatives que lors de la modélisation avec le GLM. Le package utilisé sur python est *h2o* et l'outil est le `Distributed Random Forest` (DRF) de *h2o*.

Eléments théoriques

La forêt aléatoire ou *Random Forest* en anglais a été introduite par Breiman (2001). C'est un modèle qui entraîne de nombreux arbres de décision [21] en parallèle avant de faire la moyenne de leurs résultats pour une régression ou de prendre le résultat obtenu par vote majoritaire pour une classification. Les forêts aléatoires utilisent le même principe que le *bagging* mais avec une modification considérable. Leur particularité par rapport au *bagging*, est qu'ils permettent de construire une grande collection d'arbres **décorrélés** en faisant un tirage aléatoire sur les variables d'entrée (*feature sampling*) en plus du tirage aléatoire sur les lignes.

Description de l'algorithme

Algorithme 1 : Random forest pour la régression ou la classification

Entrées : données d'apprentissage
Sorties : prédiction pour le point x
pour $b \leftarrow 1$ ***à*** B **faire**
  Dessiner un échantillon *bootstrap* Z^* de taille N à partir des données d'apprentissage;
  Construire un arbre de forêt aleatoire T_b avec les données obtenues après *bootstrap*, c'est-à-dire en répétant de manière récursive les étapes suivantes ;
  pour *chaque nœud terminal de l'arbre* **faire**
    tant que *le nœud terminal de l'arbre n'atteint pas la taille minimale de nœuds* n_{min} **faire**
      Sélectionner m variables au hasard parmi les p variables;
      Choisir la meilleure variable de division de l'arbre parmi les m variables;
      Diviser le nœud en deux nœuds filles en utilisant cette variable.
    fin
  fin
fin
Sortir l'ensemble des arbres $\{T_b\}_1^B$;
Faire une prédiction pour le point x c'est-à-dire;
si *Regression* **alors**

$$\hat{f}_{rf}^B(x) = \frac{1}{b}\sum_{b=1}^{B} T_b(x). \tag{2.17}$$

fin
sinon si *Classification* **alors**
  En notant $\hat{C}_b(x)$ la classe prédite par le b-ième arbre de la forêt aléatoire (*random forest*) alors

$$\hat{C}_{rf}^B(x) = \textit{ vote majoritaire } \{\hat{C}_b(x)\}_1^B. \tag{2.18}$$

fin

21. voir Annexe A.5

Feature sampling

Comme le bruit dans les arbres est considérable, faire la moyenne des résultats obtenus pourrait être très bénéfique. De plus, puisque chaque arbre généré par *bagging* est distribué de manière identique (i.d.), l'espérance d'une moyenne de B arbres est la même que l'espérance de n'importe lequel d'entre eux. Cela signifie que le biais des arbres générés par *bagging* est le même que celui des arbres individuels, et le seul moyen de l'améliorer est de réduire les écarts. Cela contraste avec le *boosting*, où les arbres sont obtenus de manière séquentielle pour éliminer les biais, et ne sont donc pas identiquement distribués.

Une moyenne de B indépendantes et identiquement distribuées variables aléatoires, ayant chacune une variance de σ^2, a une variance $\frac{1}{B}\sigma^2$. Cependant, dans la réalité, les variables sont simplement identiquement distribuées et présente une certaine corrélation. Si on note le coefficient de cette corrélation ρ, la variance de la moyenne est donnée par

$$\rho\sigma^2 + \frac{1}{B}\sigma^2. \tag{2.19}$$

Lorsque B augmente, le deuxième terme tend vers 0, mais le premier reste le même. Par conséquent, la corrélation des paires d'arbres générés par *bagging* limite les avantages opportés par la moyenne.
L'idée des forêts aléatoires est d'améliorer la réduction de la variance du *bagging* en réduisant la corrélation entre les arbres, sans pour autant trop augmenter la variance. Pour cela, le *feature sampling* ou sélection aléatoire des variables d'entrée est réalisé pendant le processus de construction de l'arbre.

Avant chaque division, sont sélectionnées aléatoirement $m \leq p$ des variables d'entrée comme candidats pour la division, constituant ainsi la sélection des variables. Intuitivement, réduire m revient à baisser la corrélation entre n'importe quelle paire d'arbres dans l'ensemble, et donc à diminuer la variance de la moyenne. Mais réduire m pourrait aussi éliminer certaines variables qui auraient permis d'avoir un modèle plus performant. C'est pourquoi, il faut bien choisir m. Ses valeurs usuelles sont $\sqrt{p}$ pour un problème de classication avec p variables et $\frac{p}{3}$ pour un problème de régression.

Le RF peut aussi sur-apprendre (*overfit*) ce qui entraîne qu'il aura du mal à généraliser sur des données inconnues. Son temps de calcul peut aussi être supérieur à celui du GLM. C'est pourquoi nous allons utiliser un $early - stopping$ [22] sans oublier de faire une validation croisée [23]. Pour essayer d'améliorer les performances, il est possible de faire une optimisation des paramètres.

Performances

En utilisant les métriques que nous avons définies plus tôt, nous évaluons les performances des modèles *Random Forest* sur la base test.

Rappelons que pour essayer de lutter contre le problème engendré par le déséquilibre des classes, nous avons testé :
— une méthode de sur-échantillonnage : SMOTE,
— une méthode de sous-échantillonnage : *Random Under Sampling* sur les contrats de la classe majoritaire appartenant au parcours sur-représenté ATA (afin d'essayer aussi de mieux équilibrer la représentation des parcours).
Le tableau 2.10 suivant résume les performances obtenues sur la base test avec un RF optimisé [24] (sur la base d'entraînement initiale) ainsi qu'avec les *random forests* construits sur la base de d'entraînement ré-échantillonnée et optimisés.

22. défini en Annexe A.7.2
23. définie en Annexe A.7.1
24. L'optimisation consiste à détermination des hyperparamètres permettant de maximiser une métrique - Voir Annexe A.6.1

Nous remarquons que comme avec le GLM, le modèle équilibré augmente le rappel en baissant la précision et l'AUC. Précisons que le Random Forest équilibré avec SMOTE a pris presque deux fois plus de temps pour trouver les hyperparamètres optimaux vu que la taille de la base d'entraînement obtenu avec SMOTE a augmenté d'environ 30%.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
Random forest optimisé	0.406	75.6%	43%	63%	34%
Random forest SMOTE optimisé	0.410	73%	42%	66%	30%
Random forest Random Under Sampling optimisé	0.418	75.2%	42%	65%	30%

TABLE 2.10 – Tableau présentant les valeurs des métriques avec la base test des modèles *Random Forest* de la première année construits avec ou sans ré-échantillonnage des données

Vu qu'en équilibrant les données, la représentation des parcours change, nous allons aussi regarder les performances par parcours avec les différents RF construits. Les performances suivant les parcours sont presque similaires d'un modèle à un autre ce qui fait que ce n'est pas un élément qui va influencer le choix du modèle RF. Le tableau 2.11 présente les valeurs des métriques pour chaque parcours obtenues avec le *Random Forest* sans ré-échantillonnage de la base d'entraînement pour la première année.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
ADA	0.42	73%	29%	76%	22%
ATA	0.40	75%	43%	63%	34%
MM	0.39	76%	35%	63%	28%
MA	0.39	84%	54%	66%	79%
W	0.40	91%	77%	83%	77%

TABLE 2.11 – Tableau présentant les valeurs des métriques pour chaque parcours obtenues avec le *Random Forest* pour la première année

Equilibrer les classes n'a donc pas permis d'améliorer les performances comme nous le souhaitions, nous allons donc rester sur le modèle *Random Forest* construit sans ré-échantillonnage.

2.3.3 XGBoost

Toujours dans l'optique de trouver un modèle offrant les meilleures performances, nous allons utiliser un troisième modèle qui est le XGBoost. Il va être construit avec le package *h2o* (défini en Annexe A.6) sur python (fonction `H2OXGBoostEstimator`).

Elements théoriques

XGBoost est un algorithme d'apprentissage *machine learning* basé sur un arbre de décision [25] qui utilise le gradient *boosting*. L'algorithme XGBoost a été développé dans le cadre d'un projet de recherche à l'Université de Washington. Tianqi CHEN et Carlos GUESTRIN ont présenté leur article à la conférence *SIGKDD* en 2016 (CHEN et GUESTRIN, 2016). Depuis son introduction, cet algorithme a permis de remporter de nombreux concours de *machine learning* sur Kaggle [26] et sa puissance ne

25. voir Annexe A.5

26. Kaggle est une plateforme web organisant des compétitions en science des données, les problèmes étant proposés par des entreprises qui offrent ainsi un prix aux participants obtenant les meilleures performances.

cesse d'être ventée. En conséquence, il existe une forte communauté de *data scientist* contribuant aux projets *open source* XGBoost avec environ 350 contributeurs et 3 600 commits sur GitHub. C'est pour cette raison que nous avons choisi de le tester.

Pour décrire succinctement le principe, le XGBoost comme le *Gradient Boosting Machine* (GBM) utilise l'algorithme du gradient *boosting* qui travaille de manière séquentielle contrairement au *Random Forest*. Cela va le rendre plus lent mais il va surtout permettre à l'algorithme de s'améliorer en prenant en compte l'erreur obtenue durant les exécutions précédentes. Il commence donc par construire un premier modèle et à partir de cette première évaluation, chaque individu va alors être pondéré en fonction de la performance de la prédiction, ce qui constitue le principe du *boosting*. Il peut donc déetecter et apprendre à partir de modèles de donnéees non linéaires.

Le gradient *boosting* prend comme entrées : l'ensemble d'apprentissage $\{\boldsymbol{X}_i, Y_i\}_{i=1,\ldots,n}$, avec X_i représentant le vecteur des variables explicatives de l'observation i et Y_i la variable réponse (égale à 1 s'il y a eu une résiliation ou 0 sinon) ainsi qu'une fonction de perte L. D'après Friedman (2001), la fonction de perte pour un gradient *boosting* appliqué à une classification binaire peut être égale à la logvraisemblance de la loi binomiale négative

$$L(Y, F) = \log(1 + \exp(-2YF)), \tag{2.20}$$

$$\text{avec } F(X) = \tfrac{1}{2}\log\left[\tfrac{P(Y=1|X)}{P(Y=0|X)}\right], Y \in \{0, 1\}.$$

Il suppose qu'il existe une fonction F^* qui explique le mieux la variable réponse Y par les p variables explicatives (X_1, ..., X_p) retenues dans le vecteur $\boldsymbol{X}$.

$$F^* \in \arg\min_F \mathbb{E}_X \left[\mathbb{E}_Y[L(Y, F(\boldsymbol{X})) \mid \boldsymbol{X}]\right]. \tag{2.21}$$

L'algorithme consiste donc à trouver une approximation $\widetilde{F}^*$ de la fonction F^*.
A chaque étape de l'algorithme, le gradient *boosting* essaye de faire une réduction d'erreur en prédisant les résidus par descente de gradient. Il part alors d'un modèle initial F_0 pour obtenir à chaque étape m, un modèle F_m résultant sur un modèle final $\widetilde{F}^* = F_M(\boldsymbol{X})$. Le modèle initial pour une classification binaire peut être défini comme suit

$$F_0(\boldsymbol{X}) = \frac{1}{2}\log\frac{1+\bar{Y}}{1-\bar{Y}}. \tag{2.22}$$

Le modèle obtenu en ajoutant m (=1, 2,. . ., m) base d'apprentissage ainsi que la valeur initiale constante F_0 est noté F_m et se définit comme suit

$$F_m(\boldsymbol{X}) = F_{m-1}(\boldsymbol{X}) + \rho_m h(\boldsymbol{X}, a_m). \tag{2.23}$$

Le facteur multiplicatif ρ_m assurant la réduction d'erreur est obtenu avec la formule suivante

$$\rho_m = \arg\min_\rho \sum_{i=1}^{n} L\left(Y_i, F_{m-1}\left(\boldsymbol{X}_i\right) + \rho h\left(\boldsymbol{X}_i, a_m\right)\right), \tag{2.24}$$

$h\left(\boldsymbol{X}_i, a_m\right)$ constitue le m-ième arbre de décision CART construit afin de prédire la pseudo-réponse de Friedman $\widetilde{Y}_i$ et le vecteur a_m constituent les paramètres de cet arbre. La pseudo-réponse de Friedman

parfois appelée pseudo-résidu est définie comme le gradient négatif en X_i

$$\widetilde{Y_i} = -\left[\frac{\partial L(Y_i, F(\boldsymbol{X}_i))}{\partial F(\boldsymbol{X}_i)}\right]_{F(\boldsymbol{X})=F_{m-1}(\boldsymbol{X})} = \frac{2Y_i}{1+exp(2Y_iF_{m-1}(\boldsymbol{X}_i))}. \tag{2.25}$$

Ainsi à chaque étape m, de nouvelles prédictions F_m, souvent légèrement meilleure que les précédentes, sont obtenues. Le modèle final F_M sera donc entraîné en prenant la somme pondérée de M arbres d'apprentissage de base (modélisation additive). C'est le modèle qui explique le mieux la variable réponse Y et il va donc permettre de faire les prédictions de l'acte de résiliation.

Même s'il est semblable au *Gradient Boosting Machine* (GBM), le XGBoost ajoute quelques modifications à l'algorithme du gradient *boosting* lui permettant ainsi de présenter plusieurs avantages :

— Élagage des arbres : Le critère d'arrêt du fractionnement des arbres dans le cadre du GBM est de nature gourmande et dépend de la fonction de perte au point de fractionnement. XGBoost utilise le paramètre `max_depth` (profondeur de l'arborescence) comme spécifié au lieu de la fonction de perte, et commence à élaguer les arbres vers l'arrière. Cette approche «en profondeur d'abord» améliore considérablement les performances de calcul.

— Parallélisation : le modèle s'implémente en parallèle, ce qui réduit son temps d'éxecution.

— Régularisation : XGBoost inclut différentes pénalités pour la régularisation afin d'éviter le sur-apprentissage.

Le XGBoost peut aussi sur-apprendre (*overfit*) et son temps de calcul peut aussi être supérieur à celui du GLM et du RF à cause du *boosting*. C'est pourquoi nous allons utiliser un *early_stopping* [27] sans oublier de faire une validation croisée [28].

Performances

Précisons d'abord qu'en essayant d'optimiser [29] le XGBoost, le gain de performance obtenu n'est pas très élevé peut être parce que nous n'avons trouvé qu'un optimum local (l'AUC passe de 72% à 75.7% et l'AUCPr de 40% à 42% après optimisation sans ré-échantillonnage). Il n'est pas facile d'optimiser les hyperparamètres du XGBoost avec une base d'entraînement de presque 900 000 lignes puisque l'optimisation prend beaucoup de temps (elle a pris plus de 13 heures).

Rappelons que pour essayer de lutter contre le problème engendré par le déséquilibre des classes, nous avons testé :
— une méthode de sur-échantillonnage : SMOTE,
— une méthode de sous-échantillonnage : *Random Under Sampling* sur les contrats de la classe majoritaire appartenant au parcours sur-représenté ATA (afin d'essayer aussi d'équilibrer au mieux la représentation des parcours).
Le tableau 2.12 suivant présente les valeurs des différentes métriques avec la base test obtenues avec les différents modèles XGBoost optimisés pour la modélisation de la première période d'une année.
Nous remarquons, comme avec les autres modèles, qu'équilibrer le XGBoost avec un SMOTE augmente le rappel en baissant la précision et l'AUC. Equilibrer le modèle par *Random Under Sampling* donne des performances légèrement inférieures quelle que soit la métrique. Cela pourrait s'expliquer par

27. défini en Annexe A.7.2
28. voir Annexe A.7.1
29. L'optimisation consiste à détermination des hyperparamètres permettant de maximiser une métrique - Voir Annexe A.6.1

le fait que nous avons peut-être supprimé des observations importantes lors du sous-échantillonnage. Pour le XGBoost avec SMOTE, nous n'avons pas pu tester plus de 60 combinaisons d'hyperparamètres (contre 120 combinaisons pour le XGBoost sans ré-échantillonnage) afin d'optimiser le modèle puisque la base d'entraînement ré-échantillonnée est trop grande (augmentation des données de 30% suite au SMOTE).

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
XGBoost optimisé	0.405	75.7%	42%	62%	34.87%
XGBoost SMOTE optimisé	0.409	71%	40%	64%	32.23%
XGBoost *Random Under Sampling* optimisé	0.408	74.2%	41%	62%	33.20%

TABLE 2.12 – Tableau présentant les valeurs des métriques des modèles XGBoost de la première année construit avec ou sans ré-échantillonnage des données

Vu qu'en équilibrant les données, la représentation des parcours change, nous allons aussi comparer les performances par parcours avec les différents XGBoost construits. Ceci a permis de remarquer que les performances sont assez similaires quel que soit le modèle utilisé ce qui fait que ce n'est pas un élément qui a eu une influence sur le modèle XGboost que nous avons finalement choisi. Le tableau 2.13 suivant présente les valeurs des métriques avec la base test pour chaque parcours obtenues avec le XGBoost pour la première année. Il permet de constater que le modèle performe bien sur les parcours peu représentés.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
ADA	0.42	73%	29%	76%	22%
ATA	0.40	75%	42%	62%	35%
MM	0.39	76%	35%	63%	28%
MA	0.39	84%	54%	66%	79%
W	0.40	91%	77%	83%	77%

TABLE 2.13 – Tableau présentant les valeurs des métriques avec la base test pour chaque parcours obtenues avec le XGBoost pour la première année

Equilibrer les classes n'a pas permis d'améliorer les performances comme nous le souhaitions, nous allons donc garder le XGBoost avec les classes non équilibrées.

Une fois les modèles construits (GLM, *Random Forest* (RF) et XGBoost), nous allons les comparer afin de choisir le modèle que nous souhaitons utiliser pour la modélisation de la durée de vie.

2.3.4 Comparaison des différents modèles

Nous allons d'abord comparer les 10 variables les plus importantes des 3 modèles. L'importance des variables va non seulement permettre de vérifier que les modèles fournissent des résultats cohérents mais aussi d'avoir une idée sur les variables qui influencent la résiliation pendant la première année. Comparer l'importance des variables obtenues pour chaque modèle permet aussi de relever les similarités et les dissemblances.

Importance des variables

L'importance des variables pour un GLM représente les grandeurs des coefficients du GLM. Pour le RF et le XGBoost, l'importance des variables du package *h2o* est déterminée en calculant l'influence relative de chaque variable. Celle-ci s'obtient en prenant en compte le fait que cette variable ait été sélectionnée pour une division pendant le processus de construction de l'arbre et dans quelle mesure l'erreur au carré (sur tous les arbres) s'est améliorée après cette division. Le détail des calculs est mis en Annexe A.6.3.

La figure 2.2 suivante résume les 10 variables les plus importantes pour chaque modèle. L'importance a été normalisée. Les variables retrouvées avec les 3 modèles sont coloriées en bleu. Les 10 variables les plus importantes issues du GLM représentent 63% de l'importance totale des variables. Celles qui sont obtenues avec le RF représentent 67% de l'importance totale des variables tandis que celles issues du XGBoost représentent 64%.

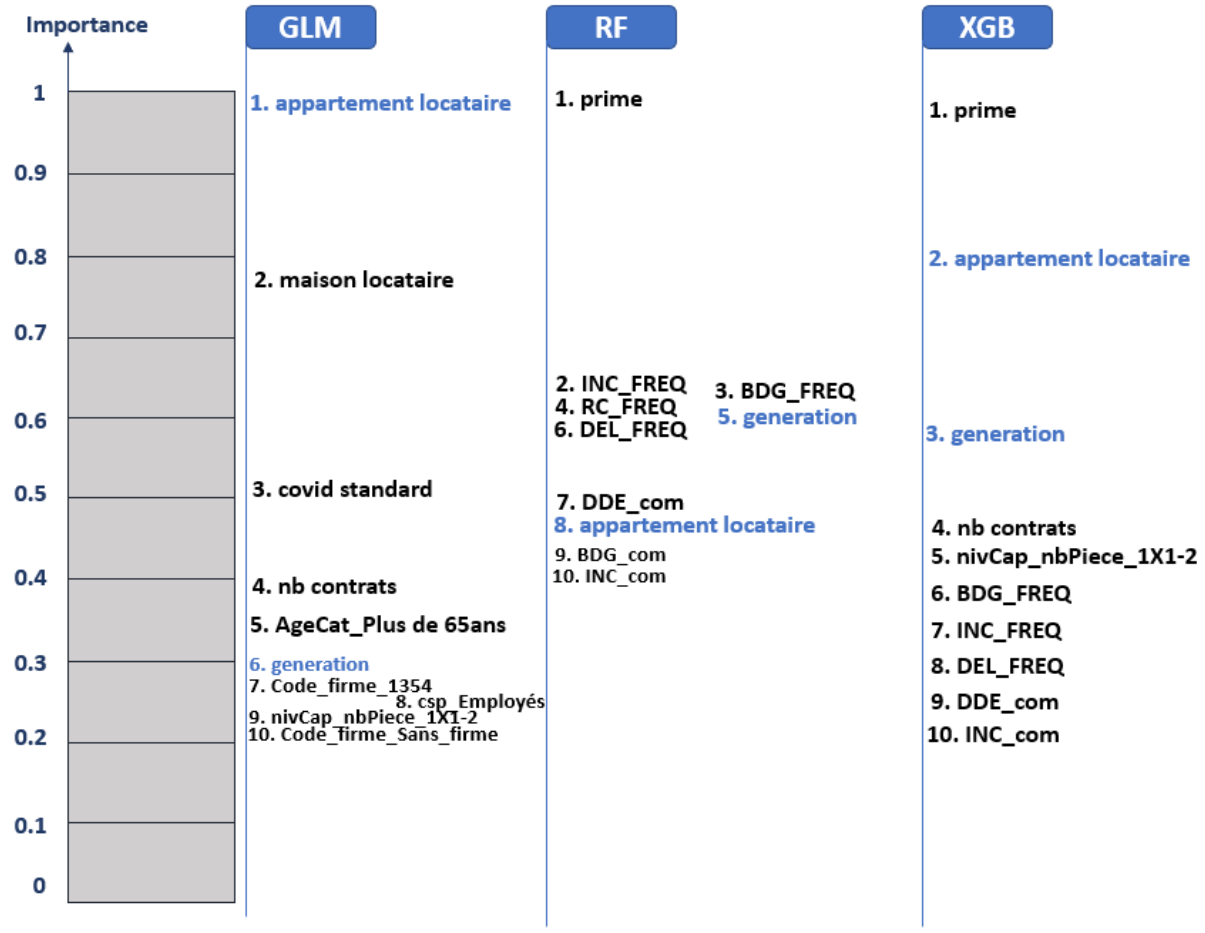


FIGURE 2.2 – Les 10 variables les plus importantes suivant les 3 modèles utilisés pour la modélisation des taux de résiliation de la première année

Nous remarquons que les variables les plus importantes ne sont pas les mêmes suivant le modèle.
— Les trois modèles ont la variable `appartement_locataire` parmi les 10 variables les plus importantes. Cependant, le classement de cette variable n'est pas le même. C'est la variable la plus importante pour le GLM, la deuxième pour le XGBoost et la huitième pour le *Random Forest*. Ils ont aussi en commum la variable *generation* coloriée en bleu.

— La prime est la variable la plus importante pour le XGBoost et le *Random Forest* alors qu'elle n'est pas parmi les 10 variables les plus importantes du GLM.
— Avec le XGBoost et le *Random Forest*, les zoniers (incendie, bris de glace et autres) représentent 50% ou plus des 10 variables les plus importantes alors qu'aucun zonier ne figure parmi les 10 variables les plus importantes du GLM.
Les différences trouvées pourraient s'expliquer par le fait que le XGBoost et le *Random Forest* n'utilisent pas toutes les variables pour construire les arbres (*feature sampling*). Lorsque deux variables sont corrélées, l'une d'entre elles est évitée lors de la sélection des variables pour construire l'arbre de décision. Le GLM sans pénalisation ne possède pas cette capacité à sélectionner les variables pour lutter contre la corrélation. Des variables corrélées ont donc parfois une importance inférieure à leur importance réelle (à cause de la corrélation). Par exemple, si deux variables A et B sont corrélées et que la variable A est la seule variable gardée, son importance est x alors que si les deux variables sont gardées, l'importance x pourrait être partagée[30] entre les variables A et B. Cela pourrait donc expliquer pourquoi la prime n'est pas aussi importante pour le GLM.
Aussi lorsque le GLM ne trouve pas de relation linéaire avec une variable, il pourrait considérer la variable comme n'étant pas importante. Ce qui pourrait expliquer pourquoi les zoniers ne sont pas parmi les variables les plus importantes du GLM.

Importance des parcours

La figure 2.3 suivante présente l'allure de l'importance des parcours obtenue quel que soit le modèle (GLM, RF ou XGBoost) pour la première année. La même allure est observée avec les modèles construits après ré-échantillonnage de la base (SMOTE et *Random Under Sampling*), le ré-échantillonnage ayant permis d'augmenter la représentation des parcours avec peu de volume.

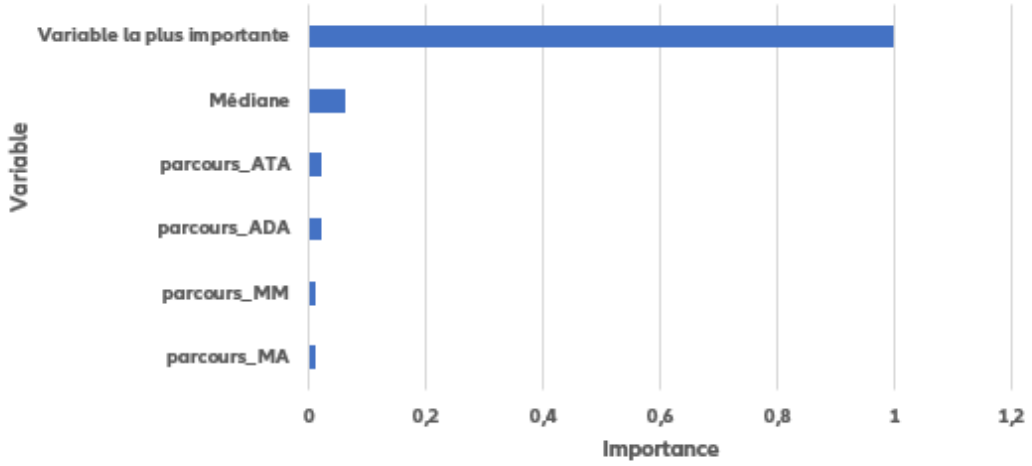


FIGURE 2.3 – Allure de l'importance des parcours quel que soit le modèle de la première année - Comparaison avec la variable la plus importante et l'importance médiane

Nous remarquons que le parcours est très peu important quel que soit le modèle pendant la première année. Nous pouvons donc dire que le parcours a très peu d'impact sur la résiliation.

30. Notons aussi que pour le *Random Forest*, l'importance de 2 variables corrélées pourrait être partagée entre elles si certains arbres choisissent la variable A et les autres la variable B.

Performances des modèles

Nous comparons les meilleures performances obtenues avec les 3 modèles afin de choisir le meilleur modèle pour le calcul des durées grâce au tableau 2.14 suivant.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
GLM	0.423	72.8%	35.17%	62.84%	31.51%
RF	0.406	75.6%	43%	63%	34%
XGBoost	0.405	75.7%	42%	62%	34.87%

TABLE 2.14 – Tableau comparant les valeurs des métriques des trois modèles utilisés pour la modélisation des taux de résiliation pendant la première année

Nous remarquons que les performances des modèles sont juste acceptables. La meilleure précision obtenue ne dépassant pas 35% pourrait s'expliquer par le déséquilibre des classes. Rappelons cependant qu'équilibrer les classes a globalement permis d'améliorer le rappel mais a baissé les valeurs des autres métriques. Les variables explicatives utilisées semblent aussi ne pas être en mesure d'expliquer à elles seules les résiliations pendant la première année.

Choix du modèle pour la première année

Les différences entre les valeurs des différentes métriques sont faibles en général. Le rappel tourne autour de 62% et 63% pour tous les modèles tandis que la précision tourne autour de 31% et 34%. l'AUCPr est la métrique qui permet la plus de différencier les modèles puisque le GLM a une AUCPr de 35% contre 43% pour le *Random Forest*. Nous avions aussi vu dans la section individuelle de chaque modèle que les performances suivant les différents parcours sont bonnes quel que soit le modèle. Il est important de s'assurer que le modèle ne néglige pas les parcours les moins bien représentés. Le tableau 2.14 permet aussi de remarquer que le *Random Forest* a un rappel et une AUCPr légérement meilleurs que ceux des autres modèles. Pour les autres métriques, il a la deuxième meilleure valeur. Ses valeurs sont cependant très proches de celles du XGBoost, ce qui nous amène à hésiter entre le *Random Forest* et le XGBoost pour modéliser les taux de résiliation à un an.

Néanmois, n'oublions pas que nous voulons calculer des durées de vie moyennes. Il faut donc que les durées de vie estimées par le modèle soit cohérentes avec celles qui sont réellement observées. Pour cela, il faut que les taux de résiliation de chaque année soient proches des taux de résiliation réels. Nous allons donc comparer les taux de résiliation observés sur la base test et les taux de résiliation prédits suivant la variable la plus discriminante commune à tous les modèles : la qualité juridique. Le tableau 2.15 suivant permet de comparer les taux de résiliation obtenus suivant le modèle et les taux de résiliation réels de la base test pour la première année.

Modèle	Locataire	Propriétaire	Total
GLM	29%	3%	21%
RF	26%	3.8%	19.3%
XGBoost	26%	4%	19.4%
taux de résiliation réél	23%	5%	17.6%

TABLE 2.15 – Tableau comparant les taux de résiliation obtenus suivant le modèle et les taux de résiliation réels de la base test pour la première année

Nous remarquons d'abord que les modèles ont tendance à sur-estimer les taux de résiliation des

locataires et sous-estimer les taux de résiliation des propriétaires. Cela pourrait s'expliquer par le fait que les propriétaires ont un taux de résiliation faible (moins de 5% pour les propriétaires alors que le taux de résiliation des locataires dépasse 20%). Cela entraîne que les modèles vont avoir plus de mal à détecter les résiliations des propriétaires vu qu'elles sont plus susceptibles d'être «ignorées» à cause de leur volume faible.
Nous remarquons aussi que le XGBoost et le *Random Forest* prédisent tous les deux les taux de résiliation les plus proches des taux de résiliation observés dans la base test. **Nous confirmons donc que le XGBoost peut être choisi aussi bien que le *Random Forest* pour modéliser les taux de résiliation pour la première période d'une année. Regarder les performances des autres périodes peut aider à choisir un modèle global pour modéliser les durées de vie.**

Après avoir effectué la modélisation des taux de résiliation pendant la première année suivant une méthodologie définie, nous allons appliquer la même méthodologie pour modéliser les taux de résiliation des périodes suivantes.

2.4 Modélisation des périodes d'une année suivantes

Nous souhaitons modéliser maintenant les taux de résiliation pendant l'année N sachant que les contrats ont survécu jusqu'à l'année $N-1$, $2 \leq N \leq 5$. Nous avions vu que plus le nombre d'années passées par le contrat dans le portefeuille augmente, plus le taux de résiliation baisse. Nous allons donc bien évidemment continuer à tester les méthodes de ré-échantillonnage utilisées pendant la première période (SMOTE et *Random Under Sampling*) afin d'équilibrer les deux classes. Ce n'est pas parce que ces méthodes n'ont pas été concluantes pendant la première période qu'elles ne le seront pas pour les autres périodes.

Rappelons que nous souhaitons modéliser la durée de vie des affaires nouvelles (AFN). Nous utilisons donc toujours les variables dont nous disposons à l'AFN. Cela entraîne que pour chaque période de modélisation, les mêmes variables explicatives sont gardées. Nous savons qu' au cours de la période ou d'une période à une autre, plusieurs variables sont susceptibles de changer de valeur :
— suite à un déménagement : changement du zonier,
— après échéance : majoration de la prime par exemple,
— face aux nouveaux besoins du client : changement de garantie (changement du niveau de la garantie vol) par exemple, etc.
Mais nous ne pouvons pas prendre en compte ces changements à moins que nous ne soyions capables de les prédire à partir des variables vues à l'affaire nouvelle. Nous ne disposons pas suffisamment de temps pour cela. Les mêmes variables que la modélisation de la première période vont donc être gardées. Cependant, il y a des traitements (exclusifs à ces périodes) sur la base de données à faire avant d'amorcer la modélisation. Nous allons donc procéder de la manière suivante afin de modéliser les taux de résiliation des autres années.

2.4.1 Préprocessing

Il faut s'assurer que les contrats que nous allons utiliser pour la modélisation des taux de résiliation d'une période vont pouvoir être entièrement observés pendant la période donnée. Les résiliations sont observées du 1er Janvier 2015 au 31 Décembre 2020. L'ancienneté des contrats est calculé au 31 Décembre 2020. Ainsi, les contrats que nous utilisons pour les 3 dernières périodes de modélisation doivent vérifier deux conditions. Pour la **période** N :
(a) il faut que les contrats n'aient pas été résiliés pendant la période $N-1$. Il faut donc supprimer les

contrats qui ont été résiliés pendant la période $N-1$.
(b) il faut que leur ancienneté soit supérieure ou égale à N années au 31 Décembre 2020. Il faut donc supprimer les contrats qui n'ont pas été résiliés mais qui ont moins de de N années entières d'ancienneté.
Par exemple, pour la **deuxième période** :
(a) il faut que les contrats n'aient pas été résiliés pendant la première année. Il faut donc supprimer les contrats qui ont été résiliés pendant la première année.
(b) il faut que leur ancienneté soit supérieure ou égale à 2 ans au 31 Décembre 2020. Il faut donc supprimer les contrats qui n'ont pas été résiliés mais qui ont moins de deux ans d'ancienneté. Nous supprimons donc les AFN de 2019.

Ainsi, (dans la base) il n'y aura plus les parcours Web qui n'ont qu'une année (entière) d'ancienneté. Comme nous avions vu que les parcours semblent ne pas avoir d'impact sur la résiliation pendant la première année, nous pouvons supposer que c'est le cas pour les autres années pour les parcours qui n'ont qu'une seule année d'ancienneté. Si le parcours reste très peu important pour les autres années, nous n'aurons donc plus besoin d'exploiter l'une des deux pistes décrites en Annexe A.8) afin de modéliser les parcours avec une ancienneté inférieure à la durée totale modélisée. Nous procédons de la même manière que la première période pour le reste de la modélisation. Nous utilisons les 3 mêmes modèles que la première période avec la même méthodologie.

Puis une fois les modèles construits, nous allons analyser l'importance des variables pour vérifier si elle est cohérente et pour visualiser les variables les plus discriminantes pour les modèles.

2.4.2 Importance des variables

Nous allons d'abord analyser l'importance des variables qui va permettre de vérifier que les modèles sont cohérents c'est-à-dire qu'ils ne donnent pas des informations allant à l'encontre de nos connaissances. Nous n'allons par exemple jamais avoir confiance à un modèle qui dit que les propriétaires ont plus de chance de résilier que les locataires. Elle va aussi nous aider à voir les variables (notamment les parcours) qui ont une influence sur la résiliation pendant l'année donnée.

Deuxième année

La figure 2.4 présente les 10 variables les plus importantes des différents modèles pendant la deuxième période. L'importance a été normalisée. Les variables retrouvées avec chaque modèle sont coloriées en bleu.
Comme pour la première année, nous remarquons que nous ne retrouvons pas les mêmes variables les plus importantes dans les 3 modèles. Cependant, la plupart des variables sont généralement les mêmes que celles de la première année. La génération (l'année de souscription) est la variable la plus importante. Cela pourrait s'expliquer par le fait que les contrats dont l'observation des résiliations est faite pendant la Covid-19 (en 2020) ont moins de chance d'être résiliés. Les autres variables deviennent aussi moins importantes surtout celles qui ont pu changer de valeur entre l'affaire nouvelle et la deuxième année. En effet, nous ne prenons pas en compte les avenants ce qui introduit un biais considérable. Par exemple, le modèle pourrait trouver une résiliation avec un niveau de capital (croisé avec le nombre de pièces) égal à $nc1$ (à l'affaire nouvelle) alors que la valeur de celui-ci a changé à $nc2$. La résiliation trouvée est réellement associée au niveau de capital $nc2$, utiliser $nc1$ "embrouillerait" le modèle et engendrerait donc un biais.
Nous remarquons ainsi que pour la deuxième année, l'importance de la prime a baissé quel que soit le modèle. Rappelons qu'à la première année, la prime est la variable la plus importante pour le *Random*

Forest et le XGBoost. Cela pourrait s'expliquer par le biais introduit par la majoration que vont subir les contrats. Cela pourrait aussi s'expliquer par le fait qu'il y ait moins de clients qui sont sensibles au prix vu qu'une bonne partie ont pu résilié pendant la première année pour aller trouver un tarif affaire nouvelle plus avantageux chez un autre assureur.
Nous ne retrouvons les variables zoniers qu'avec le XGBoost mais leur importance a baissé. Certains clients ont pu déménager et donc changer de zonier.
La variable locataire d'appartement reste toujours aussi importante parce que les locataires ont toujours plus de chance de résilier que les propriétaires quel que soit le temps passé par le contrat dans le portefeuille. Il y a aussi peu de clients qui changent de qualité juridique en avenant.
La variable nombre de contrats reste toujours importante. Elle reste cependant une variable susceptible de changer dans le temps. Rappelons que la variable *nombre de contrats* n'a que deux modalités : 1 contrat (le client n'a que le contrat d'assurance observé comme contrat chez Allianz) ou plusieurs contrats (2+). Certains clients qui n'avaient que le contrat d'assurance habitation peuvent souscrire un autre contrat chez Allianz, devenir multi-détenteurs et donc la valeur de la variable passerait de 1 à 2+ contrats.

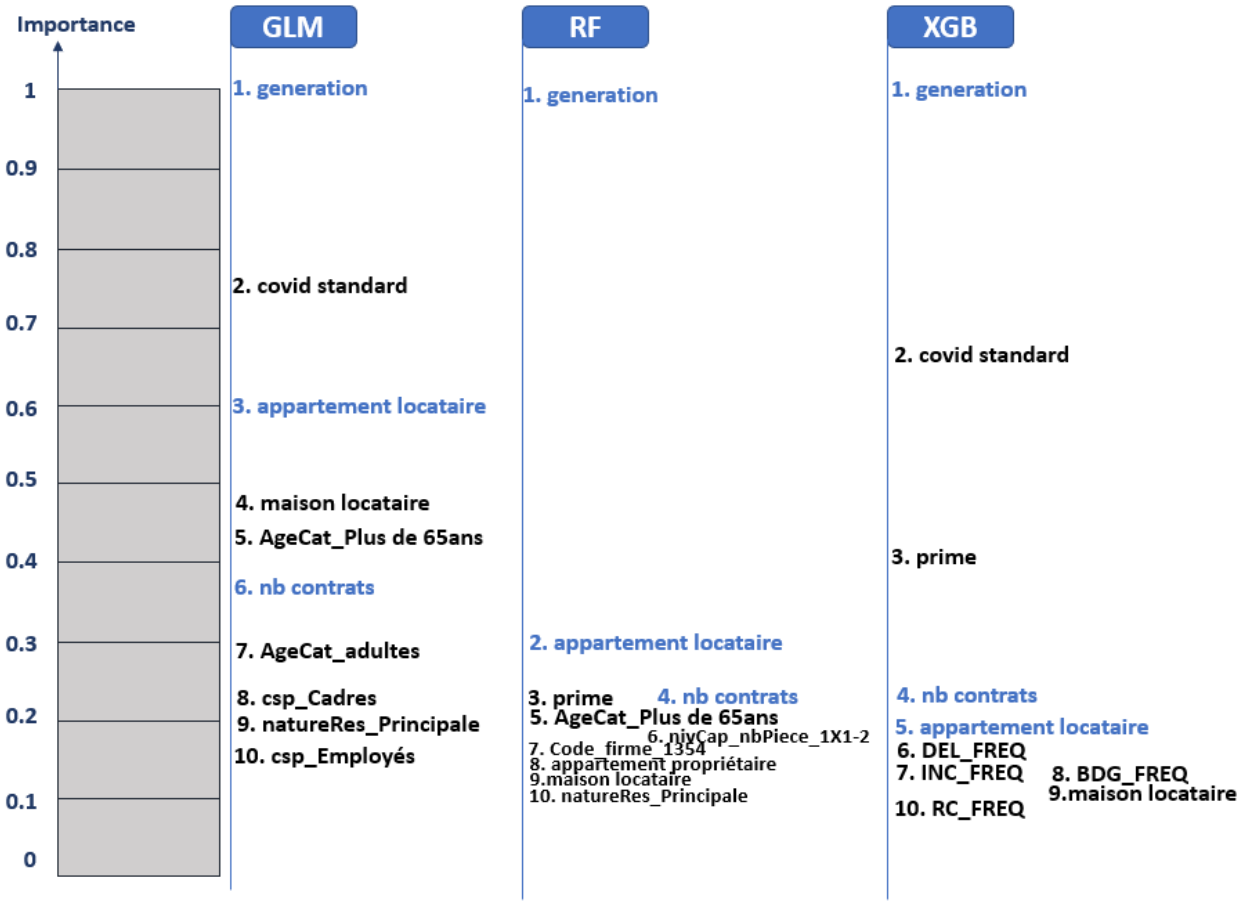


FIGURE 2.4 – Les 10 variables les plus importantes suivant les 3 modèles utilisées pour la modélisation des taux de résiliation de la deuxième année

Troisième année

La figure 2.5 suivante présente les 10 variables les plus importantes des différents modèles pendant la troisième période. L'importance a été normalisée. Les variables retrouvées avec chaque modèle sont coloriées en bleu.

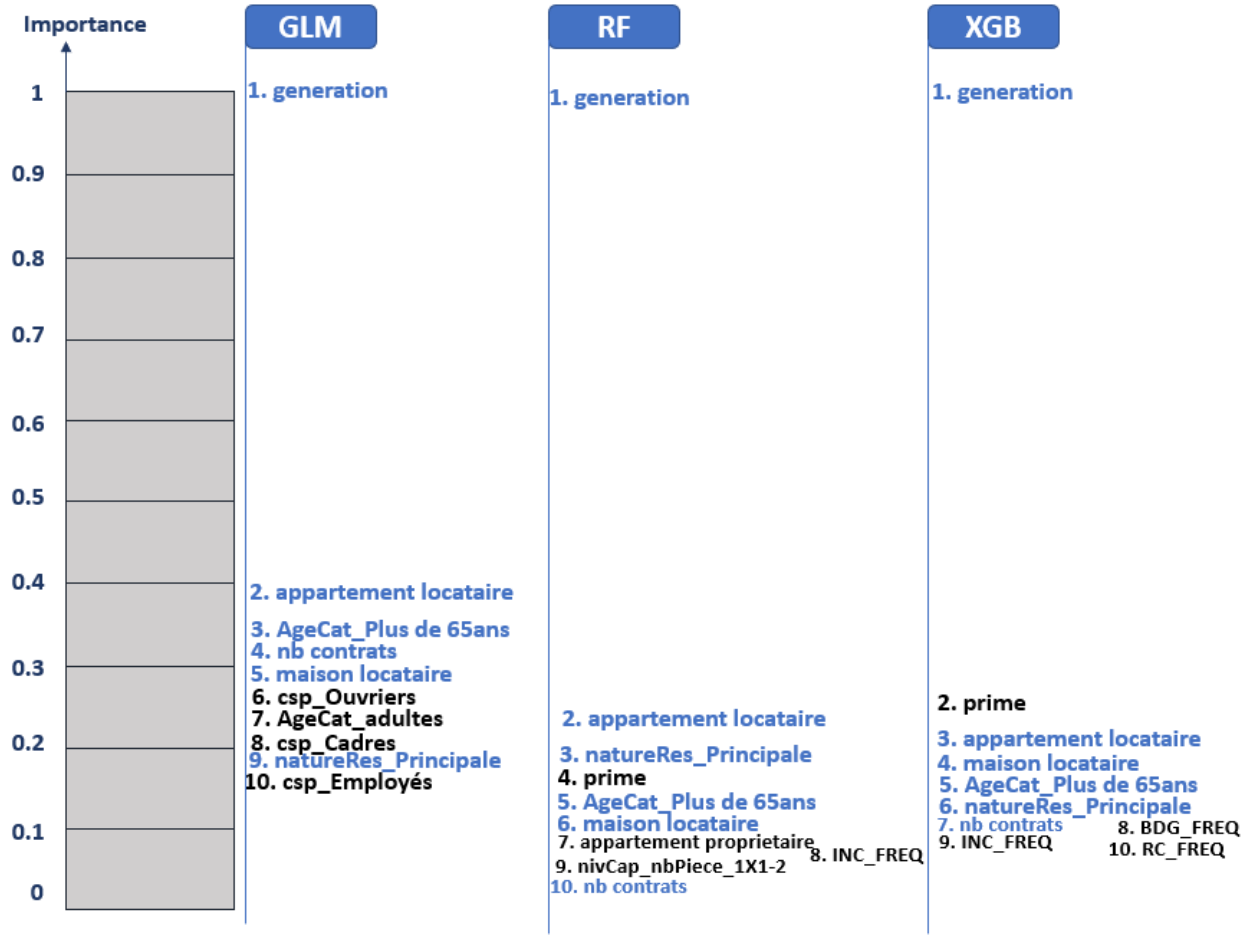


FIGURE 2.5 – Les 10 variables les plus importantes suivant les 3 modèles utilisés pour la modélisation des taux de résiliation de la troisième année

Nous remarquons que globalement, les variables les plus importantes ne changent pas beaucoup par rapport aux autres années. Elles deviennent toutes beaucoup moins importantes par rapport aux années précédentes au profit de la génération. Leur importance est inférieure à 0,4. Cela pourrait s'expliquer par le fait que nous ne prenons pas en compte les avenants ce qui introduit un biais considérable.

Quatrième année

La figure 2.6 suivante présente les 10 variables les plus importantes des différents modèles pendant la quatrième période. L'importance a été normalisée. Les variables retrouvées avec chaque modèle sont coloriées en bleu.
Nous faisons en général la même remarque que pour la période de trois ans. Nous voyons qu'à part la génération, les variables ont une importance inférieure à 0,3. Les variables voient leur importance

diminuer considérablement au profit de la génération. Cela pourrait s'expliquer par le fait que nous ne prenons pas en compte les avenants ce qui introduit un biais considérable alors qu'à la quatrième année du contrat dans le portefeuille, il y a encore plus de chance qu'il y ait des avenants.

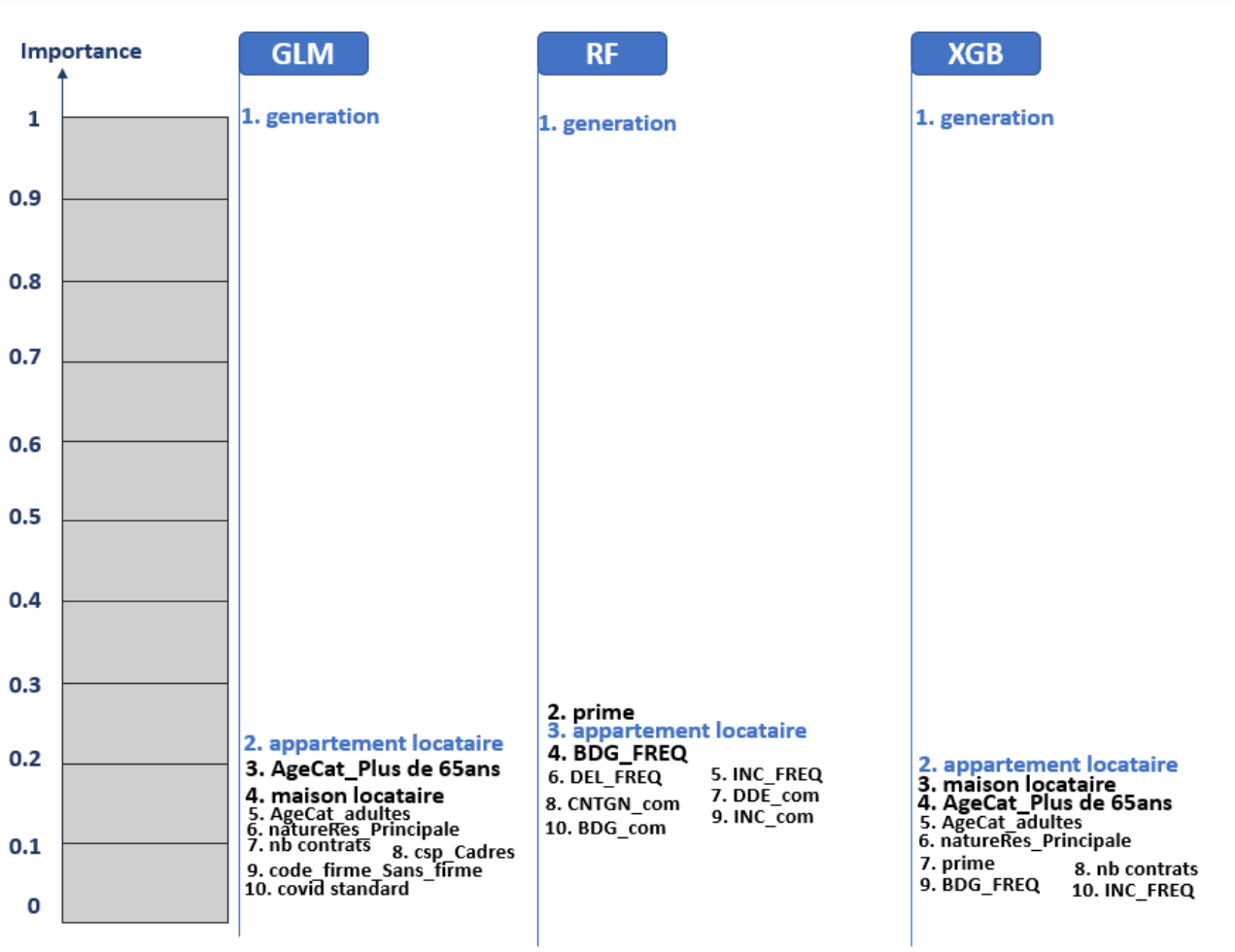


FIGURE 2.6 – Les 10 variables les plus importantes suivant les 3 modèles utilisés pour la modélisation des taux de résiliation de la quatrième année

Importance et significativité des parcours

Le tableau 2.16 suivant permet de comparer les p-value obtenus suivant les différents parcours (ayant toujours des contrats pendant la période donnée afin de construire le modèle). Le symbole «-» montre que le parcours n'est plus présent dans la base de données pour la période étudiée. La valeur *reference* indique que le parcours est pris comme référence et a donc été supprimé.

Nous remarquons qu'il n'y a que le parcours MA qui est significatif pour la deuxième année. Son *odds ratio* est 0,995 donc le parcours a très peu d'impact sur la résiliation (le contrat aurait 0,5% moins de chance d'être résilié en étant dans les parcours MA plutôt que les parcours MM). Pour les autres années, aucun parcours n'est significatif. Ainsi, d'après le GLM, le parcours a très peu d'influence sur la résiliation pendant les deuxième, troisième et quatrième années.

Parcours	P-value période 2	P-value période 3	P-value période 4
ADA	0,54	0,13	0,22
ATA	0,24	0,76	0,15
MM	reference	reference	reference
MA	0,003	0.97	-

TABLE 2.16 – P-value des parcours avec le modèle GLM pour les deuxième, troisième et quatrième années

Nous allons ensuite regarder l'importance des parcours pour chacun des modèles et pour les différentes périodes afin de vérifier si nous allons faire le même constat. La figure 2.7 suivante donne l'allure de l'importance des parcours quel que soit le modèle de la deuxième année.

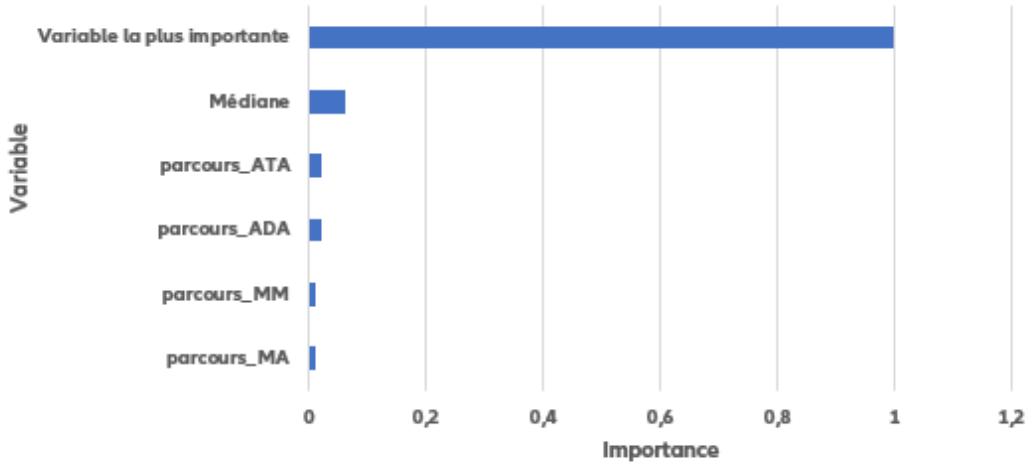


FIGURE 2.7 – Allure de l'importance des parcours quel que soit le modèle de la deuxième année - Comparaison avec la variable la plus importante et l'importance médiane

Nous constatons que quel que soit le modèle[31], l'importance des parcours est très faible pour la deuxième année. Nous retrouvons le même résultat à la troisième et quatrième année. Cela nous pousse alors à croire que les parcours n'ont pas d'impact sur la résiliation pour les deuxième, troisième et quatrième années.

2.4.3 Performances des modèles et choix du meilleur modèle

Nous utilisons les mêmes métriques que celles qui ont permis d'évaluer les modèles de la première période. Nous allons aussi procéder de la même manière afin de déterminer quel(s) modèle(s) choisir pour la modélisation des durées de vie c'est-à-dire en faisant une :

— Comparaison des valeurs des métriques des différents modèles,

— Comparaison des valeurs des métriques des différents modèles suivant les parcours pour éviter de choisir un modèle qui «néglige» certains parcours,

31. Sont inclus les modèles construits après ré-échantillonnage de la base (SMOTE et *Random Under Sampling*), le ré-échantillonnage ayant permis d'augmenter la représentation des parcours avec peu de volume.

— Comparaison des taux de résiliation prédits par les différents modèles avec ceux observés réellement dans la base test suivant la variable discriminante pour tous les modèles (la qualité juridique) pour s'assurer que les durées de vie moyennes prédites vont être cohérentes avec celles qui sont réellement observées.

Le tableau 2.17 suivant présente les valeurs des métriques calculées avec la base test suivant le modèle de taux de résiliation utilisé pour la **deuxième année**.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
GLM	0.394	70%	36%	65.2%	25.11%
RF	0.395	72%	33.7%	64.3%	27%
XGBoost	0.393	72.20%	34%	62%	27.6%

TABLE 2.17 – Tableau présentant les valeurs des métriques suivant le modèle de taux de résiliation pour la deuxième année

Le tableau 2.18 suivant présente les valeurs des métriques calculées avec la base test suivant le modèle de taux de résiliation utilisé pour la **troisième année**.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
GLM	0.293	74.1%	21%	67%	20%
RF	0.3	74.1%	20.8%	67%	20%
XGBoost	0.3	74.3%	21%	69%	20%

TABLE 2.18 – Tableau présentant les valeurs des métriques suivant le modèle de taux de résiliation pour la troisième année

Le tableau 2.19 suivant présente les valeurs des métriques calculées avec la base test suivant le modèle de taux de résiliation utilisé pour la **quatrième année**.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
GLM	0.21	76.3%	15.7%	58%	16%
RF	0.21	76%	15.2%	56%	16%
XGBoost	0.20	76.4%	15.8%	61%	16%

TABLE 2.19 – Tableau présentant les valeurs des métriques suivant le modèle de taux de résiliation pour la quatrième année

Nous remarquons d'abord que les performances se dégradent beaucoup trop à partir de la quatrième année. La précision est égale à 16% à la quatrième année quel que soit le modèle. Cela veut dire que parmi les résiliations que les modèles prédisent, seul 16% sont des résiliations réelles.
Nous nous arrêtons donc à la quatrième année pour la modélisation.

En comparant les performances suivant les différentes périodes, nous avons aussi remarqué que le logloss ne change pas beaucoup suivant la période (il varie entre 0,3 et 0,2). Ce n'est pas le cas de l'AUCPr et la précision dont les valeurs diminuent de plus en plus chaque année. Elles passent de 36% à environ 15% pour l'AUCPr et 27% à 16% pour la précision. Cela pourrait s'expliquer par le fait que chaque année, le taux de résiliation diminue considérablement. Nous avons vu dans la première partie qu'elle passe de 18% à 10% de la première à la quatrième année. Les deux classes des modèles deviennent de plus en plus déséquilibrées, ce qui expliquent que les modèles ont plus de mal à bien identifier les résiliations. De même, en avancant dans les périodes, les variables deviennent de moins en

moins explicatives. Nous l'avons déjà vu pendant l'analyse de l'importance des variables. L'importance de presque toutes les variables baisse chaque année sauf celle de la génération. Cela s'explique par le fait que nous gardons la vision des variables à l'affaire nouvelle alors qu'elles sont susceptibles de changer au cours du temps. Il y a de très forte chance que la prime augmente à cause de la revalorisation. Il y a aussi plusieurs autres variables qui pourraient changer de valeurs.

Choix du modèle

Nous remarquons que les valeurs des métriques des différents modèles sont proches quelle que soit l'année. Le logloss a une différence maximale de l'ordre de 10e-2 (quatrième année) et l'AUC de 2% (deuxième année). La plus grande différence de valeurs se note avec le rappel qui atteind 5% à la quatrième année. Nous pouvons donc globalement choisir un seul modèle pour modéliser les taux de résiliation.

Même si les performances des différents modèles sont proches dans l'ensemble (en prenant aussi en compte la première année), le meilleur modèle sur les quatre périodes semble être le XGBoost. Il a la meilleure AUC sur les 3 dernières périodes, les meilleurs AUCPr et rappel sur les deux dernières périodes. Rappelons aussi que c'est l'un des meilleurs modèles [32] pour la modélisation de la première période (avec le RF).

Avant de confirmer le choix du modèle, nous avons aussi regardé les valeurs des métriques suivant le modèle et par parcours. Elles sont assez semblables d'un modèle à l'autre. Ce n'est donc pas un élément qui a été discriminant dans le choix du meilleur modèle. Nous pouvons voir dans le tableau 2.20 suivant, les performances par parcours du modèle XGBoost pendant la deuxième [33] année.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
ADA	0.401	71%	35%	61%	24%
ATA	0.392	72%	34%	62%	27%
MM	0.313	74%	40%	65%	28%
MA	0.399	79%	41%	61%	29%

TABLE 2.20 – Tableau présentant les valeurs des métriques pour chaque parcours obtenues avec le XGBoost pour la deuxième année

Pour confirmer notre choix pour le XGBoost, il faut aussi vérifier si les taux de résiliation qu'il prédit sont proches des taux de résiliation réels. Le tableau 2.21 suivant permet de comparer les taux de résiliation prédits par les modèles et les taux de résiliation réels sur la base test pour la deuxième [34] année suivant la qualité juridique.

Nous remarquons qu'au total, les taux trouvés pour la deuxième année sont proches des taux réels observés dans la base test (même remarque pour les autres années). En faisant la distinction suivant la qualité juridique, nous nous rendons compte qu'avec le XGBoost et le *Random Forest*, les taux de résiliation prédits pour les locataires sont supérieurs aux taux de résiliation réels alors que les taux de résiliation prédits pour les propriétaires sont inférieurs aux taux de résiliation réel. Le XGBoost est un des modèles dont les taux de résiliation prédits se rapprochent le plus des taux de résiliation réels suivant la qualité juridique.

Nous choisissons donc le XGBoost pour modéliser les durées de vie par période d'une année.

32. voir sous-section 2.3.4
33. Le tableau des deux dernières années est mis en Annexe B.4.
34. Le tableau des deux dernières années est mis en Annexe B.4.

Modèle	Locataire	Propriétaire	Total
GLM	20%	5%	13%
RF	23%	1%	13%
XGBoost	24%	4%	14.8%
taux de résiliation réél	21%	7%	14.6%

TABLE 2.21 – Tableau comparant les taux de résiliation prédits par les modèles et les taux de résiliation réels sur la base test pour la deuxième année suivant la qualité juridique

Nous avons donc à présent fini de construire les modèles pour la modélisation des taux de résiliation par période d'une année et nous savons quel modèle utilisé pour modéliser les taux de résiliation par période d'une année. Nous sommes cependant tentés de voir si en utilisant une période beaucoup plus petite, nous n'aurions pas des durées de vie plus précises.

2.5 Modélisation des taux de résiliation avec un horizon plus petit

2.5.1 Détermination d'un horizon de modélisation plus petit

Nous souhaitons prédire la valeur de la variable binaire

$$\begin{aligned} &Y = 1 \text{ si le contrat a été résilié,} \\ &Y = 0 \text{ sinon.} \end{aligned} \tag{2.26}$$

en fonction des variables explicatives présentes dans la base.

Nous pouvons modéliser par période d'une année c'est-à-dire modéliser les taux de résiliation à]0,1 an] puis]1 an, 2 ans] jusqu'à]3 ans, 4 ans] comme nous venons de le faire. Mais nous aimerions bien avoir un horizon de modélisation plus petit pour que le calcul des durées soit plus précis. Nous savons que nous pourrions avoir un problème lié au déséquilibre des classes beaucoup plus sévère vu que les taux de résiliation obtenus peuvent être largement inférieurs aux taux de résiliation par année. La modélisation va également prendre plus de temps vu qu'il y aura plus de périodes à modéliser pour arriver à 4 ans. Pour déterminer cette période, nous calculons les taux de résiliation au cours de chaque mois de l'année afin de déterminer à partir de combien de mois, le taux de résiliation est suffisamment élevé pour être utilisé comme horizon de modélisation.

La figure 2.8 suivante donne les taux de résiliation par cadence mensuelle pendant la première année du contrat suivant la génération (année de souscription du contrat). Nous traçons les taux de résiliation suivant l'année de souscription des contrats afin de vérifier leur stabilité dans le temps par cadence mensuelle. Nous souhaitons que la résiliation à échéance ne soit pas comprise dans une période de modélisation ainsi il faut que la période de modélisation que nous allons choisir soit un diviseur de 12. 6 est le plus grand diviseur de 12 et à 6 mois, le taux de résiliation semble suffisamment élevé (entre 6% et 8%). **Nous choisissons donc 6 mois comme horizon de modélisation des taux de résiliation**.
Nous souhaitons alors faire la modélisation pendant plusieurs périodes de 6 mois :]0 mois, 6 mois],]6 mois, 12 mois],]12 mois, 18 mois] ...]42 mois, 48 mois].

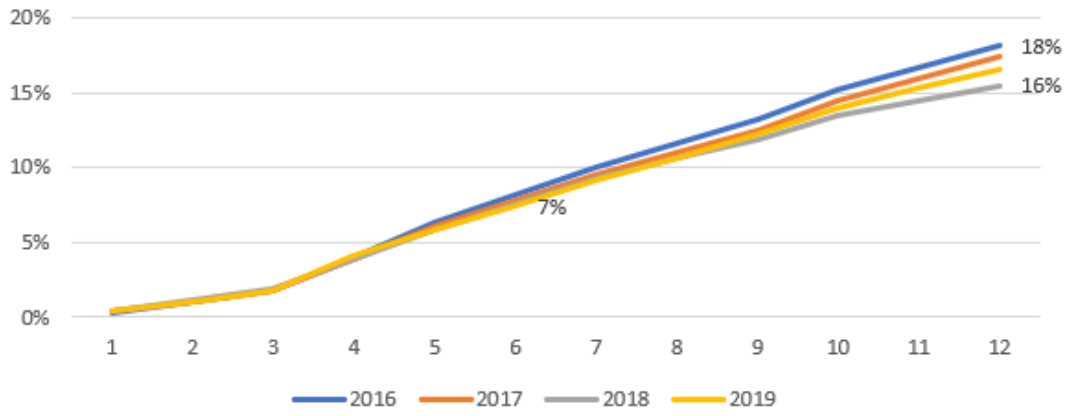


FIGURE 2.8 – Taux de résiliation et nombre d'AFN suivant l'année de souscription pendant les 12 premiers mois du contrat

2.5.2 Modélisation

Le processus est similaire à celui de la modélisation par période d'une année sauf que 1 an est remplacé par 6 mois. Nous pouvons donc garder les affaires nouvelles (AFN) de 2020 qui ont pu être observées pendant plus de 6 mois (les affaires nouvelles du 1er Janvier au 30 Juin 2020). Nous souhaitons donc modéliser :
1. les taux de résiliation pendant la première période de 6 mois :]0 mois, 6 mois].
2. les taux de résiliation pendant la N-ième période de 6 mois (par exemple avec N=2 :]6,12] mois) sachant que les contrats ont survécu jusqu'à la $N-1$ ième période.

Ainsi, les contrats utilisés pour les périodes de modélisation doivent vérifier des conditions. Pour la :

— **période]0,6] mois** :
(a) il faut que l'ancienneté des contrats soit supérieure ou égale à 6 mois au 31 Décembre 2020. Nous supprimons donc les contrats qui ont moins de 6 mois d'ancienneté (ils correspondent aux affaires nouvelles qui ont pris effet à partir du 1er Juillet 2020).

— **période N de 6 mois** :
(a) il faut que les contrats n'aient pas été résiliés pendant la période $N-1$. Pour la deuxième période de 6 mois par exemple, nous supprimons les contrats qui ont résiliés pendant leurs 6 premiers mois.
(b) il faut que l'ancienneté des contrats soit supérieure ou égale à $N \times 6$ mois au 31 Décembre 2020. Pour la deuxième période de 6 mois par exemple, les contrats qui n'ont pas été résiliés mais qui ont moins de 12 mois d'ancienneté vont être supprimés (ils correspondent aux affaires nouvelles de 2020).

Nous utilisons toujours comme modèles le GLM pour sa facilité d'interprétation, le *Random Forest* et le XGBoost. Pour les construire, nous allons suivre la même méthodologie que la modélisation par période d'une année. Nous pouvons voir sur la figure 2.8 que le déséquilibre des classes est plus présent vu que les taux de résiliation par 6 mois peuvent être égaux à 7% pour la période]0,6] mois par exemple.
Nous allons ensuite comparer pour chaque période de 6 mois les performances des 3 modèles afin de retenir le meilleur modèle. Nous allons aussi comparer ces performances à celles des modèles obtenues avec la modélisation par période d'une année. Si les performances sont jugées trop faibles comparées à

celles de la modélisation par période d'une année, nous allons garder la modélisation par période d'une année pour calculer les durées. Le but de la modélisation à 6 mois au lieu de 1 an est de permettre de pouvoir calculer des durées plus précises. Si les performances sont beaucoup trop faibles, nous préférerons garder la modélisation à un an puisque la modélisation à 6 mois risque d'aboutir sur des résultats moins exacts.

2.5.3 Performance des modèles

Période]0, 6] mois

Le tableau 2.22 compare les valeurs des métriques des trois modèles utilisés pour la modélisation des taux de résiliation pendant la période]0 mois, 6 mois]. Nous remarquons que les performances sont correctes. Nous avons même une AUC de 80% avec le XGBoost mais la précision ne dépasse pas 42%.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
GLM	0.244	77%	28.2%	55.5%	22.47%
RF	0.217	79.7%	40%	64.3%	38.9%
XGBoost	0.216	80.5%	38.96%	65%	42%

TABLE 2.22 – Tableau comparant les valeurs des métriques (obtenues avec la base test) des trois modèles utilisés pour la modélisation des taux de résiliation pendant la période]0 mois, 6mois]

Période]6, 12] mois

Le tableau 2.23 compare les valeurs des métriques des trois modèles utilisés pour la modélisation des taux de résiliation pendant la période]6 mois, 12 mois]. Nous remarquons que l'AUCPr et le rappel s'améliorent par rapport à la période précédente. Cela pourrait s'expliquer par le fait que le taux de résiliation de cette période est de 11% contre 7% pour la première période de 6 mois.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
GLM	0.31	73%	53%	60%	30%
RF	0.312	74%	54%	64%	29%
XGBoost	0.304	77%	54%	66%	31%

TABLE 2.23 – Tableau comparant les valeurs des métriques (obtenues avec la base test) des trois modèles utilisés pour la modélisation des taux de résiliation pendant la période]6 mois, 12 mois]

Période]12, 18] mois

Le tableau 2.24 compare les valeurs des métriques des trois modèles utilisés pour la modélisation des taux de résiliation pendant la période]12 mois, 18 mois]. Nous remarquons que les performances baissent beaucoup. L'AUC atteint à peine 70%, le rappel est inférieur ou égal à 48% et la précision ne dépasse pas 20%.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
GLM	0.284	66%	15%	42%	15%
RF	0.273	69%	21%	46%	17%
XGBoost	0.254	70%	22%	48%	20%

TABLE 2.24 – Tableau comparant les valeurs des métriques (obtenues avec la base test) des trois modèles utilisés pour la modélisation des taux de résiliation pendant la période]12 mois, 18 mois]

Période]18, 24] mois

Le tableau 2.25 compare les valeurs des métriques des trois modèles utilisés pour la modélisation des taux de résiliation pendant la période]18 mois, 24 mois].

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
GLM	0.236	71%	16%	42%	14%
RF	0.225	71%	15%	43%	14%
XGBoost	0.220	71%	17%	44%	14%

TABLE 2.25 – Tableau comparant les valeurs des métriques (obtenues avec la base test) des trois modèles que nous avons utilisés pour la modélisation des taux de résiliation pendant la période]18 mois, 24 mois]

A partir de la quatrième période de 6 mois, le rappel des modèles ne dépasse pas 44% alors que la précision est inférieure ou égale à 20% dès la période]12 mois, 18 mois] (troisième période). Nous remarquons que l'AUCPr a aussi beaucoup baissé à partir de la troisième période. Nous allons donc nous arrêter à cette période et comparer avec la modélisation à 1 an.

2.5.4 Choix entre la période de 6 mois et celle de 1 an

Nous allons regarder les performances des modélisations des différentes périodes de 6 mois et celles des différentes périodes d'une année. Nous allons choisir l'horizon qui va prédire les durées de vie les plus fiables et précises. Le tableau 2.26 suivant présente les valeurs des différentes métriques obtenues avec les modèles pour la dernière période de la modélisation à 1 an tandis que le tableau 2.27 suivant présente celles de la quatrième période de la modélisation à 6 mois.
Alors qu'il reste encore plusieurs périodes de 6 mois à modéliser, nous remarquons déjà que les performances obtenues avec la période]18, 24] mois (pour la modélisation à 6 mois) sont inférieures à celles obtenues avec la dernière période de la modélisation à 1 an (]3, 4] ans).

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
GLM	0.21	76.3%	15.7%	58%	16%
RF	0.21	76%	15.2%	56%	16%
XGBoost	0.20	76.4%	15.8%	61%	16%

TABLE 2.26 – Tableau comparant les valeurs des métriques des trois modèles utilisés pour la modélisation des taux de résiliation pendant la période]3 ans, 4 ans] (dernière période - **quatrième année**)

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
GLM	0.236	71%	16%	42%	14%
RF	0.225	71%	15%	43%	14%
XGBoost	0.220	71%	17%	44%	14%

TABLE 2.27 – Tableau comparant les valeurs des métriques des trois modèles utilisés pour la modélisation des taux de résiliation pendant la période **]18 mois, 24 mois]**

Nous choisissons donc de conserver la modélisation par période d'une année pour calculer les durées.

Conclusion du chapitre

Notre objectif est de modéliser la durée de vie des contrats d'assurance habitation afin d'optimiser la rentabilité de la stratégie multi-accès. Dans cette deuxième partie du mémoire, nous avons pu réaliser la modélisation des taux de résiliation qui va par la suite permettre d'estimer les durées de vie.

Nous hésitions cependant sur le choix de la durée de la période de modélisation c'est-à-dire entre faire la modélisation par période d'un an (qui pourrait permettre d'avoir de bonnes performances mais la période est trop large) et par période de 6 mois. Celle-ci pourrait permettre d'obtenir des durées plus précises puisque la période est plus petite mais elle risque d'aboutir sur des performances moins bonnes vu que les taux de résiliations pour certaines périodes de 6 mois sont très faibles et donc les classes plus déséquilibrés (7% de taux de résiliation pour la première période de 6 mois par exemple). Nous avons donc testé les deux. Pour chaque période, 3 modèles ont été évalués : un GLM, un *Random Forest* et un XGBoost. En comparant les performances des modèles pour les deux différents horizons de modélisation, nous avons finalement conservé la modélisation par période d'un an.

Pour cet horizon de modélisation, le meilleur modèle est le XGBoost. Le meilleur modèle est défini en comparant à chaque période, les différentes métriques suivant les modèles mais aussi suivant les parcours (afin de s'assurer qu'il n'y a pas que les parcours bien représentés qui vont avoir des résiliations bien prédites) et en vérifiant que les taux de résiliation observés sont proches des taux de résiliation réels. Comme nous nous y attendions, les performances se dégradent de plus en plus à chaque période. En effet, les valeurs de certaines métriques comme la précision baissent nous poussant ainsi à limiter la modélisation à 4 ans (au lieu de 5 ans). Cela est dû au fait que nous utilisons la vision des variables explicatives à l'affaire nouvelle pour faire des prédictions qui s'étendent sur quatre ans. Plusieurs de ces variables ont pu changer de valeurs dans le temps. La performance de la première période de modélisation n'est pas excellente non plus. Nous verrons un peu plus en détail, dans la dernière partie, les limites des modèles et comment nous aurions pu les améliorer. Nous jugeons les performances obtenues comme étant acceptables et allons donc calculer les durées en utilisant le meilleur modèle choisi qui est le XGBoost.

L'importance des variables des différents modèles n'est pas la même mais elle présente plusieurs similarités. La qualité juridique fait partie des variables les plus discriminantes pour chaque modèle. L'importance des variables au fil des périodes baisse, les variables les plus importantes étant globalement les mêmes. La modélisation des taux de résiliation a aussi permis de voir que l'importance des parcours est très faible. Dans la partie suivante, d'autres outils statistiques seront utilisés afin de confirmer cette remarque mais aussi afin de mieux connaître comment les variables explicatives influencent les résiliations avec le modèle XGBoost choisi.

Chapitre 3

Modélisation de la durée de vie

3.1 Algorithme utilisant les modèles de taux de résiliation

3.1.1 Description

Une fois les taux de résiliations modélisés, il est possible de prédire à la fin de chaque période quel contrat va être résilié et quel contrat va survivre. Prédire les durées de vie va permettre de pouvoir identifier les profils les plus loyaux et d'identifier les profils les plus rentables en prenant en compte le ratio sinistres à prime.
La prédiction est obtenue avec le XGBoost pour chaque période pendant les différentes périodes qui forment une durée totale de 4 ans.
Nous pouvons donc obtenir une matrice de n lignes et 4 colonnes. Chaque ligne i représente un contrat et chaque colonne j contient la prédiction de la résiliation ou non du contrat pour la $j-ème$ période. Nous appelons cette matrice, matrice de survie et la notons $\boldsymbol{S}$.

$$S = (s_{ij}) \text{ avec } 1 \leq i \leq n \text{ et } 1 \leq j \leq 4, \tag{3.1}$$

s_{ij} donne donc la prédiction de la résiliation ou de la survie du contrat i pendant la $j-ième$ année dans le portefeuille. Ses valeurs possibles sont donc 1 ou 0, 1 correspondant à la résiliation.
Par exemple, $s_{10,1} = 0$, $s_{10,2} = 1$, $s_{10,3} = 0$ et $s_{10,4} = 0$ montre que le 10-ième contrat a été résilié pendant la deuxième année.

Ainsi, pour déterminer cette matrice, il faut déterminer si le contrat va être résilié ou non grâce à la modélisation des taux de résiliation par période.
Le principe est, en résumé, décrit ci-dessous (pour le contrat i) :

1. Initialiser la ligne i de la matrice S avec 0 à chaque colonne.
2. Initialiser la période j comme étant égale à 1 (j=1) ;
3. Tant qu'une résiliation n'est pas prédite et que $j \leq 4$:
   (a) Lancer le modèle de prédiction de la j-ème période d'un an.
   (b) Si une résiliation est prédite alors
   $$s_{ij} = 1.$$
   - L'algorithme s'arrête.
   (c) Sinon on passe à la période suivante : $j = j + 1$.

4. Si le modèle ne prédit pas de résiliation au bout de 4 ans pour le contrat i, (la durée est égale à 4 ans)

$$s_{ij} = 0 \ \ \forall\, j \text{ tel que } 1 \leq j \leq 4.$$

La figure 3.1 suivante décrit cet algorithme dans le cas où la durée maximale est de 2 ans (au lieu de 4 ans) pour simplifier la lecture.

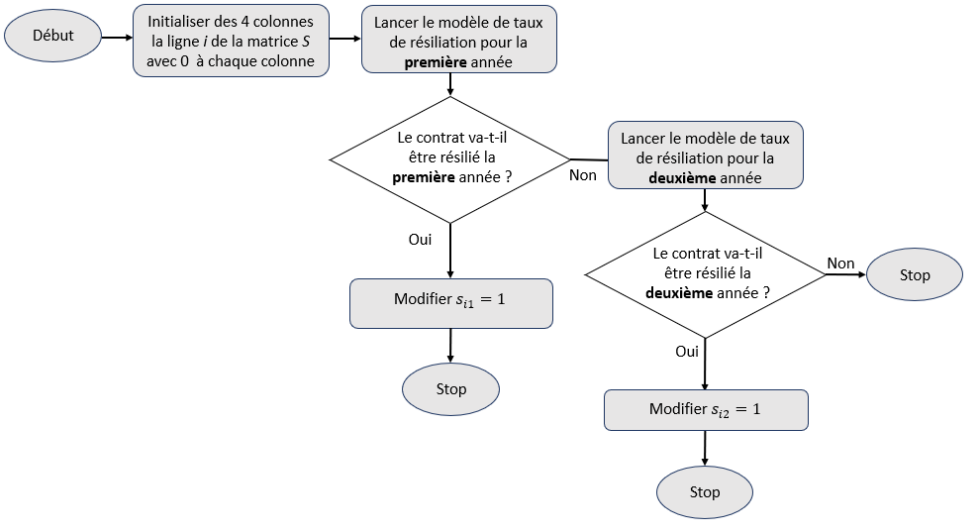


FIGURE 3.1 – Schéma résumant l'algorithme de détermination de la matrice S en utilisant les modèles de taux de résiliation. Exemple pour un total de 2 ans.

Il est donc possible de lancer l'algorithme à partir d'une année donnée. Par exemple, pour les contrats souscrits en 2019, leur durée de vie a déjà été observée sur une année. Rappelons que la fin d'observation des résiliations est le 31 Décembre 2020. Pour pouvoir déterminer la durée sur 4 ans des contrats qui n'ont pas été résiliés la première année, nous pouvons lancer l'algorithme de calcul de la durée de vie à partir de la deuxième année en utilisant comme données d'entrée, les contrats ayants survécu la première année. Dans ce cas, j est initialisé avec la valeur $j = 2$.

Pour approximer la durée des contrats qui ont été résiliés pendant une période, nous allons calculer la durée moyenne des contrats ayant été réellement résiliés au cours de cette période. Si par exemple, la période est de]1 an, 2 ans] et que la durée moyenne trouvée pour les contrats ayant été résiliés à cette période est de 1.4 ans, nous allons donc considérer, dans notre modèle de durée, que les contrats ayant été résiliés dans la période]1 an, 2 ans] ont duré 1.4 ans dans le portefeuille.

Le tableau 3.1 suivant présente la durée moyenne des contrats résiliés pendant la première, la deuxième, la troisième et la quatrième année. Par exemple, 0.565 an est la durée moyenne des contrats ayant été résiliés la première année.
Nous remarquons que plus l'année passée dans le portefeuille augmente, plus la durée de vie moyenne est éloignée de la durée à l'échéance. Les contrats ayant survécu jusqu'à 3 ans résilient en moyenne le quatrième mois de leur quatrième année dans le portefeuille alors que les affaires nouvelles résilient entre le sixième et le septième mois. Cela pourrait tout d'abord s'expliquer par le fait qu'avec la loi Hamon, l'assuré peut résilier à n'importe quel moment sans motif valable à partir de la deuxième année du contrat. La première année, il y a plus d'individus qui attendent la première échéance avant de résilier puisque les assurés ne peuvent pas résilier sans motif valable.

Année dans le portefeuille	Durée moyenne des contrats résiliés (en année)
1	0,565
2	1,418
3	2,405
4	3,34

TABLE 3.1 – Durée de vie moyenne des contrats suivant le nombre d'années passées dans le portefeuille avant de résilier

Nous allons appelé ce tableau : vecteur de durées de vie moyennes et allons donc le représenter par le vecteur $\boldsymbol{D}$ défini comme suit

$$D = (d_j) \text{ avec } 1 \leq j \leq 4, \tag{3.2}$$

d_j donne la durée de vie moyenne d'un contrat ayant été résilié la $j - ième$ année dans le portefeuille. Par exemple : $d_2 = 1,418$ ans.

Ce vecteur ainsi que la matrice de survie $\boldsymbol{S}$ vont nous permettre d'avoir un algorithme simple permettant de calculer la durée de vie moyenne sur 4 ans de n contrats observés à partir de l'affaire nouvelle. La figure 3.2 suivante décrit l'algorithme de calcul des durées de vie à partir de la matrice S et du vecteur D pour le contrat i :

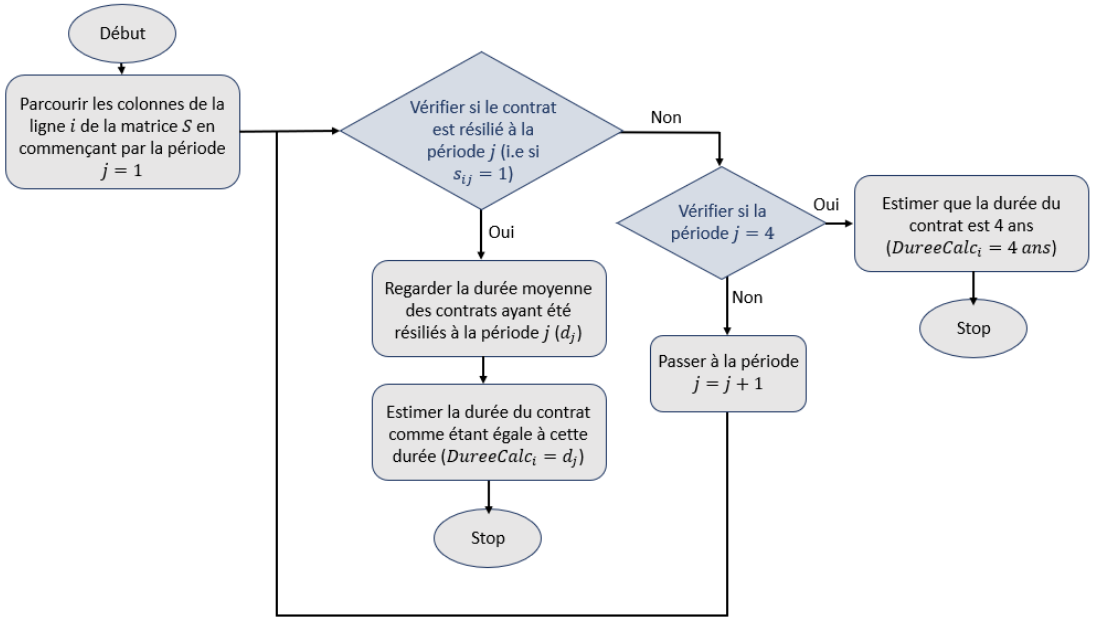


FIGURE 3.2 – Schéma résumant l'algorithme d'estimation de la durée de vie du contrat i notée $DureeCalc_i$ à partir de la matrice de survie S et du vecteur de durées de vie moyennes D

L'algorithme est défini comme suit.

Algorithme 2 : Calcul de la durée de vie moyenne sur 4 ans de n contrats à partir de l'affaire nouvelle

Entrées : Matrice de survie $\boldsymbol{S}$ (de taille $n \times 4$) et vecteur (de taille 4) de durées de vie moyennes suivant l'année de résiliation $\boldsymbol{D}$
Sorties : Vecteur de taille n contenant la durée de vie estimée des n contrats ***DureeCalc***
pour *i = 1 **à** n (pour chaque contrat)* **faire**
  j=1 ;
  tant que $s_{ij} = 0$ ***et*** $j < 4$ *(le modèle ne prédit toujours pas de résliation pour le i-ème contrat au bout de la j-ème période inférieure à 4 ans)* **faire**
    Passer à l'année suivante : $j = j + 1$;
  fin
  si $s_{ij} = 1$ *(le modèle prédit que le contrat va être résilié au bout de j an(s))* **alors**
    La durée estimée est égale à la durée moyenne des contrats ayant réellement été résiliés au bout de j an(s) : $DureeCalc_i = d_j$.
  fin
  si $s_{ij} = 0$ *(le modèle ne prédit pas de résiliation au bout de 4 ans)* **alors**
    La durée est égale à 4 ans : $DureeCalc_i = 4$ ans.
  fin
fin

La durée de vie de l'exemple précédent c'est-à-dire $s_{10,1} = 0$, $s_{10,2} = 1$, $s_{10,3} = 0$ et $s_{10,4} = 0$ est égale à $d_2 = 1,418$ ans. Le 10-ème contrat de la base a été résilié pendant la deuxième année et sa durée de vie moyenne $DureeCalc_{10}$ est 1,418 ans.

Cet algorithme va donc nous aider à calculer les durées de vie moyennes des contrats de notre portefeuille. Néanmoins, avant d'aller plus loin, nous allons d'abord vérifier si les durées de vie obtenues sont bien cohérentes avec les durées de vie observées sur 4 ans. Vu que nous avons choisi le modèle de taux de résiliation en nous assurant que les taux trouvés sont proches des taux réellement observés, nous espérons que les durées de vie moyennes estimées soient proches des durées de vie moyennes réelles.

3.1.2 Vérification

Nous allons comparer :

— les durées de vie observées pour les générations avec un recul d'au moins 4 ans (2015 et 2016).

— les durées de vie déterminées à l'aide de notre modèle pour ces mêmes générations (2015 et 2016).

Ce test permet de vérifier si l'algorithme prédit des durées de vie cohérentes, ce n'est pas un test de performance puisque nous utilisons exactement les mêmes contrats qui ont servi à entraîner (ou tester) nos modèles de taux de résiliation.

RMSE et MAE

L'erreur absolue moyenne (ou *Mean Absolute Error* MAE) et l'erreur quadratique moyenne (ou *Root Mean Squared Error* RMSE) sont deux des mesures les plus utilisées afin d'évaluer la précision d'une prédiction de variables continues. Elles expriment l'erreur de prédiction moyenne du modèle en unité de la variable d'intérêt. La MAE mesure l'ampleur moyenne des erreurs dans un ensemble de prédictions, où toutes les différences individuelles ont le même poids et sans tenir compte du signe de

l'erreur. La RMSE est la racine carrée des erreurs quadratiques moyennes. Ceci le rend moins facile à interpréter que la MAE mais aussi intéressante puisqu'elle donne un poids relativement élevé aux erreurs importantes.
La RMSE et la MAE sont obtenues grâce aux formules suivantes

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,. \tag{3.3}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{3.4}$$

avec y_i : durée (en année) observée réellement pour les contrats souscrits en 2015 et 2016, $\hat{y}_i$: durée (en année) prédite pour les contrats souscrits en 2015 et 2016 et n : nombre de contrats souscrits en 2015 et 2016 (n= 395 234 contrats).

La figure 3.2 présente les valeurs de la MAE et de la RMSE au total et en fonction de la qualité juridique [1].

Métrique	Locataire	Propriétaire	Total
MAE	0,22 an	0,45 an	0,37 an
RMSE	0,54 an	0,83 an	0,69 an

TABLE 3.2 – Valeurs de la MAE et de la RMSE au total et en fonction de la qualité juridique

Nous remarquons qu'au total la RMSE est presque 2 fois plus grande que la MAE vu qu'elle pénalise davantage les grandes erreurs (et qu'il pourrait y en avoir à cause des contrats prédits comme ayant survécu jusqu'à la quatrième année (durée prédite=4 ans) alors qu'ils ont été réellement résiliés dès la première année (durée réelle=0,9 an)). La MAE qui est de 0,37 an, n'est pas très faible et pourrait en partie s'expliquer par le fait que les durées prédites ne sont pas précises (il n'y a que 5 valeurs possibles). En faisant la distinction suivant la qualité juridique qui est l'une des variables les plus discriminantes, nous constatons que la MAE est plus faible chez les locataires (qui ont un taux de résiliation plus élévé et mieux prédit). Pour les propriétaires la RMSE s'approche dangereusement d'une année vu que leur taux de résiliation sont moins bien prédits car plus faibles.

Nous allons désormais évaluer les durées de vie moyennes prédites suivant les variables discriminantes repérées avec l'importance des variables de la partie précédente pour vérifier si elles sont cohérentes avec les durées de vie moyennes observées réellement.

Comparaison suivant une variable

Les figures 3.3, 3.4 et 3.5 suivantes comparent les durées de vie observées et trouvées par notre modèle pour les contrats souscrits en 2015 et en 2016 suivant le nombre de contrats (figure 3.3), suivant la qualité juridique (figure 3.4) et suivant l'âge du client (figure 3.5).
Les durées d'une même année sont coloriées avec la même couleur. Les durées prédites par nos modèles sont représentées en pointillés.
Nous remarquons que les durées de vie trouvées et observées sont proches.

1. l'une des variables les plus discriminantes d'après l'importance des variables du XGBoost

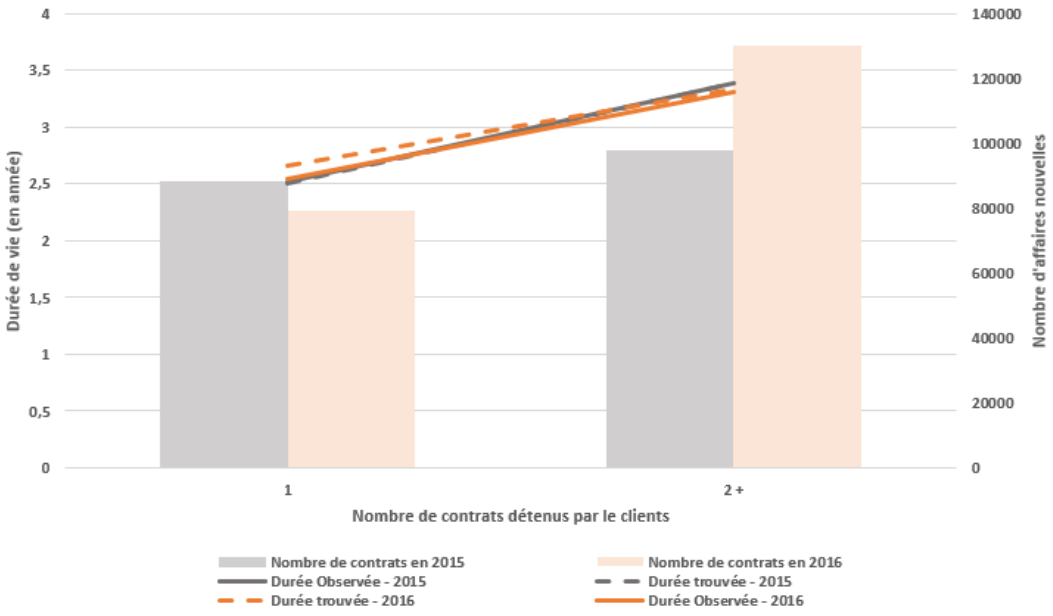


FIGURE 3.3 – Comparaison des durées de vie moyennes observées et trouvées par le modèle pour les contrats souscrits en 2015 et 2016 suivant le nombre de contrats détenus par le client

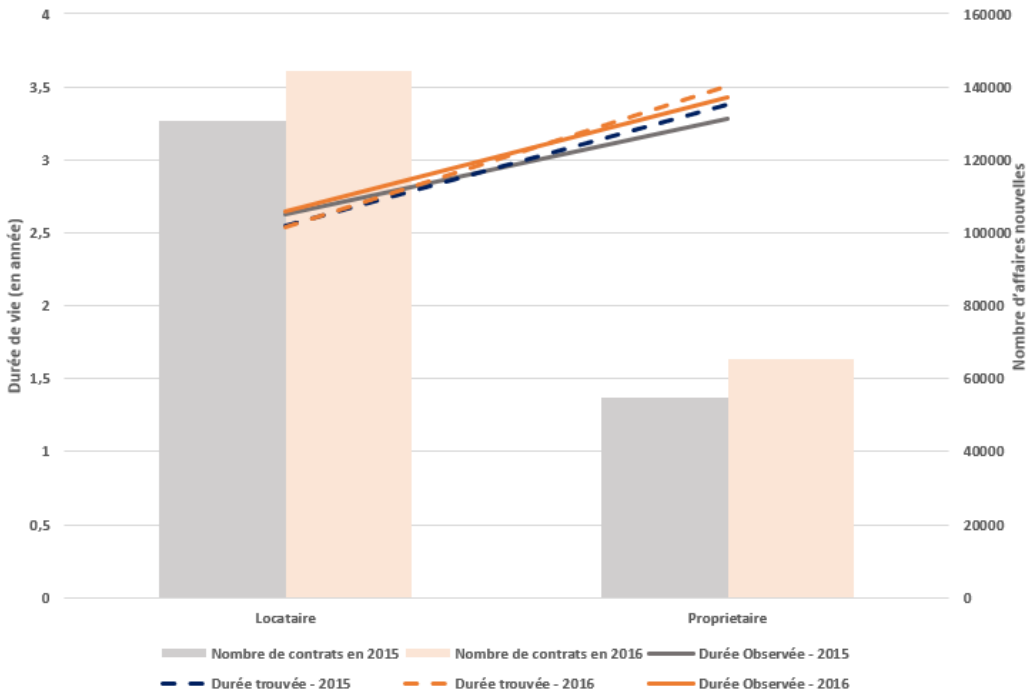


FIGURE 3.4 – Comparaison des durées de vie moyennes observées et trouvées par le modèle pour les contrats souscrits en 2015 et 2016 suivant la qualité juridique

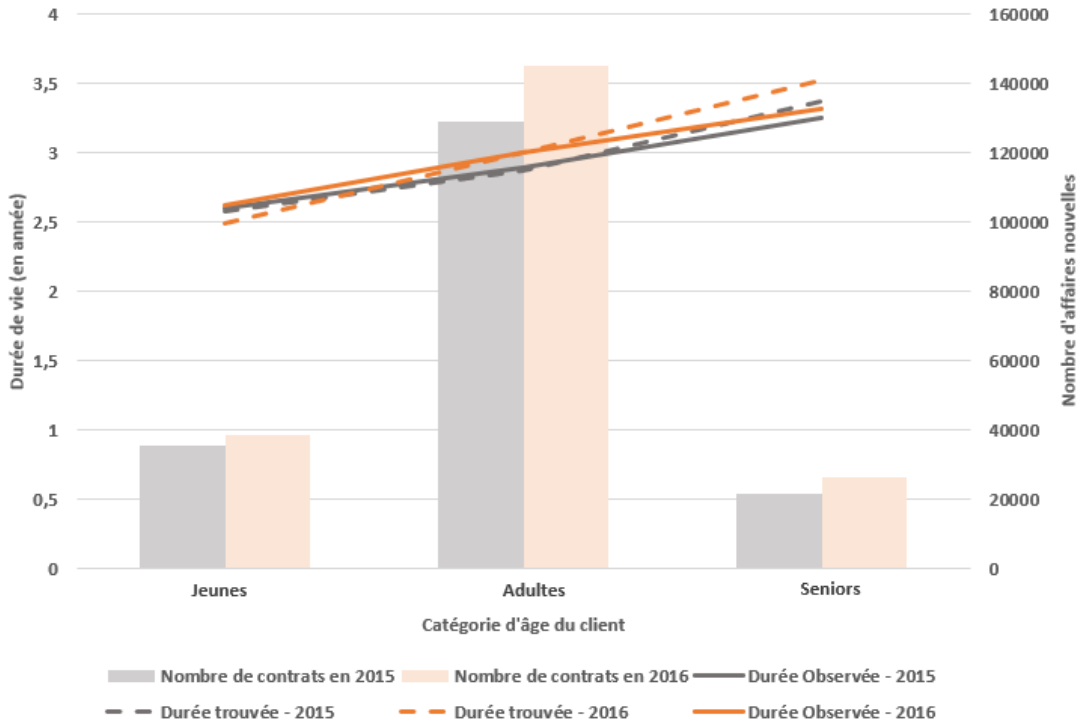


FIGURE 3.5 – Comparaison des durées de vie moyennes observées et trouvées par le modèle pour les contrats souscrits en 2015 et 2016 suivant la catégorie d'âge du client

Comparaison suivant deux variables

Les figures 3.6 et 3.7 ci-dessous permettent de comparer les durées de vie observées et les durées de vie trouvées avec notre modèle pour les contrats ayant été souscrits en 2015 et 2016 suivant le nombre de contrats croisé avec la qualité juridique (figure 3.6) et suivant l'âge du client croisé avec la qualité juridique (figure 3.7).
Les durées d'une même année sont coloriées avec la même couleur. Les durées prédites par nos modèles sont représentées en pointillés.
Nous remarquons que les durées de vie moyennes trouvées par notre modèle et observées pour les contrats souscrits en 2015 et 2016 sont proches. Cela ne veut pas forcément dire que c'est le cas pour les contrats des autres années qui n'ont pas été utilisés pour construire le modèle.

En calculant les durées de vie pour les autres générations (contrats souscrits après 2016 (que nous verrons plus tard)), des écarts sont trouvés. Ces écarts pourraient être causés par le fait que notre modèle n'est pas parfait mais aussi par la différence des profils caractérisant chaque génération. Il est donc important de bien connaître l'impact des variables sur la durée afin de non seulement pouvoir expliquer les écarts trouvés mais aussi de bien maitriser le fonctionnement de notre modèle (ce qui est nécessaire pour tout actuaire). Nous avons déjà vu dans la deuxième partie les variables qui impactent les résiliations sans creuser pour savoir comment est caractérisé cet impact. Nous allons à présent, utiliser plus d'outils statistiques afin de bien visualiser l'impact des variables pour chaque période de modélisation avec le modèle choisi (XGBoost).

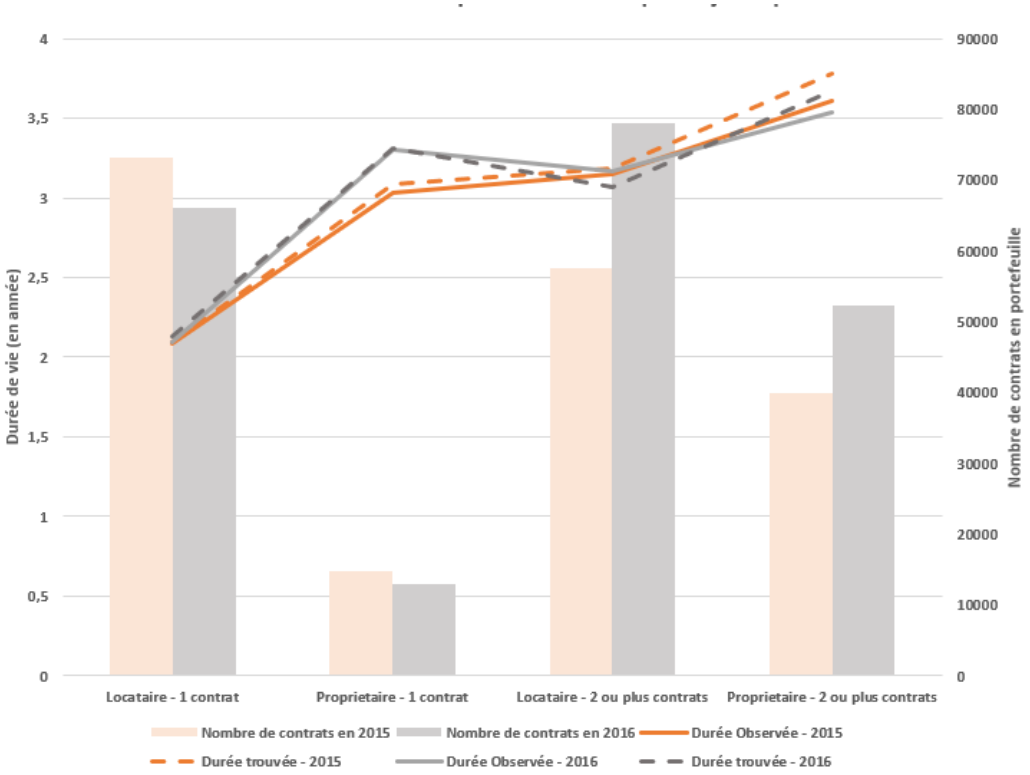


FIGURE 3.6 – Comparaison des durées de vie moyennes observées et trouvées par le modèle pour les contrats souscrits en 2015 et 2016 suivant la qualité juridique et le nombre de contrats du client

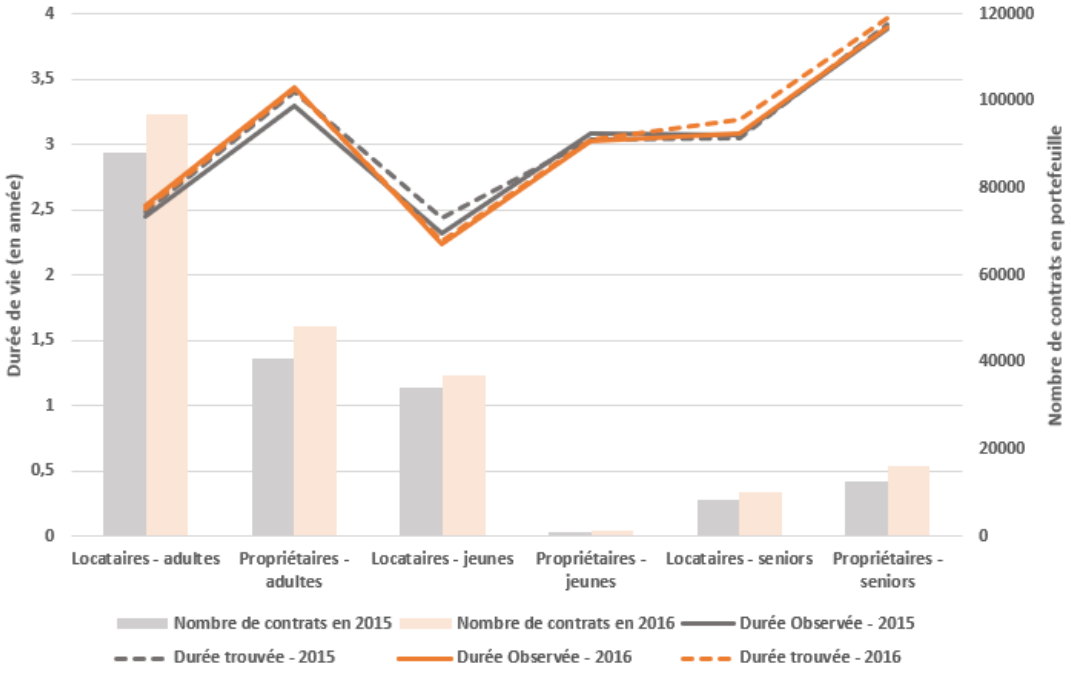


FIGURE 3.7 – Comparaison des durées de vie moyennes observées et trouvées par le modèle pour les contrats souscrits en 2015 et 2016 suivant la qualité juridique et la catégorie d'âge du client

3.2 Impact des variables sur la durée de vie

Parmi les modèles testés, il y a un modèle facile à interpréter qui est le GLM mais aussi des modèles dites à boites noires (*black box*) comme le XGBoost qui est le modèle utilisé pour estimer les durées de vie. Nous avions vu dans la partie précédente qu'avec les modèles à boites noires, la seule information disponible sur les variables explicatives par rapport à la variable à expliquer est leur importance. Nous ne savons pas comment les variables impactent la prédiction. Il y a fort heureusement plusieurs méthodes qui permettent de tirer le plus d'informations possibles sur les modèles à boites noires.

3.2.1 Outils d'interprétation utilisés

Partial Dependance plot (PDP)

Les diagrammes de dépendance partielle ou *Partial Dependance plots* (PDP) en anglais permettent de voir comment chaque variable explicative ou prédicteur affecte les prédictions du modèle (Friedman, 2001). Ces graphiques sont utiles à la fois pour extraire des informations et pour vérifier que le modèle donne bien des informations cohérentes. Ils pourraient donc être très utiles afin de déterminer l'influence de la génération et des parcours.
Le diagramme de dépendance partielle n'est calculé qu'une fois le modèle ajusté sur des données réelles. Dans le cas où l'impact des différents parcours est évalué, le modèle est utilisé pour prédire si le contrat va être résilié ou non après avoir modifié la valeur de la variable `parcours`. Pour chaque contrat de la base, la résiliation ou non de ce contrat est prédite lorsque le parcours est ATA puis lorsque le parcours n'est plus ATA. Ces prédictions vont permettre de tracer un graphique donnant la probabilité moyenne de résiliation prédite sur l'axe vertical lorsque le parcours passe de ATA aux autres parcours.

La fonction de dépendance partielle pour le modèle f et la variable explicative X^j à la valeur z est définie comme suit

$$g^j_{PD}(z) = E_{\underline{X}^{-j}}\left\{f\left(X^{j|=z}\right)\right\}. \tag{3.5}$$

Ainsi, c'est la valeur attendue des prédictions du modèle lorsque X^j est fixé à z sur la distribution (marginale) de $\underline{X}^{-j}$, c'est-à-dire sur la distribution conjointe de toutes les variables explicatives autres que X^j. Habituellement, la vraie distribution de $\underline{X}^{-j}$ est inconnue. La fonction partielle est estimée par la méthode de Monte Carlo c'est-à-dire en calculant la moyenne des n données de la base

$$\hat{g}^j_{PD}(z) = \frac{1}{n}\sum_{i=1}^{n} f\left(\underline{x}_i^{j|=z}\right). \tag{3.6}$$

Les PDP présentent donc plusieurs avantages mais aussi plusieurs inconvénients.

- **Avantages**
  - La détermination des diagrammes de dépendance partielle est intuitive : la fonction de dépendance partielle d'une valeur particulière de variable représente la prédiction moyenne obtenue en forçant tous les individus de la base de données à avoir cette valeur pour la variable correspondante.
  - Les diagrammes de dépendance partielle sont faciles à mettre en œuvre et peuvent être utilisés avec tous les modèles.
  - Dans le cas où il n'y a pas de corrélation entre la variable (dont le PDP est calculé) et les autres variables, l'interprétation est claire : le graphique de dépendance partielle montre comment la prédiction moyenne dans la base de données change lorsque cette variable est modifiée. Cependant, c'est plus compliqué lorsque les caractéristiques sont corrélées.

— **Inconvénients**
  — L'hypothèse d'indépendance (la variable pour laquelle la dépendance partielle est calculée n'est pas corrélée avec d'autres variables) est le plus grand inconvénient des graphiques PDP puisqu'elle n'est pas toujours vérifiée. Supposons que nous voulions prédire la résiliation suivant le nombre de contrats du client et la réduction. Pour la dépendance partielle de l'une des variables, par exemple la réduction, nous supposons donc qu'elle n'est pas corrélée à l'autre variable qui est le nombre de contrats, ce qui est évidemment une hypothèse non vérifiée. En calculant le PDP pour une réduction égale à 50%, la moyenne sur la distribution marginale du nombre de contrats est faite alors que cette distribution peut inclure un nombre de contrat égal à 1 alors qu'il est n'est pas possible d'avoir cette réduction avec ce nombre de contrats.
  Ainsi lorsque les variables sont corrélées, de nouveaux points dans les zones de la distribution des variables sont créés. La probabilité réelle d'avoir ces points peut être très faible (par exemple, il serait très peu probable qu'un assuré avec un seul contrat ait une réduction de 50% s'il faudrait avoir au moins 5 contrats pour bénéficier de cette réduction).
  — Il faut tenir compte de la distribution des données afin de ne pas mal interpréter un résultat faussé par un manque de données.
  — Le nombre maximum réaliste de variables dans une fonction de dépendance partielle est de deux. Ce n'est pas la faute des PDP, mais de la représentation en 2 dimensions (papier ou écran) et aussi de notre incapacité à imaginer plus de 3 dimensions.
  — Les effets hétérogènes peuvent être masqués car les graphiques PDP ne montrent que les effets marginaux moyens. Supposons que pour une variable donnée, la moitié des points de données ont une association positive avec la prédiction (plus la valeur de la variable est grande, plus la prédiction est grande) et l'autre moitié a une association négative. La courbe PDP pourrait être une ligne horizontale, car les effets des deux moitiés de l'ensemble de données pourraient s'annuler. Dans ce cas, nous aurions alors conclu que la variable n'a aucun effet sur la prédiction, ce qui n'est pas le cas.

Le dernier inconvénient mentionné nous pousse à utiliser un autre type de graphique très similaire aux diagrammes de dépendance partielle : l'ICE.

Individual Conditionnal Expectation (ICE)

Les graphiques d'espérance conditionnelle individuelle ICE introduits par Goldstein et al. (2015) affichent une ligne par instance qui montre comment la prédiction de l'instance change lorsqu'une variable varie. Un PDP est donc la moyenne des lignes d'un tracé ICE.
Nous allons alors utiliser l'ICE pour les mêmes raisons que le PDP. Même s'il a un avantage en plus par rapport au PDP, l'ICE a aussi plusieurs incovénients.

— **Avantages**
  — Contrairement aux diagrammes de dépendance partielle, les courbes ICE peuvent révéler des relations hétérogènes.
  — Les courbes d'espérance conditionnelle individuelles sont encore plus intuitives à comprendre que les diagrammes de dépendance partielle. Une ligne représente les prédictions pour une instance lorsque la variable d'intérêt varie.

— **Inconvénients**
  — Les courbes ICE souffrent du même problème que les PDP à cause de l'hypothèse d'indépendance. Si la variable d'intérêt est corrélée avec les autres caractéristiques alors certains points peuvent

être improbables selon la distribution des caractéristiques conjointes.
— Les courbes ICE ne peuvent afficher qu'une seule entité de manière significative, car deux entités nécessiteraient le dessin de plusieurs surfaces superposées et nous ne verrions rien dans le tracé.
— Si de nombreuses courbes ICE sont tracées, le graphique peut devenir surchargé et très difficile à visualiser. La solution du package *h2o* (que nous utilisons sur python) consiste à découper l'ensemble de données afin de représenter l'ICE par décile.

Un autre outil qui permet de pouvoir bien interpréter les modèles complexes sont les SHAP Values.

SHapley Additive exPlanation (SHAP)

SHapley Additive exPlanation (SHAP) de Lundberg et Lee (2017) est une approche théorique des jeux qui permet d'interpréter tout modèle de *machine learning*. L'objectif de SHAP est d'expliquer la prédiction pour toute instance x_i comme la somme des contributions de ses valeurs de variables individuelles notées ϕ_i. L'idée est de moyenner l'impact qu'une variable a pour toutes les combinaisons de variables possibles. Nous allons expliquer le concept avec un exemple simple sur la figure 3.8 ci-dessous.

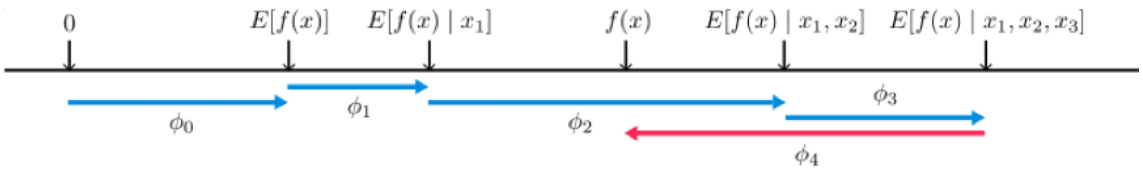


Figure 3.8 – Prédiction $f(x)$ expliquée par la somme des effets ϕ_i de chaque variable x_i - Source Lundberg et Lee, 2017

Comme vu sur la figure, les SHAP *values* ϕ_i attribuent à chaque variable x_i le changement dans la prédiction du modèle espéré après conditionnement sur cette variable. Elles expliquent comment en partant de la valeur de base $E[f(x)]$, la prédiction actuelle $f(x)$ serait obtenue. La valeur de base est la prédiction qui aurait été obtenue si aucune des variables x_i n'était connue. Lorsque le modèle est non linéaire ou que les variables ne sont pas indépendantes, l'ordre dans lequel celles-ci sont ajoutées compte, les SHAP *values* résultent alors de la moyenne des valeurs ϕ_i calculées lorsque les variables sont ajoutées dans tous les ordres possibles.

La formule générale de la SHAP *value* ϕ_i est donc

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}} \left(x_{S \cup \{i\}} \right) - f_S \left(x_S \right) \right], \tag{3.7}$$

avec F l'ensemble des variables explicatives, S un ensemble de variables, f_S la fonction de prédiction avec l'ensemble de variables S, $f_S(x_S) = E\left[f(x) \mid x_S\right]$, et i est la $i^{ème}$ variable.

En effet, pour déterminer ϕ_i, le modèle est entraîné avec tous les sous-ensembles de variables explicatives possibles $S \subseteq F$, où F est l'ensemble de toutes les variables explicatives. Il attribue une valeur d'importance à chaque variable qui représente l'effet sur la prédiction du modèle lorsque la

variable donnée fait partie des variables retenues comme variables explicatives.
Pour calculer cet effet, un modèle $f_{S\cup\{i\}}$ est entraîné avec les variables retenues (en incluant la variable dont l'effet est déterminé) tandis qu'un autre modèle f_S est entraîné en excluant la variable dont l'effet est déterminé. Les prédictions des deux modèles obtenues sont alors comparées

$$f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S\left(x_S\right), \tag{3.8}$$

où x_S représente les valeurs des entités en entrée dans l'ensemble S.

Vu que l'effet de l'inclusion de la variable dépend des autres variables du modèle, les différences sont calculées pour tous les sous-ensembles possibles $S \subseteq F\backslash\{i\}$. Les valeurs de Shapley sont ensuite calculées en faisant une moyenne pondérée de toutes les différences possibles. La pondération est donc

$$\frac{1}{|F|}\frac{1}{\left(\begin{array}{c}|F|-1\\|S|\end{array}\right)} = \frac{|S|!(|F|-|S|-1)!}{|F|!}. \tag{3.9}$$

Il y a $\left(\begin{array}{c}|F|-1\\|S|\end{array}\right) = \frac{(|F|-1)!}{|S|!(|F|-|S|-1)!}$ sous-ensembles possibles de taille $|S|$ parmi les variables explicatives de $F\backslash\{i\}$.
Ses sous ensembles ont tous un poids égal à $\frac{1}{|F|}$, $|F|$ étant le nombre de tailles possibles pour S.
Ce qui donne la formule (3.7).

Lundberg et Lee (2017) ont démontré que cette approche vérifiait les propriétés suivantes :
— Cohérence : lorsqu'un modèle change de tel sorte que l'effet d'une variable est plus important sur le modèle, la SHAP *value* de cette variable ne doit pas baisser.
— Variables nulles sans effet : lorsqu'une variable a une SHAP *value* égale à zéro pour une instance donnée, alors la variable ne doit pas avoir d'impact avec cette instance.
— Additivité : la somme des SHAP *value* des variables permet d'obtenir la prédiction du modèle pour tous les exemples.

L'additivité entraîne qu'une prédiction peut s'interpréter comme la somme des différentes SHAP values des variables à laquelle est ajoutée la valeur de base ϕ_0 qui est la moyenne de toutes les prédictions de l'ensemble de données

$$f(z') = y_{pred} = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{3.10}$$

avec y_{pred} la valeur prédite par le modèle pour l'observation donnée, ϕ_0 la valeur de base du modèle, M le nombre de variables explicatives et $z' \in \{0,1\}^M$ / quand la variable est observée : $z'_i = 1$ ou inconnue : $z'_i = 0$.

Calcul des SHAP *values* : *Tree Explainer*
Le calcul des SHAP *values* a un coût élevé puisqu'il faut faire une sommation de termes sur tous les sous-ensembles de variables possibles. Lorsque le modèle est un ensemble d'arbres de décision, il est possible d'utiliser la construction des arbres afin de baisser le coût de calcul en réduisant la sommation en un ensemble de calculs spécifiques à chaque feuille d'un arbre. Cet algorithme est appelé *Tree Explainer*. *Tree Explainer* permet le calcul exact des valeurs de Shapley en temps polynomial d'ordre faible et permet de passer d'une complexité en $O\left(TL2^M\right)$ à $\mathrm{O}\left(TLD^2\right)$, avec T le nombre d'arbres, L le nombre maximum de feuilles des arbres, M le nombre de variables et D la profondeur maximale des arbres.

Interprétabilité
Les SHAP *values* permettent d'avoir un haut niveau d'interprétabilité pour un modèle. Ils ont deux grands avantages puisqu'ils offrent à la fois une interprétabilité globale et locale.
Les SHAP *values* peuvent montrer dans quelle mesure chaque variable contribue à la prédiction de la variable à expliquer ce qui permet d'avoir une **interprétabilité globale** du modèle. Ils permettent donc de voir si la variable impacte positivement ou négativement la variable. Cette façon d'interpréter va plus loin que le graphique d'importance des variables puisqu'elle est capable de montrer la relation positive ou négative de chaque variable avec la cible.
Chaque observation a son propre ensemble de SHAP values ce qui permet d'être en mesure d'expliquer la prédiction obtenue pour un point particulier et donc d'avoir une **interprétabilité locale**. Par exemple, les SHAP *values* peuvent aider à expliquer pourquoi la prédiction du modèle pour le contrat n°546765 aboutit sur une résiliation et quelles sont les contributions des variables à cette prédiction. Les algorithmes usuels d'importance des variables ne montrent les résultats que sur l'ensemble de la base de données.

SHAP *Plots*
Les SHAP *values* permettent de tracer plusieurs graphes afin d'interpréter au mieux le modèle. Ces graphes sont appelés SHAP *Plots*. Différents graphes SHAP sont présentés dans le tableau 3.3 suivant.

SHAP Force Plots	**SHAP *Dependance Plots***	**SHAP *Summary Plots***
Diagrammes qui permettent de voir comment les variables ont contribué à la prédiction du modèle pour une observation spécifique. Ils sont donc utilisés dans le cadre d'une interprétation locale.	Diagrammes qui donnent un aperçu similaire aux *Partial Dependence Plot* (PDP) mais ils ajoutent beaucoup plus de détails en permettant de voir la distribution des effets (impact d'un point de la base) et de déduire entre autres si l'effet est assez constant ou s'il varie beaucoup en fonction des valeurs des autres variables. Ils sont donc plus difficiles à expliquer.	Diagrammes qui permettent de voir : **l'importance des variables** **l'impact des variables** : l'emplacement horizontal indique si l'effet de la valeur correspondante est associé à une prédiction supérieure ou inférieure. **la valeur des variables :** la couleur indique si la valeur de la variable correspondante est élevée (en rouge) ou faible (en bleu) pour une observation donnée.

TABLE 3.3 – Tableau présentant différents diagrammes tracés en utilisant les SHAP *values*

L'importance avec les SHAP *values* est déterminée en partant de l'idée selon laquelle plus les variables ont des valeurs Shapley absolues grandes, plus elles sont importantes. Ainsi, l'importance globale d'une variable notée I_i est obtenue en faisant la moyenne des valeurs SHAP absolues de chaque variable

$$I_i = \frac{1}{N} \sum_{j=1}^{N} \left| \phi_i^{(j)} \right|, \tag{3.11}$$

avec $\phi_i^{(j)}$ indiquant la SHAP *value* de la i-ième variable pour la j-ième prédiction et N le nombre de prédictions.
Ensuite, les variables sont triées par importance décroissante puis leur importance est tracée. Cette importance des variables est donc différente de l'importance des variables que nous avions vu dans la deuxième partie.

3.2.2 Interprétation de l'impact des variables

Le modèle choisi pour estimer les durées de vie est le XGBoost. Nous utilisons donc le SHAP *Summary Plot*[2] du package *h2o* sur python. Il va permettre de visualiser, en plus de l'importance des variables vu dans la deuxième partie, comment chaque variable impacte les durées de vie. La figure 3.9 suivante présente le *SHAP Summary Plot* obtenu avec le modèle XGBoost des taux de résiliation pour la première période d'une année. Celles des autres années sont mises en Annexe B.5.

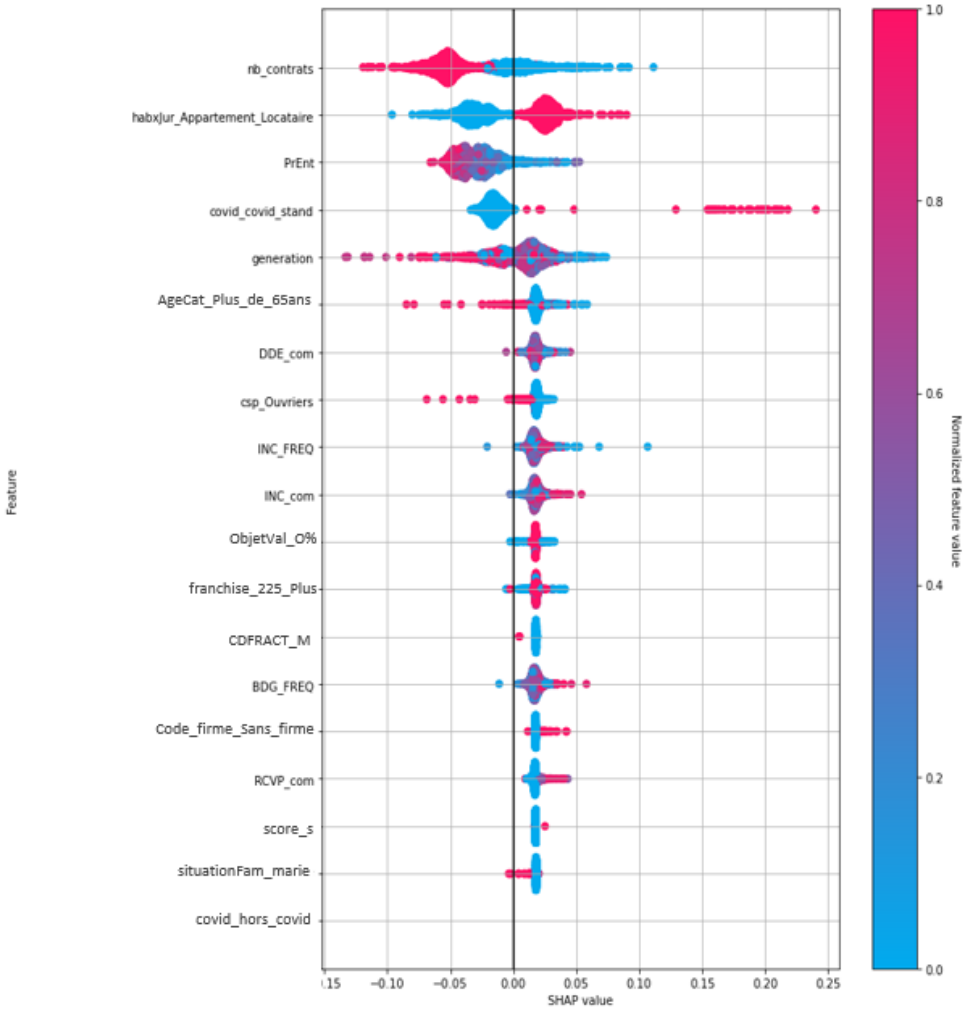


FIGURE 3.9 – SHAP *Summary Plot* du modèle XGBoost de la première année

Les valeurs des variables sont normalisées. Lorsque les points se situent à gauche de la ligne $\phi_i = 0$, ils ont une influence négative sur la variable à expliquer. Par exemple, pour la variable nombre de contrat `nb_contrats` qui a 2 modalités (1 : 1 contrat et 2 : plusieurs contrats), les valeurs rouges correspondent aux assurés qui ont plusieurs contrats tandis que les points bleu correspondent à la modalité «1 contrat». Les valeurs rouges sont trouvées à gauche ce qui signifie donc que les assurés ayant plusieurs contrats ont moins de chance de résilier que ceux n'ayant qu'un seul contrat.
Cette figure permet ainsi de confirmer les déductions faites à partir de l'importance des variables du XGBoost de la première année vue dans la deuxième partie 2.3.4 :

2. Model Explainability, h2o

— Les clients ayant 2 contrats ou plus ont moins de chance de résilier.
— Les locataires d'appartement ont plus de chance de résilier que les propriétaires.
— Les assurés qui paient une prime plus élevée ont moins de chance de résilier.
— Les assurés ont moins de chance de résilier pendant le confinement ou le couve-feu que pendant la période de Covid-19 sans confinement ni couvre-feu (standard).
— Les assurés ayant plus de 65 ans ont moins de chance de résilier que les jeunes.
— les assurés habitant dans une zone où la fréquence des sinistres (notamment celle des incendies, des dégâts des eaux, des bris de glace) est élevée ont plus de chance de résilier. Ils ont plus de chance d'avoir des sinistres et donc de résilier.

Nous avons aussi remarqué que la génération fait partie des variables les plus importantes et que les contrats de la dernière génération qui correspond à la génération 2019 (contrats souscrits en 2019) ont moins de chance d'être résiliés. Nous avons supposé que c'était un effet de la Covid-19 vu que nous avons tendance à croire que les résiliations sont stables dans le temps. Le SHAP *Summary Plot* n'a pas permis d'affirmer cela avec certitude. Afin de confirmer cette supposition, le PDP et l'ICE [3] vont être utilisés grâce au package *h2o* sur python. Nous allons d'abord utiliser l'ICE et la PDP sur la variable nombre de contrats (`nb_contrats`) afin de vérifier si ces diagrammes fournissent des informations cohérentes avec celles du SHAP *Summary Plot* et nos connaissances. La figure 3.10 représente l'ICE et la PDP du nombre de contrats avec le modèle XGBoost de la première période d'une année.

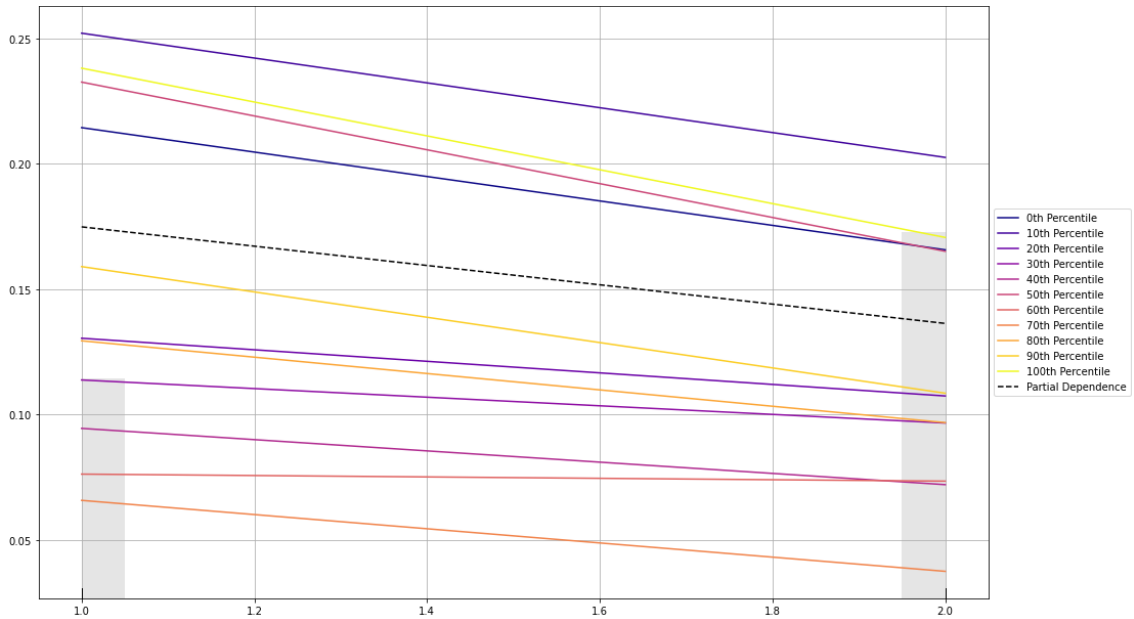


FIGURE 3.10 – *Individual Conditional Expectation* et *Partial Dependence Plot* (en pointillés) de la variable `nb_contrats` (nombre de contrats du client) avec le modèle XGBoost de la première année

La modalité 1 représente les clients avec un seul contrat tandis que la modalité 2 représente les clients avec plusieurs contrats. Les pentes des droites sont descendantes montrant ainsi que les chances de résilier baisse lorsque le nombre de contrats détenu par le client passe de 1 à plusieurs. Ce sont bien les assurés multi-détenteurs qui ont moins de chance de résilier d'après les diagrammes ICE et

3. Model Explainability, h2o

PDP qui sont donc bien cohérents. Nous pouvons alors les utiliser pour essayer de mieux comprendre l'impact de la génération.

La figure 3.11 représente l'ICE et la PDP de la génération obtenus avec le modèle XGBoost de la première période d'une année. Ces diagrammes vont donc nous permettre de voir l'impact de la génération. Nous remarquons alors que pour les trois premières générations (année de souscription), la courbe est presque une ligne droite. Ces générations n'ont pas d'impact sur la résiliation. Cependant, la pente de la droite reliant la génération 3 (2018) à la génération 4 (2019) est descendante. Les résiliations des contrats de la génération 4 (2019) ont pour la plupart été observées en 2020 et donc été perturbées par la Covid-19. C'est la dernière génération qui a moins de chance de résilier puisque certaines résiliations sont observées pendant la période de la Covid-19. La même remarque est faite pour les autres périodes modélisées.

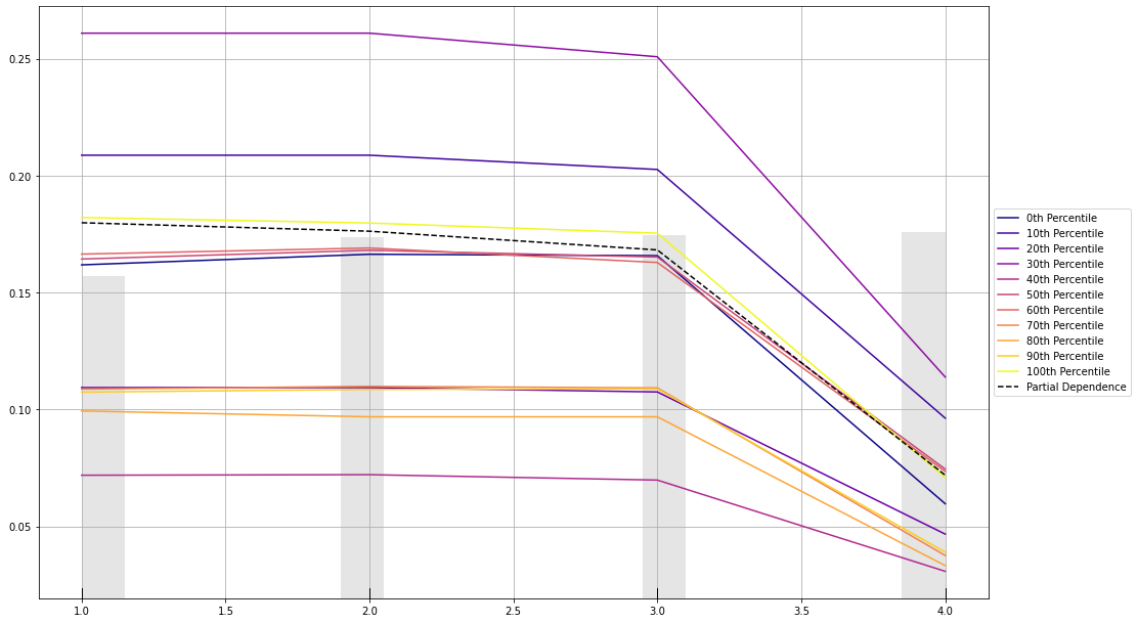


FIGURE 3.11 – *Individual Conditional Expectation* et *Partial Dependence Plot* (en pointillés) de la variable `generation` avec le modèle XGBoost de la première année

Les diagrammes de dépendance partielle PDP et d'ICE ont donc permis de confimer que sans la Covid-19, l'année de souscription n'a pas d'impact sur la durée de vie.

En regardant les variables les plus importantes afin de déterminer leur impact sur les résiliations, nous n'avons pas vu apparaître les parcours dans le classement. Nous allons donc faire un zoom sur les parcours afin de déterminer leur influence.

Zoom sur les parcours

La figure 3.12 suivante présente les diagrammes PDP et ICE du parcours ATA [4] pour la première année. La modalité : 0 représente l'appartenance au parcours ATA tandis que la modalité : 1 représente l'appartenance aux autres parcours. Nous remarquons que les lignes sont presque droites (même remarque pour les autres périodes). Cela signifie que :
Que l'assuré soit dans le parcours ATA ou non, il a presque les mêmes chances de résilier

4. parcours traditionnel - contrats souscrits et gérés en agence

(quel que soit le nombre d'années passées dans le portefeuille).
Mais comme nous pouvons le voir sur la figure, le parcours ATA est sur-représenté par rapport aux autres parcours. Il représente 90% des données. Même si nous ne pouvons complétement l'affirmer au vu des volumes des autres parcours par rapport au parcours traditionnel, nous pouvons dire que **le parcours a très peu d'influence sur les résiliations et donc la durée de vie.**

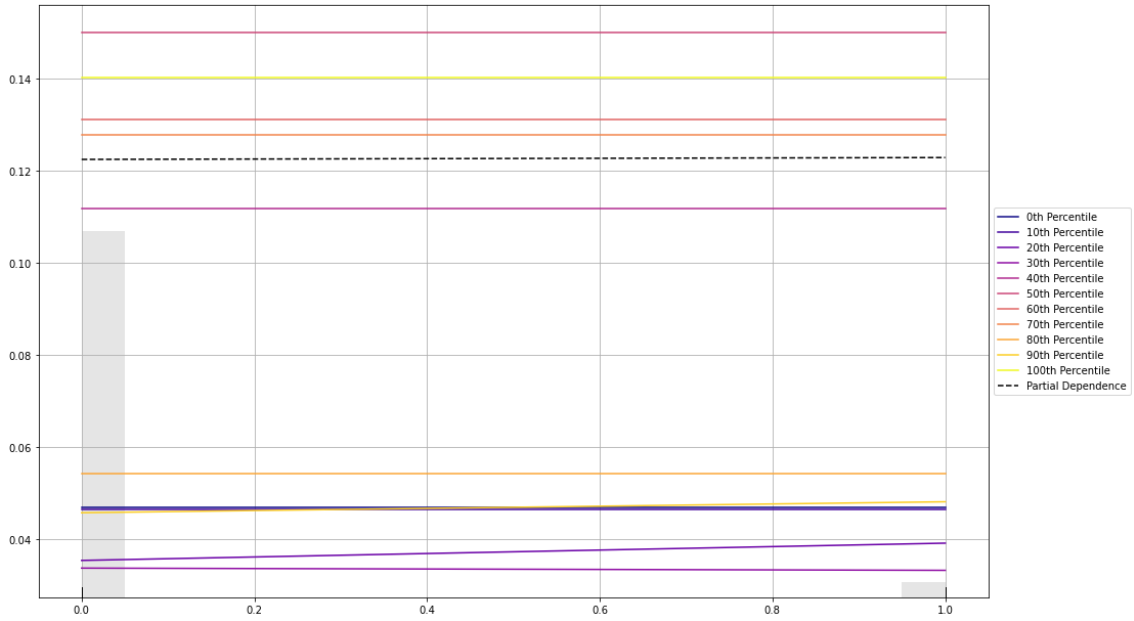


FIGURE 3.12 – PDP et ICE du parcours ATA vs les autres parcours pendant la première année avec le XGBoost

Nous n'avons donc plus besoin d'essayer de déterminer quels profils auraient duré plus longtemps s'ils changeaient de parcours.

3.2.3 Conclusion sur l'impact des variables sur la durée de vie

En comparant l'impact des variables sur les résiliations de chaque année, nous pouvons donc dire que les variables les plus discriminantes sur la durée de vie sont :
— **la qualité juridique** : les propriétaires ont plus de chance de durer dans le portefeuille que les locataires.
— **le nombre de contrats** : les assurés ayant plusieurs contrats ont plus de chance de durer.
— **la prime**.
— **l'âge du client** : les assurés de plus de 65 ans ont plus de chance de durer longtemps dans le portefeuille.
— **les zoniers** : les individus habitant dans des zones ayant des fréquences de vol, incendie... élevées ont moins de chance de durer. Cela s'explique par le fait qu'ils ont plus de chance d'avoir un ou des sinistres et donc une majoration de leur prime ce qui pourrait les pousser à résilier.

Nous avons aussi pu confirmer que l'année de souscription n'a pas d'impact sur la durée de vie à moins qu'il n'y ait eu une perturbation comme la Covid-19 au cours de l'année. La Covid-19 a baissé les résiliations grâce aux confinements et couvre-feux. Nous avons également pu confimer que le parcours a très peu d'impact sur la durée de vie.

3.3 Calcul des durées de vie moyennes

Les contrats des générations 2015 et 2016 ont pu être observés pendant au moins quatre ans, c'est pourquoi ils avaient été utilisés pour les tests. Il n'y a donc pas besoin de modéliser leur durée de vie à priori. Cependant, la quatrième année des contrats ayant été souscrits en 2016 a été impactée par la Covid-19. Il faut donc enlever l'effet de la Covid-19 puisque nous voulons déterminer les profils les plus loyaux en situation normale. Pour cela, nous allons calculer la durée de vie des contrats s'il n'y avait pas eu de Covid-19. Nous allons donc lancer le modèle de taux pour la quatrième année des contrats de 2016 en utilisant les contrats qui ont survécu jusqu'à la troisième année comme schématiser dans la figure 3.13 suivante.

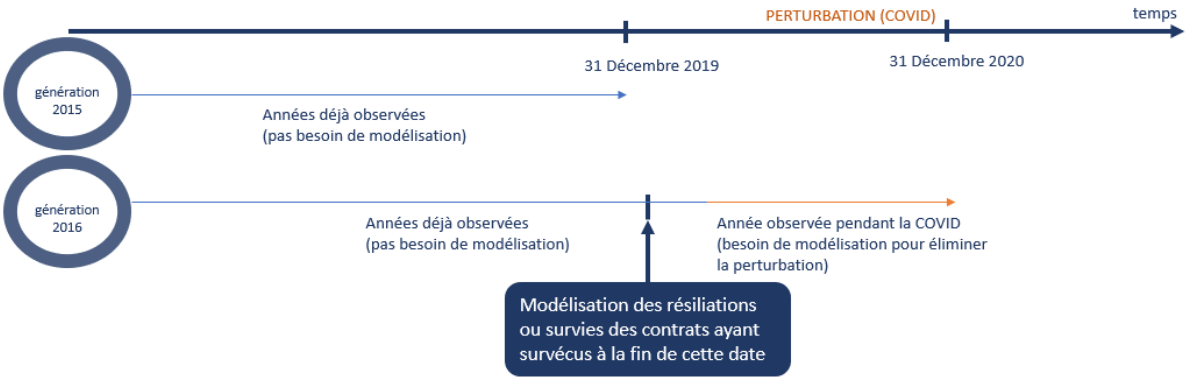


FIGURE 3.13 – Contrats souscrits en 2016 qui sont modélisés à cause de la Covid-19

Nous allons changer la variable `generation` pour qu'elle soit égale à une génération hors Covid-19 (2015) et donner à la variable `covid`[5] (résiliation observée pendant la Covid-19) la valeur `hors covid`.

Les contrats des autres générations (de 2017 à 2020) ont moins de 4 ans d'ancienneté à la fin de l'observation (31 Décembre 2020) sur le périmètre d'étude et leur dernière année d'observation (2020) a été impactée par la Covid-19. Ainsi, en situation normale (hors Covid-19), pour calculer les durées de vie de la génération :

- 2017 : le modèle de durée de vie va être lancé à partir de la deuxième annnée puisque les contrats qui ont survécu jusqu'à 2 ans ont déjà pu être réellement observés (et qu'il y a eu la Covid-19 comme perturbation pendant la troisième année d'observation).
- 2018 : le modèle de durée de vie va être lancé à partir de la première année pour les mêmes raisons que la génération 2017.
- 2019 et 2020 : le modèle de durée de vie va être lancé à partir de l'affaire nouvelle. Les contrats ayant survécu jusqu'à 1 an sont déjà connus pour la génération 2019 mais il y a eu la Covid-19 comme perturbation pendant la première année d'observation.

A la fin du processus, la base sera constituée des contrats souscrits entre 2015 et 2020 avec leur durée de vie moyenne sur 4 ans. Les contrats ont donc tous été observés sur la même durée et sans perturbation (Covid-19).

5. voir description des modalités de la variable dans le tableau 1.6

3.3.1 Détermination des durées suivant l'année de souscription

La figure 3.14 suivante donne la durée de vie moyenne prédite suivant l'année de souscription en situation normale. Nous remarquons que les durées de vie ne sont pas égales mais elles restent proches (entre 2,9 ans et 3,1 ans suivant l'année étudiée).

A partir de 2016, les durées de vie moyennes des contrats sont très proches. La durée de vie moyenne des contrats souscrits en 2015 est la plus basse. En regardant les caractéristiques du portefeuille par année (voir Annexe B.1), nous constatons qu'en 2015, il y a entre 1% et 2% moins de propriétaires que les autres années et au moins 5% moins de clients multi-détenteurs par rapport aux autres années ce qui expliquent que la durée de vie moyenne de l'ensemble des contrats de cette année est inférieure à celle des autres années.

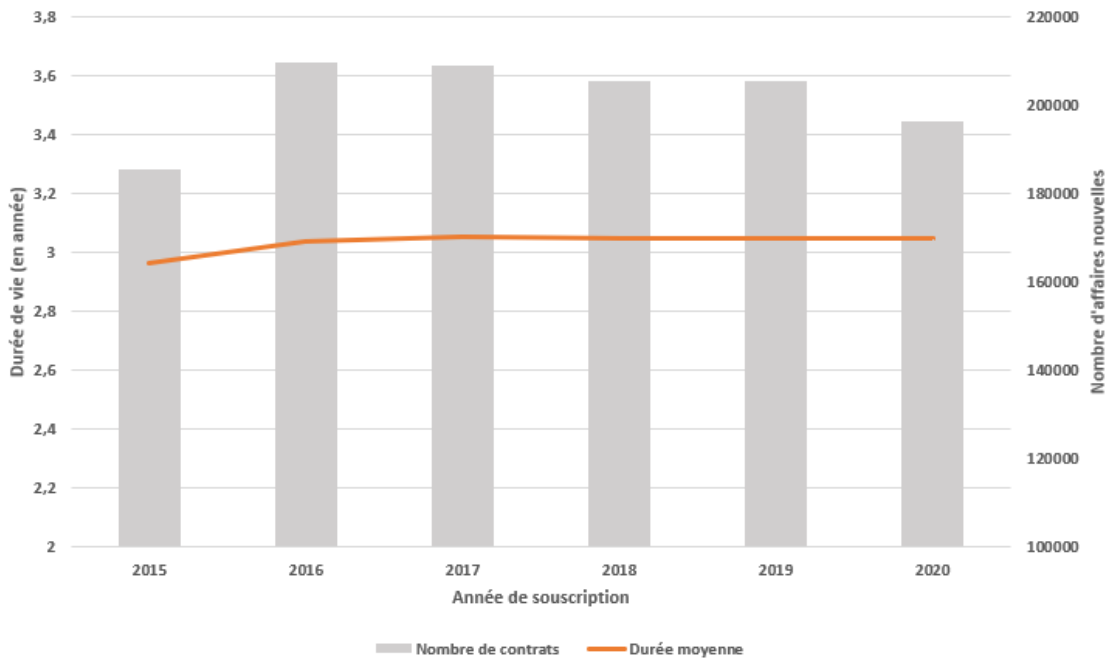


FIGURE 3.14 – Durées de vie moyennes prédites suivant l'année de souscription

3.3.2 Détermination des durées suivant les variables les plus discriminantes

Nous allons désormais regarder les durées de vie moyennes prédites des contrats suivant les variables les plus discriminantes.

Le tableau 3.4 ci-dessus présente les durées de vie moyennes prédites suivant la qualité juridique, l'âge du client et le nombre de contrats du client. Pour chaque variable, les durées sont décroissantes par modalité.

Sur 4 ans, la durée de vie des propriétaires est d'environ 3,5 ans et est supérieure à celle des locataires qui est égale à 2,6 ans soit une différence de presqu'une année. Ceci confirme que la qualité juridique est l'un des critères les plus discriminants pour la durée de vie.

Variable	Modalités	Durée de vie moyenne
Qualité juridique	Propriétaire Locataire	3,5 ans 2,6 ans
Age du client	Seniors Adultes Jeunes	3,3 ans 3 ans 2,6 ans
Nombre de contrats	Plusieurs contrats 1 contrat	3,2 ans 2,8 ans

TABLE 3.4 – Durées de vie moyennes prédites (sur 4 ans) suivant 3 variables discriminantes

La figure 3.15 suivante donne la durée de vie moyenne prédite suivant le zonier incendie des contrats souscrits en 2020.

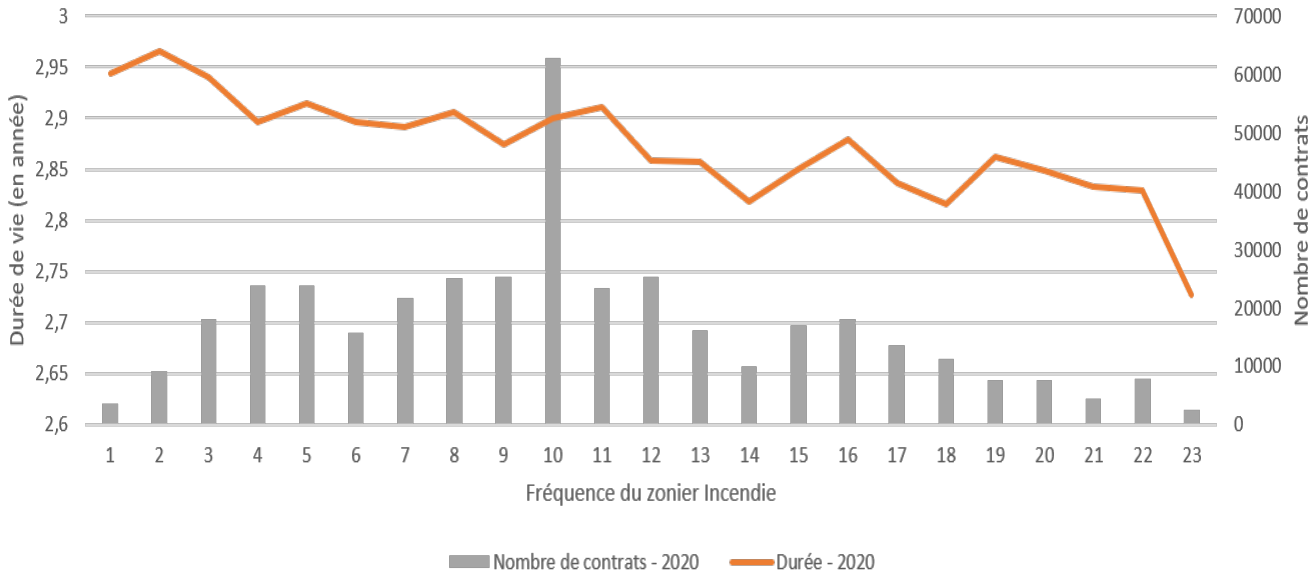


FIGURE 3.15 – Durées de vie moyennes prédites des contrats souscrits en 2020 suivant le zonier Incendie (en %)

Nous remarquons que la durée de vie moyenne passe de 2.94 ans à 2.73 ans soit une différence de 21% lorsque la fréquence du zonier incendie passe de 1% à 23%. La courbe lissée aurait une allure décroissante. Cependant, elle n'est pas lisse. Lorsque des zones ont des fréquences proches, ce n'est pas forcément la zone ayant la plus forte fréquence qui a la durée de vie la plus élevée. Par exemple, les zones de fréquences 35% et 36% ont une durée de vie plus élevée que les zones de fréquence 34%.

Nous allons à présent déterminer les durées de vie moyennes des différents parcours.

3.3.3 Détermination des durées suivant le parcours

Nous allons utiliser les contrats sosucrits en 2019 et 2020 puisque c'est à partir de 2019 que certains parcours ont débuté.

La figure 3.16 suivante présente les durées de vie moyennes des différents parcours pour les contrats souscrits en 2019 et en 2020.

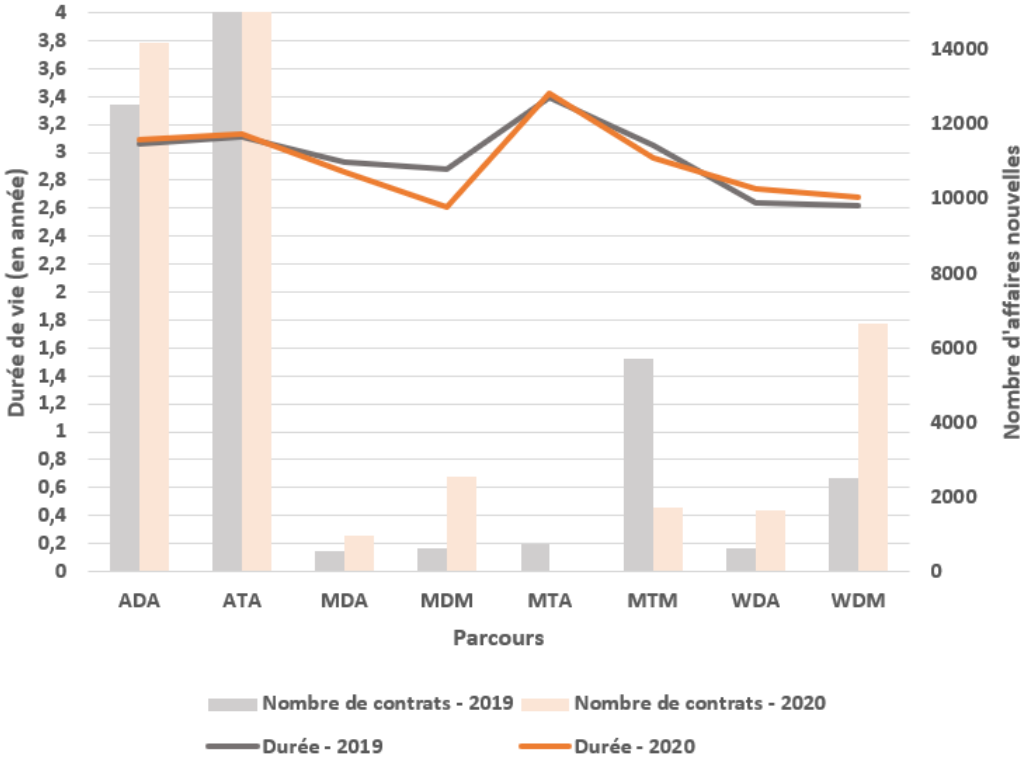


FIGURE 3.16 – Durées de vie moyennes prédites des contrats souscrits en 2019 et 2020 suivant les différents parcours

Nous remarquons que les durées de vie des contrats souscrits en 2019 et 2020 sont proches. Les parcours Web ont les durées de vie les moins élevées. Dans les parcours MDA et MTA, il y a peu de données pour conclure. Les parcours des contrats souscrits et gérés en agence ADA et ATA ont les durées de vie les plus élevées. Cela s'explique par le fait que ces parcours présentent plus de profils susceptibles de rester longtemps dans le portefeuille c'est-à-dire plus de propriétaires, de multi-détenteurs de contrats et de seniors comme nous l'avions vu dans la section analyse descriptive de la première partie 1.3. A l'opposé, les parcours Web ont les profils les plus susceptibles de rester moins de temps dans le portefeuille c'est-à-dire plus de locataires, de mono-détenteurs de contrats et de jeunes.

Le tableau 3.5 suivant présente les durées de vie moyennes des différents parcours suivant la qualité juridique (variable la plus discriminante) et au total pour les contrats souscrits en 2019 et 2020. Nous remarquons que les durées de vie moyennes des propriétaires dans les parcours agence ne sont pas très loin d'être égales à 4 ans (3,40 ans environ) tandis que celles des locataires sont proches de 3 ans pour les parcours agence (2,60 ans environ). Les parcours Web ont toujours les durées de vie les moins élevées. Les différences de durée de vie entre locataire et propriétaire peuvent atteindre 1 an dans certains parcours comme les parcours Web.

	2019			2020		
Parcours	Locataire	Propriétaire	Total	Locataire	Propriétaire	Total
ADA	2,63 ans	3,42 ans	3,06 ans	2,65 ans	3,46 ans	3,09 ans
ATA	2,65 ans	3,49 ans	3,11 ans	2,67 ans	3,52 ans	3,13 ans
MDA	2,43 ans	3,33 ans	2,93 ans	2,38 ans	3,26 ans	2,86 ans
MDM	2,43 ans	3,25 ans	2,88 ans	2,07 ans	3,01 ans	2,61 ans
MTA	2,75 ans	3,63 ans	3,39 ans	2,77 ans	3,68 ans	3,42 ans
MTM	2,61 ans	3,41 ans	3,05 ans	2,62 ans	3,22 ans	2,96 ans
WDA	2,15 ans	3,25 ans	2,64 ans	2,24 ans	3,36 ans	2,74 ans
WDM	2,12 ans	3,22 ans	2,62 ans	2,18 ans	3,29 ans	2,68 ans

TABLE 3.5 – Durées de vie moyennes prédites des différents parcours suivant la qualité juridique (variable la plus discriminante) et au total pour les contrats souscrits en 2019 et 2020

Conclusion du chapitre

Notre objectif est de modéliser la durée de vie des contrats d'assurance habitation afin d'optimiser la rentabilité de la stratégie multi-accès. Dans cette troisième partie, nous avons essayé de déterminer les durées de vie en utilisant les modèles de taux de résiliation des différentes périodes obtenus dans la partie précédente.

En essayant de déterminer l'influence des variables sur la durée de vie et mieux comprendre le modèle de durée de vie, plusieurs méthodes permettant d'interpréter des modèles boites noires (*black box*) comme le XGBoost (qui est le modèle choisi) ont été utilisées. Il s'agit du PDP (*Partial Dependence Plot*), l'ICE (*Individual Conditionnal Expectation*) et les *SHAP values* (qui a permis de tracer des *SHAP summary plots*). Nous avons pu en déduire que les locataires ont une durée de vie moyenne inférieure à celle des propriétaires, les seniors ont une durée de vie moyenne supérieure à celle des adultes et jeunes. Les assurés multi-détenteurs ont sans suprise une durée de vie moyenne plus élevée que celle des assurés n'ayant qu'un seul contrat.
Exploiter les modèles a aussi aidé à constater que la Covid-19 (à cause des confinements et des couvre-feux) a permis de réduire les résiliations mais en général les durées de vie de profils similaires sont stables dans le temps. Les écarts de durée de vie d'une année à une autre s'expliquent par les différences de profils trouvés.

En utilisant les méthodes d'interprétations des boites noires déjà mentionnées, nous avons pu confirmer la remarque faite dans la partie précédente nous poussant à croire que les parcours ont très peu d'impact sur la durée de vie puisqu'ils ont très peu d'influence sur les résiliations peu importe la période. Si les parcours ont des durées de vie moyennes différentes (sur quatre ans) c'est à cause des différents profils qui les composent, les parcours ayant plus de profils moins susceptibles de résilier ont les durées de vie les plus élevées (parcours agence). Ainsi, plus besoin de déterminer quels profils auraient duré plus longtemps s'ils étaient dans un autre parcours. La dominance du parcours traditionnelle (ATA) qui représente plus de 90% du portefeuille ainsi que les données assez faibles dans certains parcours font que cette théorie ne peut pas être complétement confirmée mais nous restons dessus pour le mémoire.

Les durées de vie calculées vont permettre d'entamer la quatrième et dernière partie du mémoire dans laquelle nous allons enfin essayer d'optimiser la rentabilité de la stratégie multi-accès.

Chapitre 4

Optimisation de la rentabilité

Il y a désormais dans chaque parcours, différents profils ainsi que leur durée de vie moyenne sur 4 ans. Nous pouvons donc essayer d'optimiser la rentabilité de ces parcours sur 4 ans par la sélection des profils à l'affaire nouvelle.
Pour cela, nous allons tout d'abord identifier les profils les plus loyaux c'est-à-dire ceux qui restent le plus longtemps dans le portefeuille afin d'orienter la souscription sur ces contrats. Nous allons par la suite prendre en compte un indicateur de rentabilité sur la durée de vie qui est le ratio sinistres à primes afin de faire ce qui pourrait être vu comme une optimisation sans contrainte avant d'entamer l'optimisation sous contraintes.

L'optimisation sans contrainte va permettre d'identifier :

— quels profils pourraient encore être acceptés en tant qu'affaires nouvelles dans chaque parcours afin de maximiser la rentabilité,

— quels profils il faudrait arrêter d'accepter dans un parcours donné (et dans quels parcours ils pourraient être acceptés).

Cette partie peut aussi constituer une aide à l'élaboration de stratégie permettant d'améliorer la rentabilité de certains profils (notamment ceux que nous serions contraints d'accepter en affaire nouvelle avec l'optimisation sous contraintes).

L'optimisation sous contraintes va permettre d'identifier :

— quels profils il faudrait avoir dans chaque parcours afin de maximiser la rentabilité tout en satisfaisant plusieurs contraintes comprenant les coûts. Il faut prendre en compte un certains nombre de contraintes lors de l'optimisation.

— quels parcours pourraient être éliminés et quels parcours devraient être plus développés (accueillir plus de clients).

4.1 Identification des profils les plus loyaux

Afin de déterminer les profils les plus loyaux, les durées de vie des différents contrats sont estimés grâce à notre modèle de durée. Ces contrats vont ensuite être regroupés suivant les variables discriminantes [1] pour la durée de vie afin de constituer des profils. Ces variables sont :

1. la qualité juridique,
2. le nombre de contrats du client,

1. vues dans la partie précédente

3. l'âge du client,
4. les zoniers,
5. la nature de la résidence,
6. le niveau du capital croisé avec le nombre de pièces,
7. la csp (catégorie socio-professionnelle),
8. le code firme [2],
9. le pourcentage d'objet de valeur,
10. la franchise,
11. la situation familiale.

Cela résulte sur 354 285 profils. Nous ne pouvons pas d'un simple coup d'oeil ni à l'aide d'analyse multivariée résumer ce qui définit un profil qui dure longtemps dans le portefeuille. Nous savons par exemple qu'un locataire a moins de chance de durer dans le portefeuille, mais il pourrait y avoir des profils de locataire qui arrivent quand même à durer jusqu'à 4 ans et nous voulons savoir ce qui caractérise ces profils de locataire. De même, un propriétaire a plus de chance de durer dans le portefeuille qu'un locataire, mais il pourrait y avoir des propriétaires qui ne durent pas aussi longtemps que certains locataires et nous voulons savoir ce qui caractérisent ces profils de propriétaire.

Nous allons alors voir ce qui définit un profil suivant sa durée de vie et donc identifier ce qui différencie les profils qui durent le plus des profils qui durent le moins. Un outil permettant facilement d'y arriver est l'arbre de décision [3]. Nous allons donc utiliser les 354 285 différents profils composant la base ainsi que les variables qui ont permis de définir ces profils afin de construire l'arbre de décision.

Construction de l'arbre de décision

Vu que nous souhaitons utiliser un arbre de décision de classification, nous allons découper les durées de vie en quatre classes :

— $classe = 1$: durée de vie moyenne ≤ 1 an.
— $classe = 2$: 1 an $<$ durée de vie moyenne ≤ 2 ans.
— $classe = 3$: 2 ans $<$ durée de vie moyenne ≤ 3 ans.
— $classe = 4$: 3 ans $<$ durée de vie moyenne ≤ 4 ans.

La variable à expliquer est donc la variable *classe* (de durée de vie). Les variables explicatives sont les différentes variables (citées plus haut) qui caractérisent les profils. Il y a donc réduction d'environ 60% des variables de la base de données de modélisation.

Les contrats de la base de données initiale ont été regroupés par profil c'est-à-dire suivant les variables mentionnées. Cela réduit donc considérablement le bruit. De même, les 4 classes de durée de vie peuvent être équilibrés par ré-échantillonnage de la base d'entraînement. Nous espérons donc obtenir un arbre de décision performant.

2. code désignant un pourcentage de réduction sur la prime
3. Voir Annexe A.5

Performance de l'arbre de décision

Les 4 classes de durée de vie ont été équilibrées par SMOTE (déjà vu dans la partie 2.1.3) puisque sans ré-échantillonnage de la base d'entrainement, les performances étaient moins bonnes.
Les taux de vrais positifs et de faux positifs vont être calculés. Le taux de vrais positifs d'une classe représente la proportion de profils appartenant effectivement à la classe prédite par l'arbre parmi les profils prédits comme appartenant à cette classe. Le taux de faux positifs d'une classe A par rapport à une classe B représente la proportion de profils appartenant réellement à la classe A mais dont l'appartenance est prédite par l'arbre comme étant à la classe B. Leur formule de calcul peut être trouvé en Annexe A.2.
Nous remarquons que pour les classes qui nous intéressent principalement (3 et 4), les taux de faux positifs sont inférieurs à 10% tandis que les taux de vrais positifs sont tous supérieurs à 80%. L'*accuracy* est égale à 92.5%. Les performances de l'arbre de décision sont donc excellentes.

Description des profils les plus loyaux [4]

La figure 4.1 suivante présente l'arbre de décision de classification permettant de donner la classe d'un profil suivant ses caractéristiques. L'arbre de décision a été construit avec le package *rpart* [5] de *R* (RDocumentation, 2019). L'arbre de décision permet donc de déterminer à quelle classe de durée de vie appartient le contrat suivant plusieurs critères.

L'arbre s'interprète de la manière suivante :

— les noeuds permettent de tester une condition dépendant de la valeur d'une des variables explicatives.

— les branches permettent de voir le résultat obtenu suite au test. Elles peuvent donc aboutir sur un autre noeud ou sur une feuille.

— les feuilles sont les noeuds terminaux. Elles prédisent le résultat de la classification et permettent donc de voir :

1. la classe à laquelle appartient le profil (première ligne de la feuille cest-à-dire du rectangle présentant le résultat),
2. la probabilité d'appartenance du profil à chacune des quatre classes (deuxième ligne). Rappelons que les durées sont celles qui sont obtenues avec le modèle de durée de vie. Les probabilités sont donc des probabilités par rapport aux prédictions du modèle de durée de vie et non par rapport aux durées de vie réelles,
3. le pourcentage de profils vérifiant les critères lus sur le(s) noeud(s) (troisième ligne).

En parcourant totalement l'arbre, nous remarquons que :
— Les **propriétaires** ont 87% de chance de se trouver dans la classe 4 et donc de durer plus de 3 ans et 13% de chance de durer entre 2 ans et 3 ans.
— Les **locataires** ont moins de chance de durer plus de 3 ans. La probabilité qu'ils durent plus de 3 ans dépend d'autres caractéristiques comme l'âge, le nombre de contrats et certaines variables du contrat comme la valeur de la franchise, le pourcentage d'objet de valeur. Ainsi,

—— les locataires ayant plusieurs contrats et plus de 65 ans ont 92% de chance de durer plus de 3 ans. 4% de profils vérifient ces critères.

—— les locataires seniors n'ayant qu'un seul contrat ont 16% de chance de durer plus de 3 ans.

4. qui durent le plus dans le portefeuille

5. Le package *R* nommé *rpart* propose une implémentation des méthodes de construction d'arbres de décision inspirées de l'approche CART de Breiman, Friedman, Olshen et Stone en 1984.

Ils représentent 3% des profils.

—— les locataires adultes ayant plusieurs contrats et un pourcentage d'objet de valeur non nul ont 67% de chance de durer plus de 3 ans et 33% de chance de durer plus de 2 ans. 5% des profils vérifient ces critères.

—— les locataires jeunes ou adultes qui n'ont qu'un seul contrat ont très peu de chance de durer plus de 3 ans (2%). Lorsque leur franchise est de moins de 225€, ils ont 1% de chance de durer plus de 3 ans et 66% de chance de durer plus de 2 ans. Avec 225€ ou plus comme franchise, ils ont 42% de chance de durer plus de 2 ans.

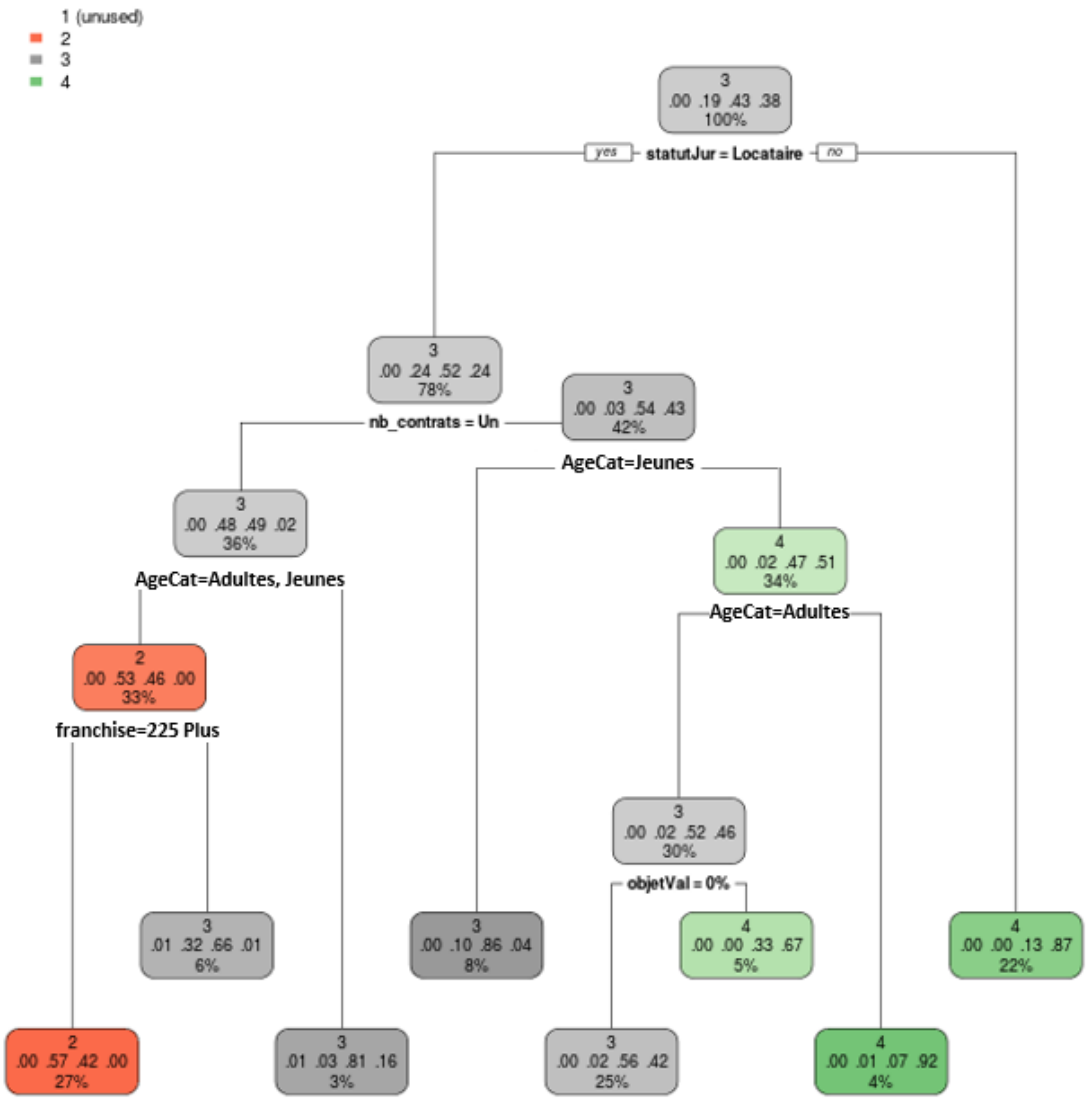


FIGURE 4.1 – Arbre de décision permettant de déterminer les caractéristiques des profils suivant leur classe de durée

Les profils les plus loyaux sont donc généralement des propriétaires, des locataires seniors ou des locataires adultes ayant plusieurs contrats. Nous encouragons la souscription sur ces profils. Nous avons déjà vu que le parcours n'a pas d'influence sur la résiliation. Les profils les plus loyaux sont donc les mêmes quel que soit le parcours. Mais à durées de vie égales, le profil ayant le meilleur S/C est le plus rentable. Nous allons donc prendre en compte le S/C sur la durée de vie comme indicateur de rentabilité afin de déterminer les profils les plus rentables se trouvant dans chaque parcours.

4.2 Indicateur de rentabilité et coût d'acquisition

4.2.1 Détermination des S/C sur la durée de vie

L'indicateur utilisé pour essayer d'optimiser la rentabilité est le rapport sinistres à primes (S/C) comme mentionné plusieurs fois.
Le ratio S/C correspond au rapport entre le montant des sinistres et celui des cotisations encaissées sur un même contrat d'assurance.

$$\text{Ratio sinistres à primes (S/C)} = \frac{\text{charge totale des sinistres}}{\text{somme totale des primes acquises}}. \tag{4.1}$$

Lorsque le rapport S/C est supérieur ou égal à 100%, la charge des sinistres est supérieure aux cotisations encaissées : le contrat n'est pas rentable. Pour notre étude, nous aurions aimé utiliser un ratio S/C prédit (PSC *Predicted* S/C) pour chaque contrat. Le *predicted* S/C (PSC) est le rapport entre la sinistralité à priori estimée par le tarif technique à horizon un an sur la prime. Il est donc obtenu en modélisant la prime pure. Cependant, nous ne disposons pas du S/C prédit.
Nous avons donc décidé de calculer les S/C en observant la sinistralité des profils de la base de modélisation sur 4 ans. L'inconvénient de l'utilisation des S/C réels est la volatilité qui va constituer un frein au calcul des S/C par profil. En effet, un individu A qui dure 4 ans dans le portefeuille peut avoir une charge des sinistres totale supérieure à celle d'un individu B qui n'a fait qu'un an dans le portefeuille parce qu'un événement naturel s'est produit pendant deux années alors que l'individu B n'était pas encore dans le portefeuille, par exemple. Si le *predicted* S/C avait été utilisé, chaque contrat ou chaque profil aurait eu un S/C représentatif et pas volatile.

Nous souhaitons alors **estimer le ratio S/C suivant la durée de vie prédite** en calculant le S/C réel de certains profils suivant leur durée de vie réelle et suivant certaines variables. Les affaires nouvelles du 1er Janvier 2015 au 31 Décembre 2019 de la base vont être utilisées afin de calculer leur S/C suivant leur durée de vie réelle. Nous allons donc observer la sinistralité de ces contrats du 1er Janvier 2015 au 31 Décembre 2019. L'année 2020 étant une année particulière à cause de la Covid-19, les sinistres survenus cette année n'ont pas été inclus.

Après avoir calculé le S/C suivant l'âge du contrat, il est possible de déterminer le S/C des profils sur leur durée de vie prédite. Cependant, la volatilité des S/C réels obtenus nous pousse à faire une segmentation. Lors de ce processus, il faut tenir en compte qu'il doit y avoir suffisamment de volume dans chaque segment. Lorsqu'un segment a peu de contrats, il pourrait arriver qu'aucun de ses contrats n'ait eu de sinistres, ce qui risque d'aboutir sur un S/C égal à 0 et donc non représentatif du risque de ce segment. Il faut alors bien choisir les variables utilisées pour la segmentation : elles doivent êtres discriminantes et non corrélées.

Pour choisir les variables, nous allons utiliser le XGBoost pour déterminer les variables les plus discriminantes. En *boosting*, lorsque deux variables A et B sont corrélées, toute l'importance sera en théorie sur l'une des deux variables. En effet, si l'algorithme choisit A (aléatoirement) pour construire l'arbre, B ne sera pas utilisée ensuite pour la construction de cet arbre vu qu'elle est redondante. En boosting, lorsqu'un lien spécifique entre variable explicative et résultat aura été appris par l'algorithme, il essaiera de ne pas se recentrer dessus (en théorie c'est ce qui se passe, la réalité n'est pas toujours aussi simple).
Les variables explicatives utilisées pour la construction du modèle sont principalement les mêmes que celles utilisées pour la modélisation des taux de résiliation. La variable à expliquer est le S/C. Le package *h2o* est utilisé sur python.
L'importance des variables obtenue est alors présentée dans la figure 4.2 ci-dessous.

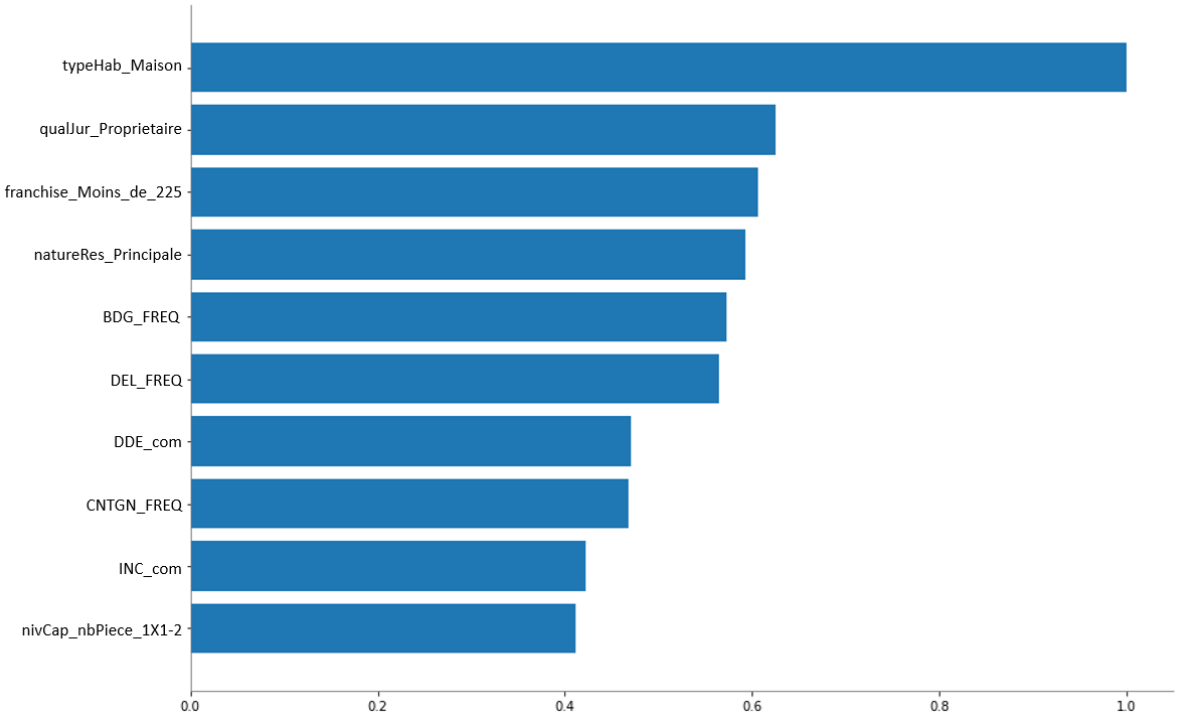


FIGURE 4.2 – Figure présentant les 10 variables les plus importantes dans l'estimation du S/C par XGBoost

Nous allons donc croiser les variables les plus importantes vu dans cette figure en les ajoutant une à une par ordre d'importance. Nous remarquons qu'à partir de 16 segments, il commence à y avoir des segments ayant peu de volume. Par soucis de volatilité, il ne faut pas faire plus de segments. N'oublions pas que le S/C est aussi calculé suivant la durée du contrat dans le portefeuille.
Voici donc les quatre variables avec le S/C le plus discriminant utilisées :

— le type d'habitation,

— la qualité juridique,

— la nature de la résidence,

— la franchise.

Le S/C suivant ces quatre variables et la durée de vie **réelle** (âge du contrat) s'obtient avec la formule suivante

$$\text{Ratio sinistres à primes}_{h,q,r,f,a} = \frac{S_{h,q,r,f,a}}{C_{h,q,r,f,a}}, \tag{4.2}$$

avec $S_{h,q,r,f,a}$ la charge totale des sinistres lorsque le type d'habitation=h, la qualité juridique=q, la nature de la résidence=r, la franchise=f et l'âge du contrat=a et

. $C_{h,q,r,f,a} = \sum \text{prime} \times \text{exposition}$ lorsque le type d'habitation=h, la qualité juridique=q, la nature de la résidence=r, la franchise=f et l'âge du contrat=a.

La charge totale des sinistres comprend les charges attritionnelles, les graves et les événements naturels. Lorsqu'il est dans un portefeuille, un contrat peut subir plusieurs avenants. Chacun de ces avenants peut être associé à une prime différente (parce que l'assuré a opté pour une garantie supplémentaire par exemple). L'assuré peut donc, par exemple, avoir une prime de 210 euros la première moitié de l'année ($exposition = 0.5$) et 250 euros la deuxième moitié ($exposition = 0.5$), il a donc payé 230 euros cette année comme prime. Ces avenants sont repérés par des dates de début et de fin de la période de changement. La différence entre ces deux dates constituent l'**exposition**. En mutipliant la prime correspondante par l'exposition, nous obtenons la prime proratisée qui correspond à la **cotisation encaissée**. Proratiser la prime permet aussi de ne pas prendre en compte les périodes de suspension du contrat.

Le tableau 4.1 suivant présente les ratios S/C obtenus pour les propriétaires d'appartement suivant la nature de la résidence, la franchise (en €) et l'âge du contrat (en année) sur 4 ans. Les chiffres ont été modifiés. Les propriétaires d'appartement en résidence principale avec une franchise de moins de 225€ et qui ont duré 1 an ont un S/C de 72% tandis que ceux qui ont duré 4 ans ont un S/C de 75%.

S/C – Appartement Propriétaire	Année dans le portefeuille				
Nature Résidence x Franchise	1	2	3	4	Total général
Principale	**84%**	**79%**	**74%**	**71%**	**76%**
225 +	88%	85%	77%	70%	80%
Moins de 225	72%	63%	66%	75%	67%
Secondaire	**65%**	**64%**	**60%**	**55%**	61%
225 +	67%	62%	56%	58%	**61%**
Moins de 225	60%	70%	67%	47%	62%
Total général	**79%**	**75%**	**71%**	**67%**	72%

TABLE 4.1 – Ratio S/C des propriétaires d'appartement suivant la nature de la résidence et la franchise (en €) ainsi que l'année passée par le contrat dans le portefeuille

De même, nous remarquons qu'en général le S/C baisse lorsque la durée de vie augmente. Pour les propriétaires d'appartement en résidence principale, elle passe de 84% pour les contrats de moins d'un an à 71% pour les contrats de plus de 3 ans. De même, les S/C des personnes assurant une résidence principale sont supérieurs à ceux des personnes assurant une résidence secondaire. Nous remarquons donc déjà que les propriétaires d'appartement en résidence principale sont moins rentables que ceux en résidence secondaire.

Lors de l'optimisation, nous devons aussi prendre en compte le coût d'acquisition client.

4.2.2 Coût d'acquisition

Le coût d'acquisition client désigne le montant moyen dépensé pour générer un client. Rappellons la valeur des coûts d'acquisition suivant le parcours en 2020 dans le tableau 4.2. **Le coût d'acquisition du parcours MTA n'a pas été calculé en 2020 puisque le parcours ne dispose que de 0.003% de données cette année. Il sera donc exclus dans la suite de l'étude**.

Parcours	Coût d'acquisition en 2020
ADA	56 €
ATA	18 €
MDA	167 €
MDM	197 €
MTM	142 €
WDA	1 €
WDM	31 €

TABLE 4.2 – Tableau présentant les différents parcours ainsi que leur coût d'acquisition en 2020

Nous remarquons donc que les parcours ont des coûts d'acquisition différents et que les parcours plateforme (commençant par la lettre M) ont les coûts d'acquisition les plus élevés. Chez Allianz, le coût d'acquisition se décompose en coût d'acquisition *sourcing* et en coût d'acquisition *hors sourcing*. Le *sourcing* désigne l'origine du contrat, le premier contact avec le futur client.
Le coût d'acquisition *sourcing* correspond aux investissements média et aux commissions des comparateurs d'assurance. Il est net de la refacturation des *leads*[6] et de celle de l'intermédiation aux agents. Le coût d'acquisition *sourcing* est donc supposé être égal à 0 pour les agents. C'est l'une des raisons pour lesquelles leur coût d'acquisition est bas.
Le coût d'acquisition hors *sourcing* se décompose en 2 parties : la première est commune à tous les parcours (elle représente les frais fixes d'Allianz France attribuable à la vente de contrats) tandis que la deuxième est spécifique aux parcours plateforme (elle constitue le coût des appels téléphoniques et est la raison pour laquelle le coût d'acquisition des parcours plateforme est aussi élevé).

4.2.3 Classification des profils suivant le S/C et le coût d'acquisition

Nous souhaitons déterminer les profils que la compagnie devrait accepter d'avoir comme affaire nouvelle dans chaque parcours afin de maximiser la rentabilité sur 4 ans sous plusieurs contraintes comme le coût d'acquisition total.
La détermination des S/C des différents profils de durée de vie de la base résulte sur 227 valeurs différentes de S/C et de coût d'acquisition. Compte tenu du fait que la maille utilisée pour calculer les S/C est trop large, il y a des risques que les S/C obtenus pour chaque profil ne soient pas représentatifs. Pour éviter cela, les S/C vont donc être regroupés par intervalle. De ce fait, plusieurs groupes de profils avec des S/C proches seront formés. Chaque groupe étant défini par un intervalle de S/C, cela entrainerait que la valeur précise du S/C des profils appartenant au groupe pourrait se trouver dans l'intervalle. De plus, nous aurions plusieurs choix de profils possibles dans un groupe ce qui fait que nous n'allons pas indexer un profil précis, ce qui serait risqué vu que les S/C ne sont pas précis.

Nous allons donc classer les profils suivant le S/C que nous venons de déterminer mais aussi le coût d'acquisition qui va représenter les différents parcours. Comment alors les classifier ? L'inconvénient

6. Les leads désignent des contacts que la compagnie espère voir se transformer en client.

de la séparation en 2 groupes pour chaque parcours est qu'il pourrait y avoir à l'intérieur d'un groupe des profils dont le S/C est très éloigné. Aussi quel seuil faut-il choisir pour séparer les groupes de S/C ? Pour trouver la réponse à ces questions, nous décidons donc de faire un partitionnement de données en utilisant l'algorithme *K-means*.

Data Clustering

Le *data clustering* (en anglais) ou partitionnement de données (en français) est une méthode d'analyse des données ayant pour but de diviser un ensemble de données en plusieurs groupes homogènes appelés *clusters*. Ainsi, les éléments d'un même *cluster* ont une ou plusieurs caractéristique(s) commune(s) correspondant le plus souvent à des critères de proximité définis en introduisant des mesures et classes de distance entre objets. Pour obtenir un bon partitionnement des données, il faut à la fois minimiser l'inertie intra-classe afin d'obtenir des *clusters* les plus homogènes possibles et maximiser l'inertie inter-classe afin que les *clusters* soient les plus différents possibles.

Détermination du nombre de *clusters*

Contrairement à l'apprentissage supervisé où il existe des métriques très pertinentes pour évaluer les performances du modèle, l'analyse de *clustering* n'a pas de métrique d'évaluation solide pouvant être utilisée afin d'évaluer le résultat de différents algorithmes de *clustering*.

Afin de déterminer le nombre optimal de *clusters*, la méthode du coude ou l'*Elbow Method* en anglais est souvent utilisé. En *clustering*, l'utilisation du «coude» ou du «genou d'une courbe» permet de choisir un certain nombre de *clusters* de tel sorte que l'ajout d'un autre *cluster* ne donne pas une bien meilleure modélisation des données. L'idée de cette méthode repose sur le fait que les premiers *clusters* expliqueront considérablement les variations et ajouteront donc fortement des informations ce qui les rend nécessaires. Cependant, une fois que le nombre de *clusters* dépasse le nombre réel de groupes formés idéalement par les données, les informations ajoutées baisseront fortement puisque les *clusters* supplémentaires ne feront que subdiviser les groupes réels. Si cela se produit, un coude aigu pourrait être identifié dans le graphique de la variation expliquée par rapport aux *clusters*. Ce coude aigu est caractérisé par une augmentation rapide jusqu'à K (zone de sous-ajustement) suivie d'une augmentation lente après K (zone de sur-ajustement). Cependant, en pratique, un tel «coude» n'est pas toujours identifié sans ambiguïté.

Sur SAS, la méthode du coude peut être utilisée avec le Critère de *Clustering* Cubique (CCC) ou *Cubic Clustering Criterion* en anglais afin d'estimer le nombre de *clusters* à l'aide de *K-means* ou d'autres méthodes basées sur la minimisation de la somme des carrés à l'intérieur du *cluster*. D'après la documentation SAS (Institute Inc. SAS, 1983), le Critère de *Clustering* Cubique est obtenu en comparant le R^2 observé au R^2 approximatif attendu en échantillonnant sur une distribution uniforme. Le R^2 s'interprète comme la proportion de variance expliquée par les *clusters*. Les valeurs positives du CCC signifient ainsi que le R^2 obtenu est supérieur à ce qui serait attendu si l'échantillonnage se faisait à partir d'une distribution uniforme et indiquent donc la présence possible de *clusters*.

La figure 4.3 suivante est un graphique montrant l'évolution du CCC en fonction du nombre de *clusters* obtenu sur SAS suivant le S/C et le coût d'acquisition du parcours.
La documentation SAS sur le CCC indique aussi comment il est interprété. Sur le graphique représentant le CCC en fonction du nombre de *clusters*, les pics avec un CCC supérieur à 2 ou 3 indiquent un bon *clustering* tandis que les pics avec le CCC entre 0 et 2 indiquent des *clusters* possibles mais doivent être interprétés avec prudence.

Les *clusters* sphériques non hiérarchiques très distincts montrent généralement une forte augmentation avant le pic suivi d'un déclin graduel tandis que des *clusters* elliptiques non hiérarchiques très distincts montrent souvent une forte augmentation jusqu'au nombre correct de *clusters* suivie d'une nouvelle augmentation graduelle et finalement d'un déclin graduel. Cependant, si les données ont une structure hiérarchique, le graphique peut présenter plusieurs pics.

Sur la figure 4.3 ci-dessous, nous remarquons qu'à partir de 16 clusters, le *Cubic Clustering Criterion* commence à considérablement baisser. De plus, le CCC du pic à 16 *clusters* est supérieur à 3 montrant qu'utiliser 16 *clusters* donne un bon *clustering*. **Nous choisissons alors K=16** *clusters*.

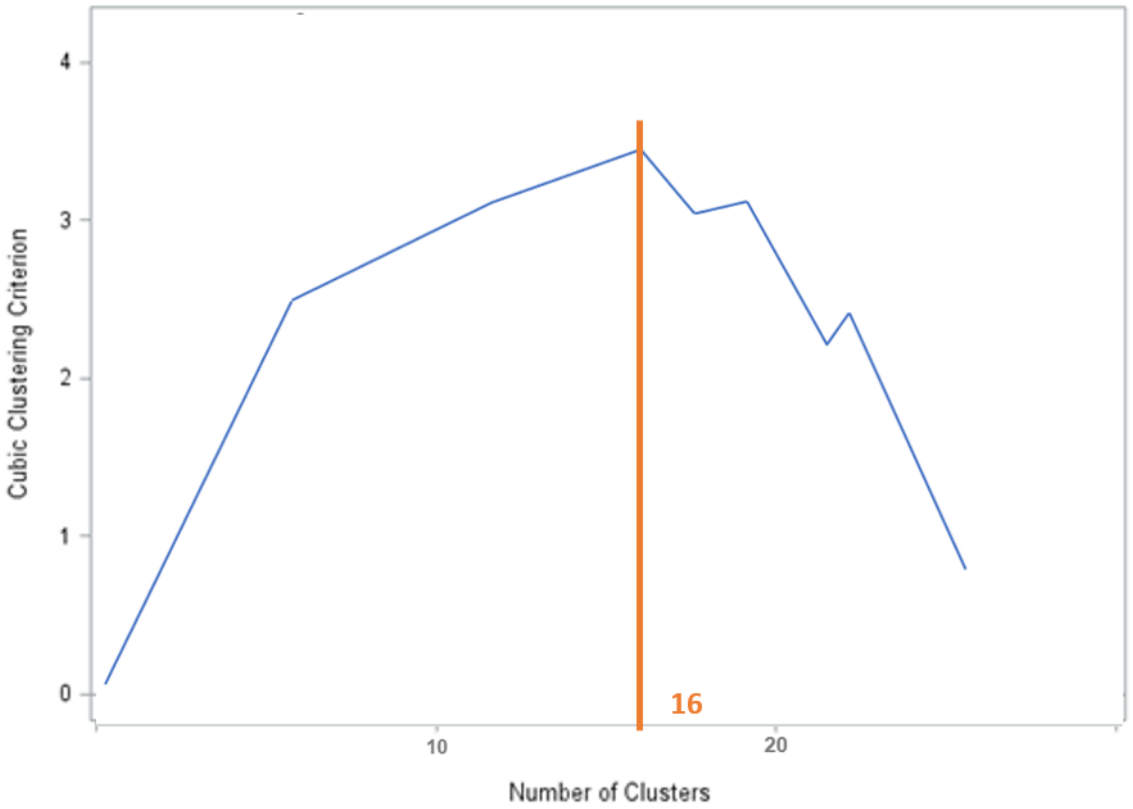


FIGURE 4.3 – Valeurs du *Cubic Clustering Criterion* en fonction du nombre de *clusters*

Description de l'algorithme K-means

Après avoir déterminé le nombre de *clusters* K comme étant égal à 16, l'algorithme des K-means est lancé sur SAS afin de définir les 16 *clusters*.

Algorithme 3 : K-means

Entrées : K (le nombre de clusters à former) et la base d'entrainement.
Sorties : K *clusters* (les K groupes formés).
Choisir aléatoirement K points de la base d'entraînement. Ces points sont les centres des *clusters* (nommé centroïdes);
tant que *il n'y a pas convergence(les centroïdes ne bougent plus lors des itérations)* **faire**
— Affecter chaque point de la base d'entraînement au groupe dont il est le plus proche du centre;
— Recalculer le centre de chaque *cluster* et modifier le centroïde.
fin

Description générale de chaque *cluster*

La figure 4.4 ci-dessous est un nuage de points qui représente les différents profils de la base suivant leur S/C (en ordonnée) et leur coût d'acquisition (en abscisse). Les points d'un même *cluster* sont coloriés avec la même couleur. Chaque coût d'acquisition représente un parcours (qui est écrit en noir). Rappelons que le parcours MTA est exclus de l'étude puisque son coût d'acquisition n'a pas été calculé en 2020 car le parcours ne dispose que de 0.003% de données cette année.

FIGURE 4.4 – Présentation des différents *clusters* obtenus en fonction de leur S/C (en ordonnée) et de leur coût d'acquisition CA (en abscisse)

Nous remarquons que chaque parcours est constitué de 2 *clusters* ou plus : un *cluster* regroupant les contrats les plus rentables, un autre avec les contrats moins rentables et éventuellement un troisième *cluster*. Les profils d'un même *cluster* ont des S/C proches. Il peut y arriver qu'il y ait certains profils

avec un S/C un peu éloigné comme les deux premiers points bleu du *cluster* 1. Les parcours ayant 3 *clusters* sont ceux ayant des profils avec les S/C les plus dispersés. ATA (parcours agence traditionnelle) et MDM (parcours plateforme initié en digital) sont les deux parcours ayant 3 *clusters*.

Nous allons ensuite voir la description des caractéristiques de chaque *cluster* dans le tableau 4.3 ci-dessous. Elle présente les différents *clusters* obtenus pour les parcours ainsi que certaines de leurs caractéristiques comme le nombre de profils qu'ils représentent, le pourcentage de contrats retrouvés dans le *cluster* par rapport au nombre total de contrats, leur S/C moyen, minimum et maximum et leur coût d'acquisition.

Cluster	Parcours	Nb profils	% portefeuille	SC	SC min	SC max	Cout Acq (en €)
9	ADA	1242	3,98%	0,53	0.2	0.67	56
4	ADA	521	1,38%	0,92	0.71	1.26	56
14	ATA	1238	17,18%	0,43	0.2	0.5	18
10	ATA	2353	57,93%	0,68	0.53	0.82	18
15	ATA	358	14,61%	1,01	0.83	1.26	18
12	MDA	88	0,11%	0,50	0.2	0.61	167
6	MDA	93	0,02%	0,86	0.64	1.03	167
16	MDM	241	0,48%	0,52	0.2	0.6	197
7	MDM	354	0,84%	0,69	0.61	0.84	197
8	MDM	147	0,06%	1,02	0.85	1.26	197
3	MTM	1137	1,95%	0,55	0.2	0.72	142
5	MTM	485	0,52%	0,93	0.73	1.26	142
11	WDA	214	0,15%	0,49	0.2	0.61	1
2	WDA	178	0,04%	0,82	0.64	1.03	1
13	WDM	340	0,66%	0,49	0.2	0.64	31
1	WDM	235	0,10%	0,86	0.65	1.08	31

TABLE 4.3 – Présentation des différents *clusters* obtenus dans chaque parcours, du nombre de profils qu'ils présentent, du pourcentage de contrats retrouvés dans le *cluster* par rapport au nombre total de contrats, de leur S/C moyen, minimum et maximum et de leur coût d'acquisition

Nous vérifions bien que chaque parcours est constitué de 2 *clusters* ou plus. Le seuil pour séparer le *cluster* plus rentable du *cluster* moins rentable est compris entre 50% et 72%. Le seuil n'est évidemment pas le même par parcours puisque des profils différents se trouvent dans chaque parcours. Des profils avec des S/C > 1 peuvent être retrouvés dans les *clusters* de profils peu rentables mais nous les conservons car sous certaines contraintes, il pourrait être intéressant de les avoir. La longueur de l'intervalle de S/C de chaque *cluster* est en moyenne 41% (différence de 41% entre le S/C minimal et le S/C maximal du *cluster*).

Nous voulons désormais découvrir les profils qui se trouvent dans chaque *cluster* tout en déterminant les profils les plus rentables sans contraintes dans chaque parcours. Cela va non seulement permettre de voir ce qui définit les profils les plus rentables, d'identifier les profils les plus rentables vers lesquels il faudrait orienter la souscription mais aussi d'avoir une idée sur quelles stratégies pourraient être adoptées pour que les profils moins rentables deviennent plus rentables.

4.3 Identification des profils les plus rentables

Dans cette partie, nous n'utilisons pas les *clusters* mais nous allons les inclure dans les figures afin de visualiser les profils qui s'y trouvent.

4.3.1 Description des profils suivant la rentabilité et la durée de vie

Nous traçons un nuage de points représentant les différents profils (de S/C et de durée de vie) de chaque parcours en fonction de leur S/C en abscisse et leur durée de vie en ordonnée.
Nous **excluons la partie où le S/C** > 1 (profils pas rentables sur 4 ans) et nous découpons le graphe en 4 parties.

Seuil du S/C
Nous calculons le S/C moyen sur 4 ans. Il est égal à 64% (chiffre modifié). Nous définissons donc comme seuil sa partie entière inférieure (60%) et considérons que les profils avec un S/C inférieur sont plus rentables et ceux avec un S/C supérieur sont moins rentables. Les commissions et autres frais chez Allianz représentent en général 30% de la prime. Le seuil que nous avons choisi permet donc de bien couvrir ces frais puisqu'il laisse 40% de marge.

Seuil de la durée
Nous supposons que lorsque le profil dure plus de 2 ans dans certains parcours où le coût d'acquisition n'est pas élevé (agence (ATA et ADA) et Web (WDA et WDM)), son coût d'acquisition sera suffisamment amorti. Par exemple, pour un profil du parcours traditionnel ATA, avec un S/C de 60% [7], une durée de 2 ans et une prime égale à la prime moyenne du parcours qui est de 196€, la marge sur les 2 ans (en excluant la partie à utiliser pour les commissions et autres frais estimée à 30%) sera égale à $2 \times 10\% \times 196 = 39.20$€ . Les 39€ sont suffisants pour amortir le coût d'acquisition du profil dans ce parcours qui est de 18€ (n'oublions pas que la revalorisation n'a même pas été prise en compte). Nous choisissons donc 2 ans comme seuil de durée pour ces parcours. Pour les parcours avec un coût d'acquisition élevé comme les parcours plateforme (MTM, MDA, MDM, MTA), nous choisissons 3 ans comme seuil puisque nous estimons qu'il faut au moins 3 ans pour amortir le coût d'acquisition.

En découpant le nuage de points suivant ces seuils, nous obtenons alors la figure 4.5.
Comme nous pouvons le voir, elle est constituée de quatre parties :

- Une partie ***verte - Plus rentable*** : où se trouvent les profils qui durent le plus (plus de 2 ou 3 ans [suivant le parcours]) et qui ont un S/C $\leq 0{,}6$. Elle constitue les profils que nous souhaitons continuer à avoir en affaires nouvelles dans le parcours.
- Une partie ***rouge - Moins rentable*** où se trouvent les profils qui durent le moins (moins de 2 ans ou 3 ans [suivant le parcours]) et qui ont un S/C $> 0{,}6$. Elle constitue les profils à éviter dans le parcours. Cependant, nous pourrions étudier sous quelles conditions, il pourrait être convenable de les avoir dans le parcours.
- Une partie ***orange - Rentable mais faible durée*** où se trouvent des contrats très rentables (SC $\leq 0,6$) mais qui durent moins de 2 ans ou 3 ans [suivant le parcours]. Elle constitue les profils que nous pourrions continuer à avoir en affaires nouvelles dans le parcours en supposant que nous pourrions les convertir en ***Plus rentable*** en utilisant une certaine stratégie (pour augmenter leur durée de vie).
- Une partie ***bleu - Peu Rentable mais durée élevée*** où se trouvent des contrats peu rentables (S/C $> 0{,}6$) mais qui durent plus de 2 ans ou 3 ans [suivant le parcours]. Elle constitue les profils

7. seuil de S/C choisi

que nous pourrions continuer à avoir en affaires nouvelles dans le parcours en supposant que nous pourrions les convertir en ***Plus rentable*** en utilisant une certaine stratégie (pour augmenter leur S/C).

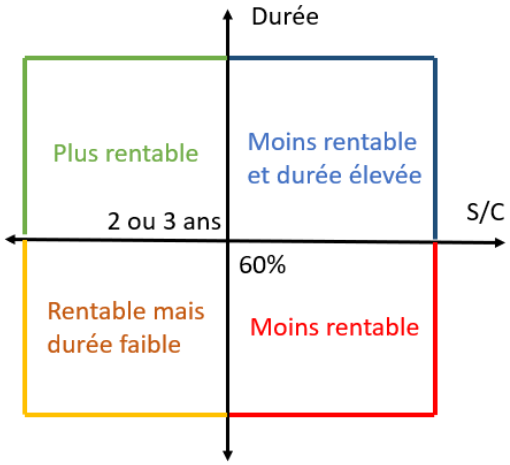


FIGURE 4.5 – Graphe permettant d'identifier les profils les plus rentables, rentables mais à durée faible, peu rentable à durée élevée et les moins rentables

Migration d'une partie à une autre
— Pour passer les profils de ***Rentable mais faible durée*** à ***Plus rentable***, augmenter leur durée de vie pourrait être envisagé. Nous pourrions réduire leur prime en leur offrant des mois gratuits vu que le S/C laisse suffisamment de marge, une réduction de prime étant susceptible de les rendre plus loyaux.
— Pour faire passer les profils de ***Peu rentable mais durée élevée*** à ***Plus rentable***, il faut faire en sorte qu'ils aient une charge des sinistres plus faible ou essayer de faire un meilleur calcul de la prime (meilleur segmentation des risques, utilisation de zoniers...).
— Il ne faut songer à faire passer les profils ***Moins rentable*** vers la partie ***Plus rentable*** que si ces profils se trouvent aussi dans la partie ***Plus rentable***. Cela montrerait qu'il faudrait essayer d'augmenter les durées de vie de ces profils afin de les faire passer dans la partie ***Plus rentable*** (puisque c'est la durée de vie qui justifie que ces profils sont dans une partie plutôt que l'autre).

Nous allons donc scinder l'analyse par parcours puisque le seuil de la durée de vie choisi dépend du parcours et que les profils qu'ils présentent peuvent différer.

Parcours ATA
La figure 4.6 suivante constitue un nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours ATA. Les quatres parties évoquées précédemment sont coloriées suivant la couleur correspondante. Le seuil de la durée de vie est de 2 ans.

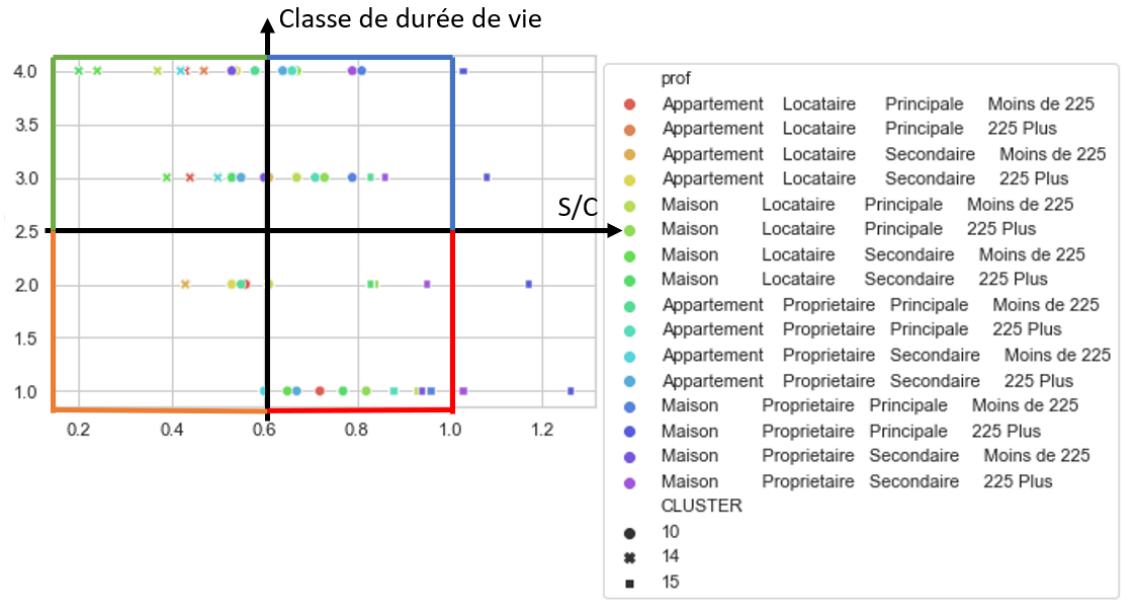


FIGURE 4.6 – Nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours ATA

Cette figure a permis de constater que :
— Les profils les plus rentables (partie verte) sont essentiellement constitués de locataires de maison et d'appartement.
— Les profils moins rentables (partie rouge) sont principalement des locataires de maison qui durent moins de deux ans. Nous retrouvons ces mêmes profils dans la partie ***Plus rentable*** avec une durée de vie plus élevée. Nous pourrions donc essayer d'augmenter leur durée de vie pour les transformer en profil plus rentables.
— Les profils qui durent longtemps mais qui sont peu rentables (partie bleu) sont constitués de propriétaires d'appartement et de maison.
— Les profils rentables mais qui ne durent pas plus de 2 ans (partie orange) sont principalement des locataires d'appartement. Ils peuvent continuer d'être reçu en affaire nouvelle dans ce parcours en essayant de trouver une stratégie pour les faire durer plus longtemps.
— Les profils pas rentables (S/C > 1) sont constitués de propriétaires de maison.

C'est dans ce parcours que se trouvent le plus de propriétaires qui s'avèrent être peu voire pas rentables. Cependant, il peut être bénéfique de continuer à les avoir en affaires nouvelles chez les agents car leur marge en euros peut être plus importante que celle d'un contrat avec un meilleur S/C (lorsque leur prime est plus élevée) et les agents peuvent multi-équiper leurs clients.

Parcours MTM

La figure 4.7 suivante constitue un nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours MTM. Les quatres parties évoquées précédemment sont coloriées suivant la couleur correspondante. Le seuil de la durée de vie est de 3 ans.

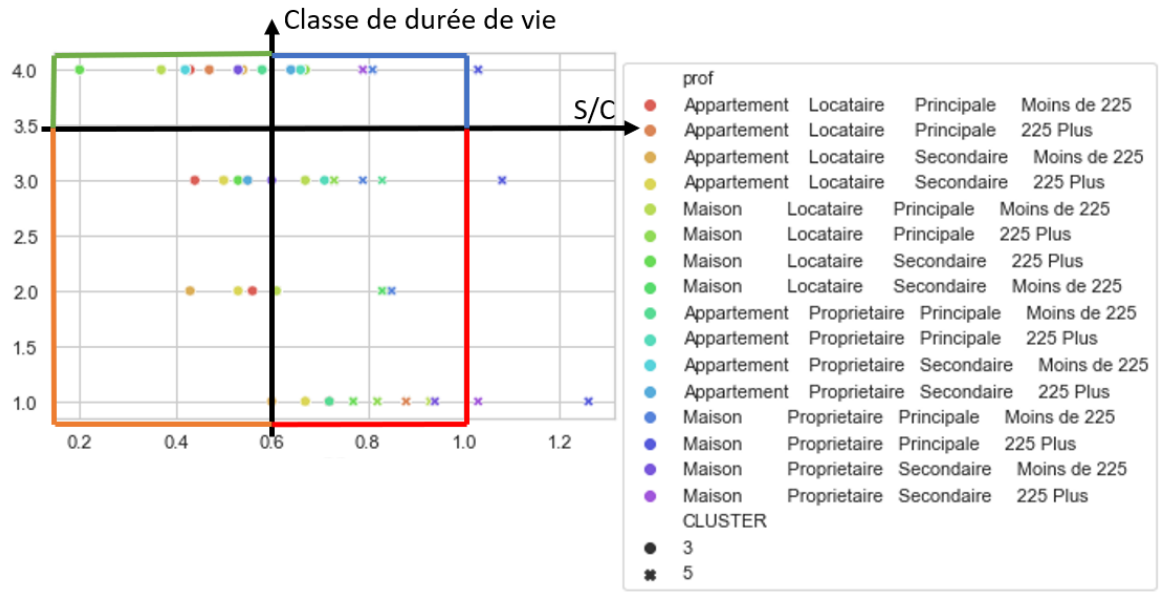


FIGURE 4.7 – Nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours MTM

Nous remarquons que dans ce parcours :
— presque tous les profils peuvent être parmi les plus rentables (partie en vert) s'ils durent plus de 3 ans.
— les profils moins rentables (partie en rouge) sont principalement des locataires de maison et des propriétaires d'appartement qui durent moins de trois ans. Ces mêmes profils peuvent aussi être retrouvés dans la partie ***Plus rentable*** avec une durée de vie plus élevée. Nous pourrions donc essayer de trouver une stratégie afin d'augmenter leur durée de vie pour les transformer en profil plus rentables.
— les profils rentables mais qui durent moins de deux ans (partie orange) sont constitués de locataires d'appartement. Nous pourrions les garder et essayer de trouver une stratégie pour les faire durer plus longtemps.
— les profils pas rentables sont des propriétaires de maison (notamment ceux avec plus de 225 euros de franchise quelle que soit leur durée de vie).

Parcours WDA

La figure 4.8 suivante constitue un nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours WDA. Les quatres parties évoquées précédemment sont coloriées suivant la couleur correspondante. Le seuil de la durée de vie est de 2 ans. Nous remarquons que :
— presque tous les profils peuvent être parmi les plus rentables (partie en vert) s'ils durent plus de 3 ans.
— les profils se trouvant dans la partie ***Moins rentable*** (partie en rouge) sont aussi présents dans la partie ***Plus rentable*** (partie en vert). Ce sont des locataires d'appartement et de maison. En tentant d'augmenter les durées de vie de ces profils, ils pourraient donc devenir plus rentables. Cependant, il n'est pas aussi simple d'y arriver. Vu que la rentabilité de ces profils dépend de leur durée de vie, baisser la prime pourrait ne pas être une bonne solution. Il faudrait trouver un autre moyen de les fidéliser comme essayer de les multi-équiper par exemple, ce qui pourrait être difficile. Les parcours Web attirent un nombre considérable de jeunes qui sont moins susceptibles d'être transformés en

clients multi-détenteurs puisqu'ils n'ont pas forcément besoin d'assurance vie et qu'il y a un nombre conséquents de jeunes qui n'ont pas de voiture ou qui utilise la même assurance automobile que leurs parents.
— les profils qui durent longtemps (plus de 2 ans) mais qui sont peu rentables (partie en bleu) sont des propriétaires d'appartement et des locataires de maison en résidence principale. Pour les garder dans ces parcours, il faudrait essayer de les faire passer dans la catégorie ***Plus rentable*** en essayant d'améliorer le calcul de la prime par exemple ou de baisser les réductions (s'il y en a). Cependant la prime avantageuse représente l'un des élément qui attire le plus les clients dans les parcours digitaux, augmenter les primes pourrait faire «fuir» ces profils.
— les profils pas rentables (S/C > 1) sont toujours des propriétaires de maison.

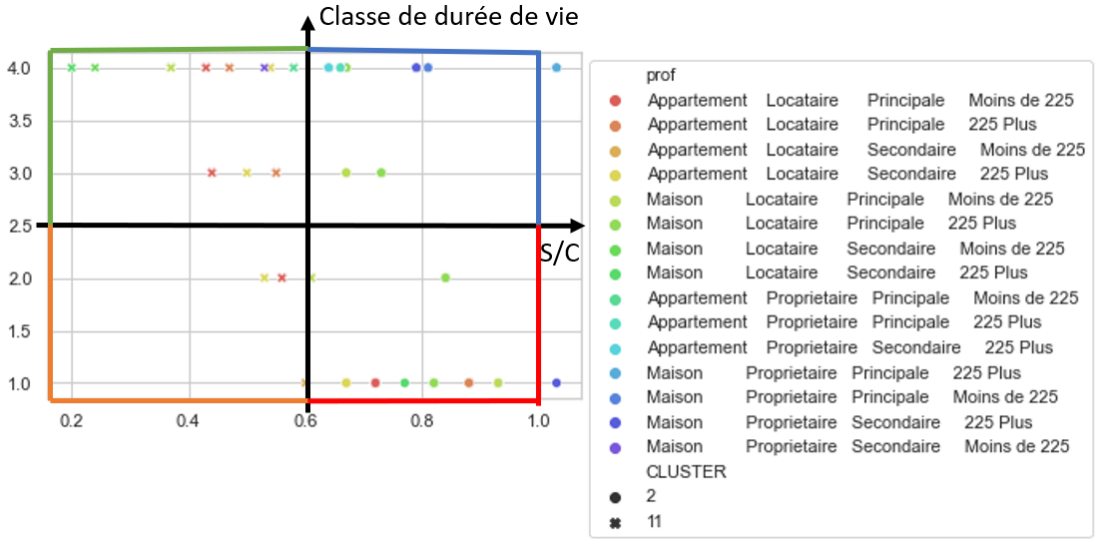


FIGURE 4.8 – Nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours WDA

Les figures des parcours restants sont mis en Annexe B.6.

4.3.2 Préconisation : quel profil faut-il continuer à accueillir dans chaque parcours ?

Suite à notre analyse, nous avons une idée sur les profils rentables et peu rentables se trouvant dans chaque parcours. Cela a alors permis de connaître les profils qu'il faudrait éviter dans certains parcours.

Nous avons conclu que nous pourrions laisser les **locataires d'appartement qui durent plus d'une année** utiliser n'importe quel parcours puisqu'ils sont rentables dans tous ces parcours. Pour ceux qui ont une durée de vie inférieure à 2 ou 3 ans (suivant le parcours), nous pouvons essayer de les rendre plus rentables en augmentant leur durée de vie (afin d'amortir le coût d'acquisition).

Cependant, il faut réfléchir plus profondément pour savoir quel parcours il vaudrait mieux laisser les autres profils utiliser.
— Les locataires de maison pourraient continuer d'utiliser les parcours agences (ATA et ADA) mais aussi dans le parcours plateforme MTM. Pour que ces profils soient très rentables il faut qu'il durent

plus de 3 ans. Nous avons vu dans la partie 4.1 que les locataires qui durent plus de 3 ans sont en général adultes avec plusieurs contrats ou seniors. En regardant les compositions des parcours ATA, ADA et MTM (vu dans la partie analyse descriptive 1.3.3), nous remarquons que ce sont les parcours qui ont le plus de clients ayant plusieurs contrats chez Allianz, ce qui montre qu'ils ont une bonne capacité à multi-équiper. Nous pouvons donc laisser les locataires de maison dans ces parcours.
Cependant, si les locataires de maison ont de forte chance de durer 3 ans (ce qui est le cas des locataires ayant déjà plusieurs contrats chez Allianz par exemple), nous pouvons les laisser utiliser n'importe quel parcours.
— Les propriétaires d'appartement (qui ne sont pas très rentables) pourraient être orientés dans les parcours agence (ATA et ADA). Ces profils durent longtemps dans le portefeuille mais sont peu rentables en S/C. Pour augmenter leur rentabilité, il faudrait adopter une stratégie afin de diminuer la charge totale des sinistres par exemple ou alors améliorer le calcul de la prime. Dans ce cas, c'est dans les parcours agence que la stratégie pourrait avoir plus de chance de réussir puisqu'augmenter les primes pourraient empêcher ces profils d'utiliser le canal direct vu que le tarif avantageux proposé par ce canal est un de ses principaux attraits.
— Les propriétaires de maison restent peu rentables voire pas rentables même après 4 ans. Quand ils sont peu rentables, ils peuvent quand même permettre d'avoir une marge supérieure à celles des profils plus rentables [8] puisque la prime est plus élevée. Quand ils ne sont pas rentables, il pourrait quand même être bénéfique de les garder dans les parcours agence (ATA et ADA). Une stratégie visant à améliorer leur S/C par la sensibilisation ou l'aide à la prévention de certains sinistres ou même un meilleur calcul des primes pourrait être établie.

En outre, les nuages de points (les points ayant une forme différente suivant le *cluster* auquel ils appartiennent) ont permis de voir que les *clusters* peuvent avoir des profils assez différents. Le *cluster* 11, étant celui du parcours Web WDA qui regroupe les profils les plus rentables, contient des locataires d'appartement ayant duré plus d'une année et même des locataires de maison.

4.4 Optimisation de la rentabilité sous contraintes

Nous retrouvons dans chaque parcours, un certain nombre de profils. Comme nous l'avons déjà vu, certains sont rentables, voire très rentables tandis que d'autres sont moins rentables, voire pas rentables du tout. Pour optimiser la rentabilité sur quatre ans, nous souhaitons qu'il n'y ait que les profils les plus rentables possibles dans chaque parcours **suivant certaines contraintes**. Cela pourrait permettre d'identifier les parcours qui pourraient éventuellement être éliminés. S'il y a par exemple des parcours n'attirant que des profils peu rentables par rapport aux autres parcours, il n'y a pas vraiment intérêt à garder un tel parcours sauf si c'est pour satisfaire certaines contraintes. Par exemple, vouloir ne pas dépasser un certain montant pour les frais annuels d'administration tout en atteignant un certain nombre de contrats en affaires nouvelles peut pousser à accepter des profils pas très rentables dans certains parcours qui ont des frais annuels d'administration faibles par rapport aux autres parcours.
Les contraintes de la sélection des profils sont donc liées :

— au **coût d'acquisition client** ($CoutAq$). Le parcours traditionnel a l'un des coûts d'acquisition client les plus bas, ce qui pourrait pousser à orienter les profils vers ce parcours plutôt qu'un autre afin d'économiser en coût d'acquisition.

— au **coût annuel d'administration** ($CoutAd$). Le coût d'administration est un ensemble de coûts récurrents qui se décompose en trois parties : la première partie est commune à tous les

8. comme les locataires d'appartement

parcours (elle représente les frais fixes d'Allianz France attribuable à la gestion des contrats, la deuxième partie est spécifique aux parcours gérés en plateforme (elle représente le coût des appels pour la gestion et leurs couts indirects) et la troisième partie est spécifique aux parcours gérés en agence (elle représente les commissions des agents). Le parcours traditionnel a cette fois-ci le coût annuel d'administration le plus élevé à cause des commissions. Cela pourrait nous pousser à orienter les profils vers des parcours autres que le parcours traditionel afin d'économiser en coût annuel d'administration.

— à **la somme totale de la prime émise** (*Primes*). En choisissant les profils, il faut aussi tenir en compte le chiffre d'affaires qu'ils vont rapporter et donc de la somme totale des primes émises par les profils choisis. Si nous voulons chaque année augmenter d'1% le chiffre d'affaires par exemple, il faudrait aussi augmenter d'un certain pourcentage la somme totale de primes émises. Certains profils pourraient être rentables mais avec une prime moyenne pas très élevée, ce qui pourrait nous pousser à accepter d'avoir des profils moins rentables mais ayant une prime moyenne plus élevée afin d'atteindre le chiffre d'affaires souhaité.

— au **nombre de contrats** (*NbContrats*). Il faudrait aussi avoir chaque année un certain nombre d'affaires nouvelles.

Nous voulons alors déterminer le nombre de contrats $(x_1, x_2 \ldots x_{16})$ à prendre dans chaque *cluster* $i \in \mathbb{N},\ 1 \leq i \leq 16$ afin d'optimiser la rentabilité en minimisant le S/C sur 4 ans comme suit

$$min_{x_1,x_2 \ldots x_{16}} f(x_1, x_2, x_3 \ldots x_{16}), \tag{4.3}$$

avec $f(x_1, x_2, x_3 \ldots x_{16}) = \dfrac{\sum_{i=1}^{K=16} x_i \times S_i}{\sum_{i=1}^{K=16} x_i \times C_i}$, le ratio sinistres sur primes à 4 ans,

x_i : le nombre de contrats du *cluster* i, $x_i \in \mathbb{N}$,

S_i : la charge moyenne des sinistres des profils du *cluster* i observée pendant 4 ans,

C_i : la prime moyenne des profils du *cluster* i observée pendant 4 ans,

sous les contraintes
$$\begin{cases} \sum_{i=1}^{K=16} x_i \times CoutAq_i = N1 \text{ euros, avec } CoutAq_i : \text{coût d'acquisition du } cluster\ i, \\ \sum_{i=1}^{K=16} x_i \times CoutAd_i <= N2 \text{ euros, avec } CoutAd_i : \text{coût d'administration du } cluster\ i, \\ \sum_{i=1}^{K=16} x_i \times Primes_i >= N3 \text{ euros, avec } Primes_i : \text{prime moyenne émise du } cluster\ i, \\ \sum_{i=1}^{K=16} x_i >= N4 \text{ contrats.} \end{cases}$$

Nous remarquons que les contraintes sont linéaires mais ce n'est pas le cas de la fonction objective f. f est une fonction continue, différentiable en tout point $x_i \in \mathbb{N}$ mais elle n'est pas convexe (voir Annexe A.9.1).
Cela entraîne qu'il est fort possible de tomber sur un minimum local mais pas forcément sur un minimum global en essayant de résoudre le problème d'optimisation.

4.4.1 Algorithmes de résolution

Vu que la fonction n'est pas convexe, des algorithmes permettant de trouver un minimum global sont utilisés sur python. Ils sont testés à l'aide du package *scipy.optimize* (SciPy, 2008-2022).

Basin-hopping

Le saut de bassin ou *basin-hopping* en anglais est une méthode d'optimisation globale décrite par Wales et Doye (1997). C'est une méthode en deux phases qui combine un algorithme pas à pas global avec une minimisation locale à chaque étape. L'optimisation locale fait référence à des algorithmes d'optimisation destinés à localiser un optimum pour une fonction objective ou à opérer dans une région où l'on pense qu'un optimum est présent alors que les algorithmes d'optimisation globale sont conçus pour localiser l'optimum global singulier parmi éventuellement plusieurs optimaux locaux (non globaux). Le saut de bassin est particulièrement utile pour l'optimisation globale dans des paysages de très haute dimension, comme la recherche de la structure d'énergie minimale pour les molécules.

L'algorithme du *basin-hopping*[9] consiste en un cycle de deux étapes, une **perturbation** des bonnes solutions candidates et l'application d'une **recherche locale à la solution perturbée**.
La **perturbation** permet à l'algorithme de recherche de sauter vers de nouvelles régions de l'espace de recherche et éventuellement de localiser un nouveau bassin conduisant à des optima différents.
La **recherche locale** permet à l'algorithme de parcourir le nouveau bassin vers les optima. Les nouveaux optima peuvent être conservés comme base pour de nouvelles perturbations aléatoires, sinon, ils sont rejetés. La décision de conserver la nouvelle solution est gérée par une fonction de décision stochastique avec une variable `temperature`. La température est ajustée en fonction du nombre d'itérations de l'algorithme. Cela permet de faciliter l'acceptation de solutions arbitraires au début de l'exécution (lorsque la température est élevée) et d'avoir une politique plus stricte consistant à n'accepter que des solutions de qualité améliorée plus tard dans la recherche (lorsque la température sera basse). L'algorithme s'exécute ainsi pour un nombre particulier d'itérations ou d'évaluations de fonctions.
De cette façon, il ressemble beaucoup à une recherche locale itérée avec différents points de départ (points perturbés).

Evolution Différentielle

L'algorithme d'évolution différentielle[10] (ED) ou *Differential Evolution* (DE) de Storn et Price (1997) est un **algorithme évolutionnaire**[11] basé sur la population et capable de trouver le minimum global de fonctions multivariées non différentiables et non linéaires. Elle est donc de nature stochastique c'est-à-dire qu'elle n'utilise pas de méthodes de gradient pour trouver le minimum et peut rechercher de vastes zones d'espace candidat mais elle nécessite souvent un plus grand nombre d'évaluations de fonctions que les méthodes basées sur le gradient.

Afin de trouver une solution optimale, un algorithme évolutionnaire commence par une population de NP individus qui constitue la **première génération**. L'adaption de ces individus au problème d'optimisation est **évaluée**. Ceux qui aboutissent à la solution la plus proche de la valeur cible sont **sélectionnés** pour créer une deuxième population, la «progéniture» qui constitue la **génération**

9. voir Annexe A.9.2
10. voir Annexe A.9.2
11. il s'inspire du mécanisme d'évolution d'une population dans son environnement. Il a connu énormément de succès depuis son apparition et fut initialement créé pour résoudre des problèmes continus. D'un côté, des opérateurs de variation apportent de la diversité à la population afin de favoriser l'exploration de l'espace de recherche. De l'autre, des opérateurs de sélection et de remplacement intensifient la recherche dans le voisinage d'une solution.

suivante. Le processus d'évolution d'une génération suit un cycle simple permettant d'améliorer séquentiellement chacun des NP individus. Ainsi, à tour de rôle chaque individu de la génération est appelé à être le vecteur cible ou solution de base.
La première étape de sélection permet de déterminer les individus qui participeront à la reproduction. Ces individus sont appelés « parents » puisqu'ils se **reproduisent** (ils sont **croisés ou recombinés**) pour donner un ensemble d'« enfants » partageant une partie des caractéristiques de leurs ascendants. Les nouveaux individus sont alors évalués [12]. Enfin, un nombre d'individus déterminé parmi l'ensemble parents + enfants est **sélectionné** afin de former la génération suivante. La «progéniture» est donc une « **mutation** » de ce meilleur ensemble de valeurs d'entrée de la première population. La deuxième population est ensuite, à son tour, évaluée afin de pouvoir créer la troisième population. Le processus est répété jusqu'à ce qu'un critère d'arrêt soit satisfait aboutissant ainsi sur la solution optimale.

Basé sur le concept de vecteur de différence [13], l'algorithme de l'ED suit ce processus en génèrant de nouveaux individus grâce à des opérations géométriques.
Par exemple, soit un vecteur cible $\mathbf{x}_i$ et trois individus $\mathbf{x}_{r1}$, $\mathbf{x}_{r2}$, $\mathbf{x}_{r3}$ générés au hasard à partir de la population actuelle tels qu'ils soient distincts les uns des autres et de l'individu x_i, i.e. $r_1 \neq r_2 \neq r_3 \neq i$ la mutation permet de former le vecteur donneur à l'aide de la formule

$$\mathbf{v}_i = \mathbf{x}_{r_1} + F\left(\mathbf{x}_{r_2} - \mathbf{x}_{r_3}\right), \tag{4.4}$$

avec F **le facteur de mutation** défini dans la plage [0,2].
Tous les éléments d'une solution de base ne sont pas mutés. Cette mutation est contrôlée via un hyperparamètre de recombinaison appelé **probabilité de croisement** et noté CR. Il est souvent défini sur une valeur élevée telle que 80%, ce qui signifie que la **plupart des variables d'une solution de base, mais pas toutes, sont remplacées**. Ainsi le vecteur d'essai $\mathbf{u}_i$ est développé à partir des éléments du vecteur cible $\mathbf{x}_i$ et des éléments du vecteur donneur $\mathbf{v}_i$ avec la probabilité CR.
La décision de conserver ou de remplacer une valeur dans une solution de base est déterminée séparément pour chaque position en **échantillonnant une distribution de probabilité telle qu'une distribution binomiale ou exponentielle**.
Les solutions de base sont remplacées par le vecteur d'essai si celui-ci a une meilleure évaluation objective de la fonction.

L'évolution différentielle a donc trois paramètres de contrôle qui sont : la **taille de la population** NP, où $NP \geq 4$, le **facteur de mutation** [14] $F \in [0, 2]$ et la **probabilité de croisement** [15] $CR \in [0, 1]$. Elle possède une nomenclature spécialisée décrivant la configuration adoptée sous la forme de **DE/x/y/z**, où x fait référence au mode de sélection du vecteur de base (solution de base) pour la mutation (*rand* si la sélection est purement aléatoire ou alors *best*, *rand-to-best* pour favoriser le meilleur vecteur). y représente le nombre de vecteurs de différence utilisés dans la perturbation de x. Enfin, z définit la distribution de probabilité pour déterminer si chaque solution est conservée ou remplacée dans la population, comme *bin* pour binôme ou *exp* pour exponentiel. Les configurations DE/rand/1/bin et DE/best/2/bin sont des exemples de configurations populaires car elles fonctionnent bien pour de nombreuses fonctions objectives.

4.4.2 Solution du problème d'optimisation

Les valeurs des coûts, des primes et des S/C ont été modifiées par soucis de confidentialité.

12. Leur valeur est mise à jour en faisant appel à la fonction objectif
13. Un vecteur de différence est la différence entre deux membres de la population choisis au hasard bien que distincts.
14. facteur de mutation ou poids différentiel ou facteur d'échelle
15. ou paramètre de contrôle

Supposons qu'à l'année N-1, 9 000 000 € de coût d'acquisition total avaient été utilisés afin d'obtenir au total 225 000 contrats dans les 8 parcours de la stratégie multi-accès. Cela a résulté sur une somme totale des primes émises de 53 200 000€ et sur une dépense de 13 000 000 € en frais annuel d'administration.

A l'année N, afin d'optimiser la rentabilité sur quatre ans, les contraintes sont alors :

— **cout d'acquisition client total**= 9 000 000 € (coût d'acquisition de l'année N-1).

— **cout annuel d'administration** $\leq$ 13 000 000 € (coût annuel d'administration de l'année N-1).

— **somme totale de la prime émise** $\geq$ 54 800 000 € (augmentation de 3% par rapport à l'année N-1).

— **nombre de contrats** $\geq$ 225 000 contrats (même nombre de contrats que l'année N-1).

Nous souhaitons donc savoir quels profils il faudrait avoir en affaire nouvelle à l'année N afin de minimiser le S/C sur 4 ans tout en :

— utilisant le même coût d'acquisition que l'année N-1,

— ne dépassant pas les frais annuels d'administration de l'année N-1,

— augmentant d'au moins de 3% la somme totale des primes émises par rapport à l'année N-1 (afin d'augmenter le chiffre d'affaires),

— ayant au moins le même nombre de contrats que l'année N-1.

Le tableau 4.4 suivant présente différentes caractéristiques des *clusters* (comme les valeurs des différents coûts fixes par parcours, les primes émises moyennes) permettant de définir les contraintes. Ainsi chaque contrat du *cluster* 9 (parcours ADA agence digitale) a un coût d'acquisition de 56€, un coût annuel d'administration de 70€ ainsi qu'une prime émise moyenne de 235€ et ces profils constituent 3,98% des contrats du portefeuille.

		Contraintes			Caractéristiques	
Parcours	Cluster	Cout d'acquisition	Prime émise moyenne	Frais annuel d'administration	S/C moyen	% portefeuille
ADA	9	56 €	235 €	70 €	53%	3,98%
ADA	4	56 €	317 €	70 €	92%	1,38%
ATA	14	18 €	293 €	76 €	43%	17,18%
ATA	10	18 €	364 €	76 €	68%	57,93%
ATA	15	18 €	473 €	76 €	101%	14,61%
MDA	12	167 €	186 €	55 €	50%	0,11%
MDA	6	167 €	244 €	55 €	86%	0,02%
MDM	16	197 €	156 €	38 €	52%	0,48%
MDM	7	197 €	212 €	38 €	69%	0,84%
MDM	8	197 €	260 €	38 €	102%	0,06%
MTM	3	142 €	202 €	38 €	55%	1,95%
MTM	5	142 €	277 €	38 €	93%	0,52%
WDA	11	1 €	128 €	55 €	49%	0,15%
WDA	2	1 €	236 €	55 €	82%	0,04%
WDM	13	31 €	126 €	38 €	49%	0,66%
WDM	1	31 €	246 €	38 €	86%	0,10%

TABLE 4.4 – Valeurs des variables permettant de définir les contraintes suivant les différents *clusters* de chaque parcours

Nous remarquons que les *clusters* les moins rentables ont les primes moyennes les plus élevées. Cela s'explique par le fait que les profils les plus rentables sont généralement des locataires d'appartement qui durent longtemps comme vu dans la partie précédente. Des profils peu rentables pourraient donc être choisis pour atteindre la somme des primes émises souhaitée.

Le ***Basin-Hopping*** et la méthode **évolution différentielle** vont être testés sur python. Les deux méthodes donnent la même solution et semblent avoir trouvé la solution globale. Les contraintes ont été tout juste satisfaites (les valeurs trouvées pour les contraintes sont égales à la valeur maximale pour le coût annuel d'administration et minimales pour les autres).
Le tableau 4.5 suivant présente l'optimum global trouvé pour le problème d'optimisation c'est-à-dire le nombre de contrats qu'il devrait y avoir dans chaque *cluster* pour minimiser le S/C sur 4 ans.

		Solution		Caractéristiques	
Parcours	Cluster	Nombre de contrats trouvés	Pourcentage	% portefeuille actuel	S/C moyen
ADA	9	0	0%	3,98%	0,53
ADA	4	0	0%	1,38%	0,92
ATA	14	54 355	24%	17,18%	0,43
ATA	10	62 750	28%	57,93%	0,68
ATA	15	0	0%	14,61%	1,01
MDA	12	0	0%	0,11%	0,50
MDA	6	0	0%	0,02%	0,86
MDM	16	0	0%	0,48%	0,52
MDM	7	0	0%	0,84%	0,69
MDM	8	0	0%	0,06%	1,02
MTM	3	31 958	14%	1,95%	0,55
MTM	5	0	0%	0,52%	0,93
WDA	11	0	0%	0,15%	0,49
WDA	2	0	0%	0,04%	0,82
WDM	13	75 936	34%	0,66%	0,49
WDM	1	0	0%	0,10%	0,86

TABLE 4.5 – Solution trouvée pour le problème d'optimisation c'est-à-dire nombre de contrats qu'il devrait y avoir dans chaque cluster pour minimiser le S/C sur 4 ans

Ainsi, la résolution permet donc de déduire qu'en dépensant 9 000 000 € comme coût d'acquisition et 13 000 000 € comme coût annuel d'administration, nous obtiendrons 54 800 000 € comme somme totale de primes émises ainsi que 225 000 affaires nouvelles et une rentabilité sur 4 ans de 53,67% en choisissant :

— **Dans le parcours ATA (traditionnel)**
24% des contrats du *cluster* 14 (cluster le plus rentanle)
28% des contrats du *cluster* 10 (cluster moyennement rentable)

— **Dans le parcours MTM (plateforme traditionnel)**
14% des contrats dans le *cluster* 3 (cluster le plus rentable)

— **Dans parcours WDM (full Web)**
34% des contrats dans le *cluster* 13 (cluster le plus rentable)

Les profils choisis ne sont presque que les profils les plus rentables de chaque parcours sauf pour le parcours ATA où des profils moyennement rentables (*cluster* 10) sont choisis.
En comparant avec la rentabilité des contrats observés sur quatre ans dans le portefeuille, cette stratégie améliorerait la rentabilité sur 4 ans de presque 10%.

Le parcours ATA a été utilisé parce que son coût d'acquisition est faible et il présente les primes moyennes les plus élevées notamment pour les profils les plus rentables qui ont aussi le S/C moyen le plus faible. Son coût d'acquisition est le deuxième le plus faible après le parcours WDA.
Le parcours MTM a aussi été utilisé. Même si son coût d'acquisition est élevé, son coût annuel d'administration est le plus faible (avec le parcours plateforme MDM et le parcours Web WDM). Il peut donc être utilisé pour limiter les frais d'administration tout comme les parcours MDM et WDM. Cependant, le coût d'acquisition du parcours MDM est plus élevé et les primes moyennes plus faibles ce qui explique le choix du parcours MTM plutôt que du parcours MDM. Le parcours MTM a peut-être aussi été utilisé à la place du parcours Web WDM pour certains profils car les primes proposées au client ne sont pas aussi basses qu'avec le parcours Web WDM. En effet, pour des profils similaires, les prix proposés dans le digital sont souvent plus bas.
Le parcours WDM a quand même été utilisé pour 34% des contrats probablement grâce aux frais faibles qu'il présente. Le fait que la prime moyenne du *cluster* le plus rentable de ce parcours soit aussi faible par rapport aux autres parcours pourrait être la raison pour laquelle plus de contrats ne sont pas orientés vers ce parcours. La stratégie d'optimisation suggère donc qu'il faudrait considérablement développer le parcours WDM qui a débuté en 2019. Les profils du *cluster* 13 sont principalement des locataires d'appartement, il pourrait donc ne pas être très difficile de les attirer vers ce parcours.

Les parcours MDA et MDM n'ont pas été choisis car leurs frais sont plus élevés que ceux des autres parcours utilisés et en plus, les primes moyennes sont aussi plus faibles (sauf par rapport à ceux des parcours Web) et les S/C moyens de leurs *clusters* font aussi partie des plus élevés.

Nous remarquons aussi que si plus de contraintes ne sont pas ajoutées, la stratégie d'optimisation suggère que seuls 3 parcours pourraient être gardés en essayant de considérablement développer 2 d'entre eux (MTM et WDM [16]). Mais dans la réalité, il y a plus de contraintes. Lorsque le client assurant son logement habite loin de l'agence, il peut être orienté vers un parcours plateforme. Ce qui entraîne qu'un certains pourcentage des clients qui auraient pu être dans le parcours traditionnel ATA vont se retrouver dans le parcours plateforme MTM [17]. Cette contrainte pourrait être ajoutée après avoir estimer le pourcentage correspondant.
Aussi afin de se lancer dès maintenant dans la course des contrats venant du digital, une contrainte liée au nombre minimal de contrats à avoir sur les parcours d'origine digitale pourrait être ajoutée, ce qui pourrait amener à utiliser le parcours ADA.
Par ailleurs, en essayant de déterminer les profils à avoir en affaires nouvelles, nous n'avons pas pris en compte le comportement et les préférences du client. Nous avons supposé que si le profil est rencontré dans un parcours alors le parcours peut attirer le profil alors que certains profils ont des préférences. Par exemple, nous avons vu dans la partie 1.1.5 les raisons qui peuvent expliquer que des assurés préfèrent les agences au canal direct. Il pourrait donc bien être utopique que de penser pouvoir attirer 34% des clients dans le parcours Web WDM même en misant fortement sur le marketing parce qu'il y a encore un nombre considérable de prospects ou clients qui préfèrent les agences.

16. Voir la figure B.1 en Annexe pour visualiser le pourcentage d'affaires nouvelles souscrits en 2019 et 2020 dans chaque parcours. Rappelons que le parcours Web WDM a commencé en 2019.
17. contrats souscrits et gérés en plateforme suite à un devis en agence

4.5 Limites et voies d'amélioration

Après avoir modélisé les taux de résiliation, les durées de vie ont pu être estimées permettant par la suite de pouvoir mettre en place une stratégie d'optimisation. Nous allons dans cette partie voir les limites de la stratégie d'optimisation ainsi que des méthodes pour l'améliorer ce qui revient aussi à déterminer les limites et les voies d'amélioration des modèles de taux de résiliation et de l'estimation des durées de vie.

4.5.1 Voies d'amélioration de la modélisation des taux de résiliation

Nous avons vu que les performances des modèles de taux de résiliation sont juste acceptables. Cela pourrait être dû au déséquilibre des classes mais aussi aux variables utilisées qui n'arrivent pas à expliquer à elles seules les résiliations. Ainsi, afin d'obtenir de meilleures performances, nous pourrions faire les améliorations suivantes.

Tester d'autres méthodes permettant de lutter contre le déséquilibre des classses
Les performances obtenues résultent en partie du déséquilibre des deux classes de prédiction. Nous avons testé certaines méthodes de ré-échantillonnage (SMOTE et *Random Under Sampling*) mais ils n'ont pas permis d'améliorer les performances. Nous pourrions tester d'autres méthodes permettant d'équilibrer les données comme la pénalisation ou d'autres méthodes de ré-échantillonnage.

Introduire de nouvelles variables explicatives
Un autre élément permettant d'expliquer les performances obtenues est le fait que seules les variables connues à l'affaire nouvelle sont utilisées et elles semblent ne pas être en mesure d'expliquer à elles seules l'acte de résiliation. En effet, si nous avions une variable donnant la majoration par exemple, nous pourrions mieux identifier les assurés qui résilient à cause de la majoration. Le problème est bien évidemment que la majoration à l'affaire nouvelle est inconnue, c'est pourquoi elle n'est pas incluse parmi les variables explicatives. Il y a cependant des variables déterminées à l'affaire nouvelle, qui aurait pu améliorer les performances mais que nous n'avons pas pu ajouter parce qu'elles n'étaient pas disponibles.

1. ETP (Ecart Tarif Portefeuille) : il représente l'écart entre le tarif actuel payé par l'assuré et le tarif portefeuille. Si l'assuré paie moins que le tarif portefeuille, il est susceptible d'être majoré à l'échéance. C'est donc une variable qui pourrait permettre de mieux détecter les résiliations dûes à la majoration.
2. Indice de compétitivité : il permet de voir le positionnement d'Allianz par rapport à ses concurrents pour un profil donné. Il aurait pu permettre de mieux identifier les clients qui résilient dans le but de bénéficier d'un tarif plus avantageux chez un concurrent.
3. Fréquence des sinistres prédites : un contrat qui risque fortement d'avoir des sinistres pourrait être résilié à cause de la majoration ou alors être résilié par l'assureur.

Prédire les changements
Comme nous gardons la vision des variables à l'affaire nouvelle et que nous ne prenons pas en compte les avenants, les performances se dégradent dans le temps puisque plus le temps passe, plus certaines variables explicatives sont susceptibles de changer de valeur comme le niveau du capital. Pour empêcher cette dégradation, nous aurions pu construire un modèle permettant de déterminer la probabilité d'avoir tel ou tel avenant comme la probabilité de changer de niveau de capital, de garantie, de déménager, de ne plus être étudiant, etc. Nous aurions aussi pu prendre en compte les changements de primes dûes à la majoration (que pourrait subir un contrat) en utilisant la fréquence des sinistres prédites et l'ETP pour estimer les éventuelles majorations.

Utiliser des données externes
Utiliser des données externes notamment les données socio-démographiques pourrait aussi être utile. Grâce aux données de l'INSEE, il est possible d'avoir, par exemple, la probabilité de déménager suivant l'iris.

4.5.2 Voies d'amélioration de la modélisation de la durée de vie

Notre modèle de durée de vie se limite à 4 ans et ne résulte que sur 5 valeurs possibles : 0,506 an ; 1,418 ans ; 2,405 ans ; 3,340 ans et 4 ans. Ainsi, les durées ne sont pas du tout précises. Cela est dû au fait que la période de mosélisation que nous utilisons est beaucoup trop grande (une année). L'idéal pour modéliser les durées de vie serait de pouvoir modéliser les taux de résiliation par mois mais le problème est que les taux de résiliation mensuels sont beaucoup trop faibles. Nous avons vu que même en utilisant une période de 6 mois, la faiblesse des taux ne permet pas d'obtenir un modèle suffisamment performant pour modéliser la durée de vie sur quatre ans.
Par ailleurs, les individus qui ont résilié la même année vont avoir la même durée de vie. Nous pourrions donc affiner la durée suivant plusieurs variables afin d'obtenir des durées moyennes plus précises.

Affinement des durées de vie suivant les variables les plus discriminantes
Au lieu de calculer les durées moyennes suivant l'année de résiliation seulement, nous pourrions utiliser aussi les variables les plus discriminantes pour la durée de vie. Ainsi un adulte locataire d'appartement ayant plus de deux contrats va avoir une durée de vie différente d'un locataire d'appartement jeune ayant un seul contrat même s'il a résilié la même année.

Utilisation d'un modèle de survie
Notre modélisation de durée de vie se limite à 4 ans. Nous avons choisi cette méthode principalement parce que certains parcours n'ont qu'une année d'ancienneté mais nous avons pu voir que les parcours semblent n'avoir que très peu d'influence sur la durée de vie. Ainsi, pour modéliser sur un plus grand horizon, nous pourrions utiliser les modèles de survie comme la régression de Cox. Le problème avec ce modèle est l'hypothèse de hazards proportionnels que certaines de nos variables ne vérifient pas. Il y a cependant d'autres modèles de survie qui ne font pas cette hypothèse comme le *Random Survival Forest*. Nous pourrions donc essayer de l'utiliser pour calculer les durées de vie.

4.5.3 Voies d'amélioration de l'optimisation de la rentabilité

Nous avons donc finalement entamé l'optimisation de la rentabilité dans la quatrième et dernière partie du mémoire. Pour ce faire, nous avons calculé les S/C suivant 4 variables (à 2 modalités chacune) et la durée de vie. Il aurait été donc mieux d'utiliser les S/C prédits (PSC) calculés en modélisant la prime pure et déterminés en divisant la prime pure obtenue par la prime observée du contrat. Nous aurions donc pu avoir un S/C par contrat ou profil. Cela aurait permis de faire une meilleure analyse des profils rentables et donc une meilleure préconisation (selon la rentabilité sur la durée de vie) sur quel profil il faudrait continuer à accepter et quel profil il faudrait éviter en affaire nouvelle dans chaque parcours. Notre analyse des profils rentables n'a permis de différencier les profils que suivant leur qualité juridique, leur type d'habitation, leur franchise, la nature de leur résidence et leur durée de vie. Utiliser les S/C prédits (PSC) aurait permis d'être beaucoup plus précis.

En faisant l'optimisation sous contraintes, nous avons utilisé des *clusters*. Les *clusters* contiennent donc des profils variés : des appartements, des maisons peuvent-être retrouvés dans un même *cluster* par exemple. Il y a en moyenne 41% de différence de S/C dans un *cluster*. Si nous disposions des S/C

prédits, nous aurions pu faire des *clusters* moins larges ou même raisonner directement par profil à condition que les performances de nos modèles de taux de résiliation aient été excellentes aussi.
De même, nous avons fait plusieurs hypothèses sur les coûts ce qui représentent une partie des limites de la stratégie d'optimisation. Nous avons gardé un coût fixe par parcours alors que celui-ci pourrait varier en fonction des profils. Les autres limites reposent sur les contraintes. Il pourrait par exemple, y avoir des parcours qui doivent absolument avoir des contrats. Comme nous savons que le canal internet pourrait se développer encore plus dans le futur, il est important de se lancer dans la course aux contrats sur internet dès maintenant. Cela nous contraindrait alors à avoir un nombre de contrats minimal pour les parcours digitaux en espérant que sur du plus long terme (plus de 4 ans) ces contrats deviennent rentables. Et nous sommes parfois obligés d'orienter en plateforme les clients habitant trop loin d'une agence.

Nous n'avons pas non plus pris en compte le comportement des clients par rapport au parcours. Ce qui pourrait nous amener à orienter les profils dans des parcours où il y a moins de chance qu'ils acceptent de l'utiliser. Ainsi, pour que le problème d'optimisation sous contraintes soit beaucoup plus réaliste, nous pourrions aussi envisager d'ajouter des probabilités afin de prendre en compte le comportement du client.

Conclusion du chapitre

Notre objectif est de modéliser la durée de vie des contrats d'assurance habitation afin d'optimiser la rentabilité de la stratégie multi-accès. Dans cette quatrième et dernière partie de notre mémoire, nous avons essayé d'optimiser la rentabilité de la stratégie multi-accès sur 4 ans.

Nous avons d'abord essayé de voir ce qui définit les profils les plus loyaux c'est-à-dire ceux qui restent le plus longtemps dans le portefeuille sur 4 ans en utilisant un arbre de décision. Ce sont majoritairement des propriétaires mais aussi des locataires seniors ou adultes multi-détenteurs.
Par la suite, nous avons essayé de déterminer les profils qu'il faudrait continuer à accueillir dans chaque parcours suivant leur S/C et leur durée de vie mais aussi les profils qu'il faudrait éviter. Ainsi, les locataires d'appartement susceptibles de durer plus d'une année pourraient continuer à utiliser n'importe quel parcours client. Cependant, il vaudrait mieux orienter la souscription des propriétaires en agence afin d'optimiser la rentabilité. Nous avons ensuite essayé de déterminer les profils qu'il faudrait idéalement avoir en affaires nouvelles à l'année N dans chaque parcours afin de minimiser le ratio sinistres sur primes sur 4 ans suivant plusieurs contraintes prenant en compte les frais annuels d'administration, le coût d'acquisition, la somme totale des primes émises et le nombre de clients. L'optimisation de la rentabilité a permis d'identifier les parcours qui pourraient être éliminés. Il y a cependant plusieurs limites à cette approche notamment sur les contraintes.

Nous avons terminé notre chapitre en déterminant les limites de la modélisation des durées de vie et de l'optimisation pour finalement donner des pistes d'amélioration de la modélisation des taux de résiliation, de calcul des durées de vie et d'optimisation de la rentabilité. Ajouter certaines variables comme la fréquence des sinistres prédite pour chaque contrat pourrait permettre d'améliorer les performances des modèles de taux de résiliation et donc du modèle de durée de vie. Quant à la stratégie d'optimisation, utiliser un S/C prédit pour chaque contrat aurait pu permettre de pouvoir choisir plus précisemment les profils au lieu de les choisir dans un groupe de profils ayant des S/C proches. De même, les coûts utilisés afin de pouvoir définir les contraintes sont des coûts fixes par parcours alors qu'ils pourraient dépendre du profil. Par ailleurs, il pourrait être interessant d'ajouter des probabilités afin de prendre en compte le comportement du client lors du problème d'optimisation sous contraintes.

Conclusion

Afin de tirer avantage du développement du canal direct tout en gardant les moyens de distribution traditionnelle, Allianz a décidé d'adopter la stratégie multi-accès offrant 8 parcours aux clients. Dans ce mémoire, nous avons essayé de modéliser les durées de vie afin d'optimiser la rentabilité de cette stratégie. Nous savons déjà que plus le contrat dure dans le portefeuille, plus le ratio sinistres à primes (S/C) baisse globalement et plus le coût d'acquisition client est amorti. Cela nous a alors amené à essayer de déterminer les profils qui durent le plus longtemps dans le portefeuille puisqu'ils sont susceptibles d'être plus rentables. Optimiser la stratégie reviendrait à aller plus loin. Il nous pousse à essayer de trouver pour chaque parcours, les profils à choisir afin d'encourager la souscription sur ces contrats et donc de maximiser la rentabilité c'est-à-dire minimiser le ratio sinistres sur primes sur 4 ans sous plusieurs contraintes prenant en compte les coûts. Et tout cela est possible si nous réussissons à calculer les durées de vie des profils de notre portefeuille.

Pour modéliser les durées de vie, nous avons utilisé une approche basée sur la modélisation des taux de résiliation par période d'une année principalement parce que certains parcours n'ont qu'une année d'ancienneté et que nous avons des pistes pour modéliser leur durée aussi longtemps que les autres parcours. Cette approche va donc permettre de voir chaque année l'impact des parcours sur la résiliation et plus globalement comment les variables (vues à l'affaire nouvelle) impactent la résiliation. Pour chaque période, 3 modèles ont été utilisés et suivant les valeurs des métriques, nous avons choisi le meilleur modèle qui s'avère être le XGBoost. Cependant, les performances des modèles ne sont pas très satisfaisantes et elles se dégradent au fil des périodes (les valeurs de certaines métriques comme la précision baissent) ce qui nous a poussé à arrêter la modélisation à 4 ans. Cela s'explique par le fait que les variables explicatives semblent ne pas être en mesure d'expliquer à elles seules les résiliations. Nous aurions donc pu ajouter des variables disponibles à l'affaire nouvelle comme l'ETP (Ecart Tarif Portefeuille) qui est égal à l'écart entre le tarif observé et le tarif portefeuille mais aussi l'indice de compétitivité et la fréquence des sinistres prédite. Comme la vision des variables à l'affaire nouvelle est gardée et que les avenants ne sont pas pris en compte, les performances se dégradent dans le temps. De même, chaque année, le taux de résiliation baisse, le déséquilibre entre les classes s'accroît, ce qui contribue à la dégradation des performances.

Les performances des modèles restent acceptables pour nous aider à atteindre notre objectif. L'importance des variables, le PDP (*Partial Dependence Plot*), l'ICE (*Individual Conditional Expectation*) et les SHAP *values* ont permis d'interpréter le modèle à boite noire choisi (XGBoost). Ceux-ci ont amené à la conclusion que **le parcours a très peu d'impact sur les résiliations**. En supposant que les autres caractéristiques du contrat restent les mêmes, le contrat a les mêmes chances d'être résilié quel que soit le parcours choisi. Il ne faut cependant pas oublier que le parcours traditionnel (ATA) représente à lui seul plus de 90% du portefeuille. Il n'y a pas encore suffisamment de données dans les autres parcours pour **totalement** confirmer cette théorie. Nous avons aussi pu voir que les propriétaires ont une durée de vie plus élevée que les locataires et les assurés ayant plusieurs contrats sont plus loyaux comme les seniors ont plus de chance de rester longtemps dans le portefeuille que

les adultes et les jeunes. Ainsi, une durée de vie plus élevée pour un parcours par rapport à un autre pourrait par exemple s'expliquer par le fait qu'il y ait plus de propriétaires dans ce parcours. A profil équivalent, les durées de vie sont stables dans le temps (hors période de Covid-19).
Pour déterminer les profils les plus loyaux, un arbre de décision a été utilisé afin de pouvoir facilement identifier ce qui les caractérise. **Les propriétaires ont, d'après l'arbre, 87% de chance de durer plus de 3 ans. Les locataires seniors ou locataires adultes ayant plusieurs contrats font aussi partie des profils les plus loyaux (durent plus de 3 ans)**. Vu que les parcours n'ont pas d'impact, nous n'avons plus besoin de déterminer les profils qui auraient duré plus longtemps s'ils étaient dans un autre parcours.

Nous avons donc finalement entamé l'optimisation de la rentabilité dans la quatrième et dernière partie du mémoire. Cependant, la maille utilisée pour l'estimation des S/C est assez large car la volatilité des S/C constitue le frein à l'obtention d'une maille plus fine. Elle n'est constituée que de 4 variables à 2 modalités. Il aurait donc été mieux d'utiliser les S/C prédits (calculés en modélisant la prime pure) qui aurait abouti sur un S/C pour chaque profil ou contrat. Vu la taille de la maille, les S/C pourraient ne pas être réprésentatifs. Nous avons alors regroupé les profils de S/C en 16 *clusters* c'est-à-dire groupes avec des S/C homogènes. Chaque parcours a alors 2 ou 3 *clusters*, un avec les profils les plus rentables et un autre pour les profils moins rentables. **Ainsi, il a été possible de déterminer les profils qu'il faudrait avoir en affaires nouvelles afin d'obtenir un S/C minimal sur 4 ans égal à 53,67% (10% de moins que le S/C sur 4 ans actuel) tout en respectant les contraintes comprenant le coût d'acquisition. Cela implique de développer certains parcours et d'en éliminer d'autres**.
Cependent, nous avons gardé des coûts fixes par parcours (pour définir les contraintes) alors qu'ils pourraient varier en fonction des profils, ce qui représentent une partie des limites de notre stratégie d'optimisation. Les autres limites reposent sur les contraintes. Il pourrait par exemple, y avoir des parcours qui doivent absolument avoir des contrats. Comme nous savons que le canal internet pourrait se développer encore plus dans le futur, il est important de se lancer dans la course aux contrats sur internet dès maintenant. Cela nous contraindrait alors à avoir un nombre de contrats minimal pour les parcours digitaux en espérant que sur du plus long terme (plus de 4 ans) ces contrats deviennent rentables. N'oublions pas que les compagnies d'assurance directes ont mis 10 années avant d'être enfin rentable.

En corrigeant les limites de notre stratégie d'optimisation et en exploitant les pistes pour améliorer le modèle de durée de vie et les modèles de taux de résiliation, nous pouvons donc établir une stratégie solide d'optimisation de la rentabilité à long terme permettant de connaître non seulement les profils qu'il faut cibler dans chaque parcours suivant plusieurs contraintes mais aussi les parcours qui pourraient être éliminés de la stratégie multi-accès. Par la suite, il pourrait être envisagé de trouver des stratégies pour améliorer la rentabilité des profils peu rentables que la stratégie d'optimisation aurait contraint à accepter dans certains parcours en partant de l'optimisation sans contrainte que nous avons essayé de faire. Celle-ci a permis d'avoir une idée sur comment améliorer la rentabilité de certains profils suivant le parcours.
Maintenant que nous sommes convaincus que le parcours n'a pas d'impact sur la durée de vie, nous pouvons essayer d'utiliser un modèle de durée comme le *random survival forest* au lieu de rester limiter à 4 ans. De même, pour que le problème d'optimisation sous contraintes soit beaucoup plus réaliste, nous pourrions envisager d'intégrer des probabilités afin de prendre en compte le comportement du client.

Bibliographie

Branco, P., Torgo, L. et Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)* 49.2, p. 1-50.

Breiman, L. (2001). Random Forests. *Machine Learning*, p. 5-32.

Caillol, T. et Giraud, C. (2017). Bancassureurs : les nouveaux champions de l'assurance ? url : https://www.insurancespeaker-wavestone.com/2017/01/bancassureurs-champions-assurance/ (visité le 19/10/2021).

Chabrier, B. (2020). Les comparateurs d'assurance à l'heure de nouveaux défis. url : https://www.argusdelassurance.com/courtiers/comparateurs/les-comparateurs-d-assurance-a-l-heure-de-nouveaux-defis.164946 (visité le 19/10/2021).

Chawla, N. V., Bowyer, K. W., Hall, L. O. et Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16, 321–357.

Chen, T. et Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

FFA (2019). Données clés 2019. url : https://www.ffa-assurance.fr/etudes-et-chiffres-cles/assurance-francaise-donnees-cles-par-annee (visité le 19/10/2021).

Friedman, J. (fév. 2002). Stochastic Gradient Boosting. *Computational Statistics Data Analysis* 38, p. 367-378.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics* 29, p. 1189-1232.

Goldstein, A., Kapelner, A., Bleich, J. et Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics* 24.1, p. 44-65.

h2o (2021). Algorithms. url : https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science.html (visité le 16/11/2021).

He, H. et Garcia, E. (2009). Learning from Imbalanced Data. *Knowledge and Data Engineering, IEEE Transactions on* 21.9, p. 54.

Institute Inc. SAS (1983). SAS Technical Report A-108, Cubic Clustering Criterion. Rapp. tech. SAS Institute Inc., p. 1-2-4-49.

Lambert, A. (1998). La distribution de l'assurance. *Assurons l'avenir de l'assurance*. Sous la dir. de Rapport d'information n° 45, S.

Les Echos (2010). 2010, année de l'émergence des comparateurs d'assurance. url : https://www.lesechos.fr/2010/08/2010-annee-de-lemergence-des-comparateurs-dassurance-429237 (visité le 19/10/2021).

Lundberg, S. M. et Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Physical Review* 10.

McCullagh, P. et Nelder, J. A. (1989). Generalized linear models (Second edition). London: Chapman & Hall, p. 500.

Molnar, C. (2021). Interpretable Machine Learning A Guide for Making Black Box Models Explainable. url : https://christophm.github.io/interpretable-ml-book/ (visité le 19/10/2021).

RDocumentation (2019). rpart: Recursive Partitioning and Regression Trees. URL : https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart (visité le 16/11/2021).

SciPy (2008-2022). Optimization and root finding (scipy.optimize). URL : https://docs.scipy.org/doc/scipy/reference/optimize.html (visité le 12/03/2022).

Storn, R. et Price, K. (jan. 1997). Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization* 11, p. 341-359.

Szumilas et Magdalena (2010). Explaining odds ratios. *Canadian Academy of Child and Adolescent Psychiatry* 19.3.

Van Rossum, G. et Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA : CreateSpace.

Waelbroeck, P., Levallois-Barth, C., Laurent, M. et Meseguer, I. (2019). Données personnelles et confiance : évolution des perceptions et usages post RGPD. *Institut Mines-Télécom* 12.

Wales, D. J. et Doye, J. P. K. (1997). Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *The Journal of Physical Chemistry A* 101.28, 5111–5116.

Weiss, G. M., McCarthy, K. et Zabar, B. (2007). Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? *DMIN*.

Zou, H. et Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67.2, p. 301-320.

Annexe A

Elements théoriques

A.1 V de Cramer

Le coefficient de corrélation vu sa définition, n'est utilisé que pour les variables numériques. Ainsi, pour déterminer la corrélation entre les variables qualitatives, nous allons utiliser le V de Cramer.
Le test V de Cramer permet de comparer le degré de dépendance entre les deux variables étudiées. Il est basé sur la statistique de test du χ^2 de Pearson.

Test d'indépendance du χ^2 de Pearson

Le test d'indépendance du χ^2 a été développé par Karl PEARSON (1857-1936). L'expression test du χ^2 fait principalement référence à 3 tests statistiques :
— le test d'homogénéité du χ^2 : permet de tester si des échantillons sont issus d'une même population.
— le test d'ajustement ou d'adéquation : permet de comparer globalement la distribution observée dans un échantillon statistique à une distribution théorique, celle du χ^2.
— Le test d'indépendance du χ^2 : permet de vérifier l'indépendance de deux variables dans une population donnée.

Le test d'indépendance du χ^2 est celui qui nous intéresse ici. Il permet de déterminer si deux variables qualitatives sont indépendantes ou non, c'est-à-dire si les réponses de l'une conditionnent les réponses de l'autre. Il ne permet toutefois pas de connaître le sens de la dépendance.

Le principe du χ^2 consiste à comparer les effectifs réels des croisements des modalités de deux variables qualitatives avec les effectifs théoriques obtenus quand les deux variables sont indépendantes. Soit X et Y deux variables aléatoires qualitatives qui prennent un nombre fini de valeurs qui est respectivement égal à I et J. Nous disposons d'un échantillonnage de N données.
Notons alors : $O_{i.}$, le nombre de données pour lesquelles $X = i$ et $O_{.j}$, le nombre de données pour lesquelles $Y = j$,
O_{ij} le nombre de données pour lesquelles X prend la valeur i et Y la valeur j et,
E_ij la valeur espérée de l'effectif O_{ij} sous l'hypothèse d'indépendance définie comme suit

$$E_{ij} = \frac{O_{i.}O_{.j}}{N}. \tag{A.1}$$

Le χ^2 a donc pour formule

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})}{E_{ij}}. \tag{A.2}$$

Le χ^2 suit asymptotiquement une loi du χ^2 à (I – 1)(J – 1) degrés de liberté.

Formule du V de Cramer

Soit X et Y deux variables qualitatives à, respectivement, I et J modalités et N le nombre d'observations comme mentionné plus haut. Le V de Cramer, basé sur la statistique de test du χ^2 de Pearson, a alors pour formule

$$V = \sqrt{\frac{\chi^2}{n \cdot min(I-1, J-1)}}. \tag{A.3}$$

Le terme $n \cdot min(I-1, J-1)$ qui est également noté χ^2_{max} représente la valeur maximale que peut prendre la statistique de test du χ^2 de Pearson.

Interprétation du V de Cramer

Le V de Cramer varie dans l'intervalle [0, 1].
— Plus sa valeur est élevée, plus la dépendance entre les deux variables est forte.
— Plus sa valeur est faible, plus les variables se rapprochent de l'indépendance.
— V=0 se rencontre dans le cas où les deux variables sont parfaitement indépendantes et V=1, dans le cas où les variables sont totalement dépendantes.
— le V ne dépend ni des effectifs ni des dimensions du tableau. Il peut donc être comparé d'un tableau à l'autre.

Le tableau A.1 ci-dessous présente l'interprétation du V de Cramer suivant sa valeur.

Valeur absolue de V	Interprétation
Entre 0 et 0.05	Absence de liaison
Entre 0.05 et 0.1	Très faible
Entre 0.1 et 0.2	Faible
Entre 0.2 et 0.4	Modérée
Entre 0.4 et 0.8	Forte
Entre 0.8 et 1	Colinéarité

TABLE A.1 – Tableau donnant l'interprétation du V de Cramer obtenu

A.2 Métrique ROC

La courbe ROC est une représentation graphique décrivant la relation entre la sensibilité [1] et la spécificité [2] du modèle pour toutes les valeurs seuils possibles s. En d'autres termes, il trace le taux de fausses alarmes par rapport au taux de réussite. L'abscisse représente le taux de vrais positifs (sensibilité) et l'ordonnée correspond au taux de faux positifs (1 - spécificité). En reprenant l'annotation

1. la sensibilité mesure la capacité du modèle à prédire un résultat positif lorsque le résultat est réellement positif.
2. la spécificité mesure la capacité du modèle à prédire un résultat négatif lorsque le résultat est réellement négatif.

de la partie 2.2.1, voici leur formule de calcul

$$\text{Spécificité}(s) = \frac{TN(s)}{FP(s) + TN(s)}.$$
$$\text{Sensibilité}(s) = \frac{TP(s)}{TP(s) + FN(s)}.$$
$$\text{Taux de TP(s)} = \frac{TP(s)}{TP(s) + FN(s)} = \text{Sensibilité}(s).$$
$$\text{Taux de FP(s)} = \frac{FP(s)}{FP(s) + TN(s)} = 1 - \text{Spécificité}(s).$$

La figure A.1 suivante présente des courbes ROC, un prédicteur excellent, moyen et sans intérêt peuvent être aperçus.

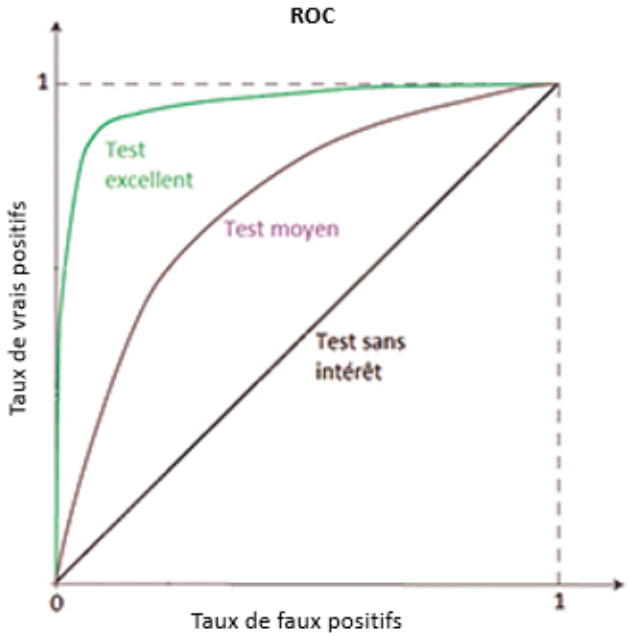


FIGURE A.1 – Exemple de courbes ROC

Nous remarquons donc que le coin supérieur gauche du graphique ROC est le point «idéal», celui qui donne un taux de faux positifs de 0 et un taux de vrais positifs de 1. Des valeurs plus petites sur l'axe des abscisses du graphique indiquent des faux positifs plus faibles et des vrais négatifs plus élevés. Des valeurs plus grandes sur l'axe des ordonnées du graphique indiquent des vrais positifs plus élevés et des faux négatifs plus faibles. La « pente » des courbes ROC est également importante, car elle est idéale pour maximiser le taux de vrais positifs tout en minimisant le taux de faux positifs. Les courbes ROC peuvent donc être aussi utilisées afin de trouver une valeur seuil, le choix de la valeur seuil dépendant de la manière dont le classificateur est destiné à être utilisé. Si notre but était de détecter coûte que coûte les résiliations afin d'établir une stratégie pour les retenir, nous aurions pu

choisir un seuil permettant de maximiser la sensibilité. Cependant, dans notre cas, l'objectif est de calculer des durées de vie moyennes. Maximiser la sensibilité pousserait le modèle à avoir un taux de faux positifs élevé et donc à prédire beaucoup de fausses résiliations. Cela pourrait amener à conclure qu'un contrat va chuter dès la première année alors qu'en réalité, il survit jusqu'à la dernière année. Bien que la courbe ROC soit un outil d'évaluation utile, il peut être difficile de comparer deux ou plusieurs classificateurs en fonction de leurs courbes. L'aire sous la courbe (AUC) peut alors être calculée pour donner un score unique pour un modèle de classificateur sur toutes les valeurs de seuil.

A.3 GLM

Soit $(y_1, y_2, \ldots, y_n)$ le vecteur d'observation obtenue avec la réalisation de la variable aléatoire $Y = (Y_1, Y_2, \ldots, Y_n)^t$ et $X_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^t$ le i-ème vecteur ligne des variables explicatives associées à l'observation i. X est donc une matrice de taille $n \times (p+1)$ dont les lignes sont les vecteurs lignes X_i^t et les variables correspondantes peuvent être quantitatives ou qualitatives.
Popularisé par McCullagh et Nelder (1989), le modèle linéaire généralisé permet de modéliser une relation non-linéaire entre la variable aléatoire $Y \in \mathbb{R}^n$ et les p variables explicatives X

$$g(E[Y|X]) = X\beta, \tag{A.4}$$

avec β le vecteur des $p+1$ paramètres.

Les Modèles Linéaires Généralisés sont caractérisés par trois composantes.

La composante aléatoire

Elle identifie la distribution de probabilités de la variable à expliquer. Supposons que l'échantillon statistique soit constitué de n variables aléatoires $Y_i;\ i = 1, ..., n$ indépendantes admettant des distributions issues d'une famille exponentielle. Cela signifie que les lois de ces variables sont dominées par une même mesure dite de référence et que la famille de leurs densités par rapport à cette mesure se met sous la forme

$$f_{Y\,i}(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - v(\theta_i)}{u(\phi)} + w(y_i, \phi)\right\}, \tag{A.5}$$

avec : $\theta_i \in \mathbb{R}$ = paramètre canonique ou de la moyenne, $\phi \in \mathbb{R}$ = paramètre de dispersion, u fonction définie sur $\mathbb{R}$ non nulle, v fonction définie sur $\mathbb{R}$ deux fois dérivable et w fonction définie sur $\mathbb{R}^2$.

La plupart des lois usuelles comportant un ou deux paramètres appartiennent donc à la famille exponentielle : gaussienne, gamma, Poisson, binomiale, etc. Nous retrouvons dans le tableau A.2 suivant les paramètres des principales lois de probabilité de la famille exponentielle.

Distribution de Y_i	θ_i	ϕ	$u_i(\phi)$	$v(\theta_i)$
$Normale(\mu_i, \sigma_i^2)$	μ_i	σ^2	ϕ	$\frac{\theta_i^2}{2}$
$Bernoulli(\mu_i)$	$\log(\frac{\mu_i}{1-\mu_i}))$	1	ϕ	$\log(1-\exp(\theta_i)$
$Binomiale\frac{1}{n_i}(n_i, \mu_i)$	$\log(\frac{\mu_i}{1-\mu_i})$	$\frac{1}{\mu_i}$	ϕ	$\log(1+\exp(\theta_i))$
$Poisson(\mu_i)$	$\log(\mu_i)$	1	ϕ	$\exp(\theta_i)$
$Gamma(\mu_i, \alpha)$	$\frac{-1}{\mu_i}$	$\frac{1}{\alpha}$	ϕ	$-\log(-\theta)$

Table A.2 – Tableau présentant les valeurs des différents paramètres définissant la famille exponentielle suivant des lois données

Pour la modélisation des taux de résiliation, la variable à expliquer (l'acte de résiliation) est binaire (elle prend les valeurs 0 ou 1). Elle suit donc une loi de Bernoulli (coloriée en bleu).

La composante déterministe
La composante déterministe du modèle aussi appelée prédicteur linéaire est le vecteur à n composantes défini comme suit

$$\eta = X\beta. \tag{A.6}$$

La fonction de lien
La troisième composante exprime une relation fonctionnelle entre la composante aléatoire et la composante déterministe. On pose

$$g(\mathbb{E}[Y \mid X]) = \eta = X\beta. \tag{A.7}$$

La fonction de lien g est supposée monotone et différentiable. La fonction de lien qui associe $\mathbb{E}[Y \mid X]$ au paramètre naturel θ peut être qualifiée de fonction de lien canonique. Dans ce cas,

$$g(E[Y|X]) = \theta = X\beta. \tag{A.8}$$

La variable aléatoire Y est distribuée selon une loi appartenant à une famille exponentielle alors $\mathbb{E}[Y] = u'(\theta)$. Ainsi en choisissant $g = (u')^{-1}$, la relation $\theta = X\beta$ est obtenue.

Estimation des coefficients β
Un GLM a pour but d'estimer les coefficients de régression $\beta_j, j = 0, ..., p$ en utilisant la méthode de maximum de vraisemblance. Si $y = (y_1, \ldots, y_n)$ est une réalisation de l'échantillon de n variables aléatoires indépendantes $Y_1, \ldots, Y_n$ dont les fonctions de densité f_{Y_i} sont issues de la famille exponentielle la vraisemblance en y s'écrit

$$L(y_1, ..., y_n; \theta, \phi) = \exp(\Sigma_{i=1}^{n} \frac{y_i\theta_i - v(\theta_i)}{u_i(\phi)} + w(y_i, \phi)). \tag{A.9}$$

L'expression de la log-vraisemblance est donc

$$log(L) = \Sigma_{i=1}^{n} \frac{y_i\theta_i - v(\theta_i)}{u_i(\phi)} + w(y_i, \phi). \tag{A.10}$$

Pour trouver β_j maximisant cette log-vraisemblance, il faut donc résoudre les équations de vraisemblance suivantes

$$\frac{\partial Log(L)}{\partial \beta_j} = 0, \forall j = 0, ..., p. \tag{A.11}$$

A.4 Critères de sélection des variables

Parmi les critères les plus utilisés pour la sélection des variables avec les méthodes *Wrapper-based* (sélection ascendante, descendante), il y a les critères suivantes.

A.4.1 *La P-value*

Pour déterminer si l'association entre la variable à expliquer et chaque variable du modèle est statistiquement significative, la p-value est souvent utilisée. Celle-ci permet d'évaluer l'hypothèse nulle et est définie comme suit

$$p = P(x|H_0). \tag{A.12}$$

L'hypothèse nulle H_0 : il n'y a pas d'association entre la variable donnée et la variable à expliquer.
L'hypothèse H_1 : il y a une association entre la variable donnée et la variable à expliquer.
Un résultat statistiquement significatif est un résultat qui serait improbable si l'hypothèse nulle H_0 était vérifiée. L'hypothèse nulle est rejetée lorsque la p-value est inférieure ou égale à un seuil noté α. Celui-ci est généralement égal à 0.05. Un seuil de 0,05 indique un risque de 5% de conclure à l'existence d'une association lorsqu'il n'y a pas d'association réelle.
En résumé :
$P - value \leq \alpha$: l'association entre la variable donnée et la variable explicative est statistiquement significative.
$P - value > \alpha$: l'association n'est pas statistiquement significative.
Si la variable n'est pas signifiative, elle peut être supprimer dans le cadre d'une sélection des variables.

A.4.2 L'*Akaike Information Criterion* (AIC)

L'AIC est un critère qui prend en compte non seulement la complexité du modèle (avec le nombre de variables retenues) mais aussi la qualité de l'ajustement à travers la fonction de vraisemblance. Il est donné par la formule suivante

$$AIC = 2\log(L) + 2p, \tag{A.13}$$

où $\log(L)$ constitue la log-vraisemblance maximisée et p représente le nombre de paramètres.

A.4.3 Le *Bayesian Information Criterion* (BIC)

L'AIC pourrait mener au choix de modèles avec un nombre considérable de variables explicatives, ce qui n'est pas le cas du BIC. En effet, dans la formule du BIC, le nombre de paramètres est multiplié par le logarithme du nombre d'observations. Cela permet d'appliquer une pénalité plus sévère permettant ainsi de pouvoir choisir un modèle avec moins de variables explicatives lorsque le BIC est le critère d'évaluation. Il est donc défini par la formule suivante

$$BIC = 2log(L) + plog(n). \tag{A.14}$$

Ces deux dernières critères permettent aussi de comparer des modèles entre eux. Un modèle pourrait être considéré comme étant meilleur s'il présente un AIC (ou BIC) plus petit par rapport aux autres modèles testés.

A.5 Arbre de décision

Un arbre de décision est une représentation simple pour classer des exemples. Il s'agit d'un modèle supervisé de *machine learning* où les données sont continuellement divisées en fonction d'un certain paramètre. L'arbre de décision se compose de :

- Nœuds : testent la valeur d'un certain attribut.
- Bords/Branches : correspondent au résultat d'un test et se connectent au prochain nœud ou feuille.

— Nœuds feuilles : nœuds terminaux qui prédisent le résultat (représentent les étiquettes de classe ou la distribution de classe).

Avantages
— Un arbre de décision est facile à interpréter et à visualiser.
— Il peut être utilisé avec des données qualitatives et quantitatives.
— Il peut facilement capturer des motifs non linéaires.
— Il nécessite moins de traitement de données de la part de l'utilisateur. Il n'est, par exemple, pas nécessaire de normaliser les données.
— Aucune hypothèse n'est faite sur la distribution de l'arbre de décision en raison de la nature non paramétrique de l'algorithme.

Inconvénients
— Il est sensible aux bruits. Il a donc tendance à sur-apprendre (*overfit*) avec les données bruitées.
— Une petite variation (ou variance) des données peut donner un arbre de décision différent. Cette instabilité peut être réduit par des algorithmes de *bagging* et de *boosting*. Ces algorithmes sont cependant beaucoup moins facile à interpréter.
— Les arbres de décision sont biaisés lorsque les données sont déséquilibrées.

A.6 Package *h2o* de python

h2o est une plate-forme Java de *machine learning* proposant des outils pour la manipulation et la préparation de données, des algorithmes de modélisation, supervisées, non-supervisées ou de réduction de dimensionnalité utilisées par des entreprises notamment des compagnies d'assurance. Il peut être utilisé avec python, R ou encore Java. La parallélisation de plusieurs algorithmes standards de *machine learning* tels que le *eXtreme Gradient Boosting Machine* ou encore le *Random Forest* est possible avec le package *h2o*, ce qui permet à l'exécution des algorithmes de prendre moins de temps. Un autre avantage de h2o est qu'il lui est possible d'augmenter la capacité et la puissance de calculs sur de larges bases de données.

A.6.1 Optimisation

Le *Random Forest* et le XGBoost sont des modèles ayant des hyperparamètres. De meilleures performances peuvent être obtenues suivant la valeur de leurs hyperparamètres. Souvent, les effets généraux des hyperparamètres pris individuellement sur un modèle sont connus, par exemple augmenter la profondeur de l'arbre *max_depth* (jusqu'à un certain seuil) permet d'améliorer les performances du modèle. Cependant, déterminer comment définir au mieux une combinaison d'hyperparamètres qui s'adapte bien aux données est un véritable challenge.
Une approche consiste à utiliser plusieurs combinaisons d'hypermaramètres parmi leurs différentes valeurs possibles (spécifiées) et à choisir la combinaison qui résulte sur un modèle atteignant les meilleures performances sur l'ensemble de données. C'est ce qu'on appelle l'optimisation des hyperparamètres. Le résultat d'une optimisation d'hyperparamètres est un ensemble unique d'hyperparamètres performants pouvant être utilisé pour configurer le modèle.

Nous cherchons donc les hyperparamètres qui vont maximiser l'AUCPr puisque nous avons remarqué que l'AUCPr des modèles avec les paramètres par défaut n'est pas très proche de 1. Pour cela,

le *gridsearch* (grille de recherche) de *h2o* va être utilisé. Celui-ci permet de trouver avec la validation croisée [3] et l'$early-stopping$ [4], les valeurs des hyperparamètres parmi les valeurs possibles [5] de chaque hyperparamètre permettant d'obtenir un *Random Forest* et un XGBoost avec une AUCPr maximal. Le problème principal de l'optimisation par *gridsearch* est le temps de calcul. Pour tester 150 combinaisons d'hyperparamètres du RF, le temps d'exécution est d'environ 9 heures. Afin d'optimiser le *Random Forest*, nous allons donc essayer de trouver la meilleure combinaison des hyperparamètres suivant :

— `ntrees` : il spécifie le nombre d'arbres à construire pour constituer la forêt. Augmenter le nombre d'arbres pourrait résulter sur un modèle plus stable et performant. Cependant, lorsque le nombre d'arbre augmente, le temps de calcul augmente ainsi que la complexité du modèle.

— `max_depth` : il spécifie la profondeur maximale de l'arborescence. Augmenter la profondeur de l'arbre peut permettre d'améliorer les performances mais il revient aussi à augmenter la complexité du modèle ainsi que le temps d'entrainement et peut pousser le modèle à sur-apprendre (*overfit*).

— `min_rows` : il spécifie le nombre minimum d'observations qu'il faut avant qu'une feuille ne se divise. Par exemple, si `min_rows` = 10 est choisi et que les données aboutissent sur 10 A et 8 B, le noeud ne se divisera pas car il nécessite 10 réponses des deux côtés. Ce paramètre peut donc permettre de lutter contre le sur-apprentissage.

— `sample_rate` : représente le taux d'échantillonnage des lignes de la base d'entraînement. Par exemple, si cet hyperparamètre a pour valeur 0,5, $random\ forest$ va collecter de manière aléatoire la moitié des instances de données pour développer des arbres. Ses valeurs possibles sont comprises entre 0,0 et 1,0. L'échantillonnage de lignes peut améliorer la généralisation et réduire les erreurs de validation et de test d'après Friedman (2002). Généralement, les bonnes valeurs pour les grands ensembles de données sont d'environ 0,7 à 0,8 (échantillonnage de 70 à 80% des données) car des valeurs plus élevées améliorent généralement la précision de l'entraînement.

En plus de ces paramètres, il y a aussi pour l'optimisation du XGBoost qui a pris 13 heures :

— `col_sample_rate` : il spécifie le taux d'échantillonnage des colonnes pour chaque division d'un noeud. Il permet donc de faire une sélection des variables (*feature sampling*). Il est similaire à la variable `sample_rate` mais l'échantillonnage se fait sur les colonnes. Sa valeur par défaut est 1,0 et ses valeurs possibles sont comprises entre 0,0 et 1,0. Des valeurs plus élevées peuvent améliorer la précision de l'entraînement. La précision du modèle peut s'améliorer lorsque des colonnes ou des lignes sont échantillonnées d'après Friedman (2002).

— `learn_rate` : il spécifie le taux d'apprentissage selon lequel il faut réduire les pondérations des variables. La réduction des poids des variables après chaque étape de *boosting* rend le processus de *boosting* plus conservateur et empêche le sur-apprentissage. Cet hyperparamètre a une valeur comprise entre 0,0 et 1,0.

Après avoir testé respectivement 150 modèles de *Random Forest* et 120 modèles de XGBoost avec le *gridsearch* de *h2o*, les meilleurs [6] modèles construits sans ré-échantillonnage de la base d'entraînement sont obtenus avec les valeurs des hyperparamètres présentées dans le tableau A.3 suivant.

3. définie en Annexe A.7.1
4. défini en Annexe A.7.2
5. valeurs que nous avons renseignées
6. le modèle maximisant l'AUCPr

Paramètre	Valeur optimale RF	Valeur optimale XGBost
`ntrees`	80	25
`max_depth`	45	30
`min_rows`	15	20
`sample_rate`	0.80	0.9
`col_sample_rate`	-	0.7
`learn_rate`	-	0.1

TABLE A.3 – Tableau présentant les valeurs des hyperparamètres du *Random Forest* et du XGBoost donnant le modèle maximisant l'AUCPr suite à l'optimisation

A.6.2 Choix du seuil de prédiction

Pour les problèmes de classification binaire, *h2o* utilise le modèle avec l'ensemble de données donné pour calculer le seuil de prédiction qui va permettre de maximiser le score F1 pour l'ensemble de données donné. Le score F1 est calculé à partir de la moyenne harmonique de la précision et du rappel comme suit

$$F1 = 2\left(\frac{(precision)\,(rappel)}{precision + rappel}\right). \tag{A.15}$$

Le score F1 fournit une mesure de la capacité d'un classificateur binaire à classer correctement les cas positifs (étant donné une valeur seuil) d'après la documentation *h2o* [7]. Choisir le seuil de prédiction permettant de maximiser le score F1 pourrait donc aider à lutter contre le déséquilibre des classes.

A.6.3 Importance des variables

L'importance [8] d'une variable du package *h2o* pour le *Random Forest* et le XGBoost est déterminée en calculant dans quelle mesure l'erreur au carré sur tous les arbres (*Squared Error* SE) a baissé lorsque cette variable a été sélectionnée pour faire la division de l'arbre. Chaque fois qu'un noeud est divisé en fonction d'une variable, la réduction de l'erreur quadratique obtenue grâce à la variable est la différence d'erreur quadratique entre ce noeud et ses noeuds fils. L'erreur quadratique pour chaque nœud individuel est la réduction de la variance de la variable réponse au sein de ce nœud. Le calcul suppose un estimateur sans biais, c'est-à-dire : $SE = MSE \times N$,

$$SE = MSE \times N = \frac{1}{N}\sum_{i=0}^{N}(y_i - \bar{y})^2 \times N = \left[\frac{1}{N} \times \sum_{i=0}^{N} y_i^2 - N \times \bar{y}^2\right] \times N = \left[\sum_{i=0}^{N} \frac{y_i^2}{N} - \bar{y}^2\right] \times N. \tag{A.16}$$

A.7 Validation croisée et *early-stopping*

Il est possible que le modèle construit fasse de l'*overfitting* c'est-à-dire qu'il apprenne "par cœur" les données d'entraînement au risque de ne pas savoir généraliser à des données inconnues. Ainsi,

7. Performance and prediction, h2o
8. Variable Importance, h2o

pour éviter l'*overfitting*, il faut réévaluer le modèle à chaque fois sur des données non vues pendant l'entraînement. Pour cela, une validation croisée a été effectuée. Un *early-stopping* a aussi été utilisé pour lutter contre l'*overfitting*.

A.7.1 Validation croisée

La validation croisée (CV) est l'une des techniques les plus utilisées pour tester l'efficacité d'un modèle machine learning et essayer d'éviter l'*overfitting*. Pour effectuer une CV, une partie des données est gardée et donc non utilisée pour former le modèle. Elle va donc être utilisée pour tester et valider le modèle. La technique des K-blocs comme validation croisée est une technique populaire et facile à comprendre. Elle aboutit généralement sur un modèle moins biaisé par rapport à une séparation en base d'apprentissage et base de test uniquement puisqu'elle garantit que chaque observation de l'ensemble de données d'origine a la chance d'apparaître dans l'ensemble d'apprentissage et de validation. Pour faire une validation croisée avec K blocs, il faut :

1. Diviser la base d'entraînement de manière aléatoire en K blocs. La valeur de K ne doit pas être trop petite ou trop élevée, généralement 5 à 10 sont choisies en fonction de la taille des données. Nous avons choisi $K = 10$ puisque certaines des modalités de nos variables explicatives ont un volume faible.
2. Ajuster le modèle en utilisant les K-1 (K moins 1) blocs et valider le modèle en utilisant le K-ème bloc restant. Les scores ou erreurs sont évaluées.
3. Répéter ce processus jusqu'à ce que chaque bloc en K serve d'ensemble de test. Ensuite, la moyenne des scores enregistrés est faite pour constituer la métrique de performance pour le modèle.

Avec *h2o*, pour faire la validation croisée[9], il faut définir la valeur du paramètre `nfolds` à une valeur supérieure à 1. Nous avons donc paramétré `nfolds`=10.

A.7.2 *Early-stopping*

Trop entraîner le modèle pourrait provoquer l'*overfitting*. Mais ne pas suffisamment entrainer le modèle fait que celui-ci ne va pas bien s'ajuster aux bases d'apprentissage et de test. Pour trouver le juste milieu, il existe une méthode qui consiste à entraîner le modèle sur l'ensemble de données d'entraînement et à arrêter l'entraînement au moment où les performances sur la base d'entrainement ou la base de validation ne s'améliore plus.

Ainsi, elle permet aussi de diminuer le temps de calcul. Elle est très utilisée en deep learning grâce à sa simplicité et à son efficacité. Elle est appelée $early - stopping$.
Avec *h2o*, il est possible de spécifier les valeurs des paramètres liés au $early - stopping$[10]. Ces paramètres sont :

— `stopping_rounds` : cette option permet d'arrêter l'entraînement du modèle lorsque la métrique choisie pour `stopping_metric` ne s'améliore pas avec le nombre spécifié pour définir la valeur de l'option `stopping_rounds`. La valeur par défaut est égale à 0 pour le Random Forest et le XGBoost. `early_stopping` est désactivé dans ce cas. Pour l'activer , il faut que `stopping_rounds >0`.

— `stopping_tolerance` : cette option spécifie la valeur de la tolérance. Le modèle doit s'améliorer de cette valeur pour que son entraînement ne s'arrête pas. La valeur par défaut est 0,001.

9. Cross validation, h2o
10. Early-Stopping, h2o

— `stopping_metric` : cette option spécifie la métrique à évaluer pour savoir s'il faut continuer ou non l'entrainement du modèle. Par défaut, sa valeur est «logloss» pour la classification et «déviance» pour la régression.

Quand la base de validation est définie parmi les paramètres du modèle, l'algorithme arrête d'apprendre lorsque la métrique choisie de la base d'entrainement est supérieure à celle de la base de validation. Si la validation croisée est activée, tous les modèles de validation croisée arrêtent l'entraînement lorsque la métrique de validation ne s'améliore pas.

A.8 Pistes pour la projection de l'impact et la durée des parcours

Les parcours n'ont pas la même ancienneté. Certains ont juste une année d'ancienneté. Dans le cas où ils auraient un impact sur la résiliation, il faudrait essayer de trouver des pistes afin de modéliser les durées de vie de ces parcours sur le même horizon que les autres parcours qui ont plus d'une année d'ancienneté. Pour cela, deux pistes auraient pu être exploitées :

Piste 1 :
Il est possible de retrouver les mêmes profils dans plusieurs parcours. Cependant, l'effet du parcours sur la résiliation, s'il existe, va être l'élément différenciant ces profils aux caractéristiques similaires. Pour modéliser les taux de résiliation d'un parcours pendant une période où il n'a plus de contrats, nous pourrions :

— (i) regarder les profils similaires à ceux du parcours manquant dans les autres parcours (utilisation des k plus proches voisins par exemple) et qui n'ont pas été résiliés avant la période que nous souhaitons modéliser,

— (ii) étudier ces profils là pour représenter le parcours manquant,

— (iii) ajouter l'impact du parcours (sur la résiliation) aux caractéristiques du profil.

Pour déterminer l'impact du parcours sur la résiliation de chaque période, nous pourrions :

— pour les parcours avec une seule année d'ancienneté, uiliser l'impact du parcours sur la résiliation pendant la première année et supposer qu'elle est constante pour les autres années ou qu'elle décroît d'un certains pourcentage chaque année.

— pour les parcours avec plusieurs années d'ancienneté, étudier l'évolution de l'impact du parcours sur la résiliation suivant l'année passée dans le portefeuille. Par exemple, si l'effet baisse d'environ 0,3 chaque année, nous pourrions supposer que l'effet du parcours sur la résiliation pendant les années où nous n'avons plus d'affaires nouvelles du parcours baisse de 0,3 à chaque année.

Piste 2 :
Pour modéliser les durées de vie des parcours avec une ancienneté inférieure à celle des autres parcours, nous pourrions utiliser l'origine du contrat et l'intermediaire (de ce fait origine=digitale et intermediaire=plateforme va représenter les parcours manquants). Des contrats remplissant ces critères et ayant 5 ans d'ancienneté sont trouvés dans le parcours MDM (qui a néanmoins peu de données). Mais cela reviendrait à regrouper tous les parcours du canal direct pour certaines années. Si un parcours a une influence particulière sur la résiliation, l'influence ne sera pas prise en compte.

A.9 Optimisation sous contraintes

A.9.1 Convexité

Une fonction est convexe si sa dérivée est croissante. Ce qui n'est pas le cas de la fonction objectif f représentant la rentabilité sur 4 ans et utilisée pour l'optimisation sous contrainte à la partie 4.4. Pour le prouver, dérivons la fonction suivant une de ses variables x_1

$$\frac{df(x_1, x_2, x_3...x_{16})}{dx_1} = \frac{S_1 \sum_{i=1}^{K=16} x_i \times C_i - C_1 \sum_{i=1}^{K=16} x_i \times S_i}{(\sum_{i=1}^{K=16} x_i \times C_i)^2}. \tag{A.17}$$

Vu $x_1 \geq 0$, $x_1 \in \mathbb{N}$, le terme $\frac{1}{(\sum_{i=1}^{K=16} x_i \times C_i)^2}$ décroît quand la valeur de x_1 augmente tandis que le terme $S_1 \sum_{i=1}^{K=16} x_i \times C_i - C_1 \sum_{i=1}^{K=16} x_i \times S_i$ reste constant (puisqu'il ne dépend pas de x_1). En effet,

$$\begin{aligned} S_1 \sum_{i=1}^{K=16} x_i \times C_i - C_1 \sum_{i=1}^{K=16} x_i \times S_i &= S_1 x_1 C_1 + S_1 \sum_{i=2}^{K=16} x_i \times C_i - C_1 x_1 S_1 - C_1 \sum_{i=2}^{K=16} x_i \times S_i, \quad (A.18) \\ &= S_1 \sum_{i=2}^{K=16} x_i \times C_i - C_1 \sum_{i=2}^{K=16} x_i \times S_i. \quad (A.19) \end{aligned}$$

Ainsi, pour que la dérivée en x_1 soit croissante, il faut que $S_1 \sum_{i=2}^{K=16} x_i \times C_i - C_1 \sum_{i=2}^{K=16} x_i \times S_i$ soit négatif $\forall x_i \in \mathbb{N},\ 2 \leq i \leq 16$. Or, avec $C_1 \neq 0$ et $\sum_{i=2}^{K=16} x_i \times C_i \neq 0$

$$S_1 \sum_{i=2}^{K=16} x_i \times C_i - C_1 \sum_{i=2}^{K=16} x_i \times S_i < 0 \iff \frac{S_1}{C_1} < \frac{\sum_{i=2}^{K=16} x_i \times S_i}{\sum_{i=2}^{K=16} x_i \times C_i}. \tag{A.20}$$

Il est possible de trouver des nombres entiers x_i ne vérifiant pas cet inégalité ce qui entraine que la dérivée de f en x_1 n'est pas croissante. Il en est de même pour les autres x_i.

Nous pouvons ainsi en déduire que la fonction objectif f n'est pas une fonction convexe.

A.9.2 Algorithmes de résolution

Evolution différentielle

Soit $\mathbf{x} \in \mathbb{R}^D$ une solution candidate dans la population courante, où D est la dimension du problème à optimiser et $f : \mathbb{R}^D \to \mathbb{R}$ est la fonction objectif à minimiser. L'algorithme DE de base, suivant le schéma «DE/rand/1», peut être décrit schématiquement comme suit.

Algorithme 4 : Pseudo code de l'algorithme d'évolution différentielle DE/rand/1

Entrées : `NP` : taille de la population, `F` : fateur de mutation, `CR` : probabilité de croisement, `MAXFES` : nombre maximal d'itérations de la fonction objectif

INITIALISATION G=0 (génération) ; FES=1 ; Initialiser tous les individus `NP` avec des positions aléatoires dans l'espace de recherche;

tant que $FES < MAXFES$ **faire**

  pour $i \leftarrow 1$ *à* NP **faire**

    GÉNÉRER trois individus x_{r1}, x_{r2}, x_{r3} *à partir de la population actuelle au hasard. Ceux-ci doivent être distincts les uns des autres mais aussi de l'individu* x_i, *i.e.* $r_1 \neq r_2 \neq r_3 \neq i$;

    MUTATION Formez le vecteur donneur à l'aide de la formule

$$\mathbf{v}_i = \mathbf{x}_{r_1} + F\left(\mathbf{x}_{r_2} - \mathbf{x}_{r_3}\right). \tag{A.21}$$

    CROSSOVER Le vecteur d'essai $\mathbf{u}_i$ *est développé à partir des éléments du vecteur cible* $\mathbf{x}_i$ *et des éléments du vecteur donneur* $\mathbf{v}_i$ *comme suit*

$$u_{i,j} = \begin{cases} \mathrm{v}_{i,j}, & si\ r_{i,j} \leq CR\ ou\ j = j_{rand}\,. \\ x_{i,j}, & sinon. \end{cases} \tag{A.22}$$

    où $i = \{1, \ldots, NP\}, j = \{1, \ldots, D\}, r_{i,j} \sim \cup(0,1)$ *est un nombre aléatoire uniformément distribué qui est généré pour chaque* j *et* $j_{rand} \in \{1, \ldots, D\}$ *est un entier aléatoire utilisé pour s'assurer que* $\mathbf{u}_i \neq \mathbf{x}_i$ *dans tous les cas ;*

    si $f(\mathbf{u}_i) \leq f(\mathbf{x}_i)$ **alors**

      remplacer l'individu $\mathbf{x}_i$ *dans la population par le vecteur d'essai* $\mathbf{u}_i$.

    fin

    $FES = FES + NP$

  fin

  G=G+1.

fin

Basin-hopping

Comme indiqué dans le pseudocode de l'algorithme 5, le cadre peut être décrit en termes de procédure de recherche locale `RECHERCHELOCALE` qui mappe un point X_i dans l'espace variable à son minimum le plus proche Y_i avant d'ajouter un mouvement de perturbation `PERTURB` qui modifie le minimum courant Y_i pour obtenir un nouveau point X_{i+1} dans l'espace variable, et un critère d'arrêt `STOP` qui met fin à ces applications répétées d'une perturbation structurelle suivie d'une optimisation locale. Les applications répétées aboutissent à une trajectoire de minima locaux Y_i. Comme le montre l'algorithme 5, seul le minimum le plus bas doit être conservé en mémoire lors de la recherche du minimum global d'une fonction f. Il est important de noter que l'algorithme 5 montre une réalisation spécifique du cadre BH, connu sous le nom de BH (MBH) monotone, où le minimum actuel n'est pas accepté s'il n'abaisse pas la valeur la plus basse obtenue pour la fonction f jusqu'ici. Dans ce cas, une autre perturbation est tentée afin d'obtenir un nouveau point de départ pour l'optimisation locale qui suit.

Algorithme 5 : Pseudo-code du *Basin-Hopping*

(1) $i \leftarrow 0$;
(2) $X_i \leftarrow$ point initial aléatoire dans l'espace variable ;
(3) $Y_i \leftarrow$ RECHERCHELOCALE (X_i) ;
tant que *STOP n'est pas satisfait* **faire**
    (5) $X_{i+1} \leftarrow$ PERTURB (Y_i) ;
    (6) $Y_{i+1} \leftarrow$ RECHERCHELOCALE (X_{i+1}) ;
    si $f(Y_{i+1}) < f(Y_i)$ **alors**
        $i \leftarrow i + 1$.
    fin
fin

Annexe B

Analyses et résultats

B.1 Caractéristiques du portefeuille

La figure B.1 suivante montre le pourcentage des affaires nouvelles souscrites en 2019 et en 2020 suivant le parcours.

Parcours	% des affaires nouvelles en 2019	% des affaires nouvelles en 2020
ADA	6,09%	7,23%
ATA	88,66%	85,87%
MDA	0,27%	0,49%
MDM	0,30%	1,31%
MTA	0,363%	0,003%
MTM	2,78%	0,87%
WDA	0,31%	0,83%
WDM	1,22%	3,39%
Total général	**100,00%**	**100,00%**

FIGURE B.1 – Pourcentage des affaires nouvelles souscrites en 2019 et en 2020 suivant le parcours

Les figures B.2 et B.3 suivantes présentent respectivement les répartitions propriétaires/locataires, du nombre de contrats du client et de l'âge du client suivant l'année de souscription.

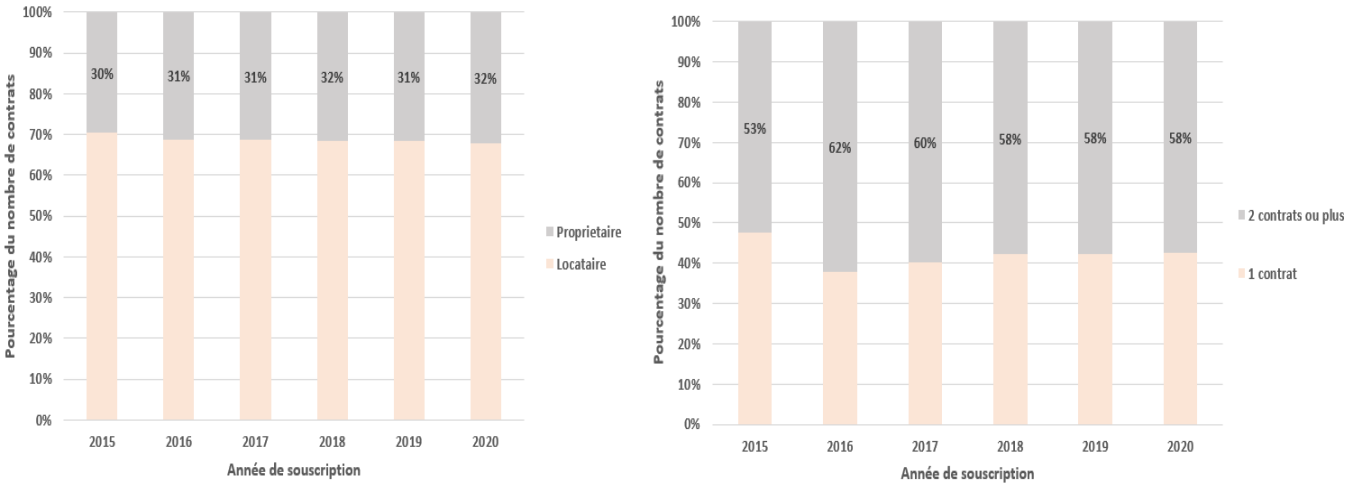


FIGURE B.2 – Pourcentage de locataires et de propriétaires suivant l'année de souscription (à gauche) et pourcentage de clients multi-détenteurs de contrats suivant l'année de souscription (à droite)

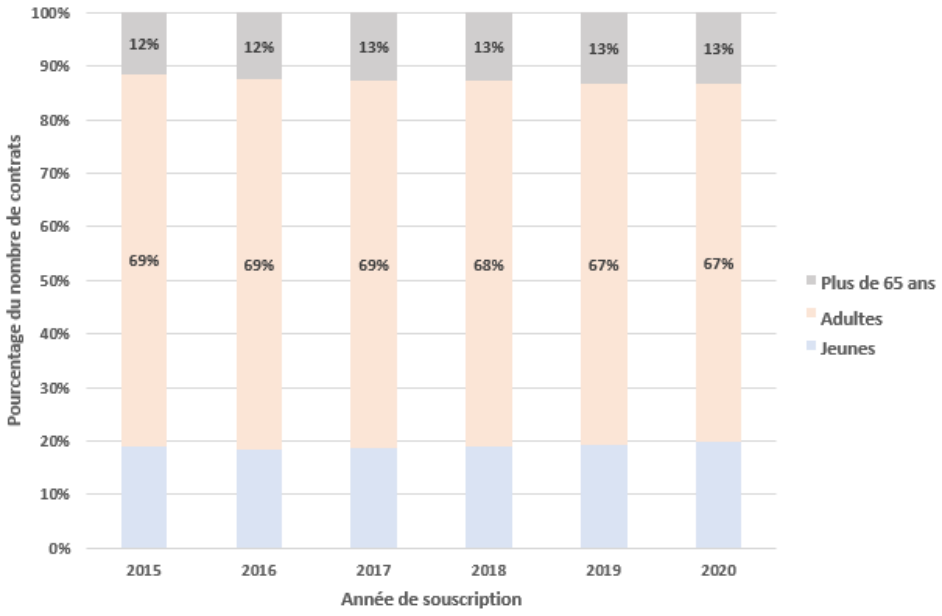


FIGURE B.3 – Pourcentage de chaque catégorie d'âge suivant l'année de souscription

Nous remarquons qu'en 2015 , il y a 1% ou 2% moins de propriétaires que les autres années, au moins 5% moins de clients multi-détenteurs par rapport aux autres années.

B.2 Variables explicatives

Le tableau B.1 suivant présente le restant des principales variables contrat ainsi que la répartition de leurs modalités et le pourcentage de valeurs manquantes de ces variables. Ces variables peuvent présenter des valeurs manquantes qui sont remplacées par la modalité la plus fréquente. Nous suppo-

sons que la valeur de la variable est manquante parce qu'elle n'est pas renseignée et que cela se justifie par le fait que l'assuré ait choisi l'option la plus fréquente. Par exemple, si la fréquence de paiement n'est pas précisée c'est parce que l'assuré a choisi la cadence mensuelle (qui est la plus courante).

Variable	Description	Modalités	Répartition des modalités	Valeurs manquantes
`ext_dommages`	extention de garantie	0 (sans) 1 (avec)	53% 47%	0%
`vol`	garantie vol	0 (sans) 1 (avec)	5% 95%	0%
`bdg`	garantie bris de glace	0 (sans) 1 (avec)	2% 98%	0%
`ass`	option assistance	0 (sans) 1 (avec)	3% 97%	0%
`enfant`	l'assuré a-t-il un ou plusieurs enfants ?	0 (non) 1 (oui)	63% 34%	3%
`nb_sin_ant`	l'assuré a-t-il eu un ou plusieurs sinistres avant de souscrire ?	0 (non) 1 (oui)	57.95% 0.05%	42%
`CDFREQ`	fréquence de paiement	Mensuelle Autre	56% 11%	33%

TABLE B.1 – Tableau présentant les principales autres variables contrat

Le tableau B.2 suivant présente des variables zonier.

Variable	Description	Valeurs manquantes
`CNTGN_FREQ`	zonier technique CNTGN	0.6%
`ATT_com`	zonier commercial attentat	0.6%
`DDE_com`	zonier commercial dégât des eaux	0.6%
`BDG_com`	zonier commercial bris de glace	0.6%
`VOL_com`	zonier commercial vol	0.6%
`RCIM_com`	zonier commercial Responsabilité Civile propriétaire d'IMmeuble	0.6%
`RCVP_com`	zonier commercial Responsabilité Civile Vie Privée	0.6%
`ELEC_com`	zonier commercial dommage électrique	0.6%
`score_s`	score sécheresse	0.6%

TABLE B.2 – Tableau présentant d'autres variables zonier

B.3 Analyse descriptive

B.3.1 Type d'habitation

La figure B.4 suivante présente la répartition des appartements/maisons des affaires nouvelles de la base de modélisation suivant le parcours. Nous remarquons que les parcours ADA, ATA et MTM sont les parcours qui ont au moins 28% d'individus assurant une maison. Cela s'explique par le fait

que les prospects voulant assurer une maison sont généralement orientés vers les parcours agence et ceux vivant loin d'une agence dans le parcours MTM.

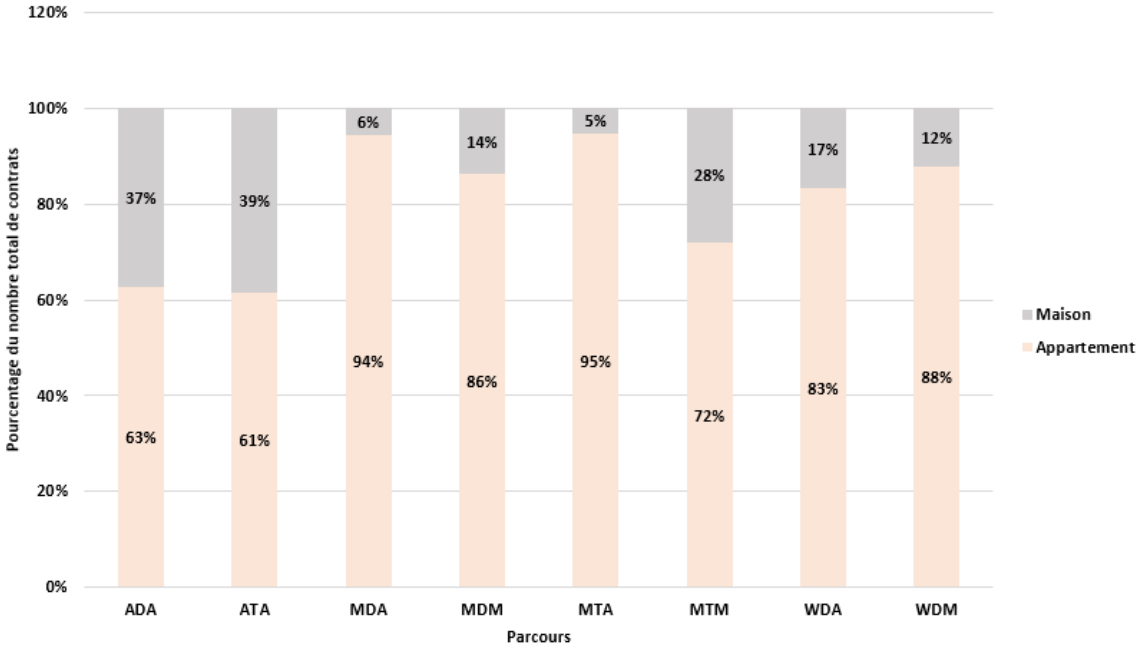


FIGURE B.4 – Répartition du type d'habitation des affaires nouvelles par parcours

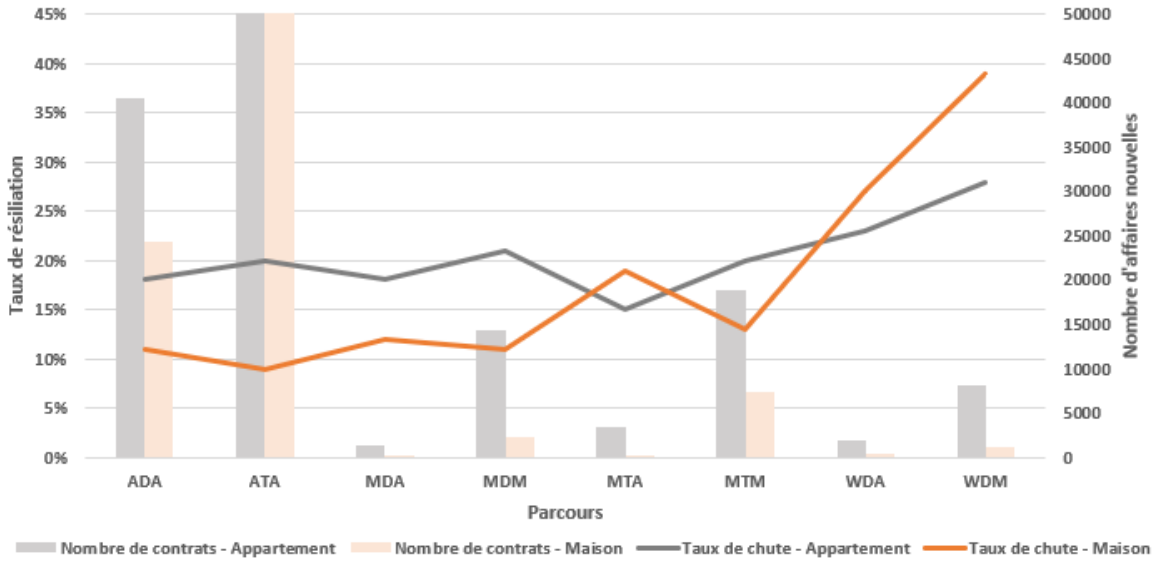


FIGURE B.5 – Nombre d'AFN et taux de résiliation suivant la type d'habitation et le parcours

La figure B.5 ci-dessus présente les taux de résiliation pendant la première année suivant le type d'habitation et le parcours. Nous observons alors que les taux de résiliation des appartements sont

plus élevés que ceux des maisons pour les parcours agence.

Une troisième variable permettant de différencier les profils est la catégorie socio-professionnelle.

B.3.2 Catégorie socio-professionnelle du client

Pour faciliter l'interprétation, nous regardons la répartition des catégories socio-professionnelles selon l'origine (digitale ou traditionnelle) du contrat.
La figure B.6 suivante présente la répartition de certaines catégories socio-professionnelles suivant l'origine digitale ou non du contrat. Nous remarquons que globalement moins de 10% des clients utilisent le digital. Les écarts de pourcentage entre les différentes catégories socio-professionnelles différenciant ceux qui utilisent le digital de ceux qui ne l'utilisent pas, ne sont pas grandes. Ce sont les retraités qui utilisent le moins Internet pour trouver leur contrat d'assurance habitation chez Allianz. Les cadres et les employés font partie de ceux qui utilisent le plus Internet mais la différence de pourcentage par rapport aux catégories socio-professionnelles mentionnées précédemment est faible (environ 2%). Les clients qui utilisent le plus le digital ne renseignent pas leur catégorie socio-professionnelle.

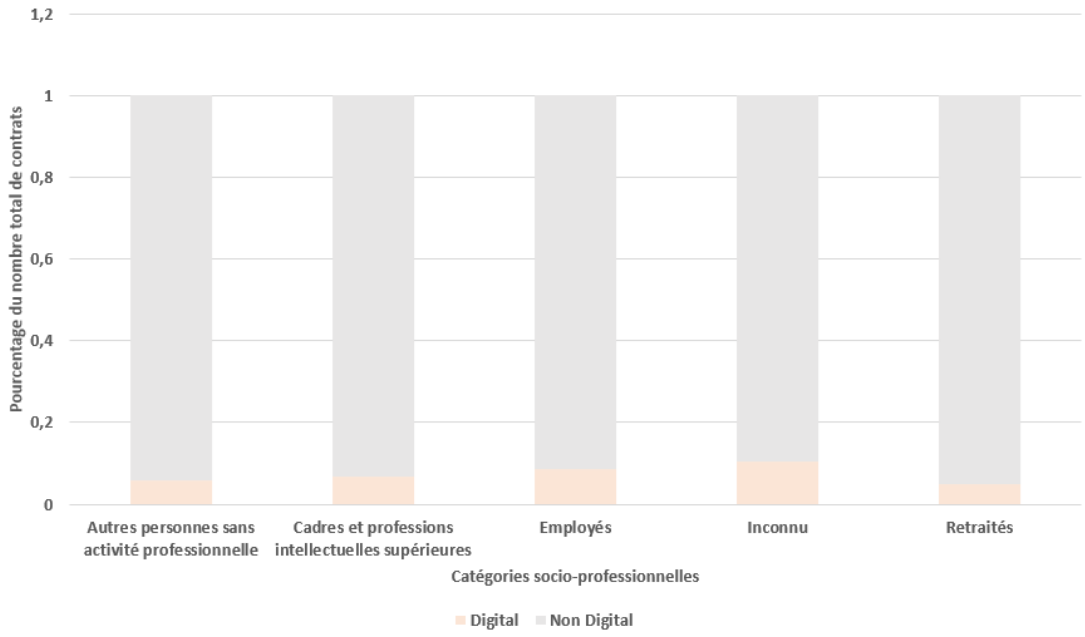


FIGURE B.6 – Répartition des AFN par parcours suivant l'origine du contrat et la catégorie socio-professionnelle

La figure B.7 ci-dessous présente le taux de résiliation pendant la première année du contrat venant du digital ou non suivant la catégorie socio-profesionnelle du client. Quand nous observons les taux de résiliation pendant la première année du contrat, nous constatons que l'origine du contrat semble avoir peu d'impact sur les taux de résiliation suivant les catégories socio-professionnelles puisque les taux de résiliation sont proches. Les retraités sont ceux qui résilient le moins. Les personnes sans activité professionnelle sont, sans surprise, celles qui résilient le plus. Elles pourraient être plus sensibles au prix puisqu'elles n'ont pas de salaires. Elles pourraient donc facilement résilier leur contrats si elles trouvent

moins chers ailleurs. Elles peuvent aussi être résiliées par l'assureur dans le cas où elles n'auraient pas pu payer leur prime.

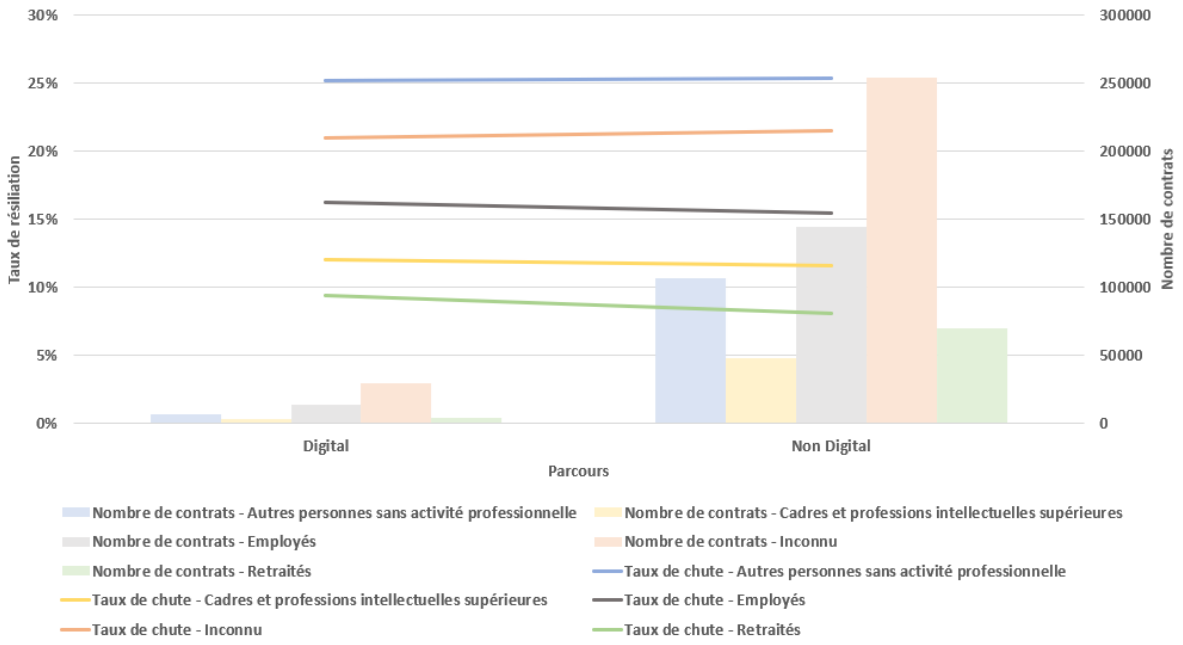


FIGURE B.7 – Taux de résiliation suivant la catégorie socio-professionnelle et l'origine du contrat

B.4 Performances des modèles

Ces éléments permettent de justifier le choix du XGBoost comme meilleur modèle pour les taux de résiliation.

B.4.1 Troisième année

Nous avons regardé les valeurs des métriques suivant le modèle et **par parcours**. Elles sont proches d'un modèle à l'autre. Ce n'est donc pas un élément qui nous a aidé dans notre choix du meilleur modèle. Le tableau B.3 suivant présente les valeurs des métriques pour chaque parcours obtenues avec le XGBoost pour la troisième année.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
ADA	0.240	70%	20%	67%	20%
ATA	0.30	74%	21%	69%	20%
MM	0.293	75%	43%	70%	21%

TABLE B.3 – Tableau présentant les valeurs des métriques pour chaque parcours obtenues avec le XGBoost pour la troisième année

Nous allons finalement vérifier si le XGBoost donne des taux de résiliation plus proches des taux de résiliation réels de la base test suivant la qualité juridique.
Le tableau B.4 suivant permet de comparer les taux de résiliation prédits par les modèles et les taux

de résiliation réels sur la base test pour la troisième année suivant la qualité juridique.

Modèle	Locataire	Propriétaire	Total
GLM	19%	1%	9%
RF	19%	1%	9%
XGBoost	17%	4%	9.7%
taux de résiliation réél	15%	5%	9.4%

TABLE B.4 – Tableau comparant les taux de résiliation obtenus suivant le modèle et les taux de résiliation réels sur la base test pour la troisième année

Nous remarquons qu'au total, les taux trouvés sont proches des taux réels observés dans la base test. Quand nous distinguons suivant la qualité juridique, nous nous rendons compte que les taux de résiliation prédits pour les locataires sont supérieurs aux taux de résiliation réels alors que les les taux de résiliation prédits pour les propriétaires sont inférieurs aux taux de résiliation réels. Le XGBoost est le modèle dont les taux de résiliation prédits se rapproche le plus des taux de résiliation réels suivant la qualité juridique.

B.4.2 Quatrième année

Le tableau B.5 suivant présente les valeurs des métriques pour chaque parcours obtenues avec le XGBoost pour la quatrième année.

Modèle	Logloss	AUC	AUCPr	Rappel	Precision
ADA	0.21	77%	14%	59%	16%
ATA	0.20	76%	16%	61%	16%
MM	0.20	75%	20%	60%	18%

TABLE B.5 – Tableau présentant les valeurs des métriques calculées avec la base test pour chaque parcours obtenues avec le XGBoost pour la quatrième année

Le tableau B.6 suivant permet de comparer les taux de résiliation prédits par les modèles et les taux de résiliation réels sur la base test pour la troisième année suivant la qualité juridique.

Modèle	Locataire	Propriétaire	Total
GLM	17%	1%	7%
RF	17%	2%	7.5%
XGBoost	16%	2%	7.2%
Taux de résiliation réél	14%	4%	7.7%

TABLE B.6 – Tableau comparant les taux de résiliation obtenus suivant le modèle et les taux de résiliation réels sur la base test pour la quatrième année

B.5 Impact des variables

B.5.1 Deuxième année

Le *SHAP summary plot* présenté dans la figure B.8 permet de voir l'impact des variables sur la résiliation pendant la deuxième année.

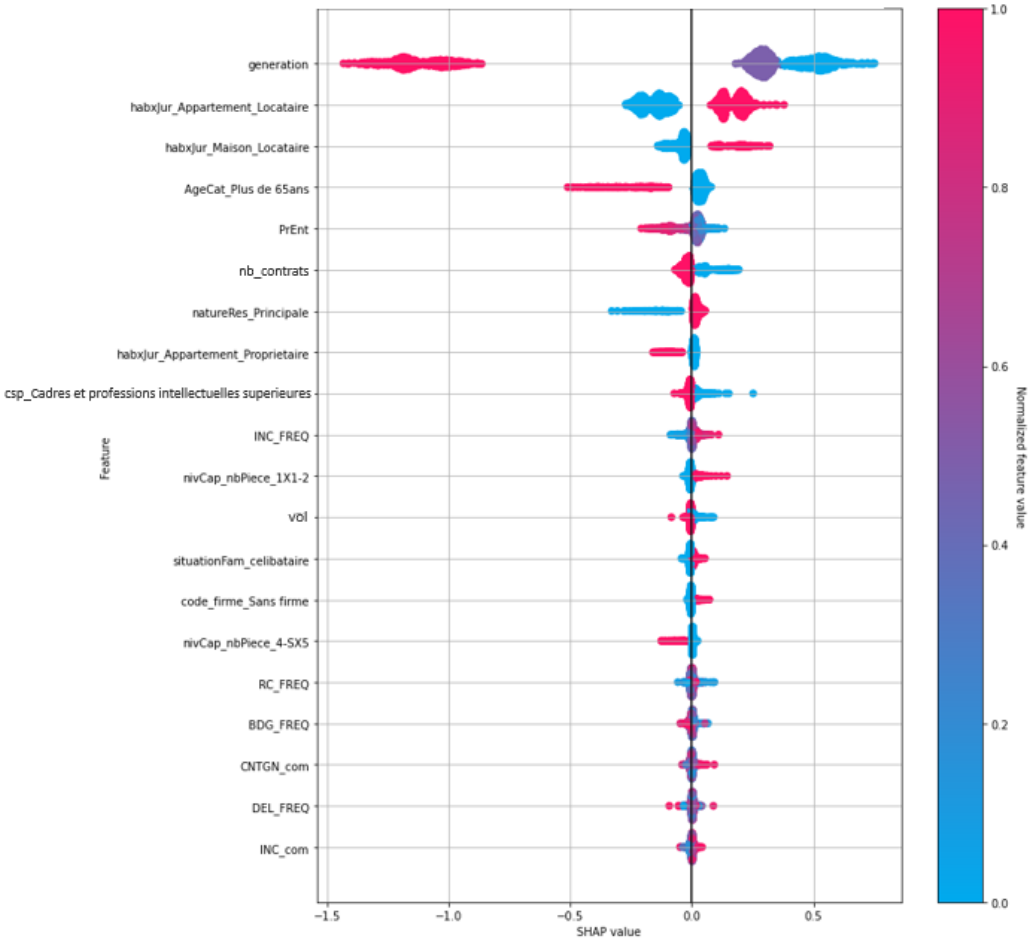


FIGURE B.8 – SHAP Summary Plot du modèle XGBoost de la deuxième année

B.5.2 Troisième année

Le *SHAP summary plot* présenté dans la figure B.9 permet de voir l'impact des variables sur la résiliation pendant la troisième année.

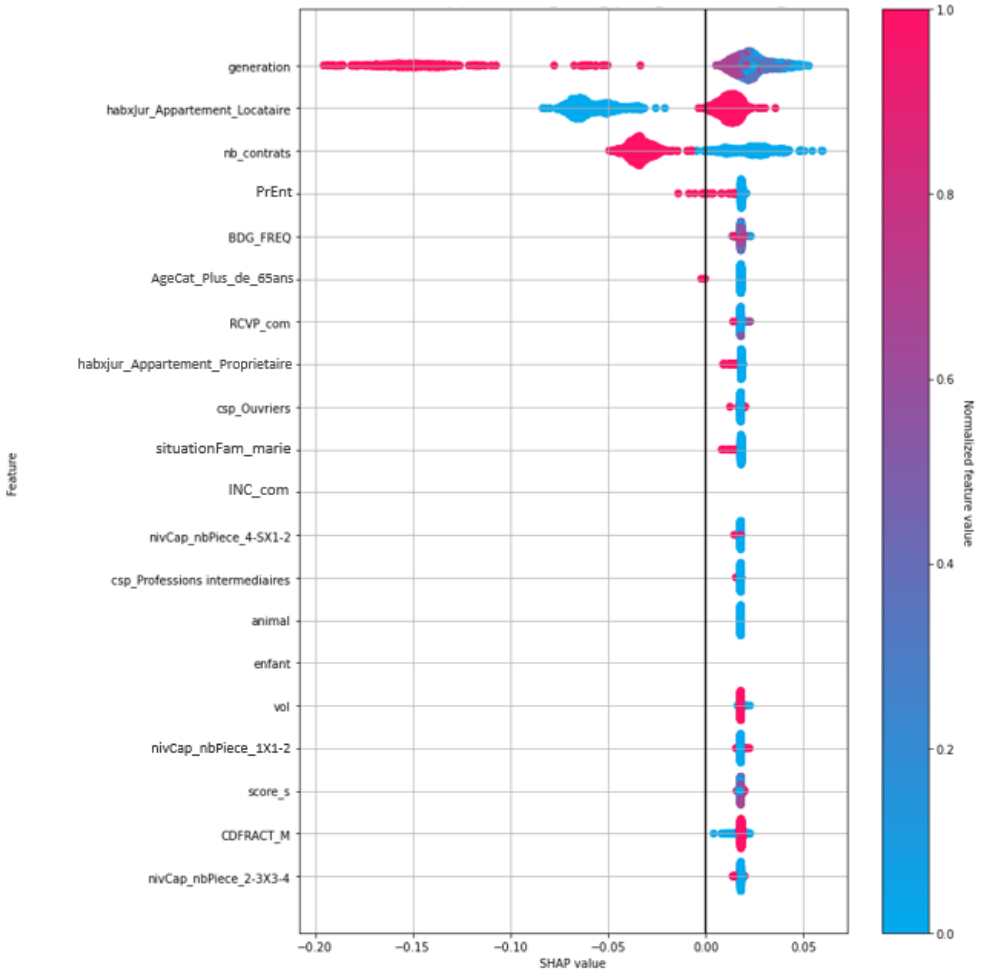


FIGURE B.9 – SHAP Summary plot du modèle XGBoost de la troisième année

B.5.3 Quatrième année

Le *SHAP summary plot* présenté dans la figure B.10 suivante permet de voir l'impact des variables sur la résiliation pendant la quatrième année.

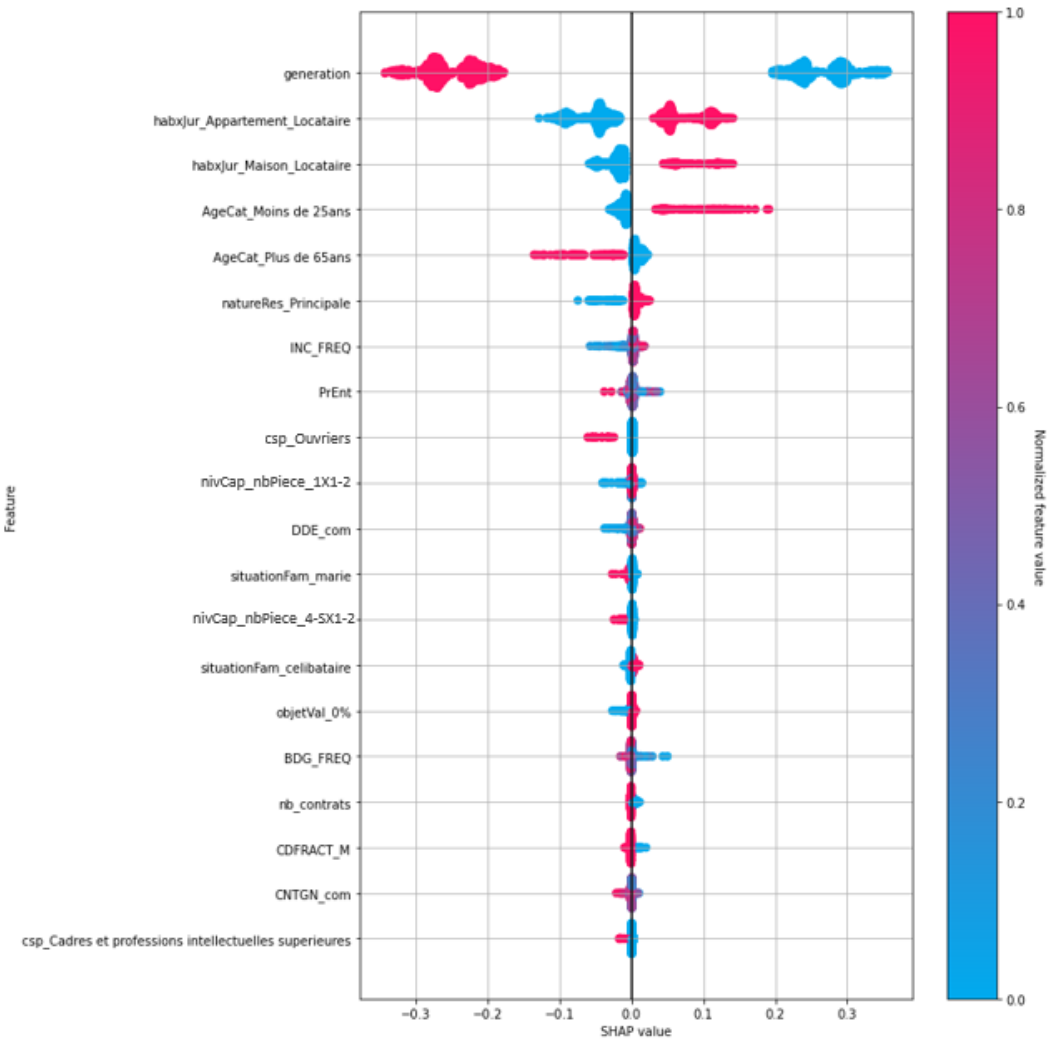


FIGURE B.10 – SHAP *Summary plot* du modèle XGBoost de la quatrième année

B.6 Identification des profils rentables par parcours

Les quatres parties évoquées dans la partie 4.3.1 (optimisation sans contrainte) sont coloriées suivant la couleur correspondante.

B.6.1 Parcours ADA

La figure B.11 suivante constitue un nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours ADA. Le seuil de la durée de vie est de 2 ans.

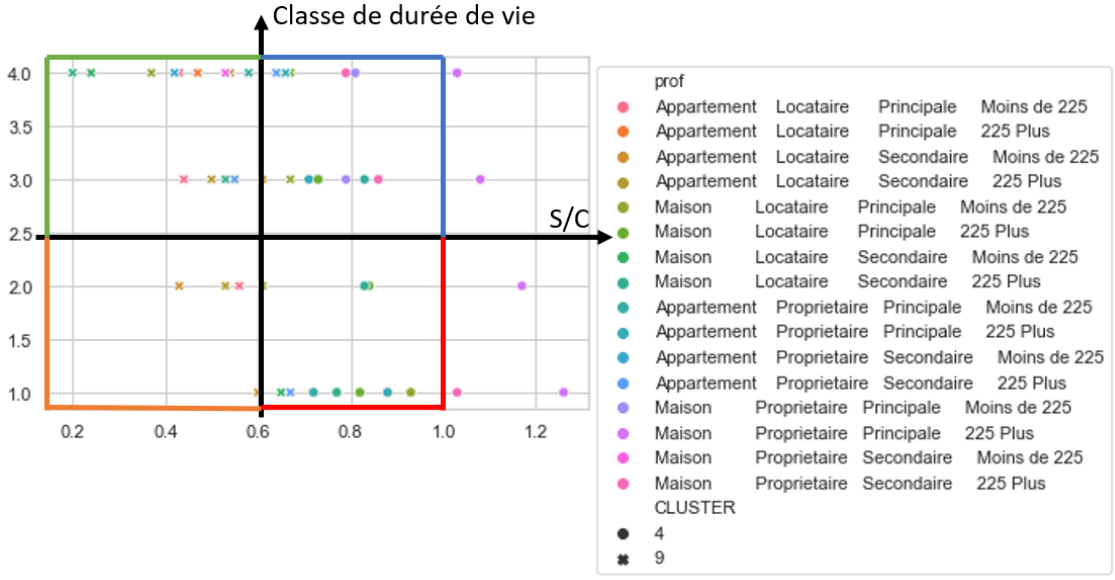


FIGURE B.11 – Nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours ADA

B.6.2 Parcours MDM

La figure B.12 suivante constitue un nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours MDM. Le seuil de la durée de vie est de 3 ans.

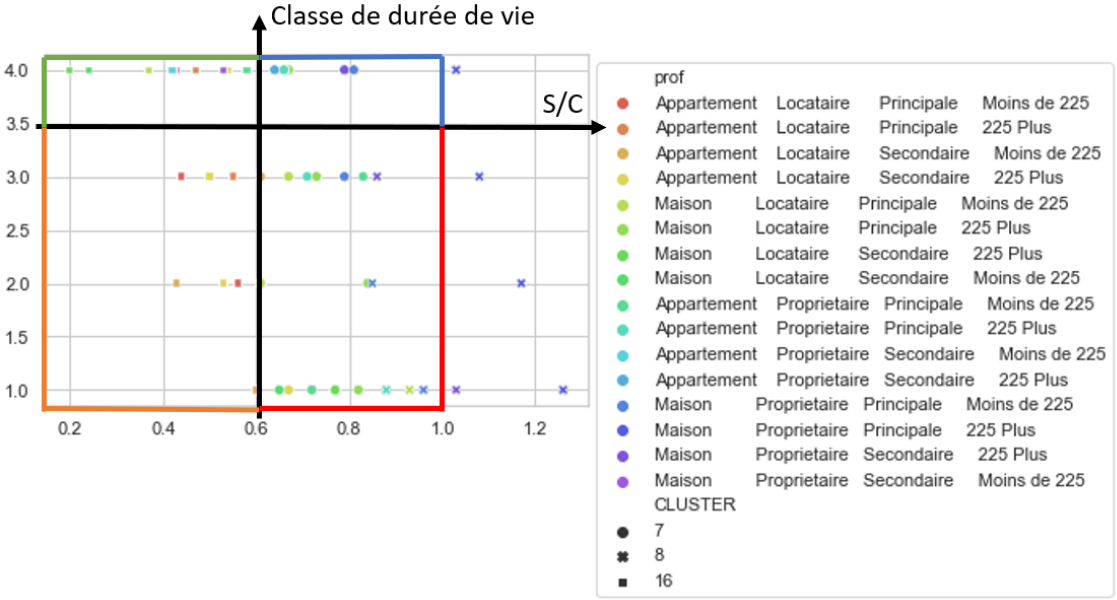


FIGURE B.12 – Nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours MDM

B.6.3 Parcours MDA

La figure B.13 suivante constitue un nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours MDA. Le seuil de la durée de vie est de 3 ans.

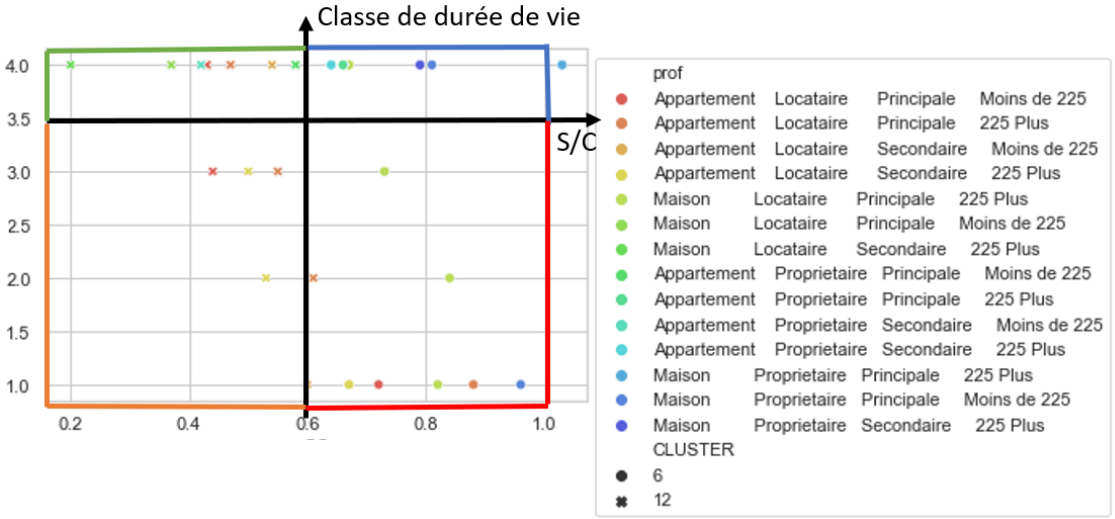


FIGURE B.13 – Nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours MDA

B.6.4 Parcours WDM

La figure B.14 suivante constitue un nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours WDM. Le seuil de la durée de vie est de 2 ans.

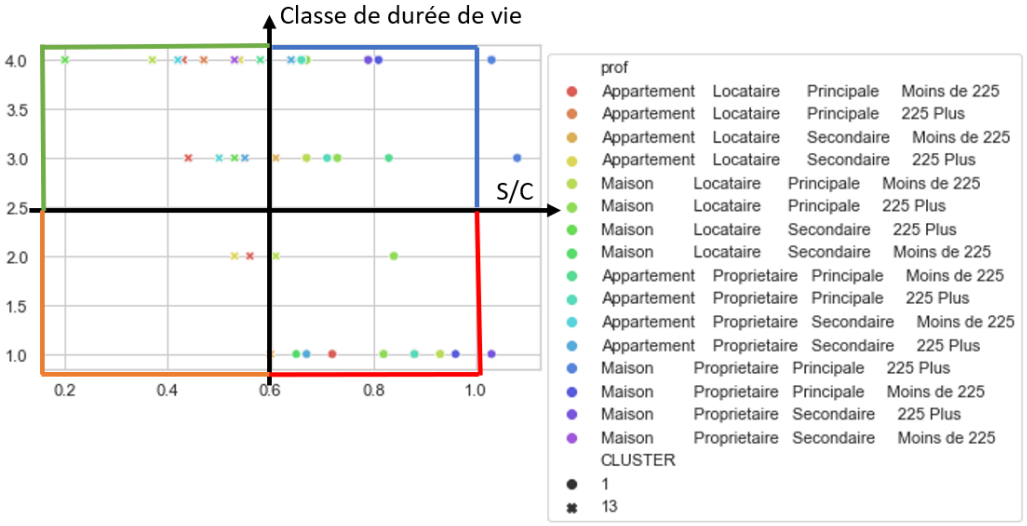


FIGURE B.14 – Nuage de points représentant les profils suivant le ratio S/C (en abscisse) et la durée de vie (en ordonnée) pour le parcours WDM