Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : LOPEZ Quentin

Titre : Étude des résiliations des contrats multirisques habitation et leur sensibilité aux majorations

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

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

Membre présents du jury de l'Institut des Actuaires signature

ZEC Nicolas

Membres présents du jury de l'ISFA

EYRAUD-LOISEL Anne

Entreprise : COVEA

Nom : WILCZIK Ewen

Signature :

Directeur de mémoire en entreprise :

Nom : RIOULT Mathieu

Signature :

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

Signature du responsable entreprise

Signature du candidat

Résumé

Le marché de l'assurance habitation a fortement évolué ces dernières années et plus particulièrement la concurrence au sein de cette branche. Pour un assureur, il est alors nécessaire d'avoir une connaissance importante de ses flux d'assurés. Premièrement, connaître, comprendre et bien conseiller un assuré lors de sa souscription. MMA a la chance de posséder l'un des meilleurs réseaux d'agents de France pour gérer les flux entrants. Deuxièmement, il est important de connaître son portefeuille, de savoir quels assurés résilient et pour quelle raison.

L'objectif du mémoire est d'apporter des connaissances sur cet second aspect, à savoir, identifier les clients susceptibles de quitter son assureur, mais aussi, de comprendre quels facteurs influent sur cette décision. L'étude s'articule donc autour de deux grandes idées.

La première consistera à élaborer le modèle de prédiction de résiliations le plus performant possible. La modélisation classique réalisée à l'aide d'un Modèle Linéaire Généralisé sera comparée à celle plus complexe réalisée au moyen de modèles de Machine Learning comme le Gradient Boosting. L'étude des sorties des modèles de Machine Learning feront l'objet d'une analyse plus approfondie afin de pouvoir les utiliser opérationnellement.

La seconde partie aura pour but d'analyser la sensibilité des résiliations aux majorations tarifaires. Chaque année l'assureur est dans l'obligation d'augmenter ses primes afin de contenir la hausse de la sinistralité. Ce choix est central pour le pôle Tarification, l'objectif est alors d'apporter des arguments supplémentaires dans le choix et la répartition des revalorisations.

Mots clés : *MultiRisques Habitation (MRH), Résiliation, Politique Tarifaire, DataLab, H2O.ai, Modèles Linéaires Généralisés (MLG), Gradient Boosting Machine (GBM), Graphique de dépendance partielle, Valeurs de Shapley, Sensibilité au prix*

Abstract

Recently, the home insurance market has significantly changed, competition has been very keen. For an insurer, it is necessary to have an important knowledge of his insurance flows. First of all, it is important to know, understand and advise the insured person when he buys insurance policy. MMA is fortunate to have one of the best agents' networks that manage incoming flows in France. Secondly, it is important to watch our portfolio and to know which insured are cancelling and for what reason.

The purpose of this study is to bring knowledge on this second aspect, which means identify the customers that are susceptible to leave their insurer but also to understand what factors influence this decision. Therefore, this dissertation is structured around two main ideas.

The first idea will be to develop the most efficient model to predict the cancellation of a contract. The classical modeling performed with a Generalized Linear Model will be compared to more complex models made with machine learning such as Gradient Boosting. The outputs of the machine learning models will be further analyzed in order to use them operationally.

The second part will aim to analyze the sensitivity of cancellations to price increases. Each year, the insurer is obligated to increase its premiums in order to contain the increase in claims. This decision is central for the Pricing Department, and the purpose is to provide additional arguments for these increases.

Key words : *Household Insurance, Termination, Pricing Policy, DataLab, H2O.ai, Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), Partial Dependance Plot (PDP), SHAP Value, Sensitivity to price*

Remerciements

Durant mon année d'alternance, j'ai eu la chance d'être entouré par de brillantes personnes qui ont su me challenger et me rassurer lorsque la situation l'exigeait.

Pour cela, je tiens à d'abord à remercier chaleureusement mon tuteur en entreprise Mathieu RIOULT. Il a toujours su se montrer disponible afin de répondre à mes questions, m'accompagner et m'encadrer dans mes travaux. Travailler à ses côtés durant une année fût un plaisir. D'un point de vue professionnel, j'ai énormément progressé aux cotés de Mathieu qui est une personne très pédagogue et nous avons ainsi pu échanger sur de nombreux sujets même annexes au mémoire.

Je souhaite également remercier Ewen WILCZIK, manager de l'équipe Tarification MRH. Tout d'abord, pour la confiance qu'il m'a accordé mais aussi pour le choix du sujet, son aide pointilleuse et sa capacité à m'aider à prendre de la hauteur sur les différents sujets.

Je tiens aussi à remercier l'ensemble des membres de l'équipe Tarification MRH pour leur accueil et les réponses qu'ils ont pu m'apporter.

Je remercie l'équipe pédagogique et administrative de l'ISFA et plus particulièrement mon tuteur académique Nicolas LEBOISNE pour leur accompagnement.

Un mot tout particulier pour l'ensemble de mes camarades maintenant devenus amis, rencontrés lors de mon parcours académique, avec une mention spéciale pour Valentin CHANTELOUP avec qui j'ai pu fortement échanger sur la rédaction du mémoire. C'est une grande chance d'avoir des personnes de confiance avec qui échanger sur le plan professionnel.

Le mot de la fin est adressé aux personnes qui me supportent et soutiennent quotidiennement, je pense à mes parents Murielle et Mariano LOPEZ, mon frère Victor LOPEZ et mon amie Perla ALKHOURY.

Table des matières

Introduction

La crise économique et sanitaire a mis à mal certains départements du secteur assurantiel, notamment le domaine des assurances des professionnels, tandis que d'autres comme l'assurance Auto ont été épargnés, voir protégés. Durant le confinement, la sinistralité automobile a fortement chuté. En revanche, la pandémie a eu peu d'effet sur la sinistralité globale de l'assurance Multirisques Habitation. La généralisation du télétravail a engendré une baisse de la sinistralité vol et incendie, mais le bénéfice de cette diminution a été neutralisé par la forte hausse des matières premières et l'augmentation des sinistres climatiques. En seulement une année, le prix de certaines matières premières comme le PVC et l'acier a doublé. De plus, un rapport de la Caisse Centrale de Réassurance relève qu'entre 2010 et 2019, le coût annuel moyen des catastrophes naturelles a doublé, passant de 850 millions d'euros par an avant 2015 à 1.6 milliard d'euros annuel entre 2016 et 2019.

La pandémie a alors ajouté une pression supplémentaire dans un climat concurrentiel déjà important. Depuis plusieurs années, avec la mise en application de la loi Hamon et l'arrivée massive des bancassureurs sur le marché de l'habitation, il est désormais plus difficile de gagner et conserver des parts de marché. Covéa est leader sur le marché de l'assurance MultiRisques Habitation (MRH), donc dans une stratégie de défense du portefeuille ce qui rend l'étude des résiliations indispensable.

Le premier objectif de ce mémoire est de fournir un nouvel outil au service Tarification, Statistiques et Pilotage (TSP) MRH afin de mieux connaître les profils susceptibles de résilier. Cette équipe est en charge de tarifer et piloter les produits Habitation et Prévoyance commercialisés par la marque MMA. L'étude du taux de résiliation portera exclusivement sur le produit majeur habitation qui s'adresse aux propriétaires et locataires occupants leur bien, maison ou appartement. MMA a en portefeuille de nombreux assurés détenant plusieurs produits assurantiels, notamment sur le marché des entreprises. De plus, le réseau des agents est très proches des clients et possède de nombreux leviers à disposition. Ces deux facteurs sont intégrés dans le modèle et font l'objet d'une attention particulière.

Actuellement, seul un suivi du nombre global d'affaires nouvelles et d'affaires perdues est effectué. L'étude du taux de résiliation a donc pour objectif, d'une part, de révéler les profils les plus susceptibles de résilier et d'autre part, d'observer les conséquences d'une évolution tarifaire plus ou moins forte sur le portefeuille.

Sur un second plan, le mémoire doit aider le service TSP à majorer les primes des assurés. Cette étape est essentielle, car une majoration élevée ferait augmenter le chiffre d'affaires et les résultats à court terme, mais entraînerait la fuite des meilleurs clients. A l'inverse, une majoration trop basse engendrerait des résultats qui ne seraient plus à la hauteur des engagements auxquels doit faire face l'assureur. Le choix

du montant de ces majorations est central pour le client et pour MMA. La majoration de la cotisation doit absorber l'inflation des matières premières, la hausse des frais et l'évolution du risque tout en restant compétitive vis-à-vis d'un marché de plus en plus ouvert avec des nouvelles lois amplifiant les transferts d'assurés telle que la loi Hamon.

Le rapport de l'étude se décompose en quatre grandes parties :

- La première partie donne au lecteur une vision d'ensemble du contexte dans lequel est réalisé ce mémoire. Les différentes spécificités de l'assurance habitation et du processus de résiliation y sont détaillées.
- La deuxième partie traite de la création de la base de données ainsi que des environnements informatiques utilisés pour l'étude. Dans cette partie, un chapitre est dédié à la présentation et à l'analyse primaire des variables explicatives. Cette analyse permet d'aider au cadrage de la partie modélisation.
- La troisième partie expose l'ensemble des travaux menés sur la modélisation du taux de résiliation. Plusieurs types de modèles ont été testés, le Modèle Linéaire Généralisé (GLM) et le Gradient Boosting Machine (GBM). Le Gradient Boosting est un modèle qui présente de meilleures performances mais sa complexité rend son utilisation opérationnelle plus difficile. Diverses méthodes de comparaison et d'interprétation des modèles de Machine Learning sont présentés afin de pallier à ces difficultés.
- La dernière partie porte sur la création d'un outil de sensibilité au prix. Cet outil est une application possible du modèle de résiliation, son but est d'aider à aiguiller les majorations tarifaires.

Première partie

Contexte de l'étude

Chapitre 1

Périmètre de l'étude

Sommaire

1.1 Contrat d'assurance habitation

1.1.1 Les risques couverts

Le contrat d'assurance Multirisques Habitation permet de se couvrir contre les risques auxquels sont exposés le logement, mais aussi toutes les personnes qui y vivent. Ce contrat couvre les dommages aux biens, la Responsabilité Civile « vie privée » et la Responsabilité Civile en tant que propriétaire ou locataire de l'habitation. Les garanties les plus courantes sont :

- La garantie dégât des eaux : fuites d'eau, ruptures des conduites, débordements de canalisations, infiltrations ;
- La garantie dégât électrique : elle couvre les dommages résultant de la chute de la foudre ou d'une surtension subie par les appareils électriques mobiliers situés dans les bâtiments assurés.
- La garantie incendie : elle couvre les dégâts causés par le feu et la fumée, mais aussi ceux provoqués par les pompiers pendant leur intervention ;
- La garantie bris de glace : les éléments de séparation avec l'extérieur ou qui délimitent une pièce (portes, fenêtres, baies vitrées. . .) ;
- La garantie vol et vandalisme ;
- La garantie catastrophes naturelles : l'événement doit avoir fait l'objet d'une déclaration par arrêté interministériel pour entraîner un dédommagement ;
- La garantie climatique : Si l'évènement ne fait pas l'objet d'un arrêté, cette garantie permet de se couvrir contre les risques d'une tempête, de la grêle, de la neige et des coulées de boues ;

En complément de ces garanties, le contrat MRH inclut généralement une assurance de protection juridique et des garanties d'assistance.

1.1.2 Obligation légale

A la différence de l'assurance Auto, l'assurance MRH n'est pas obligatoire pour tout le monde.

Un propriétaire n'a aucune obligation légale à souscrire une assurance Habitation. La seule exception est le fait de couvrir un bien appartenant à une copropriété. Auquel cas, le propriétaire doit s'assurer à minima pour garantir sa responsabilité envers ses voisins, la copropriété et les éventuels locataires.

Un locataire, quant à lui, a l'obligation légale de s'assurer à minima pour les risques locatifs. L'assureur réglera donc au propriétaire, à la place du locataire, le montant des dommages dont il est responsable. Si le locataire n'est pas assuré mais responsable, il sera tenu d'indemniser personnellement les victimes. Le propriétaire peut exiger que le locataire lui remette une attestation d'assurance lors de la remise des clés. Il a aussi le droit d'insérer dans son contrat de location une clause de résiliation pour défaut d'assurance (Loi n° 89-462 du 6 juillet 1989).

L'assurance MRH est néanmoins très fortement recommandée au regard du préjudice subi en cas de sinistre important sur un logement non assuré. Selon un rapport de la FFA, le taux de couverture des résidences principales frôle le 100% de couverture. Le portefeuille est constitué à 90% de résidences principales, il n'y aura donc pas de biais entre propriétaires et locataires dans leur façon de percevoir le produit MRH. Tous les deux l'assimilent à un produit obligatoire comme l'assurance Auto.

1.2 Objectifs de l'étude

Depuis quelques années, la concurrence s'est accrue dans le domaine de l'assurance MRH. L'arrivée des bancassureurs bousculent les pratiques tarifaires en proposant l'assurance MRH en complément d'un crédit immobilier. Les bancassureurs gagnent de plus en plus de parts de marché ce qui force les assureurs classiques à être de plus en plus compétitifs dans leur tarification. On peut noter la seconde place en MRH, juste derrière Covéa, d'un bancassureur.

De surcroît, la progression de la sinistralité MRH nécessite une revalorisation annuelle de la prime. Cette hausse est notamment dûe à la multiplication du nombre de sinistres catastrophes naturelles et dégâts des eaux, mais aussi dûe à l'augmentation continue des coûts de construction.

Récemment, MMA a commercialisé une nouvelle offre MRH 410 plus compétitive au niveau tarifaire, mais qui permet de cibler des profils moins risqués et plus conventionnels. L'étude porte sur le produit majeur du pôle MRH : le produit 410, qui est dédié aux propriétaires et locataires occupants.

L'étude du taux de résiliation a pour ambition de répondre à plusieurs problématiques :

- Cibler les profils de risque susceptibles de résilier : bien les connaître permet d'adapter les stratégies de vente ou de mettre en place des actions pour mieux les conserver en portefeuille. Le multi-équipement du client est un élément important de la stratégie de MMA. En parallèle, la compagnie met en place de nombreux leviers de fidélisation du client basé sur le postulat que les dérogations et le multi-équipement fidélisent. Aucune étude sur les résiliations en assurance habitation n'a été menée, l'objectif est alors de cibler les critères impactant le taux de résiliation.
- Étudier la sensibilité au prix des assurés afin d'adapter et d'ajuster au mieux les mesures tarifaires. Lors des revalorisations, cette sensibilité au prix n'est jamais étudiée de façon approfondie

pour pouvoir cibler les profils susceptibles de résilier suite à une hausse plus ou moins importante du tarif.

Cette étude est l'occasion de mettre en exergue les profils les plus susceptibles de résilier et de confirmer ou réfuter différents postulats sur le comportement des assurés vis-à-vis des hausses tarifaires mais aussi des leviers mis en place par MMA.

1.3 Résiliation du contrat

Depuis l'entrée en vigueur de la loi Hamon en 2015, l'assuré se réserve le droit de résilier son contrat MRH à tout moment une fois la première échéance passée.

1.3.1 Résiliation par l'assuré

Le contrat d'assurance est en tacite reconduction, c'est-à-dire qu'il est reconduit automatiquement à chaque échéance. L'échéance est généralement annuelle, mise à part dans le cas où le contrat fixe une période plus courte, le plus souvent pour être reconduit au 1er janvier.

Depuis 2005, la loi Chatel oblige les assureurs à tenir informer, avec un avis d'échéance, la date limite à laquelle l'assuré peut empêcher la reconduction tacite de son contrat. Les assureurs doivent notifier à l'assuré, au plus tôt trois mois et au plus tard un mois avant la période autorisant le rejet de la reconduction. Si l'assureur ne se plie pas à cette règle, l'assuré est dans son droit de résilier à tout moment son contrat sans pénalité.

L'entrée en vigueur en 2015 de la loi Hamon a offert beaucoup plus de flexibilité dans la résiliation des contrats habitation. L'assuré n'est plus obligé d'attendre la date d'anniversaire de son contrat pour y mettre un terme. L'assuré fait part de son souhait de résiliation par courrier recommandé. La résiliation du contrat prendra effet un mois après réception et l'assuré sera remboursé de la prime au pro rata de la période restant à courir. Si l'assuré est locataire, il est dans l'obligation de contacter un autre assureur pour souscrire un nouveau contrat qui prendra le relais du contrat résilié.

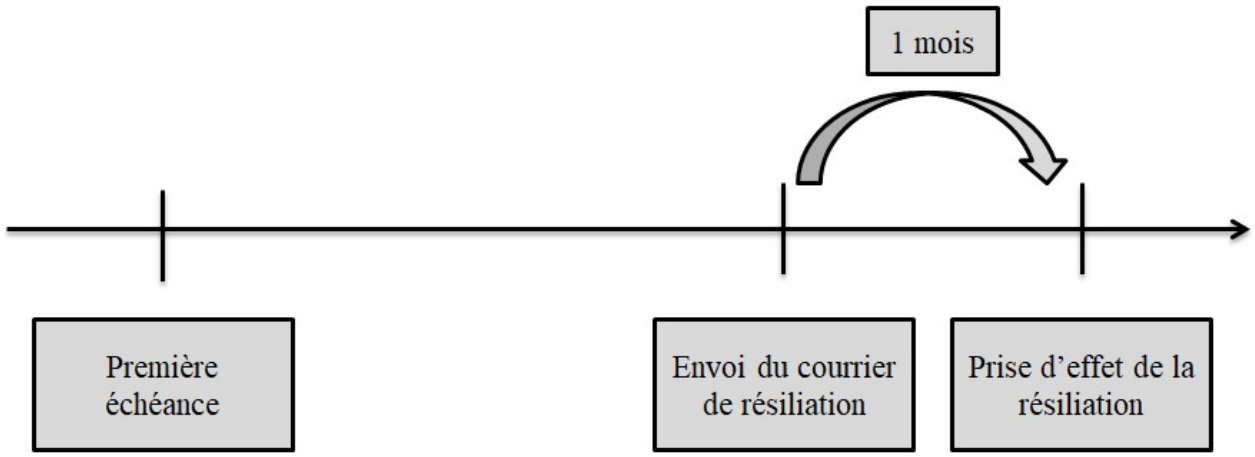

FIGURE 1.1: Résiliation dans le cadre de la loi Hamon

Cette nouvelle démarche de résiliation a également accru la concurrence entre les différents établissements d'assurance, car elle facilite grandement les transferts d'assurés. Une certaine périodicité au sein de l'année existe d'où l'importance d'étudier le taux de résiliation sur une année complète.

1.3.2 Résiliation suite à un changement de situation

L'assuré a la possibilité de résilier son contrat avant la première d'année d'assurance écoulée, en envoyant un courrier recommandé avec accusé de réception dans les cas suivants :

- Lorsque la situation personnelle de l'assuré change au point de modifier les risques couverts par le contrat d'habitation. L'assureur peut décider de conserver les mêmes conditions ou bien les modifier. Dans le second cas, l'assuré a la possibilité de ne pas les accepter et de résilier.
- Lorsque l'assureur résilie unilatéralement un autre contrat auquel l'assuré a souscrit, celui-ci peut résilier son contrat MRH.
- Lors d'un changement de bien : vente ou fin de location.
- En cas de décès de l'assuré, l'assurance habitation continue automatiquement. Les héritiers ont le choix de résilier ou non le contrat. Si les héritiers choisissent de laisser le contrat se poursuivre, ils doivent continuer à payer les cotisations. S'ils décident de résilier le contrat, ils doivent envoyer à l'assureur une lettre de résiliation par courrier recommandé. La résiliation prend effet un mois après la date de réception du courrier.

1.3.3 Résiliation du fait de l'assureur

L'assureur a la possibilité de mettre un terme au contrat en cas de faute ou manquement de l'assuré. En cas de non paiement et mise en demeure de règlement, l'assureur est en droit de résilier le contrat. Par ailleurs, si l'assureur constate une fausse déclaration ou une aggravation du risque non déclarée, il peut résilier le contrat.

L'assureur a également la possibilité de rompre le contrat habitation après un sinistre. Dans ce cas, il doit respecter un préavis de deux mois et doit informer l'assuré par courrier recommandé. Au sein de MMA, en cas de très mauvais résultats, certains assurés sont placés sous surveillance. Lors de ce processus, la compagnie lui applique diverses sanctions qui peuvent aller d'une simple hausse de la prime jusqu'à la résiliation de son contrat MRH.

Dans notre étude, seules les résiliations à l'initiative de l'assuré sont étudiées, les résiliations liées au système de surveillance sont retirées de l'étude.

Chapitre 2

Contexte TSP MMA

Sommaire

2.1 Politique tarifaire

Depuis des années, l'assurance MRH subit de fortes hausses tarifaires à cause de nombreux facteurs. Aujourd'hui, le facteur le plus marquant est la forte hausse des sinistres climatiques. On dénote également une augmentation significative des indices construction (FFB). Cette majoration, qui intervient à l'échéance du contrat pour l'assuré, est donc un enjeu majeur pour les assureurs MRH. La prime d'assurance est un levier central dans la souscription et la conservation d'un contrat d'assurance. C'est l'élément le plus important pour le client, mais il est aussi essentiel pour l'assureur qui doit maintenir un équilibre technique.

Chaque année, l'équipe TSP consacre les mois de juin et juillet à la présentation du dossier tarifaire. Ce dossier a pour objectif de présenter les résultats des années précédentes mais surtout de fixer les mesures tarifaires pour l'année suivante. La première partie de la présentation est consacrée aux études qui vont permettre d'expliquer certains résultats et prendre des décisions en conséquences. La seconde partie présente les mesures qui vont être appliquées à partir du 2 janvier de l'année suivante. On y retrouve les mesures portant sur le risque, les garanties et les zones qui forment la majoration tarifaire subie à l'échéance du contrat. Ce processus mené par le service TSP, s'effectue en concertation avec de nombreux acteurs dont la Direction Qualité Technique des Réseaux et la Direction Performance Économique.

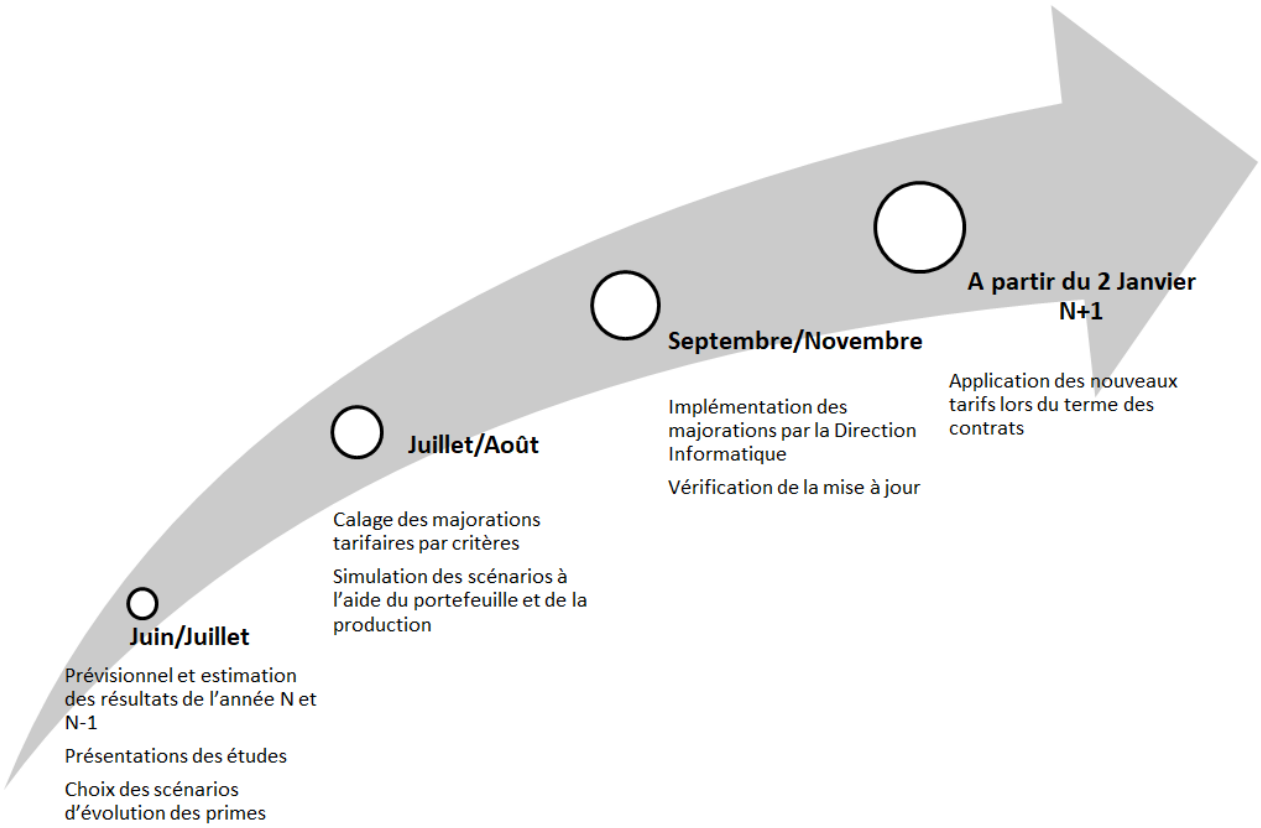

FIGURE 2.1: Processus de mise à jour tarifaire

Le calcul de la majoration à l'échéance du contrat se décompose en deux étapes :

- La mesure barème qui est la prime théorique calculée à partir des critères de l'habitation, des leviers tarifaires et du zonier. Cette cotisation barème est utilisée lors des affaires nouvelles et avenants, ainsi que pour la majoration à l'échéance du contrat. Pour les contrats déjà en portefeuille, cette cotisation sera encadrée par la mesure terme.
- La mesure terme est un encadrement de la mesure barème. Elle permet d'éviter une baisse de la prime mais aussi limite des hausses de prime trop importantes d'une année sur l'autre. Cette mesure est seulement appliquée sur les contrats déjà en portefeuille.

2.1.1 Processus des mesures barèmes

La mesure barème est l'élément clé qui offre la possibilité à MMA d'ajuster la rentabilité du produit mais aussi de discriminer certains types de biens lors de la majoration tarifaire. La Direction Performance Économique et le pôle TSP se concertent pour établir les objectifs de marge de l'année suivante. A partir de ce chiffre, le service TSP MRH segmente par critère, par garantie et par zone la majoration tarifaire à appliquer. A l'aide de l'ensemble des connaissances et des études menées durant l'année, l'équipe décide de plus ou moins augmenter le tarif de certains critères ou garanties. Par exemple, certains critères sont catégorisés comme plus risqués au vue de leurs résultats dégradés. Ces critères aggravants subissent des majorations tarifaires plus importantes, d'une part, pour rééquilibrer le ratio combiné et d'autre part, pour éviter de souscrire ces types de risques catégorisés comme mauvais. L'objectif est donc de segmenter assez finement les hausses tarifaires afin d'éviter l'anti-sélection.

Appartement		Mesure	% PTF	Impact
Évolution de base	Dégâts des eaux et Vol TGN, Bris de vitre* et RCVP Incendie Autres garanties	+4,0% +3,0% +2,0% +1,0%	50%	+3,0%
Impact Appartement				+3,0%

Maison				
Évolution de base	Incendie TGN Piscine Dégâts des eaux et Bris de vitre* Autres garanties	+3,0% +4,0% +6,0% +2,0% +2,0%	50%	+2,4%
Isolement		+1,5%	10%	+0,30%
Nombre de pièces	de 7 à 30 pièces	de +1% à +3%	5%	+0,30%
Impact Maison				+3,0%

FIGURE 2.2: Tableau fictif des mesures barèmes

2.1.2 Processus des mesures termes

La mesure terme encadre la majoration tarifaire calculée lors de l'étape précédente. Cette mesure protège les assurés déjà en portefeuille contre une hausse trop significative de leur prime habitation. Par exemple, une habitation qui cumule différents critères jugés aggravants peut, à l'issue de la mesure barème, se retrouver avec une hausse importante de la prime. Si le client est non sinistré, cette hausse sera modérée par l'encadrement fixé.

Cette mesure terme impose un traitement différent entre les différents types de clients. Les clients avec de mauvais résultats sont peu protégés par cet encadrement, à l'inverse des clients catégorisés comme rentables.

	Client Type 1		Client Type 2		Client Type 3		Client Type 4	
Groupes à l'échéance	Mini	Maxi	Mini	Maxi	Mini	Maxi	Mini	Maxi
Propriétaire de maison sans sinistre 12 mois								
Locataire de maison sans sinistre 12 mois								
Propriétaire Appartement sans sinistre 12 mois								
Locataire Appartement sans sinistre 12 mois								
Propriétaire de maison avec sinistre 12 mois								
Locataire de maison avec sinistre 12 mois								
Propriétaire Appartement avec sinistre 12 mois								
Locataire Appartement avec sinistre 12 mois								
	4,0%		3,0%		2,0%		1,5%	
	3,0%							

FIGURE 2.3: Tableau fictif des mesures termes

2.2 Notion d'élasticité au prix

Un des objectif de cette étude est de réussir à capter la sensibilité au prix des différents profils de nos assurés afin d'ajuster nos mesures tarifaires. Pour cela, nous utiliserons l'élasticité du taux de résiliation au prix. Cet indicateur mesure la sensibilité de résiliation par rapport à son prix. Plus exactement, c'est un indicateur de la réaction du taux de résiliation suite à une variation de 1% du montant de la cotisation.

$$\frac{\frac{\delta T_x(p)}{T_x(p)}}{\frac{\delta p}{p}} \tag{2.1}$$

Mesurer la sensibilité nous permettra de mieux connaître le comportement des assurés en fonction du prix et ainsi mieux cibler le profil pouvant supporter ou non une majoration tarifaire.

Deuxième partie

Environnement et base de données

Chapitre 3

Construction de la base de données

Sommaire

La construction et la finalisation de la base de l'étude se sont réalisées en deux grandes étapes. La première a été de travailler sous l'environnement SAS afin d'utiliser les bases mises à disposition pour l'équipe TSP MRH. L'objectif est d'obtenir une base très étoffée comportant un très grand nombre d'informations sur les clients assurés avant de la transférer sous un autre environnement : le DataLab. Le but est d'éviter les aller-retours coûteux entre les deux environnements. Dans un second temps, cette base sera affinée sur ce nouvel environnement.

3.1 Base socle de l'étude

Tout d'abord, il est important de rappeler que la donnée est au coeur du métier d'actuaire. Elle constitue le socle de toutes les études et participe grandement à leur qualité. Afin de répondre aux enjeux de la Data, MMA possède une équipe dédiée à la création et maintenance des bases de MMA IARD. De nombreuses bases contenant chacune des informations sur les contrats IARD et plus précisément MRH sont mises à notre disposition.

L'ensemble des opérations effectuées par le service TSP MRH est réalisé sous SAS Enterprise Guide. L'équipe MRH se charge ensuite de concaténer ces différentes bases au sein d'un même Datamart MRH mis à jour trimestriellement. Cette base contient les informations des situations d'assurances sur 4 ans. Dans le cas où le client a de l'ancienneté, celui-ci possède plusieurs situations réparties sur les exercices étudiés. Ce Datamart va constituer le socle de notre base d'étude. Les caractéristiques de l'extraction de celui-ci sont :

- 4 années d'historique : 1er janvier 2017 - 31 décembre 2020 ;
- 12 millions de lignes ;
- 430 colonnes ;
- Informations sur la situation (Type de bien, qualité juridique, montant cotisation, garanties souscrites ...) ;

— Informations sur la sinistralité (Type de sinistre, montant du sinistre ...) ;
— Informations sur le client (Numéro, nom, prénom, équipement, cible marketing ...) ;
— Informations cartographie / INSEE (Commune, IRIS « Ilots Regroupés pour l'Information Statistique » , zonier ...) ;

Dans le but de répondre à un large panel de demandes, le Datamart coupe au 31 décembre les lignes de chaque situation afin d'avoir les résultats par année civile. Dans le cadre de notre étude, cette ligne supplémentaire n'est pas nécessaire. Les deux lignes de chaque année civile sont fusionnées pour ne conserver qu'une ligne par situation. Sur l'exemple ci-contre, les lignes sont concaténées deux à deux.

NU_AFFA	DT_DEBU_SITU_RCCL	DT_FIN_SITU_RCCL	CD_ADHE	NATU_SITU	MOTI_FIN_SITU	MT_COTISATION	SSAA_ISO
100007091	01MAY2017	31DEC2017	410	TER	TER	654 €	2017
100007091	01JAN2018	30APR2018	410	TER	TER	654 €	2017
100007091	01MAY2018	31DEC2018	410	TER	AVT	671 €	2018
100007091	01JAN2019	30APR2019	410	TER	AVT	671 €	2018
100007091	01MAY2019	31DEC2019	410	AVT	TER	694 €	2019
100007091	01JAN2020	30APR2020	410	AVT	TER	694 €	2019
100007091	01MAY2020	31DEC2020	410	TER	#	735 €	2020

FIGURE 3.1: Datamart 2 lignes situation

NU_AFFA	DT_DEBU_SITU_RCCL	DT_FIN_SITU_RCCL	CD_ADHE	NATU_SITU	MOTI_FIN_SITU	MT_COTISATION	SSAA_ISO
100007091	01MAY2017	30APR2018	410	TER	TER	654 €	2017
100007091	01MAY2018	30APR2019	410	TER	AVT	671 €	2018
100007091	01MAY2019	30APR2020	410	AVT	TER	694 €	2019
100007091	01MAY2020	31DEC2020	410	TER	#	735 €	2020

FIGURE 3.2: Datamart 1 ligne situation

A la suite de ces retraitements le Datamart ne compte plus que 6.5 millions de lignes.

3.2 Informations supplémentaires

Dans le but d'améliorer notre connaissance des profils susceptibles de résilier, nous allons venir greffer des informations issues d'autres bases et ainsi créer de nouveaux indicateurs. De plus, plusieurs données vont être calculées pour disposer de données dynamiques comme l'évolution tarifaire et l'évolution du nombre de contrats. L'objectif est d'obtenir une base regroupant beaucoup d'informations sur la situation, mais aussi sur la vie du contrat afin de tester leur pouvoir explicatif par la suite.

3.2.1 Datamart Production

La première opération consiste à fiabiliser la base issue du Datamart MRH en la rapprochant à la base production. Cette base production nous permet d'obtenir des informations sur toutes les opérations

réalisées entre deux dates : Affaires Nouvelles (AN), Affaires Perdues (AP), avenants, devis ... Dans notre cas, seules les opérations concernant les Affaires Perdues entre le 1er janvier 2017 et le 31 décembre 2020 sont rapprochées au Datamart MRH. Le motif de résiliation est également rapproché. Cependant, certaines Affaires Perdues ne sont pas rattachées à une situation. Un travail de fiabilisation de la donnée a dû être effectué en plusieurs étapes :

1. Récupérer l'ensemble des Affaires Perdues (AP) sur la période 2017-2020. Un total de 608k affaires ont été perdues sur cette période.
2. Rapprocher ces AP avec le Datamart situation MRH. Lors de ce rapprochement, 584k AP ont bien été rapprochées à leur ligne situation. Un total de 24k AP restent à rapprocher.
3. Extraire les AP non apparentées à une situation. Certaines AP sont des annulations d'AN ce qui signifie que ces AP ne peuvent pas être rattachées à une ligne situation car aucune situation n'a été enregistrée pour ce contrat. De plus, de nombreuses AP ont pour date d'effet le 1er Janvier 2017, donc considérées comme des résiliations de situation de 2016. Nous avons alors rapproché les AP sur la période du 2 janvier 2017 au 1er janvier 2021 inclus (cf. figure 3.3). L'Affaire Perdue enregistrée le 1er janvier 2021 est comptabilisée comme une AP de 2020. Au stade de cette étape, seules 3k AP ne sont pas rattachées à une ligne situation.

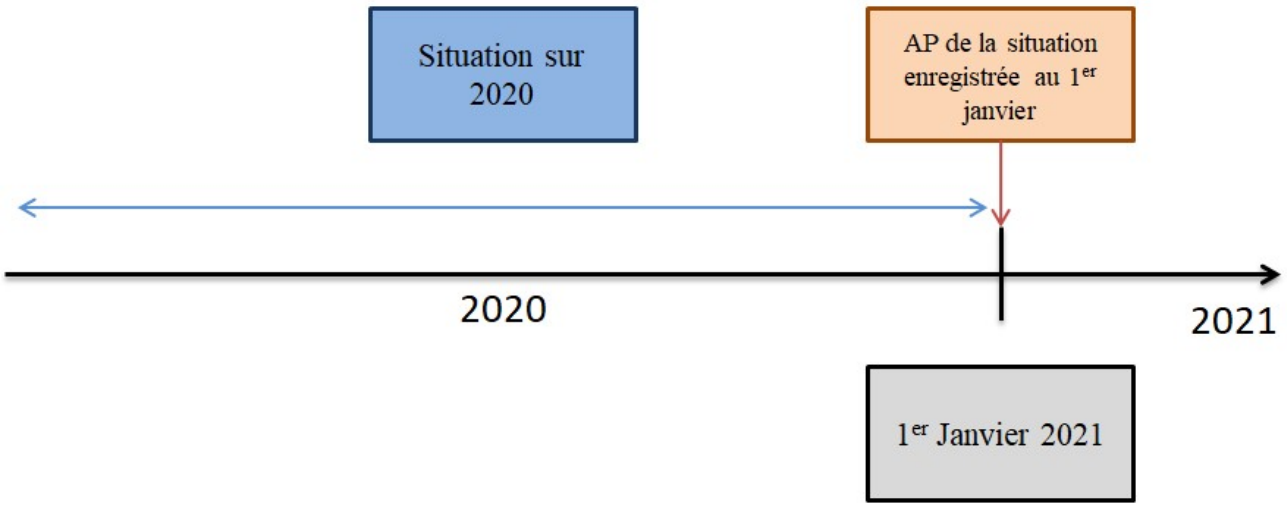

FIGURE 3.3: Gestion des AP déclarées au 1er Janvier 2021

4. Vérifier les 3k AP non apparentées à une ligne qui ne sont pas des AN annulées. Ces AP possèdent généralement une ligne situation toujours d'actualité, nous considérerons alors ces 3k AP comme des erreurs de saisi et par conséquent des situations non résiliées.

3.2.2 Base évolution

L'objectif de cette base est d'apporter des informations antérieures à la situation que nous allons observer. Chaque situation, si l'historique le permet, se voit attribuer des informations de 4 situations antérieures. Par exemple, une situation entre 2017 et 2018 se voit lui rapprocher des informations des années 2014-2015-2016-2017. Les informations suivantes sont incluses dans la base :

— L'évolution des dérogations tarifaires ainsi que leur montant ;

— La présence d'un avenant ;
— Le nombre et le coût des sinistres ;
— L'évolution de la cotisation ;

La variable d'évolution de la cotisation indique la hausse tarifaire subie lors du terme. Dans notre étude, nous nous affranchissons de l'effet avenant, l'indicateur d'évolution est calculé uniquement sur l'effet terme. Un historique de 4 évolutions est calculé, pour chaque situation les évolutions n, n1, n2, et n3 sont calculées si possible. Dans le cas contraire, la modalité prend la valeur NR ce qui signifie qu'elle n'est pas renseignée.

Dans l'exemple ci-dessous (cf figure 3.4), l'évolution de la cotisation est calculée à chaque ligne, mais seules les évolutions en vert sont conservées. Pour la dernière situation de ce contrat :

— L'évolution tarifaire n = 6.0%.
— L'évolution tarifaire n1 = 2.5%. L'évolution de 3.4% liée à l'avenant n'est pas comptabilisée pour la variable d'évolution tarifaire. C'est la précédente qui est alors affectée.

NU_AFFA	DT_DEBU_SITU_RCCL	DT_FIN_SITU_RCCL	CD_ADHE	NATU_SITU	MOTI_FIN_SITU	MT_COTISATION	EVOL_COT	SSAA_ISO
100007091	01MAY2017	30APR2018	410	TER	TER	654 €		2017
100007091	01MAY2018	30APR2019	410	TER	AVT	671 €	2,5%	2018
100007091	01MAY2019	30APR2020	410	AVT	TER	694 €	3,4%	2019
100007091	01MAY2020	31DEC2020	410	TER	#	735 €	6,0%	2020

FIGURE 3.4: Création de la variable évolution tarifaire

3.2.3 Base client

Cette base a pour finalité d'étudier le contexte global de l'assuré MRH au sein de MMA. Nous allons raccorder à la base de notre étude des informations sur le nombre de contrats détenu par le client sur d'autres produits MMA : Auto, Protection Juridique, Pro, mais aussi son nombre de contrats MRH. Un assuré possédant plusieurs biens peut détenir plusieurs contrats MRH. Afin d'obtenir l'évolution du nombre de contrats, quatre années d'historique sont rattachées à la ligne situation de notre base d'étude.

3.2.4 Indicateurs supplémentaires

Afin de compléter et étendre notre connaissance sur les biens assurés quelques indicateurs supplémentaires sont intégrés à notre base d'étude :

— Indicateurs de rentabilité à l'iris ;
— Le niveau des zoniers appliqué ;
— Le nombre d'agences MMA à proximité. Cet indicateur a été créé en croisant les coordonnées X et Y de l'habitation avec ceux des agences MMA ;

Chapitre 4

Le DataLab

Sommaire

4.1 Présentation de l'environnement

4.1.1 Contexte

SAS est un magnifique outil de manipulation, de formatage, d'extraction et de visualisation de tableaux de données, néanmoins, ce logiciel n'est pas conçu pour la modélisation avancée. En effet, seuls les modèles linéaires sont disponibles et interprétables sous ce logiciel chez Covéa.

Aujourd'hui, les langages les plus adaptés à la création de modèles en statistiques plus ou moins avancés sont R et Python. Ces deux outils, en licence libre, offrent la possibilité aux utilisateurs de partager et de collaborer efficacement en bénéficiant de solutions déjà créées par d'autres utilisateurs.

La suite des travaux se déroule sous Python. Il offre la possibilité de réaliser toutes sortes d'opérations grâce aux modules mis à disposition par Python ou des utilisateurs.

Depuis quelques années, Covéa développe une nouvelle interface nommée DataLab qui permet notamment l'utilisation de Python. La force de cet environnement réside dans l'utilisation de serveurs dédiés. Plus précisément, les calculs ne sont pas effectués à l'aide du processeur de notre machine personnelle, mais sur des machines beaucoup plus performantes dédiées à cet effet. La taille conséquente de notre base de données, presque 15Gb, nécessite l'utilisation de ces performances supérieures.

4.1.2 Parallélisation des calculs

L'accès au serveur du Datalab optimise les temps de traitement à l'aide de la parallélisation des calculs. Généralement, Python utilise du calcul séquentiel, c'est-à-dire que le programme exécute les instructions étape par étape, que les opérations soient indépendantes ou non. Le DataLab donne accès à un environnement Sparkling Water. Sparkling Water permet aux utilisateurs de combiner les algorithmes

d'apprentissage automatiques rapides et évolutifs de H2O avec les capacités de Spark. Les tables de données sont distribuées en mémoire favorisant la parallélisation des calculs.

H2O donne accès à des modèles d'apprentissages plus élaborés, rapides et open-source. Avec H2O, il est possible d'obtenir des résultats très rapidement avec la parallélisation des calculs. Des algorithmes avancés tels que l'apprentissage profond, le Boosting et le Bagging sont disponibles sous H2O.

4.2 Implémentation de la base dans le DataLab

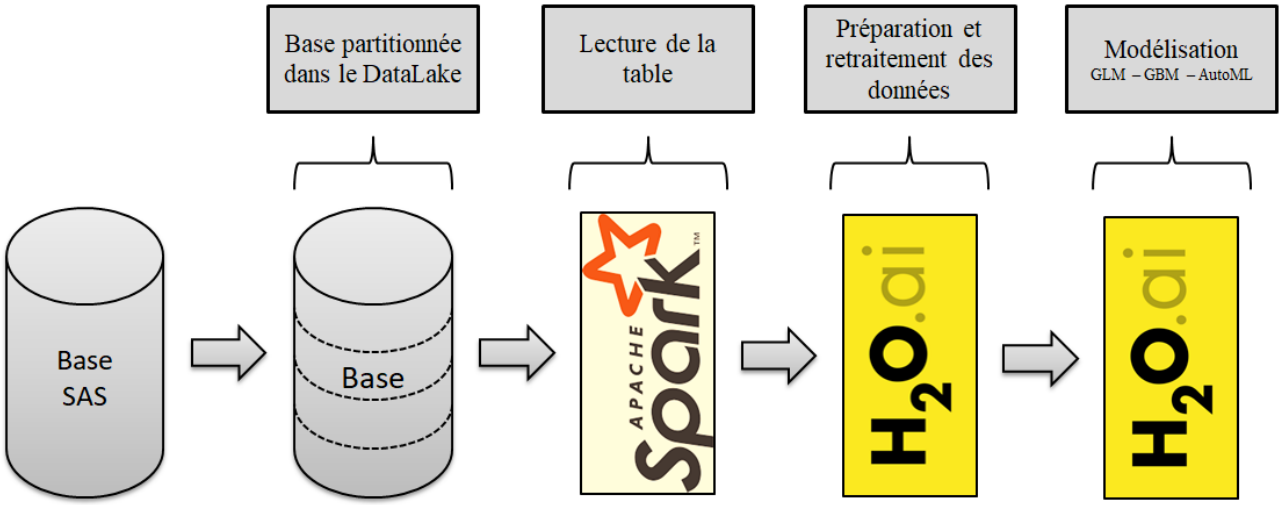

FIGURE 4.1: Processus d'utilisation de l'environnement du DataLab

4.2.1 Lecture de la base d'étude

L'outil Spark lit la table stockée dans le DataLake que nous transformons directement en DataFrame H2O. Sur les 430 colonnes, seulement 170 sont lues. La majeure partie des colonnes donnent des informations très spécifiques. Par exemple, les niveaux de chaque zonier, la part de la population entre deux âges, les indicatrices de chaque garantie souscrite, seules les plus importantes sont conservées. De plus, nombre de ces colonnes ont servi à la création d'autres. Le tableau de données H2O contient alors différents types d'informations repartis sur environ 6.5 millions de lignes pour et colonnes :

- Les informations classiques de la situation (Numéro d'affaire, exposition, dates de situation, numéros de produit et d'évènement ...) ;
- Les informations sur le bien assuré (Type, qualité juridique, nombre de pièces, dépendance, critère isolement ...) ;
- Les informations tarifaires (Montant de la cotisation, Bonus-Malus, montant des leviers et dérogation ...) ;
- Les informations sur les garanties souscrites (Formule, présence des garanties ou non, niveau souscrit ...) ;
- Les évolutions de situation sur 4 ans (Tarifaire, leviers, avenants, nombre de contrats MRH, Auto et Pro ...) ;
- Les informations à l'iris (Zoniers, nombre de résidences principales, score à l'iris) ;

— Les informations client (Tranche d'âge, génération du contrat ...) ;
— Les informations sur les résiliations (Résiliation ou non, date de résiliation) ;
— Les informations sinistres (Nombre et coût des sinistres de la situation ainsi que sur les 4 dernières années) ;

Avant de débuter l'étude, certaines informations sont retraitées ou supprimées. En effet, notre base de données balaye un grand nombre de profils différents qui ne nécessite pas tous d'être inclus dans l'étude. Les profils supprimés sont :

— Les clients qui possèdent un grand nombre de contrats MRH. Certains de nos assurés sont des associations ou des organismes détenant un nombre conséquent de bâtisses. Tous les clients détenant plus de 10 contrats MRH sont retirés pour une suppression totale de 30k lignes.
— L'ensemble des résiliations à l'initiative de la compagnie, c'est-à-dire celles liées au système de surveillance sont retirées de la base, 3.5k lignes sont supprimées.

Nous avons fait le choix de conserver les résiliations liées au décès. Nous considérons le décès comme un fait de vie qui engendre la disparition du risque. Ce motif joue un rôle non négligeable dans le taux de résiliation, en effet sur ces 4 années près de 2% des résiliations sont liées à un décès.

Les contrats étudiants et très récents (moins d'un an) sont également conservés. De notre point de vue, ils font intégralement partie du portefeuille et représentent un risque supplémentaire de résiliation à ne pas négliger.

4.2.2 Choix des variables d'étude

Pour étudier le taux de résiliation, l'ensemble des 170 colonnes ne sont pas nécessaires. Une première sélection à dire d'expert a été réalisée. 49 variables explicatives pour l'étude ont été retenues.

Plusieurs de ces variables regroupent en partie des informations de nombreuses autres colonnes. Par exemple, la variable FORMULE, synthétise l'information donnée par la présence ou non de certaines garanties. Pour compléter cette variable FORMULE, la variable NB_OPTION, qui dénombre les renforts souscrits en plus de la formule, est créée. De plus, certaines données sont mal renseignées ou bien jugées trop spécifiques pour être incluses dans l'étude. Enfin, nous avons décidé de ne conserver que la dernière année d'évolution pour les variables nombre de contrats, multi-équipement et leviers tarifaires.

Ces 49 variables retenues sont regroupées en 5 catégories :

— 8 variables d'informations sur le bien assuré ;
— 11 variables sur l'information tarifaire ;
— 16 variables sur l'environnement tarifaire ;
— 7 variables sur la sinistralité ;
— 7 variables classées dans une catégorie autre ;

Info	Nom de la variable	Description de la variable
Habitation	CD_TYPE_HABI	Type du logement
	CD_QLTE_ASSU_HABI	Qualité juridique de l'habitant
	NB_PIEC	Nombre de pièces
	CD_USAG_RISQ	Résidence principale ou secondaire
	NB_PIEC_SUPE	Nombre de pièces supérieures à 40m²
	CD_ISOL	Critère isolement
	flag_DPDC	Présence d'une dépendance
	Taille_DPDC	Taille de la dépendance
Tarifaire	cot_net_n	Montant de la cotisation nette
	cot_bar_n	Montant de la cotisation barème
	flag_surtarif	1 si la cotisation nette est supérieure à la cotisation barème
	evol_net_fin_n	Evolution annuelle de la cotisation nette entre la situation actuelle et la précédente
	evol_net_fin_n1	Evolution annuelle de la cotisation nette entre la situation n-1 et n-2
	evol_net_fin_n2	Evolution annuelle de la cotisation nette entre la situation n-2 et n-3
	evol_net_fin_n3	Evolution annuelle de la cotisation nette entre la situation n-3 et n-4
	evol_net_fin_n1-n3	Evolution de la cotisation nette sur 3 ans entre la situation n-1 et n-4
	VA_COEF_BM_MRH	Bonus-Malus MRH
	evol_levi_n	Evolution du nombre de leviers tarifaires par rapport à la situation précédente
	MT_DRGT_MBA_TTC	Montant de dérogation MBA appliqué par les agents généraux
Environnement commercial	nb_affa_auto_n	Nombre d'affaires AUTO du client
	evol_AUTO	Evolution du nombre d'affaires par rapport à la situation précédente
	nb_affa_mrh_n	Nombre d'affaires MRH du client
	evol_MRH	Evolution du nombre d'affaires par rapport à la situation précédente
	nb_pro_ent_n	Nombre d'affaires PRO du client
	DETENTION_n	Multi équipement du Client
	FORMULE	Formule souscrite en MRH
	nb_option	Nombre de renforts souscrit en MRH
	CAPI_MOBI	Niveau de capital mobilier souscrit
	nb_1km	Nombre d'agences MMA à moins de 1km de l'habitation
	nb_2km	Nombre d'agences MMA à moins de 2km de l'habitation
	nb_5km	Nombre d'agences MMA à moins de 5km de l'habitation
	nb_10km	Nombre d'agences MMA à moins de 10km de l'habitation
	nb_20km	Nombre d'agences MMA à moins de 20km de l'habitation
	nb_30km	Nombre d'agences MMA à moins de 30km de l'habitation
	SEGM_PART	Segmentation personnalisée des clients en fonction de leur tranche d'âge et de l'unité urbaine.
Sinistralité	nb_sin_n1	Nombre de sinistres survenus pendant la situation
	nb_sin_4ans	Nombre de sinistres survenus durant les 4 dernières situations
	COUT_OBS_n1	Coût observé des sinistres de la situation
	cout_sin_4ans	Coût observé des sinistres des 4 dernières situations
	NB_SINI_OUV	Nombre de sinistres ouverts
	CD_CTGR_SINI_ANTE	Type de sinistre survenu
	MAJO_SINI	En fonction du type et nombre de sinistre une majoration barème est appliquée.
Autres	Generation	Nombre d'année du contrat MRH
	NATU_SITU	Nature du début de situation
	Flag_avt_n	1 si la situation à débutée à cause d'un avenant
	TARIF_B100	Tarif moyen par iris basé sur 6 cas fictifs
	SCORE	Score de rentabilité par iris obtenu avec la part de marché, la densité et de l'évol ptf
	EVOL_BAR_1AN	Evolution barème de l'iris
	NBP_MOY_MAISON_RP	Score de résidences principales par iris

FIGURE 4.2: Liste des 49 variables

Chapitre 5

Premières analyses

Sommaire

5.1 Analyse macro du portefeuille

Le portefeuille MRH MMA est stable depuis quelques années, mais la hausse de la prime moyenne fait croître le chiffre d'affaires de plus de 1% par an en moyenne. Sur le marché habitation, MMA est un acteur important avec 4.2% des parts de marché, et plus globalement, Covéa est le leader du segment avec 18% des parts du marché.

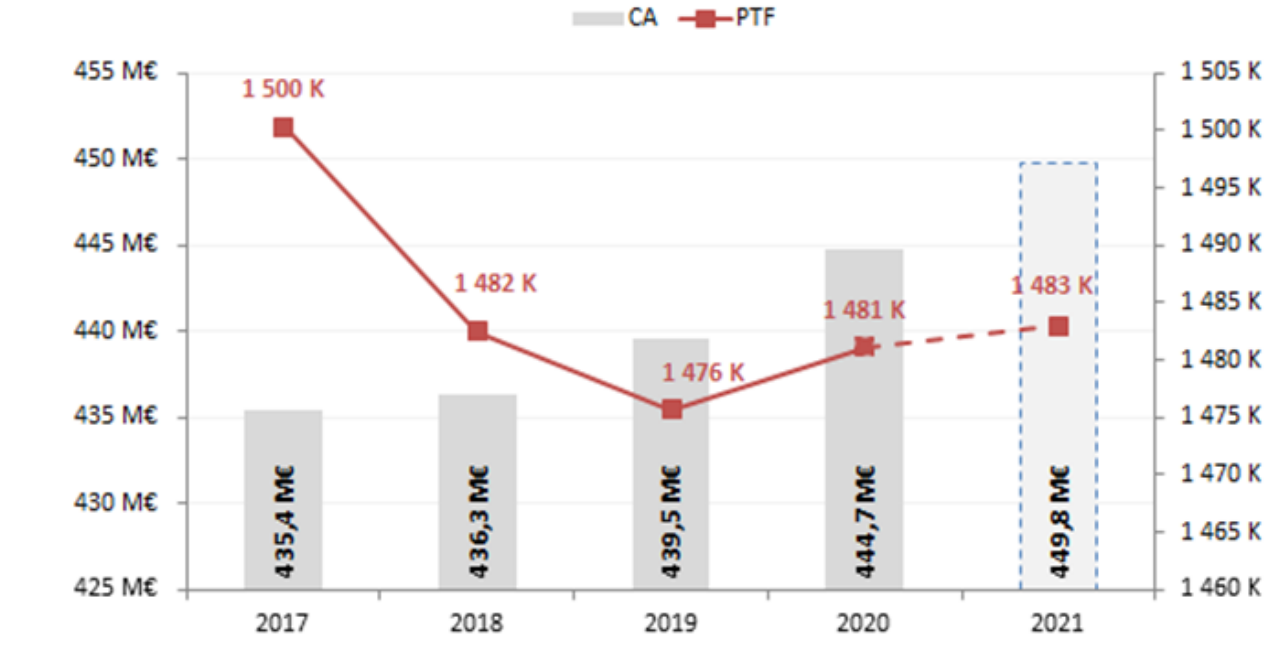

FIGURE 5.1: Évolution du portefeuille et du CA MRH

Cette force est également une faiblesse car il est plus difficile de récupérer des parts de marché et inversement, plus facile d'en perdre. Le suivi du portefeuille est donc un point essentiel pour le service TSP MRH. Il existe deux grands leviers pour faire croître le portefeuille. D'une part, dynamiser la production et d'autre part, mieux maîtriser les affaires perdues.

Du fait de la situation particulière avec les restrictions imposées, cette année 2020 a vu le nombre d'AN diminuer de 5% et le nombre d'AP diminuer de 12%. Cette variation reste toutefois légère, car un rattrapage a eu lieu à la suite du premier confinement. Les personnes souhaitant résilier ont décalé leur souhait de quelques mois. Malgré l'atypisme de cette année, l'année 2020 est conservée dans l'étude, car la baisse semble homogène sur l'ensemble des profils de risque.

5.2 Analyse des variables de la base

Dans cette partie, nous allons étudier le comportement du taux de résiliation en fonction des différentes variables explicatives majeures. Les graphiques comportent deux axes qui offrent deux types d'informations chacun :

- La graduation majeure de l'axe des abscisses renseigne les modalités des variables explicatives.
- La graduation mineure de l'axe des abscisses représente l'année.
- L'axe n°1 des ordonnées indique l'exposition de la modalité en Risque Année (RA).
- L'axe n°2 des ordonnées représente le taux de résiliation.

5.2.1 Taux de résiliation annuel

Pour chacune des situations, une variable présence de résiliation ainsi que sa date sont indiquées. L'étude porte sur les 4 dernières années complètes, de 2017 à 2020. Ce premier graphique représente l'évolution du taux de résiliation annuel tous critères confondus ainsi que la moyenne de ce taux sur 4 ans. L'exposition en Risque Année (RA) est représentée par les histogrammes.

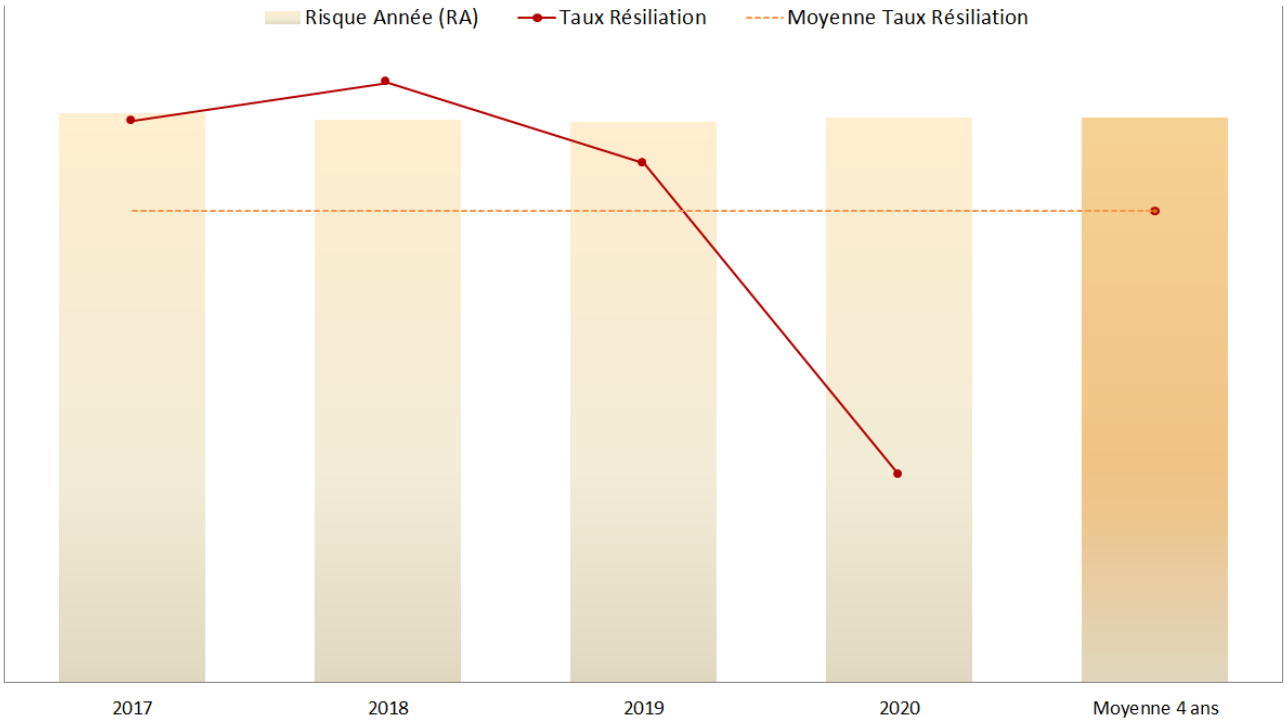

FIGURE 5.2: Taux de résiliation annuel

La moyenne du taux de résiliation sur cette période est légèrement diminuée par l'année atypique 2020. Les années 2017, 2018 et 2019 ont un taux de résiliation plus fort que l'année 2020. Nous avons toutefois conservé l'année 2020 dans cette étude. Le principal objectif est d'identifier les profils qui ont une tendance forte à résilier et mesurer leur sensibilité. Or, nous constatons sur les graphiques suivants une baisse homogène sur l'ensemble des profils du portefeuille. Pour chaque modalité, le taux de résiliation de l'année 2020 sera inférieur à celui de 2019. Cette baisse de résiliations sur l'année 2020 n'influe donc pas sur les profils qui résilient.

5.2.2 Variables explicatives univariées

Type de bien et qualité juridique de l'occupant : Les deux critères les plus utilisés en assurance habitation sont le type de bien : Maison, Appartement et Mobil-Home ainsi que la qualité juridique de l'occupant : Propriétaire et Locataire. Ces deux variables sont croisées afin d'étudier l'impact de chacune d'entre elle sans être biaisée par l'autre.

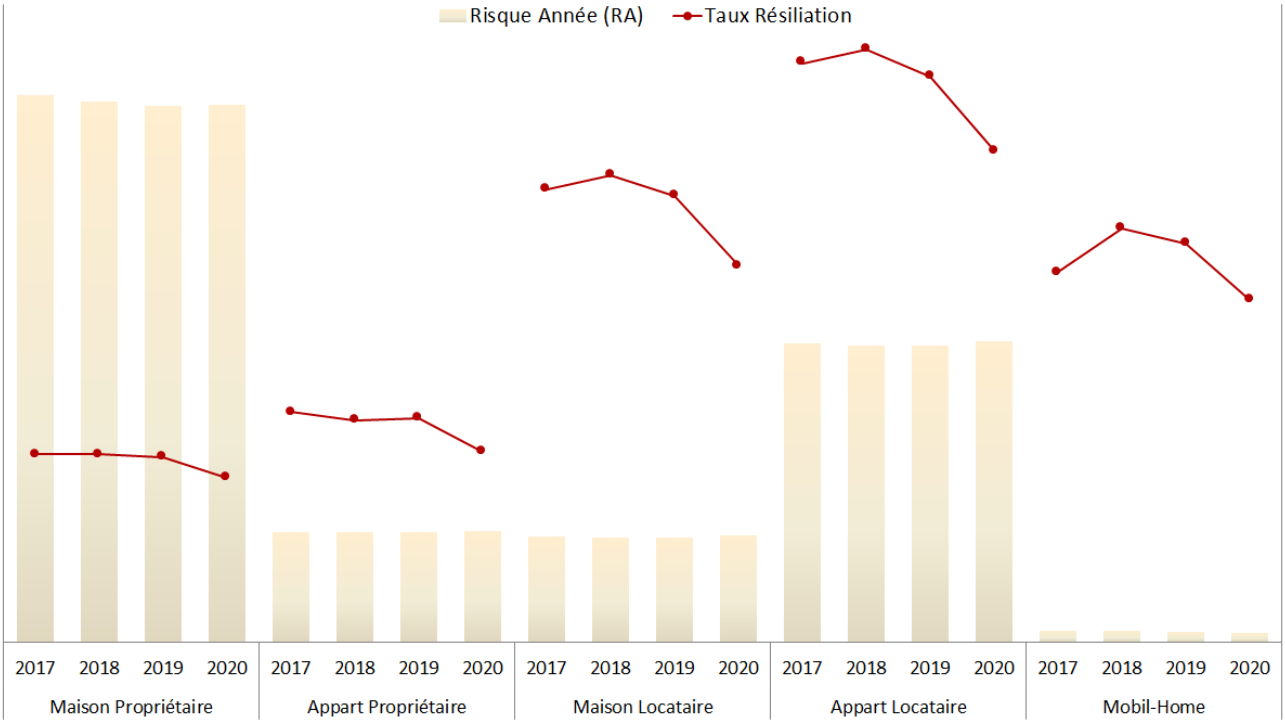

FIGURE 5.3: Taux de résiliation en fonction du type de bien et de la qualité de l'occupant

La première remarque à noter est la baisse des taux de résiliation sur l'ensemble des profils sur l'année 2020. Cette baisse est homogène, car les 4 premiers profils ont tous vu leur nombre de résiliations chuter de 11% à 14% entre 2019 et 2020. La baisse du nombre de résiliations sur les mobile-homes est légèrement supérieure, autour de 18%, mais n'a qu'un poids infinitésimal sur le portefeuille.

La qualité juridique a un impact très important sur le taux de résiliation. Le taux de résiliation des deux premières modalités : Maison Propriétaire et Appartement Propriétaire est très faible par rapport au taux des locataires qui est plus de deux fois supérieur. En effet, le taux de turnover est réputé pour être plus important sur les locations.

A contrario, le type de bien n'impacte que faiblement le taux de résiliation. Cette information va à l'encontre des préjugés. Lorsque nous observons un écart important entre maison et appartement, cet écart est majoritairement dû à la part importante de propriétaires en maisons et de locataires en appartements.

Nombre de pièces : Ce critère renseigne le nombre de pièces de l'habitation. Il est couramment utilisé pour identifier les biens atypiques de très grande taille.

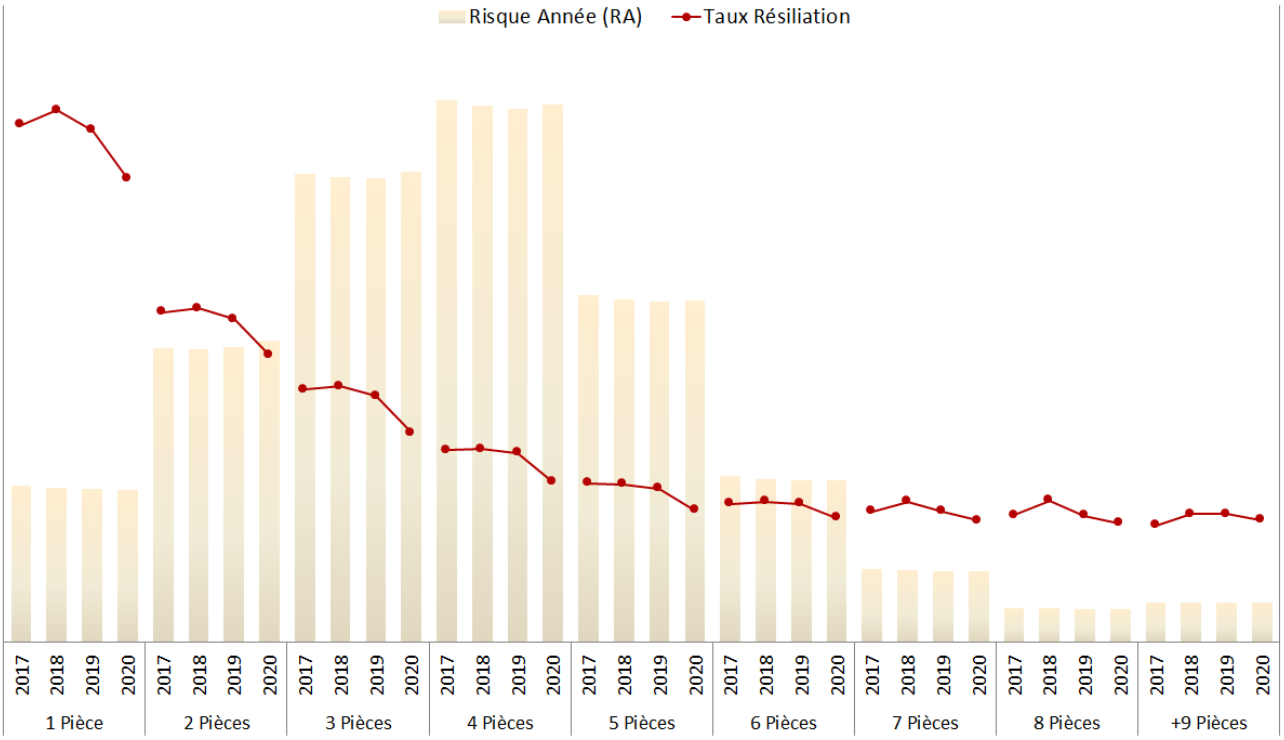

FIGURE 5.4: Taux de résiliation en fonction du nombre de pièces

Le taux de rétention est meilleur sur les biens de grande taille. Généralement, ces biens atypiques sont occupés par des propriétaires et sont moins exposés à la concurrence. L'information donnée par ces graphiques est alors toujours à nuancer. De plus, les habitations d'une pièce ont un taux de résiliation très élevé contrairement au reste des biens. Ce taux élevé est expliqué par le turnover très important engendré par les étudiants sur ce type de bien. Sur le graphique de la variable montant de cotisation (cf. figure 5.5), nous remarquons effectivement que le taux de résiliation des contrats de moins de 100€ est extrêmement fort.

Montant de la cotisation annuelle : La cotisation annuelle payée par l'assuré a été découpée en plusieurs tranches pour faciliter la restitution.

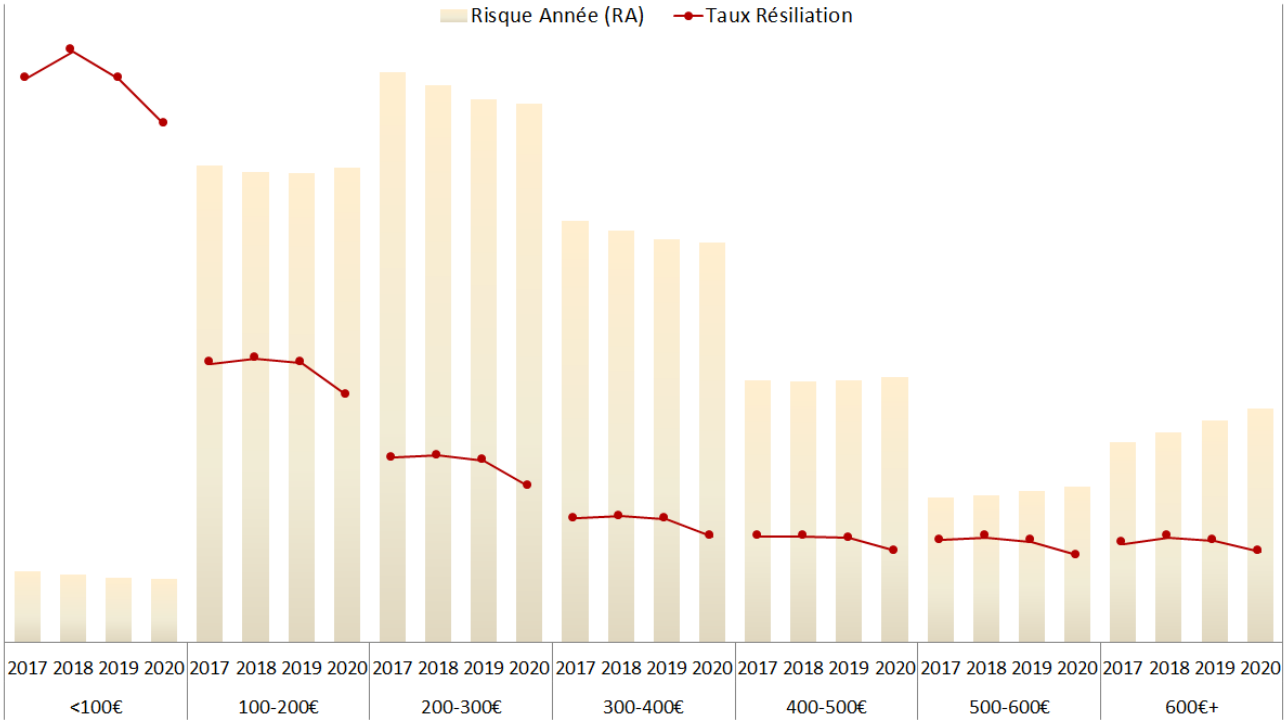

FIGURE 5.5: Taux de résiliation en fonction de la cotisation annuelle

Comme évoqué précédemment, la tranche inférieure à 100€ possède un taux de résiliation extrêmement fort, environ 3.5 fois plus élevé que la moyenne tous critères confondus. Cette tranche de cotisation est essentiellement composée d'étudiants louant de petits appartements, ce qui explique le taux de turnover très important de cette modalité.

Le montant de la cotisation est étroitement lié au nombre de pièces du bien. Le nombre de pièces est un critère très discriminant dans la tarification du produit habitation, donc il influe fortement le montant de la cotisation. Nous retrouvons alors le même phénomène que sur le nombre de pièces, les contrats avec un prime annuelle élevée ont moins de résiliations dû à la forte part de maisons propriétaires dans les tranches de cotisations élevées.

Évolution tarifaire sur 1 an : Cette variable quantitative a été catégorisée pour simplifier sa représentation. Elle représente l'évolution tarifaire de la cotisation annuelle nette entre la situation actuelle et la précédente. Lors de la création de cette variable les évolutions liées à un avenant n'ont pas été comptabilisées. En effet, si l'assuré effectue un avenant suite à une extension, le prix de la cotisation va automatiquement bondir. Cet avenant va alors fausser l'évolution appliquée. Nous avons seulement sélectionné les évolutions au terme du contrat, à savoir celles à l'initiative de MMA. La variable possède une modalité NR pour les situations pour lesquelles il n'est pas possible de calculer l'évolution tarifaire. D'autres variables explicatives ont également cette modalité non renseignée. Cela constitue un point d'attention pour la partie modélisation.

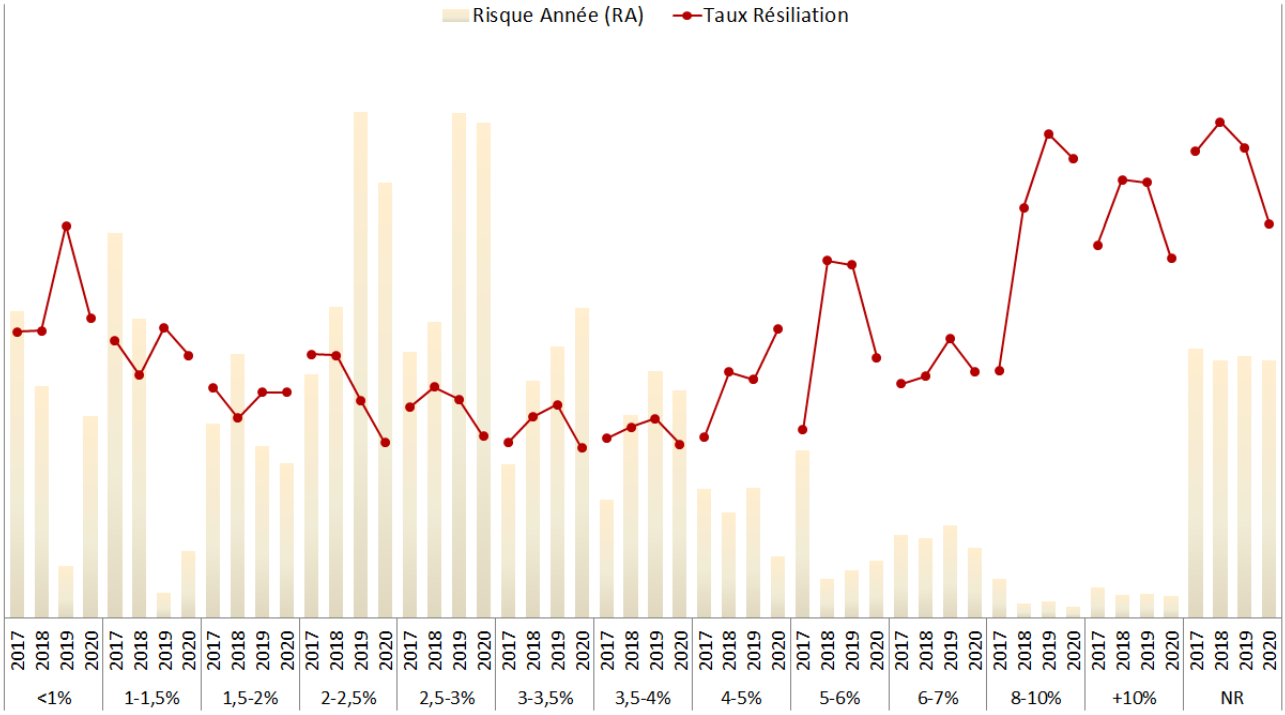

FIGURE 5.6: Taux de résiliation en fonction de l'évolution tarifaire 1 an

Une tendance est très visible pour le montant de la cotisation annuelle mais aucune tendance n'est réellement visible pour l'évolution tarifaire. Néanmoins, le taux de résiliation en fonction de l'évolution tarifaire semble avoir une courbe en forme de « U ». En effet, deux observations ressortent de ce graphique :

- Malgré la plus faible volumétrie le taux de résiliation augmente significativement pour les majorations de plus de 8%.
- Le taux de résiliation est élevé pour les contrats ayant subi une majoration inférieure à 1.5%.

La catégorie NR regroupe l'ensemble des contrats dont la variable n'est pas renseignée. Ces contrats sont essentiellement des affaires nouvelles et des avenants dont la situation précédente est une affaire nouvelle. La modalité NR donne alors de l'information sur l'ancienneté du contrat.

Nous allons étudier en détail la variable explicative d'évolution tarifaire. En effet, nous avons précédemment vu qu'une tendance de résiliation est difficilement observable sur ce critère (cf. figure 5.6).

Nous avons étudié plus en détails ce critère en distinguant les maisons des appartements. Les maisons sont représentées en bleu et les appartements en orange. Après la segmentation par type de bien une tendance plus visible se dégage. La volatilité entre chaque année est assez forte et créait du bruit sur le précédent graphique. Pour pallier à cette difficulté, le graphique ci-dessous (cf. figure 5.7) est une moyenne du taux de résiliation sur les 4 années d'étude.

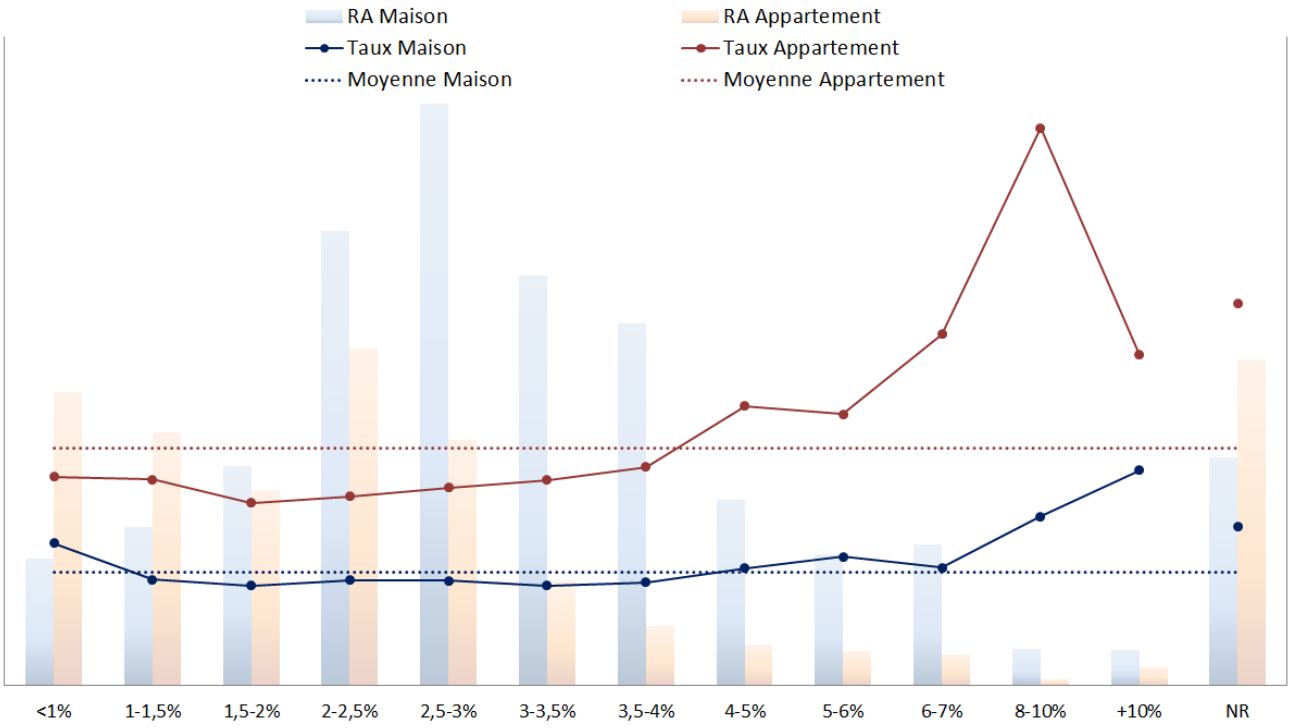

FIGURE 5.7: Taux de résiliation moyen sur 4 ans en fonction de l'évolution tarifaire

Les appartements semblent plus sensibles aux évolutions tarifaires. A partir d'une majoration de 4%, nous observons une nette augmentation du taux de résiliation. Pour les maisons, le taux de résiliation ne subit qu'une nette accentuation à partir d'une majoration de 8%.

Cependant, le taux de résiliation est encore élevé sur les situations ayant subi une hausse de moins de 1.5%. Ce taux de résiliation élevé semble contre intuitif, donc nous allons comparer la répartition des biens entre le portefeuille dans sa globalité et la répartition de situations ayant subi une majoration inférieure à 1.5%.

Sur la figure suivante (cf. figure 5.8), nous remarquons l'inversion des proportions entre les propriétaires de maison majoritaires sur le portefeuille global et les locataires d'appartement majoritaires sur les situations avec une faible majoration tarifaire. Les appartements en location ont un turnover plus important d'où la légère hausse du taux de résiliation observée sur les situations avec une majoration tarifaire inférieure à 1.5%.

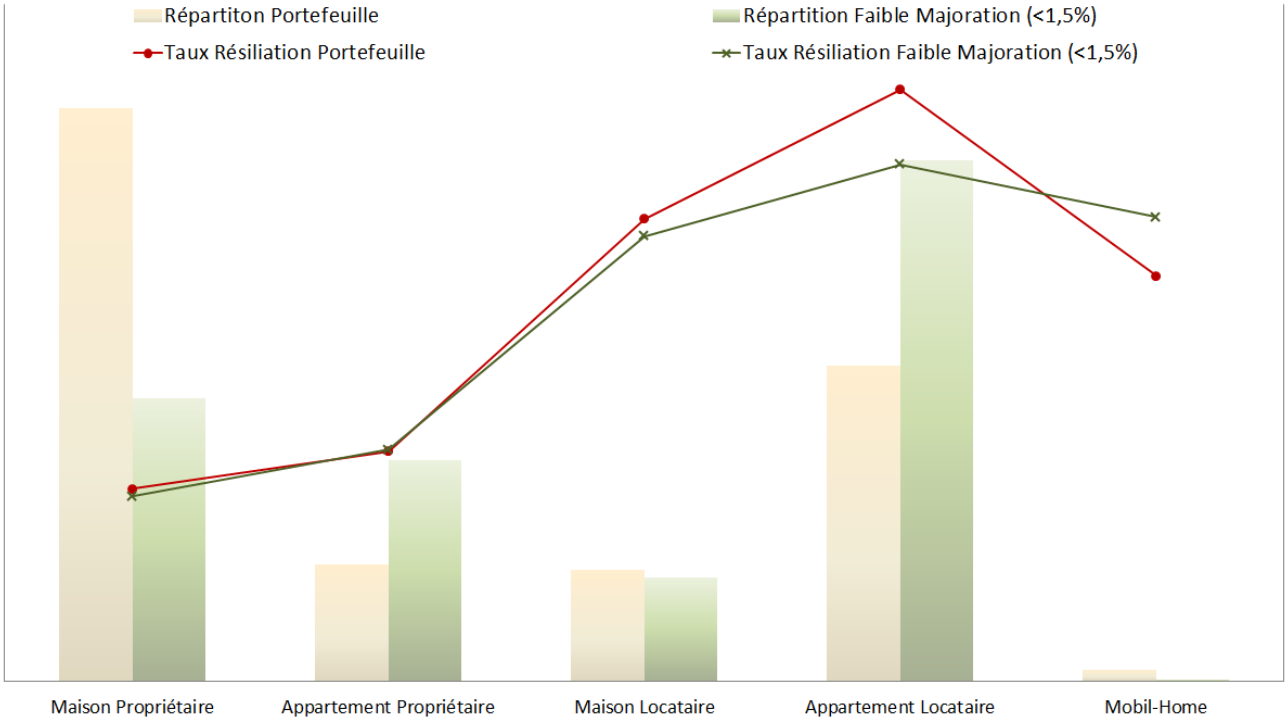

FIGURE 5.8: Taux de résiliation moyen sur 4 ans du portefeuille et des situations faiblement majorées

Conclusion : Cette partie d'analyse des variables explicatives aide à mieux cerner et saisir l'ensemble du périmètre de l'étude ainsi que de s'assurer de la qualité des données. Certaines variables se sont révélées être en adéquation avec les préjugés, comme la diminution du taux de résiliation en fonction du montant de la cotisation. D'autres variables se sont révélées moins impactantes que prévu, comme le type d'habitation qui est finalement moins discriminant que la qualité juridique de l'occupant.

A l'aide de statistiques à plat, il est difficile de confirmer des hypothèses. L'objectif dans la partie modélisation va être de vérifier si les tendances ne sont pas affectées par d'autres variables comme nous l'avons constaté précédemment sur la variable d'évolution tarifaire (cf. figures 5.7 et 5.8).

Troisième partie

Modélisation du taux de résiliation

Chapitre 6

Régression logistique

Sommaire

L'objectif de ce chapitre est de prédire les valeurs prises par la variable aléatoire Y modélisant le taux de résiliation d'une ligne situation. Dans le cadre d'une régression logistique binaire, cette variable ne prend que deux modalités, ici $\{0, 1\}$. Dans la base d'étude, la variable Y est la variable FLAG_RESIL et les variables explicatives citées précédemment sont notées $\{X_1, X_2, ..., X_n\}$.

6.1 Rappels théoriques

6.1.1 Apprentissage supervisé

Dans le cadre de l'apprentissage supervisé, l'objectif est de prédire ou d'expliquer une variable Y en fonction d'un vecteur de variables explicatives $\{X_1, X_2, ..., X_n\}$. Il s'agit de faire le lien entre ce vecteur et la variable à prédire sous la forme :

$$Y = f(X, \alpha)$$

La fonction $f(.)$ est le modèle de prédiction tandis que α est le vecteur des paramètres de la fonction.

Le classifieur Bayesien est celui qui répond de manière optimale aux spécifications de notre problème. Pour un individu ω, il s'agit de calculer les probabilités conditionnelles pour chaque modalité y_k de Y. Ici Y, prend les valeurs : 0 = situation non résiliée ou 1 = situation résiliée :

$$\mathbb{P}[Y(\omega) = y_k | X(\omega)] \tag{6.1}$$

A partir de cette probabilité, on associe l'individu à sa classe la plus probable. Dans le cas de la modélisation binaire, un seuil est déterminé. Ce seuil est la frontière entre prédire une situation résiliée ou non. La formule de la probabilité conditionnelle s'écrit comme suit :

$$\mathbb{P}(Y = y_k|X) = \frac{\mathbb{P}(Y = y_k) \times \mathbb{P}(X|Y = y_k)}{\mathbb{P}(X)} = \frac{\mathbb{P}(Y = y_k) \times \mathbb{P}(X|Y = y_k)}{\sum_k \mathbb{P}(Y = y_k) \times \mathbb{P}(X|Y = y_k)} \tag{6.2}$$

Dans le cadre de notre étude des résiliations, il suffit de comparer $\mathbb{P}(Y = 1|X)$ et $\mathbb{P}(Y = 0|X)$.

6.1.2 Modèle logit

Plusieurs modèles de régression existent, par exemple :

- Le modèle Logit
- Le modèle Probit
- Le modèle Tobit

Néanmoins, nous appliquerons la transformation Logit pour déterminer le taux de résiliation car les coefficients α des variables explicatives sont facilement interprétables et cette transformation est adaptée à la modélisation dans le cadre binaire.

Pour un individu ω sa transformation Logit est :

$$ln\left[\frac{\pi(\omega)}{1 - \pi(\omega)}\right] = \alpha_0 + \alpha_1 X_1 + ... + \alpha_n X_n \tag{6.3}$$

π représente la probabilité de résilier. Pour l'obtenir, posons $C(X) = \alpha_0 + \alpha_1 X_1 + ... + \alpha_n X_n$. Le Logit, $C(X)$ est défini entre $-\infty$ et $+\infty$, tandis que π est bien compris entre 0 et 1, avec la fonction logistique π qui s'écrit de la manière suivante :

$$\pi = \frac{e^{C(X)}}{1 + e^{C(X)}} = \frac{1}{1 + e^{-C(X)}} \tag{6.4}$$

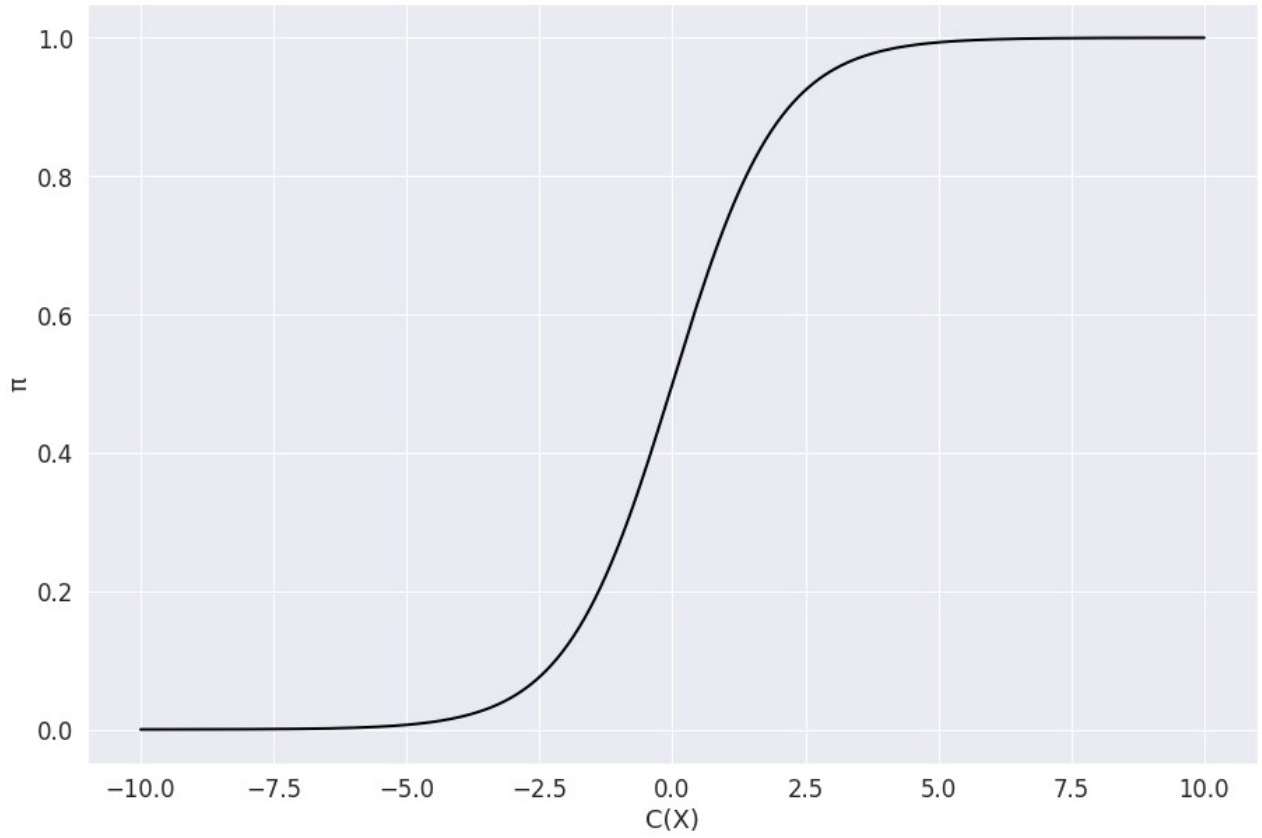

FIGURE 6.1: Fonction Logistique

6.1.3 Estimation des paramètres

Les paramètres de la régression logistique s'estiment par la méthode du maximum de vraisemblance. Il est d'abord nécessaire de déterminer la loi de distribution de $\mathbb{P}(Y|X)$. Par chance, dans notre étude Y est une variable binaire définie dans $\{0, 1\}$. Pour un individu ω, on modélise la probabilité à l'aide de la loi binomiale $\mathcal{B}(1, \pi)$:

$$\mathbb{P}(Y(\omega)|X(\omega)) = \pi(\omega)^{y(\omega)} \times (1 - \pi(\omega))^{1-y(\omega)} \tag{6.5}$$

— Si $y(\omega) = 1$, alors $\mathbb{P}(Y(\omega) = 1|X(\omega)) = \pi$
— Si $y(\omega) = 0$, alors $\mathbb{P}(Y(\omega) = 0|X(\omega)) = 1 - \pi$

La vraisemblance correspond à la probabilité d'obtenir l'échantillon souhaité à partir d'un tirage dans la population. Cette mesure varie entre 0 et 1. La méthode du maximum de vraisemblance consiste à estimer les paramètres α de la régression logistique de sorte à maximiser cette probabilité. La vraisemblance s'écrit comme suit :

$$\mathbf{L} = \prod_{\omega} \pi(\omega)^{y(\omega)} \times (1 - \pi(\omega))^{1-y(\omega)} \tag{6.6}$$

Le modèle H2O est ajusté en maximisant la fonction qui résulte de la log-vraisemblance suivante :

$$\max_{C(X)} \frac{1}{N} \sum_{i=1}^{N} \Big[y_i \times C(X) - ln(1 + C(X)) \Big] \tag{6.7}$$

Le vecteur $\alpha = (\alpha_0, \alpha_1, ..., \alpha_n)$ qui maximise la vraisemblance et la log-vraisemblance est le même car le logarithme népérien est une fonction monotone. Cet estimateur du maximum de vraisemblance $\hat{\alpha}$ possède les propriétés suivantes :

— Il est asymptotiquement sans biais ;
— Il est asymptotiquement gaussien ;
— Il est de variance minimale ;

La log-vraisemblance est une fonction convexe, il existe alors une solution unique pour le vecteur $\hat{\alpha}$. Néanmoins, il n'existe pas de solution analytique, nous devons passer par un algorithme pour approcher cette valeur. Chaque logiciel possède ses propres algorithmes. H2O a implémenté différentes méthodes de calcul. L'utilisation des différentes méthodes n'influe pas ou très peu sur les résultats. Nous avons conservé l'algorithme par défaut, IRLSM : « Iteratively Reweighted Least Squares Method ».

Dans le cadre de la régression logistique la probabilité d'être résilié est calculée comme suit :

$$\hat{y} = \mathbb{P}(y = 1|X) = \frac{e^{C(X)}}{1 + e^{C(X)}} \tag{6.8}$$

6.1.4 Évaluation du modèle

Le critère de qualité le plus connu et utilisé est le critère des moindres carrés ordinaires (MCO). Cette métrique mesure la moyenne des carrés des erreurs entre la prédiction du modèle et la valeur réelle.

$$MCO = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{6.9}$$

Néanmoins, ce critère n'est pas très optimisé pour estimer l'erreur ou la qualité d'un modèle logistique binaire. La valeur observée prend les modalités 0 ou 1, tandis que la valeur prédite par le modèle est une probabilité qui ensuite est transformée en $\{0, 1\}$. La distance mesurée par les MCO est alors soit de 0, soit de 1. Une autre mesure semble plus adaptée pour la classification binaire.

La métrique utilisée dans l'étude pour comparer les modèles est la courbe ROC (Receiver Operating Characteristic). Elle présente des caractéristiques très intéressantes pour l'évaluation et la comparaison des performances des modèles :

— C'est un outil graphique donc très intuitif et visuel auquel on attache un indicateur numérique : la métrique AUC. L'AUC, compris entre 0 et 1, est l'aire sous la courbe ROC.
— La courbe est indépendante des coûts de mauvaise affectation. Elle permet par exemple de déterminer si un modèle surpasse un autre, quelle que soit la combinaison de coûts utilisée.
— Cette métrique est opérationnelle même lorsque les distributions sont déséquilibrées.

La courbe ROC est obtenue à l'aide de la matrice de confusion. Quatre classes constituent cette matrice :
— Vrai Négatif : les valeurs non résiliées bien prédites ;
— Faux Positif : les valeurs non résiliées prédites comme résiliées ;
— Faux Négatif : les valeurs résiliées mais prédites comme non résiliées ;
— Vrai Positif : les valeurs résiliées correctement prédites ;

		Valeur prédite	
		0 : Non Résilié	**1 : Résilié**
Valaur observée	**0 : Non Résilié**	Vrai Négatif (VN)	Faux Positif (FP)
	1 : Résilié	Faux Négatif (FN)	Vrai Positif (VP)

FIGURE 6.2: Matrice de confusion

La courbe ROC met en relation le taux de vrais positifs et le taux de faux positifs dans un graphique. Sur la figure ci-dessous, la courbe verte représente une discrimination parfaite c'est-à-dire que le modèle prédit parfaitement les valeurs observées. L'inverse est représenté par la courbe rouge. Dans ce cas il n'y pas de discrimination, le modèle à la même capacité de prédiction que le hasard. La capacité discriminatoire d'un modèle est jugée acceptable lorsque le critère AUC est supérieur à 0.7

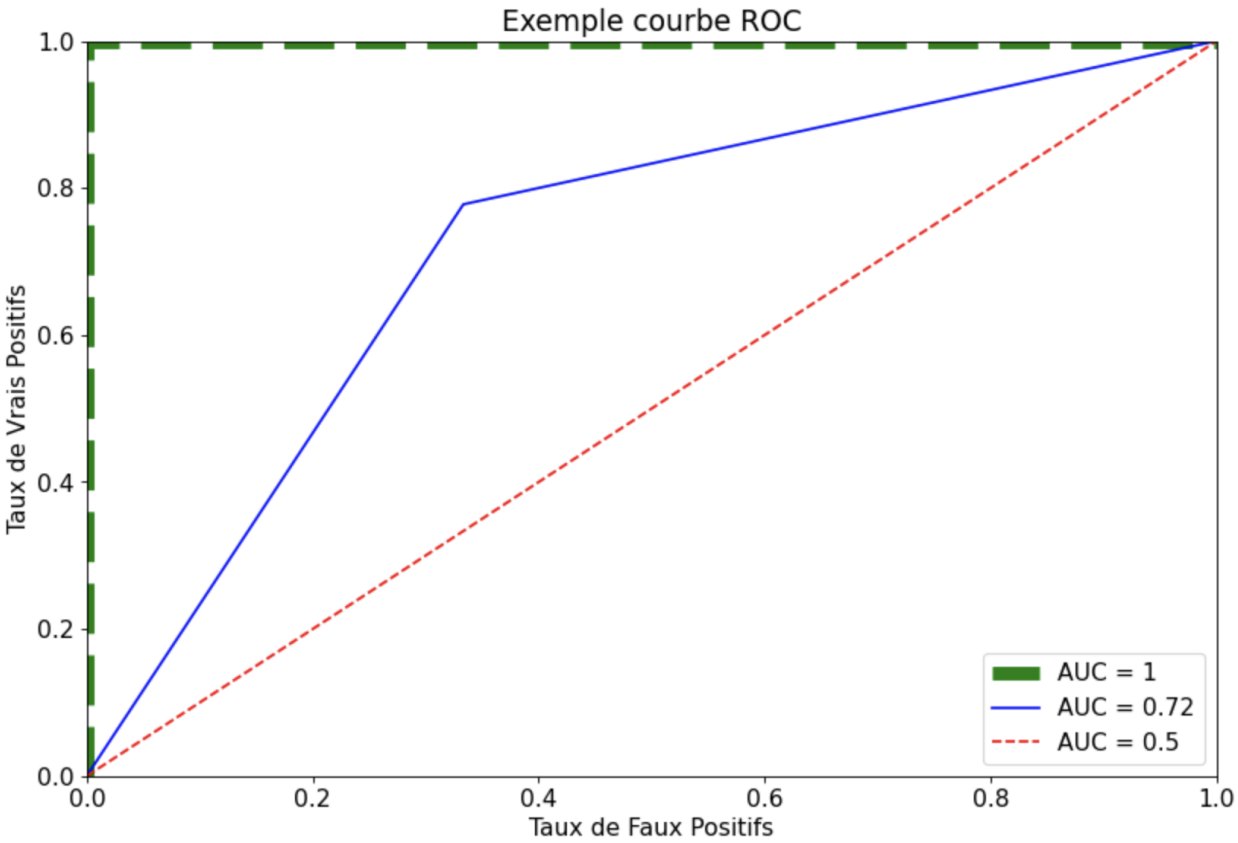

FIGURE 6.3: Exemple de courbe ROC

6.2 Sélection 1 : Matrice de corrélation

6.2.1 Colinéarité et corrélation

Dans une régression, la colinéarité des variables explicatives est un problème qui survient lorsque certaines mesurent le même phénomène. Une colinéarité des variables importante s'avère problématique, car elle peut augmenter la variance des coefficients de régression, les rendre instables et difficiles à interpréter. La colinéarité ne biaise pas les prédictions du modèle, mais les coefficients individuels associés à chaque variable explicative ne peuvent pas être interprétés de façon fiable. Or, la grande force des régressions logistiques est l'utilisation de ces coefficients individuels, car nous souhaitons identifier les profils susceptibles de résilier.

Dans cette partie, nous étudierons la corrélation des variables et non pas leur colinéarité. Néanmoins, des variables colinéaires sont obligatoirement fortement corrélées entre elles.

La corrélation est une analyse bivariée qui mesure la force de l'association entre deux variables et la direction de la relation. La valeur de ce coefficient de corrélation varie entre +1 et -1. Une valeur de $\pm$ 1 indique un degré parfait d'association entre les deux variables. Plus la valeur du coefficient de corrélation se rapproche de 0, plus la relation entre les deux variables est faible. Le sens de la relation est indiqué par le signe du coefficient : un signe + indique une relation positive et un signe - une relation négative. Les trois corrélations les plus répandues sont la corrélation de Pearson, la corrélation de Kendall et la corrélation de Spearman.

— La corrélation de Pearson est la statistique de corrélation la plus largement utilisée pour mesurer le degré de relation linéaire entre des variables.
— Les corrélations de rang de Kendall et de Spearman sont des tests non paramétriques qui mesurent la force de la dépendance entre deux variables.

Les corrélations issues du tau de Kendall et du rho de Spearman donnent des résultats similaires. A l'inverse de la corrélation de Pearson, ces deux mesures permettent de capter la corrélation non-linéaire. Nous avons fait le choix du rho de Spearman : ρ, car son temps de calcul est beaucoup plus faible. L'idée du coefficient de Spearman est d'utiliser le rang des observations pour mesurer leur relation. Dans cet exemple, nous mesurons la corrélation entre deux variables issues des colonnes X et Y. On pose R_i le rang de l'observation x_i et S_i le rang de l'observation y_i.

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}} \tag{6.10}$$

Dans la suite, lorsque deux variables ont une corrélation jugée raisonnable, les deux variables sont conservées pour la modélisation du GLM. A l'inverse, si deux variables ont une corrélation supérieure à 0.6, c'est-à-dire une corrélation forte, un choix entre ces deux variables doit être effectué. Enfin, lorsque deux variables sont corrélées modérément, nous étudions au cas par cas. Certaines variables explicatives sont opérationnellement importantes et il est alors difficile de les retirer totalement de l'étude.

corrélation : ρ	Relation	Variable jugée importante	Variable jugée accessoire
ρ < 0,4	Corrélation raisonnable	Conservation	Conservation
0,4 ≤ ρ < 0,6	Corrélation modérée	Conservation	Suppression ou retraitement
ρ ≥ 0,6	Corrélation forte	Suppression ou retraitement	Suppression ou retraitement

FIGURE 6.4: Tableau de corrélation décisionnel

6.2.2 Suppression des variables corrélées

49 variables ont été préalablement sélectionnées pour la modélisation (cf. figure 4.3). Seulement, certains des critères sélectionnés sont fortement liés entre eux. L'objectif de cette partie est de présenter les différentes modifications et suppressions réalisées pour améliorer la qualité et l'interprétabilité du modèle logistique.

Variable croisée TYPExQLTExPIECE : 49-2=47 variables La corrélation entre les variables CD_TYPE_HABI, CD_QLTE_ASSU_HABI et NB_PIEC est assez importante. Ces 3 variables sont omnipotentes dans l'assurance habitation, il est donc impossible d'en retirer une. Une variable croisée est ainsi créée pour conserver l'ensemble des informations fourni par les 3 variables. Il est dangereux de conserver ces 3 variables en l'état, car leur coefficients GLM peuvent être biaisés par leurs fortes interactions.

	flag_resil	CD_TYPE_HABI	CD_QLTE_ASSU_HABI	NB_PIEC
flag_resil	1.00	-0.11	-0.14	-0.11
CD_TYPE_HABI	-0.11	1.00	0.57	0.57
CD_QLTE_ASSU_HABI	-0.14	0.57	1.00	0.44
NB_PIEC	-0.11	0.57	0.44	1.00

FIGURE 6.5: Corrélation entre le type d'habitation, la qualité juridique et le nombre de pièces

Variables dépendance : 47-1=46 variables Deux variables offrent de l'information sur la dépendance de l'habitation. La première : FLAG_DPDC indique si l'habitation possède une dépendance ou non, tandis que la seconde, TAILLE_DPDC indique la taille de celle-ci. Cette variable est transformée en variable catégorielle pour aider à la discrimination des habitations sans dépendance. La corrélation entre ces deux variables est logiquement extrêmement forte, nous décidons alors de ne conserver que la variable qui offre le plus de détails : TAILLE_DPDC.

Variables cotisation : 46-1=45 variables Deux variables existent pour donner le montant de la cotisation. La première est le montant net annuel de la cotisation et la seconde le montant annuel de la cotisation barème. Le montant net est la prime réelle payée par le client, tandis que le second montant est celui calculé avant l'application des dérogations et des encadrements administrés à l'échéance du contrat. Ces deux variables ont alors logiquement une corrélation très proche de 1, ce qui nous pousse à ne conserver que la variable COT_NET_N qui est la prime réellement connue et payée par le client.

Variables d'évolution tarifaire : 45-3=42 variables Dans la base d'étude, 5 variables offrent de l'information sur la majoration tarifaire subie au terme. Les variables EVOL_NET_FIN_N* indiquent l'évolution tarifaire entre une situation et sa précédente. La variable EVOL_NET_FIN_N1-N3 est un cumul des évolutions de la situation n-4 à n-1.

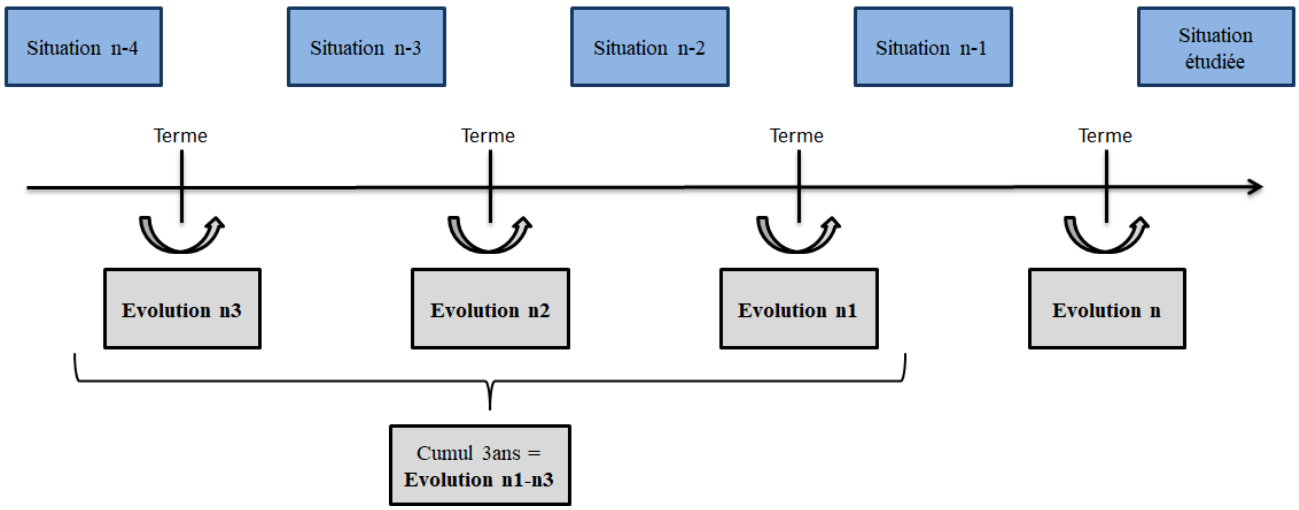

FIGURE 6.6: Construction de la variable d'évolution cumulée

Du fait de sa construction, la variable EVOL_NET_FIN_N1-N3 est corrélée avec les 3 variables d'évolution qui la constituent. La dernière variable d'évolution annuelle est conservée afin de mesurer l'impact à

court terme d'une majoration, ainsi que la variable cumulée sur 3 ans pour mesurer l'impact d'une accumulation de hausses tarifaires.

	flag_resil	evol_net_fin_n	evol_net_fin_n1	evol_net_fin_n2	evol_net_fin_n3	evol_net_fin_n1-n3
flag_resil	1.00	-0.01	-0.00	-0.01	0.00	-0.03
evol_net_fin_n	-0.01	1.00	0.08	-0.03	0.03	0.06
evol_net_fin_n1	-0.00	0.08	1.00	0.11	-0.00	0.52
evol_net_fin_n2	-0.01	-0.03	0.11	1.00	0.10	0.55
evol_net_fin_n3	0.00	0.03	-0.00	0.10	1.00	0.74
evol_net_fin_n1-n3	-0.03	0.06	0.52	0.55	0.74	1.00

FIGURE 6.7: Corrélation entre les variables d'évolution tarifaire

Variables nombre d'agences : 42-5=37 variables Ces variables renseignent sur le nombre d'agences dans les x kilomètres autour du risque assuré. Logiquement, les variables qui dénombrent les agences dans un rayon proche sont corrélées entre elles. Nous choisissons de conserver la variable NB_2KM car elle est la plus corrélée avec la résiliation. Ensuite, la seule variable nombre d'agences corrélée modérément avec la précédente est NB_30KM. Le reste des variables nombre d'agences est retiré de la modélisation Logit ainsi que la variable TARIF_B100.

	flag_resil	nb_1km	nb_2km	nb_5km	nb_10km	nb_20km	nb_30km	TARIF_B100
flag_resil	1.00	0.04	0.05	0.03	0.02	0.01	0.01	0.03
nb_1km	0.04	1.00	0.69	0.36	0.23	0.18	0.16	0.11
nb_2km	0.05	0.69	1.00	0.68	0.52	0.43	0.38	0.30
nb_5km	0.03	0.36	0.68	1.00	0.92	0.81	0.73	0.51
nb_10km	0.02	0.23	0.52	0.92	1.00	0.94	0.87	0.54
nb_20km	0.01	0.18	0.43	0.81	0.94	1.00	0.97	0.54
nb_30km	0.01	0.16	0.38	0.73	0.87	0.97	1.00	0.53
TARIF_B100	0.03	0.11	0.30	0.51	0.54	0.54	0.53	1.00

FIGURE 6.8: Corrélation entre les variables nombre d'agences

Variables sinistralité : 37-4=33 variables Plusieurs variables donnent des informations sur la sinistralité actuelle et passée du contrat. Les premières sont le nombre et le coût des sinistres sur la situation étudiée et sur 4 années cumulées. Nous avons fait le choix de conserver les variables calculées sur la période des 4 années passées, car très peu de situations ont au moins un sinistre sur la situation étudiée.

De plus, deux variables ont aidé à la construction de la variable MAJO_SINI. Ces deux variables, corrélées avec cette dernière, sont alors retirées des variables explicatives.

Variables nature de situation : 33-1=32 variables Dans la base d'étude, la variable NATU_SITU est l'évènement qui conduit à la création de la situation. On y retrouve les modalités affaire nouvelle, terme, avenant et rectification. En plus de cette variable, un flag avenant est renseigné. La corrélation

assez importante entre les deux, ainsi que le fait que la NATU_SITU englobe le flag avenant, nous pousse à retirer ce dernier de l'étude.

	flag_resil	flag_avt_n	NATU_SITU
flag_resil	1.00	-0.02	-0.05
flag_avt_n	-0.02	1.00	-0.51
NATU_SITU	-0.05	-0.51	1.00

FIGURE 6.9: Corrélation entre les variables nature de situation et avenant

6.2.3 Performances du premier GLM

A l'issue de cette sélection, un premier modèle GLM est appliqué sur la base. Pour cela nous avons scinder notre base d'étude en deux échantillons :

— Un échantillon de 80% de la base d'étude pour la base d'apprentissage. C'est à l'aide de cette base que seront estimés les paramètres du modèle à l'aide du maximum de vraisemblance.

— Un échantillon de 20% pour la base validation. Cette base permet de mesurer la performance du modèle sur des données non apprises.

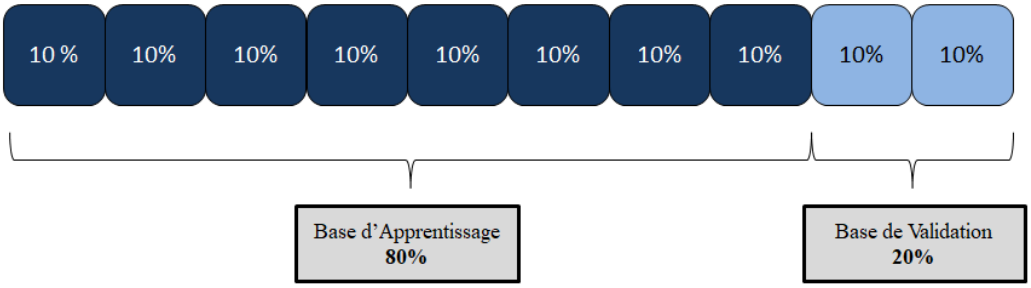

FIGURE 6.10: Découpage base d'apprentissage, de validation et de test

Les résultats des modèles présentés sont ceux calculés sur l'échantillon de validation. Mesurer la qualité d'un modèle sur son échantillon d'apprentissage n'est pas optimal et favorise le sur-apprentissage dans les modèles.

Les résultats de ce premier GLM sont extrêmement bons que ce soit sur l'échantillon d'apprentissage ou de validation.

— AUC base d'apprentissage : 80.21%

— AUC base de validation : 79.94%

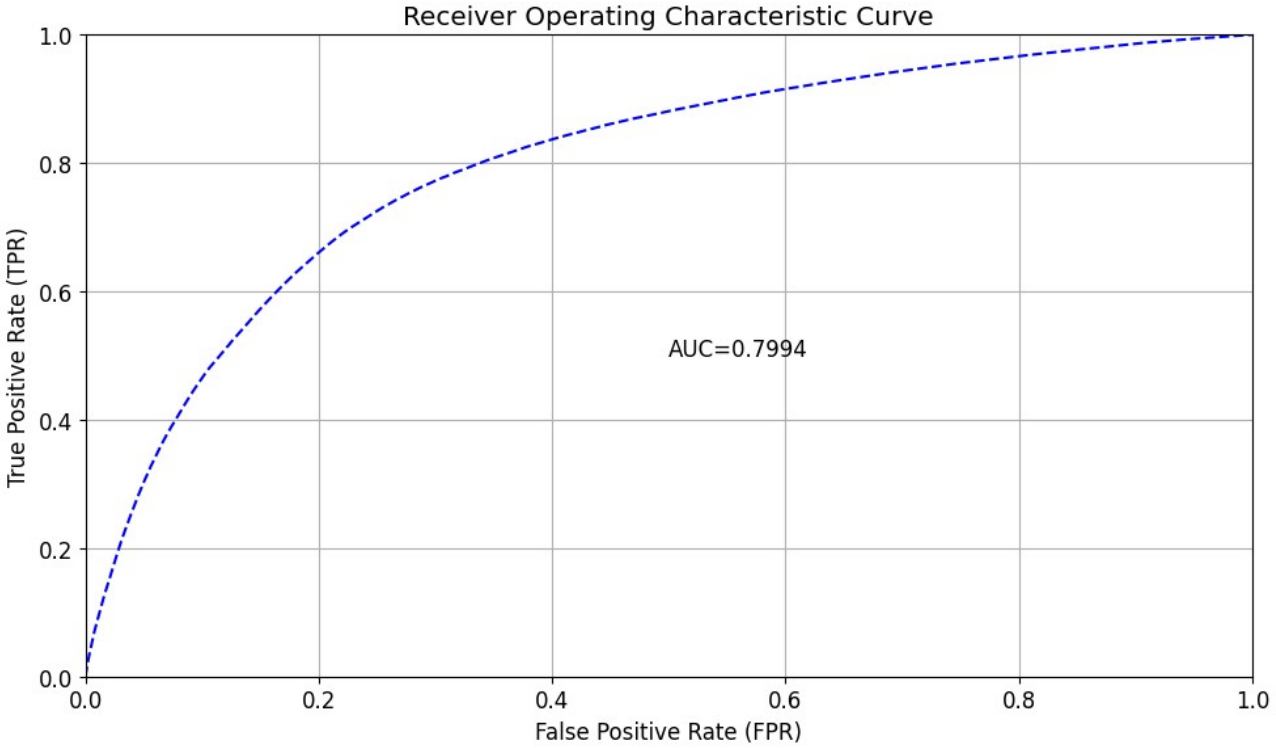

FIGURE 6.11: AUC GLM avec les variables non retraitées

6.3 Raffinement de la base

La base compte désormais 32 variables explicatives. Cette partie développe les derniers ajustements réalisés afin de parfaire la modélisation logistique.

6.3.1 Catégorisation des variables numériques

Catégoriser certaines variables permet à la fois de les rendre plus discriminantes mais aussi plus facilement interprétables. En effet, du fait de la construction du GLM, conserver certaines variables sous leur forme numérique peut faire perdre de l'information ou biaiser le coefficient associé à la variable. Par exemple, les générations de contrat de plus de 30 années ont été regroupées au sein d'une même catégorie afin d'éviter au modèle d'ajuster le coefficient jusqu'aux plus anciens contrats MRH. D'autres variables ont été catégorisées comme le montant de la cotisation ou l'évolution tarifaire. L'objectif est d'aider le modèle à capter des informations autres que linéaires sur ce genre de critère. Lorsque la variable est continue le GLM estime un unique paramètre, tandis que pour une variable catégorielle le GLM estime un paramètre pour chaque modalité de la variable catégorielle.

Les résultats sur ce modèle avec les 32 variables transformées en catégorie sont encore meilleurs. L'AUC de ce nouveau modèle est d'environ 1.3% supérieur au précédent.

— AUC base d'apprentissage : 81.48%
— AUC base de validation : 81.25%

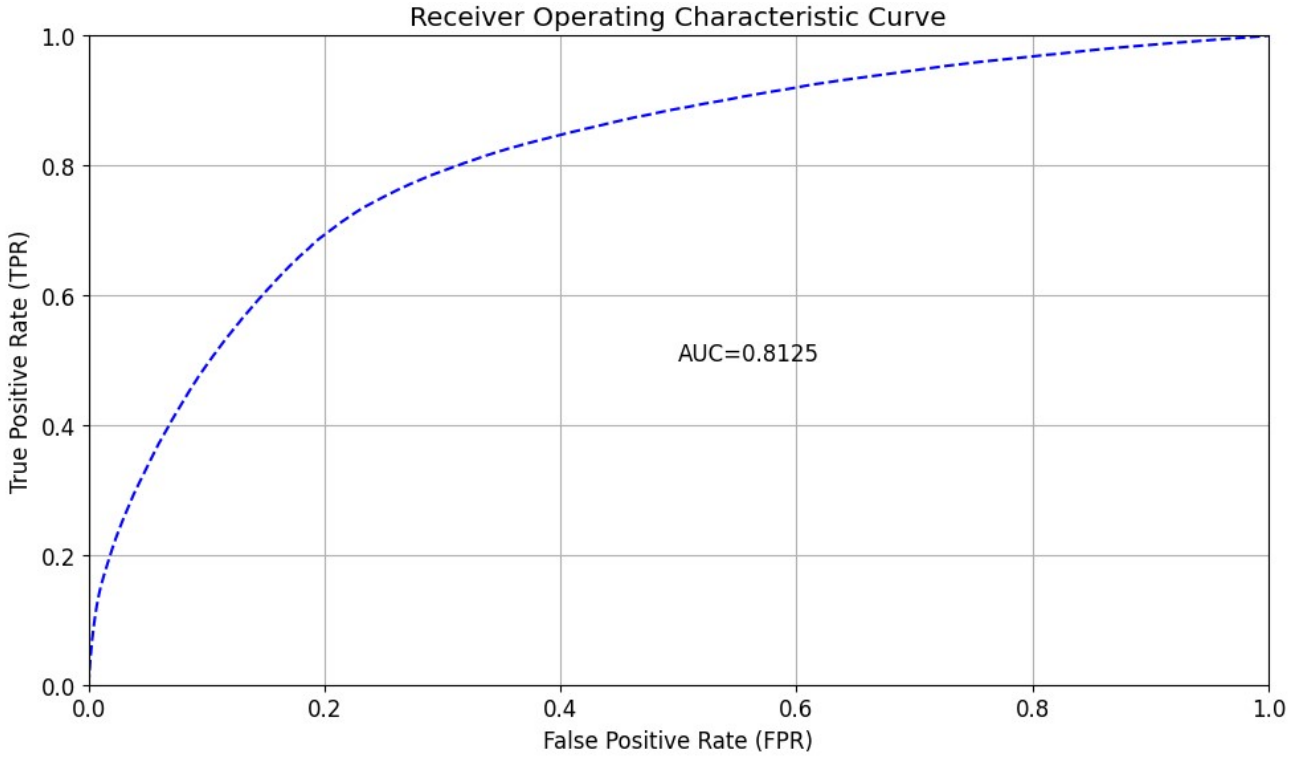

FIGURE 6.12: AUC GLM avec les variables catégorisées

Les performances de ces modèles Logit sont excellents, nous décidons alors d'étudier plus en détail les performances de chaque variable. Étonnamment, la variable la plus discriminante est le segment du client. Cette donnée, pour la première fois incluse dans une étude, est créée à partir de la classe d'âge et de l'unité urbaine du client. La force discriminante de cette variable a été fortement sous-estimée.

Le graphique de dépendance partielle montre l'effet marginal d'une variable sur la prédiction d'un classifieur. Il traduit l'impact d'une variable en moyennant l'influence de toutes les autres variables explicatives. Le graphique ci-dessous met en lumière un biais engendré par la modalité NR (Non-Renseignée) de la variable CD_SEGM_PART_AGG. La réponse moyenne de cette modalité est bien supérieure aux autres. Le modèle utilise donc cette modalité des non renseignées pour aider à la prédiction de la résiliation.

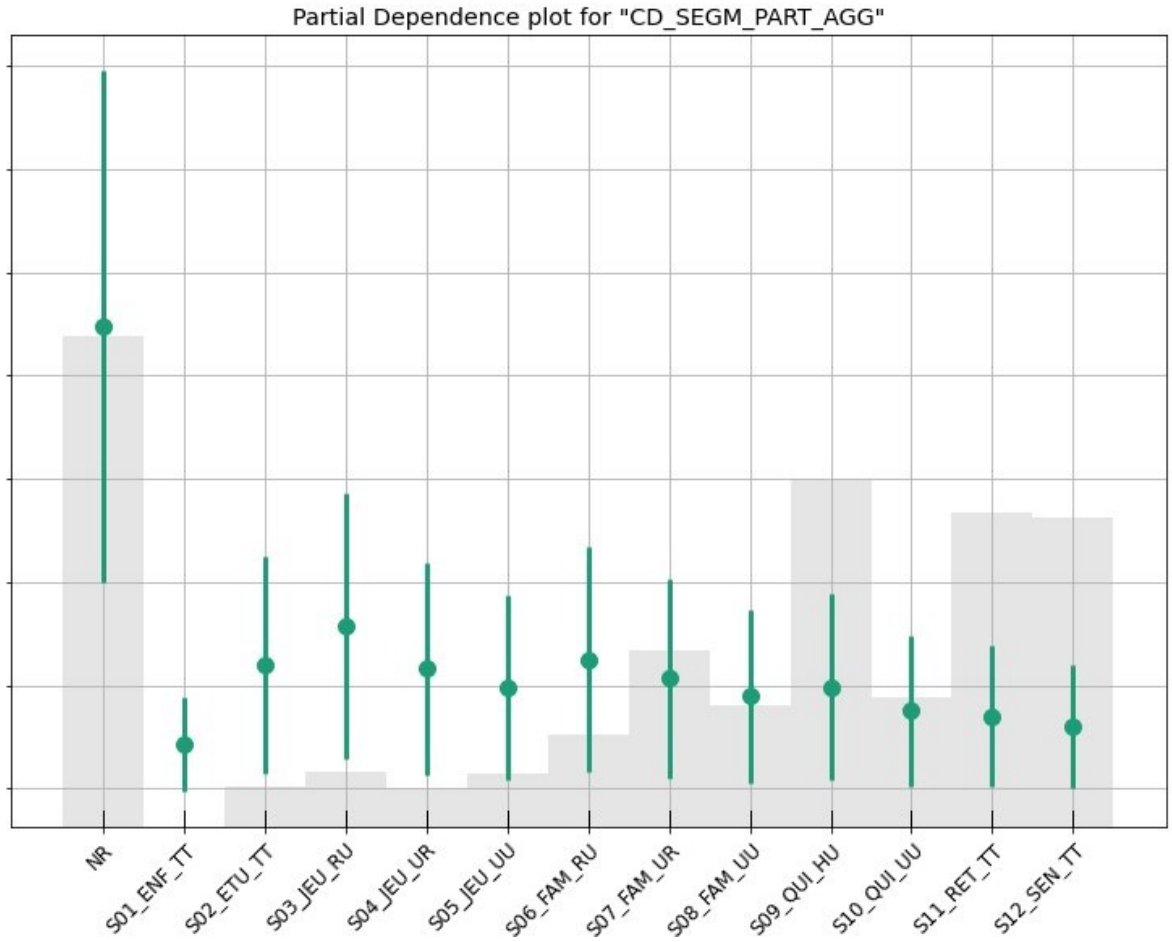

FIGURE 6.13: Graphique de dépendance partielle de la variable segment du client

6.3.2 Traitement de la variable segment client

La variable du segment client mis en place par la Direction Marketing s'est révélée très significative dans la modélisation logistique. Néanmoins, ce descripteur compte 1.45 millions de lignes non renseignées qui biaisent le modèle. Cette variable est très récente et n'a jamais été intégrée dans nos études. L'intégration de celle-ci a été réalisée en utilisant seulement la dernière vision (2020) de la base client. En effet, en observant de plus près cette variable, nous remarquons que les situations résiliées ont un fort taux de non renseignées. Environ 81% des situations non résiliées ont la variable de renseignée contre seulement 38% des situations résiliées.

SEGM_PART	Non résilié	Résilié
Renseigné	81%	38%
Non-Renseigné	19%	62%

FIGURE 6.14: Graphique de dépendance partielle de la variable segment du client

Les premières solutions envisagées sont de supprimer ou recoder les lignes non renseignées. Seulement, 60% des résiliations de la base d'étude n'ont pas cette variable de renseignée. Ces deux solutions sont écartées, il est impensable de supprimer ou recoder 60% des résiliations de la base. Le risque de créer un biais est trop important.

Pour résoudre le problème, nous avons affiné le rapprochement de cette variable dans notre base d'étude sur SAS. Le rapprochement entre cette variable et notre base d'étude s'effectue à l'aide du numéro client. La variable CD_SEGM_PART_AGG évolue dans le temps, nous avons alors récupéré plusieurs visions de la variable pour affecter la vision la plus proche de la date de début de situation. Si cette vision n'est pas renseignée la seconde la plus proche est rapprochée et ainsi de suite. Avec ce nouveau rapprochement, la modalité NR est passée de 1.45 à 0.35 millions de lignes. De plus, la répartition des lignes non renseignées entre situations résiliées et non résiliées s'est presque équilibrée.

Dans le but de corriger le dernier biais, nous voulons réduire à 0 le nombre de lignes non renseignées. Pour cela, un algorithme d'imputation est nécessaire. Ce type d'algorithme peut être plus ou moins complexe. Dans notre étude nous avons utiliser une imputation simple dans sa conception, qui consiste à attribuer la valeur la plus représentée du profil auquel est affecté la situation dont la valeur est non renseignée. Les variables CD_TYPEXQLTEXPIEC, GENERATION, FORMULE , DETENTION_N et BONUS-MALUS sont croisées afin de créer les différents profils. Au total, 70.000 profils sont créés.

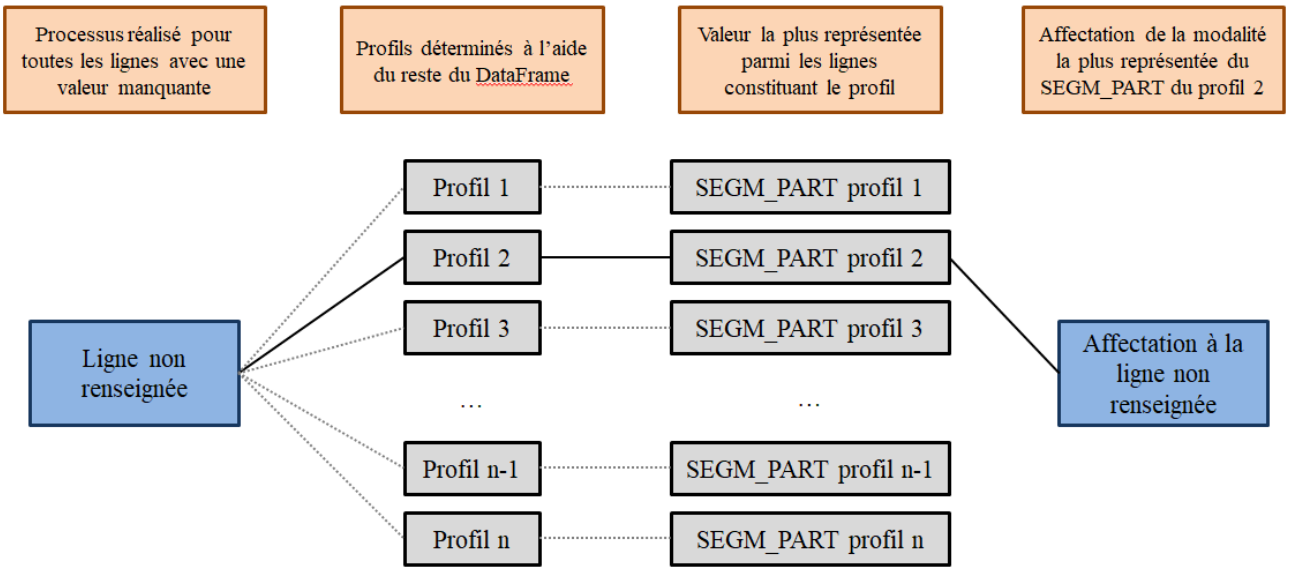

FIGURE 6.15: Algorithme d'imputation

Deux grandes étapes de retraitement de la variable CD_SEGM_PART_AGG ont été réalisées. La première a permis de passer de 22.5% de NR à 5.5% en affinant la récupération de cette variable sur SAS. La seconde étape réduit à 0 le nombre de lignes non renseignées. Le graphique ci-dessous conforte les retraitements effectués. Tout d'abord, l'enrichissement de la variable a permis de hausser le taux de résiliation à son niveau réel. De nouvelles discriminations intéressantes en ressortent, nous remarquons par exemple que peu importe la classe d'âge les clients en zone urbaine ont un taux de résiliation plus important que les ruraux. Cette observation est l'inverse de celle constaté avant le retraitement. Enfin, la partie imputation de données n'a entraîné aucune dérive sur le taux de résiliation. Le taux de résiliation avant imputation, en vert, est le même, quelque soit la modalité, que le taux de résiliation après imputation, en bleu. Les deux courbes se superposent parfaitement.

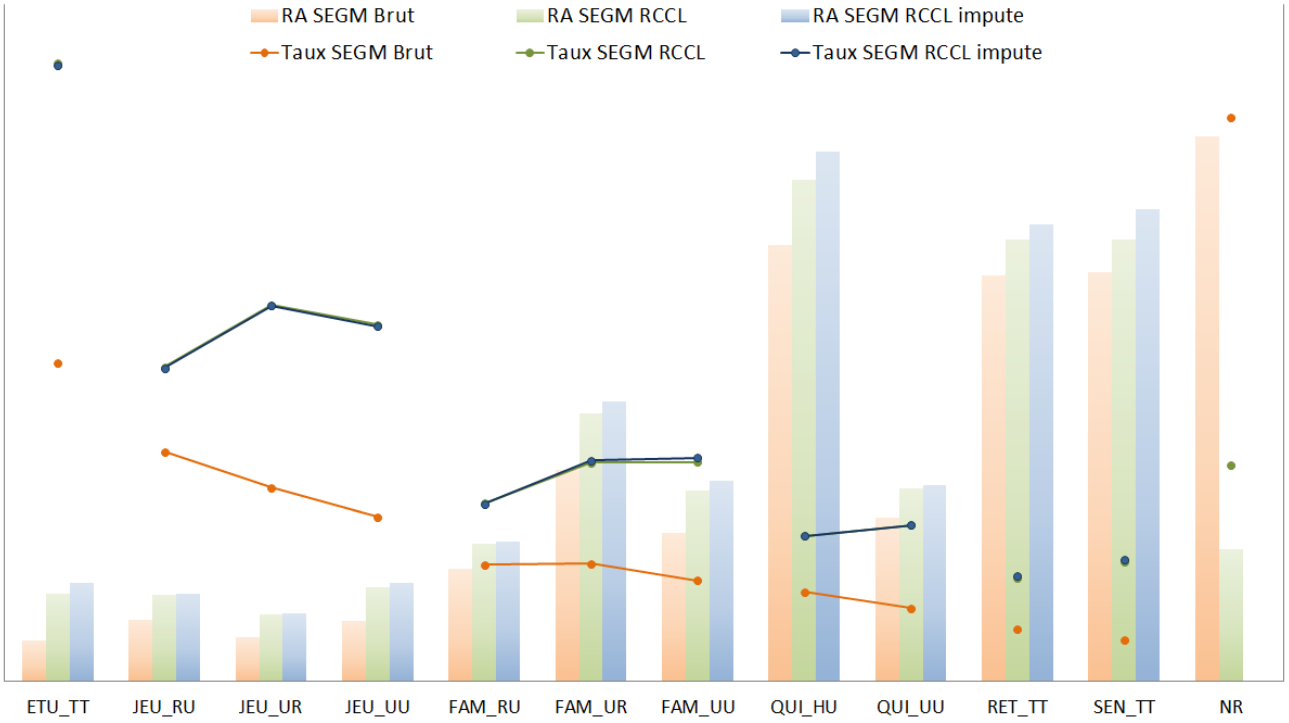

FIGURE 6.16: Comparaison du taux de résiliation de la variable segment client après différents retraitements

6.3.3 Traitement de la variable d'évolution tarifaire

D'autres variables retenues dans la sélection des 32 comportent également des lignes non renseignées, c'est le cas de la variable d'évolution tarifaire. Environ 800k lignes ne sont pas renseignées. Après avoir observé les résultats de l'imputation précédente, notre première intuition a également été d'imputer ces valeurs manquantes à l'aide du même algorithme. Au lieu d'affecter la valeur la plus représentée, l'algorithme affecte la valeur moyenne des profils pour les variables continues comme l'évolution tarifaire.

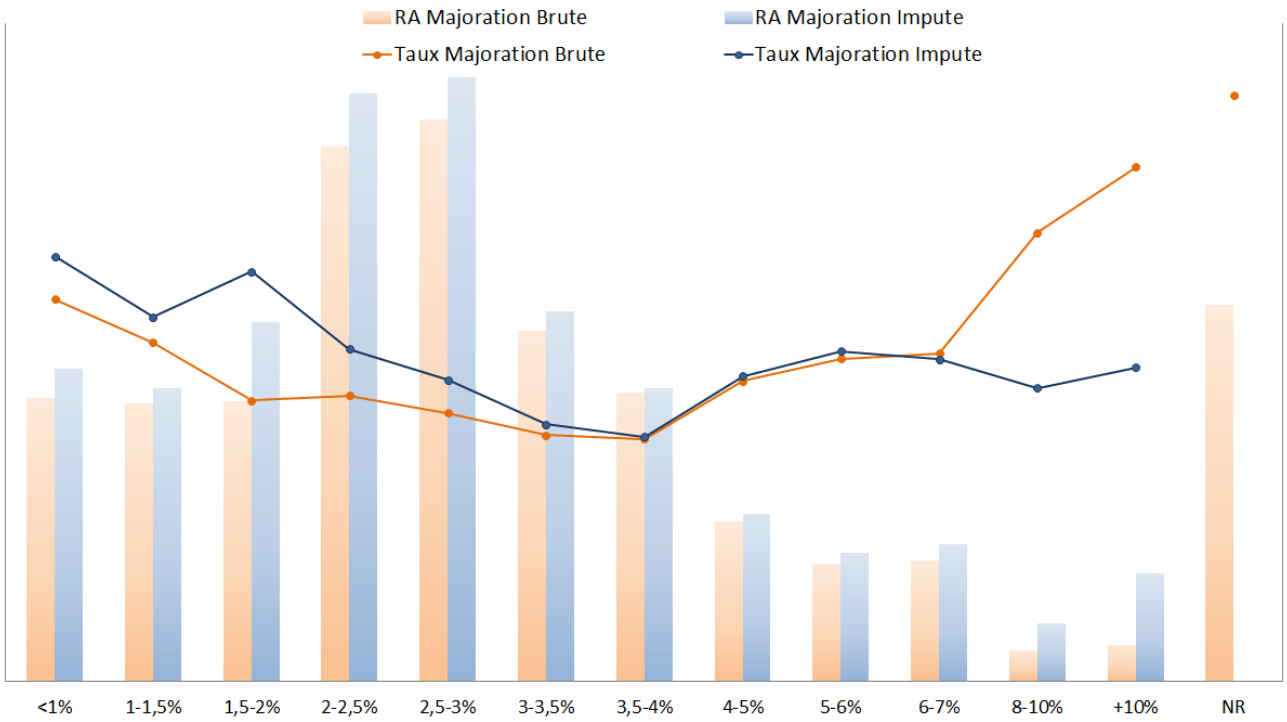

FIGURE 6.17: Comparaison du taux de résiliation de la variable évolution tarifaire avant et après imputation

Les résultats sont bien moins concluant que pour la variable segment client. Le taux de résiliation moyen de certaines modalités de la variable a été désaxé de son taux moyen observé avant l'imputation. La modalité 1.5-2% a subi une hausse importante, tandis que le taux moyen des modalités avec une évolution tarifaire de plus de 8% a fortement diminué.

Ce biais créé par l'imputation s'explique par le fait que la modalité NR n'est pas une modalité non renseignée aléatoirement comme pouvait l'être le segment client. En effet, les 800k lignes non renseignées sont des Affaires Nouvelles (AN) (640k) et des AVeNants (AVT) (160k). La façon dont est créée la variable d'évolution tarifaire (cf. figure 3.4) nécessite que le contrat ait déjà eu une majoration à son terme. Il est alors impossible pour une AN d'avoir cette variable de renseignée. De même pour les situations avenantées dont la seule situation d'historique est une affaire nouvelle.

La variable d'évolution tarifaire n'est donc pas imputée et la modalité NR est conservée, car celle-ci donne de l'information sur la situation du client. L'algorithme d'imputation ne peut pas déterminer de profil type pour affecter une évolution tarifaire correcte.

6.3.4 Traitement de la variable majoration sinistre

La variable MAJO_SINI possède également un nombre important de valeurs non renseignées. Environ 1.2 millions de lignes ne sont pas renseignées soit 17% de la base. Cependant, parmi les valeurs non renseignées, aucune information n'en ressort à l'inverse de l'évolution tarifaire. Nous retirons cette variable du modèle afin d'éviter les biais que celle-ci peut engendrer comme a pu le faire la variable segment client. Le GLM compte désormais 31 descripteurs.

6.3.5 Performances du GLM avec 31 variables retraitées

Le GLM exécuté sur la nouvelle base paraît à première vue moins performant en terme d'AUC. La nouvelle variable client n'est plus biaisée par le flag résiliation. Précédemment, du fait du mauvais rapprochement, le fait d'être résilié accroissait les chances d'avoir la variable segment client non renseignée. Cette information NR était ensuite utilisée par le modèle pour estimer le taux de résiliation. Un biais endogène existait, car le fait d'être résilié induisait une forte probabilité d'avoir la modalité NR qui était par la suite utilisée par le modèle. Ce processus gonflait les performances du GLM. Ce nouveau modèle est donc de meilleure qualité que l'ancien malgré son AUC plus faible.

— AUC base d'apprentissage : 72.46%

— AUC base de validation : 72.12%

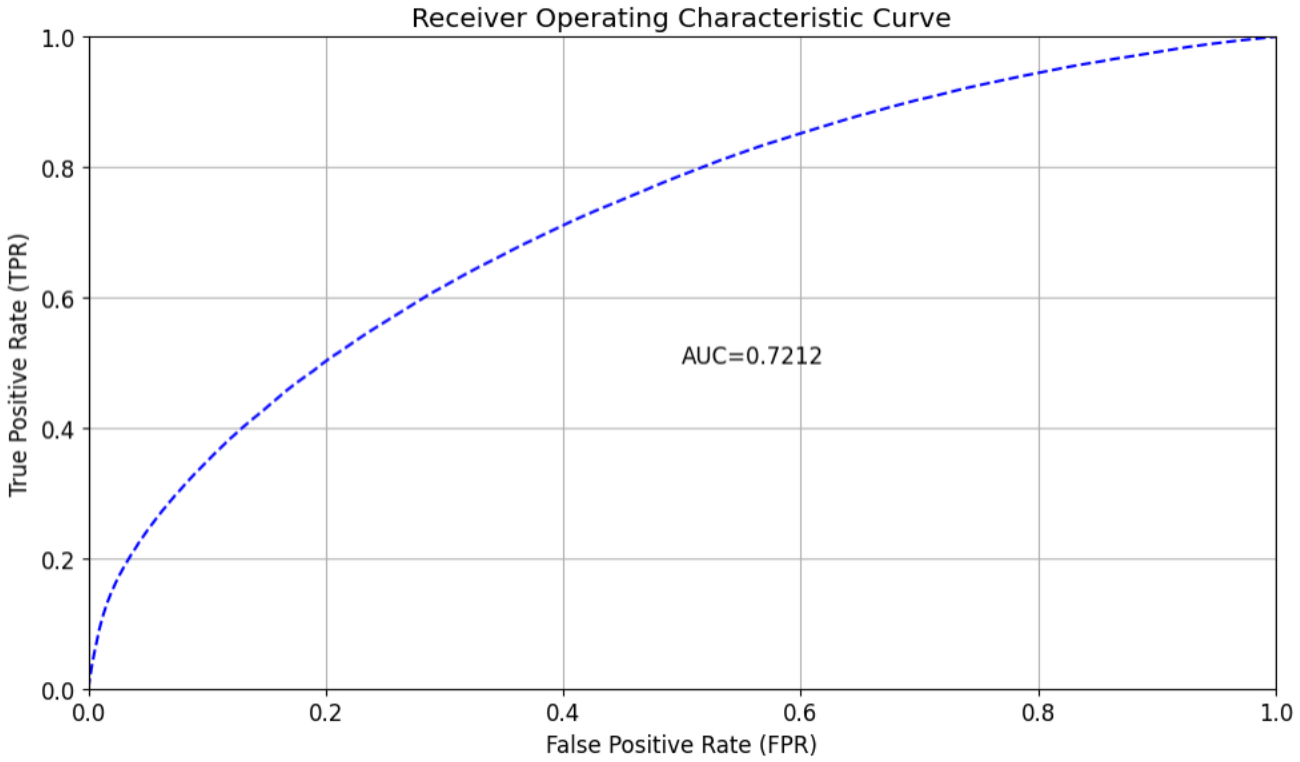

FIGURE 6.18: AUC GLM avec les 31 variables catégorisées et retraitées

6.4 Sélection 2 : Modélisation pénalisée

6.4.1 Principes de la pénalisation

La deuxième façon de réduire le nombre de variables est d'utiliser un modèle pénalisé. Un trop grand nombre d'estimateurs peut entraîner une variance élevée de l'estimation, car les coefficients estimés sont très erratiques. La régression pénalisée repose sur la fonction de vraisemblance précédemment énoncée avec un terme de pénalisation supplémentaire.

$$\max_{C(X)} \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left[y_i \times C(X) - ln(1+C(X)\right]}_{Vraisemblance} - \underbrace{\lambda\left[a||\alpha||_1 + \frac{1}{2}(1-a)||\alpha||_2^2\right]}_{Fonction\ de\ penalite} \tag{6.11}$$

Ce terme supplémentaire impose une contrainte sur les paramètres estimés de la régression. Plusieurs méthodes de pénalisation existent :

— LASSO : Cette méthode utilise la norme L1 pour pénaliser les coefficients
— RIDGE : Cette méthode utilise la norme L2 pour la pénalisation.
— ELASTIC-NET : C'est une combinaison des deux précédentes méthodes.

Chacune de ses méthodes à la même fonction : réduire la taille des estimateurs.

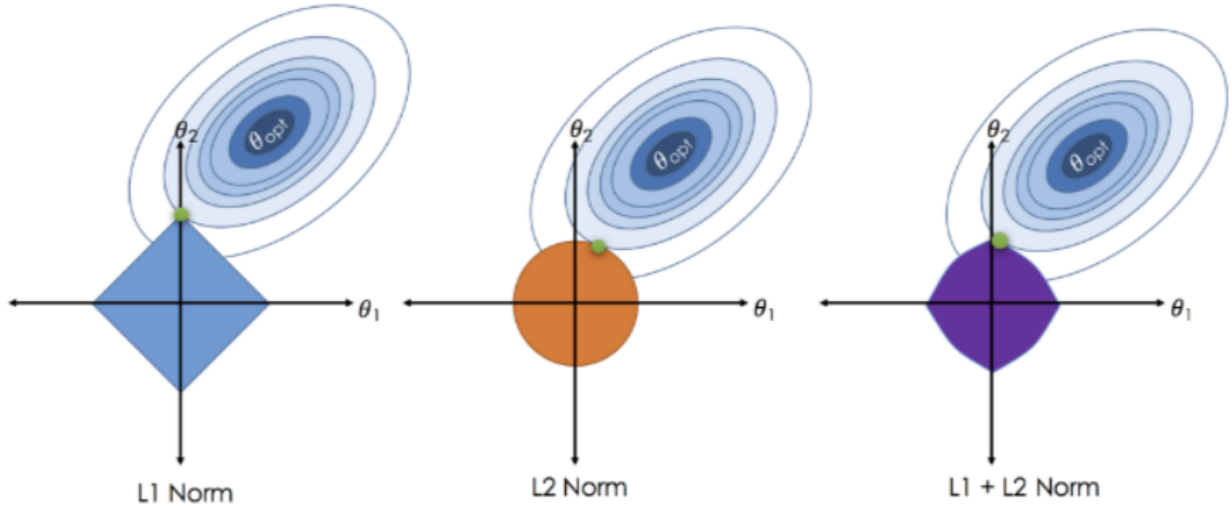

FIGURE 6.19: Pénalisation en fonction de la méthode choisie

Dans H2O, deux paramètres contrôlent la pénalisation du modèle :

— λ : Le coefficient de pénalité qui ajuste l'impact de la pénalisation. Plus λ est grand, plus les coefficients α tendent vers 0.
— a : Le paramètre de choix de méthode entre Lasso ($a = 1$) et Ridge ($a = 0$).

Afin d'optimiser la sélection de variables, nous avons testé différentes combinaisons de paramètres. La régression pénalisée réduit l'AUC du fait de l'ajout du terme de pénalisation dans la fonction coût. Nous avons dû trancher entre conserver un modèle avec une AUC importante tout en pénalisant un minimum la fonction de coût pour réduire les coefficients des variables explicatives.

La régression Ridge offre de meilleur résultats en terme d'AUC que la regression Lasso ou Elastic-Net. Néanmoins, la régression Ridge se charge de réduire la taille des paramètres, tandis que Lasso va encore plus loin en permettant de réduire facilement certains coefficients à zéro, ce qui les élimine du modèle. En effet, la norme L2, utilisée pour la pénalisation Ridge, est plus conciliante envers les coefficients. Sur la figure précédente, la norme L1 pousse le coefficient θ_1 à 0 et impose une faible restriction sur θ_2, alors que la norme L2 impose une restriction modérée sur les deux coefficients. Contrairement à la méthode Ridge, la méthode Lasso permet de retirer la variable liée au coefficient θ_1.

La régression Lasso est donc privilégiée pour la sélection de variables. le paramètre a est fixé à 1 et le paramètre λ est fixé à 0.001.

6.4.2 Application de la pénalisation

Les variables explicatives dont les coefficients sont devenus nuls ont été retirées de la base d'étude. De plus, le modèle affiche l'importance relative des variables dans la modélisation. A partir de celle-ci, les variables avec un pourcentage d'importance très faible sont soustraites de la base. Un total de 15 variables est ainsi retiré :

- Critères isolement ;
- Usage risque (RS ou RP) ;
- Flag surtarif (1 si la cotisation nette est supérieure à la cotisation barème) ;
- Nombre de pièces supérieures à 40m² ;
- Dernière évolution des leviers tarifaires ;
- Montant MBA ;
- Nombre d'affaires MRH ;
- Nombre d'affaires AUTO ;
- Nombre d'affaires PRO ;
- Nombre de sinistres sur 4 ans ;
- Coût des sinistres sur 4 ans ;
- Nombre d'agence dans les 2 km ;
- Score de rentabilité à l'iris ;
- Dernière évolution 2021 par rapport à 2020 Barème calculée à l'iris ;
- Nombre de résidences principales à l'iris ;

6.4.3 Performances du GLM avec 16 variables retraitées

A la suite de ces suppressions, le modèle est légèrement moins performant. Néanmoins, malgré la suppression de 15 variables, soit près de la moitié, l'AUC du GLM avec seulement 16 variables sur la base validation a diminué de seulement 0.18%.

- AUC base d'apprentissage : 72.29%
- AUC base de validation : 71.94%

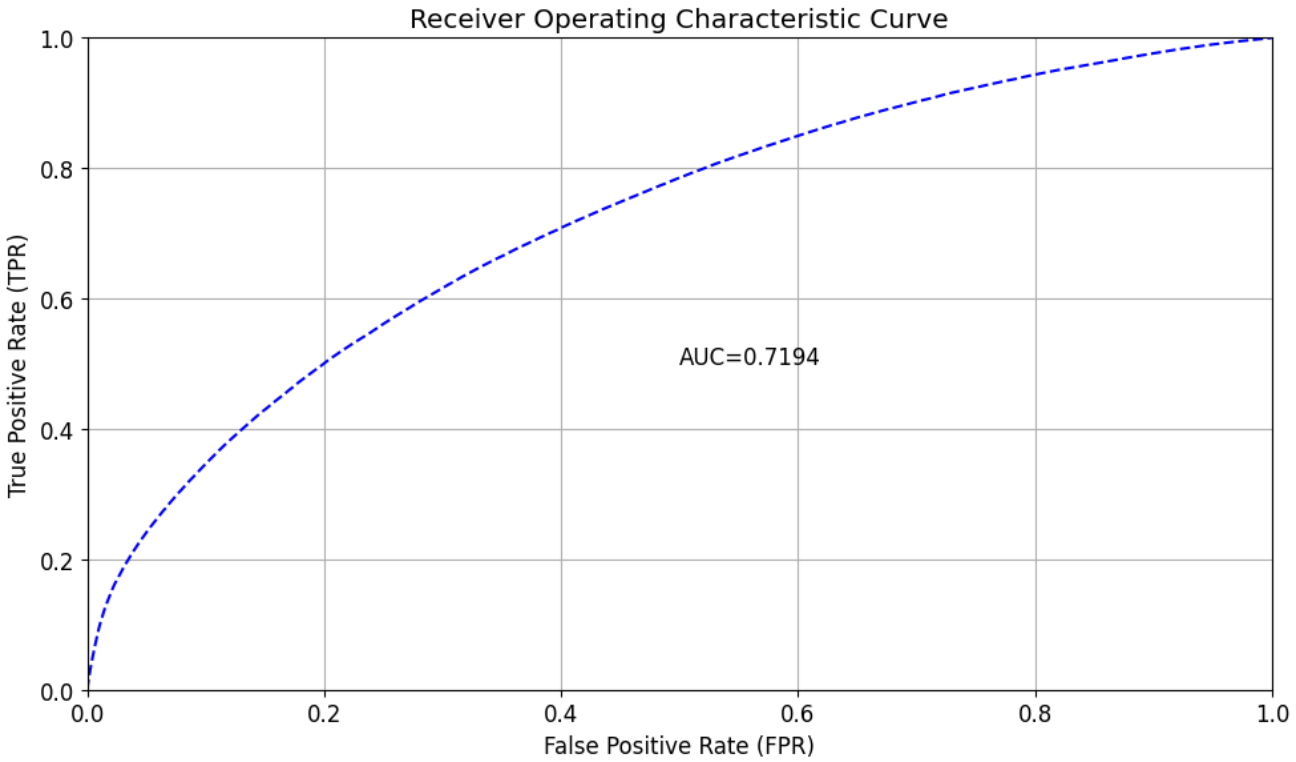

FIGURE 6.20: AUC GLM 16 variables catégorisées et retraitées

6.4.4 Validation avec le Machine Learning

Un grand nombre de variables a été soustrait de la base suite à la modélisation pénalisée. A l'aide d'un modèle de Gradient Boosting Machine (GBM) avec des paramètres basiques, nous vérifions si aucune potentielle variable explicative n'a été retirée. Le modèle est lancé avec les même 31 variables explicatives mais cette fois-ci brute, les variables ne sont pas catégorisées comme pour le GLM. Les modèles de Machine Learning retraitent généralement eux même les variables explicatives.

	variable	relative_importance	scaled_importance	percentage
0	Generation	63334.148438	1.000000	0.280317
1	CD_TYPExQLTExPIECE	46002.324219	0.726343	0.203606
2	VA_COEF_BM_MRH	23830.519531	0.376267	0.105474
3	NATU_SITU	17332.320312	0.273665	0.076713
4	SEGM_PART_PERSO	15294.788086	0.241494	0.067695
5	CAPI_MOBI	14124.625000	0.223018	0.062516
6	evol_net_fin_n	12368.297852	0.195286	0.054742
7	DETENTION_n	7552.062012	0.119242	0.033425
8	cot_net_n	7087.791992	0.111911	0.031371
9	nb_affa_mrh_n	5970.715820	0.094273	0.026426
10	evol_MRH	1845.608765	0.029141	0.008169
11	nb_pro_ent_n	1761.411743	0.027811	0.007796
12	evol_AUTO	1670.138916	0.026370	0.007392
13	nb_option	1560.828491	0.024644	0.006908
14	MT_DRGT_MBA_TTC	1452.166748	0.022929	0.006427
15	FORMULE	1315.886963	0.020777	0.005824

FIGURE 6.21: Première moitié du tableau d'importance des variables du GBM avec 31 variables

Le classement des descripteurs les plus importants confirme les suppressions réalisées par la régression pénalisée, excepté pour la variable nombre d'affaires MRH qui ressort dans le top 10. Dans la suite, nous croisons les deux sélections pour réduire de nouveau le nombre de variables explicatives.

Info	Nom de la variable	Sélection Pénalisée	Sélection GBM	Test significativité	Modèle final (12vars)
Habi	CD_TYPExQLTExPIECE	X	X		X
	Taille_DPDC	X			
Tarif	cot_net	X	X		X
	evol_net_fin_n (Dernière évolution)	X	X		X
	evol_net_fin_n1-n3	X			
	VA_COEF_BM_MRH	X	X		X
Environnement commercial	nb_affa_mrh_n		X		X
	evol_AUTO (Dernière évolution)	X		X	X
	evol_MRH (Dernière évolution)	X			
	DETENTION	X	X		X
	FORMULE	X			
	nb_option	X		X	X
	CAPI_MOBI	X	X		X
	nb_30km	X			
	SEGM_PART	X	X		X
Autre	Generation	X	X		X
	NATU_SITU	X	X		X

FIGURE 6.22: Tableau synthèse de la sélection de variable

Les 10 variables les plus significatives issues du GBM sont conservées pour l'étude. La variable NB_AFFA_MRH est la seule à être de nouveau utilisée, hormis celle-ci les sélections par régression pénalisée et Machine Learning sont en adéquation.

La significativité des autres descripteurs est testée une à une. La suppression de certaines variables n'entraîne que très peu de baisse de performance sur la base validation en terme d'AUC :

- EVOL_MRH : Baisse de 0.01%
- TAILLE_DPDC : Baisse de 0.01%
- FORMULE : Baisse de 0.02%
- EVOL_NET_FIN_N1-N3 : Baisse de 0.03%
- NB_30KM : Baisse de 0.03%

Ces 5 variables (en orange) sont retirées de l'étude, tandis que les variables (en bleu) NB_OPTION et EVOL_AUTO sont conservées. Le dernier GLM est composé des 12 variables de la dernière colonne du tableau ci-dessus (cf. figure 6.22).

6.4.5 Performances du GLM avec 12 variables

Le GLM final est composé de seulement 12 variables. Cette baisse drastique se justifie par le fait que la suppression de 19 descripteurs a diminué de seulement 0.0028 l'AUC par rapport au GLM à 31 variables.

- AUC base d'apprentissage : 72.19%
- AUC base de validation : 71.84%

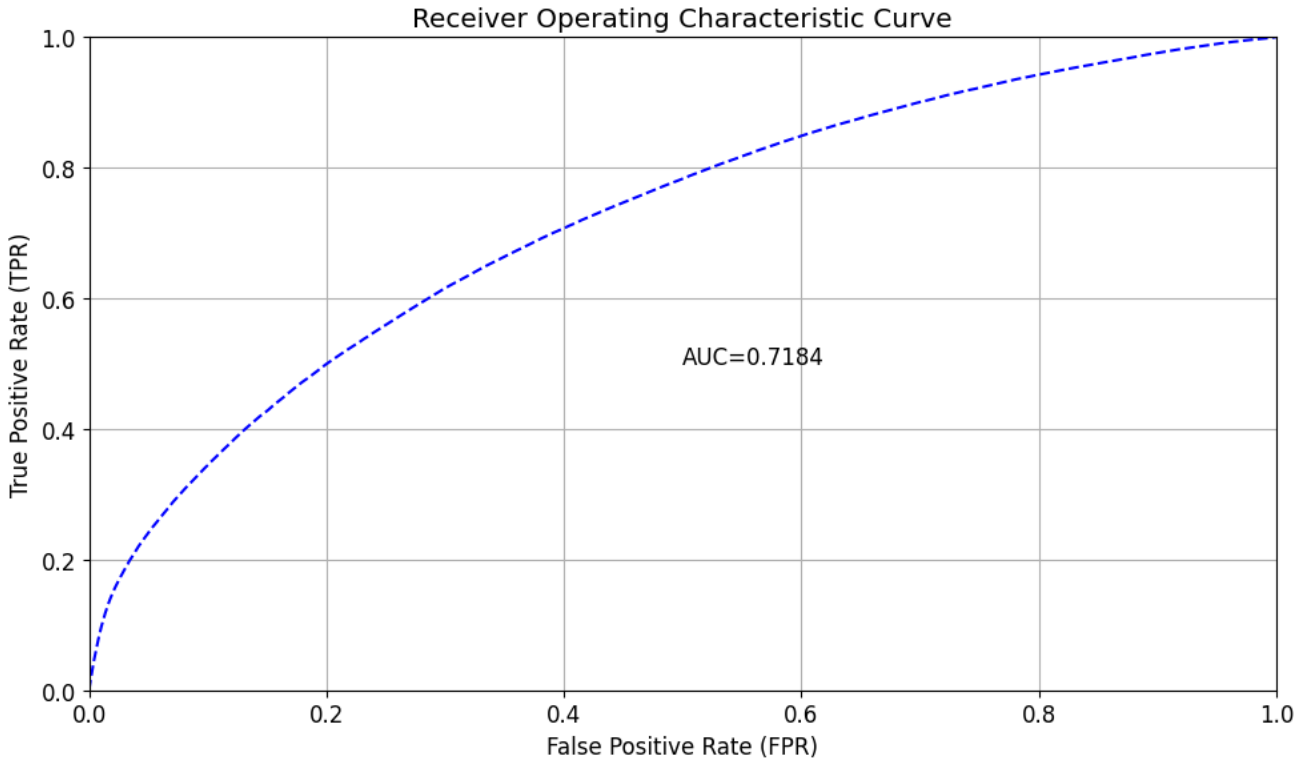

FIGURE 6.23: AUC GLM 12 variables catégorisées et retraitées

6.5 Modélisation et résultats du GLM

Au total 5 GLM avec chacun leurs particularités ont été lancés sur la base d'étude. Deux GLM avant le retraitement de certaines variables et trois après ce retraitement avec plus ou moins de descripteurs. Au regard de la performance seule, les modèles sans le retraitement des variables sont bien meilleurs. Néanmoins, actuariellement il est impensable de retenir un modèle avec un tel biais pour nos futures études. Le choix se dirige alors vers les trois modèles après retraitements. Le faible écart de performance et la simplification du nombre de descripteurs rendent la sélection du modèle à 12 variables logique. Le fait de retirer les variables peu explicatives, rend les restantes plus discriminantes, or les 12 descripteurs restants sont des critères fortement utilisés opérationnellement.

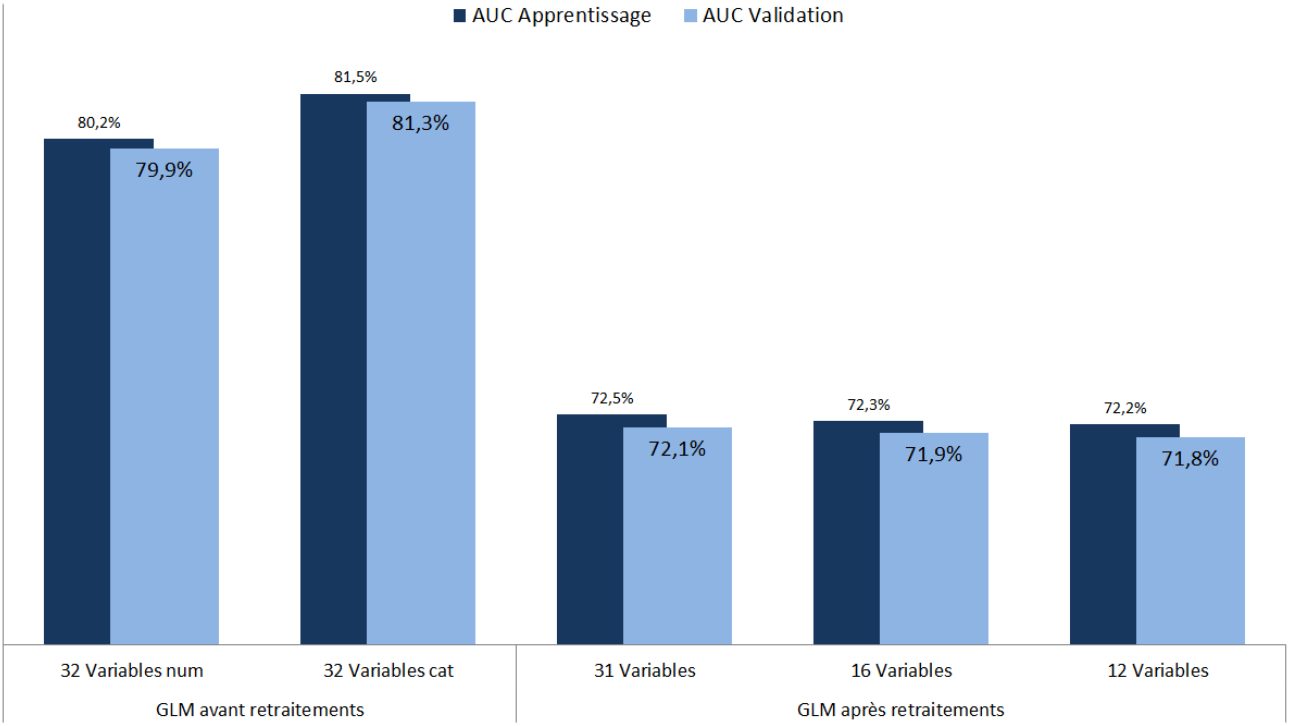

FIGURE 6.24: AUC GLM 12 variables catégorisées et retraitées

6.5.1 Probabilité de résiliation GLM

Le GLM avec 12 variables est utilisé pour prédire les résiliations sur la base de validation. Les prédictions sont comparées avec les résiliations réellement observées. La base validation contient environ 1.2 millions de lignes, ce qui est largement suffisant pour mesurer la qualité de prédiction du modèle. La première chose que le modèle prédit est la probabilité de résiliation. C'est à partir de cette probabilité que celui-ci va ainsi classer les situations entre résiliées ou non. Pour déterminer ce score le modèle utilise un seuil, le score prend alors les modalités :

— 0 : Situation prédite comme non résiliée ;
— 1 : Situation prédite comme résiliée ;

Dans la suite, la probabilité de résiliation va être comparée avec le taux de résiliation observé. Cette comparaison permet de mesurer l'adéquation de notre modèle avec la réalité.

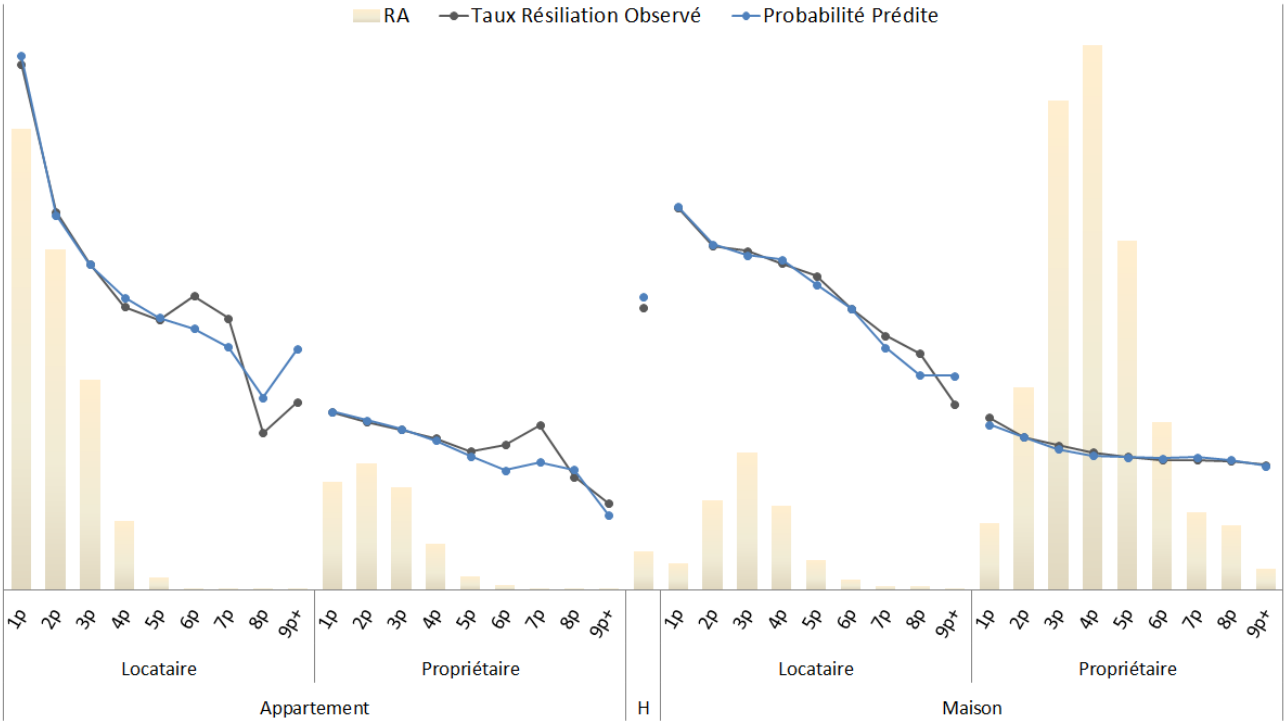

FIGURE 6.25: Probabilité prédite par le GLM vs les résiliations observées en fonction de la variable croisée

Sur le graphique ci-dessus, nous remarquons que les probabilités prédite par le GLM sont extrêmement proches des résiliations observées. Deux phénomènes expliquent ces écarts minimes :

— Le GLM se base sur la moyenne pour calculer les probabilités de résilier et estimer ses coefficients. Lorsque l'on agrège les résultats on a alors des résultats très proches du taux observé.
— Le fait d'agréger les probabilités qui sont comprises entre 0 et 1 plutôt que d'agréger le score permet de lisser les erreurs. La probabilité prédite est beaucoup moins segmentante que le score prédit.

6.5.2 Classification GLM

Toute la difficulté d'une classification binaire réside dans la prédiction du score, c'est-à-dire de prédire 0 ou 1. Les résultats précédents sont parfaitement ajustés sur le taux de résiliation observé. La probabilité va lisser les erreurs, elle est beaucoup moins discriminante que le score prédit par le GLM. C'est le score qui permet de mesurer la vraie qualité d'un modèle. Pour le prédire le modèle fonctionne de la manière suivante :

1. Entraînement et estimation des coefficients du modèle logistique
2. Prédiction d'une probabilité de résiliation pour chaque ligne de la base validation
3. Détermination du seuil à partir duquel sera prédit le score
4. Calcul du score : si la probabilité prédite de la situation est supérieure à ce seuil le modèle prédit 1, c'est-à-dire une résiliation. A l'inverse, si la probabilité est inférieure au seuil le modèle prédit 0.

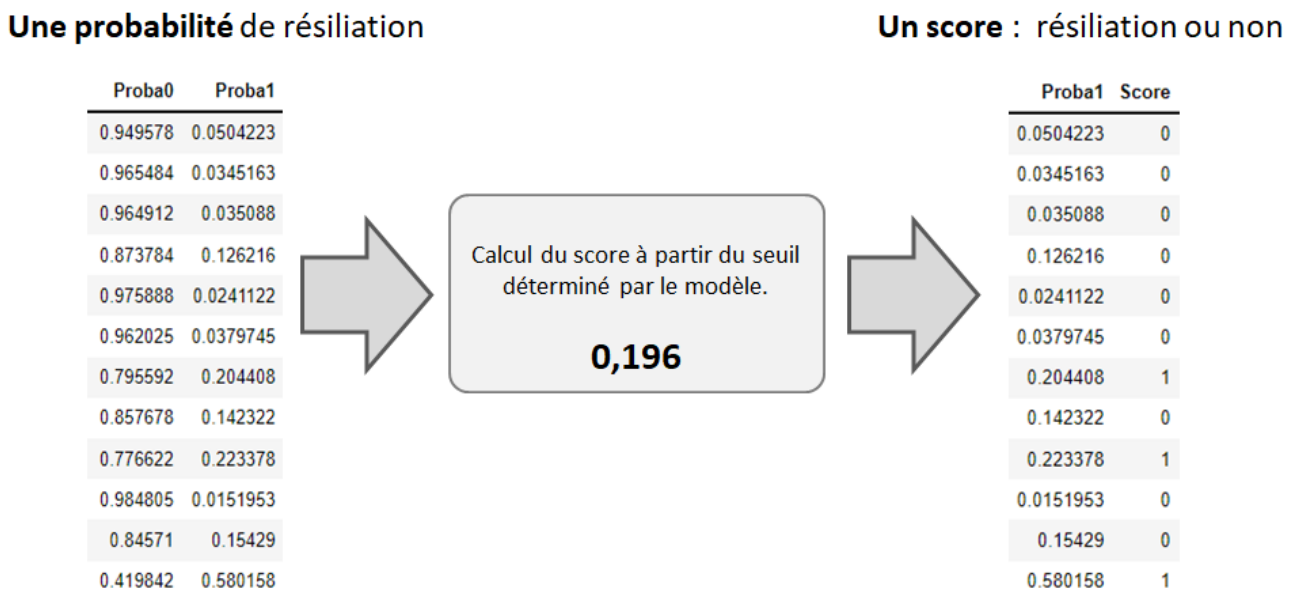

FIGURE 6.26: Calcul du Score à partir de la probabilité prédite

Pour estimer la qualité du modèle la matrice de confusion est un bon outil. Elle permet de comptabiliser les scores prédits par le modèle avec les résiliations observées. En observant cette matrice nous remarquons que le nombre de résiliations prédites par le modèle est bien plus important que les résiliations observées.

Le seuil joue un rôle essentiel dans la quantité de résiliations prédites. Le seuil F1 est le plus utilisé, ce seuil accorde plus d'importance aux erreurs sur la prédiction de situations résiliées. Le nombre de situations résiliées étant plus faible que le nombre de situations non résiliées, minimiser les erreurs sur les prédictions pousse le modèle à prédire plus de résiliations pour augmenter les chances de bien prédire une situation réellement résiliée.

Le second seuil utilisé est le MCC. Ce seuil minimise l'erreur sur l'ensemble de la base : situation résiliées et non résiliées. Le nombre de situations non résiliées étant plus important, ce seuil MCC est plus grand que le seuil F1. Il prédit alors un nombre moins important de situations résiliées.

Nous avons également déterminé un seuil personnalisé afin de prédire le même nombre de prédictions que la réalité. Ce seuil est alors situé entre les deux précédents. Aucune métrique sous H2O permet d'obtenir un seuil qui prédit autant de résiliations que la réalité.

		Nombre résiliations	Ecart avec l'observé
Observé	Nombre observé sur la base de validation	116 264	-
Seuil F1	Minimise l'erreur sur les prédictions c'est-à-dire sur les 1	187 131	70 867
Seuil MCC	Minimise l'erreur globale	89 497	- 26 767
Seuil Perso	Seuil déterminé pour obtenir le nombre de résiliations observées	116 468	204

FIGURE 6.27: Nombre de résiliations avec les différents seuils de prédiction

Les écarts entre les seuils F1 / MCC et la résiliations observées sont assez conséquents. Le seuil personnalisé est très proche de la réalité, néanmoins ce seuil varie à chaque nouveau modèle et dépend de la base de validation. Il est alors très difficile d'industrialiser ce seuil, car il est fixé manuellement contrairement aux deux précédents fixés par le modèle. Le seuil personnalisé peut aussi être un outil pour augmenter ou diminuer le nombre de résiliations prédits en fonction des tendances observées.

Pour estimer plus efficacement le GLM, il est nécessaire de comparer le score avec les résiliations observées sur la base de validation. Sur les graphiques suivants, le taux de résiliation prédit n'est plus la probabilité issue du modèle mais le nombre de prédictions divisé par les risques années.

— En noir : Taux de résiliation observé.
— En bleu : Taux prédit à l'aide du seuil F1 qui prédit un nombre plus important de résiliations.
— En vert : Taux prédit à l'aide du seuil MCC qui prédit un nombre moindre de résiliations.
— En bleu : Taux prédit à l'aide du seuil personnalisé qui prédit le même nombre de résiliations que celles observées.

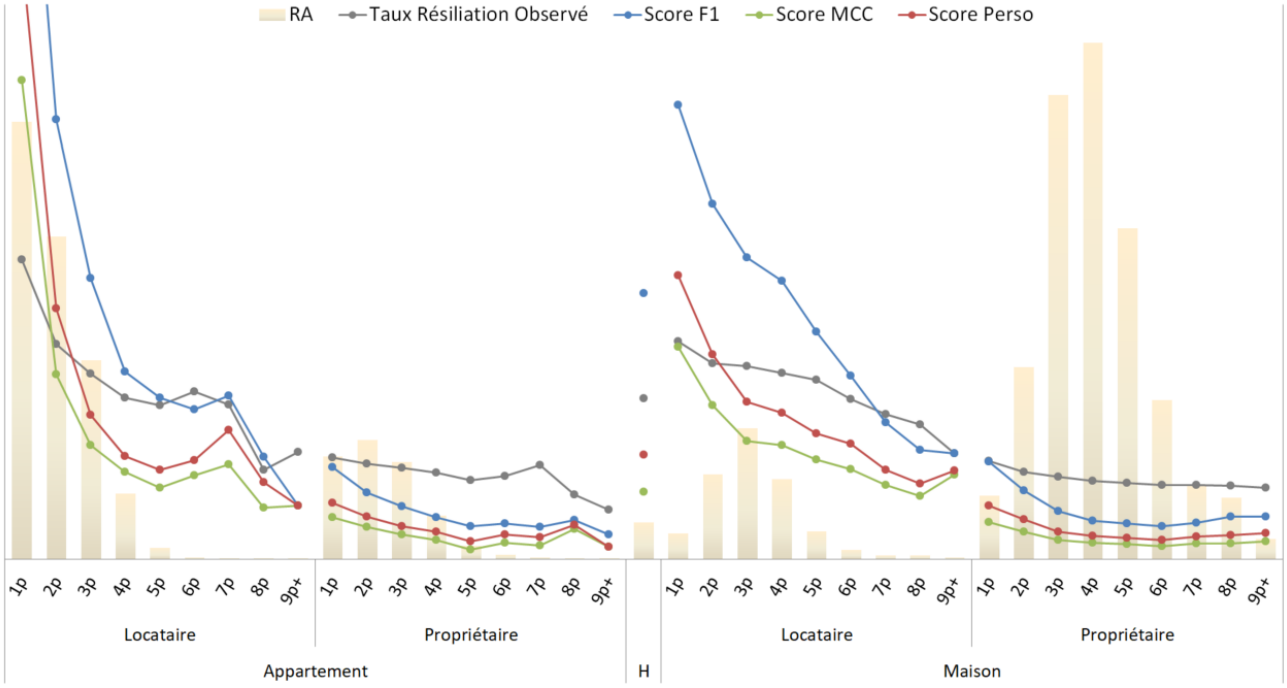

FIGURE 6.28: Score prédit par le GLM vs les résiliations observées en fonction de la variable croisée

Le modèle a une forte tendance à amplifier le phénomène de résiliation sur les modalités avec un taux de résiliation assez important. L'amplification se remarque sur les biens en location avec peu de pièces : Appartement Location 1 pièce, Appartement Location 2 pièces, Maison Location 1 pièce, Maison Location 2 pièces. A l'inverse, le modèle sous estime les résiliations sur les modalités avec un faible taux de résiliation observé. Plus le taux de résiliation observé est éloigné du seuil, plus l'erreur de prédiction sera importante. Le modèle logistique a alors une forte tendance à pénaliser les profils qui résilient et inversement protéger les profils fidèles.

Ce phénomène de pénalisation et de protection se retrouvent quelque soit le critère étudié. Sur la variable croisée, ce phénomène est très important, tandis que sur des critères avec moins de volatilité entre les modalités celui-ci l'est beaucoup moins. Par exemple, sur le critère d'évolution tarifaire, le

modèle arrive à prédire collectivement les résiliations.

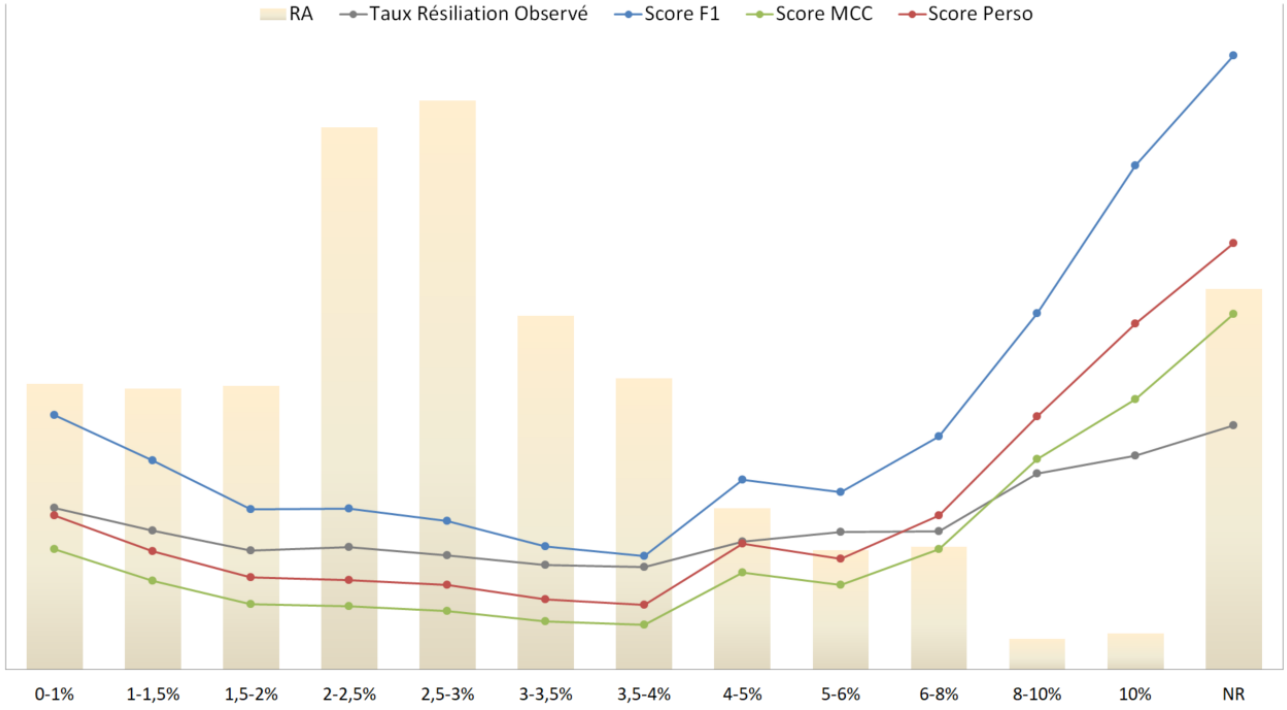

FIGURE 6.29: Score prédit par le GLM vs les résiliations observées en fonction de la variable d'évolution tarifaire

Le taux de résiliation entre le score et l'observé sont ici très proche. Il y a seulement une légère dérive à partir de 8% d'évolution tarifaire. Cette dérive s'observe surtout pour le seuil F1 qui a tendance a fortement pénaliser ces profils qui résilient plus que la moyenne.

Le modèle logistique est le modèle le plus couramment utilisé mais aussi le plus simpliste dans sa compréhension. Les performances de celui-ci restent toutefois très acceptables. L'AUC du modèle est proche de 72%, ce qui représente un score d'un modèle qui arrive a bien discriminer les profils qui résilient. A ce stade, le modèle logistique est difficilement perfectible. Ce dernier modèle est assez simpliste, il compte seulement 12 variables explicatives et malgré ce faible nombre la performance de celui-ci est tout à fait acceptable. L'objectif de la suite des travaux est de challenger ce modèle logistique avec des modèles de Machine Learning. Les modèles de Machine Learning ont une architecture plus complexe ce qui leur permet de gagner en performance au prix de la simplicité.

Chapitre 7

Gradient Boosting

Sommaire

Cette partie s'intéresse à un autre type de modélisation plus avancé que la régression logistique : le Gradient Boosting Machine (GBM). Cette méthode utilise de manière judicieuse les arbres de décisions pour former un modèle plus complexe mais généralement plus performant que le GLM.

7.1 Rappels théorique

7.1.1 Bagging et Boosting

Le modèle GBM utilise un algorithme qui repose sur le Boosting pour apprendre et affiner ses prédictions. Pour comprendre le Boosting, il est plus simple de définir le Bagging en premier lieu.

Dans notre cas, le Bagging est une technique qui consiste à créer indépendamment, puis assembler un grand nombre d'arbres de classification afin de construire un modèle plus performant. Ces arbres de petite taille et donc de faible performance sont nommés "weak learners". Leur force est qu'une fois mis en relation, ces weak learners forment un seul et grand modèle capable d'être très performant. Chaque arbre va émettre une prédiction et la réponse finale du modèle sera la moyenne de toutes les prédictions des weak learners. L'algorithme de Bagging le plus célèbre est le Random Forest.

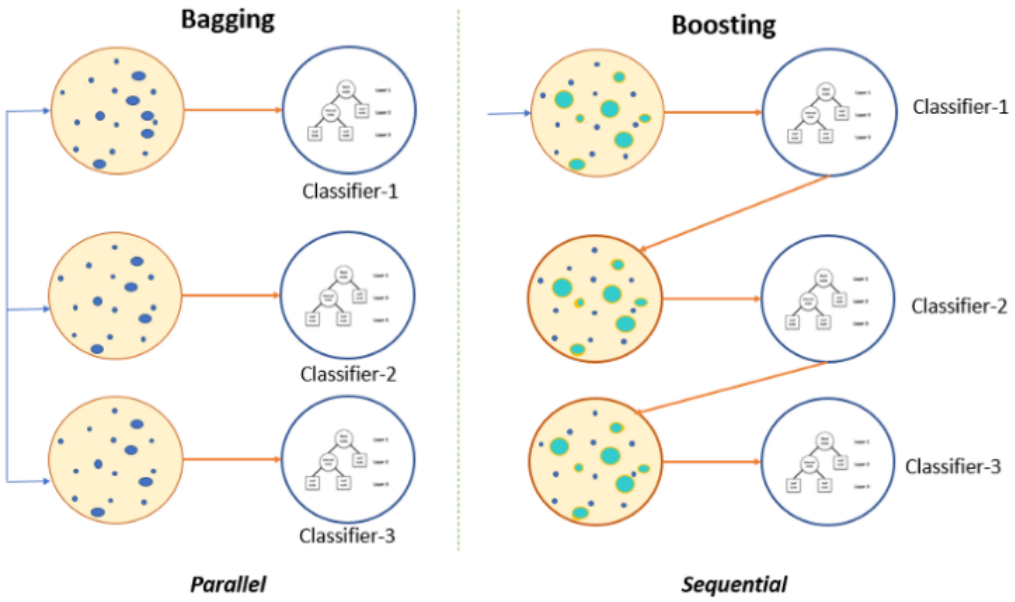

FIGURE 7.1: Algorithme de Bagging vs Boosting

Le Boosting repose sur le même principe que la Bagging, à savoir l'agrégation de plusieurs arbres de classification, cependant pour le Boosting la création de ces arbres n'est plus indépendante. Les modèles de Boosting travaillent alors de manière séquentielle, itération par itération. Entre chacune de ces itérations, l'algorithme analyse les données pour augmenter le poids des observations difficiles à classifier, ou à l'inverse diminuer le poids de celles dont la classification est simple. Par conséquent, l'arbre suivant va s'attarder à corriger les erreurs de classification du modèle d'ensemble construit à partir des weak learners précédents.

7.1.2 Gradient Boosting Machine

L'algorithme de Gradient Boosting Machine est un cas particulier d'algorithme de Boosting.

Le premier arbre construit se base simplement sur la moyenne des observations, il est donc peu efficace. Par la suite, l'algorithme calcule le résidu, à savoir l'écart entre la prédiction et la valeur observée.

Le second arbre est ensuite entraîné pour prédire le résidu calculé avec le premier weak learner. Les prédictions réalisées sont multipliées par un facteur inférieur à 1 afin d'être précis dans la recherche de l'optimum.

Ensuite cette opération est répétée pour créer de nouveaux arbres jusqu'à ce qu'une condition d'arrêt soit atteinte, comme par exemple un nombre d'itérations maximum fixé préalablement. Chaque itération se décompose de la manière suivante :

1. Calcul des résidus avec le modèle d'ensemble ;
2. Construction d'un arbre de classification faible (weak learner) ;
3. Calibration de cet arbre pour prédire les résidus ;
4. Ajout de l'arbre au modèle d'ensemble ;

Le schéma ci-dessous résume bien la construction itérative le modèle d'ensemble :

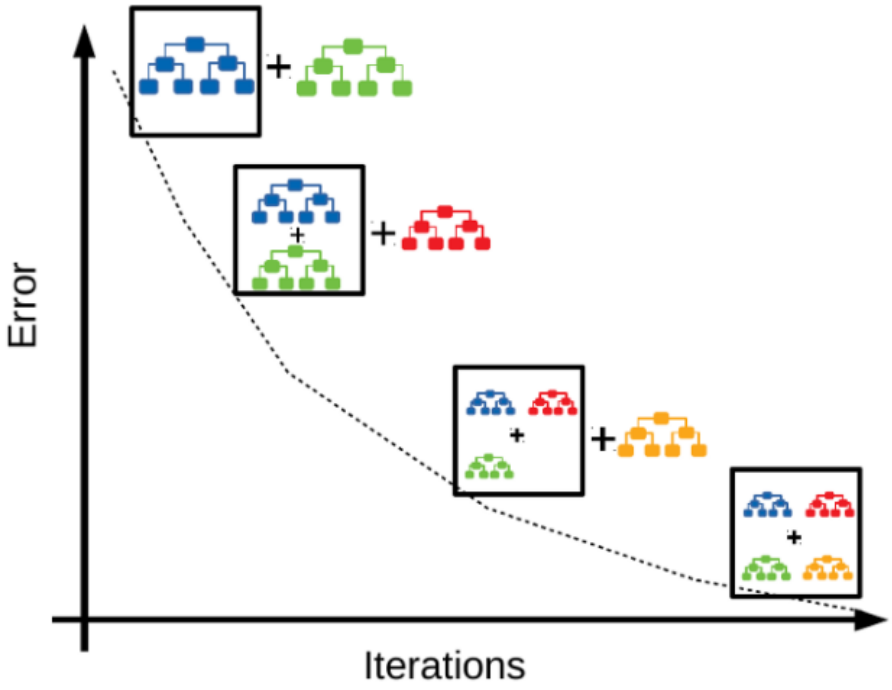

FIGURE 7.2: Agrégation des arbres Gradient Boosting

Sur H2O l'algorithme du Gradient Boosting se décompose en plusieurs étapes. Dans notre cas binaire $K = 2$.

1. Initialisation pour chaque classe k :

$$f_{k0} = 0, k = 1, 2, ..., K$$

Pour l'ensemble des itérations de $m = 1$ à M :

2. Calcul de la probabilité conditionnelle d'appartenir à la classe K

$$p_k(x) = \frac{e^{f_k(x)}}{\sum_{l=1}^{K} e^{f_l(x)}}, k = 1, 2, ..., K \tag{7.1}$$

3. Pour $k = 1$ à K :
   (a) Calcul des résidus

$$r_{ikm} = y_{ikm} - p_k(x_i), i = 1, 2, ..., N \tag{7.2}$$

   (b) Entraînement d'un arbre de classification sur les résidus r_{ikm} avec comme noeuds terminaux $R_{jim}, j = 1, 2, ..., J_m$
   (c) Calcul

$$\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{x_i \in R_{jkm}} r_{ikm}}{\sum_{x_i \in R_{jkm}} |r_{ikm}|(1 - |r_{ikm}|)}, j = 1, 2, ..., J_m \tag{7.3}$$

(d) Calcul de la prédiction à l'itération m

$$f_{km}(x) = f_{k,m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jkm} I(x \in R_{jkm}) \tag{7.4}$$

Fin de la boucle sur m

4. Le résultat final du modèle d'ensemble est $\hat{f}_k(x) = f_{kM}(x), k = 1, 2, ..., K$

7.2 Optimisation des paramètres

Pour l'optimisation du modèle de Gradient Boosting, nous avons utilisé les mêmes données que pour la régression linaire.

7.2.1 Validation croisée

L'échantillon d'apprentissage représente 80% de la base et l'échantillon de validation regroupe les 20% restants (cf. figure 6.10). L'algorithme d'optimisation intègre de la validation croisée, en d'autres termes le modèle utilisera une partie de la base d'apprentissage afin de valider ses résultats. La base d'apprentissage est alors découpée en 4 parts égals. Une de ces parties sera utilisée comme échantillon de validation et le reste conservera ces fonctions de données d'entraînement.

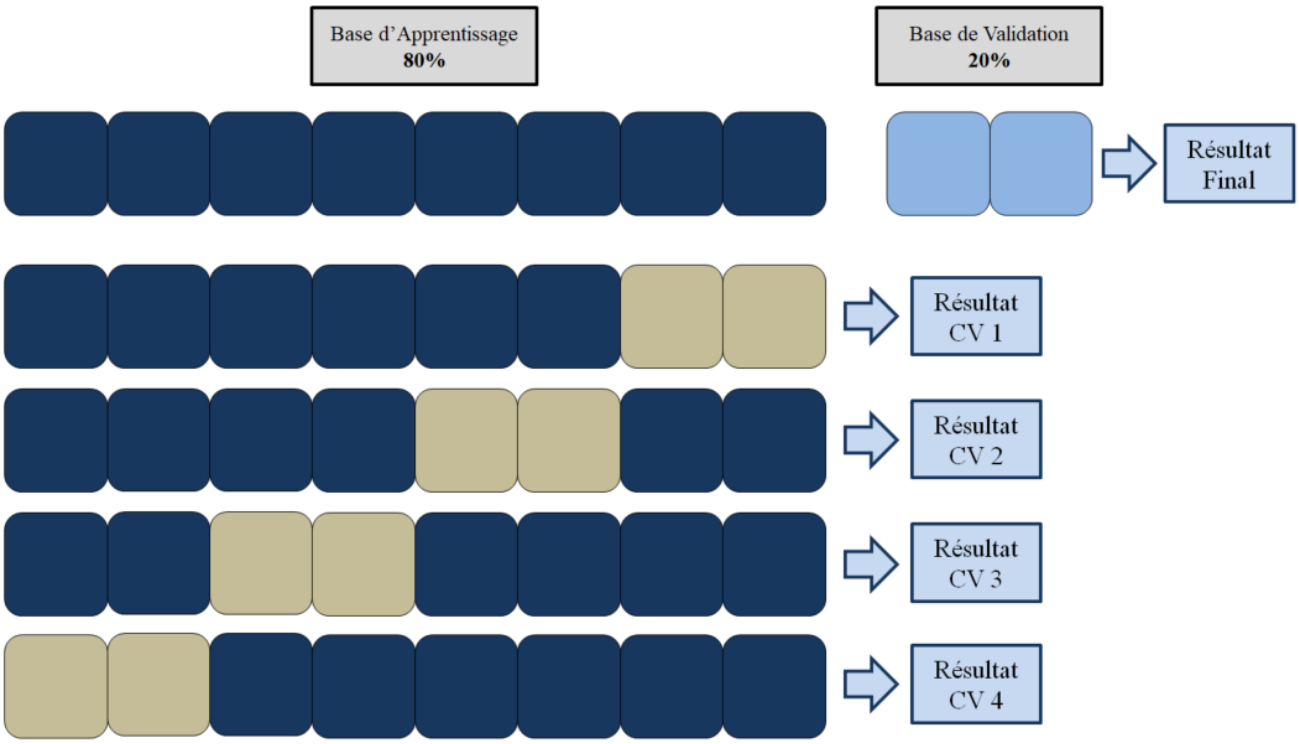

FIGURE 7.3: Validation croisée

Cette méthode permet de rendre les résultats plus robustes car le modèle est entraîné diverses fois sur des bases plus modestes. De plus, la validation est déterminée sur plusieurs échantillons ce qui évite ainsi que le modèle s'accommode trop à un échantillon en particulier.

7.2.2 Sélection des variables

Une des spécificités du modèle GBM est qu'il est capable de traiter un grand nombre de variables mêmes si celles-ci sont corrélées entre elles. Le problème qui se posait au chapitre précédent et qui nous a obligé à réduire le nombre de variables à 31 ne se pose plus.

Toutefois, plusieurs GBM avec un nombre différent de variables ont été testés. Un premier avec 12 variables, un deuxième avec 31 variables et un dernier avec l'ensemble des 49 variables. Les résultats des GBM à 31 et 49 variables étaient assez proches en terme de performance. Cependant, l'importance de la variable d'évolution tarifaire est bien plus faible avec le GBM à 49 variables. De plus, le modèle de régression linaire a été calibré sur ces 31 variables ce qui permet d'avoir une base de données identique entre GLM et GBM. Nous retiendrons alors le modèle à 31 variables pour la suite de ce chapitre.

Info	Nom de la variable	Description de la variable
Habitation	CD_TYPExQLTExPIECE	Type d'habitation, Qualité Juridique de l'occupant et Nombre de pièces
	CD_USAG_RISQ	Résidence principale ou secondaire
	NB_PIEC_SUPE	Nombre de pièces supérieures à 40m²
	CD_ISOL	Critère isolement
	Taille_DPDC	Taille de la dépendance
Tarifaire	cot_net_n	Montant de la cotisation nette
	flag_surtarif	1 si la cotisation nette est supérieure à la cotisation barème
	evol_net_fin_n	Evolution annuelle de la cotisation nette entre la situation actuelle et la précédente
	evol_net_fin_n1-n3	Evolution de la cotisation nette sur 3 ans entre la situation n-1 et n-4
	VA_COEF_BM_MRH	Bonus-Malus MRH
	evol_levi_n	Evolution du nombre de leviers tarifaires par rapport à la situation précédente
	MT_DRGT_MBA_TTC	Montant de dérogation MBA appliqué par les agents généraux
Environnement commercial	nb_affa_auto_n	Nombre d'affaires AUTO du client
	evol_AUTO	Evolution du nombre d'affaires par rapport à la situation précédente
	nb_affa_mrh_n	Nombre d'affaires MRH du client
	evol_MRH	Evolution du nombre d'affaires par rapport à la situation précédente
	nb_pro_ent_n	Nombre d'affaires PRO du client
	DETENTION_n	Multi équipement du Client
	FORMULE	Formule souscrite en MRH
	nb_option	Nombre de renforts souscrit en MRH
	CAPI_MOBI	Niveau de capital mobilier souscrit
	nb_2km	Nombre d'agences MMA à moins de 2km de l'habitation
	nb_30km	Nombre d'agences MMA à moins de 30km de l'habitation
	SEGM_PART	Segmentation des clients en fonction de leur tranche d'âge et de l'unité urbaine.
Sin	nb_sin_4ans	Nombre de sinistres survenus durant les 4 dernières situations
	cout_sin_4ans	Coût observé des sinistres des 4 dernières situations
Autres	Generation	Nombre d'année du contrat MRH
	NATU_SITU	Nature du début de situation
	SCORE	Score de rentabilité par iris obtenu avec la part de marché, la densité et de l'évol ptf
	EVOL_BAR_1AN	Evolution barème de l'iris
	NBP_MOY_MAISON_RP	Score de résidences principales par iris

FIGURE 7.4: Liste des 31 variables du GBM

7.2.3 GridSearch

L'algorithme de Gradient Boosting possède de nombreux paramètres pouvant influer sur la construction du modèle d'ensemble et par conséquent sur les performances de celui-ci. Afin d'optimiser le modèle GBM, nous avons utilisé GridSearch. GridSearch permet de tester une série de paramètres pour ensuite comparer les performances des modèles résultants des différentes combinaisons.

Avant d'optimiser le modèle GBM avec ses paramètres, il est nécessaire de définir les conditions d'arrêt de la création des weak learners. Nous avons décidé d'utiliser la métrique AUC qui a été préalablement utilisée dans le chapitre précédent. Le modèle GBM s'arrête lorsque que 5 fois consécutivement la création d'arbre n'améliore pas l'AUC de 0.01%. De plus, dans le but d'éviter que l'algorithme tourne indéfiniment ou presque le nombre d'arbres maximum est limité à 2000.

Pour chaque paramètre, nous déterminons alors un ensemble de valeurs à tester. Dans notre étude, nous avons testé les différents paramètres suivants afin d'optimiser le GBM :

- **Max Depth** : Profondeur maximum des arbres / weak learners
- **Learn rate** : Cette option est utilisée pour spécifier le taux d'apprentissage de GBM lors de la construction d'un modèle. Des taux d'apprentissage plus faibles pénalisent davantage l'ajout de nouveaux arbres, mais les résultats sont généralement meilleurs.
- **Learn rate annealing** : Coefficient (entre 0 et 1) par lequel est multiplié le Learn rate à chaque itération. Il permet donc d'augmenter la finesse d'apprentissage entre chaque itération. Ainsi, le taux d'apprentissage s'affine au fur et à mesure des itérations.
- **Sample rate** : C'est un facteur compris entre 0 et 1 qui permet de sélectionner de manière aléatoire un échantillon de lignes égal à la proportion définie.
- **Col Sample rate** : C'est un facteur compris entre 0 et 1 qui permet de sélectionner de manière aléatoire un échantillon de colonnes égal à la proportion définie.

Les temps de traitement sont relativement coûteux, nous avons donc limité notre recherche d'optimum sur ces critères. En effet, si nous déterminons 3 niveaux à tester pour chacun de ces 5 critères, le nombre de modèles est déjà de $3^5 = 243$ modèles.

La figure ci-dessous permet de représenter visuellement le fonctionnement de la fonction Grid Search. Dans l'exemple ci-dessous, nous avons testé différentes valeurs sur deux paramètres du GBM : le taux d'apprentissage et la profondeur maximale des arbres. L'optimum avec la métrique de l'AUC se trouve aux alentours de 76,9% avec comme paramètre une profondeur d'arbre de 12 ou 15 et un taux d'apprentissage de 1% ou 5%. Un détail assez contre intuitif, est la baisse de performance à partir d'une profondeur de 18. En effet, intuitivement une profondeur d'arbre supérieure devrait donner de meilleurs résultats, néanmoins une trop grande complexité peu créer du bruit entre les performances sur la base d'apprentissage et de validation et entraîner une sous performance.

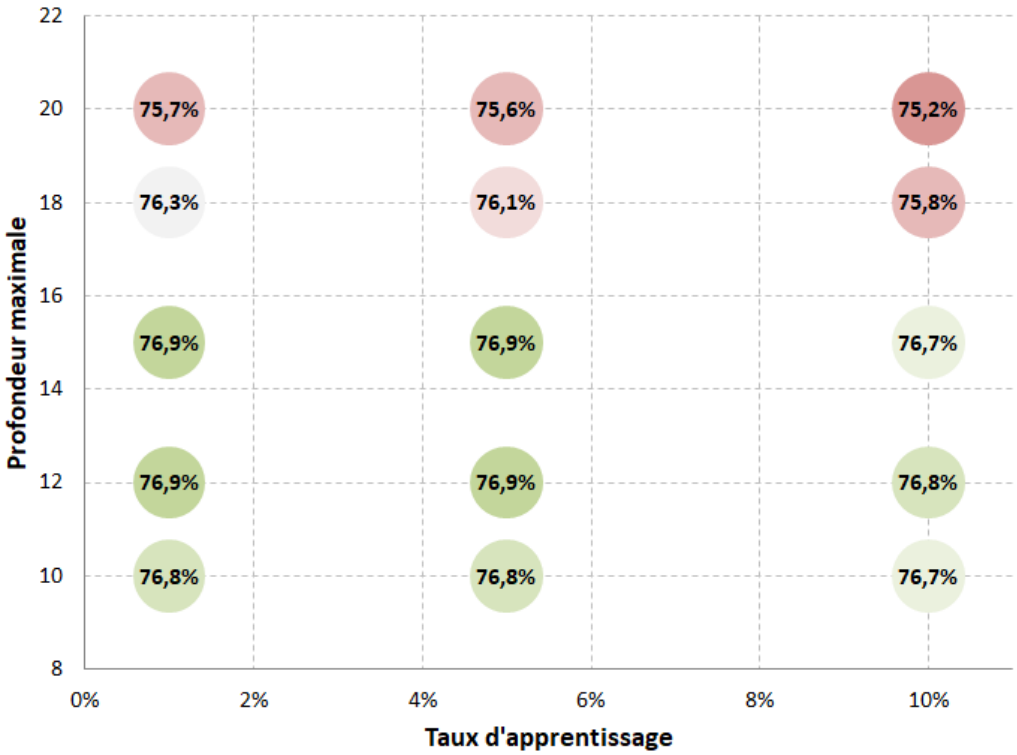

FIGURE 7.5: Exemple Matrice Grid Search

A l'issu des différents GridSearch, les optimums de chaque paramètre sont fixés.

Paramètre	**Définition**	**Valeurs testées**	**Optimum**
Max Depth	Profondeur maximum de l'arbre	$\{10, 12, 15, 18, 20\}$	15
Learn Rate	Taux d'apprentissage	$\{0.01, 0.05, 0.1\}$	0.01
Learn rate annealing	Facteur dégressif	$\{0.98, 0.99, 1\}$	1
Sample rate	Échantillonage des lignes	$\{0.6, 0.7, 0.8, 0.9, 1\}$	0.8
Col Sample rate	Échantillonage des colonnes	$\{0.6, 0.7, 0.8, 0.9, 1\}$	0.8

7.3 Modélisation et résultats du GBM

7.3.1 Performances

Les performances du modèle Gradient Boosting sont bien meilleures que celles du GLM. Si on compare les deux modèles à 31 variables, l'AUC du GBM est supérieur de 5,1 points.

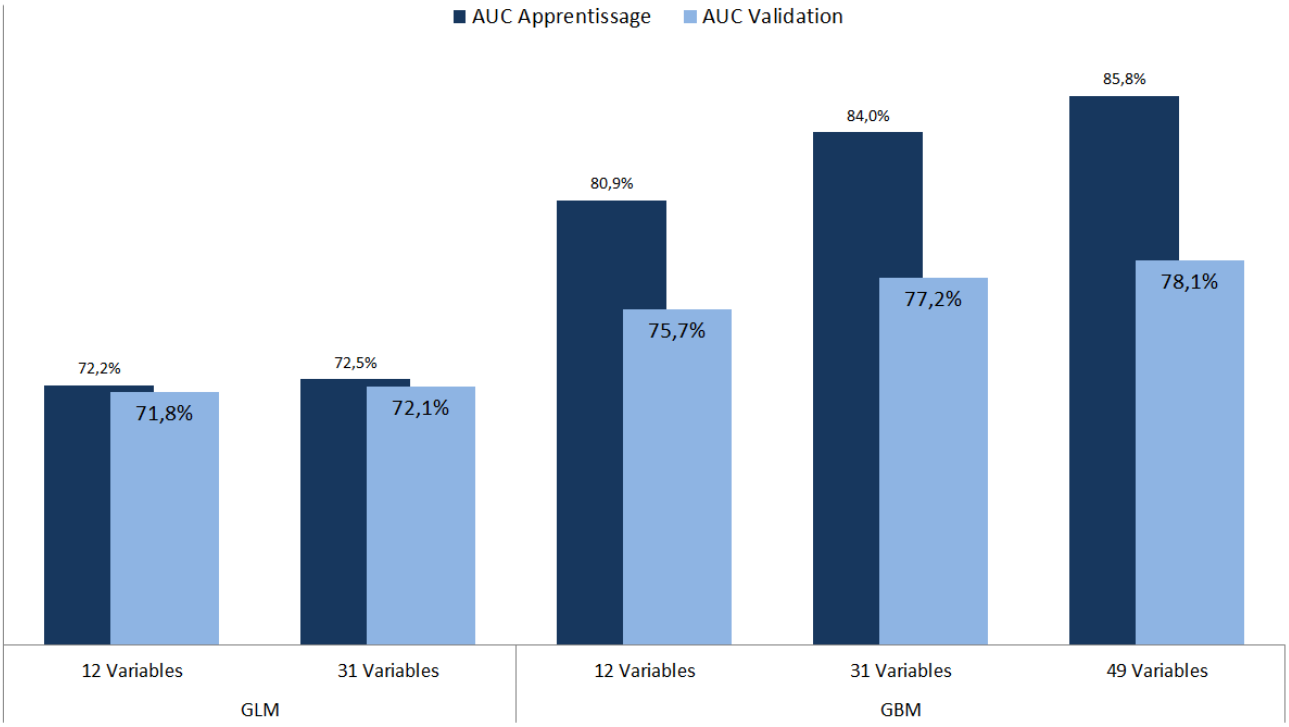

FIGURE 7.6: AUC des GLM et GBM

Le seul fait notable est l'écart de performance entre l'échantillon d'apprentissage et de validation qui n'existait pas sur les GLM. La complexité de certains modèles peuvent créer un phénomène nommé le sur-apprentissage. Ce sur-apprentissage apparaît lorsque le modèle capture tous les détails des données d'apprentissage et sur-performe sur cet échantillon. Par conséquent, lors de prédictions sur de nouvelles données, celui-ci reproduira des spécificités propres à l'échantillon d'entraînement ce qui va créer un biais.

Pour éviter ce phénomène, plusieurs actions sont mise en place :

— La création d'un échantillon d'entraînement et de validation. Le modèle est alors entraîné sur la première base mais les performances sont mesurées sur la seconde.

— Le paramétrage du modèle qui permet d'arrêter celui-ci lorsque l'apprentissage devient trop profond.

Dans notre étude, aucun signe de sur-apprentissage n'est présent malgré l'écart de performances entre la base d'apprentissage et de validation. Une trace de sur-apprentissage serait visible sur le graphique si le taux d'erreur sur la base d'apprentissage (bleu) continuait de décroître tandis que le taux d'erreur sur la base de validation (orange) augmenterait.

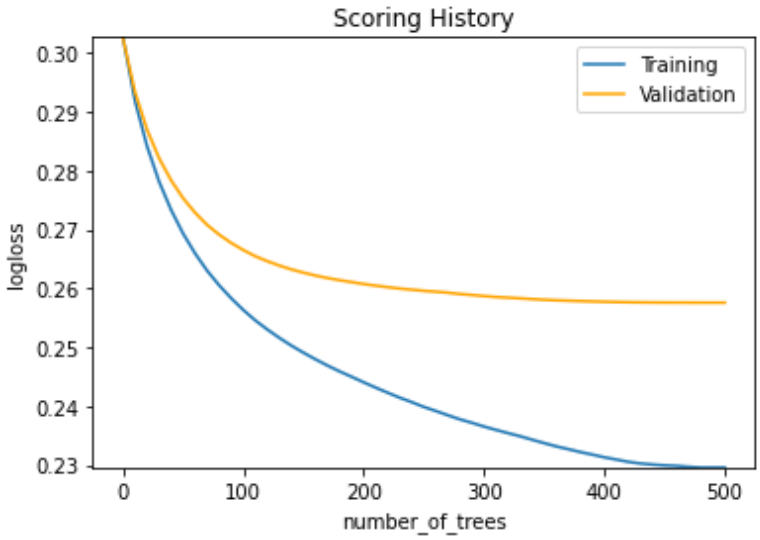

FIGURE 7.7: Taux d'erreur du modèle en fonction du nombre de arbres weeak learners

7.3.2 Classification GBM

La classification binaire du GBM fonctionne comme celle du GLM (cf. Section 6.5.2). Le modèle prédit une probabilité de résiliation et à partir d'un seuil détermine le score, c'est-à-dire résiliation ou non. Comme le GLM (cf. figure 6.25), la probabilité de résiliation prédite par le GBM est extrêmement proche du taux observé. Le modèle Gradient Boosting possède néanmoins les mêmes biais que le modèle Logit. Sur le graphique ci-dessous, les scores de prédictions du GLM avec le seuil F1 (bleu) et du GBM (orange) sont comparés aux résiliations observées (noir). Les deux modèles ont tendance à pénaliser les profils qui résilient beaucoup en prédisant un nombre plus fort que le nombre observé. Inversement, le taux prédit pour les propriétaires est toujours inférieur à celui observé.

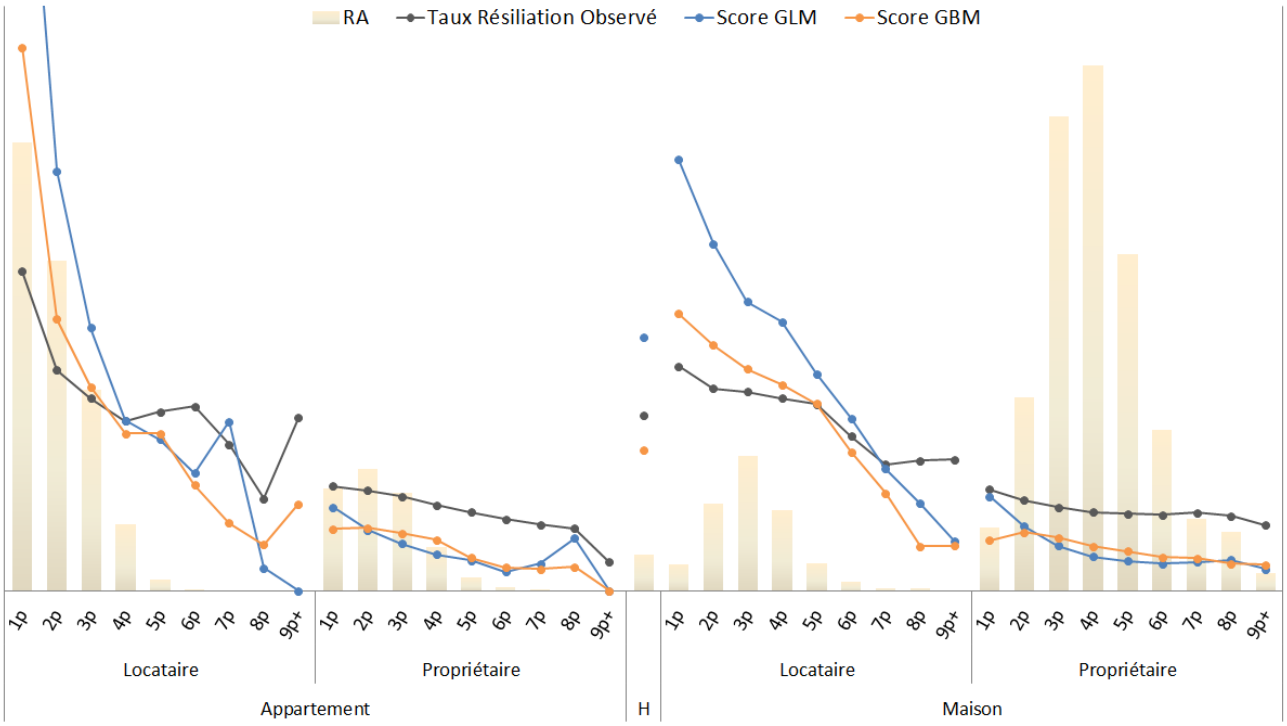

FIGURE 7.8: Score prédit par le GLM et le GBM vs les résiliations observées en fonction de la variable croisée

Toutefois, le seuil du GBM est mieux adapté que celui du GLM. Ce seuil permet de prédire, à quelques milliers près le même nombre de résiliations que celles observées. En plus, des performances supérieures au niveau de l'AUC le Gradient Boosting prouve une nouvelle fois sa capacité à être meilleur qu'une régression logistique.

		Nombre résiliations	Ecart avec l'observé
Observé	Nombre observé sur la base de validation	110 948	-
GLM avec le seuil F1	Minimise d'avantage l'erreur sur les prédictions c'est-à-dire sur les 1	165 892	54 944
GBM avec son seuil	Minimise l'erreur globale	119 713	8 765

FIGURE 7.9: Nombre de résiliations du GLM et GBM avec le seuil de prédiction

7.3.3 Auto Machine Learning

Récemment, une nouvelle façon de concevoir des modèles est apparue. Cette méthode permet d'automatiser la création d'un modèle de Machine Learning. L'algorithme se charge alors de développer et paramétrer plusieurs modèles pour ensuite tester leur efficacité sur les données. L'Auto ML permet alors à des personnes non-expertes de développer des modèles de Machine Learning, mais également d'orienter et challenger une modélisation réalisée pas à pas comme nous l'avons faite.

Avant de comparer les résultats des modèles façonnés par Auto ML, nous devons rapidement fixer le cadre d'apprentissage automatique sous H2O. La recherche de modèle a été limitée aux modèles suivants : GLM, GBM, Forêt Aléatoire (DRF), Arbre de décisions aléatoires (XRF) et Réseaux de

Neurones (Deep Learning). De plus, afin de limiter les temps de traitement, le seuil d'arrêt a été augmenté à 0.1% d'AUC.

model_id	auc	logloss	aucpr
GBM_grid__1_AutoML_20210831_090924_model_8	0.762659	0.259376	0.309913
GBM_2_AutoML_20210831_090924	0.762495	0.259314	0.310327
GBM_3_AutoML_20210831_090924	0.76225	0.259408	0.309909
GBM_4_AutoML_20210831_090924	0.761853	0.259551	0.308733
GBM_grid__1_AutoML_20210831_090924_model_2	0.761806	0.259592	0.3092
GBM_1_AutoML_20210831_090924	0.761793	0.259625	0.30884
GBM_grid__1_AutoML_20210831_090924_model_11	0.761374	0.260008	0.306536
GBM_grid__1_AutoML_20210831_090924_model_10	0.761061	0.260528	0.305076
GBM_grid__1_AutoML_20210831_090924_model_7	0.760918	0.260145	0.306535
GBM_5_AutoML_20210831_090924	0.760665	0.259828	0.307471
GBM_grid__1_AutoML_20210831_090924_model_3	0.755546	0.262078	0.298594
GBM_grid__1_AutoML_20210831_090924_model_4	0.752612	0.263062	0.294489
GBM_grid__1_AutoML_20210831_090924_model_9	0.752053	0.264611	0.284638
GBM_grid__1_AutoML_20210831_090924_model_6	0.749834	0.264297	0.285653
GBM_grid__1_AutoML_20210831_090924_model_5	0.743691	0.266748	0.281829
DRF_1_AutoML_20210831_090924	0.743167	0.266454	0.278265
GBM_grid__1_AutoML_20210831_090924_model_1	0.742262	0.268083	0.269274
DeepLearning_grid__1_AutoML_20210831_090924_model_2	0.735871	0.267986	0.278122
XRT_1_AutoML_20210831_090924	0.733511	0.270441	0.2697
DeepLearning_1_AutoML_20210831_090924	0.707665	0.278354	0.222182
DeepLearning_grid__1_AutoML_20210831_090924_model_4	0.705511	0.277651	0.25829
GLM_1_AutoML_20210831_090924	0.704939	0.2803	0.19605

FIGURE 7.10: Classement modélisation Auto ML

La métrique de comparaison utilisée est toujours l'AUC calculée sur la base de validation. Ici, nous remarquons que le classement est très largement dominé par les modèles de Gradient Boosting. Le premier modèle hors GBM est un algorithme de forêt aléatoire (DRF) avec une AUC de 74,2%. Les arbres de décision sont efficaces pour la modélisation d'une variable binaire en conséquence les modèles construits à partir le sont également. En terme de précision, le GBM est incontestablement le meilleur algorithme pour notre étude.

Les GBM les plus performants possèdent des caractéristiques similaires à celui obtenu à l'aide de GridSearch. Les seules différences entre le modèle en première position et celui calibré à la main se trouvent au niveau du nombre maximum d'arbres et du coefficient d'apprentissage.

7.4 Interprétabilité du GBM

Le modèle de Gradient Boosting fournit une alternative plus performante aux méthodes plus classiques comme le GLM. Toutefois, ce gain de précision se fait au détriment de l'interprétabilité et de la transparence. Il est difficile d'obtenir une interprétation claire de l'influence de chaque variable explicative ainsi que le cheminement qui a mené à la prédiction. Du fait de cette nécessité, de plus en plus d'outils se développent pour aider à la compréhension des modèles de Machine Learning.

7.4.1 Variable Importance

La manière la plus simple et explicite de déterminer l'importance d'une variable dans un modèle est d'utiliser l'algorithme de calcul présent sous H2O. Cet algorithme calcule l'influence relative de chaque variable. Il détermine, sur l'ensemble des arbres weak lerners, dans quelle mesure l'erreur quadratique a été modifiée en perturbant les valeurs de la variable. Si la prédiction du modèle est grandement affectée par le mélange de ces modalités, l'erreur quadratique subira une variation importante et l'algorithme jugera que la variable en question joue un rôle prépondérant dans le modèle. Inversement, une faible variation de l'erreur quadratique induira que la variable n'est pas importante.

Dans notre modèle de Gradient Boosting, les 3 variables les plus discriminantes sont :

— **GENERATION** : Ancienneté du contrat
— **CD_TYPExQLTExPIECE** : Le type de logement, la qualité juridique de l'occupant et le nombre de pièces.
— **SEGM_PART_PERSO** : Segment du client déterminé avec l'âge et l'urbanisation.

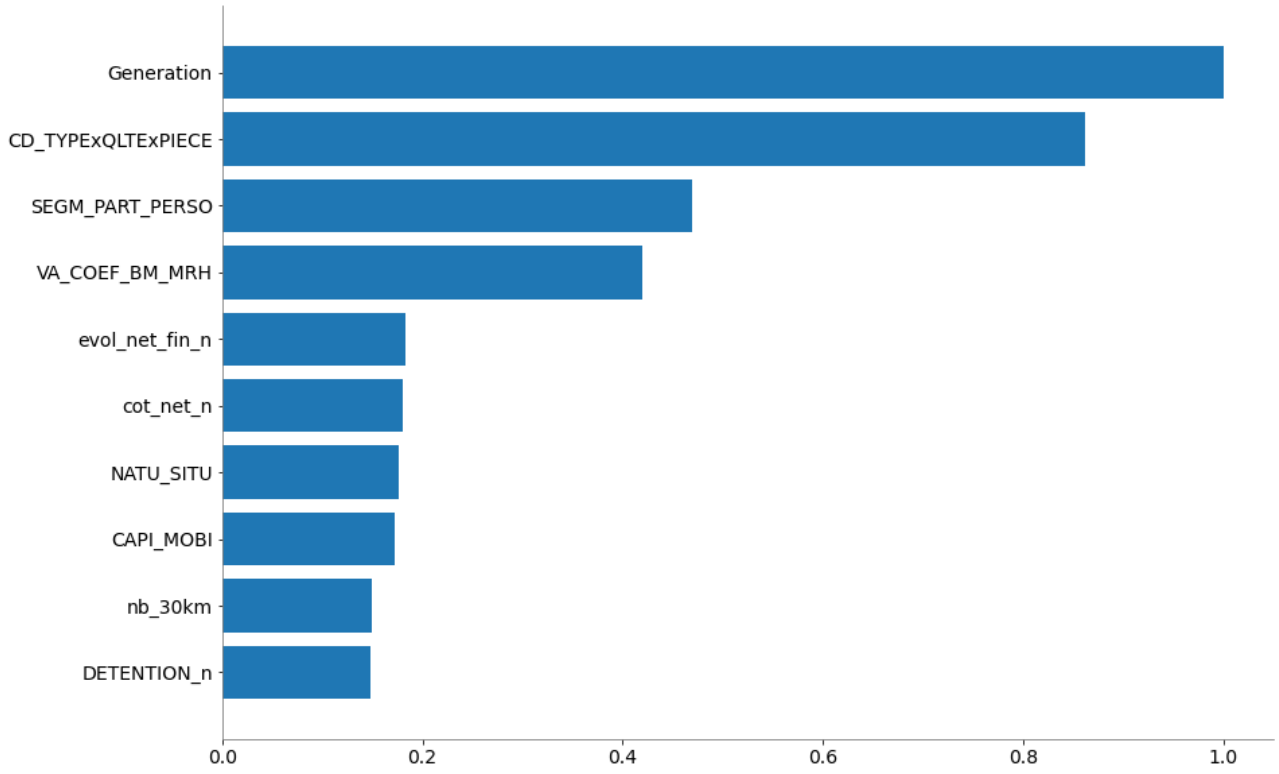

FIGURE 7.11: Classement des variables les plus importantes

7.4.2 Partial Dependance Plot (PDP)

Le graphique de dépendance partielle est un outil d'interprétation de modèle complexe de Machine Learning. A l'inverse du GLM, le Gradient Boosting ne fournit pas de coefficient pour chaque modalité de chaque variable. Le PDP pallie à ce manque, en mesurant l'effet marginal d'une ou deux variables explicatives sur la prédiction du modèle. Par exemple, lorsqu'ils sont appliqués à un modèle de régression linéaire, les diagrammes de dépendance partielle montrent toujours une relation linéaire entre les caractéristiques et la prédiction.

Si l'on note S l'ensemble des variables explicatives que l'on souhaite étudier, et C l'ensemble des autres variables du modèle, le graphique de dépendance partielle se définit ainsi :

$$\hat{f}_S(x_S) = \mathbb{E}_{XC}[\hat{f}(x_S, X_C)] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C) \tag{7.5}$$

Cet outil va nous permettre d'analyser le comportement du taux de résiliation en fonction des variables les plus importantes de notre modèle. Nous allons pouvoir confronter cet effet marginal avec le taux de résiliation observé et la probabilité prédite par le modèle. En réalité, ces deux derniers indicateurs peuvent être influencés par des effets de corrélation entre les variables ce qui rend difficile la mesure exacte de l'impact de la variable sur la prédiction du taux de résiliation.

Les variables les plus importantes sont analysées par la suite, à savoir la génération du contrat, la variable croisée et le segment client. Ces trois variables sont les plus discriminantes selon le GBM et présentent donc un intérêt tout particulier d'un point de vue Data. Pour l'entreprise, il est plus intéressant d'analyser des variables comme le multi-équipement, le montant des dérogations et l'évolution tarifaire car elles sont centrales pour le pôle TSP.

Génération du contrat Comme nous l'avons vu précédemment avec le GLM (cf. figure 6.25), en moyenne la probabilité prédite par le modèle Gradient Boosting est très proche du taux de résiliation observé. Ce résultat a l'avantage de conforter la bonne calibration du modèle, mais les résultats des prédictions ne peuvent être interprétés ou extrapolés.

L'effet marginal calculé par le graphique de dépendance partielle révèle l'impact réel de chaque valeur prise par la variable. Pour l'ancienneté de contrat, le PDP montre que l'impact d'un contrat de génération inférieure à 5 ans est beaucoup plus important que le montre la vision observée. Nous savions que la génération du contrat jouait un impact prépondérant sur la prédiction du taux de résiliation, mais le modèle nous indique que nous avions tendance à sous estimer cet impact.

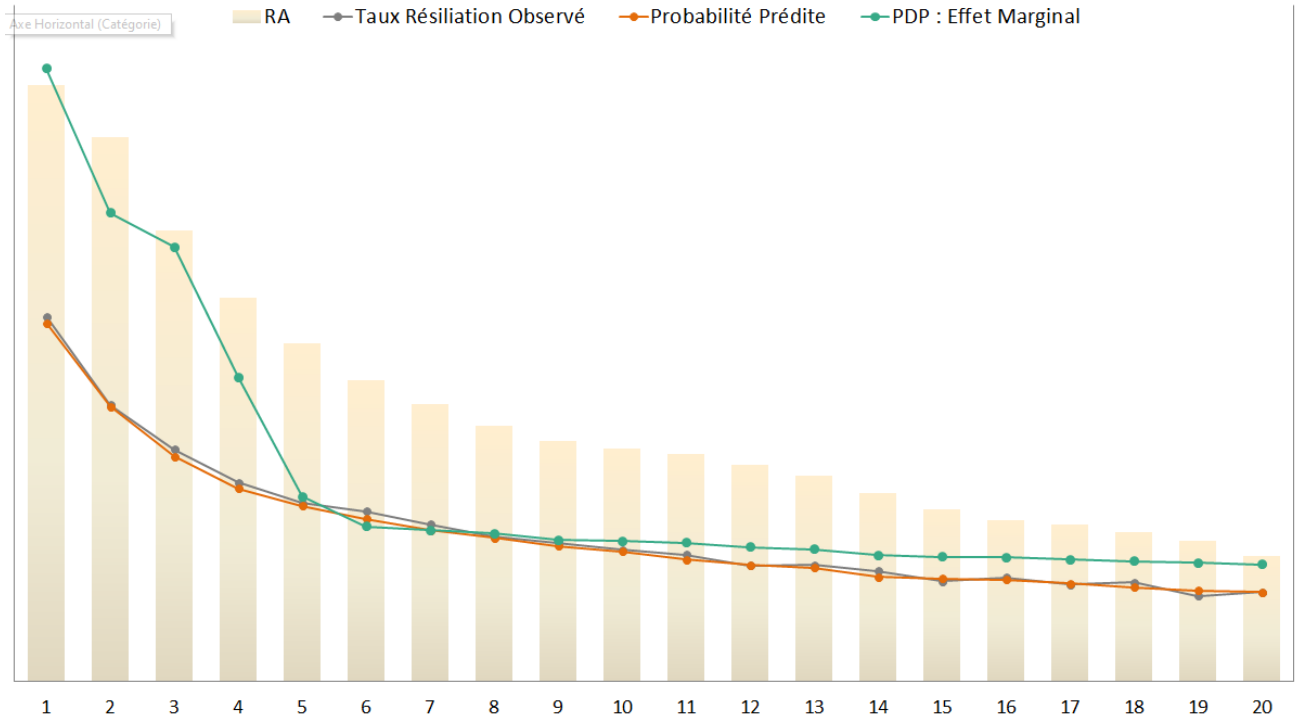

FIGURE 7.12: Graphique de dépendance partielle de la variable Génération

Variable Croisée : Type x Qualité x Nombre de pièces Sur certaines modalités de la variable les résultats de la prédiction peuvent être légèrement éloignés du taux de résiliation observé, toutefois cet écart est principalement concentré sur les modalités avec peu de volume.

L'effet marginal ne se comporte pas de la même manière vis à vis du taux observé entre les propriétaires et les locataires. Sur les locataires, l'effet marginal donné par le graphique de dépendance partielle est moins important que le taux de résiliation observé. Inversement pour les propriétaires, l'effet marginal est supérieur au taux de résiliation observé. Avec les données observées, nous avions tendance à sur-estimer l'impact de la modalité locataire sur la hausse des résiliations, et inversement nous avions tendance à sur-estimer l'impact de la modalité propriétaire sur la baisse des résiliations.

Dernièrement, l'effet marginal montre que le nombre de pièces influe peu sur le taux de résiliation. L'effet marginal a tendance à aplatir la courbe ce qui signifie que la différence de résiliation entre une habitation avec peu et beaucoup de pièces n'est pas aussi importante que le montre le taux observé.

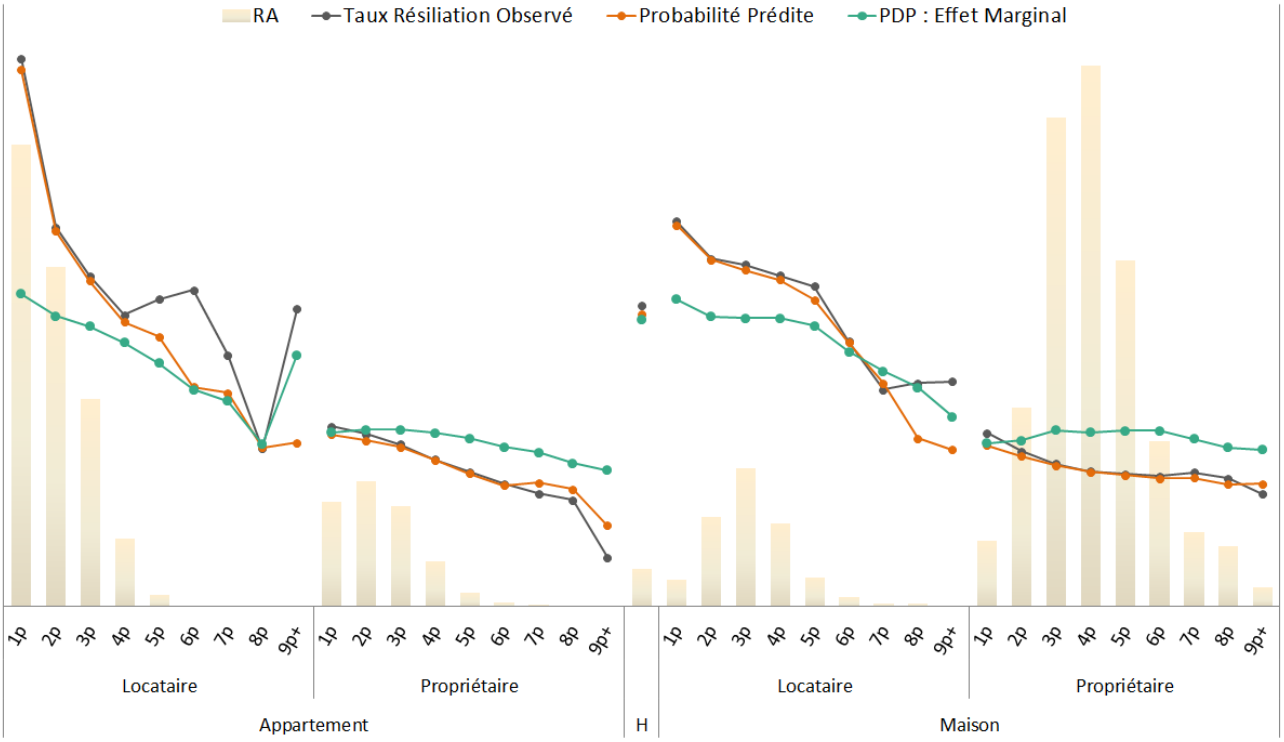

FIGURE 7.13: Graphique de dépendance partielle de la Variable Croisée : Type x Qualité x Nombre de pièces

Segment du client : Classe d'âge x Urbanisation En moyenne et sur chacune des classes, les prédictions du GBM sont en parfaite adéquation avec le taux de résiliation observé.

L'effet marginal atténue fortement les impacts des modalités sur le taux de résiliation. Tout d'abord, l'effet lié à l'urbanité de l'habitation est le même pour chaque classe d'âge. Nous remarquons seulement un taux légèrement plus élevé pour la modalité Rural. De plus, l'effet lié à la classe d'âge est bien moins important, spécialement pour la classe Jeune. En effet, cette classe regroupe une forte proportion de locations et de contrats récents en portefeuille ce qui biaise la vision observée et donc l'impact réel de cette variable est surestimé sur la vision observée.

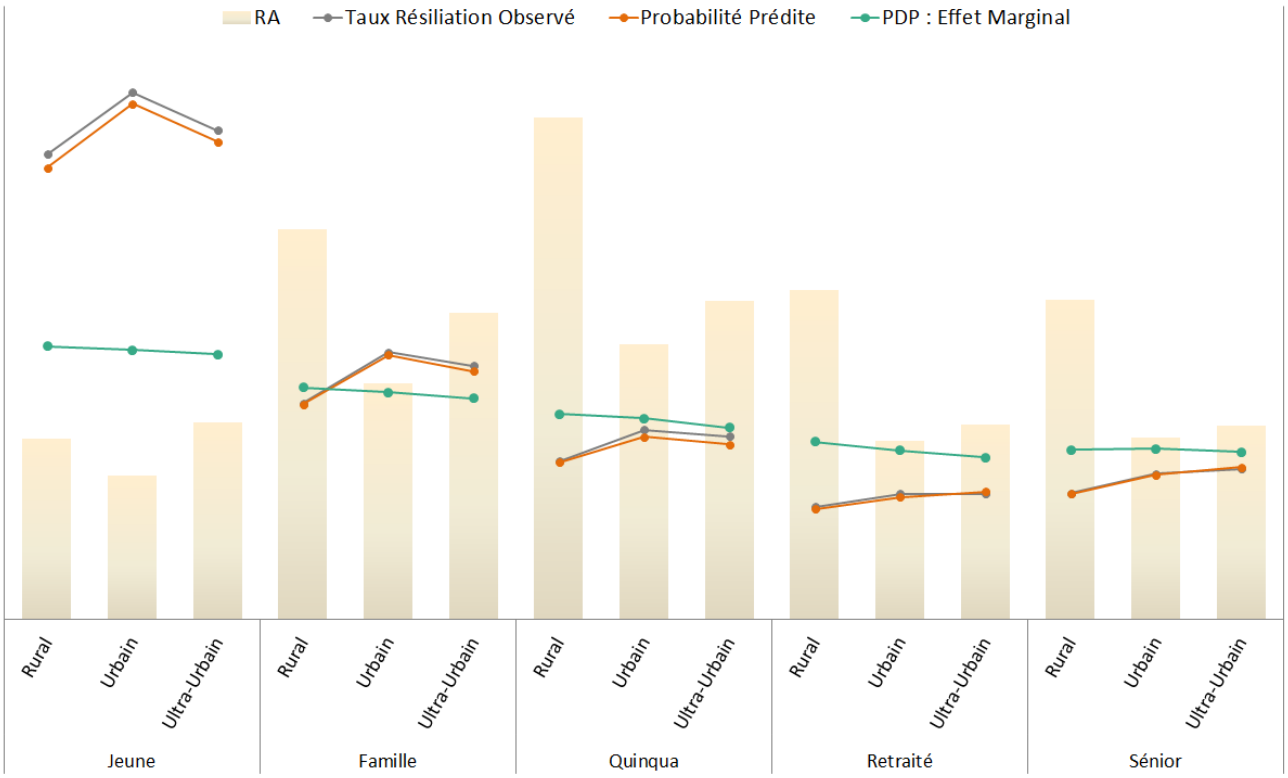

FIGURE 7.14: Graphique de dépendance partielle de la variable Segment du client : Classe d'âge x Urbanisation

Évolution tarifaire Lors des premières analyses au chapitre 5, le taux de résiliation semblait assez volatile et prenait une courbure assez particulière. Nous avions essayé de séparer appartements et maisons afin de réduire cette courbe en « U ». L'effet marginal de la variable d'évolution tarifaire ne possède pas du tout cet aspect courbé. Le PDP montre une légère augmentation du taux de résiliation en fonction de la revalorisation tarifaire. Cet effet semble bien plus cohérent avec les préjugés que nous avons tous, à savoir que l'augmentation de la cotisation entraîne une hausse des résiliations.

Néanmoins, la réponse moyenne calculée par le PDP ne croît que faiblement lorsque la majoration tarifaire augmente. Le taux de résiliation ne semble alors pas très sensible à la hausse de la cotisation.

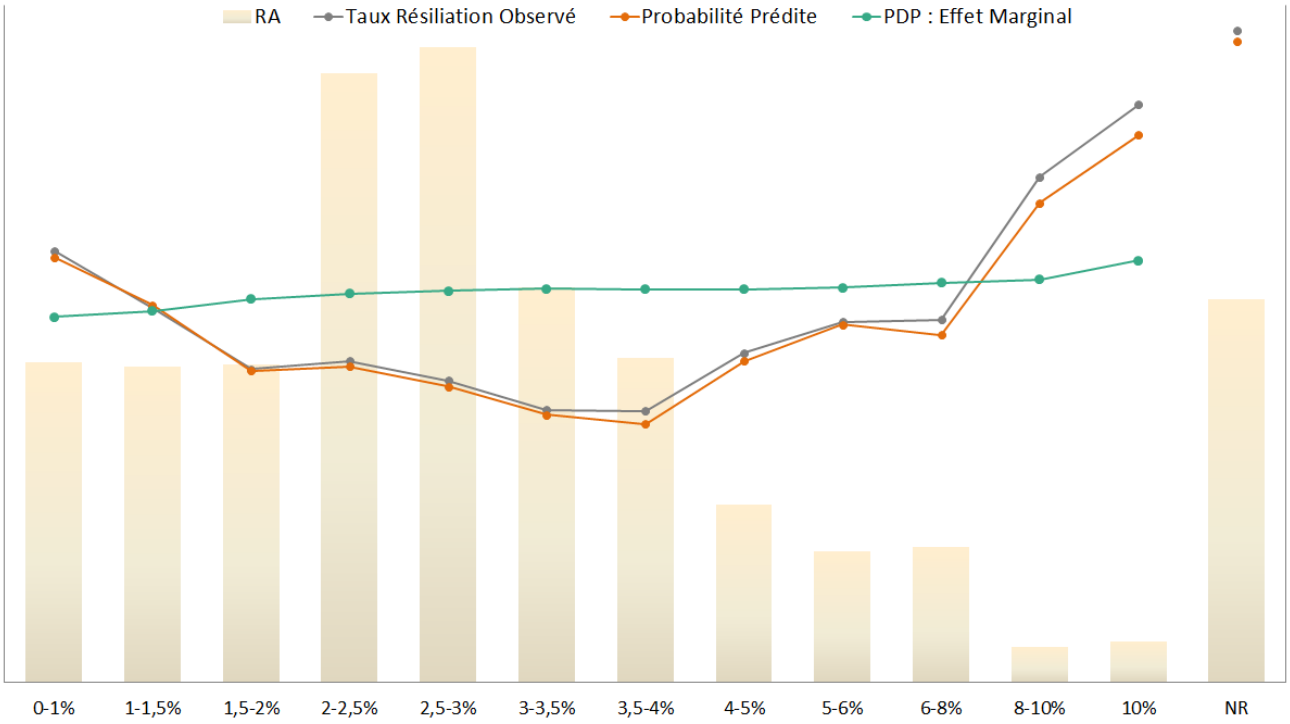

FIGURE 7.15: Graphique de dépendance partielle de la variable d'évolution tarifaire

Bilan Les résultats du graphique de dépendance partielle se rapprochent fortement de ceux que peuvent apporter les coefficients des Modèles Linéaires Généralisés. Historiquement, ces coefficients sont utilisés pour estimer la contribution unitaire de chaque variable au résultat sortie par le modèle linaire. Les résultats du PDP ont le même objectif : estimer la contribution unitaire des variables. Cette méthode possède un avantage sur les odds ratio, c'est qu'elle arrive à capter des contributions non linéaires. Sur la figure 7.12, l'effet marginal de l'ancienneté de contrat avec un GLM serait linéaire, le modèle ne capterait pas les nuances pour les générations inférieures à 5 ans.

Si la caractéristique pour laquelle est calculé le graphique de dépendance partielle n'est pas corrélée avec les autres caractéristiques, alors le PDP représente parfaitement la façon dont la caractéristique influence la prédiction en moyenne. Dans le cas non corrélé, l'interprétation est claire : le graphique de dépendance partielle montre exactement comment la prédiction moyenne de l'ensemble de données change lorsque la j-ième caractéristique est modifiée.

Néanmoins, le graphique de dépendance partielle possède un inconvénient, à savoir de ne pas retourner les interactions entre les variables. Dans sa construction et dans son calcul de la prédiction, le modèle

Gradient Boosting introduit des interactions, cependant il est impossible pour le PDP en univarié de les retourner. Cette incapacité à retourner ces interactions créées lorsque les variables ne sont pas indépendantes peuvent biaiser les résultats du PDP.

Il est possible de demander au graphique de dépendance partielle d'analyser deux variables à la fois. De ce fait, le PDP peut retourner des résultats en incluant les interactions entre ces deux variables. Néanmoins, les temps de traitement peuvent rapidement devenir très long et les sorties peuvent devenir complexes à représenter.

Une alternative qui se développe est le graphique des effets locaux accumulés (ALE). Cette méthode corrige le PDP, notamment lorsque les variables explicatives sont fortement corrélées entre elles. L'ALE va décrire comment les variables influent sur la prédiction du modèle en moyenne mais sans biais. Cependant, les graphiques des effets locaux accumulés ne sont pas disponibles avec la librairie H2O, il est donc seulement possible de les appliquer sur des bases de taille plus modeste sans passer par H2O. Pour nos résultats, nous nous contenterons d'utiliser les graphiques de dépendance partielle tout en ayant conscience que l'effet marginal de certaines caractéristiques fortement corrélées à d'autres peut être biaisé. Un autre indicateur robuste est disponible sous H2O : Les SHAP Values.

7.4.3 SHAP

La valeur de Shapley a été introduite en théorie des jeux dans un système de jeu coopératif. Son objectif est de répartir équitablement des gains entre plusieurs joueurs ayant travaillés ensemble en fonction de leur contribution aux résultats. Cette valeur, appliquée à l'interprétabilité des modèles, a pour objectif de quantifier le rôle de chaque variable dans la prédiction finale du modèle. On translate la théorie des jeux en supposant que chaque variable est un joueur dans un jeu où le résultat est la prédiction du taux de résiliation.

Le but de SHAP est alors d'expliquer la prédiction en calculant la contribution de chaque variable explicative à cette réponse. Les valeurs des variables explicatives agissent comme des joueurs dans une coalition. Les valeurs de Shapley nous indiquent comme répartir équitablement le gain, ici la prédiction, entre ces variables. SHAP retranscrit la contribution des variables de la manière suivante :

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j \tag{7.6}$$

où $g(.)$ est le modèle d'explication, $z' \in \{0,1\}^M$ le vecteur de coalition, M la taille maximal de la coalition et ϕ_j est la contribution de la caractéristique j.

Dans le cas particulier où toutes les valeurs de caractéristiques sont incluses dans le vecteur de coalition, z' est toujours égal à 1, donc la formule se simplifie comme suit :

$$g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j \approx f(x) \tag{7.7}$$

Afin de mesurer les contributions des variables, la valeur de Shapley utilise la moyenne des prédictions

comme référentiel : $\mathbb{E}[f(x)]$. La méthode de SHAP explique l'écart entre cette moyenne et la prédiction réalisée par le modèle $\hat{y} = f(x)$ en décomposant la contribution de chaque variable explicative. Par exemple, pour obtenir la contribution de la variable x_1 sur la prédiction du modèle nous calculons :

$$\phi_1 = \mathbb{E}[f(x)|x_1] - \mathbb{E}[f(X)]$$

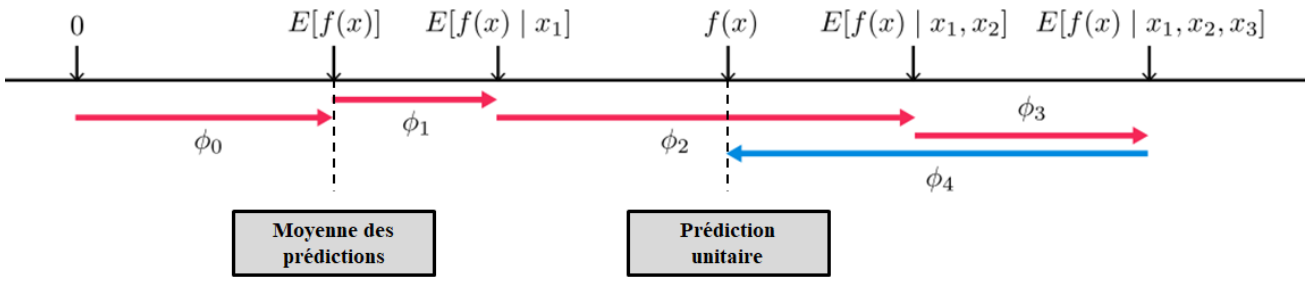

FIGURE 7.16: Exemple de SHAP values avec 4 variables explicatives

Néanmoins, le graphique ci-dessus est une des nombreuses manières de calculer la contribution des variables. Les variables x_i ne sont généralement pas indépendantes donc l'ordre de calcul des contributions ϕ_i influe sur leur valeur. Afin d'éviter que l'ordre d'ajout des variables ne soit un facteur influant, les SHAP values sont calculées à partir de toutes les combinaisons possibles et c'est la moyenne de tous les ϕ_i possibles qui est renvoyée.

Prédictions Individuelles La première utilisation des SHAP values consistent à analyser une prédiction individuelle à l'aide d'un diagramme de forces. Ce diagramme illustre la contribution de chaque variable explicative à la prédiction du modèle GBM. Les flèches rouges retranscrivent la force des variables qui contribuent à la hausse, c'est-à-dire celles qui poussent le modèle à prédire une résiliation, tandis que les flèches bleues représentent les variables qui contribuent à une baisse, soit à pousser le modèle à prédire une non résiliation. La longueur de ces flèches correspond à la valeur absolue de la SHAP Value, plus celle-ci est grande plus la variable explicative influe sur la prédiction.

Nous allons analyser deux situations, une première lorsque le modèle prédit une résiliation et une seconde lorsque le modèle prédit une non résiliation. Pour chaque situation, les deux sorties expliquent la prédiction du modèle à l'aide des SHAP values.

Sur cette première situation, le premier graphique : le diagramme de force montre comment chaque variables contribuent au résultat : $f(x)$. Sur cet exemple les variables : GENERATION, CD_TYPEXQLTEXPIECE, NATU_SITU, SEGM_PART_PERSO et DETENTION_N sont les variables qui influent les plus sur la prédiction de cette observation. Tous ces impacts sont positifs et favorisent alors la prédiction = 1 soit une résiliation.

Le second graphique représente les 20 variables les plus influentes sur la prédiction du modèle ainsi que leur modalité. Ce second graphique est certes moins élégant, mais il permet de facilement relier l'impact à la modalité de la variable. Par exemple, le fait que le contrat a seulement 1 an d'ancienneté

et soit un appartement en location (AL_2) accroît fortement le résultat. Le résultat sur cet exemple est identique aux effets marginaux déterminés par les graphiques de dépendance partielle. Le fait d'avoir un contrat récent et d'assurer un appartement en location sont des modalités avec des réponses moyennes de résilier élevées.

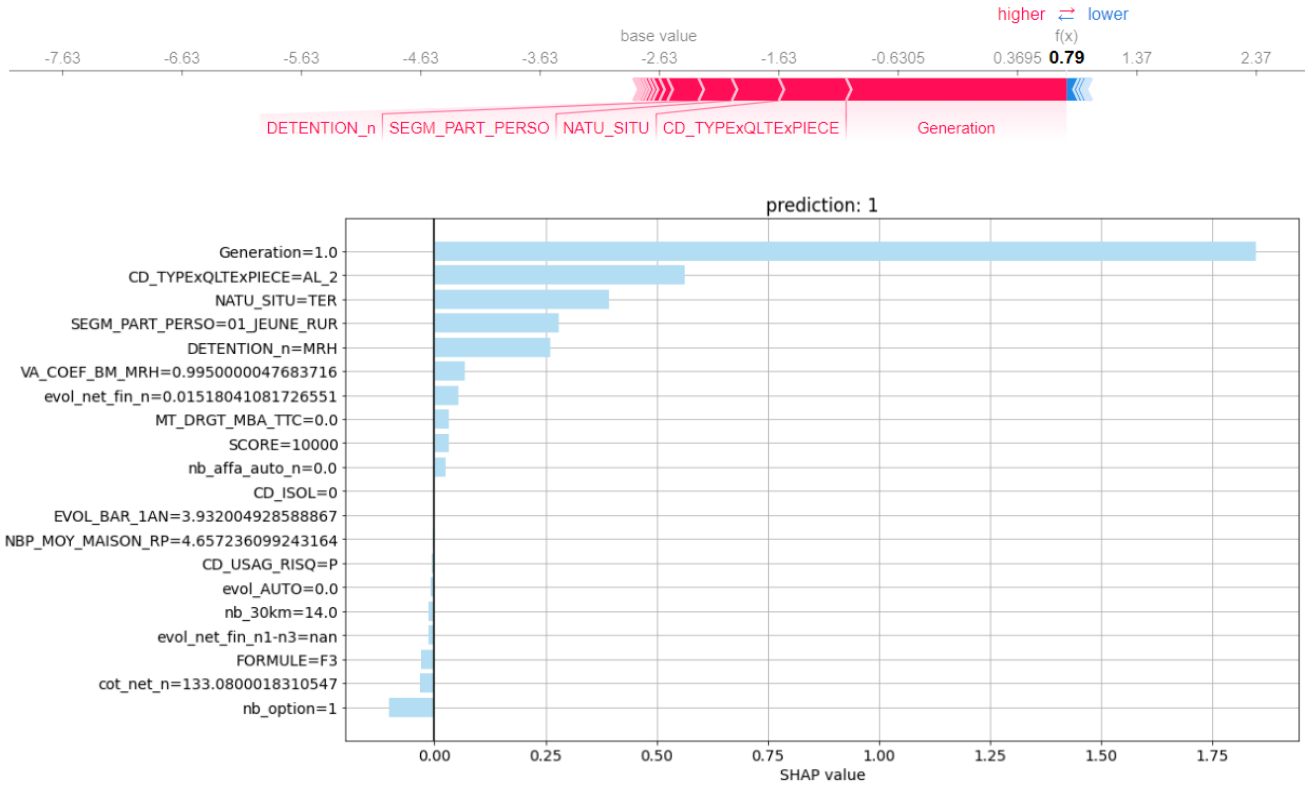

FIGURE 7.17: SHAP value pour une prédiction = 1

Sur la seconde situation, le modèle prédit que le contrat restera en portefeuille. Le diagramme de force est beaucoup plus équilibré que le précédent, aucune variable n'a un effet aussi écrasant que la modalité GENERATION = 1 du précédent exemple.

Deux variables ont un effet négatif important : l'ancienneté de contrat de 8 ans et le fait que l'habitation soit une maison propriétaire de 4 pièces (MP_4). A l'inverse, la variable détention et segment client influencent positivement le résultat.

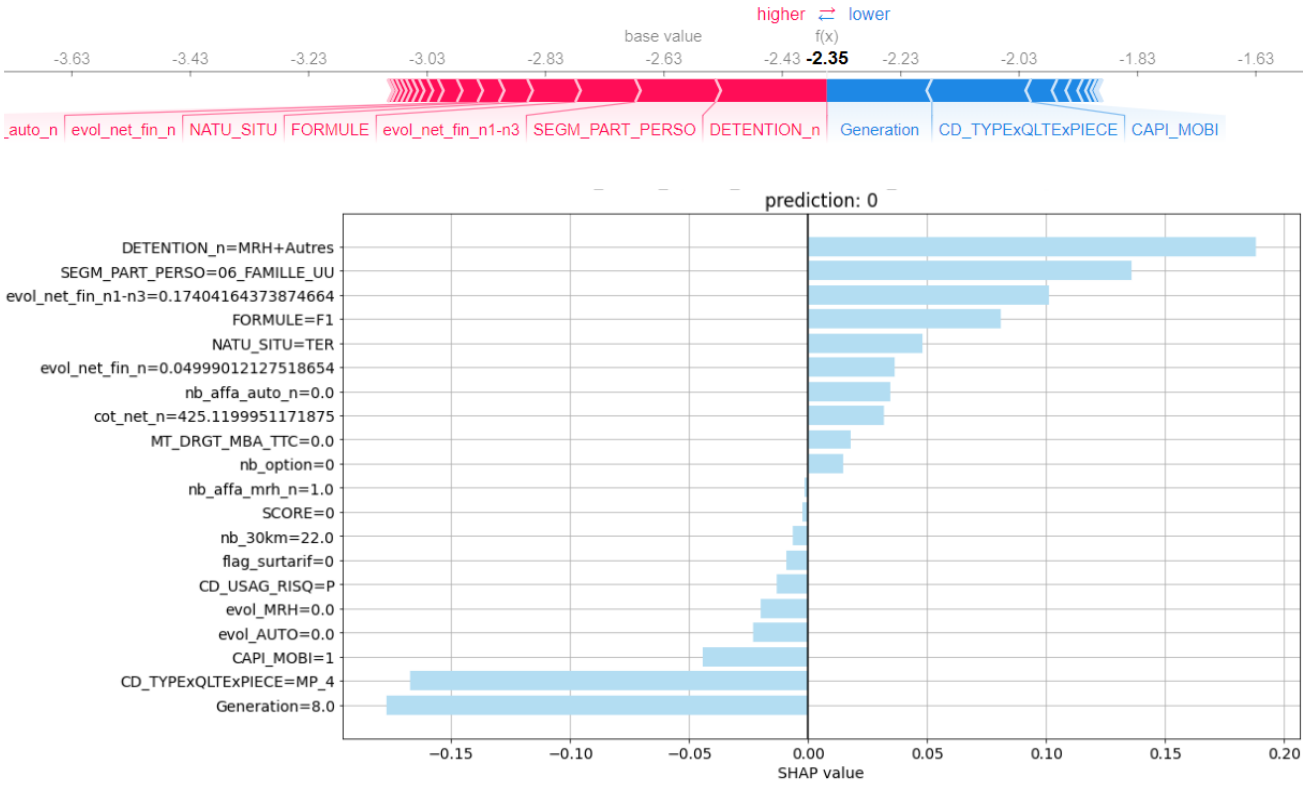

FIGURE 7.18: SHAP value pour une prédiction = 0

Interprétation globale La seconde utilisation des SHAP values est d'agréger les interprétations individuelles afin d'obtenir une idée générale des contributions des variables explicatives.

Premièrement, il est possible de déterminer l'importance des variables à l'aide des valeurs de Shapley. Pour cela, l'algorithme mesure l'impact moyen des variables explicatives sur les prédictions. Plus cet impact mesuré est fort plus la variable sera considérée comme importante au sens des SHAP values. Donc les caractéristiques ayant de grandes valeurs absolues de Shapley sont importantes. Puisque nous voulons l'importance globale, nous faisons la moyenne des valeurs absolues de Shapley par variables explicatives sur l'ensemble des données. L'importance selon SHAP est donnée par :

$$I_j = \frac{1}{n}\sum_{i=1}^{n}|\phi_j^{(i)}| \tag{7.8}$$

Cette méthode confirme que la variable d'ancienneté de contrat et la variable croisée sont très influentes. Toutefois, la variable du segment client est moins impactante avec cette mesure plutôt qu'avec celle réalisée par le modèle GBM (cf. figure 7.11).

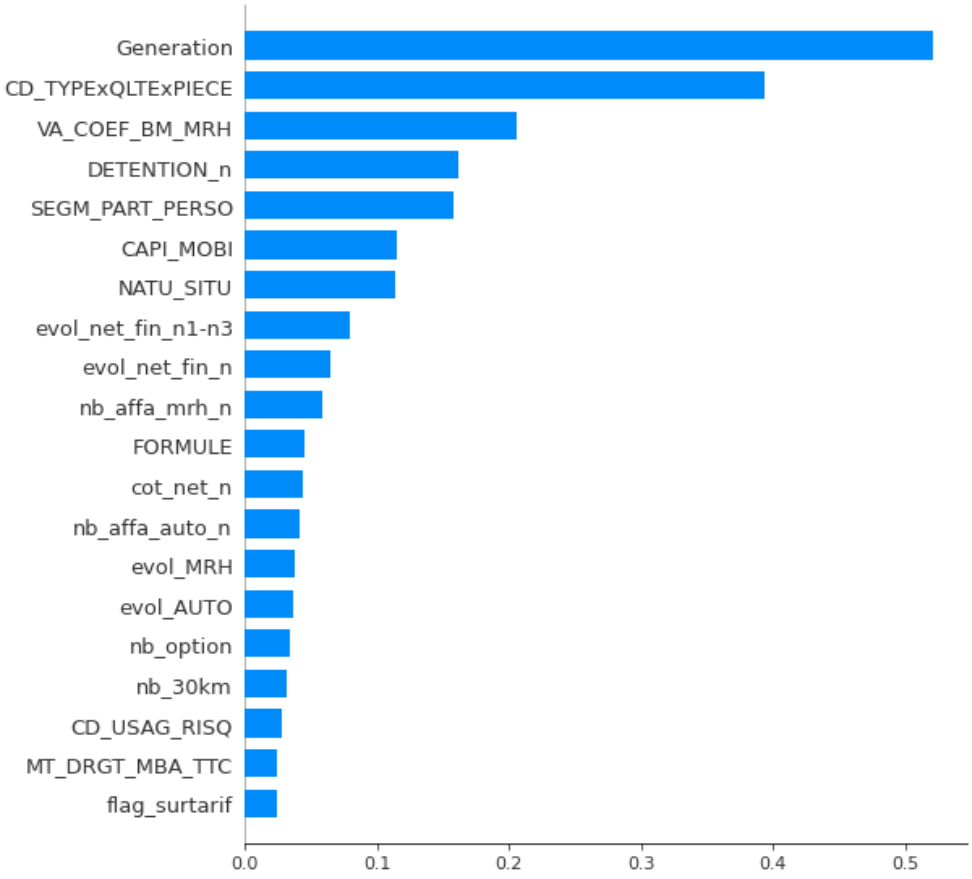

FIGURE 7.19: Importance des variables explicatives selon les valeurs de Shapley

En plus de l'importance des variables, un graphique synthétique permet de résumer l'ensemble des contributions des variables aux prédictions du modèle. La lecture se réalise de la manière suivante :

- Chaque point représente une observation.
- En ordonnée se trouve la liste des variables explicatives dans l'ordre d'importance au sens de Shapley.
- L'abscisse représente la SHAP value donc l'impact de la variable sur la prédiction.
- La couleur des points renseigne sur la valeur prise par la variable explicative. Un point bleu indique que la valeur prise est faible tandis qu'un point rouge indique que la valeur prise est élevée.

Pour des variables quantitatives et ordonnées ce graphique est parfaitement interprétable. Par exemple, pour la variable génération, les points bleus représentent des contrats récents tandis que les points rouges représentent des contrats anciens en portefeuille. Comme évoqué lors des résultats précédents, les points bleus, à savoir les contrats récents, se trouvent à droite de l'axe ce qui signifie qu'ils ont un impact positif sur le taux de résiliation prédit par le modèle.

Pour une variable catégorielle, les résultats sont plus difficilement interprétables car le modèle ordonne toujours les modalités pour les insérer sur l'échelle de couleur. Pour la variable croisée les maisons sont considérées comme des valeurs élevées (rouge) et les appartements comme des valeurs faibles (bleu). Des groupes se forment, car en moyenne les maisons ont un taux de résiliation moins important. Cependant, pour la variable DETENTION_N, qui est catégorielle et non ordonnée il est très difficile d'interpréter le code couleur établi par l'algorithme.

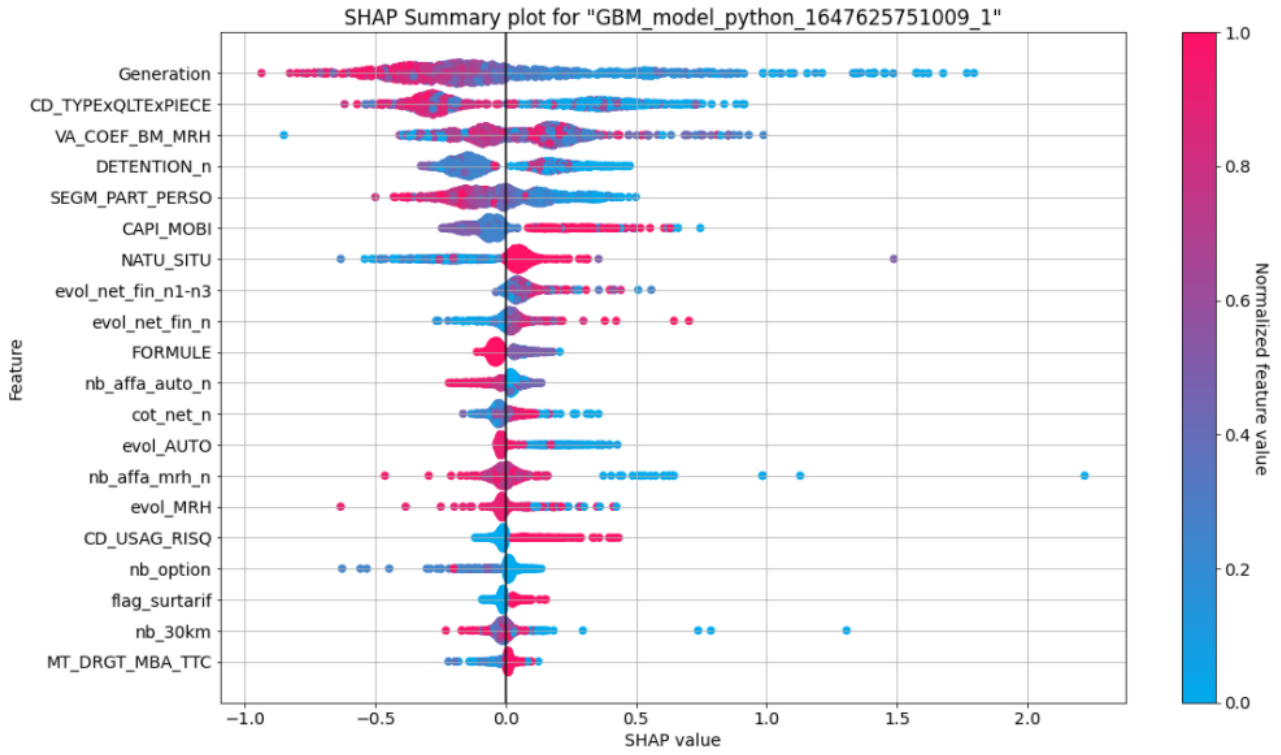

FIGURE 7.20: Synthèse globale des SHAP values

7.4.4 Bilan de l'interprétation

L'interprétabilité des modèles de Machine Learning est incontournable dans le monde de l'actuariat. Ces modèles offrent des performances supérieures aux traditionnels Modèles Linéaires Généralisés, mais leur effet boite noire est un frein à leur utilisation. En effet, il est inenvisageable de mettre en production des modèles boites noires, d'une part, pour des raisons réglementaires avec le droit à l'explication, et d'autre part, pour des raisons de confiance. Comment peut-on avoir confiance en des modèles qui nous offrent un résultat sans explication ?

Les graphiques de dépendance partielle fournissent des informations facilement interprétables, qui ressemblent aux sorties fréquemment utilisées comme les odds ratio. Nous avons la possibilité d'obtenir un coefficient pour chaque modalité d'une variable. Néanmoins, cet outil suppose l'indépendance entre elles. Utiliser les coefficients comme les odds ratio n'est possible que si l'indépendance entre les variables est avérée. Si la variable explicative n'est pas corrélée avec les autres caractéristiques, alors le PDP est une représentation parfaite de la façon dont la variable explicative influence la prédiction en moyenne. Dans le cas d'une non-corrélation, l'interprétation est claire : le graphique de dépendance partielle montre comment la prédiction moyenne dans votre ensemble de données change lorsque la caractéristique est modifiée.

Les valeurs de Shapley ont pour avantage d'être une méthode très robuste qui repose sur une base mathématique solide. Les SHAP Values permettent alors de justifier pleinement les segmentations pouvant être mis en place. Toutefois les sorties de cette approche manque parfois de lisibilité et les sorties sont difficilement exploitables opérationnellement.

Quatrième partie

Exploitation du modèle de résiliation

Chapitre 8

Sensibilité du taux de résiliation

Sommaire

L'objectif de ce chapitre est de mesurer l'impact du prix sur le taux de résiliation à l'aide de notre modèle. A ce titre, nous allons faire varier les majorations tarifaires pour évaluer l'incidence sur les résiliations.

8.1 Méthodologie

8.1.1 Modèle utilisé

Pour ce chapitre, nous allons utiliser le modèle à 31 variables calibré précédemment. Néanmoins, et afin de nous placer en situation réelle, nous allons introduire une base test. Cette base n'est alors jamais lu par le modèle, elle servira exclusivement aux prédictions.

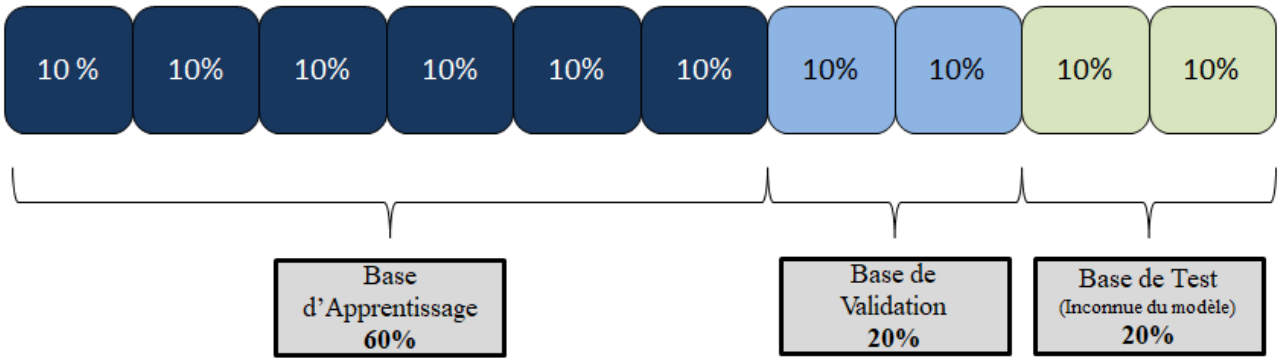

FIGURE 8.1: Découpage base d'apprentissage, de validation et de test

Les résultats de ce nouveau modèle : « GBM Exploitation », sont très proches du GBM calibré lors de la partie modélisation. L'AUC du « GBM Exploitation » de 1 point supérieur sur la base d'apprentissage

et similaire sur la base validation. L'AUC sur la base test est presque identique à celui de la base de validation. Les performances sur cette dernière base sont seulement de 0.1% inférieures.

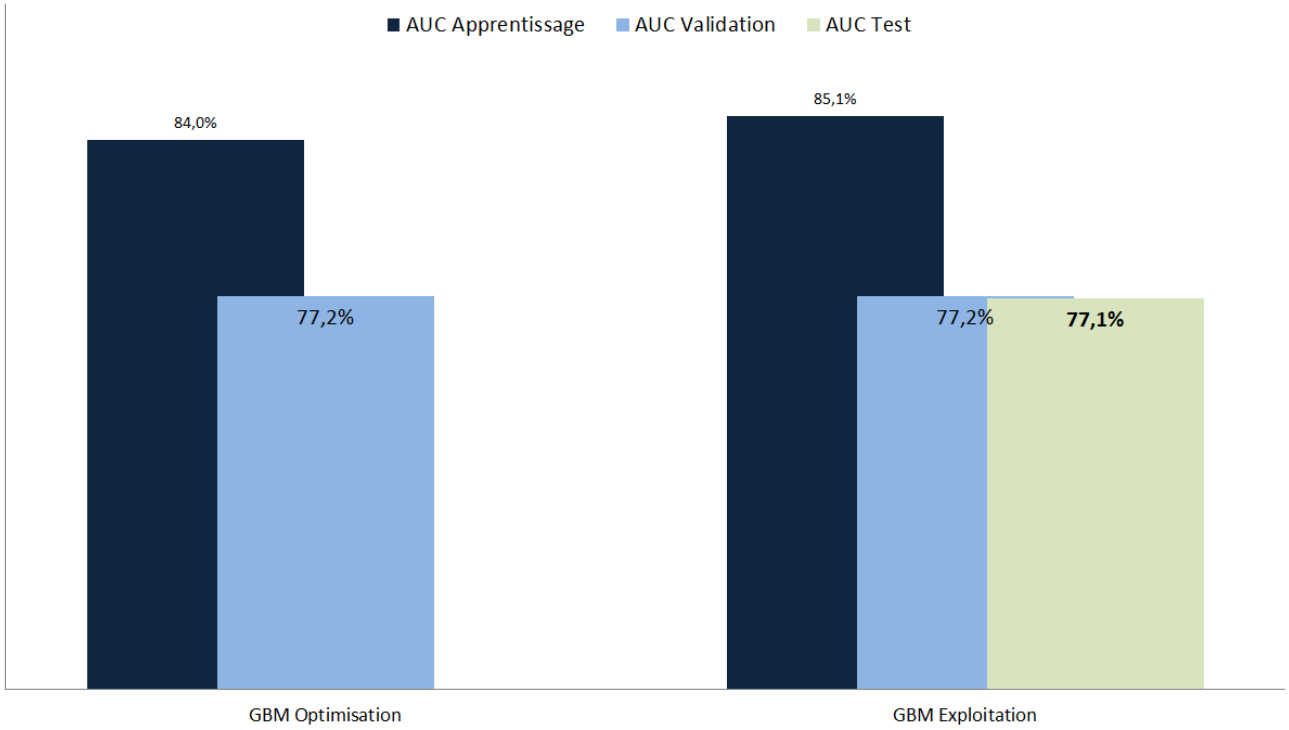

FIGURE 8.2: AUC des GBM d'optimisation et d'exploitation

Les 31 variables choisies lors du chapitre 6, sont assez peu corrélées entre elles. Les corrélations les plus fortes se trouvent sur les variables les plus importantes de notre modèle. Nous voulons un minimum d'interactions entre nos variables afin d'éviter des biais trop importants sur les graphiques de dépendance partielle.

	CD_TYPExQLTExPIECE	Generation	NATU_SITU	VA_COEF_BM_cat	SEGM_PART_PERSO	DETENTION_n	cot_net_n	evol_net_fin_n
CD_TYPExQLTExPIECE	1.00	0.30	0.14	0.03	0.23	0.09	0.55	0.16
Generation	0.30	1.00	0.31	0.06	0.43	0.01	0.28	-0.01
NATU_SITU	0.14	0.31	1.00	0.43	0.19	0.04	0.15	-0.01
VA_COEF_BM_cat	0.03	0.06	0.43	1.00	0.06	0.01	-0.03	-0.15
SEGM_PART_PERSO	0.23	0.43	0.19	0.06	1.00	-0.03	0.24	-0.01
DETENTION_n	0.09	0.01	0.04	0.01	-0.03	1.00	0.11	0.01
cot_net_n	0.55	0.28	0.15	-0.03	0.24	0.11	1.00	0.12
evol_net_fin_n	0.16	-0.01	-0.01	-0.15	-0.01	0.01	0.12	1.00

FIGURE 8.3: Matrice de corrélation des variables les plus importantes du GBM

Nous n'observons aucune corrélation très forte, hormis entre la variable cotisation COT_NET_N et la variable croisée CD_TYPE_QLTE_PIECE. En particulier, la variable d'évolution tarifaire EVOL_NET_FIN_N est très peu corrélée avec d'autres variables ce qui nous permettra d'utiliser les graphiques de dépendance partielle avec une bonne robustesse. Les résultats obtenus avec ces derniers seront comparés avec les méthodes réalisées à la main en relançant le modèle pour mesurer la sensibilité au prix.

8.1.2 Méthode 1 : Fixer la variable explicative

La première méthode utilisée afin de mesurer l'impact de la majoration tarifaire sur le taux de résiliation est de fixer différents niveaux d'évolution puis de mesurer les prédictions du modèle. Nous allons nous intéresser aux évolutions tarifaires entre 0% et 10% avec un pas de 0.5%. Pour chaque observation l'évolution est fixée comme suit :

$$x_{\%,i} = pourcentage fixé$$

Pour chaque pas d'évolution tarifaire fixée, le modèle prédit le taux de résiliation. Le tableau de sortie se présente comme suit :

Evol 0%	**Prédict 0%**	**Evol 0.5%**	**Prédict 0.5%**	**...**	**Evol 10%**	**Prédict 10%**
$x_{0\%,1}$	$\hat{f}(x_{0\%,1})$	$x_{0.5\%,1}$	$\hat{f}(x_{0.5\%,1})$	...	$x_{10\%,1}$	$\hat{f}(x_{10\%,1})$
$x_{0\%,2}$	$\hat{f}(x_{0\%,2})$	$x_{0.5\%,2}$	$\hat{f}(x_{0.5\%,2})$	...	$x_{10\%,2}$	$\hat{f}(x_{10\%,2})$
...	...	...	...	...	...	...
$x_{0\%,N}$	$\hat{f}(x_{0\%,N})$	$x_{0.5\%,N}$	$\hat{f}(x_{0.5\%,N})$	...	$x_{10\%,N}$	$\hat{f}(x_{10\%,N})$

8.1.3 Méthode 2 : Stresser la variable explicative

La seconde méthode consiste à modifier la majoration tarifaire déjà appliquée et observer l'impact sur les prédictions du modèle. Pour chaque observation l'évolution est fixée comme suit :

$$x_{choc_j,i} = x_i + choc_j$$

Nous allons appliquer des chocs allant de -2% à +5% sur la majoration. Le tableau de sortie se présente comme suit :

Choc -2%	**Prédict -2%**	**...**	**Initial**	**Prédict Initiale**	**...**	**Choc +5%**	**Prédict +5%**
$x_{choc_{-2},1}$	$\hat{f}(x_{choc_{-2},1})$	...	x_1	$\hat{f}(x_1)$	...	$x_{choc_{+5},1}$	$\hat{f}(x_{choc_{+5},1})$
$x_{choc_{-2},2}$	$\hat{f}(x_{choc_{-2},2})$	...	x_2	$\hat{f}(x_2)$	...	$x_{choc_{+5},2}$	$\hat{f}(x_{choc_{+5},2})$
...	...	...	...	...	...	...	...
$x_{choc_{-2},N}$	$\hat{f}(x_{choc_{-2},N})$	...	x_N	$\hat{f}(x_N)$	...	$x_{choc_{+5},N}$	$\hat{f}(x_{choc_{+5},N})$

8.2 Sensibilité à la variable d'évolution tarifaire

8.2.1 Méthode 1 : Fixer la majoration

Nous avons voulu étudier le comportement des résiliations en fonction des évolutions tarifaires appliquées. Pour aller plus loin dans l'analyse, en plus de l'étude sur l'ensemble du portefeuille, 4 classes ont été constituées :

— **Le Global (Somme des 4 classes)** : représenté en pointillé gris ;
— **Maison Propriétaire** : représentée en bleu marine ;
— **Maison Locataire** : représentée en bleu ciel ;

— **Appartement Propriétaire** : représentée en rouge ;
— **Appartement Locataire** : représentée en orange ;

La création de ces différentes classes va nous permettre de déterminer si ces populations se comportent de la même manière face à une hausse tarifaire. Cette information pourra ensuite nous aiguiller pour ajuster les majorations sur notre portefeuille.

Pour chacune de ces classes, nous allons mesurer l'augmentation du nombre de résiliations à partir du point pivot calculé pour chacune des classes. Ce point pivot est le nombre de résiliations prédit par le modèle lorsque la majoration tarifaire est fixée à 0.5%. Par exemple, pour une majoration de 10% sur le global portefeuille nous observons une augmentation de près de 11.5% du nombre de résiliations par rapport au point pivot.

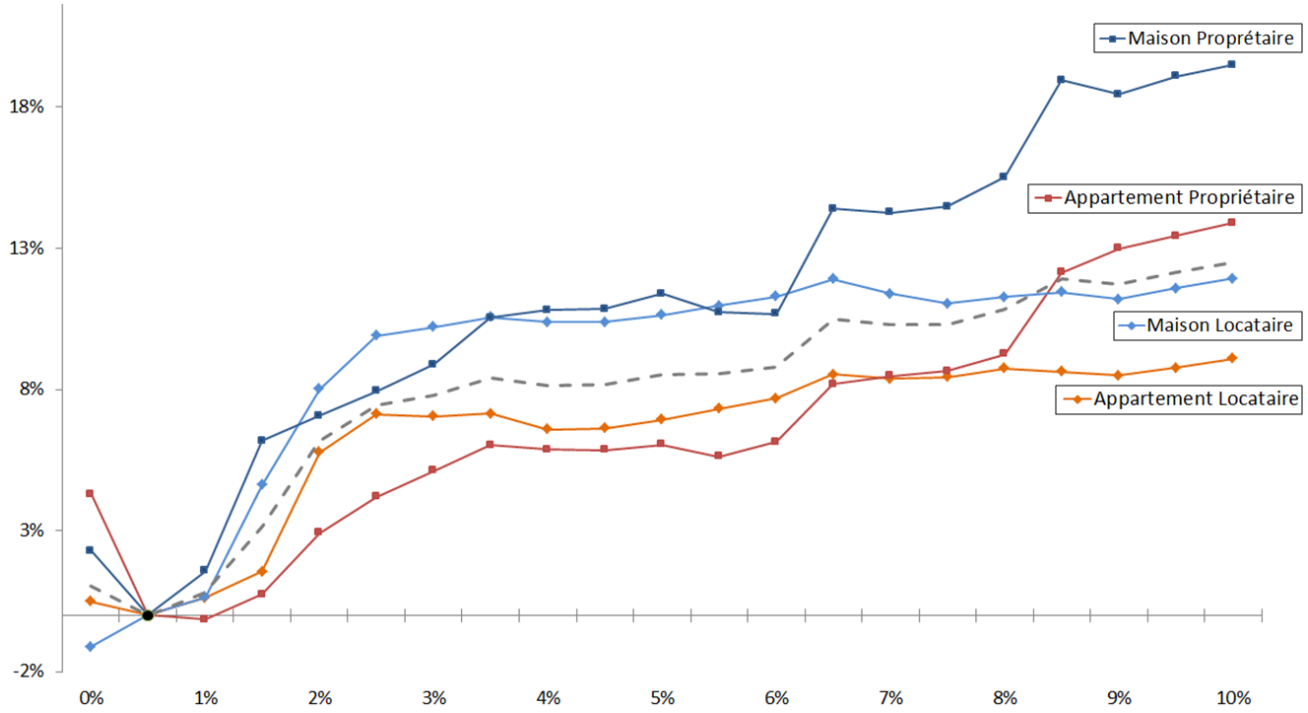

FIGURE 8.4: Évolution du nombre de résiliations en fonction de la majoration tarifaire fixée

Les plages de majorations peuvent être découpées en 3 segments, le premier de 0.5% à 3.5%, le deuxième de 4% à 6% et enfin le dernier de 6.5% à 10%.

— **Segment 1** $[0.5\%; 3.5\%]$: Sur ce segment, les maisons ont une sensibilité plus importante que les appartements. Les résiliations ont augmenté de 10% pour les maisons contre 7% pour les appartements.
— **Segment 2** $[4\%; 6\%]$: Sur ce segment, les résiliations stagnent quelque soit la classe observée. Nous remarquons même des diminutions locales pour les locataires.
— **Segment 3** $[6.5\%; 10\%]$: Sur ce segment, ce sont les propriétaires qui subissent une forte hausse, tandis que les résiliations sur les locataires stagnent.

Pour résumer les graphiques et les commentaires :

Classe	[0.5%; 3.5%]	[4%; 6%]	[6.5%; 10%]
Maison Propriétaire	Forte Hausse	Stable	Hausse
Maison Locataire	Forte Hausse	Stable	Stable
Appartement Propriétaire	Hausse	Stable	Hausse
Appartement Locataire	Hausse	Stable	Stable

8.2.2 Méthode 2 : Stresser la majoration

Cette seconde méthode, ne fixe pas les majorations tarifaires mais fait varier celles déjà appliquées. Elle permet alors d'étudier la sensibilité au prix des différentes classes mais d'un autre point de vue. En moyenne, les populations ne subissent pas les mêmes évolutions chaque année, donc cette méthode compare les sensibilités avec un point pivot qui n'est pas calculé avec le même pourcentage de revalorisation.

Sur le graphique, la première observation frappante est la linéarité de l'évolution du nombre de résiliations pour l'ensemble des classes. Sur les maisons propriétaires, la linéarité est presque parfaite.

Sur la partie gauche du graphique, nous pouvons étudier l'impact d'une diminution des revalorisations. Pour toutes les classes hormis les appartements propriétaires, la réduction de la majoration tarifaire entraîne une baisse des résiliations avec la même force qu'une augmentation de majoration. Pour sa part, la population d'appartements propriétaires est beaucoup moins sensible à une diminution de son niveau de majoration. Appliquer -2% sur cette population affecte presque nullement les résiliations.

La partie droite du graphique montre comment les classes sont sensibles aux augmentations des revalorisations. Le classement des populations les plus sensibles est identique à la méthode numéro 1. Les maisons propriétaires semblent être les plus vulnérables aux évolutions positives, tandis que les appartements locataires semblent les moins vulnérables.

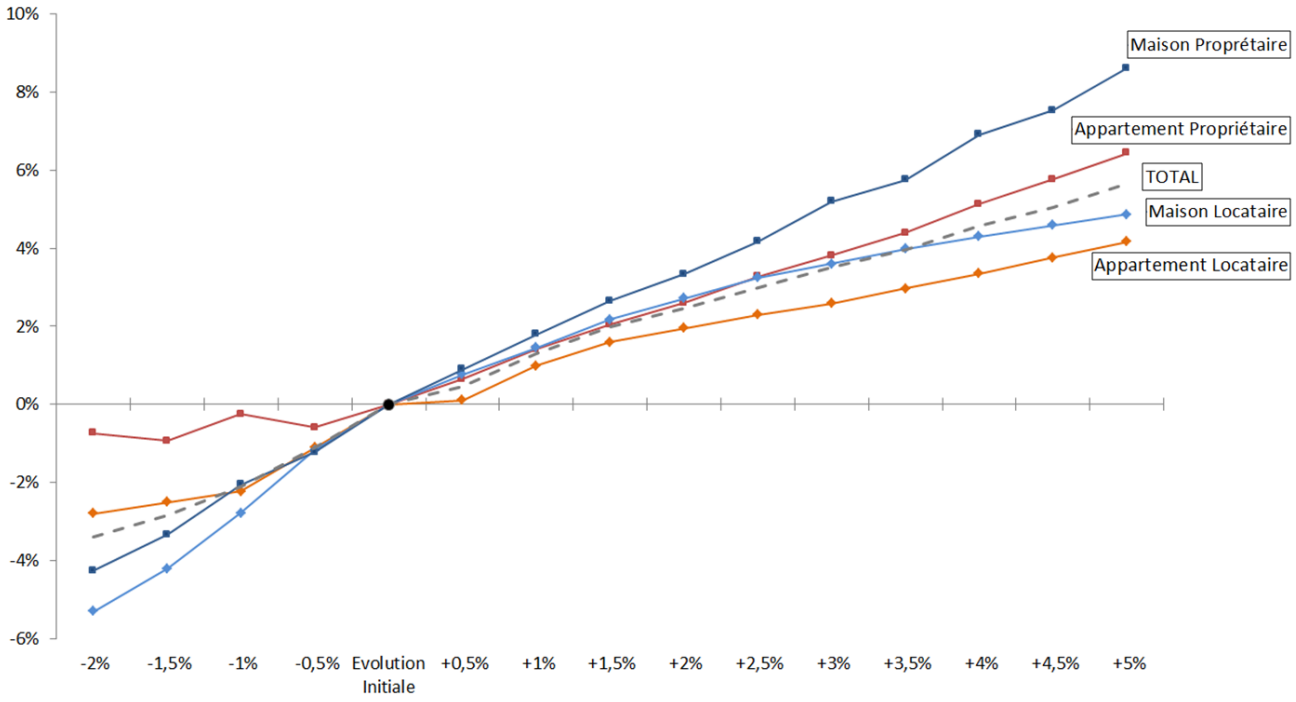

FIGURE 8.5: Évolution du nombre de résiliations en fonction du choc sur la majoration tarifaire

Les résultats des deux méthodologies sont cohérents. Il est possible de retrouver des résultats semblables à la première méthode en uniformisant les points pivots de chaque classe. Par exemple, en le déterminant uniformément entre les classes avec une évolution tarifaire à 0.5%.

8.2.3 Comparaison avec le PDP

Les résultats obtenus précédemment à l'aide de méthodes réalisées à la main sont comparés aux résultats du graphique de dépendance partielle. Les principes sont identiques dans les deux cas, l'objectif est de mesurer l'effet marginal de la variable explicative sur le taux de résiliation.

En moyenne, les résultats doivent être identiques car les sorties sont agrégées. En effet, la méthode 1 peut capter les interactions entre les variables explicatives car le modèle est entièrement relancé. En revanche, le graphique de dépendance partiel ne capture pas ces interactions. Les prédictions individuelles peuvent alors être différentes entre la méthode 1 et la méthode des graphiques de dépendance partielle.

Sur la figure ci-dessous, l'évolution du nombre de résiliations est comparée à l'effet marginal calculé par le PDP. En moyenne sur le portefeuille, les évolutions sont identiques entre la méthode 1 et l'effet marginal issu du PDP.

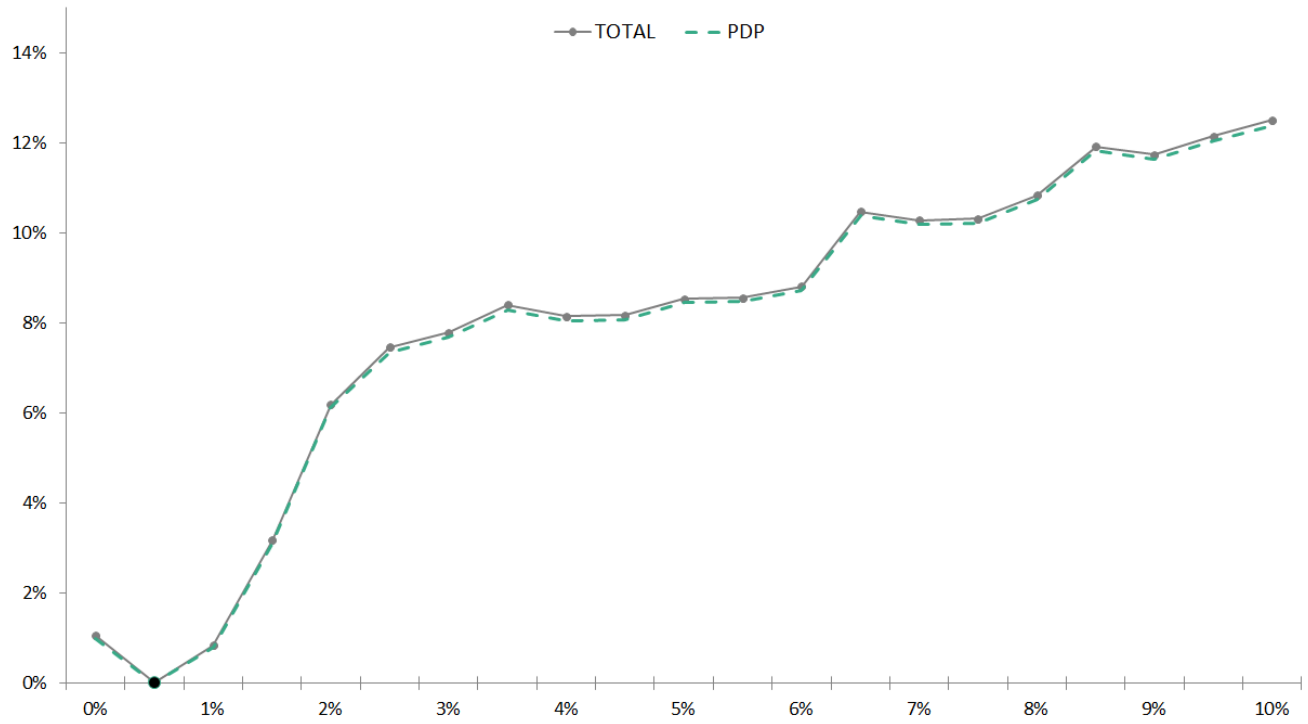

FIGURE 8.6: Évolution du nombre de résiliations en univarié en fonction de la variable d'évolution tarifaire : Méthode 1 vs PDP

8.3 Sensibilité à la génération du contrat

A l'inverse de la génération du contrat, la variable évolution tarifaire n'est pas une variable explicative très impactante selon le GBM. La variable explicative de l'ancienneté du contrat est la plus importante du modèle, donc celle qui devrait majoritairement influer sur les résultats. Nous avons alors réalisé le même exercice que précédemment.

8.3.1 Méthode 1 : Fixation de la génération

Comme pour l'évolution tarifaire, les 4 mêmes classes vont être étudiées afin de mesurer leur sensibilité à l'ancienneté. La lecture des graphiques est identique à la section précédente. Le nombre de résiliations pivot est le nombre prédit par le modèle lorsque la génération est fixée à 1.

Sur ce premier graphique, nous avons donc fixé pour l'ensemble de nos situations, l'ancienneté du contrat. Logiquement et comme le montrait les résultats Figure 7.12, plus le contrat est ancien en portefeuille, moins celui-ci à de chance de résilier.

Hormis, sur le point de la génération 3 ans, toutes les classes se comportent de la même manière. Lorsque le contrat est récent, à savoir inférieur à 5 ans, le fait d'augmenter d'un an la génération diminue très fortement le taux de résiliation. Toutefois pour les contrats plus anciens, la génération n'influe pas sur les résiliations. Le modèle transcrit aucune différence entre un contrat de 10 ou 11 ans par exemple.

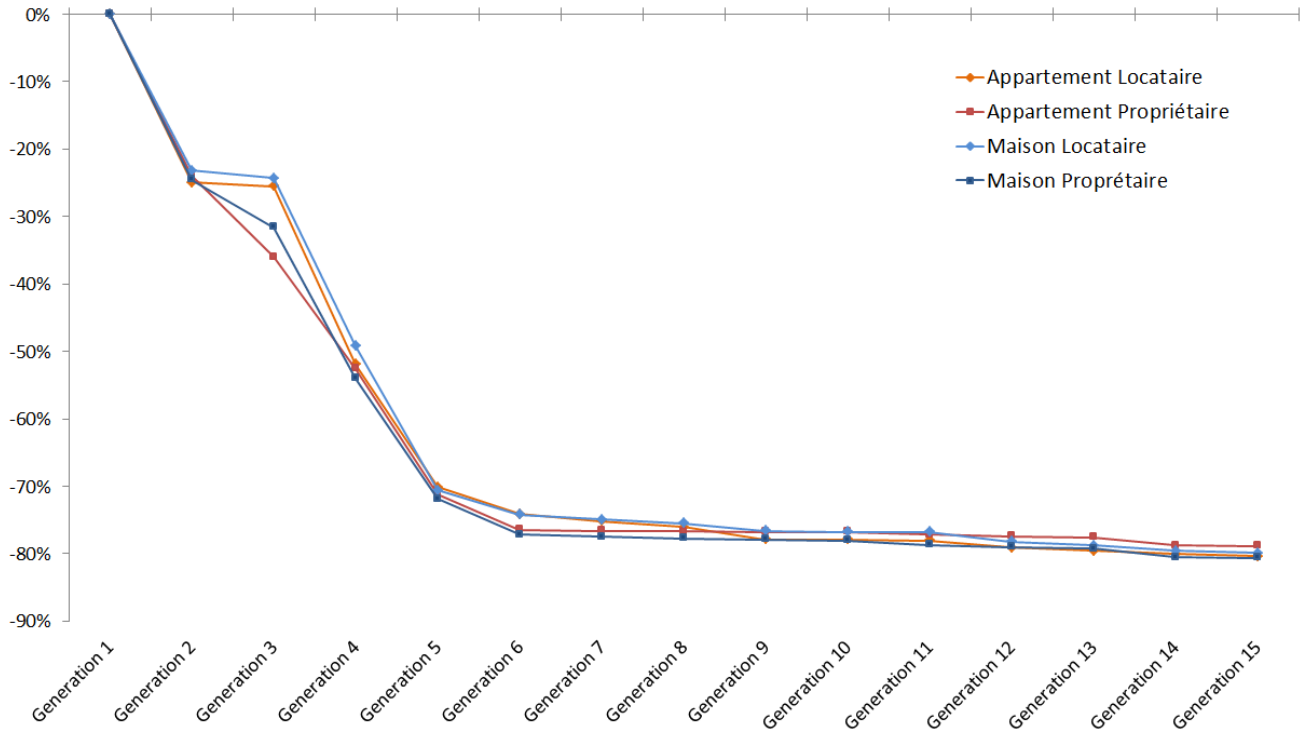

FIGURE 8.7: Évolution du nombre de résiliations en fonction de la génération du contrat fixée

8.3.2 Méthode 2 : Choc sur la génération

Avec cette seconde méthode, les caractéristiques du portefeuille sont prises en compte. Par conséquent, les propriétaires ont une sensibilité plus faible à une augmentation de l'ancienneté car celle-ci est généralement élevée. A l'inverse, les locataires subissent une forte baisse de leur résiliations.

Ce phénomène s'explique à l'aide du graphique 8.7, du fait que la pente est beaucoup plus abrupte sur les contrats avec peu d'ancienneté. Ces contrats à faible ancienneté sont majoritairement des locations.

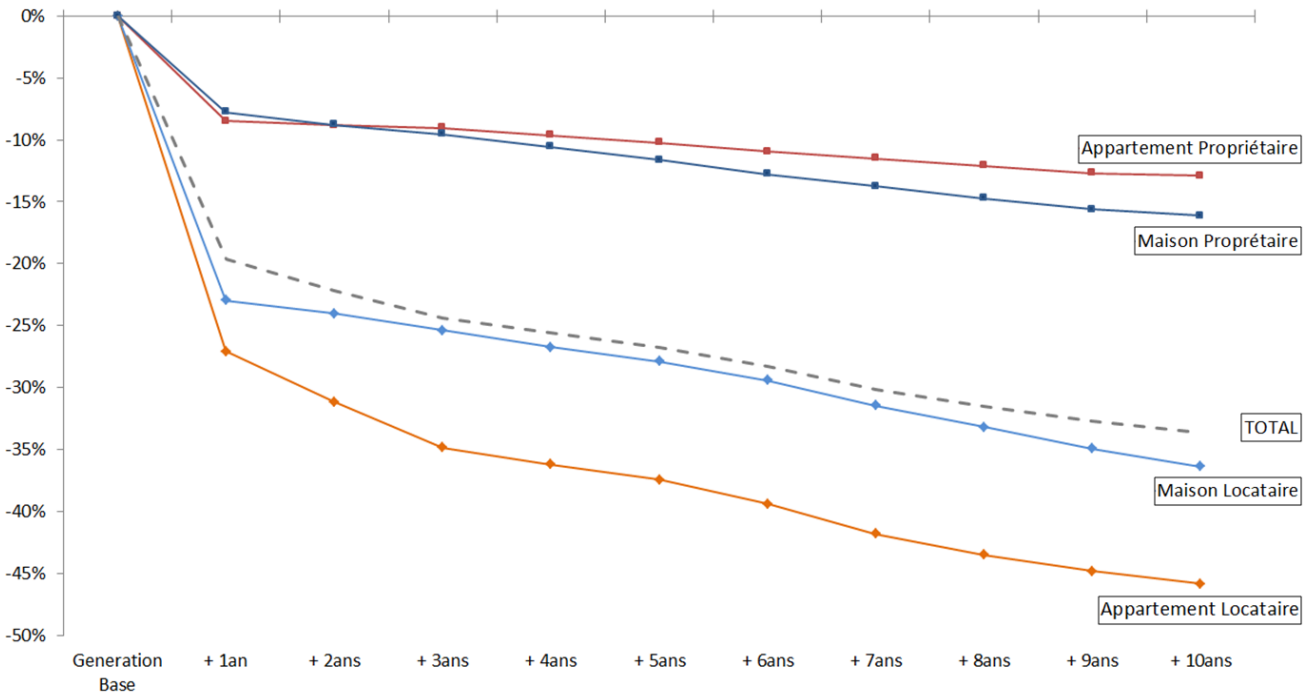

FIGURE 8.8: Évolution du nombre de résiliations en fonction du choc sur la génération du contrat

8.3.3 Comparaison avec le PDP

Comme pour la variable d'évolution tarifaire, les résultats du graphique de dépendance partielle sont identiques à ceux de la méthode réalisée à la main où l'ancienneté est fixée. Cet exemple montre une nouvelle fois la robustesse des résultats obtenus à l'aide du graphique de dépendance partielle. Les effets des interactions avec les autres variables n'influent pas ou très peu sur la variable explicative de la génération du contrat. Le PDP représente presque parfaitement l'effet de cette caractéristique sur les résiliations au global.

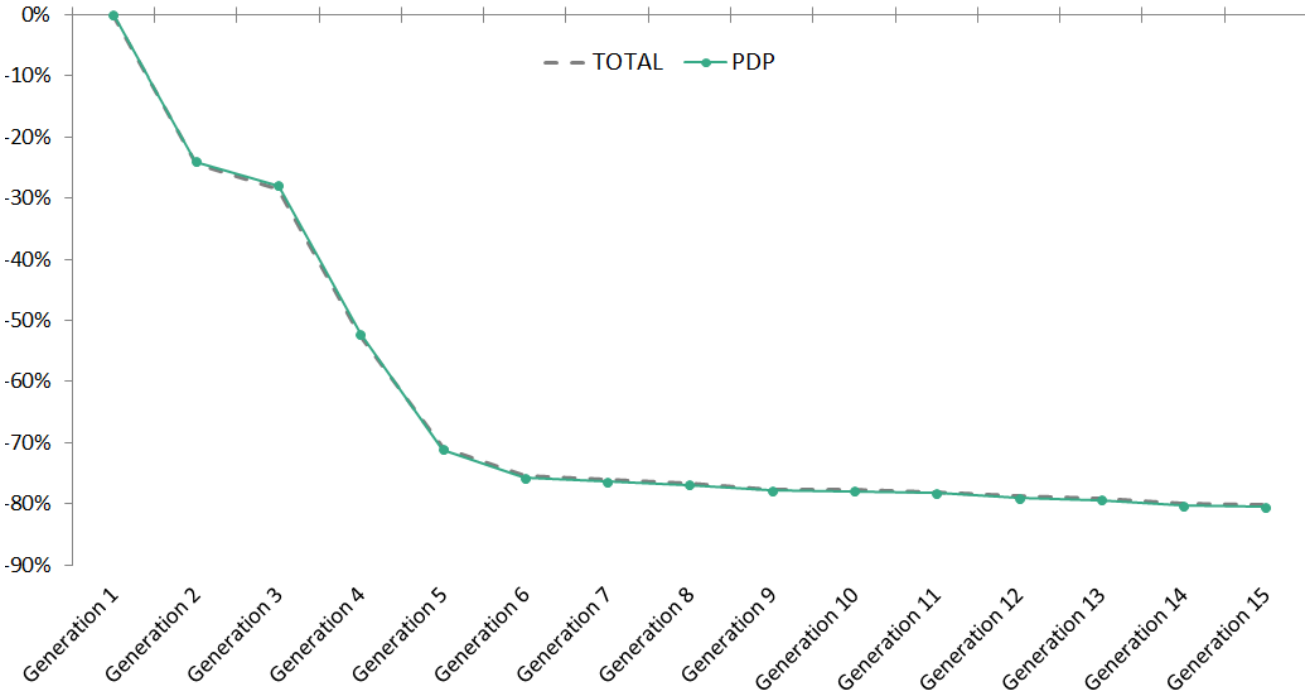

FIGURE 8.9: Évolution du nombre de résiliations en fonction de la variable génération : Méthode 1 vs PDP

Les résultats sur la variable GENERATION sont bien plus segmentants, néanmoins le fait que cette variable soit difficile à modifier rend les résultats moins intéressants d'un point de vue opérationnel. En effet, étudier les variables les plus importantes du GBM est très intéressant car les résultats en sorties seront très impactés. Pourtant, opérationnellement l'étude peut être peu profitable comme nous l'avons vu avec la variable GENERATION. Il y a très peu de levier qui permettent de modifier cette ancienneté de contrat, mise à part essayer de conserver au maximum les assurés au sein de MMA.

De surcroît, le modèle montre que les variables explicatives les plus impactantes ne sont pas modifiables : La génération du contrat, le type d'habitation, la qualité juridique de l'occupant et le segment du client. Toutes ces variables sont des caractéristiques propres à l'assuré et à son bien. Ces résultats doivent être couplés avec une étude sur les types de profils en entrée afin d'optimiser le coût d'entrée avec la probabilité de sortie.

8.4 Exploitation du graphique de dépendance partielle

8.4.1 Sorties du PDP

Afin d'exploiter plus en profondeur les résultats du graphique de dépendance partielle, nous avons étudié la sensibilité du taux de résiliation aux majorations non pas au global, mais avec les 4 classes constituées précédemment. Le PDP nous donne pour chaque profil la réponse moyenne du modèle. Dans notre cas par exemple :

Type x Qlte x Pièces	**Évolution tarifaire nette**	**Réponse moyenne**
AL_1	0%	0.378
AL_1	0.5%	0.326
...	...	...
MP_9	9.5%	0.121
MP_9	10%	0.108

8.4.2 Comparaison GBM vs PDP croisé

Pour comparer les résultats obtenus à la main avec le Gradient Boosting et ceux du graphique de dépendance partielle, nous allons reprendre les sorties de la Figure 8.4. Ce graphique mesure les augmentations du nombre de résiliation pour le GBM qui vont être comparées aux augmentations calculées à l'aide des sorties du PDP.

Les résultats du GBM sont représentés par les courbes continues tandis que les résultats du PDP le sont par des courbes en pointillé. Globalement, nous observons que contrairement aux figures 8.6 et 8.9 les résultats par classe ne sont pas parfaits.

A gauche, pour la classe maison propriétaire le PDP estime que celle-ci est moins sensible aux majorations que la méthode réalisée à la main. Sur les appartements propriétaires les résultats sont relativement proches entre les deux calculs.

A droite, le PDP sous estime toujours la sensibilité aux revalorisations, que ce soit pour les maisons ou les appartements locataires.

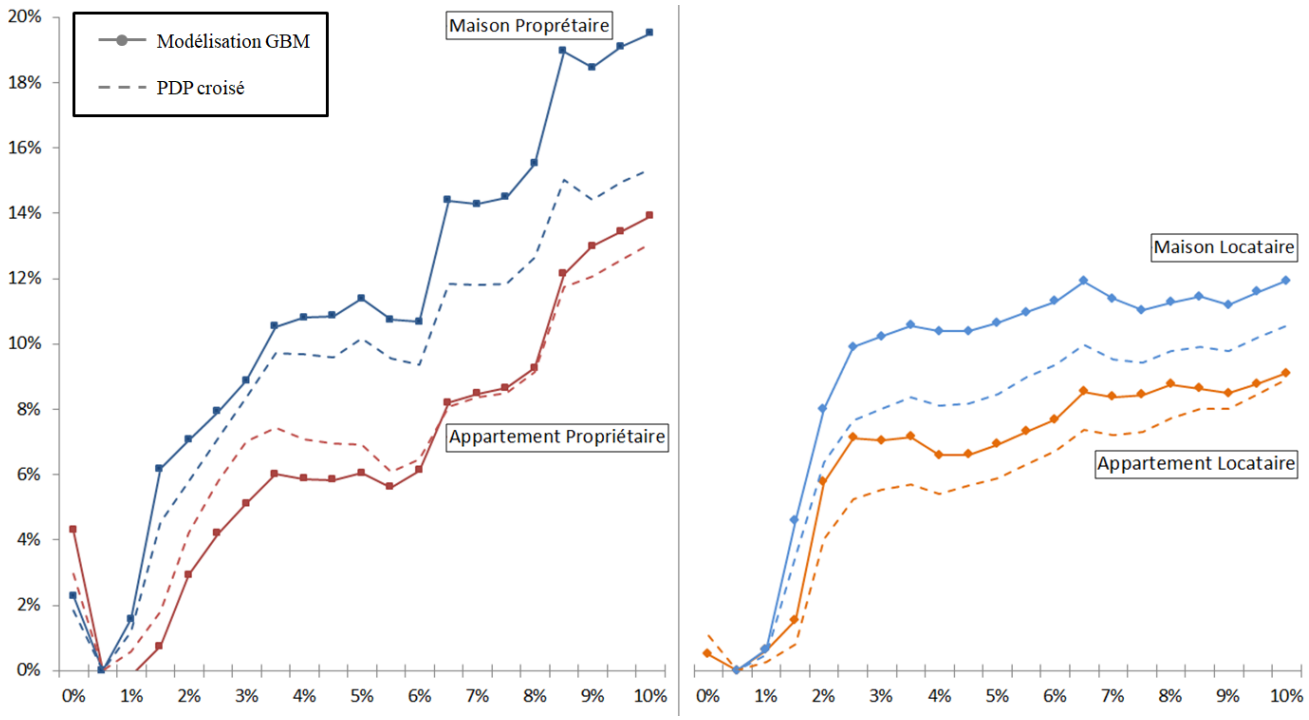

FIGURE 8.10: Évolution du nombre de résiliations sur les différentes classes en fonction de la variable des revalorisations : GBM vs PDP croisé

Ces différences viennent du fait que le graphique de dépendance partielle n'affiche que l'effet marginal des variables explicatives étudiées. A l'inverse, la méthode faite main en relançant le GBM pour chaque niveau d'évolution tarifaire capte les effets de toutes les variables explicatives et les interactions entre elles. Par exemple, l'effet de la variable GENERATION n'est pas mesuré par PDP ce qui entraîne un biais.

Afin d'illustrer partiellement ce biais, nous pouvons comparer non pas les évolutions à partir d'un point pivot mais le taux de résiliation entre les deux méthodes. Pour illustrer parfaitement le biais, il faudrait comparer le PDP croisé à l'agrégation de PDP univariés (cf. section 8.4.3).

Sur le graphique ci-dessous, le PDP capture bien les fluctuations du taux de résiliation en fonction de la majoration. En effet, sur la partie gauche, le même creux est visible pour le GBM et le PDP autour de 0.5% et les différents plateaux le sont également.

Néanmoins, la réponse moyenne du graphique de dépendance partielle se retrouve très décadrée pour les maisons propriétaires et les appartements locataires.

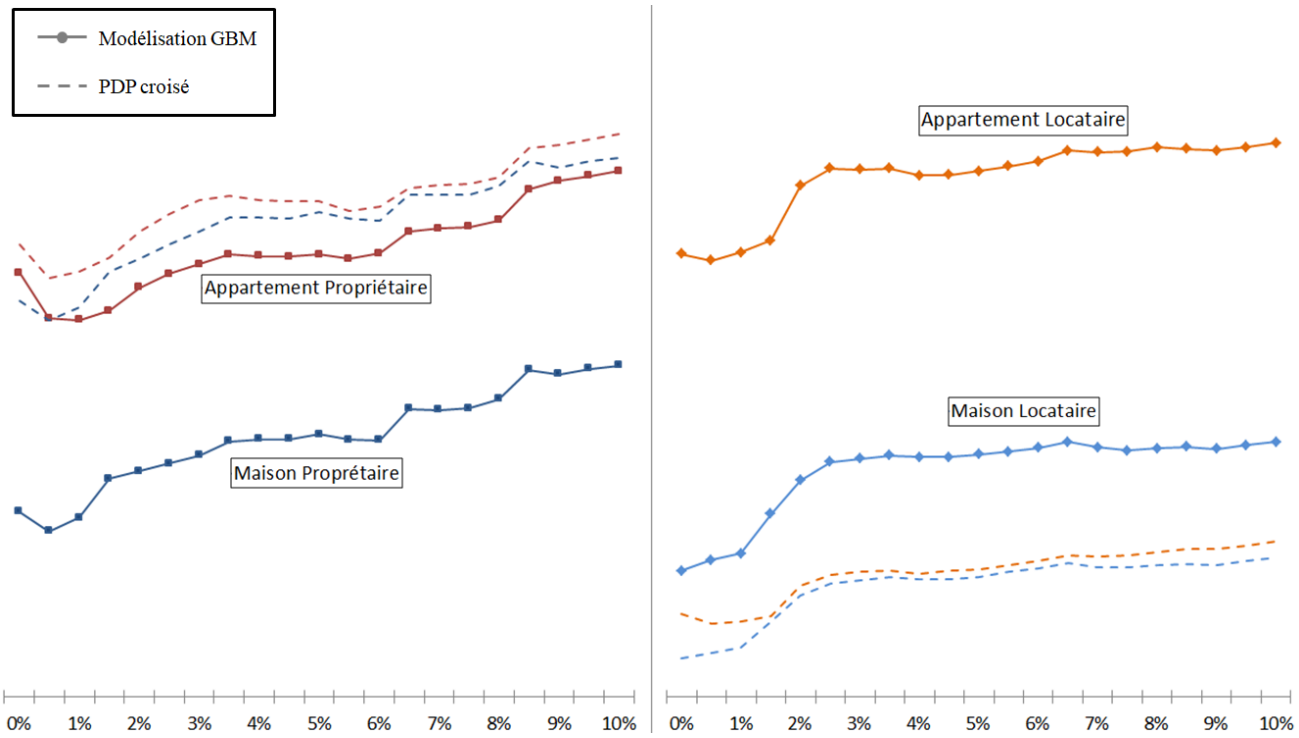

FIGURE 8.11: Taux de résiliations sur les différentes classes en fonction de la variable des revalorisations : GBM vs PDP croisé

Afin d'obtenir des résultats parfaits, il est indispensable de créer des PDP croisés avec l'exhaustivité des variables explicatives. Cette méthode est inapplicable à cause de la limitation à deux variables croisées à la fois. Toutefois, même si nous avions la possibilité d'utiliser plus de deux variables explicatives les temps de traitement rendrait l'exercice impossible. A notre échelle, avec nos serveurs et la taille de la base le calcul des réponses moyennes du PDP croisés entre la variable CD_TYPExQLTExPIECE et d'évolution tarifaire nécessite près de 5H. Avec 29 autres variables à croiser l'exercice est insurmontable.

8.4.3 Comparaison GBM vs PDP croisé vs PDP créé manuellement

Les graphiques de dépendance partielle sont restreints par la limite à deux croisements. Cependant il est possible d'agréger les résultats de plusieurs PDP en univarié afin de capturer les effets de toutes les variables explicatives du modèle. L'ennui de cette méthode est la perte des interactions entre les variables. En effet, la force du PDP croisé était de capturer différentes sensibilités en fonction de la population étudiée. Par exemple, les maisons propriétaires n'ont pas la même sensibilité au prix que les 3 autres classes.

Les résultats des trois méthodes sont comparés :

- **Modélisation GBM** : Résultats obtenus en exécutant le modèle de Gradient Boosting avec plusieurs niveaux d'évolution tarifaire.
- **PDP croisé** : Réponse moyenne calculée par le modèle en prenant compte des interactions entre les deux variables explicatives étudiées.
- **PDP manuel** : Agrégation des graphiques de dépendance partielle univariés des deux variables explicatives étudiées

Comparaison de l'évolution du nombre de résiliations Sur le graphique ci-dessous, les résultats du PDP créé manuellement ont été ajoutés. Nous remarquons que les courbes se chevauchent car nous avons perdu les effets d'interaction entre la variable CD_TYPEXQLTEXPIECE et la variable d'évolution tarifaire. C'est-à-dire que quelque soit la population étudiée le PDP manuel capture les mêmes sensibilités aux majorations.

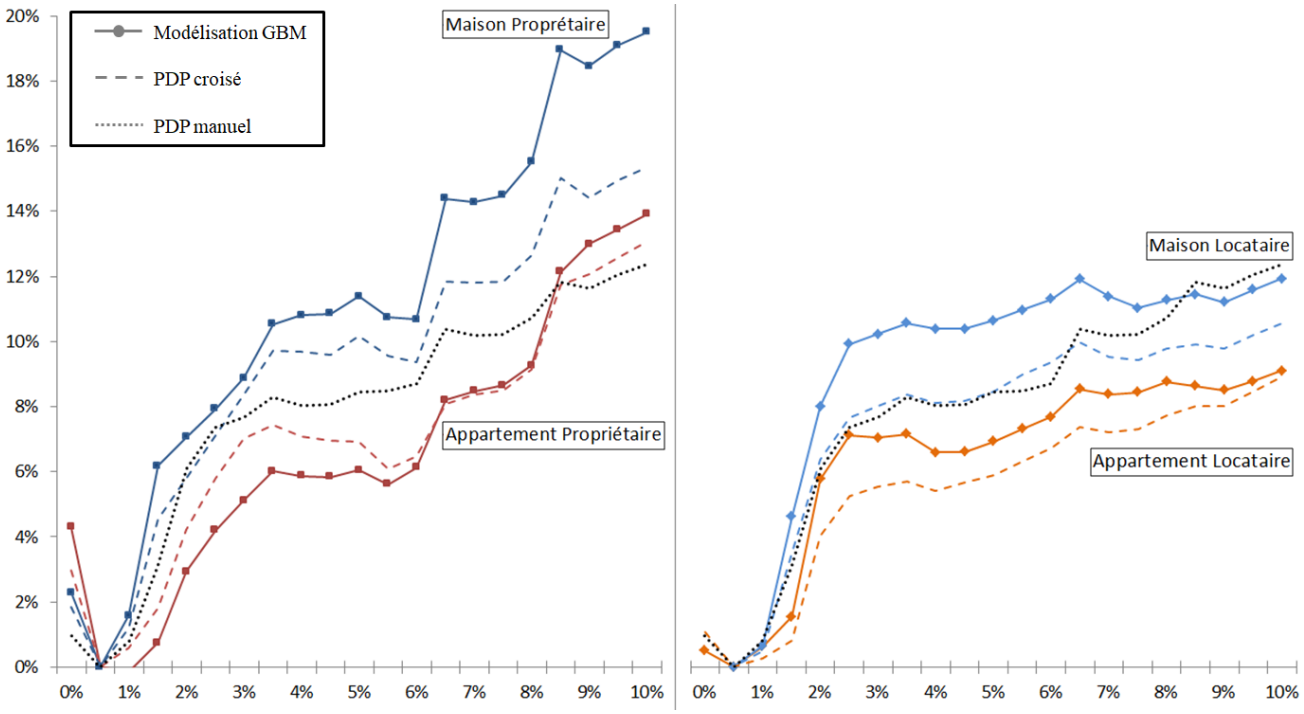

FIGURE 8.12: Évolution du nombre de résiliations sur les différentes classes en fonction de la variable des revalorisations : GBM vs PDP croisé vs PDP manuel

8.4.4 Mise en production opérationnelle du PDP

Opérationnellement les modèles linéaires généralisés sont préférés aux modèles de Machine Learning car ces premiers fournissent des coefficients facilement exploitables. En effet, un modèle Logit est dit multiplicatif, c'est-à-dire que le produit de tous ses coefficients permet d'obtenir parfaitement le résultat prédit par le modèle. Par conséquence, des « calculatrices » sont créées à l'aide des coefficients et permettent donc de prédire les résultats du modèle sans avoir à lancer celui-ci.

L'objectif de cette section est d'exploiter les réponses moyennes du PDP comme les coefficients d'un GLM afin de retrouver le résultat du modèle GBM sans avoir à l'exécuter. Pour cela, nous mesurons pour nos différentes classes l'écart à la moyenne des résiliations. L'utilisation de points pivots dans l'agrégation des réponses moyennes des PDP univariés créent des biais si nous utilisons les réponses moyennes brutes.

Ce premier graphique montre l'écart à la moyenne globale en pourcentage pour les 4 populations et cela pour les 3 résultats :

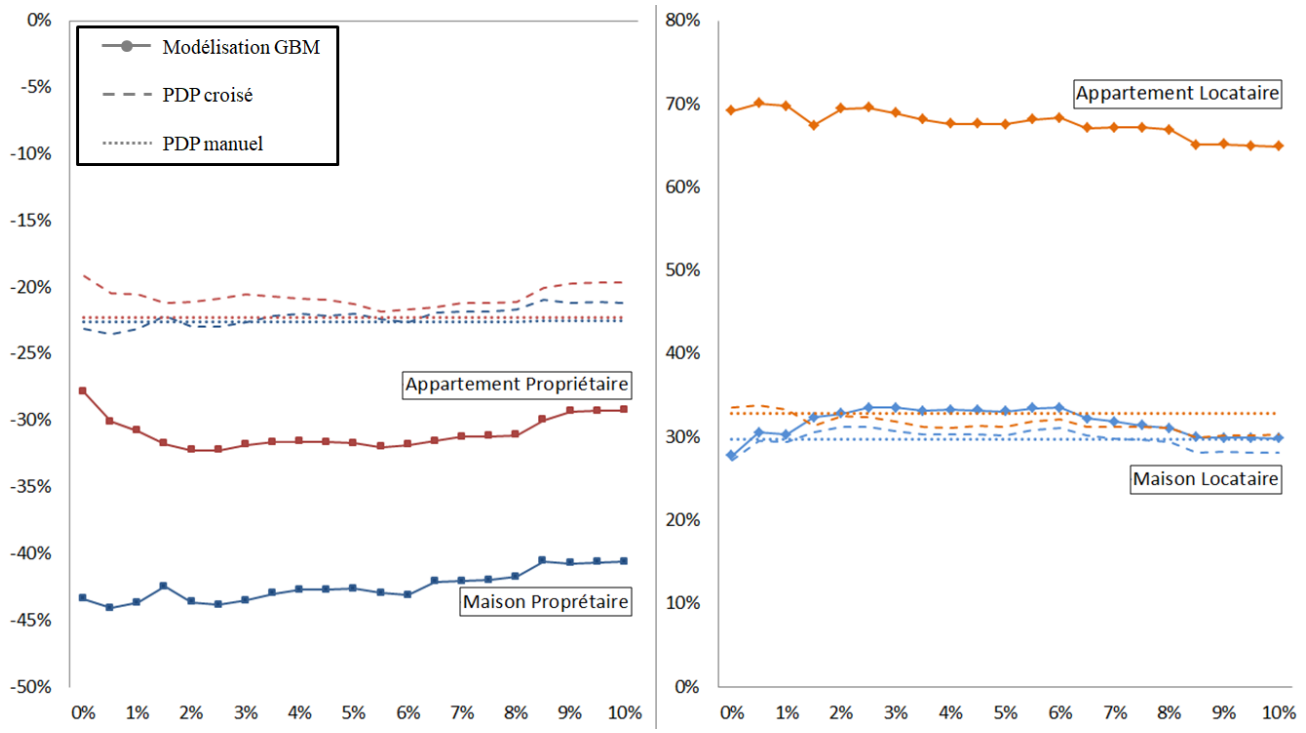

FIGURE 8.13: Écart à la moyenne des différentes classes en fonction de la variable des revalorisations : GBM vs PDP croisé vs PDP manuel

Les résultats ne sont pas convaincants. Hormis pour les maisons locataires, les réponses moyennes des PDP sont décadrées. La figure 8.13 compare seulement les effets marginaux des deux variables CD_TYPEXQLTEXPIECE et EVOL_NET_N contre les réponses du modèle calculées à l'aide de l'ensemble des 31 variables explicatives. Afin de réduire l'écart, nous allons utiliser le graphique de dépendance partielle en univarié de la variable GENERATION.

Entre les figures 8.13 et 8.14 l'écart s'est fortement réduit. Sur les populations de maisons propriétaires (bleu marine) et d'appartements locataires (orange), la différence de résultats est assez faible. Cependant sur les deux autres classes la différence est non négligeable.

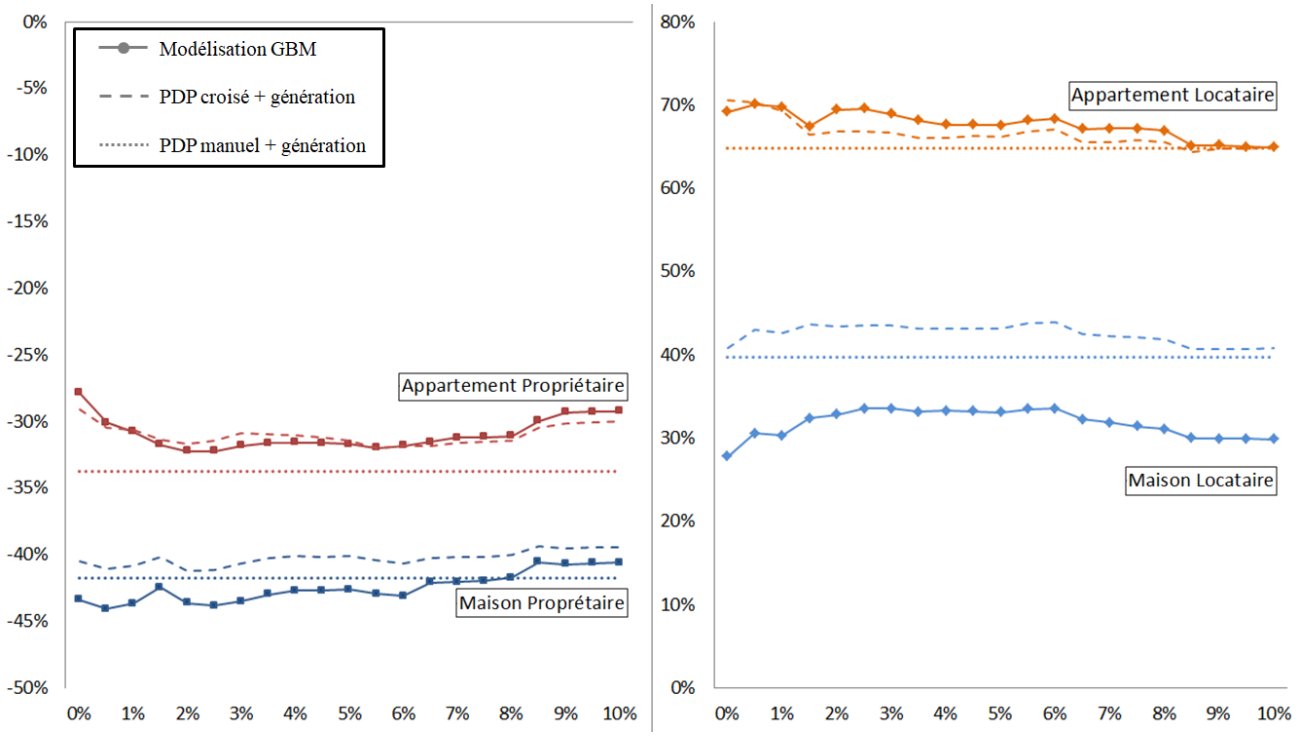

FIGURE 8.14: Écart à la moyenne des différentes classes en fonction de la variable des revalorisations avec l'ajout de la variable génération du contrat : GBM vs PDP croisé vs PDP Manuel

Pour essayer de mesurer parfaitement l'écart entre la prédiction du modèle et les réponses moyennes du PDP, il faudrait ajouter les 28 autres variables comme réalisé pour la génération du contrat. Dans le cas de variables indépendantes, le résultat entre l'agrégation des PDP et la prédiction du modèle serait parfaite. Cependant, en cas pratique cette hypothèse n'est jamais vérifiée. L'utilisation des graphiques ALE (cf. section 7.4.2) pourrait pallier au problème d'hypothèse d'indépendance du PDP et donc faire office de « calculatrice » au même titre que les odds ratio obtenus à l'aide des GLM.

Afin de pouvoir utiliser le PDP il serait nécessaire de pouvoir mesurer mathématiquement cet écart dans le but de prouver qu'en deçà d'un certain seuil l'agrégation des PDP pourra être appliquée opérationnellement. L'utilisation du Gradient Boosting a l'avantage de capturer des effets non linéaires ce qui est impossible à faire à l'aide d'un GLM. Sur la variable d'évolution tarifaire par exemple le coefficient du GLM ne donne qu'une information linéaire. Nous le remarquons parfaitement sur le graphique ci-contre.

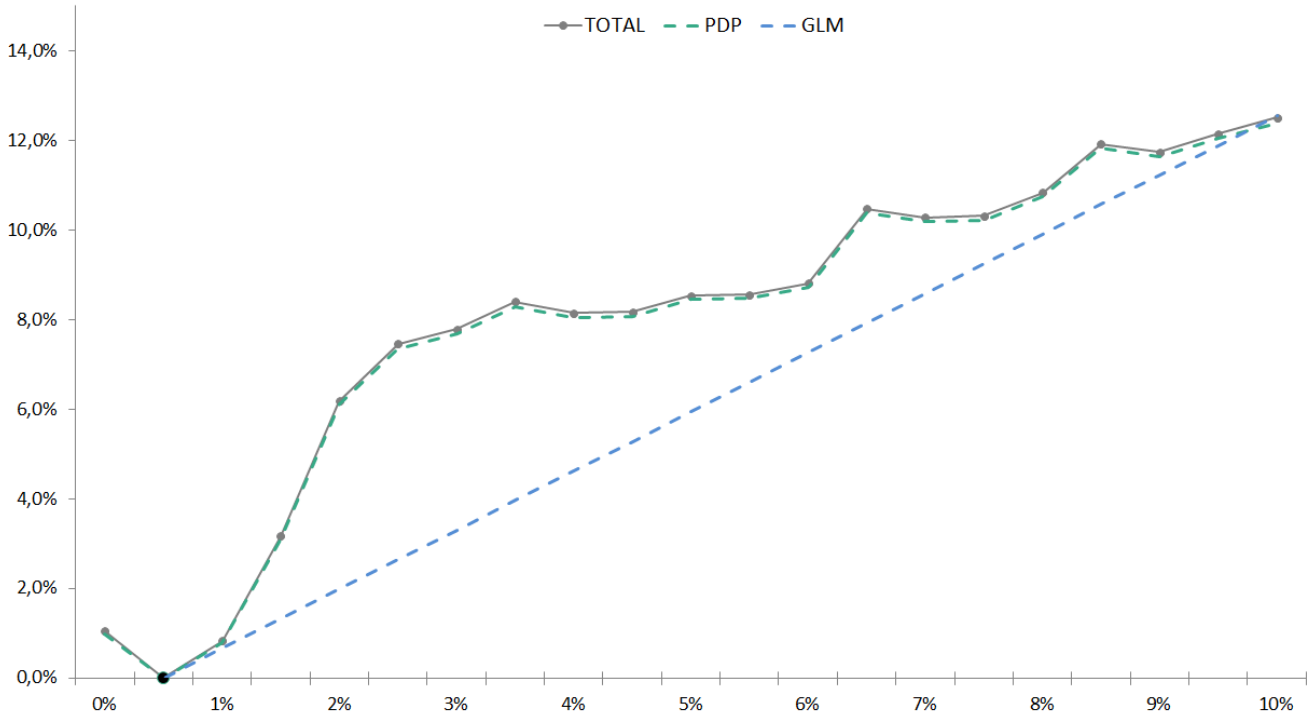

FIGURE 8.15: Évolution du nombre de résiliations en fonction de la variable d'évolution tarifaire : Méthode 1 vs PDP vs GLM

Chapitre 9

Optimisation de la majoration tarifaire

Sommaire

Ce dernier chapitre a pour vocation d'exploiter un maximum le modèle de résiliation, à savoir essayer de prédire le chiffre d'affaires en fonction de la majoration tarifaire.

9.1 Cadre de l'exercice d'optimisation

9.1.1 Modèle utilisé

Le modèle utilisé est identique à celui du chapitre précédent. Les résultats présentés sont calculés sur la même base de test. Sur ce chapitre, nous n'allons pas raisonner en terme d'évolution du nombre de résiliations mais bien en taux de résiliation déterminé à partir du portefeuille.

Afin de mesurer la sensibilité du chiffre d'affaires et tenter de trouver un optimum en fonction des revalorisations, nous utiliserons les mêmes méthodes et outils du chapitre précédent.

A noter, qu'aucune hypothèse n'est posée sur les affaires nouvelles. Pour aller plus loin dans cette simulation et créer un environnement où le portefeuille se rapproche d'une situation réelle, il est nécessaire de créer un modèle pour les affaires nouvelles.

9.1.2 Méthode 1 : Fixer de la majoration

La valeur qui va être étudiée est donc le chiffre d'affaires. Ce chiffre d'affaires est la somme de toutes les cotisations de notre portefeuille. Nous considérons alors que la base test est notre portefeuille à l'instant t. Plusieurs simulations du taux de résiliation sont réalisées avec différents niveaux de majorations tarifaires. A l'aide de ces taux, nous déterminons alors le chiffre d'affaires prévisionnel qui est calculé comme suit :

$$CA_{ccl} = \sum_{n=1}^{N} (1 - \mathbb{1}_{résil_i}) \times Cotisation_i \times (1 + majoration)$$

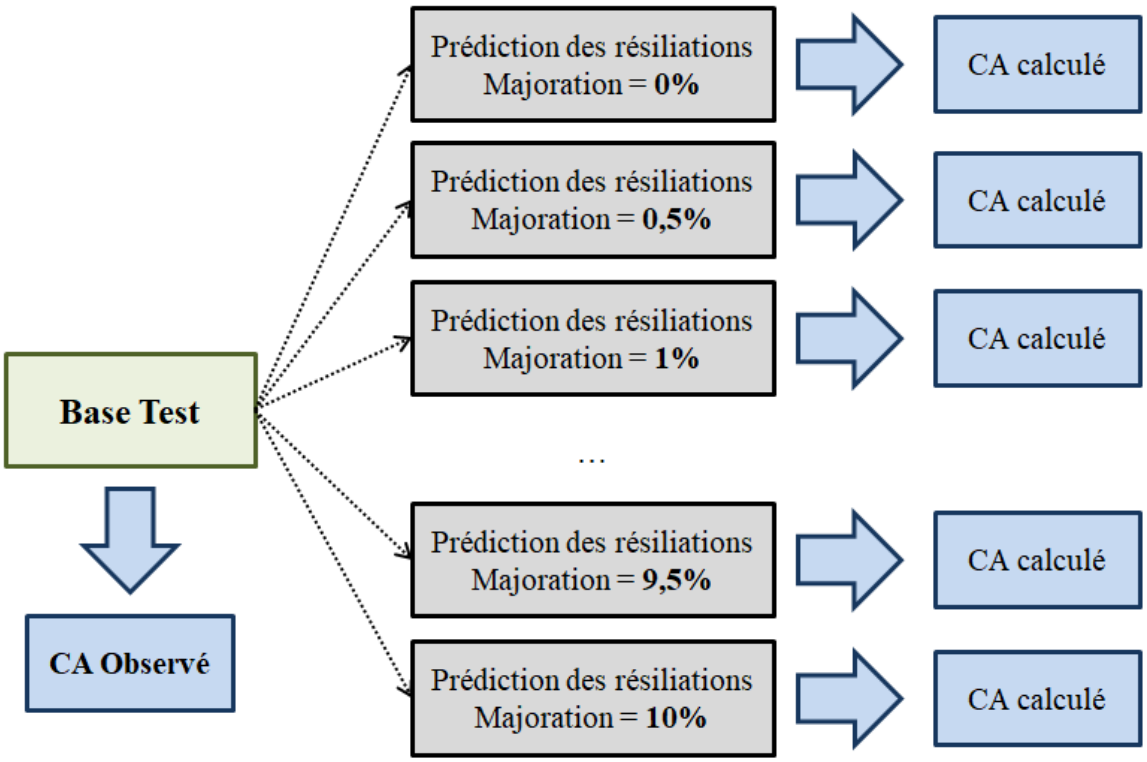

FIGURE 9.1: Schéma du calcul du chiffre d'affaires en fonction des majorations tarifaires fixées

9.2 Sensibilité du chiffre d'affaires

9.2.1 Résultats sur l'ensemble du portefeuille

Sur le graphique ci-dessous, la courbe représente le taux de résiliation prédit par le Gradient Boosting et en histogramme le chiffre d'affaires qui en découle.

L'augmentation du taux de résiliation ne compense pas la hausse sur les cotisations. C'est-à-dire que la sensibilité du taux de résiliation à l'augmentation de la majoration tarifaire n'est pas assez forte pour compenser la hausse du chiffre d'affaires dûe à l'augmentation des cotisations. Le modèle nous pousse alors à appliquer des revalorisations tarifaires très élevées afin d'augmenter le chiffre d'affaires MRH.

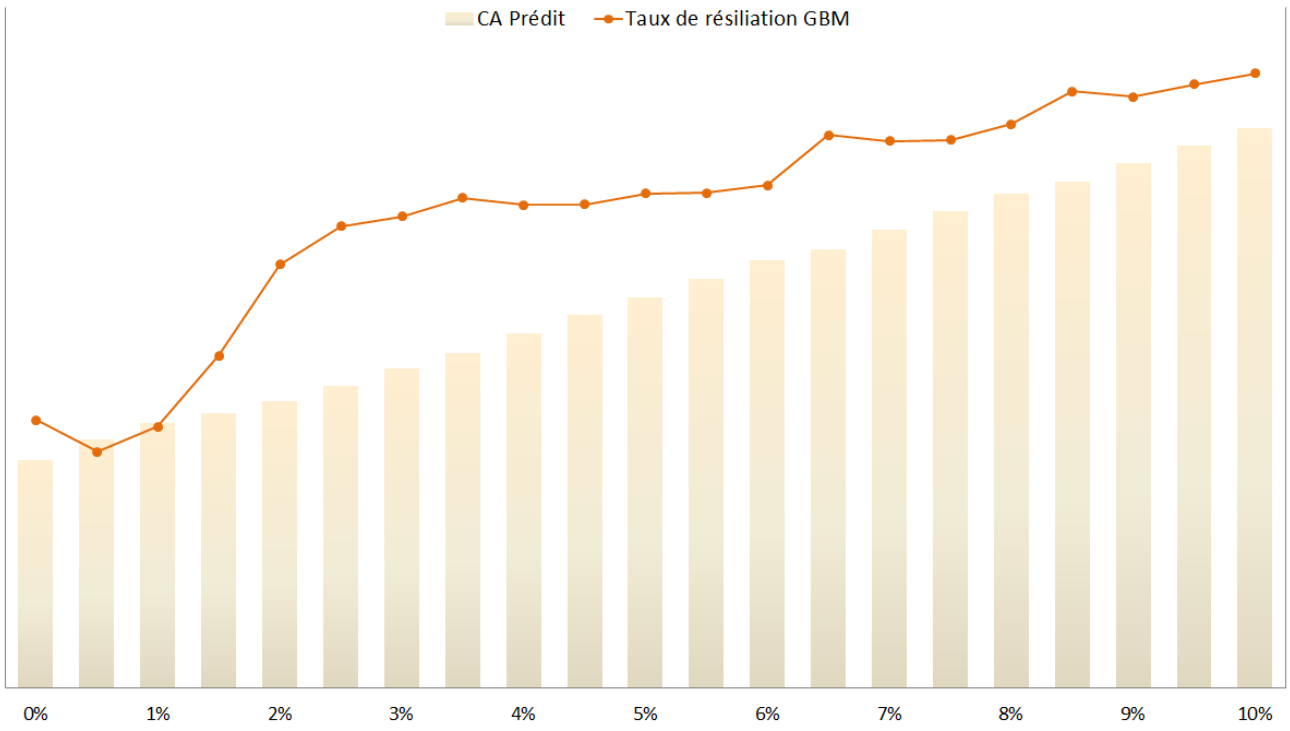

FIGURE 9.2: Taux de résiliation et chiffres d'affaires prédits à l'aide GBM

Afin de challenger le modèle de Gradient Boosting, un Modèle Linéaire Généralisé a été construit sur la même base. Le taux de résiliation prédit par ce dernier est légèrement inférieur. De plus, à l'inverse du GBM, le GLM ne capte aucune variation de sensibilité entre les différentes plages de majorations. Le GBM apporte alors une connaissance bien plus fine du comportement de nos assurés vis à vis des majorations, notamment sur la plage $[1\%; 2.5\%]$ où la sensibilité est bien plus forte.

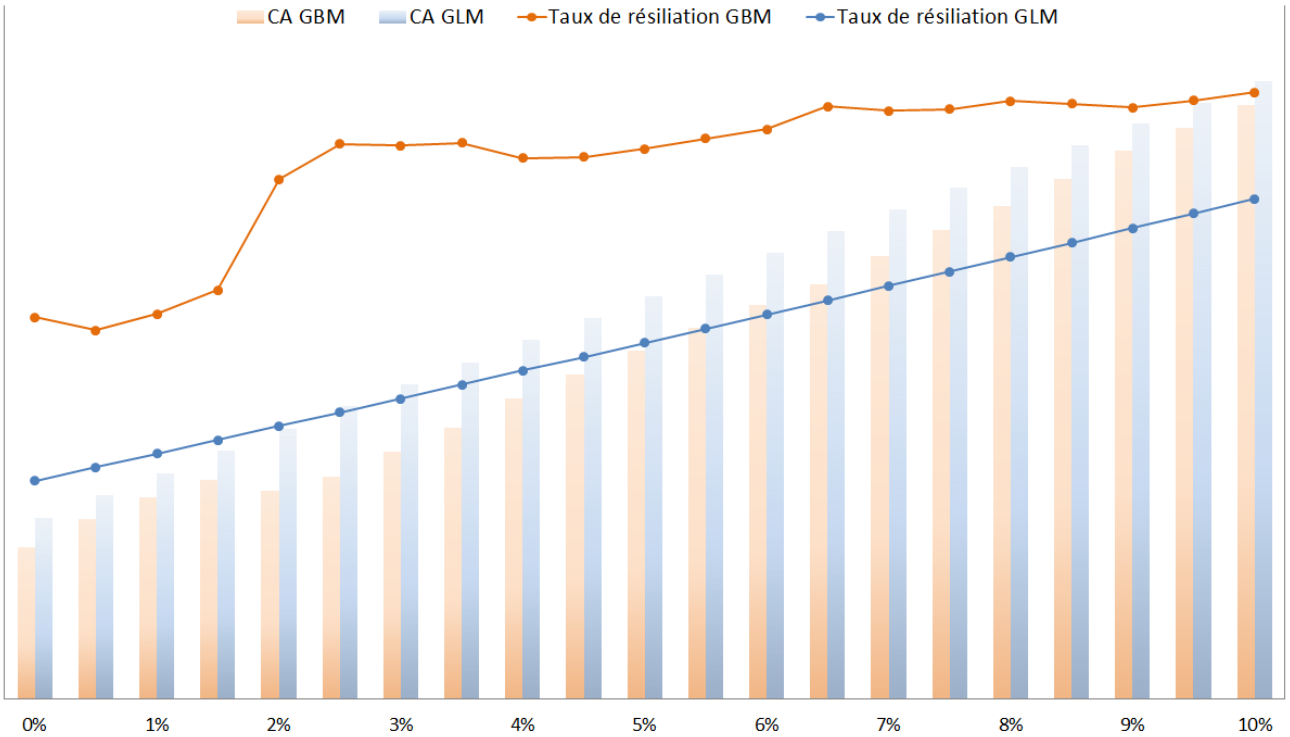

FIGURE 9.3: Taux de résiliation et chiffres d'affaires prédits à l'aide GBM vs GLM

9.2.2 Résultats sur l'ensemble du portefeuille

A noter, la seule plage de majorations qui comporte de l'intérêt est la plage $[1\%; 2.5\%]$. Sur cette plage le chiffre d'affaires ne croît que très peu avec le GBM. Dans la seconde section nous allons alors zoomer sur le comportement des différentes classes construites lors du chapitre 8.

- **Maison Propriétaire** : représentée en bleu marine
- **Maison Locataire** : représentée en bleu ciel
- **Appartement Propriétaire** : représentée en rouge
- **Appartement Locataire** : représentée en orange

Le graphique ci-dessous montre l'évolution du chiffre d'affaires pour chacune des classes. Les locataires ont un comportement spécifique sur la plage $[1\%; 2.5\%]$ tandis que l'évolution des cotisations totales pour les propriétaires est proche d'être linéaire.

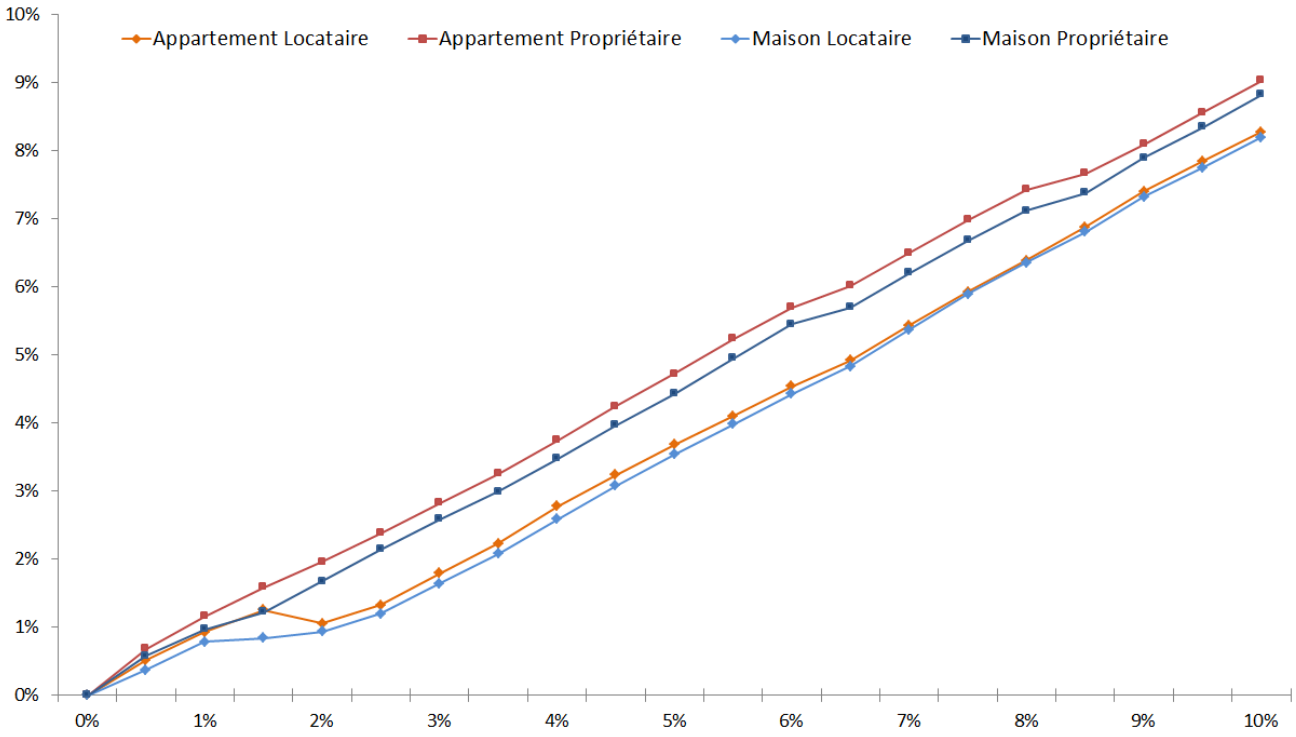

FIGURE 9.4: Évolution du chiffre d'affaires

Les résultats du dernier graphique montrent une nouvelle fois la supériorité du Gradient Boosting pour ce type d'analyse. Précédemment, nous avons remarqué que le GBM est très complexe à utiliser opérationnellement pour des travaux de tarification par exemple, néanmoins, pour des exercices d'analyses du comportement des résiliations ce dernier se révèle particulièrement intéressant. Le GBM capte ces différences de comportement entre les divers profils du portefeuille, en particulier ici sur les assurés en location.

Toujours sur cette plage $[1\%; 2.5\%]$ et sur cette population, la hausse des résiliations compense celle du chiffre d'affaires. Localement, des arbitrages peuvent être réalisés pour les appartements en location. Sur cette plage un maximum local est visible pour une majoration tarifaire égale à 1.5%.

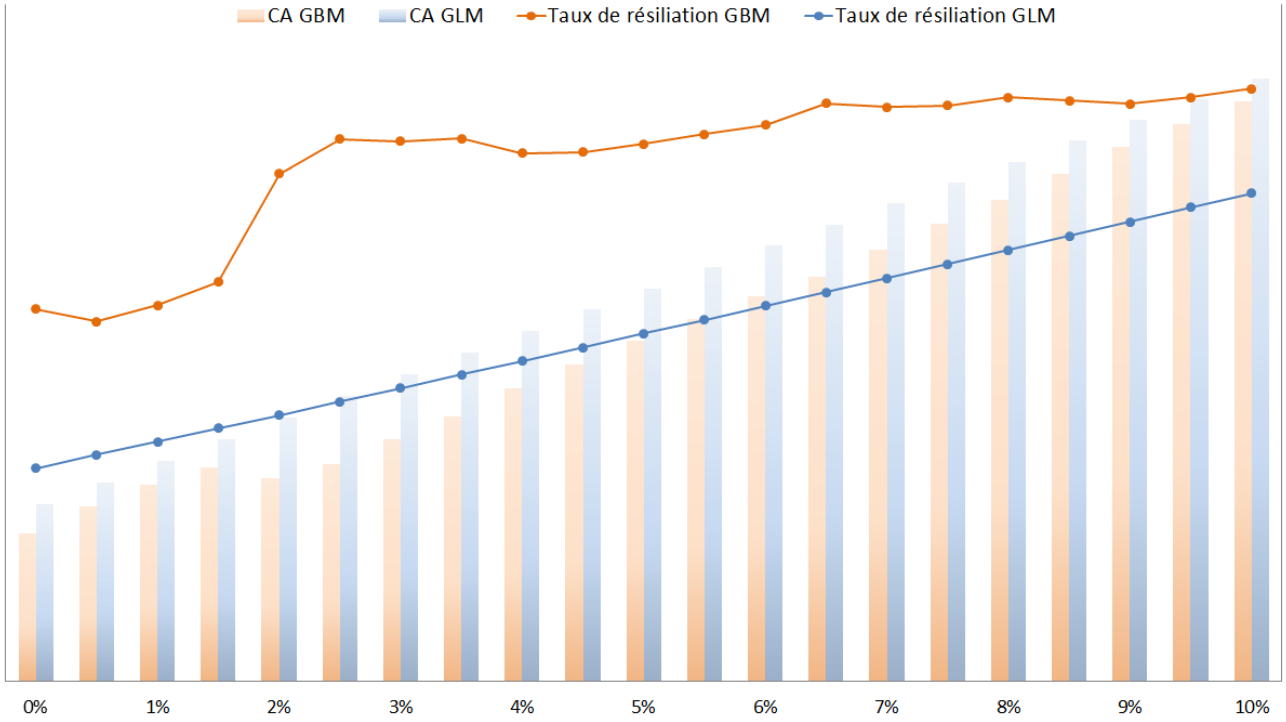

FIGURE 9.5: Taux de résiliation et chiffres d'affaires prédit à l'aide GBM vs GLM sur la classe Appartement Locataire

Les résultats des autres populations ne sont pas présentés, mais les comportements des maisons en location est similaire à celui des appartements en location. Les propriétaires quant à eux ont très peu de particularité, et le taux de résiliation évolue de manière plus linéaire que pour les locataires.

9.3 Bilan de la modélisation

9.3.1 Bilan du modèle de prédiction

Le modèle de prédiction des résiliations est très performant. En effet, celui-ci arrive à obtenir des scores en terme d'AUC très satisfaisants. Il est rare de trouver des modèles opérationnels capable d'atteindre ce score de 77.1% d'AUC.

Cette performance a permis d'apporter de nombreux éléments de compréhension au service tarification MRH sur les profils susceptibles de résilier. Nous avons approfondi l'étude sur plusieurs variables explicatives, notamment les plus importantes et influentes selon le modèle, comme la génération du contrat. D'autres variables suscitent beaucoup d'intérêt opérationnellement comme le multi-équipement des assurés qui selon les SHAP values est la 4ème caractéristique la plus impactante (cf. figure 7.19)

Toutefois et à notre plus grande déception, la variable d'évolution tarifaire n'est pas une variable clé selon le modèle de Gradient Boosting. Cette révélation va alors fortement impacter notre étude sur la sensibilité au prix des résiliations.

Plusieurs changements du modèle de prédiction sont possibles afin d'améliorer sa capacité à expliquer le comportement des différents profils vis à vis des évolutions tarifaires.

— **Majorations en pourcentage et en euros** : Au début de notre étude, nous avons choisi

d'utiliser la majoration en pourcentage de la prime comme variable. Cependant, la question d'utiliser la revalorisation en euros s'est posée. Les résultats des deux modèles sont proches en terme d'AUC, mais nous n'avons jamais été au bout du processus pour tester la sensibilité avec la variables explicative d'évolution tarifaire en euros. De surcroît, le croisement du pourcentage et du montant en euros de la majoration n'a pas été exploité.

— **Ajouter une notion de temporalité** : Avec la nouvelle loi Hamon, l'assuré n'est plus dans l'obligation d'attendre le terme pour résilier. Il peut alors résilier à tout moment entre deux échéances après la première année donc après la première majoration. Une notion de temps entre la dernière modification de la cotisation et la résiliation peut être introduite. Cette temporalité pourra être exploitée en identifiant les résiliations susceptibles d'être liées à la revalorisation.

— **Introduire des facteurs externes** : La sensibilité au prix de la résiliation n'est pas seulement une question de tarif brut mais est également une question de tarif relatif. L'ajout du contexte concurrentielle au niveau tarifaire peut être judicieux. Seulement, l'information du prix de notre assuré vis à vis de la concurrence est extrêmement difficile à obtenir.

9.3.2 Bilan de l'optimisation du chiffre d'affaires

Les résultats de la modélisation du chiffre d'affaires sont assez contrastés. En effet, le modèle de résiliation ne considère pas la variable explicative d'évolution tarifaire comme une caractéristique clé dans la prédiction de la résiliation. Le chapitre 7.4 qui traite de l'interprétabilité du modèle, illustre plusieurs fois que le lien entre le pourcentage de majoration et le taux de résiliation n'est pas aussi important que nous l'imaginions. Lorsque nous étudions le comportement des résiliations et du chiffre d'affaires en fonction de cette variable les résultats ne fluctuent alors que très peu.

Le premier graphique de ce chapitre (cf. figure 9.2) illustre parfaitement ce problème : Aucun optimum du chiffre d'affaires n'est visible sur la plage des majorations étudiées. Le chiffre d'affaires ne cesse de croître pour des revalorisations allant de 0% à 10%.

En plus de ce problème, le cadre de cette étude sur le chiffres d'affaires est trop simpliste pour que l'étude soit appliquée opérationnellement.

— **Complexification du cadre de l'étude** : Dans notre exemple simpliste, nous avons simplement étudié un portefeuille en autarcie. C'est-à-dire, que nous avons simulé les sorties du portefeuille, à savoir le nombre de résiliations, mais aucune autre notion n'a été exploitée.
  — Introduire un modèle d'affaires nouvelles qui serait également sensible aux revalorisations. Le modèle aura la tâche inverse du modèle de résiliation, autrement dit, à faire entrer des assurés en portefeuille et prédire leur profil.
  — Introduire une notion de rentabilité. Cela nécessiterait de construire un modèle de prime pure afin de prédire la sinistralité du portefeuille et la confronter aux cotisations acquises.

— **Est que l'optimisation sur le seul critère du chiffre d'affaires est judicieux ?** Le chiffre d'affaires est un élément essentiel dans le pilotage d'un portefeuille, cependant d'autres notions comme la rentabilité et l'importance du client sont centrales. L'optimisation doit alors être réalisée sur plusieurs critères.
  — La rentabilité est un élément primordial car sans elle la compagnie ne peut survivre. Elle a

l'atout d'être un élément très facilement quantifiable donc facile à intégrer dans l'optimisation.

— Le concept d'importance client est plus nébuleux. A la vision MRH, un propriétaire est plus important qu'un locataire, car le propriétaire a de grandes chances de rester en portefeuille et possède généralement une cotisation élevée. De plus, aux yeux de la compagnie certains clients sont plus précieux, par exemple ceux qui possèdent plusieurs contrats sur différents périmètres comme l'Auto ou les Pros ce qui apporte un chiffre d'affaires important à la compagnie au global. D'autres facteurs peuvent également être pris en compte pour mesurer cette importance client et alors bâtir un score à optimiser.

L'ajout de ces briques permettrait de simuler finement le pilotage d'un portefeuille. Néanmoins cela nécessite tout de même la construction de plusieurs modèles et de réfléchir à la façon de les synchroniser pour aider au pilotage.

Conclusion

Le premier objectif du modèle de résiliation est d'identifier les profils qui ont une forte tendance à rompre leur contrat MRH. En effet, les résiliations de ces contrats n'ont jamais fait l'objet d'une étude approfondie, pourtant, avec le taux d'affaires nouvelles, c'est l'indicateur de production le plus suivi par le service Tarification et Pilotage MRH.

Les statistiques descriptives révèlent de forts écarts de taux de résiliations entre différentes populations du portefeuille. Par exemple, l'écart observé entre les propriétaires et les locataires est bien plus important que l'écart entre les maisons et appartements. Néanmoins ces informations peuvent comporter des biais et donc altérer notre jugement. Afin de conduire une analyse plus poussée, nous avons eu recourt à un modèle afin d'observer les effets de chacune des variables indépendamment des autres.

La modélisation sous le logiciel SAS est limitée aux Modèles Linéaires Généralisés. Afin d'étendre le choix du type de modélisation et de sorties, la modélisation est réalisée sur le DataLab de Covéa. Ce DataLab offre un compromis parfait entre la puissance de calcul nécessaire et l'agilité des langages comme Python ou R. Ce choix s'est avéré payant, car la modélisation par Gradient Boosting offre des performances supérieures aux GLM.

Les sorties du GBM peuvent souffrir de leur manque d'interprétabilité, cependant les graphiques de dépendance partielle et les valeurs de Shapley sont deux outils qui proposent des alternatives très satisfaisantes. Les PDP dévoile l'effet marginal pour chaque modalité, de chaque variable explicative, à savoir son influence réelle sur le taux de résiliations, décorrélée donc des effets structures. Cette information est cruciale pour utiliser les résultats opérationnellement. L'utilisation des méthodes de Machine Learning plus performantes que les modélisations classiques offre un modèle de meilleure qualité et donc, une meilleure connaissance du profil des résiliations.

Afin de fidéliser le portefeuille, des actions peuvent être misent en place. Néanmoins, sans ce calcul des effets marginaux, il est difficile de mettre en production des politiques de fidélisation si nous sommes incapable de mesurer leur impact à terme. A l'aide du modèle et des interprétations, nous arrivons alors à quantifier le gain de résiliations sur le portefeuille et donc projeter avec plus de fiabilité les effets sur le portefeuille.

Toutefois, le modèle de résiliations est perfectible. Les performances sont très satisfaisantes mais les variables explicatives les plus importantes ne sont pas des éléments que l'assureur peut modifier et optimiser. Les résultats seuls offrent alors peu de marge de manoeuvre mise à part sur certaines variables comme le multi-équipement client. Cette caractéristique influe significativement le taux de résiliation et offre une marge de manoeuvre pour l'assureur.

En combinant les résultats d'un modèle d'affaires perdues et d'affaires nouvelles, nous pourrions affiner le ciblage des profils à entrer en portefeuille, c'est-à-dire, ceux avec un faible coût d'entrée et qui restent fidèles sur le long terme.

La seconde partie de l'étude s'attarde sur les majorations tarifaires. Le modèle de majoration au terme de MMA est complexe car il se décompose en deux niveaux : le premier qui consiste à calculer un tarif barème et le second qui permet d'encadrer ce même tarif. Cette complexité permet alors de segmenter finement le portefeuille pour protéger certaines populations. Trois facteurs influent grandement le processus de revalorisation et sa segmentation :

— L'optimisation de la rentabilité : Une mauvaise rentabilité du portefeuille ou d'une partie de celui-ci entraîne généralement une hausse plus importante des cotisations sur le segment en difficulté.
— L'optimisation de la croissance du portefeuille : En cas de recherche de croissance du portefeuille, la baisse des cotisations est un levier fréquemment utilisé.
— L'importance client : Certains clients sont jugés primordiaux pour MMA et bénéficient d'avantages, notamment tarifaire.

La rentabilité et l'importance des clients sont deux sujets très suivis au sein de la marque MMA. Cependant, aucune étude de sensibilité au prix n'est menée pour aider à la croissance du portefeuille. Actuellement, le choix des populations favorisées par le processus se base uniquement sur l'expertise des équipes, mais aucun modèle n'aide à la décision. L'étude de la sensibilité des résiliations au prix permet de confirmer ou réfuter ces actions de protection mises en place.

Néanmoins, l'étude de la sensibilité des résiliations aux évolutions tarifaires ne s'est pas montrée très concluante. Nous avons pu identifier si certains profils sont plus sensibles aux majorations que d'autres, seulement, au global la sensibilité au prix mesurée par le modèle n'est pas élevée. Le dernier chapitre illustre parfaitement cette difficulté, où l'on remarque que quelque soit l'augmentation de la cotisation, la hausse des résiliations n'est pas assez importante et entraîne alors toujours une augmentation du chiffre d'affaires. Différentes pistes ont été évoquées dans le but d'améliorer la capture de la sensibilité au prix.

Personnellement, je pense que les études sur le sujet des résiliations sont souvent trop orientées chiffres et pas assez client. Dans nos bases de données le motif de résiliation n'est pas assez fiable pour pouvoir être utilisé dans une modélisation. Néanmoins, posséder cette donnée et l'exploiter serait très intéressant. En effet, nous n'observons pas une explosion de résiliations dans les jours suivants le terme du contrat, donc à la suite d'une majoration tarifaire. La hausse tarifaire subit par l'assuré ne provoque pas de vague massive de résiliations.

D'autres facteurs influent alors sur les résiliations comme des faits de vie. Un assuré qui change de lieu de résidence a sûrement plus de chances de réaliser des devis chez plusieurs autres assureurs afin de comparer les prix et garanties proposés. Améliorer l'accompagnement lors de ce changement peut être un piste à étudier.

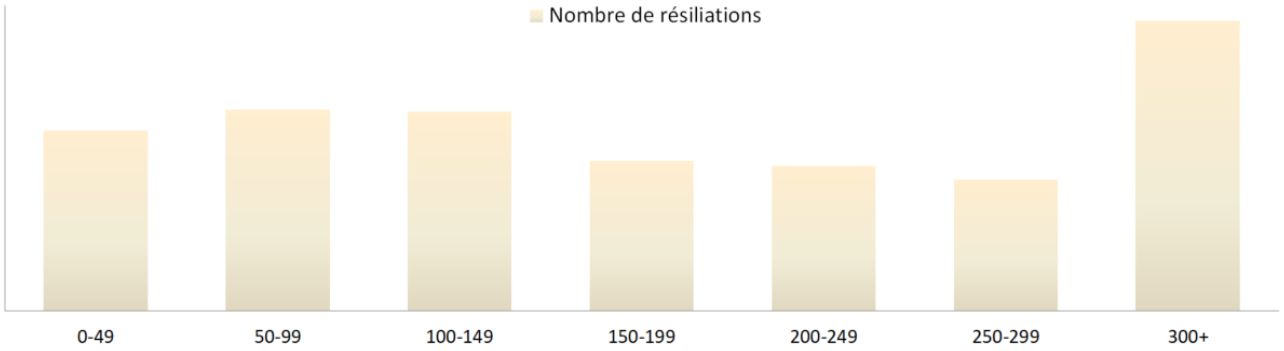

FIGURE 9.6: Nombre de résiliations en fonction du nombre de jours écoulés depuis le terme

Bibliographie

Documents internes (2020-2021). Covéa - MMA.

MOLNAR, Christoph (2022). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*.

DUTANG, Christophe (2011). *Regression models of price elasticity in non-life insurance*. Institut des Actuaires.

ADICEOM, Clara (2019). *Optimisation de la stratégie de majoration des primes de contrats d'assurance habitation au terme*. Institut des Actuaires.

COUILLAUX, Alexandre (2019). *Sensibilité au prix des contrats Habitation et optimisation de la stratégie de renouvellement*. Institut des Actuaires.

NAKACHE Jean-Pierre, CONFAIS Josiane (2003). *Statistique explicative appliquée*. Editions TECHNIP.

FFA - Les garanties du contrat multirisques habitation (2019). URL : `https://www.ffa-assurance.fr/infos-assures/les-garanties-du-contrat-multirisques-habitation`.

FFA - Résiliation du contrat (2021). URL : `https://www.ffa-assurance.fr/infos-assures/comment-resilier-votre-contrat-assurance`.

Argus de l'assurance (2020-2022). URL : `https://www.argusdelassurance.com/`.

GitHub (2008-2022). URL : `https://github.com/`.

Freakonometrics (2007-2022). URL : `https://freakonometrics.hypotheses.org/`.

Pratique de la Régression Logistique (2017). URL : `http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf`.

Regularized regression (2017). URL : https://eric.univ-lyon2.fr/~ricco/cours/slides/regularized_regression.pdf.

Ridge and Lasso Regression (2018). URL : https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b.

Gradient Boosting Trees for Classification (2020). URL : https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea.

Consistent Individualized Feature Attribution for Tree Ensembles (2018). URL : https://arxiv.org/pdf/1802.03888.pdf.

Explain Your Model with the SHAP Values (2019). URL : https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d.

Table des figures

Annexe A

Graphiques descriptifs

A.1 Statistiques descriptives supplémentaires

Variable Détention : Cette variable renseigne sur le multi-équipement du client. Plus précisément, si celui-ci possède des contrats autres que MRH au sein de MMA. Le multi-équipement est généralement un levier important dans les campagnes marketing et dans le système de surveillance. Aujourd'hui près de 3/4 du portefeuille MRH est multi-équipé au sein de MMA.

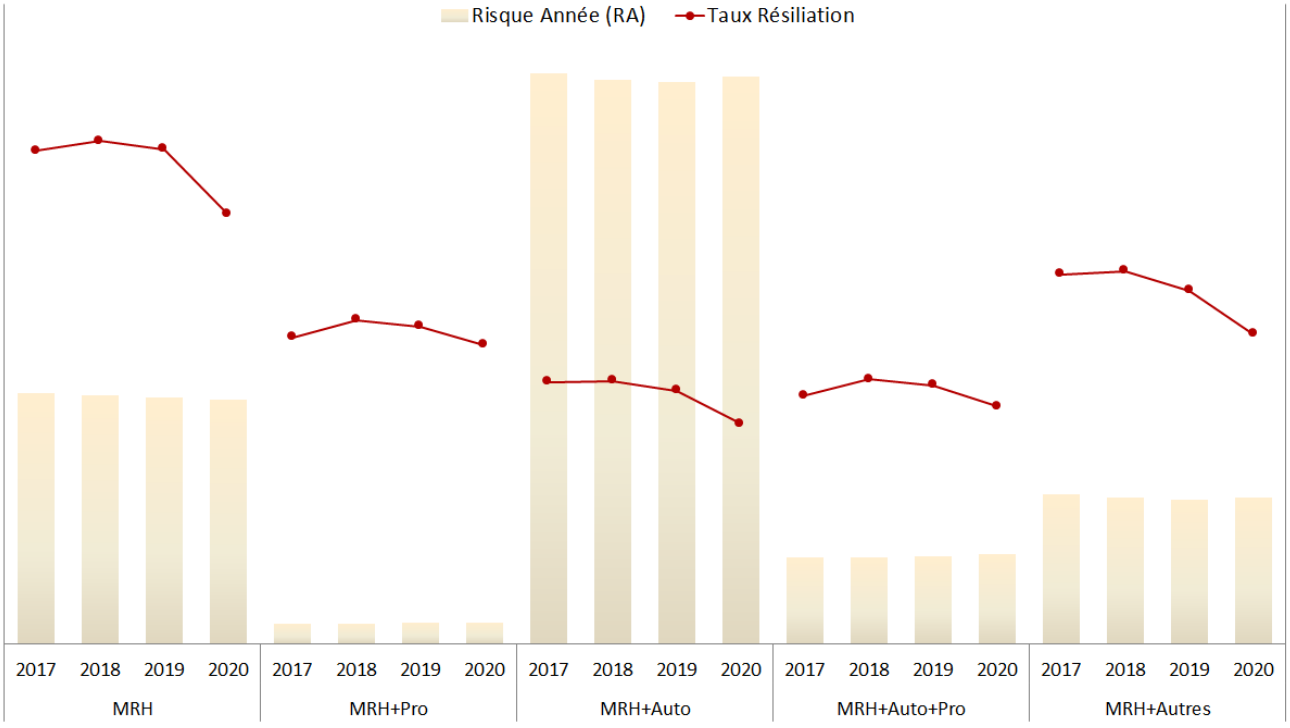

FIGURE A.1: Taux de résiliation en fonction du multi-équipement

Génération du contrat : Cette variable renseigne sur l'ancienneté du contrat MRH du client. C'est un âge calculé en année qui mesure le temps passé entre le premier contrat souscrit par l'assuré et l'année de situation étudiée. Les contrats de 30 ans et plus ont été regroupés au sein d'une même modalité.

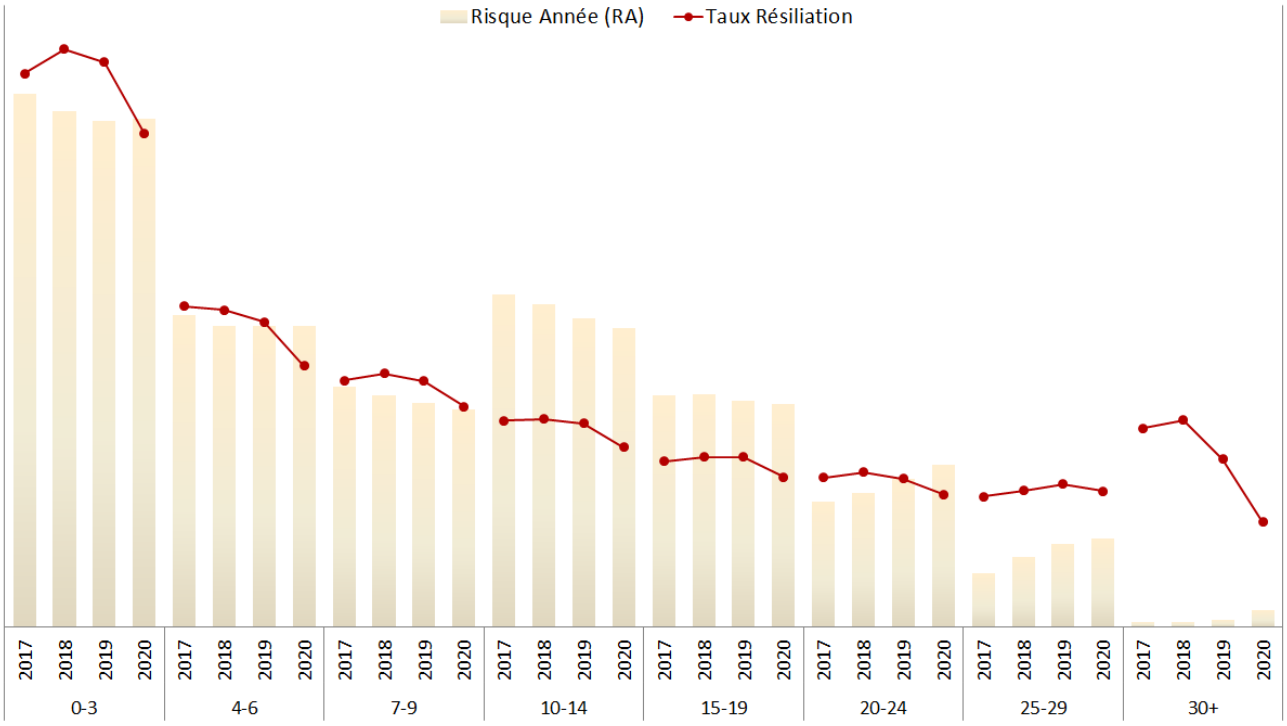

FIGURE A.2: Taux de résiliation en fonction de l'ancienneté du contrat

En regardant l'exposition au risque, nous remarquons que le nombre de contrats de plus de 20 ans a tendance à croître depuis 4 années. Le portefeuille est légèrement vieillissant, c'est-à-dire que l'âge moyen des assurés a également tendance à croître.

Concernant le taux de résiliation, un contrat plus ancien à moins de chance de résilier qu'un récent. Le taux de résiliation est assez important sur les contrats très récents. Dans cette catégorie se trouve les profils les plus opportunistes. Ces profils ont tendance à changer fréquemment d'assureur pour bénéficier d'offres et d'avantages de bienvenue.

A première vue, une génération de contrat importante protège de la résiliation et nous pousserait à discriminer les anciens contrats, néanmoins lors de l'étude de prime pure menée au début de l'alternance nous avons observé que les clients les plus anciens sont également les moins risqués. Trop majorer les clients les plus anciens pourrait créer un phénomène d'anti-sélection.

Variable SEGM PART : Cette variable a été récemment créée par la Direction Marketing de MMA à partir de la tranche d'âge du client et de son unité urbaine : Rurale (RU) / Urbain (UR) / Ultra-Urbain (UU) / Tout confondu (TT).

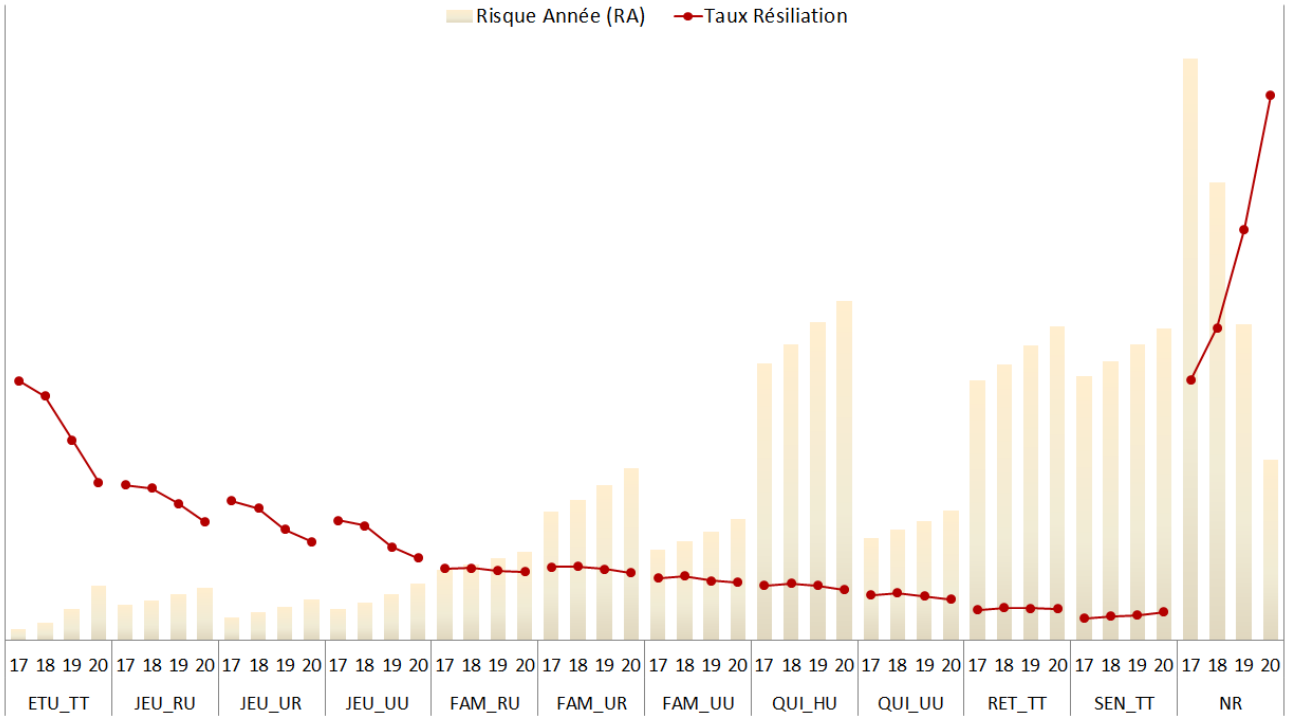

FIGURE A.3: Taux de résiliation en fonction du segment du client

Le premier constat frappant est la nette diminution du taux de résiliation en fonction des tranches d'âge du client. En moyenne, un jeune à un taux de résiliation bien plus élevé qu'un senior. Le second constat est l'observation du taux de résiliation moins important pour les biens situés dans une zone urbaine.

Néanmoins, une part assez importante des données est non renseignée. Sur les 6.5 millions de lignes de la base 1.45 millions ne sont pas renseignées, soit environ 22% de la base. De plus, le taux de résiliation de cette modalité NR est anormalement élevé en comparaison des autres modalités. Un point d'attention sur la qualité de cette donnée est à relever.

A.2 Les critères aggravants

Lors des épisodes tarifaires, plusieurs critères sont jugés aggravants et subissent une majoration tarifaire plus importante. Un des objectif de cette étude est également de déterminer si les profils possédant ces critères ne souffrent pas d'un taux de résiliation trop important. Pour étudier plus en détail le taux de résiliation, nous avons ajouté un 3ème niveau à l'axe des abscisses pour distinguer les maisons des appartements.

Critère dépendance : Ce critère renseigne sur la présence d'une dépendance ainsi que de sa taille.

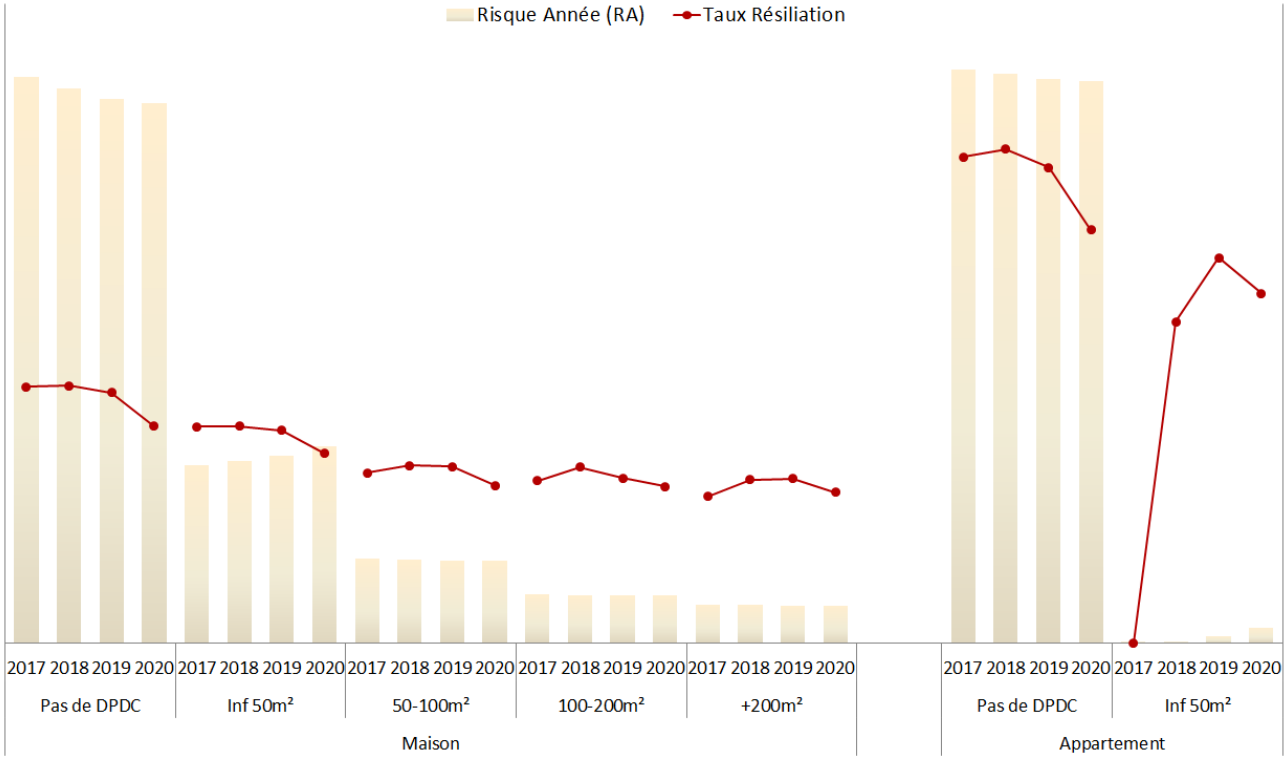

FIGURE A.4: Taux de résiliation en fonction du critère dépendance

Le taux de résiliation a tendance à diminuer en fonction de la taille de la dépendance. Néanmoins, cette diminution est logiquement liée à d'autres critères comme la taille du bâtiment principal, le montant de la cotisation et la qualité juridique de l'occupant.

L'objectif des mesures est de diminuer légèrement l'exposition au risque de ces critères sans toutefois décupler le taux de résiliation. Le problème de ce critère est que malgré les majorations, l'exposition au risque des biens avec dépendance augmente.

Critère nombre de pièces supérieures à 40m² : Ce critère renseigne sur le nombre de pièces de plus de 40m² présent au sein de l'habitation. Lors des dernières études, les habitations avec de grandes pièces possédaient des résultats dégradés.

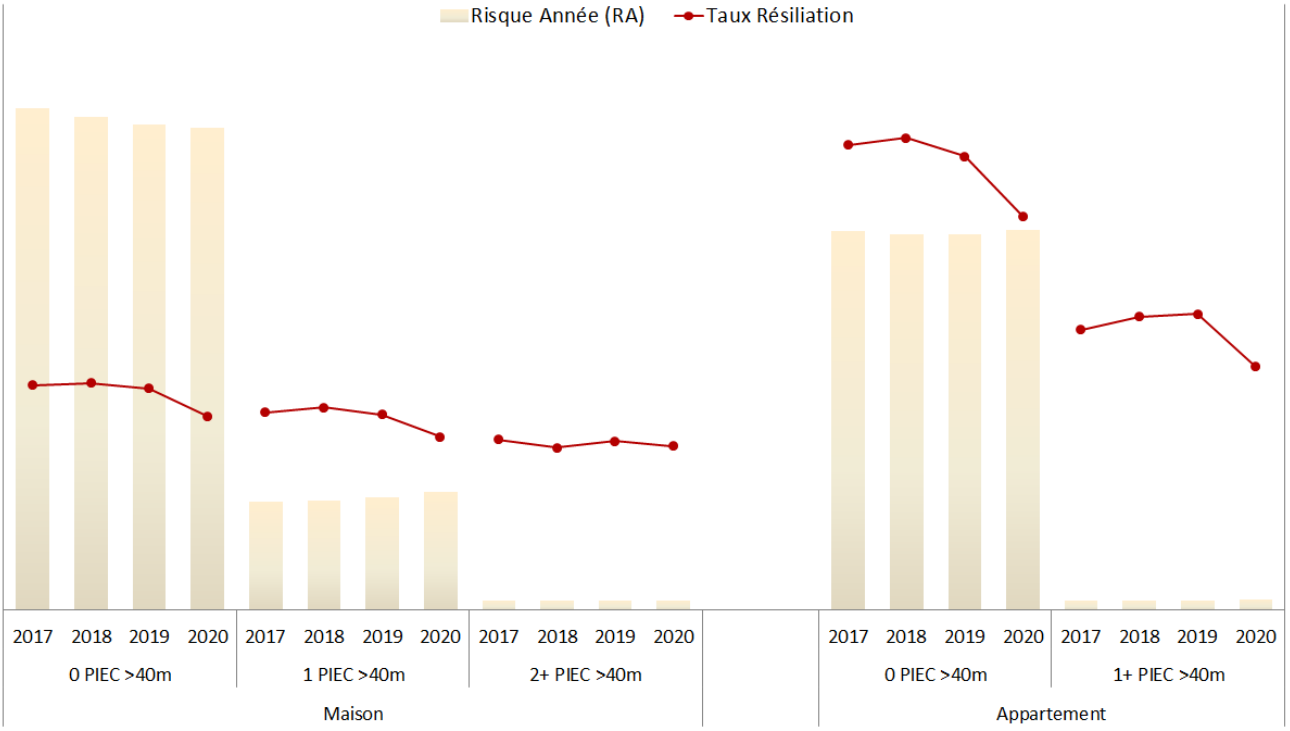

FIGURE A.5: Taux de résiliation en fonction du critère nombre de pièces de plus de 40m²

Le constat est le même que pour le critère dépendance, le taux de résiliation diminue en fonction du nombre de grandes pièces présent. Mais cette baisse est corrélée à d'autres critères. De plus, la part des maisons avec 1 pièce de plus de 40m² a tendance à croître depuis 4 ans.

Critère isolement : Ce critère permet de repérer les habitations isolées des autres. La variable prend la modalité 1 lorsqu'aucune autre habitation se trouve à moins de 50m de celle-ci. Une habitation isolée a plus de chances d'être la cible de cambriolages et donc est potentiellement plus risqué. Depuis quelques années des mesures spécifiques ont été mises en place afin d'augmenter la cotisation de ces habitations isolées.

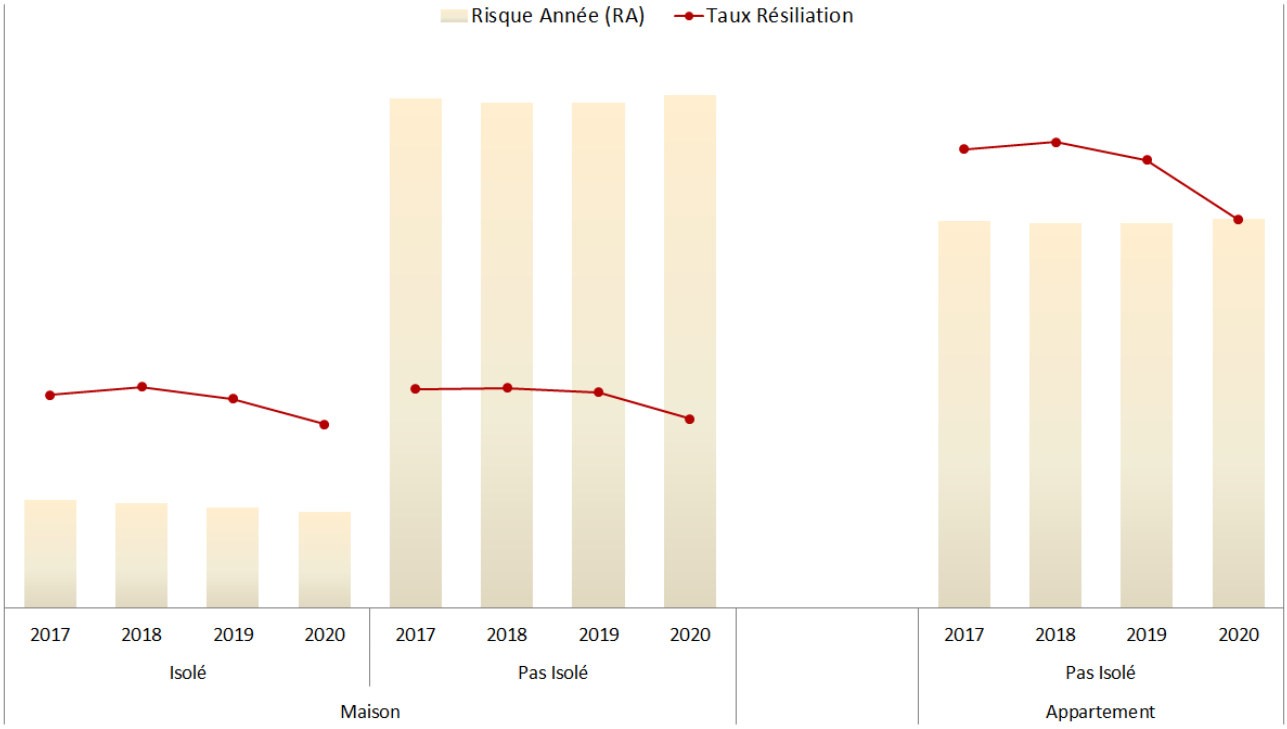

FIGURE A.6: Taux de résiliation en fonction du critère isolement

En segmentant les maisons des appartements, nous remarquons que les maisons isolées n'ont pas un taux de résiliation plus important que les non isolées. Les majorations supplémentaires n'ont pas un impact significatif sur le taux de résiliation. Cette information est une bonne nouvelle car nous avons pu rééquilibrer nos résultats sur cette modalité sans impacter significativement les résiliations. Nous pouvons également noter que l'exposition sur le critère isolement diminue légèrement depuis 4 ans.

Annexe B

Matrice de corrélation

	flag_resil	CD_TYPExQLTE	CD_USAG_RISQ	evol_levi_n	flag_surtarif	NATU_SITU	BC_ENVI_COMM	CD_ISOL	FORMULE	CD_SEGM_PART_AGG	nb_sin_n1	nb_sin_4ans	nb_2km	nb_30km	DETENTION_n	MAJO_SINI	NB_PIEC_SUPE	NB_PIEC	Taille_DPDC	CAPI_MOBI	cot_net_n
flag_resil	1.0	-0.1	0.0	-0.0	0.0	-0.0	0.0	-0.0	-0.1	-0.3	-0.0	-0.0	0.0	0.0	-0.0	0.0	-0.0	-0.1	0.1	0.1	-0.1
CD_TYPExQLTE	-0.1	1.0	0.1	-0.0	-0.1	0.1	-0.0	0.3	0.2	0.2	0.0	0.0	-0.4	-0.3	0.1	-0.1	0.3	0.6	-0.4	-0.1	0.5
CD_USAG_RISQ	0.0	0.1	1.0	0.0	0.0	0.0	-0.0	0.1	-0.3	-0.0	-0.0	-0.0	-0.1	-0.1	-0.0	0.0	-0.0	-0.1	-0.0	0.2	-0.0
evol_levi_n	-0.0	-0.0	0.0	1.0	0.0	0.0	-0.1	-0.0	-0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0
flag_surtarif	0.0	-0.1	0.0	0.0	1.0	0.3	-0.2	-0.2	-0.1	0.0	-0.0	-0.0	0.1	0.0	-0.0	-0.0	-0.1	-0.1	0.1	0.0	-0.1
NATU_SITU	-0.0	0.1	0.0	0.0	0.3	1.0	-0.1	0.0	0.0	0.2	0.0	0.0	-0.0	0.0	0.1	-0.4	0.0	0.1	-0.0	-0.1	0.1
BC_ENVI_COMM	0.0	-0.0	-0.0	-0.1	-0.2	-0.1	1.0	0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.1	-0.0	0.1	-0.0	-0.0	0.1	-0.1
CD_ISOL	-0.0	0.3	0.1	-0.0	-0.2	0.0	0.0	1.0	0.0	0.0	0.0	0.0	-0.2	-0.1	0.0	-0.0	0.1	0.1	-0.1	-0.0	0.2
FORMULE	-0.1	0.2	-0.3	-0.0	-0.1	0.0	0.0	0.0	1.0	0.1	0.0	0.0	-0.1	-0.2	0.1	-0.1	0.1	0.3	-0.1	-0.1	0.3
CD_SEGM_PART_AGG	-0.3	0.2	-0.0	0.0	0.0	0.2	-0.0	0.0	0.1	1.0	0.0	0.0	-0.1	-0.0	0.1	-0.1	0.0	0.2	-0.1	-0.1	0.2
nb_sin_n1	-0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.5	-0.0	0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0
nb_sin_4ans	-0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	-0.0	0.0	0.0	-0.0	0.0	0.1	-0.0	-0.0	0.1
nb_2km	0.0	-0.4	-0.1	0.0	0.1	-0.0	-0.0	-0.2	-0.1	-0.1	-0.0	-0.0	1.0	0.4	-0.0	0.0	-0.1	-0.3	0.2	0.1	-0.2
nb_30km	0.0	-0.3	-0.1	0.0	0.0	0.0	-0.0	-0.1	-0.2	-0.0	0.0	0.0	0.4	1.0	-0.0	-0.0	-0.1	-0.2	0.1	-0.0	-0.0
DETENTION_n	-0.0	0.1	-0.0	-0.0	-0.0	0.1	0.1	0.0	0.1	0.1	0.0	0.0	-0.0	-0.0	1.0	-0.1	0.1	0.1	-0.0	-0.0	0.1
MAJO_SINI	0.0	-0.1	0.0	0.0	-0.0	-0.4	-0.0	-0.0	-0.1	-0.1	-0.0	-0.0	0.0	-0.0	-0.1	1.0	-0.1	-0.1	0.0	0.1	-0.1
NB_PIEC_SUPE	-0.0	0.3	-0.0	-0.0	-0.1	0.0	0.1	0.1	0.1	0.0	0.0	0.0	-0.1	-0.1	0.1	-0.1	1.0	0.2	-0.1	-0.1	0.4
NB_PIEC	-0.1	0.6	-0.1	-0.0	-0.1	0.1	-0.0	0.1	0.3	0.2	0.0	0.1	-0.3	-0.2	0.1	-0.1	0.2	1.0	-0.2	-0.3	0.5
Taille_DPDC	0.1	-0.4	-0.0	0.0	0.1	-0.0	-0.0	-0.1	-0.1	-0.1	-0.0	-0.0	0.2	0.1	-0.0	0.0	-0.1	-0.2	1.0	0.1	-0.2
CAPI_MOBI	0.1	-0.1	0.2	0.0	0.0	-0.1	0.1	-0.0	-0.1	-0.1	-0.0	-0.0	0.1	-0.0	-0.0	0.1	-0.1	-0.3	0.1	1.0	-0.1
cot_net_n	-0.1	0.5	-0.0	-0.0	-0.1	0.1	-0.1	0.2	0.3	0.2	0.0	0.1	-0.2	-0.0	0.1	-0.1	0.4	0.5	-0.2	-0.1	1.0
evol_net_fin_n	0.0	0.1	0.0	-0.2	-0.1	-0.0	0.0	0.1	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0	-0.2	0.1	0.0	-0.1	0.0	0.1
evol_net_fin_n1	-0.0	0.2	0.0	-0.0	-0.1	-0.0	-0.0	0.1	0.0	-0.0	-0.0	0.0	-0.1	-0.0	0.0	-0.0	0.1	0.1	-0.1	0.0	0.1
evol_net_fin_n2	-0.0	0.2	0.1	-0.0	-0.1	-0.0	-0.0	0.1	0.0	-0.0	-0.0	0.0	-0.1	-0.1	0.0	-0.0	0.1	0.1	-0.1	0.0	0.1
evol_net_fin_n3	0.0	0.1	0.0	-0.0	-0.1	-0.0	-0.0	0.0	-0.0	-0.0	0.0	0.0	-0.1	-0.0	0.0	-0.0	0.0	0.0	-0.0	0.0	0.0
VA_COEF_BM_MRH	0.0	0.1	-0.1	-0.0	-0.1	-0.0	0.0	0.0	0.1	0.0	0.0	0.1	-0.0	0.0	0.0	-0.4	0.1	0.1	-0.1	-0.0	0.2
NBP_MOY_MAISON_RP	-0.0	0.1	-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.1	0.0	-0.0	0.0	-0.2	-0.2	0.0	-0.0	0.1	0.1	-0.1	-0.0	0.1
COUT_OBS_n1	-0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	-0.0	-0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0
cout_sin_4ans	-0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	-0.0	-0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0
nb_affa_auto_n	-0.0	0.2	-0.0	-0.0	-0.1	-0.0	0.3	0.1	0.1	0.1	0.0	0.0	-0.1	-0.1	0.0	-0.0	0.1	0.1	-0.1	0.0	0.1
evol_AUTO	-0.0	-0.0	-0.0	-0.0	-0.0	-0.1	0.0	-0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0
nb_affa_mrh_n	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0
evol_MRH	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0	-0.0
nb_pro_ent_n	-0.0	0.0	0.0	-0.0	-0.1	-0.0	0.1	0.1	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.2	-0.0	0.1	0.0	-0.0	0.0	0.1
nb_option	-0.0	0.2	-0.0	-0.0	-0.1	-0.1	0.1	0.0	0.2	0.0	0.0	0.1	-0.1	-0.0	0.1	-0.0	0.2	0.2	-0.2	-0.1	0.3
Generation	-0.1	0.3	0.0	0.1	0.2	0.3	-0.3	0.0	0.2	0.3	0.0	0.0	-0.1	-0.0	0.0	-0.1	0.0	0.2	-0.1	-0.1	0.3
TARIF_B100	0.0	-0.2	-0.0	-0.0	-0.1	-0.0	-0.0	-0.1	-0.2	-0.0	0.0	0.0	0.3	0.5	0.0	-0.0	-0.0	-0.1	0.1	0.0	0.1
SCORE	0.0	-0.0	-0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.1	0.1	-0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0
EVOL_BAR_1AN	-0.0	0.2	0.0	-0.0	-0.0	0.0	-0.0	0.1	0.1	0.0	0.0	0.0	-0.1	-0.1	0.0	-0.0	0.1	0.1	-0.1	-0.0	0.1
VA_COEF_LEVI	0.0	-0.1	-0.0	-0.1	-0.2	-0.2	0.2	0.0	0.0	-0.1	0.0	0.0	0.0	-0.0	0.1	0.1	0.1	-0.0	0.0	-0.0	-0.1
MT_DRGT	0.0	-0.2	0.0	0.0	0.4	0.0	-0.1	-0.2	-0.1	-0.0	-0.0	-0.1	0.1	0.0	-0.1	0.1	-0.3	-0.2	0.2	0.0	-0.3

FIGURE B.1: Matrice complète de corrélation (1)

evol_net_fin_n	evol_net_fin_n1	evol_net_fin_n2	evol_net_fin_n3	VA_COEF_BM_MRH	NBP_MOY_MAISON_RP	COUT_OBS_n1	cout_sin_4ans	nb_affa_auto_n	evol_AUTO	nb_affa_mrh_n	evol_MRH	nb_pro_ent_n	nb_option	Generation	TARIF_B100	SCORE	EVOL_BAR_1AN	VA_COEF_LEVI	MT_DRGT
0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.1	0.0	0.0	-0.0	0.0	0.0
0.1	0.2	0.2	0.1	0.1	0.1	0.0	0.0	0.2	-0.0	-0.0	-0.0	0.0	0.2	0.3	-0.2	-0.0	0.2	-0.1	-0.2
0.0	0.0	0.1	0.0	-0.1	-0.0	0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0	-0.0	0.0	-0.0	-0.0	0.0	-0.0	0.0
-0.2	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.1	-0.0	0.0	-0.0	-0.1	0.0
-0.1	-0.1	-0.1	-0.1	-0.1	-0.0	-0.0	-0.0	-0.1	-0.0	0.0	-0.0	-0.1	-0.1	0.2	-0.1	0.0	-0.0	-0.2	0.4
-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.1	0.0	-0.0	-0.0	-0.1	0.3	-0.0	0.0	0.0	-0.2	0.0
0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.1	0.1	-0.3	-0.0	-0.0	-0.0	0.2	-0.1
0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1	-0.0	-0.0	-0.0	0.1	0.0	0.0	-0.1	-0.0	0.1	0.0	-0.2
0.0	0.0	0.0	-0.0	0.1	0.1	0.0	0.0	0.1	0.0	-0.0	-0.0	0.0	0.2	0.2	-0.2	-0.0	0.1	0.0	-0.1
-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.1	0.0	-0.0	-0.0	-0.0	0.0	0.3	-0.0	0.0	0.0	-0.1	-0.0
0.0	-0.0	-0.0	0.0	0.0	-0.0	0.1	0.1	0.0	0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0
0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.0	-0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	-0.1
-0.0	-0.1	-0.1	-0.1	-0.0	-0.2	-0.0	-0.0	-0.1	0.0	0.0	0.0	-0.0	-0.1	-0.1	0.3	0.1	-0.1	0.0	0.1
-0.0	-0.0	-0.1	-0.0	0.0	-0.2	-0.0	-0.0	-0.1	-0.0	0.0	0.0	-0.0	-0.0	-0.0	0.5	0.1	-0.1	-0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	0.0	0.0	0.2	0.1	0.0	0.0	-0.0	0.0	0.1	-0.1
-0.2	-0.0	-0.0	-0.0	-0.4	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0	-0.0	-0.0	-0.1	-0.0	-0.0	-0.0	0.1	0.1
0.1	0.1	0.1	0.0	0.1	0.1	0.0	0.0	0.1	0.0	-0.0	-0.0	0.1	0.2	0.0	-0.0	-0.0	0.1	0.1	-0.3
0.0	0.1	0.1	0.0	0.1	0.1	0.0	0.0	0.1	-0.0	-0.0	-0.0	0.0	0.2	0.2	-0.1	-0.0	0.1	-0.0	-0.2
-0.1	-0.1	-0.1	-0.0	-0.1	-0.1	-0.0	-0.0	-0.1	-0.0	0.0	0.0	-0.0	-0.2	-0.1	0.1	0.0	-0.1	0.0	0.2
0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0	-0.1	-0.1	0.0	0.0	-0.0	-0.0	0.0
0.1	0.1	0.1	0.0	0.2	0.1	0.0	0.0	0.1	-0.0	-0.0	-0.0	0.1	0.3	0.3	0.1	-0.0	0.1	-0.1	-0.3
1.0	0.1	-0.0	0.0	0.2	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	-0.1
0.1	1.0	0.1	-0.0	0.2	0.0	0.0	0.0	0.1	-0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	-0.1
-0.0	0.1	1.0	0.1	0.1	0.0	-0.0	0.0	0.1	-0.0	-0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	-0.1
0.0	-0.0	0.1	1.0	0.1	0.0	0.0	0.0	0.0	-0.0	0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	-0.1
0.2	0.2	0.1	0.1	1.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	-0.2
0.0	0.0	0.0	0.0	0.0	1.0	-0.0	0.0	0.1	0.0	-0.0	-0.0	0.0	0.0	0.0	-0.2	-0.1	0.0	0.0	-0.1
0.0	0.0	-0.0	0.0	0.0	-0.0	1.0	0.5	0.0	0.0	0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0
0.0	0.1	0.1	0.0	0.0	0.1	0.0	0.0	1.0	0.2	0.1	0.1	0.2	0.1	-0.0	-0.1	-0.0	0.0	0.1	-0.2
-0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.2	1.0	0.1	0.1	0.0	0.0	-0.1	-0.0	-0.0	-0.0	0.0	-0.0
-0.0	-0.0	-0.0	0.0	-0.0	-0.0	0.0	0.0	0.1	0.1	1.0	0.7	0.2	-0.0	0.0	0.0	0.0	-0.0	0.0	-0.0
-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	0.1	0.1	0.7	1.0	0.1	-0.0	-0.0	0.0	0.0	-0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.2	0.1	1.0	0.1	-0.1	0.0	-0.0	0.0	0.2	-0.2
0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	-0.0	-0.0	0.1	1.0	-0.1	0.0	-0.0	0.0	0.1	-0.3
-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	-0.0	-0.1	0.0	-0.0	-0.1	-0.1	1.0	-0.1	0.0	0.1	-0.2	0.0
0.0	0.0	0.0	0.0	0.0	-0.2	0.0	0.0	-0.1	-0.0	0.0	0.0	0.0	0.0	-0.1	1.0	0.1	-0.0	0.0	-0.1
-0.0	-0.0	-0.0	-0.0	0.0	-0.1	0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0	0.1	1.0	-0.0	-0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	0.1	-0.0	-0.0	1.0	-0.0	-0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.1	-0.2	0.0	-0.0	-0.0	1.0	-0.2
-0.1	-0.1	-0.1	-0.1	-0.2	-0.1	-0.0	-0.0	-0.2	-0.0	-0.0	0.0	-0.2	-0.3	0.0	-0.1	0.0	-0.0	-0.2	1.0

FIGURE B.2: Matrice complète de corrélation (2)