Institut des ACTUAIRES

ENSAE IP PARIS

Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 08/03/2023

Par : **Margaux REGNAULT**

Titre : **Modélisation des comportements de fidélité des assurés : sensibilité au prix et applications à l'optimisation tarifaire sur un portefeuille auto**

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

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

Membres présents du jury de la filière

Nom : Caroline HILLAIRET

Membres présents du jury de l'Institut des Actuaires

Entreprise : ADDACTIS France

Signature :

ADDACTIS France
Espace WeWork
4, rue Jules Lefebvre - 75009 Paris
Tél. 01 56 89 07 70
Siren 413 611 344 - NAF 7022 Z
TVA Intracommunautaire FR53 413 611 344
SAS au capital de 100.000 €

Directeur du mémoire en entreprise :

Nom : Linda KROLIKOWSKI
Signature :

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

Signature du responsable entreprise

Secrétariat :

Signature du candidat

Bibliothèque :

Regnault

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

Résumé

Le secteur de l'assurance automobile, par l'entrée de nouveaux acteurs sur le marché, le nombre de législations facilitant la résiliation et le développement de comparateurs de prix, fait face à une concurrence accrue. Dans ce contexte d'autant plus compétitif, fidéliser les clients acquis et s'assurer de cibler les prospects à rétention élevée se révèle nécessaire à l'obtention d'un portefeuille résilient. Ainsi, l'appréhension des comportements des assurés face à la résiliation est un enjeu majeur pour garantir la pérennité économique des organismes d'assurance. L'objet de ce mémoire est de proposer différents outils permettant de mieux capter les comportements des assurés et les facteurs conduisant à une rétention, ou au contraire à une résiliation.

Afin de répondre à cette problématique de rétention des assurés, deux modèles sont mis en œuvre. Dans un premier temps, la modélisation de la probabilité de résiliation à un an repose sur une analyse fine des décisions prises par l'assuré dans un laps de temps réduit. Ensuite, la volonté de disposer d'un indicateur prenant plus de recul temporel conduit à la modélisation de la durée de vie a priori des contrats.

A partir des modèles construits, plusieurs outils d'aide à la prise de décision sont soumis. Une cartographie du portefeuille est mise en place dans une démarche de segmentation des assurés en fonction de leur degré de fidélité, de leur sensibilité au prix et de leur marge. De plus, l'étude approfondie de certains profils permet la mise en place d'un scénario tarifaire, visant à augmenter le profit de l'assureur en maintenant ses autres indicateurs de rentabilité. Ce mémoire constitue une base à l'optimisation tarifaire au renouvellement d'une part, et au calcul de l'indicateur de la valeur client d'autre part.

Mots-clés : *Assurance non-vie*, *Modèle linéaire généralisé*, *Modèle à risques proportionnels de Cox*, *Résiliation*, *Durée de vie*, *Optimisation tarifaire*, *Valeur client.*

Abstract

The auto insurance industry is facing increased competition due to the entry of new players in the market, the number of legislation facilitating termination and the development of price comparisons. In this increasingly competitive environment, retaining current customers and ensuring that high retention prospects are targeted is necessary to obtain a resilient portfolio. Thus, understanding policyholders' behavior in the face of termination is a major challenge to guarantee the economic sustainability of insurance organizations. The purpose of this thesis is to propose different tools to better understand policyholder behavior and the factors leading to retention or, on the contrary, to termination.

In order to address this issue of policyholder retention, two models are proposed. First, the modeling of the probability of cancellation at one year proposes a detailed analysis of the decisions taken by the client in a reduced period of time. Secondly, the desire to have an indicator that takes more time into account leads to the modeling of the lifetime of contracts.

Based on the models built, several decision-making tools are proposed. A portfolio mapping is set up to segment policyholders according to their degree of loyalty, their price sensitivity and their profit margin. In addition, the in-depth study of certain profiles allows the implementation of a pricing scenario, aiming at increasing the insurer's profit while maintaining its other profitability indicators. This report constitutes a basis for the optimization of the premium at renewal, and for the calculation of the customer value indicator.

Keywords : *Non-life insurance*, *Generalized linear model*, *Cox Proportional-Hazards Model*, *Termination*, *Lifetime*, *Price optimization*, *Customer lifetime value.*

Remerciements

Mes premiers remerciements vont vers l'équipe Pricing & Analytics P&C d'ADDACTIS France pour son accueil bienveillant et chaleureux.

Je tiens à remercier tout particulièrement Linda KROLIKOWSKI, directrice de ce mémoire, pour la confiance accordée en m'offrant l'opportunité de réaliser ce stage. Son expertise, son engagement et son goût pour la transmission m'ont permis de conduire sereinement ce travail de recherche. Merci Linda d'être le manager qui soutient mes premiers pas dans la vie professionnelle.

Je remercie Guillaume ROSOLEK, Médérick BESARABOV et Nabil RACHDI pour leurs appréciations et conseils actuariels. Merci à Pierre CHATELAIN pour sa disponibilité, ses relectures et conseils techniques avisés.

Merci à mes collègues, Cédric DENIEL, Mathilde ROCHELLE, Thomas PAIN et Markéta KRÚPOVÁ pour la pertinence de leurs relectures. Mon dernier remerciement appuyé s'adresse à Markéta, pour nos échanges et notre complicité professionnelle.

Sommaire

Introduction

L'assurance automobile, par son caractère obligatoire et sa part de cotisation, est un acteur majeur du secteur assurantiel. Bien qu'indispensable au bon fonctionnement de la société, l'assurance reste néanmoins soumise aux contingences du marché. L'entrée de nouveaux acteurs, les évolutions réglementaires et le développement de comparateurs de tarifs ont contribué à rendre le marché de l'assurance automobile très concurrentiel. La mutualisation des risques étant au cœur du concept même d'assurance, détenir des portefeuilles larges pour permettre une résilience en cas de stress est crucial pour les assureurs. Ces derniers redoublent alors d'efforts pour proposer des offres attractives et capter de nouveaux prospects. Toujours est-il que conserver des contrats déjà acquis reste moins onéreux que de chercher à en fédérer de nouveaux. Être en mesure d'appréhender les comportements des clients face à la résiliation se révèle alors primordial à la bonne santé économique d'un assureur.

Ainsi, ce mémoire s'attache à la modélisation de ces comportements. Dans une démarche de rentabilité à court terme mais également à moyen terme, deux analyses seront menées. A court terme, l'étude de la probabilité de résiliation à un an fournit des éléments essentiels à la compréhension des décisions immédiates prises par les assurés. Pour une appréhension des comportements à plus long terme, l'analyse de la durée de vie des clients en fonction de leurs caractéristiques à la souscription est essentielle. Cette dernière permet d'une part, d'identifier les profils plus ou moins fidèles a priori et d'autre part, de disposer du recul que n'offre pas l'étude de la résiliation à un an. Ces deux éléments, que sont la probabilité de résiliation à un an et la durée de vie a priori, se complètent et fournissent des indicateurs quantitatifs du degré de fidélisation des assurés.

Au-delà de l'identification des facteurs conduisant à l'acte de résiliation, plusieurs applications seront proposées. Entre autres, une cartographie du portefeuille, en fonction du niveau de marge appliqué, de la sensibilité au prix, et du degré de fidélité des assurés est proposée. Celle-ci permet une première identification des segments sur lesquels l'assureur bénéficierait à réadapter légèrement sa politique. Ensuite, une analyse ciblée de certains profils d'assurés constituera une base à l'élaboration d'une stratégie tarifaire lors du renouvellement annuel des contrats.

Chapitre 1

Contexte, enjeux et périmètre de l'étude

Sommaire

Ce premier chapitre introduit quelques notions fondamentales de l'assurance, avant de détailler brièvement le principe de l'assurance automobile. Dans le marché concurrentiel qui est celui de l'assurance des véhicules, la stratégie de fidélisation des clients sera discutée. Cela permettra de présenter les motivations du mémoire, les axes de modélisation abordés puis de fournir quelques exemples d'applications. Finalement, le périmètre des analyses menées sera défini.

1.1 Le cadre de l'assurance

1.1.1 Généralités

La mutualisation et le transfert des risques sont des concepts qui remonteraient aux premières civilisations. Formalisée par un cadre juridique en France à partir du $XVII^e$ siècle, augmentée par les avancées mathématiques dans le domaine des probabilités puis devenue nécessaire avec les évolutions économiques et la complexification des échanges, l'assurance se positionne aujourd'hui en pilier de notre structure sociale et économique.

Bien que le Code des Assurances ne fournisse pas de définition juridique de l'assurance, cette dernière est cependant délimitée par le dictionnaire du Larousse comme étant un « contrat par lequel l'assureur s'engage à indemniser l'assuré, moyennant une prime ou une cotisation, de certains risques ou sinistres éventuels ». Autrement dit, l'assuré transfère vers l'assureur, en échange d'une prime, tout ou partie du risque qu'il supporte. Les engagements des deux parties sont formalisés par un contrat d'assurance qui stipule le risque couvert, la prestation en cas de survenance de ce risque et la prime que l'assuré verse à l'assureur. Une des caractéristiques majeures de l'assurance est celle dite de l'inversion du cycle de production. L'assureur fixe et perçoit les primes, avant de connaître le montant des sinistres à venir, et des prestations qu'il devra verser. Cela nécessite, entre autres, l'utilisation d'outils statistiques et probabilistes sophistiqués. De plus, la mise en œuvre mathématique du calcul de prime doit équilibrer deux paramètres : la mutualisation et la segmentation. La mutualisation permet de considérer un grand nombre d'individus et donc de réduire la volatilité du risque, quand la segmentation demande à réaliser des sous-groupes d'assurés, aux seins desquels les risques sont homogènes.

L'assurance est constituée de deux grandes familles. L'une est celle de l'assurance des personnes, dont l'indemnisation est généralement forfaitaire. Elle regroupe notamment les assurances en cas de décès et les assurances maladie. L'autre est celle de l'assurance de biens et de responsabilité. Elle inclut les assurances automobile, habitation, de responsabilité civile ou encore de protection juridique. Aussi appelée assurance dommage, elle donne droit à une indemnisation lors de la réalisation d'un évènement aléatoire, appelé sinistre. La Figure 1.1, issue des données publiées en 2021 par France Assureurs [24], schématise ces deux grandes catégories d'assurance et les branches qui en découlent. Sont représentées, la part des cotisations de chacune des branches. L'assurance automobile, sur laquelle se porte le mémoire, représente plus de 10% du chiffre d'affaires de l'assurance.

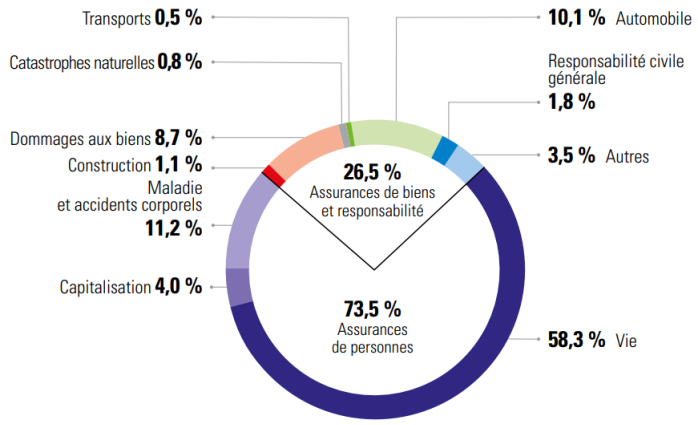


FIGURE 1.1 – Parts de cotisation en fonction des différentes branches d'assurance en France, source : France Assureurs, 2020

1.1.2 Le caractère obligatoire de l'assurance automobile

Le mémoire se concentre sur l'assurance dommage et plus particulièrement sur l'assurance automobile. En légère, mais constante évolution, le parc automobile français compte près de 44 millions de véhicules assurés en 2020. Le chiffre d'affaires de l'assurance automobile, lui aussi en constante hausse, s'élève en 2020 à 23,5 milliards d'euros [24], et représente un tiers du chiffre généré par l'assurance dommage.

Un contrat d'assurance automobile a pour objectif premier de couvrir les dommages matériels et corporels causés à autrui lors de l'utilisation d'un véhicule. Ce premier niveau de couverture est appelé assurance au tiers, ou de responsabilité civile. Les assureurs proposent également des contrats augmentés, disposant d'une protection face à un plus grand nombre de risques tels que le vol, le bris de glace ou les catastrophes naturelles.

L'assurance au tiers est obligatoire en France depuis 1958 : c'est une dépense incompressible pour une majorité des ménages. Étant un service de consommation essentiel, les assurés cherchent à optimiser cette dépense afin de limiter son impact sur le budget global. Par conséquent, l'assurance est un marché en proie à la concurrence. Les assurés comparent les prix et les couvertures pour arriver au contrat le plus attractif, en fonction de leurs besoins et de leurs moyens. Cette volatilité des clients est accentuée par la hausse brutale de l'indice des prix à la consommation (IPC), observée depuis la sortie de la crise sanitaire du covid-19. La Figure 1.2, issue des travaux de l'Insee, témoigne de cette tendance. Depuis la fin de l'année 2020, le glissement annuel, c'est-à-dire la variation observée entre les mêmes mois de deux années consécutives, de l'IPC est en constante hausse et dépasse les 5% en mai 2022. Les consommateurs sont d'autant plus sujets à optimiser leurs différents postes de dépenses.

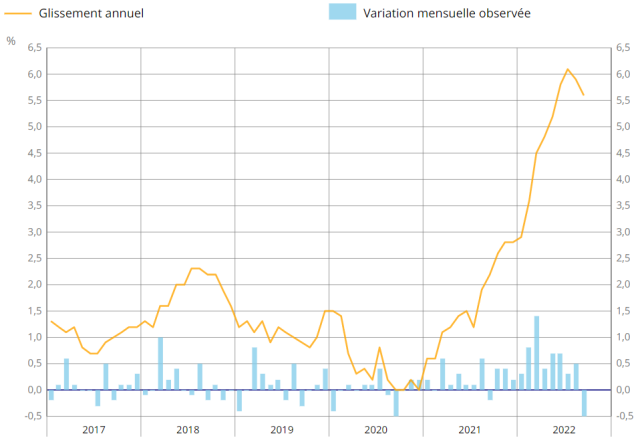


FIGURE 1.2 – Évolutions de l'indice des prix à la consommation, champs : France hors Mayotte, source : Insee

1.2 Quelle stratégie dans un marché toujours plus compétitif ?

1.2.1 Une concurrence accrue : nouveaux acteurs, digitalisation et législations

L'assurance automobile, au même titre que les autres branches de l'assurance, est un marché en perpétuelle évolution. Il en résulte un marché de plus en plus concurrentiel. Dans le cadre de l'assurance automobile, trois phénomènes tendent à accroître significativement la concurrence. Premièrement, l'entrée de nouveaux acteurs tels que les bancassureurs sur le marché de l'assurance dommage vient perturber l'équilibre anciennement établi. De plus, la rapidité de la digitalisation de la société n'échappe pas au secteur de l'assurance. Finalement, le contexte législatif, allégeant notamment les procédures de résiliation pour l'assuré, contribue à l'augmentation de la concurrence entre assureurs.

Dans les années 1990, les bancassureurs apparaissent sur le marché de l'assurance automobile et captent depuis continuellement plus de parts de marché. Encore inexistants en assurance dommage il

y a trente ans, les bancassureurs détiennent environ 10% des parts de marché en 2010. La Figure 1.3 permet d'observer qu'en 2020, plus de 15% des cotisations liées à l'assurance automobile sont réalisées par des bancassureurs. De plus, des acteurs dont l'offre est totalement en ligne font leur entrée sur le marché. Bien que la part détenue par ces acteurs exclusivement digitaux reste moindre pour l'instant, leur introduction sur le marché de l'assurance témoigne de la transformation de ce dernier, certains parlent même d'ubérisation de l'assurance.

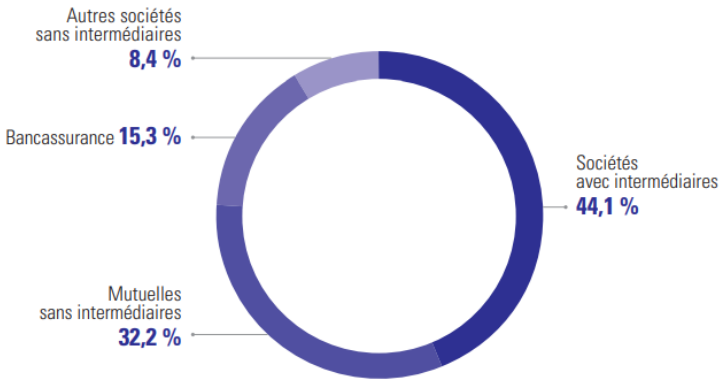


FIGURE 1.3 – Répartition des cotisations de l'assurance automobile par mode de distribution, source : France Assureurs, 2020

Un autre point à mentionner est celui de la digitalisation du secteur. Au-delà de la multiplication des offres dématérialisées, les consommateurs disposent d'un accès à l'information de plus en plus simplifié. En effet, la présence de sites comparateurs d'offres d'assurance permettent aux clients potentiels de confronter les différentes propositions et de sélectionner la plus compétitive. Quand les prospects sont sur-sollicités, les assureurs doivent s'adapter à cette transparence nouvelle et proposer des prix d'entrée d'autant plus attractifs.

De nouvelles législations offrant à l'assuré des facilités de résiliation contribuent à tendre un marché déjà concurrentiel. Les deux lois référentes dans ce domaine sont la loi Châtel et la loi Hamon. Entrée en vigueur en 2008, la loi Châtel impose à l'assureur de notifier l'assuré du renouvellement tacite de son contrat dans les semaines précédant la date limite de résiliation. La loi Hamon, mise en application depuis 2015, permet aux assurés, une fois la date du premier anniversaire passée, de résilier leur contrat d'assurance à tout moment. En outre, la loi du 16 août 2022 portant sur le pouvoir d'achat simplifie également l'acte de résiliation en ligne. Elle prévoit la possibilité de résilier en ligne les contrats souscrits par ce mode. De plus, les assureurs proposant de souscrire par voie électronique à la date de la résiliation se doivent de permettre à leurs assurés de résilier par internet. L'ensemble de ces mesures renforcent la concurrence du marché de l'assurance automobile.

Finalement, l'ensemble de ces nouveaux acteurs, outils, et réglementations, ont contribué à accroître la concurrence du secteur de l'assurance, et au sein de l'assurance automobile tout particulièrement. Les différents acteurs tentent alors de capter des parts de marché complexes à acquérir. Pour séduire de nouveaux clients, les assureurs, d'une part se voient investir dans des campagnes marketing, et d'autre part se doivent de présenter des offres attractives aux prospects, de diversifier leurs offres, de proposer des garanties additionnelles, des services gratuits ou des réductions tarifaires à certains profils. Cela induit un accroissement notable des coûts d'acquisition, à amortir sur une période de rétention d'autant plus courte.

1.2.2 Fidélisation ou acquisition ?

Les assureurs doivent fournir des efforts plus conséquents qu'auparavant pour se différencier de leurs nombreux concurrents et capter des parts de marché difficiles à atteindre. Cependant, une stratégie à ne pas négliger est celle de la fidélisation des assurés déjà présents en portefeuille.

De nombreuses recherches se sont penchées sur la question de la fidélisation des clients. Les premiers travaux mesurant les effets de la rétention sont proposés par Dawkins et Reichheld [29]. Publiés en 1990, leurs recherches tendent à montrer que mettre en place des stratégies de fidélisation d'un client coûte jusqu'à cinq fois moins cher que de chercher à en capter un nouveau. Le secteur de l'assurance n'échappe pas à cette règle et conquérir de nouveaux assurés est sensiblement plus cher que de conserver ceux déjà acquis. Les avantages de la rétention client sont multiples. Dans une démarche de rentabilité à court terme, il est plus efficace d'investir sur la rétention des clients déjà acquis. Cette stratégie est moins coûteuse que celle de l'acquisition de nouveaux prospects, qui engendre, entre autres, des dépenses marketing ou la proposition de tarifs d'entrées très attractifs. A plus long terme, la fidélisation des assurés permet une résilience aux crises conjoncturelles : les clients acquis et fidélisés tendent à moins résilier leurs contrats, et ce même dans un contexte économique difficile.

1.3 Motivations et exemples d'applications

Dans le contexte fortement concurrentiel auquel les assureurs font face et après avoir mis en évidence l'enjeu de la fidélisation des clients, l'implémentation d'outils prédictifs des comportements des assurés se révèle essentiel au bon fonctionnement économique des assureurs.

Chaque fin d'année, la période de renouvellement impose de prendre une décision quant au niveau de majoration à appliquer aux contrats. Un exemple de politique classique vise à attribuer une majoration plus importante aux assurés ayant déclaré un sinistre, selon leur degré de responsabilité. Bien entendu, les stratégies commerciales adoptées impactent directement la rétention client. Hormis les critères tarifaires, d'autres causes sont à l'origine d'une résiliation, certaines peuvent être citées :

- Le changement de véhicule. Il s'agit d'un moment clef pouvant entraîner un départ de l'assuré car il lui offre l'opportunité de questionner son contrat actuel. Le nouveau tarif appliqué associé au véhicule de remplacement peut également ne pas correspondre aux attentes de l'assuré et aux couvertures et services proposés.
- Le changement de vie familiale tel qu'un divorce, un nouveau conjoint ou un enfant devenu majeur peuvent modifier la demande. Ces changements peuvent générer l'introduction d'un conducteur secondaire, la création d'un contrat multi-équipement, le déménagement, un changement des revenus, ou la vente du véhicule par exemple.
- La mauvaise gestion d'un sinistre déclaré : le temps pour la perception de l'indemnisation, le montant finalement obtenu après expertise, la mauvaise relation avec l'assistance peuvent engendrer de l'insatisfaction client.
- Le changement de catégorie socioprofessionnelle : les nouveaux revenus peuvent questionner l'assuré sur l'adéquation du contrat ou du véhicule à sa situation.
- Le décès de l'assuré.

Pour pouvoir répondre aux besoins client, il est primordial pour un assureur de comprendre les mouvements qui s'opèrent au sein de son portefeuille, et d'anticiper les moments clefs entraînant la résiliation. L'identification des profils ayant de fortes chances de résilier à court terme ou ayant une faible durée de rétention permet d'entamer des plans d'action de nature marketing ou commerciale, afin de prolonger la vie du contrat et d'améliorer l'image de marque. La probabilité de rétention à un an, ainsi que la durée de vie a priori, constituent deux éléments permettant d'évaluer quantitativement la fidélisation client. L'objectif, à travers leur modélisation, est de pouvoir disposer de tous les outils permettant l'élaboration de scénarios tarifaires au moment du renouvellement. A partir d'une hypothèse d'évolution tarifaire, il sera indispensable de déterminer l'impact sur la durée de conservation du contrat, et les rentabilités à court terme et à moyen terme associées.

Deux modèles seront donc réalisés :

1. La probabilité $\hat{f}_i$ de résiliation à un an : connaissant les caractéristiques de l'assuré i et de son contrat au début de l'année, quelle est sa probabilité de résilier avant la fin de l'année ?
2. La durée de vie a priori $\hat{T}_i$: sachant les caractéristiques de l'assuré i au moment de sa souscription, combien de temps va-t-il rester au sein du portefeuille ? En pratique, ce n'est pas la durée de vie qui sera estimée mais la probabilité de résiliation $\hat{f}_i(t)$ pour un continuum d'instants t.

Ces modèles peuvent paraître redondants mais ils se complètent. Ensemble, ils constituent un outil abouti pour la réponse aux problématiques de rétention à court et moyen termes. En effet, les variables sur lesquelles ils s'appuient diffèrent et ainsi, leurs portées n'ont pas vocation à être les mêmes. Succinctement, le modèle de rétention à un an $\hat{f}_i$, s'appuie sur un jeu de variables large. Au-delà de caractéristiques classiques, à chaque instant de la vie du contrat, telles que l'âge de l'assuré, de son véhicule, ou son montant de cotisation, des variables d'avenants et tarifaires précises sont prises en compte. Les changements de domicile, de catégorie socioprofessionnelle, de véhicule ou encore la survenance et la nature des sinistres sont considérés. De plus, des variables tarifaires telles que la prime pure, la marge et le positionnement tarifaire de l'assureur sur le marché viennent affiner la modélisation. Ainsi, $\hat{f}_i$ fournit une estimation précise de la probabilité de rétention à partir des informations de début d'état de l'assuré. A l'inverse, le modèle de durée s'appuie sur un ensemble de variables plus restreint. A partir des informations concernant l'assuré, son véhicule et son contrat à la souscription, une probabilité de résiliation $\hat{f}_i(t)$, dépendante du temps, est estimée pour $t \in [0, T]$, avec T un horizon fixé. Autrement dit, $\hat{f}_i(t)$ propose une vision à long terme de la rétention des clients, à partir de leurs caractéristiques lors de la souscription.

Simulation d'indicateurs de profit

Une fois les quantités $\hat{f}_i$ et $\hat{f}_i(t)$ estimées pour chacun des individus i en portefeuille, plusieurs indicateurs de profit spécifiques, captant le risque porté par la résiliation de l'assuré, peuvent être calculés. Le profit à un an et la valeur client sont cités en exemple :

$$\text{Profit à un an} = \sum_{i=1}^{n} \left[P_{i,HT}(1) - S_i(1) \right] \cdot \left[1 - \hat{f}_i(P_{i,HT}(1), P_{i,\text{pure}}(1), X_i(1)) \right]$$

$$\text{Valeur client future de l'assuré } i = \sum_{t=1}^{T} \frac{P_{i,HT}(t) - S_i(t)}{1 + r(t)} \cdot \left[1 - \hat{f}_i(t; P^0_{i,HT}, X^0_i) \right]$$

où :
— n : nombre d'assurés en portefeuille ;
— $P_{i,HT}(t)$: tarif commercial hors taxes de la police d'assurance du client i au temps t ;
— $P^0_{i,HT}$: tarif commercial hors taxes de la police d'assurance du client i à sa souscription ;
— $P_{i,\text{pure}}(t)$: la prime pure de l'assuré i au temps t ;
— $S_i(t)$: montant des sinistres, additionné des frais, de l'assuré i l'année t. Pour $t = 1$ ce montant est déterministe, obtenu à partir de la prime pure, pour $t > 1$, $S(t)$ est aléatoire et est calculé au travers de modèles de projection de prime ;
— $X_i(t)$: caractéristiques de l'assuré i et de son contrat telles que son âge, le type de véhicule assuré et le niveau de couverture, au temps t ;
— X^0_i : caractéristiques de l'assuré i à la souscription ;
— T : horizon de calcul de la rétention, fixé à 5 ou 10 ans par exemple ;
— $r(t)$: taux d'intérêt à l'instant t.

Optimisation de la stratégie tarifaire à la reconduction

L'optimisation des valeurs décrites permettent à l'assureur la mise en place de stratégies visant à optimiser la rentabilité à plus ou moins long terme de son portefeuille, tout en prenant en compte, à l'aide de modèles prédictifs, le risque de résiliation. Par exemple, la stratégie tarifaire de l'assureur lors de la reconduction des contrats peut être affinée par maximisation du profit à un an :

$$\max_{P_{i,HT}(1) \quad i=1,\dots,n} \left\{ \sum_{i=1}^{n} \left[P_{i,HT}(1) - S_i(1) \right] \cdot \left[1 - \hat{f}_i(P_{i,HT}(1), P_{i,\text{pure}}(1), X_i(1)) \right] \right\} \tag{1.1}$$

En augmentant le tarif $P_{i,HT}(1)$ appliqué, l'assureur viendra augmenter son profit brut de résiliation. Cependant, l'augmentation du tarif commercial engendrera également une baisse de la rétention $1 - \hat{f}_i(P_{i,HT}(1), P_{i,\text{pure}}(1), X_i)$. Ainsi, ces deux effets sont à équilibrer en optimisant le programme sur les $P_{i,HT}(1)$. Dans une démarche d'optimisation tarifaire, l'Équation 1.1 est à optimiser certes, mais sous diverses contraintes. Par exemple, les volumes d'assurés en portefeuille à court, moyen et long termes doivent être considérés. A cet égard, le sujet s'avère particulièrement complexe, et c'est en ce sens que l'examen de modèles de rétention à différents termes s'avère essentiel à la mise en place d'une stratégie consistante.

Par l'étude de la probabilité de la résiliation à un an et de la durée de vie a priori, les travaux menés visent à :

- mieux appréhender le comportement des assurés face à la décision de résilier et ainsi améliorer la relation client ;
- identifier les profils rentables et fidèles ;
- améliorer les indicateurs de profit clefs du portefeuille et optimiser la marge de l'assureur ;
- obtenir les résultats escomptés suite à une stratégie tarifaire lors de la reconduction annuelle des contrats.

1.4 Périmètre

Comme évoqué précédemment, l'étude porte sur l'assurance automobile. Deux catégories sont à distinguer dans l'assurance des véhicules : celle des flottes d'une part et celle des particuliers d'autre part. Un contrat d'assurance de flotte automobile, généralement souscrit par une entreprise, permet de couvrir plusieurs véhicules. Ainsi, le contrat est souscrit par l'entreprise, désignée comme une personne morale, et couvre l'ensemble des véhicules de la compagnie. En revanche, un contrat souscrit par un particulier, désigné comme une personne physique, couvre un seul véhicule. Ces deux types d'assurance sont à distinguer tant par leur nature que par leurs spécificités. Les tarifs appliqués sont distincts et les clients n'adoptent pas les mêmes comportements, d'où la nécessité de ne pas étudier conjointement et de la même manière les contrats souscrits par les personnes morales et physiques. Par conséquent, il a été choisi de mener les travaux sur les contrats de particuliers uniquement.

La période d'étude est de 13 ans : la base de données comprend des contrats d'assurance automobile entre 2009 et 2021. Les travaux d'analyse et de modélisation seront réalisés sur les 12 premières années et l'année 2021, constituée de contrats en cours, appuiera les applications d'optimisation tarifaire.

Chapitre 2

Présentation et traitements des données

Sommaire

Cette partie s'attache dans un premier temps à décrire les données utilisées pour la réalisation de l'étude. Ensuite, seront abordés les différents traitements appliqués aux données. Ce travail de préparation de la base constitue le point de départ pour l'obtention de modèles fiables. Finalement, la création de nouvelles variables explicatives permettra une meilleure appréhension des problématiques de résiliation et de durée de vie. Les deux aspects de l'étude, qui sont la résiliation à un an et la durée de vie a priori, induiront des traitements spécifiques à chacune des finalités et donc à deux bases de données distinctes.

2.1 Description des données

Les travaux sont réalisés sur une base de données d'assureur, utilisée par le cabinet à des fins de recherche et de développement. Cette base de contrats d'assurance automobile comprend différentes informations sur l'assuré, son comportement, son véhicule et sur son montant annuel de cotisation. La base de données obtenue et utilisée par la suite, est le résultat d'un travail préalable. Celui-ci permet de disposer, en complément des informations classiques concernant l'assuré et son contrat, de données précises sur le véhicule et sur le niveau de risque. En effet, la base de contrat est additionnée d'une part, du véhiculier de l'association SRA (Sécurité Réparation Automobile) renseignant diverses caractéristiques techniques du véhicule assuré et d'autre part, d'un zonier donnant des informations de risque inhérentes au lieu de vie du client. Une seconde base de données permet d'intégrer des renseignements sur les sinistres subis par l'assuré tels que leur date de survenance et leur nature. Le portefeuille à disposition pour la suite de l'étude comporte environ 1,4 million de contrats, répartis sur plus de 6 millions de lignes. La Figure 2.1 schématise les données ainsi présentées et expose quelques-unes des variables employées dans l'étude.

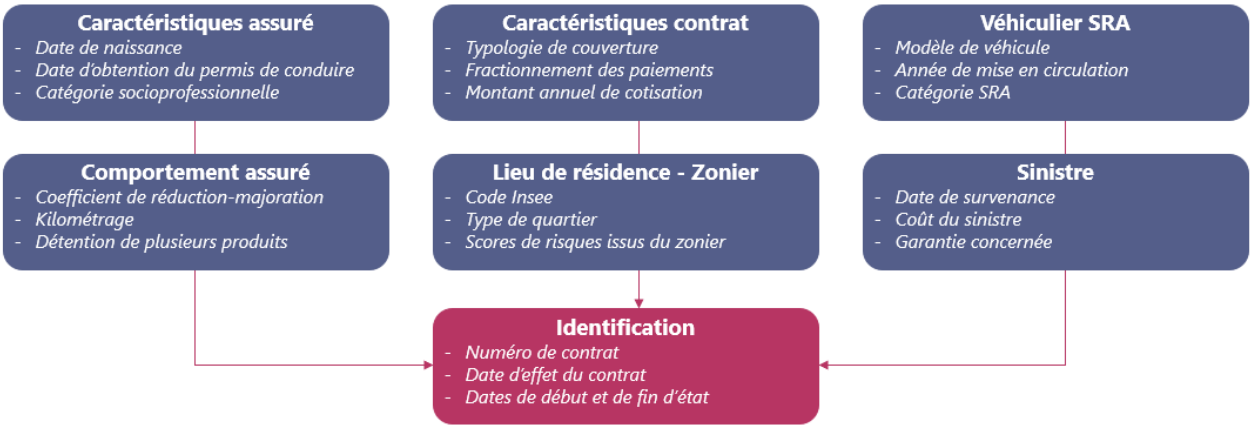


FIGURE 2.1 – Description de la base de données

Comme représenté dans la cellule rouge de la Figure 2.1, chaque ligne est identifiée par un numéro de contrat, une date d'effet et des dates de début et de fin de l'état concerné. Un même contrat peut vivre plusieurs états, appelés également images, et donc apparaître sur plusieurs lignes dans la base de données. Un changement d'état peut survenir pour deux raisons :

— lors de l'échéance du précédent état : au bout d'un an de vie, le contrat est tacitement reconduit sur un nouvel état ;
— lors de l'apparition d'un avenant dans le contrat. C'est-à-dire que si l'assuré vient à changer de situation, entraînant une modification des clauses de son contrat, un nouvel état, et donc une nouvelle ligne, informe de la situation actuelle du client.

Le Tableau 2.1 présente un extrait simplifié de la base de données et permet d'illustrer cette notion de contrat et d'état. Le contrat 18125, créé le 10/02/2017, y est observé. Le premier état de ce contrat

Numéro de contrat	Date début état	Date fin état	Date effet contrat	Identifiant SRA	...
18125	10/02/2017	09/02/2018	10/02/2017	RE23008	...
18125	10/02/2018	25/09/2018	10/02/2017	RE23008	...
18125	26/09/2018	25/09/2019	10/02/2017	VO25001	...

TABLE 2.1 – Extrait simplifié de la base de contrats

dure une année, puis le contrat est reconduit dans un nouvel état. L'assuré change de véhicule au cours

du deuxième état et en informe son assureur, un troisième état est alors créé pour rendre compte de ce changement de situation.

Un premier point concernant la qualité des données est à mentionner : la base ne renseigne pas les causes de résiliation des contrats. Dans le cadre de la résiliation, il est seulement observé qu'un contrat n'est pas reconduit à la fin d'un état ou arrêté en cours d'état, mais les causes associées à la rupture du contrat ne sont pas précisées. Bien que la cause de la résiliation ne soit pas mentionnée, cette dernière pourra se retranscrire au travers des variables à disposition. L'âge du véhicule traduit l'imminence de son remplacement, souvent accompagné d'un changement d'assureur et donc d'une résiliation. L'âge du conducteur permet de cerner les étapes majeures de la vie de l'assuré, telles que l'entrée dans la vie active, ou l'arrivée à la retraite par exemple. Ces étapes peuvent être synonymes de changement de situation et de besoins, qui induisent la réévaluation du contrat en cours, et potentiellement la résiliation de ce dernier. Les résiliations dues à l'insatisfaction du client suite à la gestion d'un sinistre pourront être captées à l'aide des variables de sinistralité, indiquant notamment la nature dudit sinistre. En outre, les données, ne disposant pas de variables concurrentielles telles que le tarif proposé par les autres assureurs du marché, ne permettent pas de capter correctement les causes des résiliations liées au mauvais positionnement tarifaire. L'apport externe de données concurrentielles, retranscrivant les résiliations dues à la recherche par l'assuré du tarif le plus compétitif, est alors essentiel.

2.2 Traitements généraux

Les enjeux des premiers traitements appliqués à la base d'étude sont multiples. Premièrement, il est nécessaire de disposer d'une base nettoyée pour mener à bien les différentes étapes de visualisation statistique puis de modélisation. Ensuite, les objectifs des travaux rendent primordiale la qualité du suivi de la vie des contrats. En effet, les données doivent présenter une conformité suffisante pour identifier l'acte de résiliation et calculer la durée de vie avec certitude.

2.2.1 Nettoyage

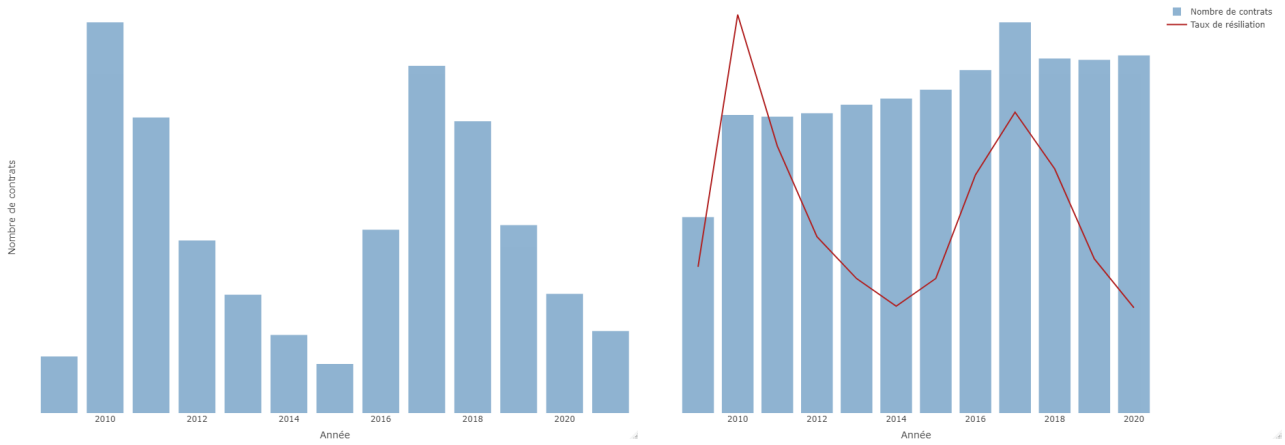
Dans un premier temps, un travail de gestion des anomalies de saisie est réalisé. Lors de données incohérentes ou manquantes, il existe deux possibilités de retraitement :

— la complétion ;
— la suppression parcimonieuse de la donnée non exploitable.

Quand cela est possible, un travail de complétion est mis en œuvre. Par exemple, il peut arriver que la date de naissance de l'assuré soit renseignée pour un des états mais pas sur l'entièreté du contrat. Dans ce cas, il faut appliquer cette date de naissance à toutes les lignes composant le contrat puis recalculer l'âge de l'assuré au moment du début de chacun des états. Cependant, il n'est pas possible de traiter de la sorte certaines anomalies. Dès lors qu'une information est nécessaire à la suite de l'étude et que cette dernière n'est pas, ou mal, renseignée, le contrat n'est pas exploitable et doit être supprimé. Par exemple, si le montant de cotisation annuelle de l'assuré n'est pas disponible ou est nul, ne serait-ce que pour un seul état, le contrat n'est pas exploitable et est exclu de l'étude. De nombreux retraitements similaires ont été appliqués pour identifier et remédier aux incohérences et aux valeurs manquantes des données. De plus, les contrats dont la durée de vie est inférieure à deux mois sont ôtés de la base d'analyse. Il peut s'agir de contrats souscrits par erreur, de contrats annulés par l'assuré qui trouve une couverture plus attractive chez un concurrent, d'individus qui assurent un véhicule dans le cadre d'une vente uniquement, ou de contrats que l'assureur résilie lui-même pour fausse déclaration. Ces résiliations sont en quelque sorte incompressibles et ne sont pas l'objet de l'étude. La présence de ces lignes impacte le taux de résiliation et les données associées ne reflètent pas nécessairement les mouvements des assurés en portefeuille. Finalement, environ 7% des contrats sont ainsi écrêtés des données.

2.2.2 Gestion de la migration des contrats

A la suite de refontes tarifaires ou de lancements de nouveaux produits, il arrive que des contrats existants soient reconduits sous un autre numéro. Ce processus est celui de la migration. Il n'est alors plus possible d'identifier l'origine et de suivre le parcours du contrat migré. Environ 25% des contrats de la base d'origine sont concernés. Poursuivre l'analyse sans corriger la migration conduirait à des erreurs d'identification de souscription et de résiliation. Cela induirait des taux de résiliation et de souscription plus élevés, ainsi que des durées de vie plus courtes que réellement observées. Figure 2.2 sont représentés, par année, le nombre de nouveaux contrats dont l'indicateur de migration est positif d'une part, et le taux de résiliation par année d'autre part. L'analyse de ces deux graphiques permet de constater que les années où sont observés le plus de contrats migrés sont également celles où les taux de résiliation sont les plus élevés. Par conséquent, il est nécessaire de retravailler les données afin de faire le lien entre les contrats migrés et les contrats initiaux. Pour ce faire, les informations autres que le numéro de contrat seront utilisées afin d'identifier les assurés.



(a) Nombre de contrats souscrits migrés par année

(b) Taux de résiliation par année

FIGURE 2.2 – Erreur dans l'identification de la résiliation due aux contrats migrés

L'exemple Tableau 2.2 illustre plus précisément le principe de la migration de contrat. Il est possible d'observer le contrat 217 pendant deux ans. Sur sa troisième année de vie, le contrat semble observer une migration : suite à un changement de calculette tarifaire, par exemple, les informations du contrat 217 réapparaissent sous un numéro de contrat différent. De plus, l'indicateur de migration est positif sur cette dernière année. Au vu des informations d'identification que partagent les deux contrats de l'exemple, il est très probable que ces deux numéros de contrat n'en représentent qu'un seul. Ainsi, les traitements de données suivants vont viser à identifier les contrats migrés, et à les raccorder à leurs contrats initiaux respectifs.

Numéro de contrat	Début état	Fin état	Effet contrat	Naissance	Permis de conduire	Code Insee	Id SRA	Migration	...
217	01/01/2012	31/12/2012	01/01/2012	03/01/1980	12/11/1999	60620	RE23008	0	...
217	01/01/2013	31/12/2013	01/01/2012	03/01/1980	12/11/1999	60620	RE23008	0	...
18125	01/01/2014	31/12/2014	01/01/2014	03/01/1980	12/11/1999	60620	RE23008	1	...

TABLE 2.2 – Exemple de contrat migré

Toujours dans l'exemple simplifié du Tableau 2.2, l'objectif est d'obtenir le Tableau 2.3. Le numéro du contrat et sa date d'effet sont corrigés, permettant, avec fiabilité :

— d'identifier l'acte de résiliation ;
— de calculer la durée de vie du contrat.

Numéro de contrat	Début état	Fin état	Effet contrat	Naissance	Permis de conduire	Code Insee	Id SRA	Migration	...
217	01/01/2012	31/12/2012	01/01/2012	03/01/1980	12/11/1999	60620	RE23008	0	...
217	01/01/2013	31/12/2013	01/01/2012	03/01/1980	12/11/1999	60620	RE23008	0	...
217	01/01/2014	31/12/2014	01/01/2012	03/01/1980	12/11/1999	60620	RE23008	1	...

TABLE 2.3 – Exemple de contrat migré corrigé

Comme dans le cadre des exemples Tableaux 2.2 et 2.3, un contrat migré est raccordé à son contrat initial par des critères concernant :

— l'assuré (date de naissance et date d'obtention du permis de conduire) ;
— l'enchaînement des dates des états (la fin du contrat initial doit précéder d'un jour le début du contrat migré) ;
— la typologie de couverture ;
— la voiture (identifiant SRA et date de mise en circulation) ;
— le lieu de résidence (code Insee).

Bien que l'utilisation de l'ensemble de ces critères d'identification induise une confiance élevée dans les raccords créés entre contrats, cette méthode est également trop restrictive pour permettre l'identification d'un grand nombre de contrats initiaux. Ainsi la méthodologie employée, illustrée par le Tableau 2.4, est itérative :

1. identification des contrats initiaux et des contrats migrés, à l'appui des critères cités ;
2. mise de côté des contrats identifiés ;
3. relance du même processus avec un ensemble de critères moins stricts.

Assuré	Dates des états	Typologie de couverture	Voiture	Lieu de résidence	Contrats migrés raccordés
x	x	x	x	x	31%
x	x		x	x	6%
x	x		x		11%
x	x			x	33%
x	x				7%
				Total	88%

TABLE 2.4 – Processus de raccords successifs de contrats migrés selon différents critères d'identification

Dans un premier temps, en utilisant tous les critères proposés, 31% des contrats migrés sont raccordés à leurs contrats initiaux respectifs. Ensuite ces numéros de contrats déjà identifiés sont mis de côté pour ne travailler que sur les contrats restants. En relâchant le critère de typologie de couverture, sont raccordés, en plus, 6% des contrats migrés. Le processus continue ainsi. Finalement, 88% des contrats migrés sont raccordés. Les 12% restants ne sont pas utilisables et sont donc supprimés.

A l'issue de ce travail, les taux de résiliation observés Figure 2.3 ne présentent plus de pics [1]. De plus, ils se situent autour de 14%, ce qui est cohérent avec le marché de l'assurance automobile des particuliers.

1. L'échelle est la même que celle utilisée Figure 2.2 pour permettre une comparaison adéquate.

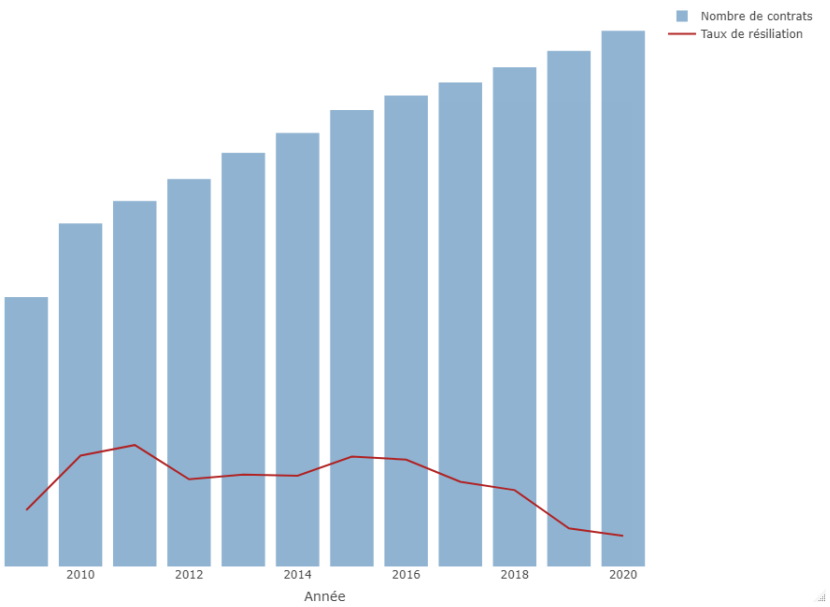


FIGURE 2.3 – Taux de résiliation par année après traitement des migrations

2.3 Traitements spécifiques à l'étude de la résiliation

La modélisation de l'acte de résiliation nécessite de disposer de données particulièrement nettoyées. Par souci de simplification du problème, les contrats doivent présenter des états qui se succèdent sans interruption. Ensuite, les états sont tenus d'être observés sur une période de temps équivalente, ce qui induit un travail de gestion des expositions. Ce travail n'est généralement pas réalisé, notamment dans la modélisation de la fréquence. L'ajustement lié à l'exposition peut se faire par un offset, qui correspond à l'ajout de l'exposition comme variable explicative avec un coefficient fixé à 1 dans les modèles. Cela est dû aux caractéristiques de certains modèles linéaires généralisés, que ne partage pas la régression logistique utilisée par la suite.

2.3.1 Cohérence de l'enchaînement des états

La nécessité pour la suite de l'étude de disposer de contrats observables sur toute leur durée de vie induit également de supprimer les contrats dont les informations ne sont pas renseignées sur certaines périodes. Le contrat 2327, dont les informations sont synthétisées Tableau 2.5, ne satisfait pas ce critère. Ce contrat est souscrit en 2014 et reconduit jusqu'en juin 2015. Bien que le contrat apparaisse en 2016 également, la trace de l'assuré est perdue entre juin 2015 et janvier 2016. Ainsi, la donnée disponible sur ce contrat ne peut pas être jugée comme fiable et ne pourra pas être utilisée dans les prochaines étapes de modélisation. Dans un souci de qualité des données, l'ensemble des états de ce contrat sont supprimés. Autrement dit, les contrats ayant observés des périodes sans être renseignés dans la base sont identifiés et supprimés.

Numéro de contrat	Date début état	Date fin état	Date effet contrat	...
2327	01/01/2014	31/12/2014	01/01/2014	...
2327	01/01/2015	05/06/2015	01/01/2014	...
2327	01/01/2016	25/09/2016	01/01/2014	...

TABLE 2.5 – Exemple de contrat sans information sur une période

2.3.2 Gestion des expositions

L'exposition est une valeur comprise entre 0 et 1, qui désigne en assurance la proportion de temps où un état est présent par rapport à un certain référentiel. Ici, le référentiel utilisé sera annuel, c'est-à-dire que si un état est observé du 1er janvier au 31 décembre, son exposition sera de 1. En revanche, si un état ne dure que six mois, son exposition sera de 0,5. Lors de la modélisation de la résiliation, il ne sera pas possible de considérer de la même manière un contrat ayant une exposition de 3, 1, ou 0,3 années. Par conséquent, chaque état doit avoir une exposition aussi proche de 1 que possible afin de prédire correctement la probabilité de la résiliation. Trois cas sont à traiter : celui où l'exposition est négative, celui où l'exposition est supérieure à 1 et celui où l'exposition est inférieure à 1.

Exposition négative

Les états présentant une exposition négative se révèlent après analyse comme étant uniquement liés à des défauts de saisie lors de l'extraction des données. Ces états sont supprimés et cette suppression ne vient pas perturber la fiabilité des données.

Exposition supérieure à 1

Lorsqu'un état présente une exposition supérieure à 1, ce dernier est découpé en plusieurs états distincts présentant tous les mêmes caractéristiques, de sorte que chacune des expositions soit inférieure ou égale à 1.

Exposition comprise entre 0 et 1

Le cas de figure où l'exposition est inférieure à 1 est très présent dans les données. En effet, dès lors que l'assuré observe un changement de situation, celui-ci doit prévenir son assureur et cet avenant conduit à la création d'un nouvel état, et ce même si l'état n'est pas arrivé à échéance. Le Tableau 2.6 décrit cette situation : suite à un changement de véhicule, un nouvel état est créé ce qui conduit à l'observation d'expositions inférieures à 1.

Numéro de contrat	Date début état	Date fin état	Date effet contrat	Identifiant SRA	Exposition	...
18125	10/02/2017	09/02/2018	10/02/2017	RE23008	1	...
18125	10/02/2018	25/09/2018	10/02/2017	RE23008	0.62	...
18125	26/09/2018	09/02/2019	10/02/2017	VO25001	0.38	...
18125	10/02/2019	09/02/2020	10/02/2017	VO25001	1	...

TABLE 2.6 – Exemple de contrat à l'exposition inférieure à 1

Afin d'obtenir des expositions égales à 1, il faut regrouper au sein du même état les états dont l'exposition est marquée en rouge. Dans ces circonstances, uniquement les caractéristiques du dernier état sont conservées. Le Tableau 2.7 présente alors les données traitées.

Numéro de contrat	Date début état	Date fin état	Date effet contrat	Identifiant SRA	Exposition	...
18125	10/02/2017	09/02/2018	10/02/2017	RE23008	1	...
18125	10/02/2018	09/02/2019	10/02/2017	VO25001	1	...
18125	10/02/2019	09/02/2020	10/02/2017	VO25001	1	...

TABLE 2.7 – Exemple de contrat après traitement de l'exposition inférieure à 1

Le Tableau 2.8 permet de se rendre compte du travail effectué sur les expositions. Il présente, hors états résiliés et états en cours, la répartition des valeurs prises par l'exposition, avant et après travail de cette dernière. Malgré les retraitements effectués, il reste quelques états dont l'exposition est

strictement inférieure, ou supérieure, à 1. Cependant, leur nombre maintenant très restreint ne viendra pas perturber les modèles.

Intervalle de valeurs	< 0.8	$[0.8; 1[$	1	> 1
Avant traitement	47%	12%	40,5%	0,5%
Après traitement	5%	13%	81,9%	0,1%

TABLE 2.8 – Répartition de l'exposition avant et après traitement

Focus sur la loi Hamon : résiliation à et hors échéance

Au-delà de la nécessité de détenir des contrats dont les états observent des expositions équivalentes pour la fiabilité des modèles mis en œuvre par la suite, le travail réalisé permet également la mise en évidence des effets directs de la loi Hamon. Bien que le mémoire ne porte pas sur ce sujet, dans le sens où les résiliations à et hors échéance ne seront pas distinguées par la suite, il reste pertinent d'étudier l'évolution des comportements de résiliation suite à l'entrée en vigueur d'une telle législation, dont les appréhensions et les impacts ont été retentissants dans secteur de l'assurance dommage. La Figure 2.4 propose d'observer, au sein des contrats résiliés, ceux qui ont été résiliés à l'échéance, en bleu, et hors échéance en orange. La courbe rouge représente le taux de résiliation hors échéance. Certes, le nombre de contrats résiliés reste stable au fil des années, et aucun choc de résiliation n'est observé en 2015, suite à la loi Hamon. Cependant, les comportements de résiliations sont différents. La proportion de contrats résiliés hors échéance augmente subitement en 2015, puis se stabilise au-delà de 2016. Plus de cinq ans après l'application de la loi Hamon, les impacts de cette dernière se dessinent clairement : les résiliations hors échéance ont augmenté de plus de 30%. Ces nouveaux comportements de résiliation contribuent à rendre incertains les revenus de l'assureur, qui se doit d'autant plus d'être en maîtrise du volume de son portefeuille.

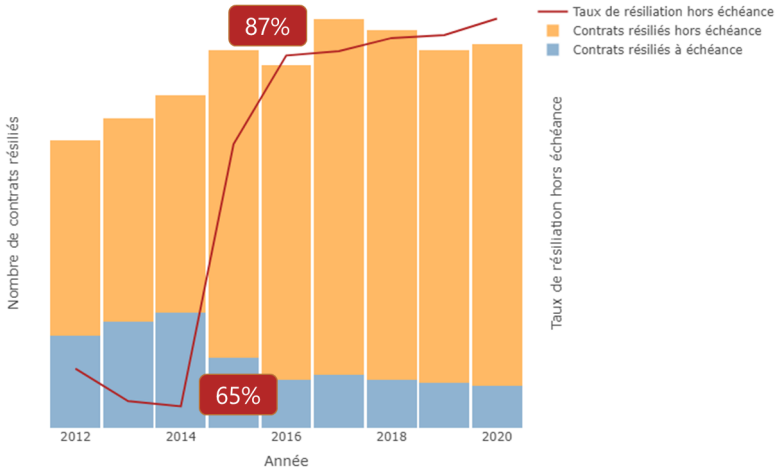


FIGURE 2.4 – Taux de résiliation - À échéance et hors échéance

2.4 Traitements spécifiques à l'étude de la durée de vie

L'étude de la durée de vie a priori consiste à estimer, à partir des informations à la souscription, la probabilité que l'assuré n'ait pas résilié à plusieurs instants $t \in [0; T]$. Ainsi, les contrats dont les informations lors de leur entrée dans le portefeuille ne sont pas disponibles ne peuvent être exploités. Le contrat 9237 est pris en exemple Tableau 2.9. Le contrat est souscrit le 01/01/2009, cependant les premières informations disponibles datent du 01/01/2012, trois ans après. Les caractéristiques de

l'assuré, de son véhicule ou de sa couverture lors de sa souscription ne sont pas disponibles. Par conséquent, ce contrat ne pourra être mis à profit dans le cadre de la modélisation de la durée de vie a priori.

Numéro de contrat	Date début état	Date fin état	Date effet contrat	...
9237	01/01/2012	31/12/2012	01/01/2009	...
9237	01/01/2013	31/12/2013	01/01/2009	...
9237	01/01/2014	31/12/2014	01/01/2009	...

TABLE 2.9 – Exemple de contrat non exploitable pour la modélisation de la durée de vie

2.5 Création de variables

Une fois les données retraitées, un travail de création de variables permet de répondre au mieux à la problématique de résiliation et de durée de vie. Dans un premier temps, il est essentiel d'identifier les contrats toujours en cours et les contrats résiliés. Ensuite, différentes variables explicatives sont créées afin de mieux appréhender le comportement des assurés.

2.5.1 Censure et résiliation

Donnée de durée, censure et troncature

L'étude de la durée de vie induit des difficultés spécifiques : quand des variables usuelles peuvent être mesurées à un instant t discret, la collecte de données de durée nécessite d'une part de disposer d'une période d'observation suffisante, et induit d'autre part des problèmes propres. La notion de données incomplètes dans le cadre de durées est introduite en 1958 par Kaplan et Meier [19]. Les auteurs exposent notamment l'idée que lors de l'étude de durées, des données incomplètes peuvent survenir suite à la nécessité de réaliser l'étude en question dans un temps raisonnable. Dans le cas de la durée de vie des contrats d'assurance automobile, il n'est pas possible de disposer d'une période d'observation assez longue pour constater de la résiliation de tous les contrats en cours. Ainsi, une partie des contrats présente des durées de vie C_i inférieures à leurs durées de vie réelles T_i, non observées pour l'instant.

Précisément, ces phénomènes sont ceux de la censure et de la troncature. Cette problématique de données manquantes est vaste et a suscité de nombreuses recherches. La littérature sur ce sujet est riche, tant sur la définition de ces notions que sur les méthodes d'estimations de durées en présence de ces singularités. Entre autres, Klein et Moeschberger [20] spécifient plusieurs catégories de censure, telles que la censure à droite, la censure de gauche et la censure d'intervalle. Les auteurs abordent également les concepts de troncature à gauche et à droite.

Le terme de censure à droite illustre le fait que l'individu n'est plus observé à partir d'un certain moment, à droite de la frise chronologique. Une censure à droite peut être de type I, II ou III :

1. La censure de type I, aussi appelée censure fixe, intervient lorsque la variable d'intérêt n'est observée que si elle survient avant un certain moment spécifié. Ce type de censure est très répandu, c'est celle à laquelle les données utilisées sont confrontées : si l'acte de résiliation ne se produit pas avant la fin de la période d'étude, ce dernier ne sera pas observé. Concrètement, la censure de type I concerne l'ensemble du portefeuille en cours. Formellement, soient $(T_1, ..., T_n)$ les durées de vie des n contrats observés. Soit $(C_1, ..., C_n)$ les variables caractérisant la date à laquelle les données sont extraites, et donc date à laquelle les individus ne sont plus suivis. Les $(T_1, ..., T_n)$ ne peuvent être observées, seules les variables $((Y_1, \delta_1), ..., (Y_n, \delta_n))$ sont constatées. Ces variables sont telles que :
$$\begin{cases} Y_i = \inf(T_i, C_i) \\ \delta_i = \mathbb{1}_{T_i \leq C_i} \end{cases}$$

2. La censure de type II intervient lorsque l'étude s'achève dès lors qu'un certain nombre d'individus est sorti de l'observation. Par exemple, la période d'observation s'arrête quand r individus ont résilié leur contrat d'assurance, avec $r < n$ un entier défini par avance.

3. Finalement, le type de censure III, dit aléatoire ou de risques concurrents, intervient dans le cas où les individus sortent aléatoirement de l'étude lorsqu'ils subissent un risque autre que celui étudié.

La censure à gauche concerne les individus ayant vécus l'évènement d'intérêt avant le début de la période d'observation. Des données peuvent également être censurées par intervalles, c'est-à-dire qu'il existe des périodes au cours desquelles les sujets ne sont pas observés : il n'est alors pas possible de déterminer précisément la date de survenance de l'évènement d'intérêt. Les données dont nous disposons ne présentent pas ce type de censure. La troncature, qui peut se présenter à gauche ou à droite, est un phénomène dans la collecte et l'étude de durées qui est à différencier de la censure. A l'inverse de la censure, un individu tronqué ne sera pas observé, il ne sera donc pas possible de déceler sa présence au sein des données.

Enjeux sur le portefeuille d'étude

L'enjeu de l'étude menée sera celui de la censure à droite de type I. En effet, les données sont extraites fin 2021. A cette période, des contrats sont toujours en cours : l'assuré n'a pas résilié son contrat d'assurance et de ce fait sa durée de vie T n'est pas observée. Les autres types de censure et de troncature ne viennent pas perturber l'observation de l'acte de résiliation et donc de la durée de vie. Néanmoins, une partie des contrats constatent des défauts de qualité qui conduisent à perdre l'information de l'assuré pendant un certain laps de temps. Deux cas de figure se présentent : celui où la base ne renseigne pas les informations de l'assuré à sa souscription et celui exposé dans la Section 2.3.1, où l'assuré est absent de la base pendant un certain temps. La Figure 2.5 recense les différentes situations présentes dans le jeu de données.

FIGURE 2.5 – Exemples de contrats présents dans les données

Les trois cas identifiés sont les suivants :

1. En orange est représenté un contrat censuré à droite : il n'est pas résilié avant la fin de la période d'observation.

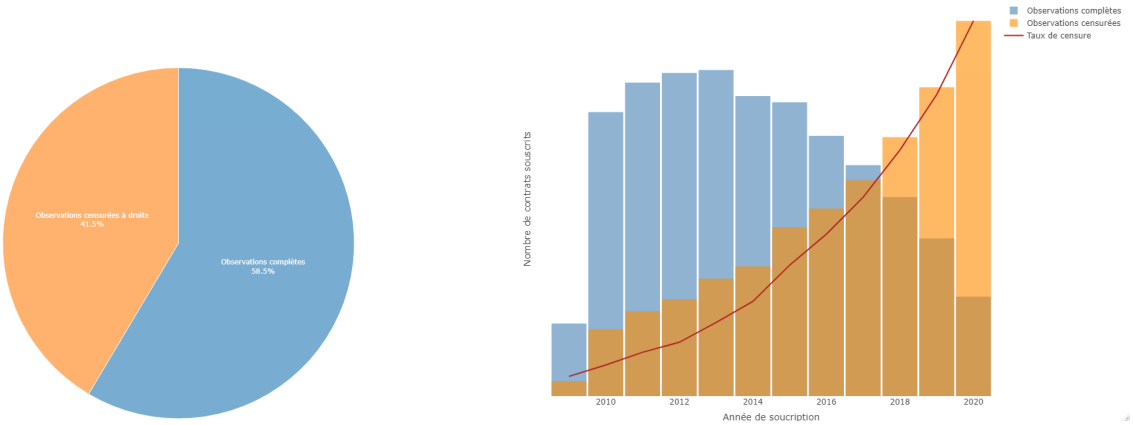
2. En rouge sont présentés des contrats dont des informations manquent sur une période. Dans la situation du premier contrat, bien que la date de souscription soit connue, les informations de l'assuré à l'ouverture du contrat ne sont pas disponibles. Le contrat suivant subit une période, au cours de sa vie, où ses informations ne sont plus renseignées. Pour ces deux contrats, il est possible d'identifier leurs dates de résiliation et leurs durées de vie. Malgré cela, les contrats de ce type poseront des problèmes dans la modélisation :

(a) les contrats dont les informations à la souscription ne sont pas connues ne seront pas utilisés dans le cadre de la durée de vie : ces modèles nécessitent de disposer des caractéristiques initiales du client ;

(b) les contrats qui présentent des discontinuités ne seront pas exploitables dans le contexte des modèles de résiliation. Il est en effet judicieux de pouvoir identifier les avenants et les évolutions effectives d'un état sur l'autre.

3. Le contrat bleu représente la donnée idéale : toutes les informations du contrat sont disponibles et sa durée de vie est connue.

Un travail d'identification de l'acte de résiliation a été mené dans les données. Au moment de l'extraction des données, deux configurations sont possibles : soit le contrat a été résilié en amont, soit il est encore en cours. Si la date de fin du dernier état d'un contrat est antérieure à la date d'extraction des données, ce dernier est résilié. Dans le cas contraire, le contrat est encore en cours et donc censuré à droite. La figure 2.6 permet de visualiser la proportion de contrats censurés et résiliés dans la base de données.



(a) Proportion d'observations censurées

(b) Taux de censure par année

FIGURE 2.6 – Taux de censure, au global et par année de souscription

Sont considérés uniquement les contrats des cas 1. et 3. (en orange et en bleu dans la Figure 2.5) de la liste précédente. Au global, 58,5% des contrats observés dans la base sont résiliés dans la période d'observation. Les 41,5% des contrats restants sont toujours en cours et sont donc dits censurés à droite. Il est également possible d'examiner le nombre de contrats résiliés et censurés selon leur année de souscription. Cette première analyse assure la cohérence de l'identification ainsi réalisée : le nombre de contrats censurés va croissant avec l'année de souscription. Certes, plus un contrat a été souscrit récemment plus sa probabilité d'être encore en cours est grande.

2.5.2 Variables explicatives

Identification des avenants

Un facteur important dans la vie d'un contrat d'assurance est celui des changements de situation, aussi appelés avenants. Un certain nombre de changements de situation tels que les changements de domicile, de situation maritale ou la vente du véhicule assuré constituent un motif suffisant pour que le client résilie son contrat. Cependant dans le cas des données utilisées, le motif de la résiliation n'est pas renseigné : être en mesure d'observer ces changements permettra alors une meilleure compréhension de la fidélisation des clients. Par exemple, un assuré qui cède son véhicule mais ne change pas d'assureur aura tendance à moins résilier en moyenne qu'un client n'observant pas de changement de véhicule. En effet, ce dernier ne profitant pas de son changement de véhicule pour résilier son contrat est un client plus fidélisé que la moyenne.

Ainsi, plusieurs changements de situation sont considérés :

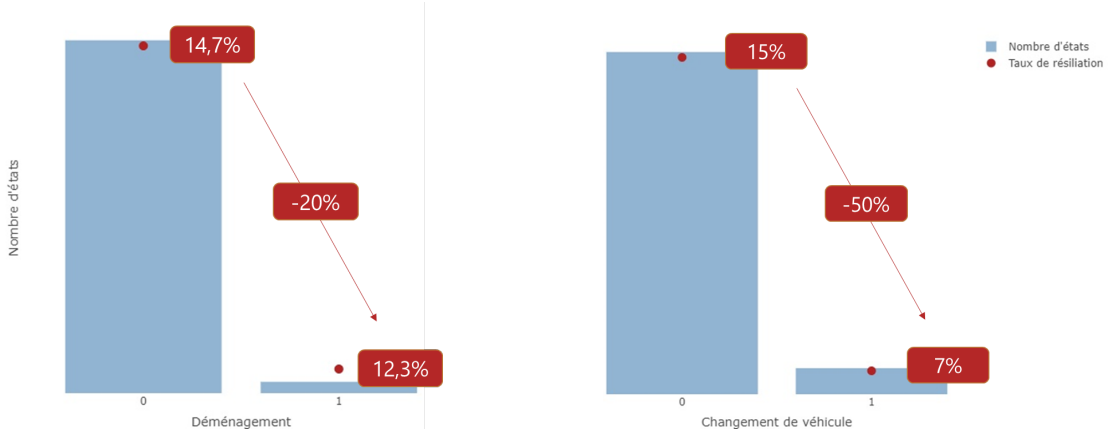
— véhicule ;
— adresse ;
— niveau de garantie.

Le Tableau 2.10 illustre l'identification des changements de situation. D'un état à l'autre, l'assuré change de véhicule : cela s'observe par le changement d'identifiant SRA du véhicule. La variable indicatrice du changement de véhicule est créée et prend pour valeur 1 lorsque l'identifiant SRA de l'état n'est pas le même que celui du précédent.

Numéro de contrat	Date début état	Date fin état	Identifiant SRA	Changement de véhicule	...
1293	01/01/2014	31/12/2014	RE23008	0	...
1293	01/01/2015	31/12/2015	VO25001	1	...
1293	01/01/2016	31/12/2016	VO25001	0	...

TABLE 2.10 – Identification d'un changement de véhicule

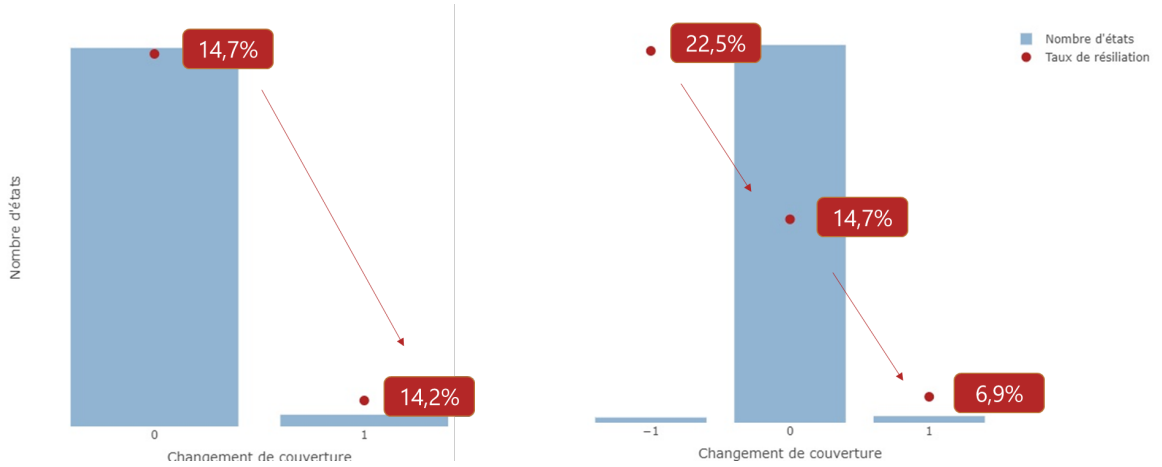
Les figures 2.7 et 2.8 permettent de visualiser l'impact d'un changement de situation d'un état $n-1$ à l'état n sur la résiliation de l'assuré à la fin de l'état n. Il y est représenté le taux de résiliation moyen en rouge, en fonction de l'observation, ou non, d'un changement de situation. L'intuition sur la fidélisation de l'assuré se vérifie ici, notamment sur le changement de voiture et de domicile. Bien qu'ayant observé un changement, l'assuré ne résilie pas son contrat d'assurance automobile. Il est alors plus enclin à rester chez son assureur l'état suivant qu'un autre client en moyenne. Un assuré qui a déménagé a une probabilité de résiliation 20% inférieure à un assuré qui n'a pas changé de domicile. Ce phénomène est encore plus manifeste dans le cas du changement de véhicule : un assuré qui vient de changer de véhicule résilie en moyenne deux fois moins qu'un autre assuré.



(a) Changement de domicile (b) Changement de véhicule

FIGURE 2.7 – Impact des changements de domicile et de véhicule sur la résiliation

Un changement de couverture est aussi un facteur de modification du comportement des assurés. La figure 2.8 présente les taux de résiliations moyens pour les individus ayant changé, ou non, de couverture. Une analyse plus fine permet de mieux capter la fidélisation des clients. En effet, un assuré qui change de couverture et qui se dirige vers un contrat moins complet résiliera 1,5 fois plus qu'un client qui ne change pas de situation, et 3 fois plus qu'un client qui opte pour une couverture plus complète.

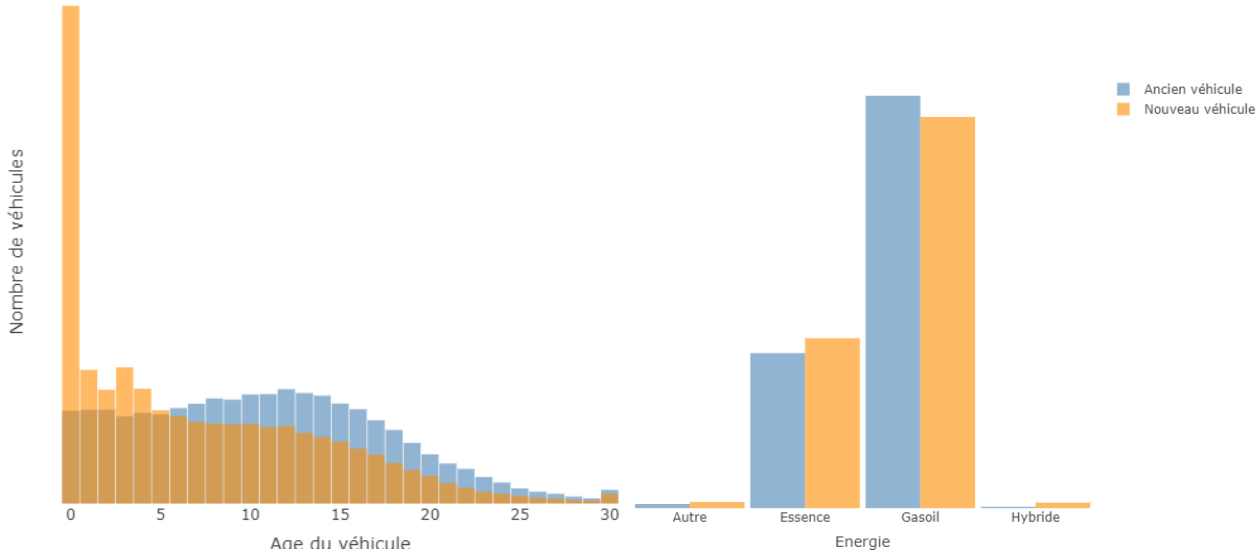


(a) Changement de couverture (b) Changement de couverture - distinction de cas

FIGURE 2.8 – Impact des changements de couverture sur la résiliation

Focus sur le changement de véhicule

Céder son véhicule pour un autre est une des causes principales de la résiliation. Comme expliqué précédemment, les causes de la résiliation ne sont pas explicitées dans la base de données. Cependant, à partir des caractéristiques du véhicule assuré, le modèle de résiliation sera en mesure de capter l'imminence du changement de véhicule et du risque lié. Cela peut s'observer Figure 2.9, où sont représentées les caractéristiques des véhicules avant changement, en bleu, et après changement, en orange. Les véhicules assurés après un changement sont plus récents et tendent à être plus alimentés à l'essence et à l'électrique que les véhicules précédant le changement.



(a) Âges des véhicules (b) Énergie des véhicules

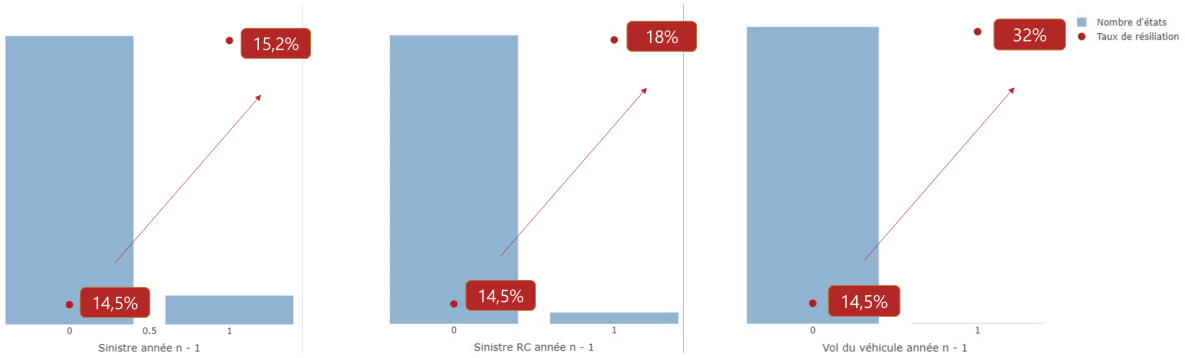
FIGURE 2.9 – Caractéristiques des véhicules avant et après leur changement

Sinistralité

La survenance d'un sinistre peut induire un mécontentement de la part de l'assuré, entre autres de par la gestion de celui-ci par l'assureur. Il est primordial pour l'étude de la résiliation et de ses facteurs de disposer d'une variable recensant les éventuels sinistres subis. Une base de sinistre, indépendante

de la base des contrats est disponible. En fusionnant ces deux bases, trois variables indicatrices sont créées, elles indiquent de la présence au cours de l'état :

— d'un sinistre, toutes garanties confondues ;
— d'un sinistre concernant la garantie de responsabilité civile ;
— d'un vol du véhicule, conduisant à la disparition du bien assuré.



(a) Sinistre - Toutes garanties (b) Sinistre - Responsabilité civile (c) Sinistre - Vol

FIGURE 2.10 – Impact de la survenance d'un sinistre, année $n-1$, sur la résiliation, année n

Toutefois, les modèles construits dans la suite de l'étude serviront à prédire les probabilités de résiliation à un an des assurés. Leur sinistralité sur l'année à venir n'est, de fait, pas connue. Ainsi, pour une année n, les variables retenues pour la construction des modèles sont celles de la survenance d'un sinistre l'année $n-1$. Pour des raisons prospectives, les informations portées par ces variables restent parlantes dans l'étude des facteurs de la résiliation. Figure 2.10 sont proposés les taux de résiliation observés en fonction de la survenance de différents types de sinistres. Un sinistre, toutes garanties confondues, augmente de 10% le taux de résiliation, quand il est augmenté de 20% environ par un sinistre touchant à la responsabilité civile ou au vol. L'exposition très faible de l'état ou l'indicatrice de vol est positive rendra son appréhension difficile par les modèles.

2.5.3 Variables tarifaires

Revalorisation de tarif

La variation de tarif est un facteur d'insatisfaction des assurés, et donc de résiliation, notamment lorsqu'elle est observée à la hausse. Le calcul de la revalorisation annuelle permet de capter la fluctuation des tarifs. En notant MC_n le montant de la cotisation de l'assuré toutes taxes comprises au cours de l'année n, la revalorisation RC_n observée l'année n s'obtient :

$$RC_n = \frac{MC_n - MC_{n-1}}{MC_{n-1}}$$

Il est à relever que cette variable de revalorisation n'est disponible qu'à partir du second état des contrats, elle ne pourra donc pas être utilisée dans les modèles. De surcroît, la résiliation de l'année n aura tendance à être impactée par RC_{n-1} et non par RC_n. Bien que l'assuré observe l'année $n-1$ un changement tarifaire, ce dernier sera enclin à ne pas réagir immédiatement, et à résilier son contrat au cours de l'année n. Figure 2.11, il est possible d'observer la croissance du taux de résiliation à l'état n, en fonction de la variable RC_{n-1}.

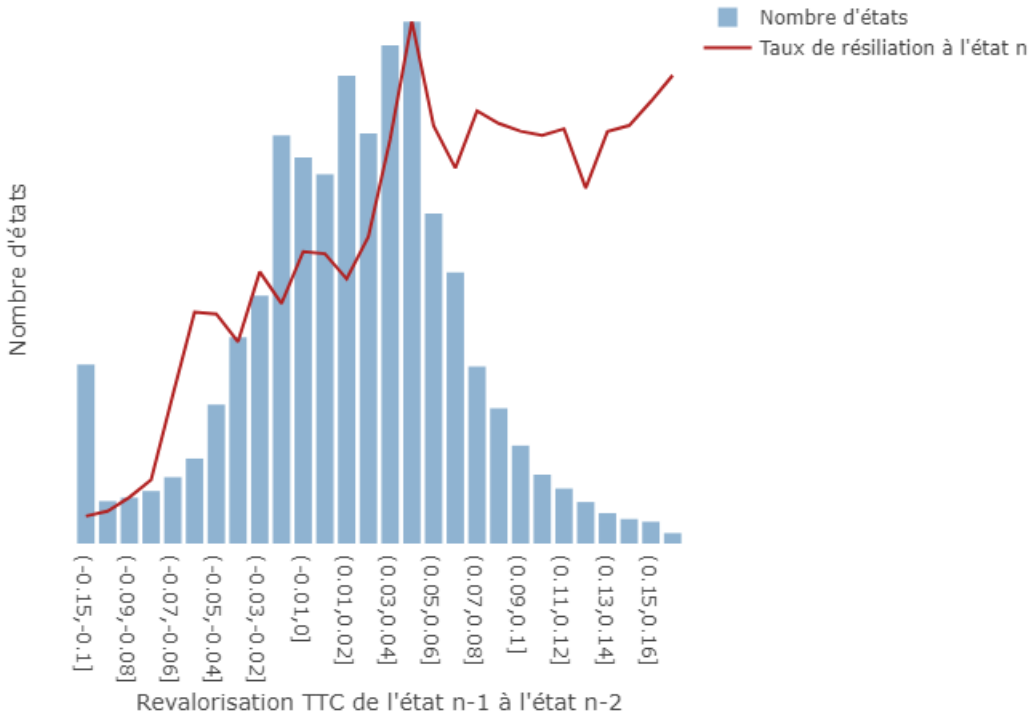


FIGURE 2.11 – Taux de résiliation année n en fonction de la revalorisation de tarif RC_{n-1}

Prime pure

La prime pure correspond à l'espérance des sinistres, c'est-à-dire à une estimation du montant attendu du risque. Elle est un élément central de la tarification et doit répondre à deux principes, qui bien que antagonistes, sont fondamentaux dans l'assurance non vie : la mutualisation et la segmentation. Charpentier, Denuit et Elie discutent de ces notions en abordant différents niveaux de segmentations [5]. Le principe de mutualisation repose sur l'idée de considérer un grand nombre de risques, afin de réduire la moyenne et la variance du risque global. Ce prince fait allusion au concept de la loi des grands nombres, bien que cette analogie soit à considérer avec précaution [4]. En revanche, la segmentation implique que l'assureur doit constituer des classes de risques homogènes, à partir des informations dont il dispose sur les assurés et sur leurs véhicules. Le coût précis et individuel de chaque assuré ne peut être déterminé en amont. Cependant le risque collectif est en partie prévisible, et ce à partir de l'expérience passée.

L'approche classique du calcul de la prime pure est celle de fréquence sévérité. Elle consiste à estimer la fréquence des sinistres ainsi que leur sévérité moyenne puis à réaliser le produit des deux. Formellement, soit X la variable aléatoire telle que :

$$X = \sum_{k=1}^{N} B_k$$

où :

- N est une variable aléatoire de comptage qui recense le nombre de sinistres ;
- les B_k sont les variables aléatoires, indépendantes et identiquement distribuées, modélisant le coût de chacun des N sinistres ;
- N est supposée indépendante de B_k, $\forall k$.

Ainsi, la prime pure correspond à :

$$\begin{aligned} \Pi(X) &= \mathbb{E}[X] \\ &= \mathbb{E}[N] \cdot \mathbb{E}[B] \end{aligned}$$

Des travaux de tarification, réalisés en amont par le cabinet, permettent d'obtenir la prime pure de chaque contrat. Pour ce faire, chaque assuré se voit associer ses coefficients, en cohérence avec sa classe de risque. Ensuite, la mise en place d'une calculette tarifaire permet d'obtenir la prime pure sur chacune des garanties (responsabilité civile, bris de glace, etc.). Finalement, la prime pure globale est obtenue en sommant les primes de chaque garantie, en fonction de la typologie de couverture de l'assuré. Des tests de cohérence sont réalisés et permettent de s'assurer de la fiabilité des calculs menés. Par exemple, la Figure 2.12 présente l'évolution de la prime pure en fonction de l'âge des assurés. La prime pure est cohérente et observe la décroissance attendue. Les conducteurs les plus jeunes observent une prime pure élevée en raison de leur grande sinistralité. Cette dernière décroît jusqu'à noter une légère bosse au niveau des profils ayant 40 à 50 ans : une partie de ces contrats présentent des conducteurs secondaires novices, souvent les enfants du conducteur principal. Ensuite, une légère augmentation de la prime pure survient au-delà de 70 ans. Cela est dû à une augmentation de la fréquence des sinistres chez les assurés les plus âgés, mais ce phénomène est en partie compensé par la détention de véhicules plus anciens et donc moins onéreux.

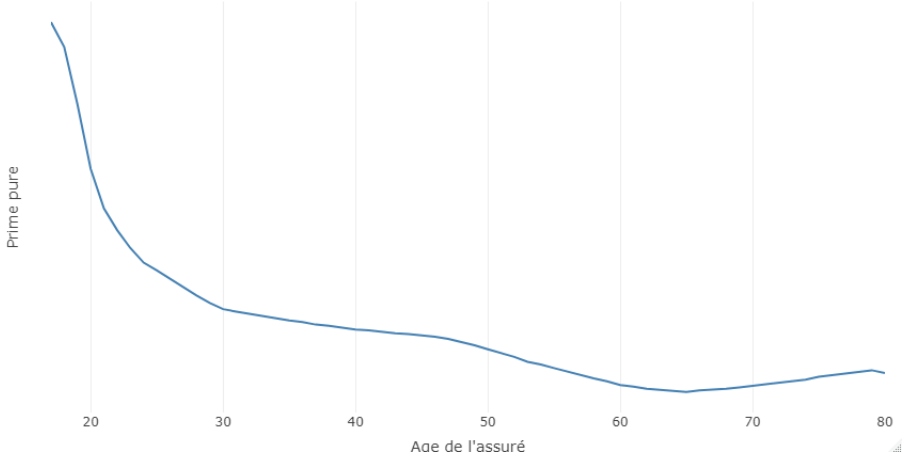


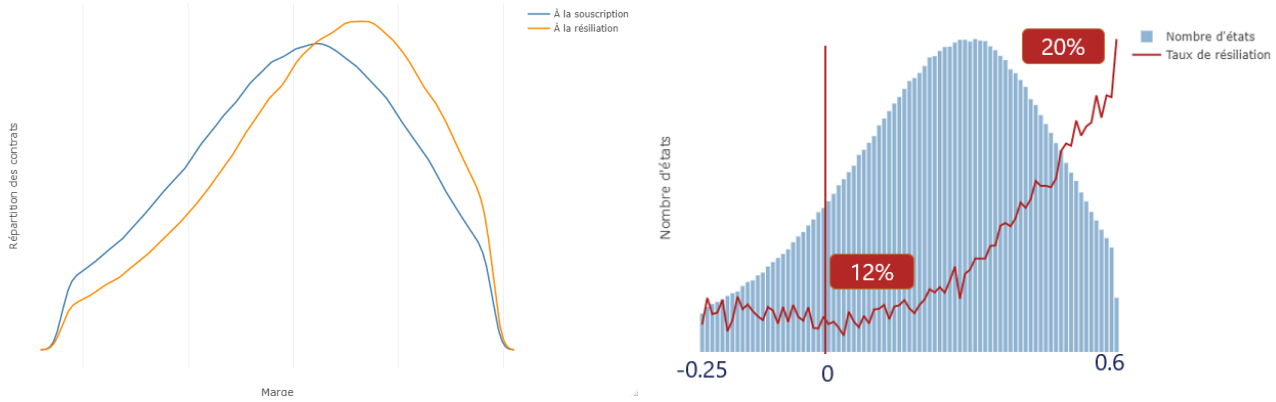
FIGURE 2.12 – Évolution de la prime pure avec l'âge des assurés

Marge

Une fois la prime pure calculée, il est possible de lui ajouter un taux de chargement - déterminé en fonction du niveau de couverture - puis de calculer la marge. La marge correspond au ratio :

$$\text{Marge} = \frac{\text{Montant cotisation HT} - \text{Prime pure} \cdot [1 + \text{Taux de chargement}]}{\text{Montant cotisation HT}}$$

Différentes analyses sont menées sur la marge et son évolution en fonction de différents facteurs. Il est intéressant, entre autres, de représenter la répartition de la marge à la souscription (en bleu) et à la résiliation (en orange). Figure 2.13a est observé que la marge réalisée est plus faible au moment de la souscription, ce qui est rationnel dans la démarche de fidélisation de nouveaux clients. Une partie de ce décalage peut également être expliqué par le fait que la prime pure a été calculée sans prise en compte directe de l'inflation. La figure 2.13b permet d'observer la croissance nette de la résiliation avec celle de la marge sur la prime pure. Dès que la marge devient positive, les taux de résiliation vont croissant.



(a) Marge à la souscription et à la résiliation

(b) Résiliation en fonction de la marge

FIGURE 2.13 – Étude de la marge sur la prime pure

Tarif médian : intégration de données concurrentielles

Le positionnement de l'assureur sur le marché est un facteur non négligeable dans l'analyse des comportements des clients. Les prix proposés par les concurrents influent sans commune mesure les décisions des assurés, qui cherchent naturellement à bénéficier du tarif qu'ils jugent comme étant le plus juste sur le marché. Par conséquent, il est proposé d'augmenter les données dont nous disposons en intégrant, pour chaque contrat et chaque année d'exercice, le tarif médian du marché. La méthodologie [2] pour l'obtention de ce tarif est la suivante :

1. constitution de profils d'assurés cohérents et représentatifs du marché de la demande d'assurance ;
2. collecte automatisée des tarifs proposés sur un site comparateur de contrats d'assurance automobile ;
3. correspondance entre la base marché constituée par les deux étapes précédentes et la base de contrats de l'étude ;
4. modélisation du tarif médian, calculé à partir de l'ensemble des tarifs proposés à la souscription.

Des tests de cohérence du tarif médian sont réalisés et permettent de s'assurer de la conformité des résultats. De nouveau, le test réalisé sur l'âge de l'assuré est présenté Figure 2.14 : le tarif est consistant.

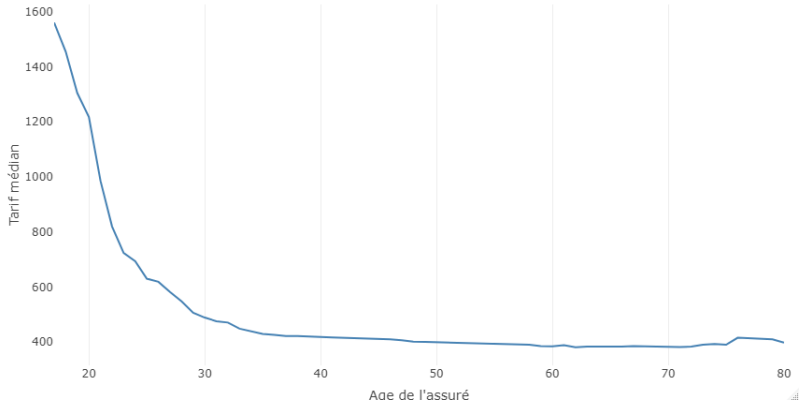


FIGURE 2.14 – Tarif médian en fonction de l'âge des assurés

2. Cette méthode est proposée et a été réalisée en amont par Mme. Linda Krolikowski, directrice de ce mémoire. La collecte et l'intégration de données concurrentielles s'appuient exhaustivement sur ses travaux [21].

La période d'extraction du tarif médian est 2020, ce qui engendre une limite dans son utilisation. Ce tarif, en cohérence avec les tendances du marché en 2020, ne peut être appliqué aux années antérieures : une actualisation est nécessaire. Elle est réalisée à partir de l'évolution de l'indice Insee des prix de l'assurance auto, présentés par la FFA (Fédération française de l'assurance) [23] et disponibles Tableau 2.11.

Année i	Indice r_i
2019	+3,0%
2018	+3,2%
2017	+1,4%
2016	+1,3%
2015	+1,7%
2014	-0,2%
2013	-1,5%

TABLE 2.11 – Indice Insee des prix de l'assurance auto entre 2013 et 2019, source : FFA, 2019 [23]

L'augmentation des prix de l'assurance auto est supposée de +3,0% également entre l'année 2019 et 2020. En notant r_i l'indice Insee de l'année i, le tarif médian de l'année n s'exprime :

$$\text{Tarif médian}(n) = \frac{\text{Tarif médian}(2020)}{\prod_{i=n}^{2020}(1+r_i)}$$

La variable utilisée dans les modèles sera l'écart relatif du montant de cotisation proposé au tarif médian du marché :

$$\text{Écart relatif tarif médian} = \frac{\text{Montant cotisation TTC} - \text{Tarif médian}}{\text{Montant cotisation TTC}}$$

La Figure 2.15 représente le taux de résiliation en fonction de l'écart relatif au tarif médian. Un comportement similaire à celui observé lors de l'étude de la marge se dessine : une fois l'écart au tarif médian positif, le taux de résiliation va croissant. Autrement dit, plus le montant de cotisation est élevé relativement au tarif médian du marché, et plus le taux de résiliation observé en moyenne est élevé également.

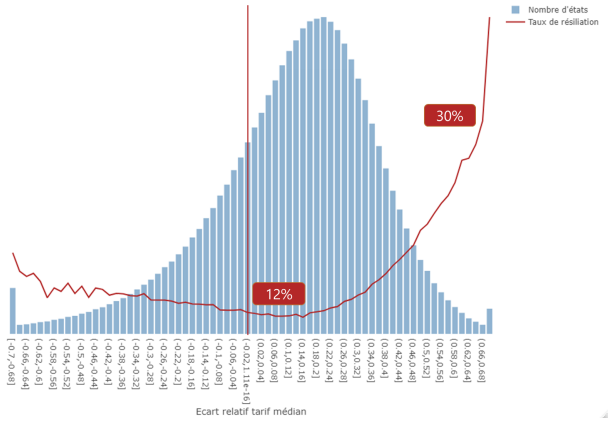


FIGURE 2.15 – Taux de résiliation en fonction de l'écart relatif au tarif médian

2.6 Synthèse des bases de données obtenues

Finalement, deux bases distinctes sont obtenues à l'issue du travail de gestion des anomalies, de retraitement et de création de variables, présenté dans cette partie. Les deux axes de l'étude requièrent des données aux attributs spécifiques ce qui induit la nécessité de disposer de deux bases, préparées légèrement différemment. Les deux bases ont été nettoyées et corrigées de la migration comme détaillé dans la Section 2.2, ensuite des traitements distincts ont été appliqués aux données.

2.6.1 Résiliation à un an

Les traitements spécifiques à la résiliation sont explicités Section 2.3. Pour rappel, La modélisation de la résiliation à un an nécessite :

— que les états d'un même contrat se suivent afin d'observer avec justesse ses avenants et ses évolutions ;
— que les états aient une exposition, c'est-à-dire une durée d'observation par rapport à un référentiel annuel, aussi proche de 1 que possible ;
— de ne pas considérer les états encore en cours dont l'éventuelle reconduction, ou non, n'est pas observée.

La base obtenue pour la modélisation de la résiliation à un an comporte environ 2,7 millions de lignes (autrement dit d'états), réparties sur un peu moins de 700 mille contrats. Les contrats sont souscrits, et résiliés, entre 2012 et 2020. La Figure 2.16 représente, par an, les états qui ont été résiliés en rouge et ceux qui ont été reconduits en bleu.

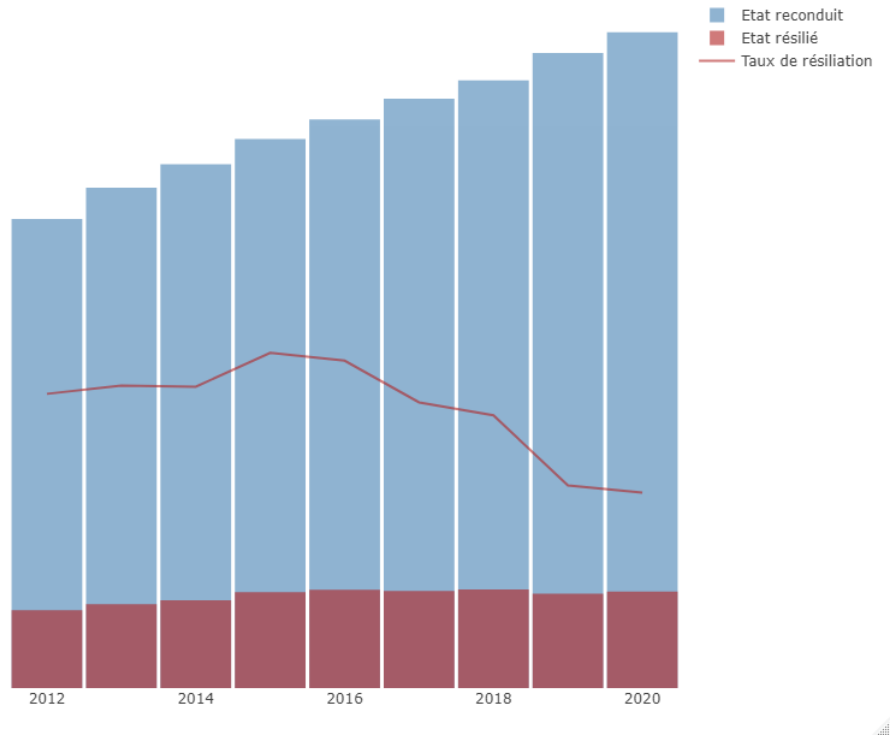


FIGURE 2.16 – Répartition des états reconduits et résiliés, taux de résiliation par année

2.6.2 Durée de vie a priori

Comme décrit Section 2.4, l'étude de la durée de vie a priori nécessite de disposer uniquement des caractéristiques des contrats lors de leur souscription. La base utilisée pour ce pan de l'étude comporte environ 650 mille affaires nouvelles, souscrites entre 2009 et 2021. La Figure 2.17 permet la visualisation des contrats souscrits par année. Les affaires nouvelles dont la durée de vie est connue, c'est-à-dire dont la résiliation a été observée, sont représentées en bleu. En revanche, les contrats dont seul un minorant de la durée de vie est observé sont représentés en orange.

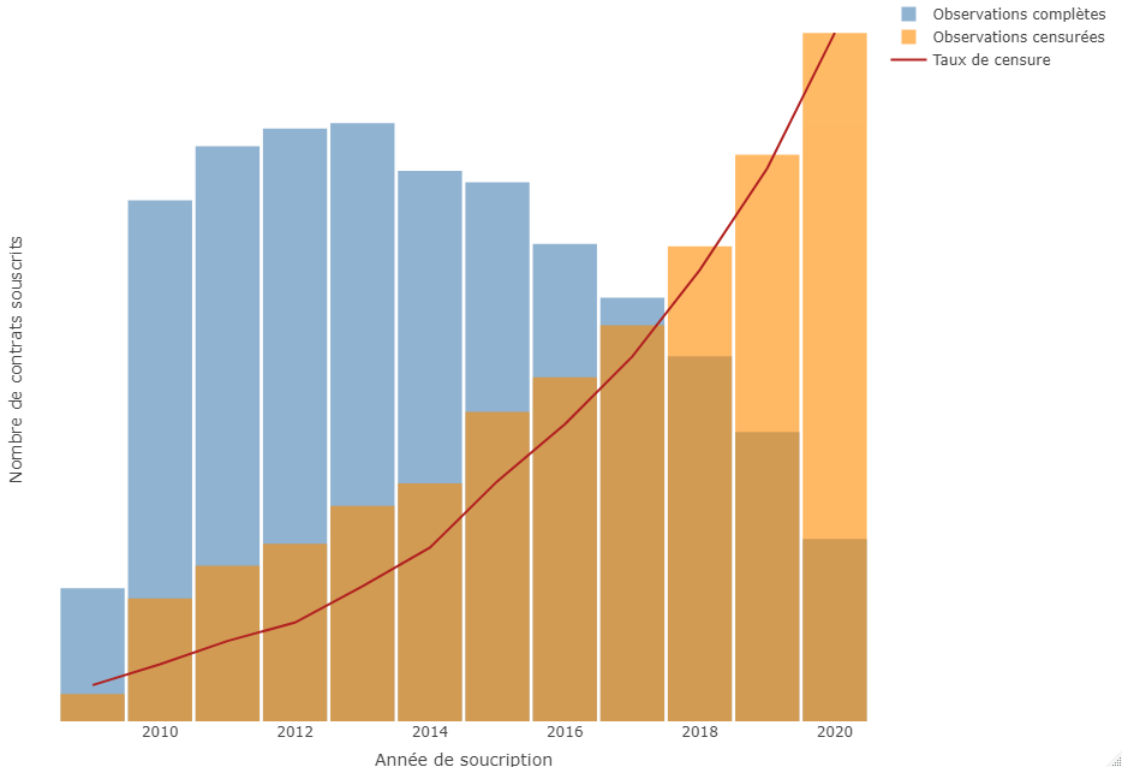


FIGURE 2.17 – Répartition des contrats souscrits, taux de censure par année

Chapitre 3

Probabilité de résiliation à un an

Sommaire

La décision prise par l'assuré, à chaque instant de la vie de son contrat, de résilier ou non l'accord qui le lie avec son assureur peut en partie s'expliquer par ses caractéristiques. Ainsi, il est possible d'appréhender son comportement à partir des informations dont l'assureur dispose. Ce chapitre s'attache à modéliser, par différentes méthodes de machine learning, les résultats de l'arbitrage réalisé par l'assuré. Dans un premier temps, le cadre théorique nécessaire est posé. Celui-ci comprend les éléments essentiels pour la mise en œuvre et la compréhension de la régression logistique et du modèle eXtrem Gradient Boosting (XGBoost). Le problème du choix des métriques d'évaluation sera discuté puis les métriques adaptées à la prédiction d'une telle probabilité seront présentées. Une fois cette formalisation effectuée, cela sera appliqué sur le portefeuille d'étude, où les différentes étapes de sélection des variables, d'évaluation et d'interprétation des modèles seront exposées. A l'issue de ce chapitre, un modèle de prédiction de la résiliation à un an est obtenu.

3.1 Le modèle linéaire généralisé

Les modèles linéaires généralisés (GLM), introduits en 1972 par Nelder et Wedderburn [28] proposent une extension au modèle linéaire classique. Les GLM sont communément utilisés en assurance IARD et permettent la modélisation de variables de fréquence, de sévérité ou encore de taux. Dans un premier temps, le cadre du modèle linéaire gaussien est posé. Ensuite, sa généralisation sera présentée, avant de détailler le cas particulier de la régression logistique.

3.1.1 Le modèle linéaire gaussien

Soit un échantillon d'observations de taille n, $(Y_i, (X_{i,j})_{j=1}^m)_{i=1}^n$ avec $Y_i \in \mathbb{R}$ et $(X_{i,j})_j \in \mathbb{R}^m$ où :
— Y_i, la variable à expliquer, est aléatoire ;
— $(X_{i,j})_j$, les m variables explicatives, sont déterministes.

Le modèle linéaire établit un lien entre Y_i et une transformation linéaire de $(X_{i,j})_j$:

$$Y_i = \beta_0 + \sum_j \beta_j X_j + \epsilon_i$$

avec :
— $(\beta_j)_{j=0}^m$, les paramètres du modèle, à estimer, appartiennent à l'espace $\mathbb{R}^{m+1}$ et sont supposés fixes ;
— $(\epsilon_i)_i$, les termes d'erreurs, aussi appelés résidus, sont supposés composer un ensemble de variables indépendantes et identiquement distribuées de loi normale centrée $\mathcal{N}(0, \sigma^2)$.

Sous forme matricielle, $Y = X\beta + \epsilon$. Les paramètres β sont estimés par méthode des moindres carrés par minimisation du terme $||Y - X\beta||^2$. Il peut être démontré que le vecteur de paramètres ainsi estimé $\hat{\beta}_{MC}$ dispose d'une expression explicite : $\hat{\beta}_{MC} = (X^T X)^{-1}(X^T Y)$.

Bien que facile à mettre en œuvre et à interpréter, le modèle linéaire gaussien n'est pas adapté à la modélisation de la probabilité de résiliation. Premièrement, la variable à expliquer est binaire : l'assuré peut résilier ou ne pas résilier. Or, les sorties d'un modèle linéaire ne sont pas forcément comprises dans l'intervalle $[0, 1]$. De plus, l'hypothèse de normalité des résidus ne peut être satisfaite par les données. Le modèle linéaire généralisé, permettant de relâcher cette hypothèse de normalité, est plus adapté au cas considéré.

3.1.2 Le modèle linéaire généralisé

Le modèle linéaire généralisé, proposé par Nelder et Wedderburn [28], repose sur trois composantes : une composante aléatoire, une composante déterministe et une fonction de lien. Une fois ces trois composantes définies, il sera possible d'estimer les paramètres du modèle.

Composante aléatoire

Les n variables aléatoires $(Y_i)_i$ sont supposées indépendantes et de distribution de probabilité appartenant à la famille exponentielle naturelle sans transformation. C'est-à-dire que la densité de Y peut s'écrire sous la forme :

$$f_Y(y, \theta, \phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} - c(y, \phi)\right\}$$

où :
— $a(.)$, $b(.)$ et $c(.)$ sont des fonctions différentiables, propres à la loi considérée. De plus, $b(.)$ est trois fois différentiable et sa première dérivée $b'(.)$, est inversible ;
— θ est le paramètre naturel de la loi, dont le support dépend de la loi considérée ;
— ϕ est le paramètre de dispersion.

De nombreuses lois usuelles appartiennent à la famille exponentielle : les lois normale, binomiale ou exponentielle peuvent être citées. Une des spécificités de cette famille de lois est que :

— $\mathbb{E}[Y] = b'(\theta)$;
— $\mathbb{V}[Y] = a(\phi)b''(\theta)$.

Composante déterministe

Le prédicteur linéaire, aussi appelé composante déterministe du modèle, est le vecteur $\eta \in \mathbb{R}^{m+1}$:

$$\eta = X\beta = \beta_0 + \sum_j \beta_j X_j$$

où les β_j représentent, comme dans le cadre du modèle linéaire gaussien, les paramètres à estimer.

Fonction de lien

La fonction de lien $g(.)$, établit une relation entre la composante linéaire déterministe η, et la composante aléatoire Y. Cette fonction est monotone, différentiable, inversible, et telle que :

$$\mathbb{E}[Y|X] = g^{-1}(\eta) = g^{-1}(\beta_0 + \sum_j \beta_j X_j)$$

Estimation des paramètres

L'estimation du vecteur de paramètres β et du paramètre ϕ se fait par méthode du maximum de vraisemblance. Les observations sont supposées indépendantes, ce qui permet d'écrire la vraisemblance du modèle :

$$\mathcal{L}(y; \beta, \phi) = \prod_i f_{Y_i}(y_i, \beta, \phi)$$

puis la log-vraisemblance :

$$\begin{aligned} l(y; \beta, \phi) &= \sum_i \ln(f_{Y_i}(y_i, \beta, \phi)) \\ &= \sum_i \left\{ \frac{\theta_i(\beta) y_i - b(\theta_i(\beta))}{a(\phi)} - c(y_i, \theta) \right\} \end{aligned}$$

où

$$\theta_i(\beta) = (b')^{-1} g^{-1}(\eta_i) = (b')^{-1} g^{-1}(x_i \beta)$$

Ainsi, les paramètres $\hat{\beta}_{EMV}$ et $\hat{\phi}_{EMV}$ sont obtenus par résolution du système d'équations différentielles :

$$\begin{cases} \frac{\partial l(y;\beta,\phi)}{\partial \beta_j} = 0, \quad \forall j \in \{1, ..., m\} \\ \frac{\partial l(y;\beta,\phi)}{\partial \phi} = 0 \end{cases}$$

Hormis dans le cas où $g(.)$ est la fonction identité, c'est-à-dire où le lien est canonique et où les équations se simplifient, il n'existe pas d'expression explicite des estimateurs $\hat{\beta}_{EMV}$ et $\hat{\phi}_{EMV}$. Ces derniers peuvent être déterminés par un algorithme itératif des moindres carrés pondérés (IRLS), décrit par Green [15].

Tests de significativité des variables

La p-valeur, attribuée à Fisher [11], quantifie la probabilité d'obtenir un effet égal ou plus extrême que celui observé si l'hypothèse nulle est vraie selon Gibbons [12]. Autrement dit, la p-valeur fournit une indication quantitative de preuve de rejet de l'hypothèse nulle : plus la p-valeur est faible, moins l'hypothèse nulle est probable. Une fois les paramètres $\hat{\beta}_{EMV}$ obtenus, de par les propriétés asymptotiques des modèles linéaires généralisés, il est possible de de réaliser le test d'hypothèses suivant :

$$H_0 : \beta_j = 0 \quad \text{contre} \quad H_1 : \beta_j \neq 0$$

La matrice d'information de Fisher $I(\hat{\beta}_{EMV})$ peut être calculée et ainsi l'hypothèse nulle est rejetée dans le cas où :

$$\frac{\hat{\beta}_{EMV}}{\sqrt{I(\hat{\beta}_{EMV})}} < q_{1-\frac{\alpha}{2}}$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$ et où le réel α est à fixer.

Hypothèse de relation linéaire

Les modèles linéaires généralisés reposent sur l'hypothèse forte que la variable à expliquer Y, est liée aux variables explicatives $(X_j)_j$, par une relation linéaire. Cependant, il arrive que des liens plus sophistiqués soient observés par les composantes. Le modèle linéaire peut s'avérer limité pour capter la subtilité de certaines relations. Dans le cas de relations complexes, une transformation $f(.)$ doit être appliquée à X, telle que la variable $Z = f(X)$ examine une relation linéaire avec Y. Quelques exemples peuvent être cités :

- $f(X) = X$: permet de retrouver le modèle linéaire premier ;
- $f(X) = \sum_{k=1}^{K} \beta_k X^k$: la variable est transformée en une fonction polynomiale, le modèle doit estimer pour chaque terme un coefficient β_k ;
- $f(X) = \sum_{k=1}^{K} \beta_k \mathbb{1}_{\{X \in I_k\}}$, où les I_k sont des intervalles disjoints de valeurs prises par X : autrement dit, les modalités de X sont regroupées en plusieurs classes.

Cette approche sera privilégiée pour les variables dont la dérivation est nécessaire lors d'une étape postérieure d'optimisation de programme.

Lors de relations particulièrement complexes entre X et Y, qu'une fonction du type de celles présentées précédemment ne saurait appréhender, il est possible d'avoir recours aux fonctions splines. Une spline est une fonction composée de plusieurs polynômes, généralement cubiques, qui se raccordent de manière régulière de sorte que la fonction ainsi définie soit continûment dérivable sur son support. Une telle fonction permet d'approcher des formes alambiquées.

3.1.3 Le cas particulier de la régression logistique binaire

Définition

La régression logistique binaire est couramment utilisée dans la modélisation de probabilités, notamment lorsque le modèle nécessite une grande interprétabilité tout en présentant une bonne performance. L'objectif de la modélisation du taux de résiliation se prête parfaitement à son utilisation. La régression logistique appartient à la famille des modèles linéaires généralisés où :

- Y suit une loi de Bernoulli $\mathcal{B}(\mu(x))$, où $\mu(x)$ représente la probabilité de résiliation de l'individu en fonction de ses caractéristiques X ;
- la fonction de lien appropriée est la fonction logit : $g(x) = \log(\frac{x}{1-x})$.

La fonction logit représentée Figure 3.1 est symétrique en $x = 0,5$ et est son support est $[0,1]$, ce qui rend cette fonction de lien adaptée dans le cadre de la prédiction de l'état binaire résiliation ou rétention.

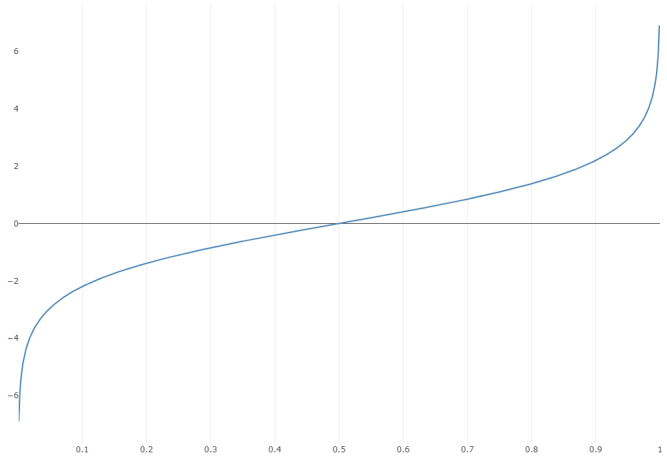


FIGURE 3.1 – Fonction logistique

Le cas particulier du support de Y permet, par la formule de Bayes, d'exprimer :

$$\begin{aligned}\mathbb{E}[Y|X] &= \mathbb{P}(Y=1|X)\\ &= \mu(X)\end{aligned}$$

et ainsi, en utilisant la fonction de lien, l'expression de la moyenne est obtenue :

$$\mu(x) = \mathbb{E}[Y|X] = \frac{\exp\eta}{1+\exp\eta}$$

Le vecteur de paramètres β est estimé par maximisation de la vraisemblance comme explicité dans le cadre général Section 3.1.2.

Odds ratios

La spécificité principale de la régression logistique réside dans l'interprétabilité des coefficients du modèle. Les odds ratios constituent un moyen simple et direct pour analyser les résultats. Basés sur la formule de Bayes, ils mesurent l'effet d'une variable continue, ou le contraste entre les effets d'une variable qualitative, sur la variable à expliquer.

Soient deux individus X_1 et X_2. L'odds ratio s'exprime de la manière suivante et se simplifie, toujours dans le cadre de la régression logistique binaire :

$$\begin{aligned}OR(x_1,x_2) &= \frac{\mathbb{P}(Y=1|X=x_1)/\mathbb{P}(Y=0|X=x_1)}{\mathbb{P}(Y=1|X=x_2)/\mathbb{P}(Y=0|X=x_2)}\\ &= \frac{\mu(x_1)/(1-\mu(x_1))}{\mu(x_2)/(1-\mu(x_2))}\\ &= \exp((x_1-x_2)\beta)\end{aligned}$$

Trois cas peuvent se présenter :
— $OR(x_1,x_2)=1$: la probabilité de résiliation est la même en présence des caractéristiques de l'individu X_1 et de celles de l'individu X_2 ;
— $OR(x_1,x_2)>1$: les caractéristiques présentées par l'individu X_1 tendent à augmenter la probabilité de résiliation par rapport aux caractéristiques de l'individu X_2 ;
— $OR(x_1,x_2)<1$: situation inverse à celle du point précédent.

Cette propriété des odds ratios dans le cas du modèle logistique binaire permet d'interpréter directement les coefficients β_j en sortie du modèle : un coefficient positif indiquera que la variable, en croissant, si elle est continue, ou sa modalité, si elle est catégorielle, a un impact négatif sur la rétention.

3.2 Le modèle XGBoost

Bien que la régression logistique soit, de par sa facilité d'interprétation, le modèle privilégié pour la prédiction de résiliation, il n'est pas toujours le plus performant. De nombreux modèles complexes peuvent être plus adaptés à des données dont la structure est non linéaire. Afin de mettre au défi le modèle linéaire, un modèle de machine learning sera également mis en place. Le modèle eXtrem Gradient Boosting (XGBoost), introduit par Chen et Guestrin en 2016 [6], repose sur une méthode de boosting appliquée à des arbres de classification. Il compte parmi les modèles de classification les plus performants et permet donc de fixer un élément de comparaison pour la régression logistique.

3.2.1 Définition

Ce modèle rassemble des modèles simples (des arbres de décision) entraînés conjointement. Cela conduit généralement à des modèles moins biaisés. La méthode de boosting sur laquelle repose XGBoost entraîne successivement les prédicteurs de façon adaptative, c'est-à-dire qu'elle donne plus d'importance aux observations qui ont été le moins bien prédites par le modèle précédent. Pour ce type de méthodes, des prédicteurs simples produisant une faible variance mais un biais important sont choisis, afin d'améliorer de façon conséquente leurs performances et de limiter le coût de calcul résultant de la phase de boosting. Il existe plusieurs méthodes de boosting qui diffèrent de par la façon dont elles intègrent les informations des prédicteurs et la façon dont elles les combinent. Pour sa part, le boosting par la méthode du gradient met à jour les valeurs des observations à chaque nouveau modèle construit et est adapté à toute fonction de perte.

Formellement, pour un jeu de données tel que présenté 3.1.1, un ensemble de modèles prédit Y_i à partir de K modèles simples tels que :

$$\hat{Y}_i = \phi(X_i) = \sum_{k=1}^{K} f_k(X_i), \qquad f_k \in \mathcal{F}$$

où $\mathcal{F}$ est l'espace des arbres de décision et f_k retourne le poids associé à la feuille que l'arbre k assigne à X_i. La fonction de perte $\mathcal{L}$, à minimiser par le modèle, comprend une partie dépendant de la mesure de perte choisie l et une autre comptant pour la régularisation :

$$\mathcal{L}(\phi) = \sum_i l(\hat{Y}_i, Y_i) + \sum_k \Omega(f_k)$$

où $\Omega(f) = \nu T + \frac{1}{2}\delta||\mathbf{w}||^2$, avec T, le nombre de feuilles de l'arbre f, et $\mathbf{w}$, le vecteur des poids associés aux feuilles de f. Cette fonction de régularisation pénalise la complexité du modèle et l'incite à limiter le nombre de feuilles et le vecteur de poids dans chaque arbre. Cette fonction est optimisée de façon additive en regardant, étape par étape, les constructions optimales pour l'arbre considéré.

Pour ce genre de modèles sophistiqués, un grand nombre d'hyperparamètres peuvent jouer sur la performance. Deux familles de paramètres se dégagent. La première concerne les paramètres de boosting tels que la profondeur maximale de l'arbre, ou le seuil du taux d'apprentissage. La deuxième famille concerne les paramètres d'apprentissage qui comprennent la fonction de perte et la métrique d'évaluation. Pour répondre à la problématique, la distribution utilisée est une loi de Bernoulli et la fonction de perte est une logloss qui permet de mesurer l'écart entre ce qui est observé dans les données et prédit par le modèle.

3.2.2 Interprétabilité

Bien que le XGBoost fasse partie des modèles dits *black box*, dans le sens où son fonctionnement est opaque, plusieurs outils de visualisation ont été introduits ces dernières années afin de mieux comprendre les prédictions fournies par un tel modèle. Ainsi, des graphiques d'importance pourront être

considérés. Ils permettent de visualiser les variables identifiées comme essentielles par l'algorithme. De plus, l'analyse des shap values (SHapley Additive exPlanations) [27], permet d'observer la contribution marginale moyenne de chaque variable dans la réduction de l'erreur de prédiction, et ce, pour chaque arbre construit. Les shap values permettent une interprétation locale d'un modèle complexe.

3.3 Métriques d'évaluation adaptées

Dans cette section, la notion de problème de classification est introduite. Bien que l'étude se porte sur ce type de problème, il sera expliqué en quoi les métriques d'évaluation usuelles ne sont pas adaptées au problème considéré. Enfin, les métriques d'évaluation retenues et utilisées par la suite seront définies.

3.3.1 Le problème de classification

En apprentissage statistique, le terme de classification regroupe l'ensemble des méthodes permettant de dégager des groupes à partir d'un échantillon de données. L'apprentissage peut être supervisé ou non supervisé [2]. Dans le premier cas, le modèle apprend à partir de données étiquetées dont on connaît par avance le groupe. Dans la deuxième configuration, le modèle cherche à dégager des groupes par lui-même, à partir des données présentées. Le cas de la modélisation de la résiliation est un cas d'apprentissage supervisé de classification. L'indicatrice de résiliation est la variable d'intérêt et cette dernière est renseignée dans les données. Le modèle est entraîné sur une partie des données (généralement 70% de l'échantillon), et est testé sur les données restantes. Le modèle cherche à classer les individus de la base d'entraînement de la manière la plus correcte possible.

Une fois le modèle choisi — les éventuels paramètres de ce dernier optimisés — et entraîné sur une partie du jeu de données, il reste à étudier la performance de la prédiction établie. Généralement sur l'ensemble de test, plusieurs mesures sont calculées afin de jauger de la qualité de la prédiction. Naturellement, ces dernières vont chercher à établir si le modèle permet d'identifier correctement les classes des individus. Bien que nous soyons dans le cas d'un problème de classification supervisée, où le modèle cherche à prédire si un individu, au vu de ses caractéristiques, est enclin à résilier ou non, attribuer aux individus des catégories binaires n'est pas l'objectif premier de l'étude. Estimer un taux de résiliation en fonction des caractéristiques de l'assuré et de son contrat est plus pertinent dans ce cadre. L'intérêt de l'étude tend alors à se porter sur le bon ordonnancement des probabilités prédites, et non sur la classe éventuellement associée à la prédiction. Les métriques d'évaluation utilisées devront considérer cet objectif.

3.3.2 Métriques usuelles et limites

Lors de l'étude de la performance d'un modèle de classification, certaines métriques sont très utilisées. Par exemple, l'accuracy indique le pourcentage d'individus bien classés. La matrice de classification, plus informative que l'accuracy, est également très répandue. Chaque colonne de la matrice correspond à une classe réelle et chaque ligne à une classe estimée. La cellule (i, j) contient le nombre d'observations de classe j classées dans la classe i. La matrice Tableau 3.1 illustre à quoi ressemblerait une telle métrique dans le cas de la prédiction de la résiliation. Aussi appelée matrice de confusion,

Classe prédite \ **Classe réelle**	**Pas de résiliation**	**Résiliation**
Pas de résiliation	Vrais négatifs (VN)	Faux négatifs (FN)
Résiliation	Faux positifs (FP)	Vrais positifs (VP)

Table 3.1 – Matrice de classification

elle permet de détecter rapidement si le modèle classe les individus correctement. Dans le cas de la

classification binaire notamment, elle donne lieu à l'observation du nombre de faux positifs et de faux négatifs.

Certaines métriques, plus précises, se penchent sur les observations positives, souvent critiques dans un modèle. Les plus communes sont le recall et la précision. Le recall (= VP/(FN+VP)) représente la proportion d'individus identifiés correctement comme positifs parmi tous les individus effectivement positifs. La précision (=VP/(FP+VP)), représente la proportion d'individus identifiés correctement comme positifs parmi tous les individus classés comme positifs par le modèle. Pour analyser correctement un modèle, il faut évaluer conjointement la précision et le recall. Le score F1, moyenne harmonique de la précision et du recall, est un indicateur prenant en compte les deux concepts. Plus le F1 score est proche de 1, plus le modèle est performant.

Toutes ces métriques considèrent la classification, en termes de groupe prédit, des individus. Particulièrement, elles demandent de définir un seuil de Bayes, seuil compris entre 0 et 1, et à partir duquel une observation est classée comme étant positive selon sa probabilité prédite par le modèle. Censé déterminer une frontière claire entre les différentes classes, il sera abordé dans la suite de la difficulté à établir ce seuil. De plus, le jeu de données est déséquilibré : le groupe des résiliés est clairement sous représenté. Un modèle qui classe toutes observations dans la catégories des états renouvelés, bien que inefficace, aurait une très bonne accuracy, proche de 86% (le reste étant des individus résiliés). Le jeu de données pourrait être artificiellement rééquilibré par des méthodes dites d'over ou d'under sampling, mais cela conduirait à des probabilités de résiliations prédites bien trop élevées. Cela appuie la décision de ne pas utiliser les métriques citées. En conclusion, les métriques définies dans cette partie, bien qu'étant d'usage, ne pourront pas être représentatives de la qualité des modèles dans le cadre de l'étude : des mesures en partie basées sur l'ordonnancement des probabilités doivent être considérées.

3.3.3 Métriques d'évaluation adaptées

Afin de pallier les défauts que présentent les métriques présentées Section 3.3.2, cette partie s'attache à définir les mesures d'évaluation des modèles qui seront adoptées dans la suite de l'étude.

Analyse des prédictions par segment

La qualité des modèles sera principalement jugée par l'étude, segment par segment, de la prédiction faite par le modèle en comparaison avec ce qui est réellement observé. Il sera représenté graphiquement pour chacune des variables, les taux de résiliation observés et prédits sur chacune de ses modalités. Cette méthode sera privilégiée car elle permet notamment de s'assurer de la cohérence des prédictions, tant en termes de moyenne que d'évolution, de comparer l'exactitude de différents modèles et de détecter les profils sur lesquels le modèle nécessiterait d'être mieux ajusté.

Densité des probabilités prédites

Les modèles associent à chaque individu une probabilité de résiliation, il est alors possible de tracer les densités des probabilités prédites pour les contrats effectivement résiliés d'une part, et pour ceux retenus d'autre part. Un bon modèle arrive à capter les distinctions qui s'opèrent entre un état où une résiliation a lieu d'un état renouvelé. Dans ce cadre, les probabilités prédites sur les états résiliés seront nécessairement plus élevées, en moyenne, que sur les états non résiliés. En revanche, ces densités ne sont pas disjointes et empêchent, de ce fait, la détermination d'un seuil de Bayes : il existe en effet des profils qui détiennent une rétention plus forte, pour lesquels les états résiliés disposeront d'un taux de résiliation plus faible que les profils à rétention faible, sur les états non résiliés. En d'autres termes, la comparaison s'effectue entre les états résiliés et non résiliés d'une même catégorie d'assurés pour identifier le moment clef de la résiliation et ses causes (avenant, évolution de tarif et des caractéristiques d'assuré, période de rétention moyenne atteinte, etc.). Un exemple de très bon modèle est présenté Figure 3.2 : le décalage des densités démontrent bien que le modèle réussit à discriminer les états où une résiliation a eu lieu.

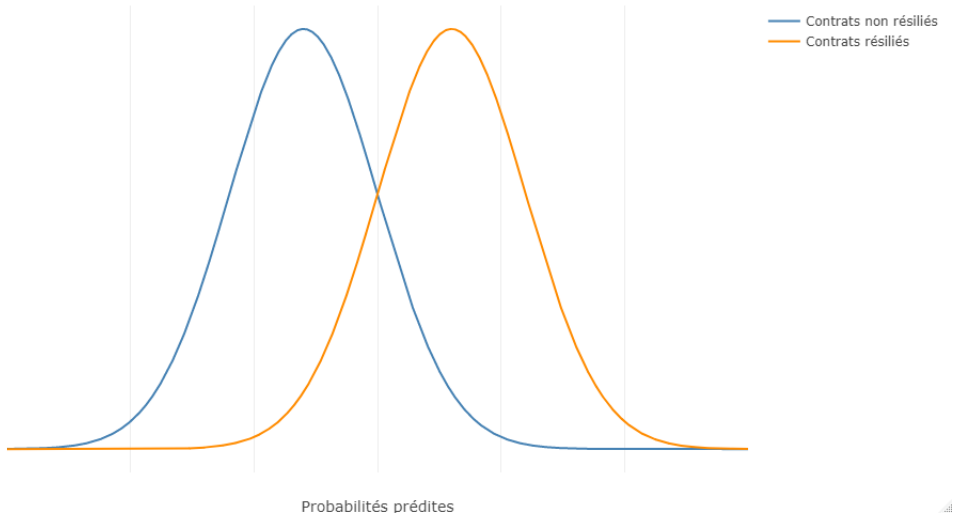


FIGURE 3.2 – Exemple de densités de probabilités prédites

Area under the curve (AUC) et coefficient de Gini

L'AUC est une mesure de performance de modèle comprise entre 0 et 1, qui correspond à l'aire sous la courbe ROC (receiver operating characteristic). La courbe ROC est formellement introduite en 1953 par Woodward [34], lors de travaux sur l'évaluation de la précision de radars. Elle permet de représenter graphiquement la capacité d'un classifieur binaire, en faisant varier le seuil de Bayes. En abscisse est représenté le taux de faux positifs, et en ordonnée le taux de vrais positifs. Plusieurs exemples de courbes ROC sont représentés Figure 3.3. En noir est tracée la droite $y = x$, elle correspond à un modèle complètement aléatoire. Plus la courbe se rapproche du point de coordonnées $(0, 1)$, meilleur est le modèle. Ainsi, dans l'exemple graphique proposé, le modèle dont la courbe ROC est représentée en rouge est le plus performant.

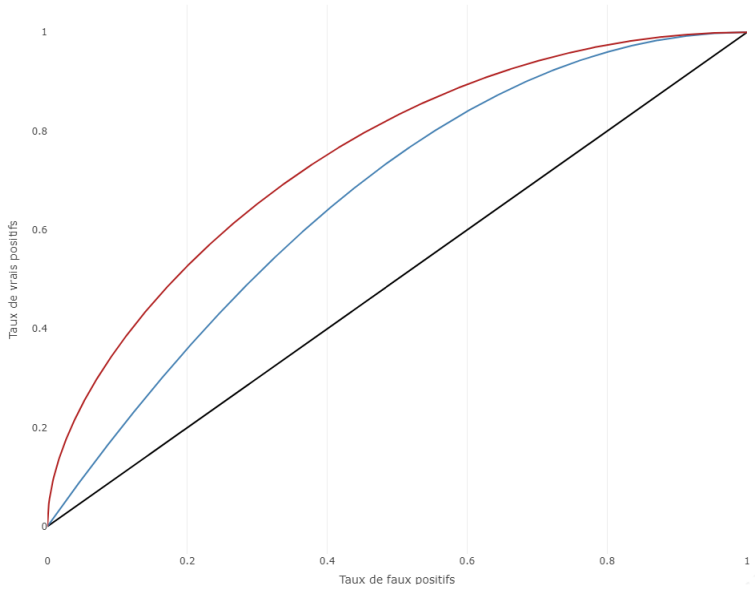


FIGURE 3.3 – Exemples de courbes ROC

L'AUC mesure en une seule valeur ce que la courbe ROC permet de visualiser. Un modèle qui classe les individus aléatoirement aura une AUC proche de 0,5, et plus un modèle sera performant plus son

AUC se rapprochera de 1. Le coefficient de Gini [26] est construit sous le même principe que l'AUC. Ce dernier mesure l'inégalité des probabilités prédites entre les deux classes. Les deux métriques sont liées par la relation :

$$\textbf{Gini} = 2\textbf{AUC} - 1$$

Ainsi, le Gini, défini sur l'intervalle $[-1, 1]$, correspond à un ajustement de l'AUC tel qu'un modèle aléatoire obtienne un score de 0 et qu'un modèle parfait obtienne un score de 1. Mieux le modèle différencie les individus des deux classes et plus l'indice de de Gini sera proche de 1.

Mean absolute error (MAE) et logloss

Finalement, des métriques telles que la MAE et la perte log permettront d'affiner la comparaison des modèles. Ces deux mesures sont des métriques de perte que le modèle doit chercher à minimiser. L'erreur moyenne absolue correspond à la moyenne des erreurs absolue de la prédiction :

$$\textbf{MAE}(\hat{y}_i, y_i) = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

La perte log est définie par :

$$\textbf{Logloss}(\hat{y}_i, y_i) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

3.4 Application sur le portefeuille

Le Chapitre 2 a permis d'introduire les données utilisées, de détailler les corrections qui ont été appliquées à la base et finalement, Section 2.6.1, de présenter, une fois traitée, la base employée. Les parties précédentes de ce Chapitre 3 ont posé le cadre théorique nécessaire à la réalisation d'un modèle de prédiction de la probabilité de résiliation à un an. Cette dernière section s'applique à la construction, sur le portefeuille d'étude, de modèles de prédiction à partir d'une régression logistique et d'un XGBoost. Dans un premier temps, les analyses exploratoires, ainsi que la sélection des variables explicatives seront exposées. Ensuite, la calibration du modèle linéaire généralisé sera présentée avant d'étudier les différentes évaluations de performance des modèles. Finalement, les résultats en sortie des modèles seront interprétés.

3.4.1 Analyse exploratoire

Une fois les données nettoyées et corrigées, il est possible par une étape de visualisation statistique, de mieux envisager les profils d'assurés prédisposés à la résiliation. Une analyse du taux de résiliation, modalité par modalité de chacune des variables, permet une première compréhension des comportements des clients.

Caractéristiques des assurés

Figure 3.4 sont représentés les taux de résiliations observés sur le portefeuille, en fonction de l'âge des assurés et de leurs catégories socioprofessionnelles. Une tendance très nette se dégage autour de l'âge. Les assurés de 18 ans résilient en moyenne moins que le reste de la population. Cependant, dès 19 ans et jusqu'à 40 ans environ, les résiliations sont fréquentes. Les jeunes actifs, entre 20 et 30 ans, enclins à de nombreux changements de situation, constituent de loin la population la plus risquée en termes de résiliation. Le taux de résiliation est décroissant de 25 à 75 ans, avant de recroître sur les profils plus âgés. Cette visualisation permet également de faire un choix sur la transformation appliquée à l'âge, en amont du modèle linéaire. Ce processus est expliqué Section 3.1.2, dans le cas de l'âge une fonction spline sera utilisée.

Les salariés, très représentés dans les données, ont un taux de résiliation avoisinant la moyenne générale. Les retraités, suivis des professions agricoles, des fonctionnaires puis des cadres ont des situations stables qui induisent des taux de résiliation plus faibles. Finalement, les artisans, commerçants et surtout les assurés sans profession, dont les étudiants font partie, sont les assurés qui résilient le plus.

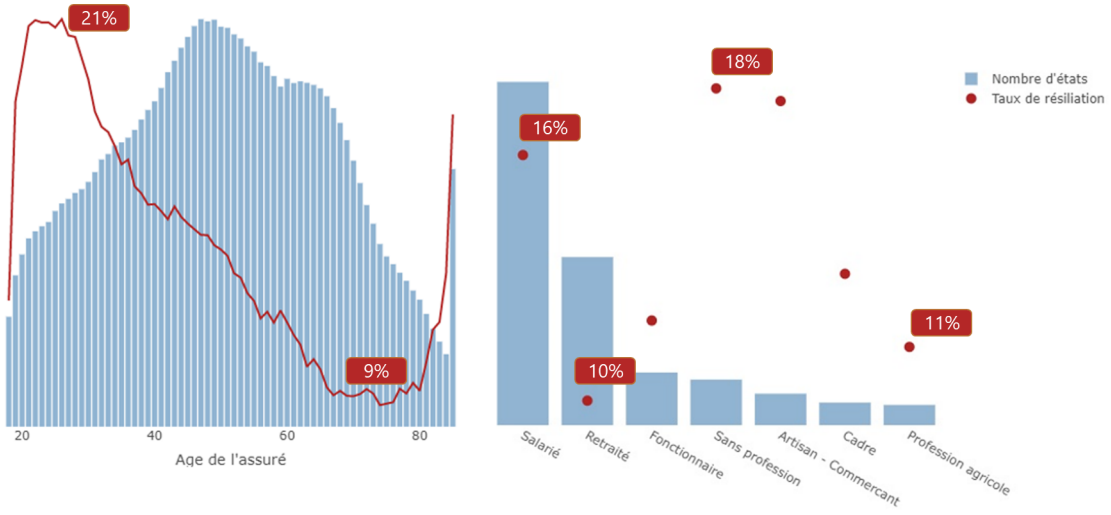


FIGURE 3.4 – Résiliation en fonction de l'âge et de la catégorie socioprofessionnelle

L'ancienneté du permis de conduire et l'indicatrice de conducteur novice sont également disponibles, cependant, étant très corrélées à l'âge, elles n'apportent pas plus d'information.

Caractéristiques du véhicule

Les caractéristiques du véhicule assuré présentées Figure 3.5 constituent également une source d'information concernant les facteurs de résiliation. Les taux de résiliation passent du simple au triple entre un véhicule neuf et un véhicule de 15 ans. Un véhicule âgé tendra à être remplacé, et par la même occasion, le client peut se questionner sur son contrat actuel et trouver plus attractif de changer d'assureur, ce qui explique en partie la forte résiliation observée sur ces segments. A partir d'un véhicule d'une vingtaine d'années, le taux de résiliation observe une décroissance. Les expositions de ces segments sont plus faibles, ce qui induit une confiance limitée dans ces résultats. La forme prise par les taux de résiliation sur cette variable induit l'utilisation d'une fonction spline dans la régression logistique.

L'étude des taux de résiliation, observés en moyenne, selon groupe SRA (Sécurité et Réparation Automobile), présente des résultats moins tranchés : une différence maximale de moins de 2% est observée entre les deux taux de résiliation les plus distants. Les conducteurs de véhicules en entrée de gamme ne sont pas les mêmes que les propriétaires de véhicules haut de gamme, les tarifs et les marges appliqués à ces contrats diffèrent également. Autrement dit, une analyse plus poussée, croisant les caractéristiques des assurés et de leur contrat, au groupe SRA, permettrait une meilleure compréhension de l'effet pur porté par cette variable. Bien qu'en moyenne, les comportements de résiliation ne semblent pas être liés au groupe SRA du véhicule assuré, il ne peut être conclu que cette variable ne porte pas d'information pertinente.

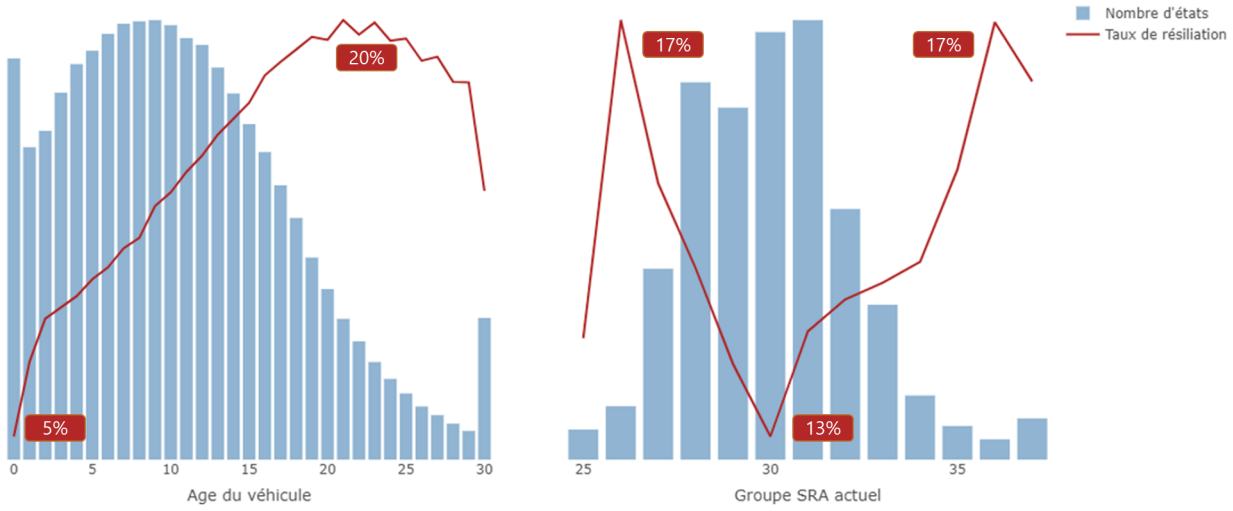


FIGURE 3.5 – Résiliation en fonction des caractéristiques du véhicule assuré

L'analyse des taux de résiliation en fonction des autres caractéristiques concernant le véhicule ne permettent pas de dégager plus d'information ou d'affiner les profils de risque. Cela se justifie notamment par les fortes corrélations observées entre les variables caractérisant le véhicule. L'étude des corrélations sera réalisée dans la Section 3.4.2 suivante.

Comportement de l'assuré

Le premier indicateur du comportement de l'assuré est le coefficient de réduction majoration. Aussi appelé coefficient bonus malus, ce dernier est dégressif et induit, suite à un comportement responsable de l'assuré qui ne subit pas de sinistre responsable, une réduction sur le tarif appliqué. Comme observé Figure 3.6, plus le coefficient de bonus malus est faible, c'est-à-dire plus la réduction appliquée est

importante, et moins les assurés résilient leur contrat. Le taux de résiliation est multiplié par 2,5 entre les assurés dont le coefficient est à 0,5 et les assurés dont le coefficient est à 1. Il est à noter que ce coefficient est très dépendant de l'âge de l'assuré, puisqu'il diminue avec le temps.

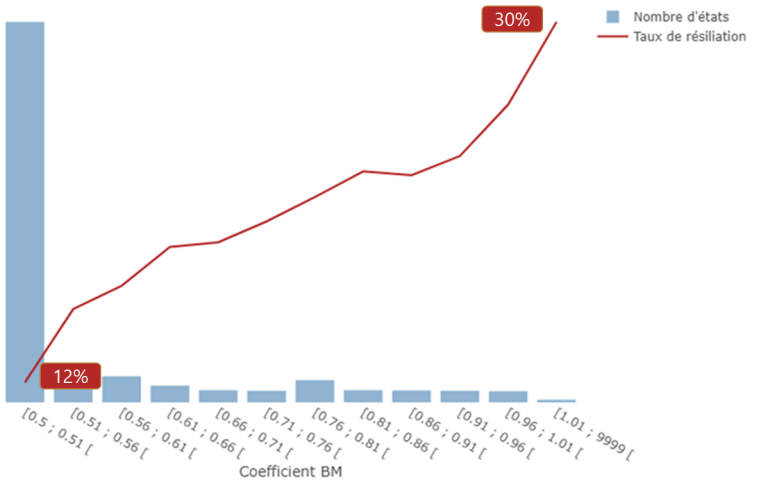


FIGURE 3.6 – Résiliation en fonction du coefficient de réduction majoration

Deux autres variables témoignent du comportement de sinistralité des assurés. La première est un bonus supplémentaire, et la deuxième est un taux de dégressivité sur la franchise. Représentées Figure 3.7, elles sont toutes deux liées au fait que l'assuré n'observe pas de sinistre responsable pendant une certaine période. Plus le bonus supplémentaire est dans une catégorie élevée, moins la réduction est conséquente, et moins les assurés sont fidèles. Ensuite, plus le taux de dégressivité appliqué sur la franchise est élevé et plus les assurés sont fidèles. Les taux de résiliation doublent entre la première et la dernière catégorie des deux variables. Ces deux variables renseignent du même phénomène que le coefficient de réduction majoration : plus un assuré observe des réductions importantes, liées à son comportement responsable, et plus il tend à être fidèle à son assureur. Cela soulève la pertinence de l'effort réalisé par l'assureur en termes de remise commerciale : ces assurés sont plus fidèles grâce à une politique qui récompense leur bon comportement. Encore une fois, ces réductions sont d'autant plus grandes que l'assuré est âgé, et comme analysé précédemment, un assuré est en moyenne d'autant plus fidèle qu'il est âgé. Ainsi, les effets constatés sur ces réductions peuvent aussi être expliqués par l'âge de l'assuré.

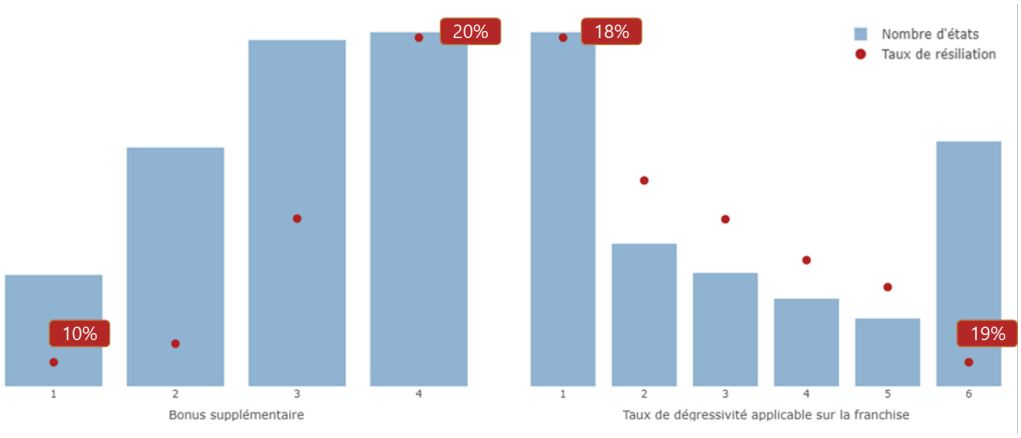


FIGURE 3.7 – Résiliation en fonction d'autres indicateurs du comportement

Le niveau d'équipement de l'assuré est également un facteur de différents comportements de fidélité. Il est d'usage qu'un assuré détenteur de plusieurs produits d'assurance au sein de la même compagnie y soit plus fidèle. Cela peut se vérifier sur l'indicatrice de présence d'un deuxième contrat automobile et sur le nombre de contrats, sur des produits différents, détenus par un même assuré. Un assuré qui bénéficie de plus de deux contrats différents verra ses chances de résilier divisées par 1,5 environ. Ces variables et les taux de résiliation associés à chaque modalité sont représentés Figure 3.8.

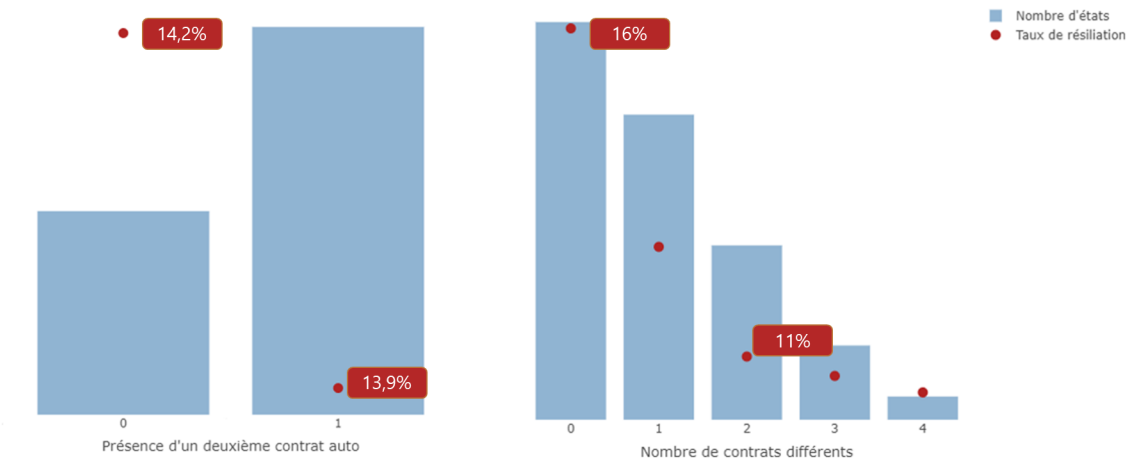


FIGURE 3.8 – Résiliation et détention de plusieurs produits

Utilisation du véhicule et niveau de couverture

L'usage du véhicule et le kilométrage effectué par l'assuré sont deux variables, Figure 3.9 qui renseignent sur l'utilisation que l'assuré fait de son véhicule. Un grand rouleur tendra à moins résilier, et les assurés qui font un usage uniquement privé ou uniquement professionnel de leur véhicule résilient moins que les autres.

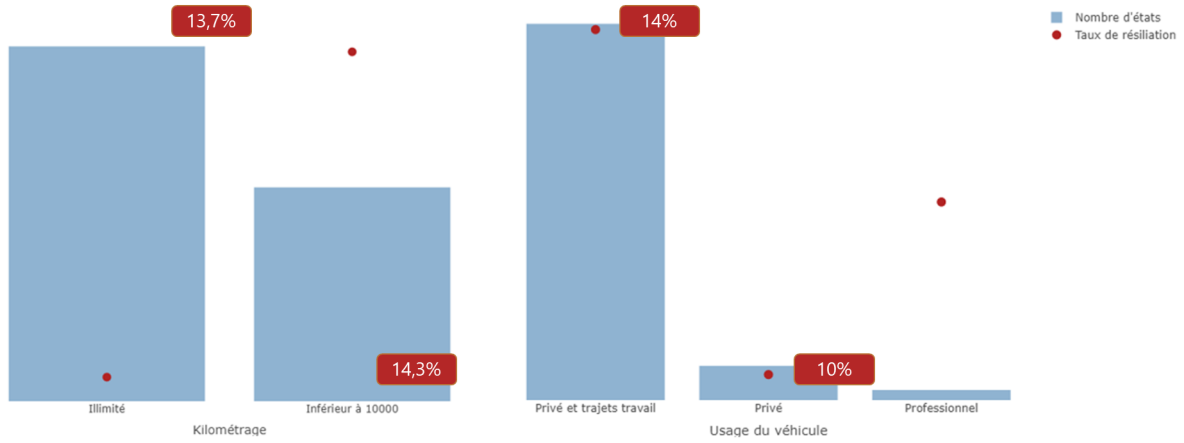


FIGURE 3.9 – Résiliation et utilisation du véhicule

La résiliation est nettement liée à la typologie de couverture choisie par le client. Figure 3.10, il est observé que plus le contrat choisi est complet, et moins les assurés résilient. Un assuré qui a choisi la formule 1 résiliera en moyenne trois fois plus qu'un assuré couvert par la formule 4.

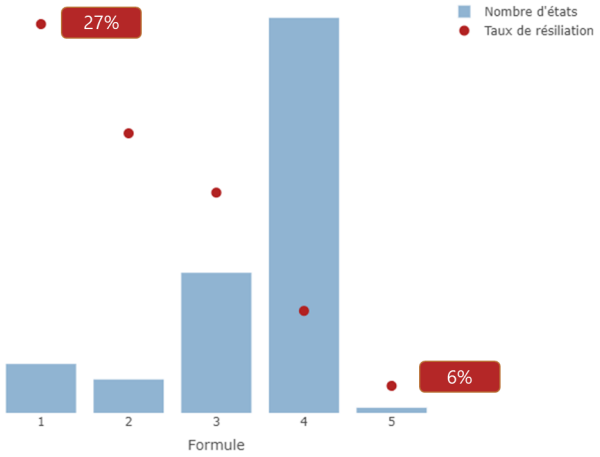


FIGURE 3.10 – Résiliation et niveau de couverture du contrat

Une analyse plus complète demande à être menée sur ces assurés couverts au tiers d'une part, et sur les assurés dont le kilométrage est limité d'autre part. Cela permettrait de comprendre quelles stratégies pourraient être mises en place par l'assureur, en vue d'améliorer son offre et d'augmenter la satisfaction de ces profils.

Tarif

Finalement, l'analyse exploratoire met en évidence que les comportements des assurés sont très dépendants aux différentes variables de tarif. Sont représentées Figure 3.11 les taux de résiliation en fonction du montant de la cotisation et de la marge. Les taux de résiliation en fonction du montant de cotisation observent une forme polynomiale, qui devra être prise en compte en amont de la régression logistique. Dans le cas de la marge, les taux évoluent, en croissant, à partir du moment où celle-ci est positive.

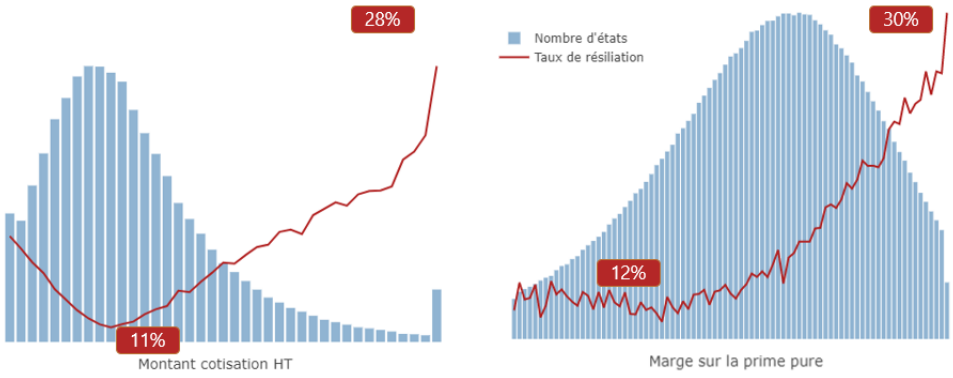


FIGURE 3.11 – Résiliation et tarif

3.4.2 Sélection des variables explicatives

La sélection des variables utilisées dans la modélisation est essentielle à la qualité de la prédiction. Réaliser la prédiction sur un jeu de variables parcimonieux, choisi avec soin, permet, par la suite, l'obtention de modèles plus justes et plus interprétables.

Analyse exploratoire

L'analyse, segment par segment, présentée Section 3.4.1, couplée à une expertise actuarielle, a permis de dégager un premier échantillon de variables potentiellement explicatives de la résiliation. Par exemple, les variables dont les écarts de taux de résiliation sont trop légers entre les différentes modalités ne seront pas intégrées aux modèles. De plus, des variables intéressantes, mais dont certaines modalités sont peu représentées dans le portefeuille sont exclues des modèles. Par exemple, il a été observé Figure 2.10 que la survenance d'un sinistre touchant aux garanties responsabilité civile ou vol, au cours de l'année $n-1$, augmente de 20% le risque de résiliation l'année n. Cependant, les expositions de ces segments sont faibles et ne pourront être captées correctement par la régression logistique employée. De ce fait, seule l'indicatrice de survenance d'un sinistre, toutes catégories confondues, sera conservée.

Analyse des corrélations

Ensuite, un travail d'analyse des corrélations permet d'identifier les groupes de variables très liées et dont l'information est redondante. Le coefficient de Pearson est une mesure de la corrélation linéaire entre deux variables. Formellement, soient X_1 et X_2 deux variables explicatives. Le coefficient de corrélation de Pearson, ρ_{X_1,X_2}, s'exprime :

$$\rho_{X_1,X_2} = \frac{cov(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}$$

Le coefficient ρ_{X_1,X_2} est compris entre -1 et 1 : plus la valeur absolue de ρ_{X_1,X_2} est proche de 1, et plus l'association linéaire entre X_1 et X_2 est forte. La Figure 3.12 permet une visualisation des corrélations entre un certain nombre des variables explicatives à disposition. Plus une cellule tend vers le bleu foncé, et plus les variables sont liées et évoluent dans le même sens. Lorsqu'une cellule tend vers le rouge foncé, les variables observent une corrélation négative, autrement dit, quand l'une augmente, l'autre diminue. Les cellules claires sont le témoin de faibles corrélations linéaires entre les variables.

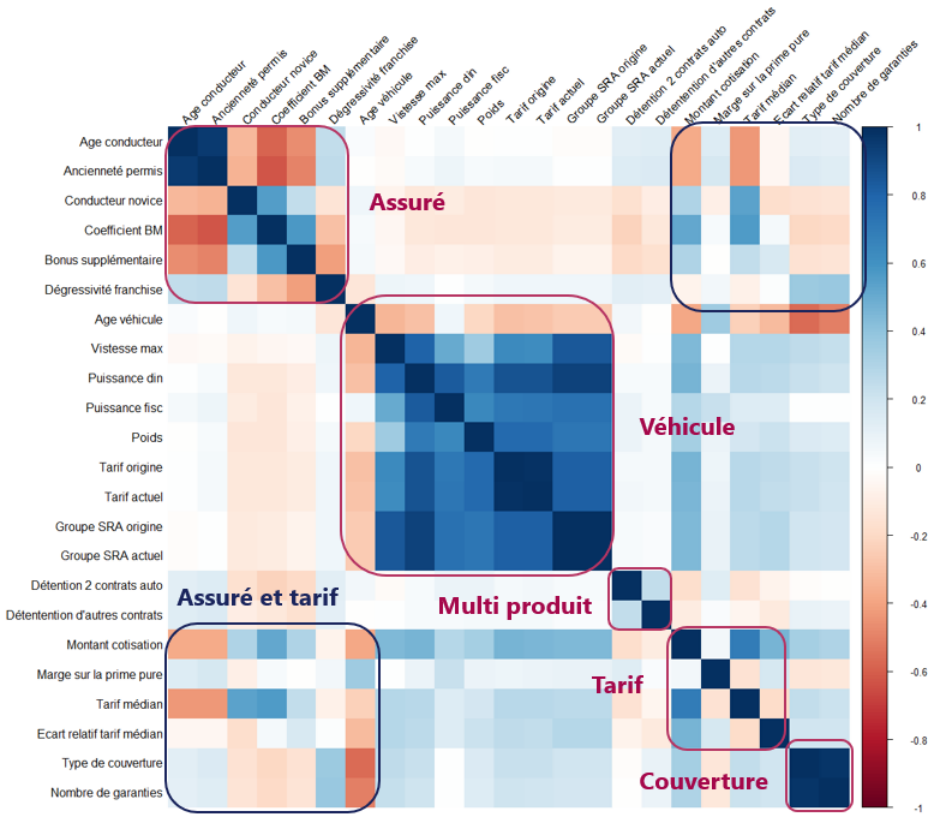


FIGURE 3.12 – Corrélations de Pearson

Par catégorie d'information, des groupes sont clairement identifiables :

— Les caractéristiques des assurés sont fortement liées. L'ancienneté du permis de conduire ou l'indicatrice de conducteur novice sont en redondance avec l'âge et ne seront donc pas considérées par la suite.
— Au centre de la figure, les informations du véhicule sont très corrélées, une sélection parmi ces variables devra être faite.
— Les variables de multi produit, de tarif et de couverture présentent des corrélation parfois élevées intra-classe.
— Finalement, des corrélations inter-classes marquées sont présentes entre l'assuré et le tarif notamment, mais également entre le véhicule et le tarif.

Sélection supervisée des variables

Une étape de sélection supervisée des variables est réalisée. Pour ce faire, un modèle XGBoost est entraîné sur l'ensemble du jeu de données. Ce modèle est très simple : la profondeur de ses arbres est fixée à 1. Le modèle doit identifier des variables significatives dans l'explication de la résiliation, mais sans créer de liens trop complexes que la régression logistique ne saurait reconnaître. La Figure 3.13 présente les variables les plus importantes retenues par le modèle. Le nombre de garanties et la modalité 1 de la formule, entourés en orange, figurent toutes deux parmi les variables les plus importantes. Cependant, comme observé lors de l'étude des corrélations, ces deux informations sont très liées. Ainsi, étant plus significatif, seul le nombre de garanties sera retenu dans la suite de la modélisation. Le même type de raisonnement aurait pu être mené pour faire un choix entre l'âge de l'assuré et l'ancienneté de son permis de conduire. Néanmoins, c'est l'âge de l'assuré qui sera conservé par la suite, bien que présentant une importance moins élevée que l'ancienneté du permis selon le XGBoost. Dans une démarche d'interprétabilité des modèles, ces deux variables renseignant foncièrement la même chose, il sera préférable de retenir l'âge du conducteur, dont les valeurs sont plus intuitives. De plus, l'âge du conducteur renseigne également sur l'étape de vie dans laquelle se situe l'assuré. Quand un assuré de 20 à 25 ans est probablement en sortie d'études, un assuré de 60 à 70 ans est susceptible de partir à la retraite. Il est intéressant de constater que le coefficient de réduction majoration n'apparaît pas dans cette représentation. Cela s'explique par le fait que d'autres variables, telles que l'âge de l'assuré, captent déjà l'expérience du conducteur et son comportement au volant. En revanche, le bonus supplémentaire et le taux de dégressivité de la franchise y figurent en bleu, ce qui souligne l'importance d'effectuer des plans d'action pour favoriser la rétention client.

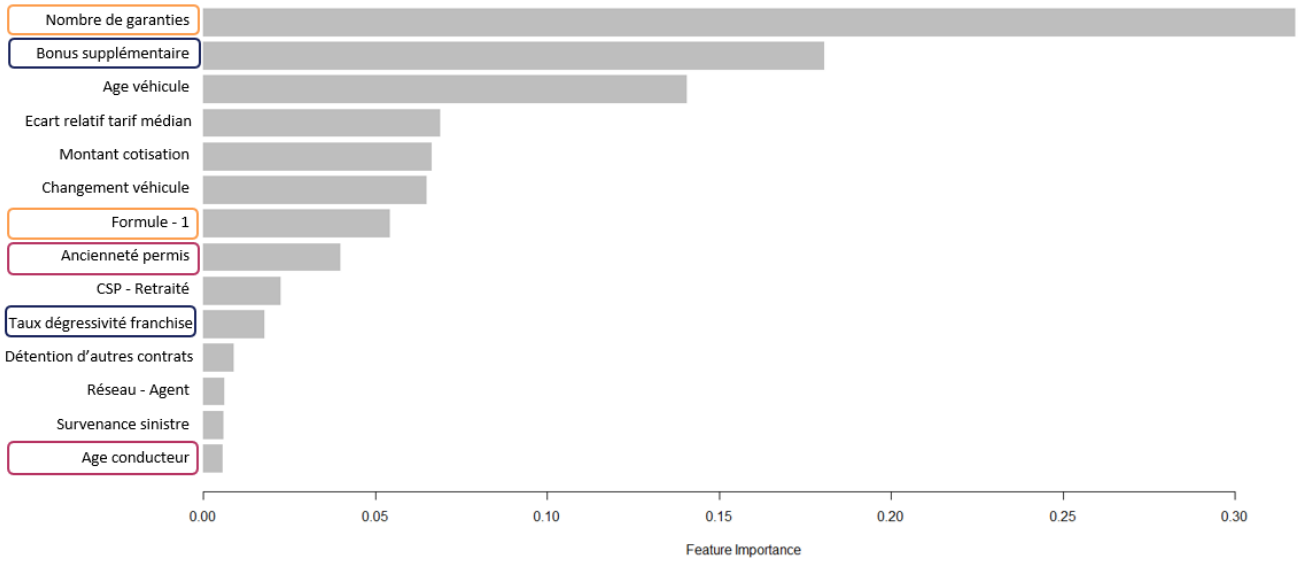


FIGURE 3.13 – Importance des variables - XGBoost

D'autres méthodes de sélection supervisée de variables auraient pu être utilisées. Par exemple, les régressions pénalisées telles que la régression Lasso (régression pénalisée d'ordre 1) ou les méthodes de type stepwise permettent également d'identifier un sous échantillon de variables explicatives dans le cadre d'un modèle d'apprentissage.

A l'issue de ces trois étapes (analyse exploratoire combinée au sens actuariel, corrélation et sélection supervisée), un premier ensemble de variables explicatives est dégagé. Les variables ainsi retenues sont disponibles Tableau 3.2.

Assuré - Véhicule	Comportement	Couverture	Contrat	Tarif	Avenant
Âge assuré	Bonus supplémentaire	Nombre garanties	Fractionnement paiements	Montant cotisation	Véhicule
CSP	Dégressivité franchise		Usage véhicule	Marge prime pure	Domicile
Âge véhicule	Autres produits d'assurance		Kilométrage	Tarif médian	Formule
Groupe SRA			Réseau (agent/courtier)	Écart relatif tarif médian	Sinistre
			Ancienneté		

TABLE 3.2 – Première sélection de variables

Sélection forward au sein de la régression logistique

Finalement, les variables du Tableau 3.2 sont ajoutées une à une au modèle sous la forme d'un algorithme qui pourrait être nommé *forward stepAUC*. Les variables sont ajoutées itérativement dans la régression logistique. Une variable est considérée comme pertinente dans l'analyse si l'ensemble de ces conditions sont respectées :

- l'AUC du modèle augmente à son ajout ;
- son coefficient est cohérent avec ce qui est observé dans l'étape d'analyse exploratoire ;
- son ou ses coefficients ont une p-valeur qui indique le rejet de l'hypothèse nulle ;
- son ajout ne vient pas perturber la cohérence des coefficients et la significativité des autres variables. Dans le cas contraire, un choix devra être fait entre les deux variables qui se font concurrence.

3.4.3 Calibration du modèle linéaire généralisé : l'exemple de l'âge des assurés

Une fois un jeu adéquat de variables explicatives établi, le modèle peut être entraîné puis testé. De par sa facilité d'interprétation et ses propriétés autorisant la dérivation de variables, ce qui est primordial dans le travail d'optimisation tarifaire, la régression logistique sera le modèle utilisé par la suite. Néanmoins, l'utilisation d'un tel modèle linéaire peut présenter certaines limites. Il sera moins en capacité de cerner certaines subtilités, quand d'autres modèles, plus complexes (tels que le XGBoost), de par leurs structures non linéaires, saisissent sans difficulté. Obtenir une modélisation satisfaisante à partir d'un modèle linéaire peut ainsi générer plus de traitements. Le travail mené pour obtenir des résultats cohérents sur la population des moins de trente ans est pris en exemple dans cette partie pour expliquer le type d'analyses supplémentaires qu'un modèle linéaire requiert.

Un modèle qui capte mal la résiliation des jeunes assurés

Le premier modèle linéaire entraîné présente des métriques d'évaluation correctes. Cependant, l'analyse des prédictions par segment, qui se révèle d'autant plus essentielle, met en évidence une faiblesse du modèle. La Figure 3.14 propose d'observer, en fonction de l'âge des assurés, les taux de résiliations observés, en bleu, et les taux de résiliation prédits, en noir. Le modèle ne capte ni la concavité, ni le niveau des résiliations sur la population des 18-30 ans. D'autres variables viennent manifestement perturber les prédictions de la régression sur cette population. Plusieurs analyses sont alors menées pour identifier la source et pallier cette incohérence.

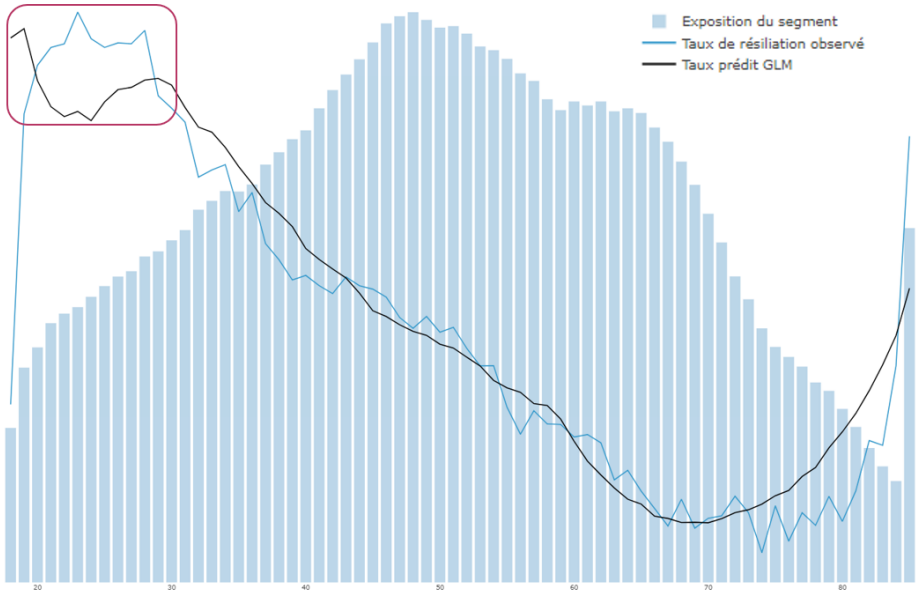


FIGURE 3.14 – Résiliation observée et prédite - Âge des assurés

Vers un modèle plus adapté

Il est observé que la résiliation des assurés de 18 à 25 ans est tirée vers le haut en grande partie par les travailleurs, les étudiants résiliant à peine plus que la moyenne. En outre, des analyses plus poussées mettent en évidence que les étudiants tendent à changer d'assureur une fois leur entrée dans la vie active. Ainsi, plus un étudiant est âgé relativement à sa classe d'âge et plus sa probabilité de résiliation augmentera. A partir de ces informations, une variable qui correspond au produit entre l'âge de l'assuré et l'indicatrice étudiant, habilite le modèle à mieux appréhender les comportements de cette classe d'âge. Une fois ces corrections appliquées, les taux de résiliation prédits sur les jeunes assurés Figure 3.15, bien que restant imparfaits, sont plus cohérents avec les observations.

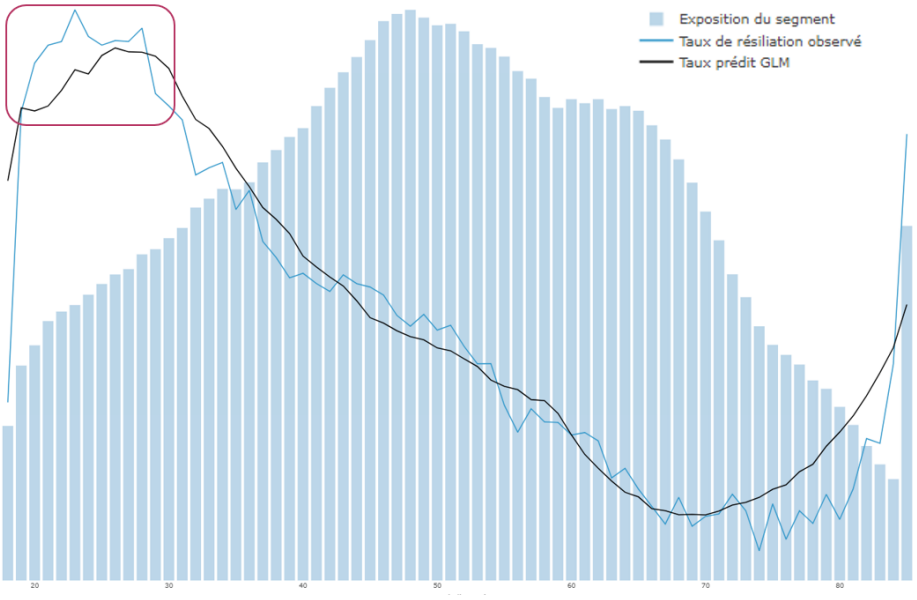


FIGURE 3.15 – Résiliation observée et prédite après correction - Âge des assurés

3.4.4 Évaluation des modèles

L'étude des métriques d'évaluation, puis des prédictions segment par segment, permettront de jauger de la qualité des modèles et de la validité de la sélection des variables. Pour rappel, deux modèles sont obtenus. La régression logistique permettra, de par son interprétabilité et sa dérivabilité, l'application du modèle de résiliation dans une démarche d'optimisation tarifaire. Parallèlement, un modèle XGBoost, réputé comme très performant, permet de juger de la qualité de la régression. Une optimisation des paramètres du XGBoost a été réalisée, mais elle reste parcimonieuse dans le souci que le modèle ne soit pas trop complexe. Pour les deux modèles, le sur apprentissage a été contrôlé et les performances sur l'échantillon de test et d'entraînement sont proches. Dans cette partie, seules les métriques obtenues sur l'échantillon de test sont présentées.

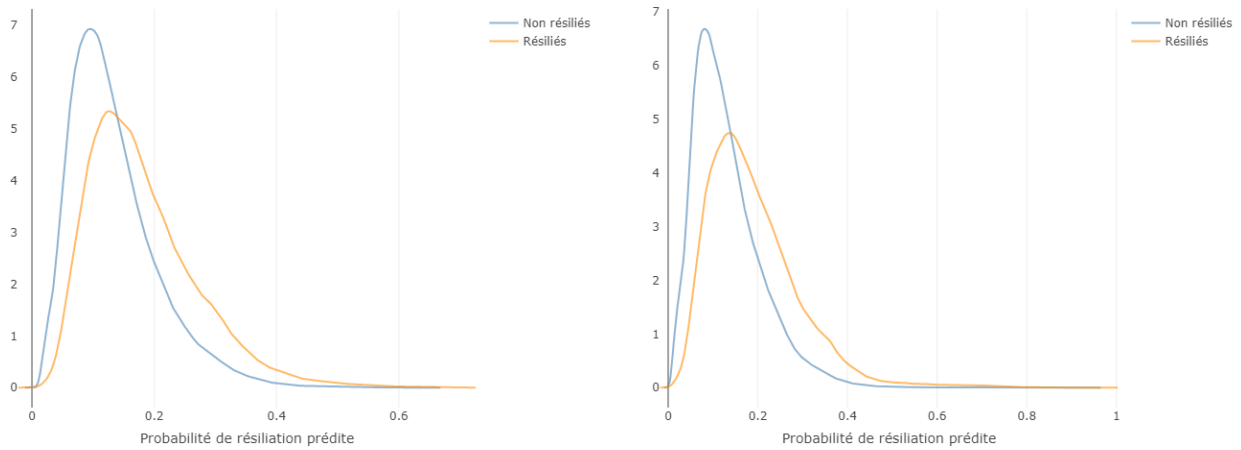
Métriques d'évaluation

Le Tableau 3.3 présente certaines des métriques d'évaluation des deux modèles, réalisées sur l'échantillon de test. L'AUC des deux modèles, proche de 0.7 est correcte. Le XGBoost présente naturellement de meilleures performances que la régression logistique, toutes métriques confondues, mais les écarts sont relativement faibles ce qui indique de la qualité du modèle linéaire.

	Gini	**AUC**	**MAE**	**Logloss**
GLM	0.337	0.669	0.229	0.382
XGBoost	0.376	0.688	0.226	0.376

TABLE 3.3 – Métriques d'évaluation - Échantillon de test

La Figure 3.16 propose d'étudier les densités des probabilités de résiliation prédites, pour les états effectivement résiliés, en orange, et pour les états renouvelés, en bleu. La visualisation de ces densités appuie le fait qu'il n'est pas possible de déterminer un seuil de Bayes efficace dans ce cadre, comme détaillé Section 3.3. Cependant, les distributions ne sont pas identiques et la différence entre les individus résiliés et non résiliés est présente. Le modèle XGBoost distingue légèrement mieux les deux groupes que le modèle linéaire.



(a) Modèle linéaire (b) Modèle XGBoost

FIGURE 3.16 – Densités des probabilités de résiliation prédites - Échantillon de test

La Figure 3.17 propose d'observer les courbes ROC obtenues sur l'échantillon de test, pour le GLM en bleu et pour le XGBoost en orange. Comme attendu après analyse de l'AUC, les deux courbes sont proches mais le XGBoost présente une courbe légèrement supérieure à celle obtenue par la régression

logistique. Le gain de performance obtenu avec le modèle XGBoost n'est pas suffisant pour justifier l'utilisation d'un modèle *black box*, ce qui rassure sur la qualité de prédiction obtenue par la régression logistique.

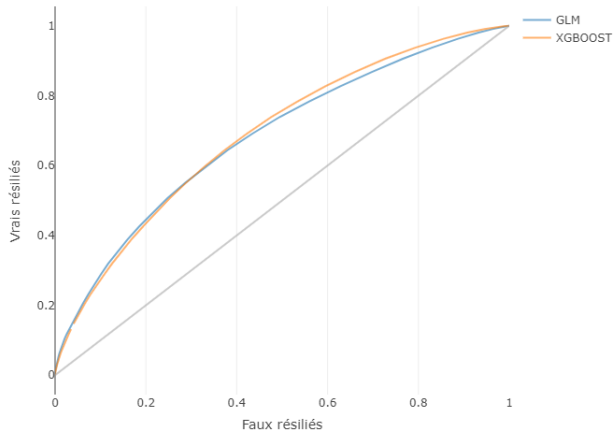


FIGURE 3.17 – Courbes ROC - Échantillon de test

Analyse segment par segment

L'analyse segment par segment permet de comparer les prédictions faites par le modèle de ce qui est réellement observé, et donc de s'assurer de la justesse de l'apprentissage. Les analyses sont menées sur les variables utilisées dans la prédiction, mais également sur celles qui n'ont pas été retenues, pour s'assurer que le modèle capte, au travers des autres variables, les informations nécessaires.

Concernant les caractéristiques des assurés, seules les variables de l'âge et de la CSP ont été retenues dans les modèles. Ainsi, comme observé Figure 3.18, les prédictions sur ces variables sont naturellement bonnes. En bleu sont représentés les taux de résiliation observés sur chacun des segments, en noir les taux prédits par la régression logistique et en rouge ceux prédits par le XGBoost. Bien que les prédictions sur l'âge de l'assuré aient été corrigées dans la Section 3.4.3, le modèle linéaire reste moins performant que le modèle plus complexe pour capter la dynamique des résiliations chez les assurés de moins de 25 ans et de plus de 75 ans. Au niveau de la catégorie socioprofessionnelle, le prédit et l'observé constatent de légers écarts sur les catégories *Artisan - Commerçant* et *Profession agricole*, qui sont les catégories les moins représentées dans les données. Le modèle linéaire s'avère plus proche de la réalité sur ces segments que le modèle complexe. Le Tableau 3.4 présente les coefficients obtenus pour chacune des catégories socioprofessionnelles, en prenant pour référence les salariés. Les coefficients sont cohérents.

Modalité	Coefficient	Ecart-type	p-valeur
Salarié	-	-	-
Artisan - Commerçant	1.6e-01	9.7e-03	$<$ 2e-16
Cadre	-6.2e-02	1.3e-02	8.3e-07
Fonctionnaire	-1.1e-01	8.8e-03	$<$ 2e-16
Profession agricole	-1.2e-01	1.4e-02	$<$ 2e-16
Retraité	-3.0e-01	9.8e-03	$<$ 2e-16
Sans profession	1.7e-01	1.1e-02	$<$ 2e-16

TABLE 3.4 – Coefficients régression logistique - Catégorie socioprofessionnelle

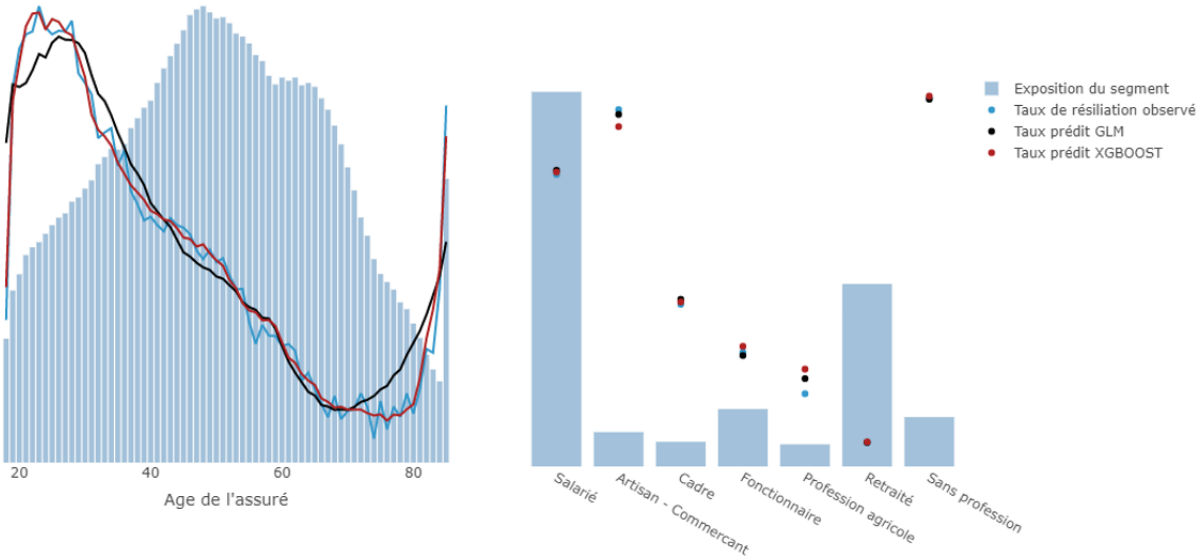


FIGURE 3.18 – Prédictions des probabilités de résiliation sur l'âge et sur la CSP des assurés

Ensuite, il est intéressant de remarquer Figure 3.19 que les modèles fournissent de très bonnes prédictions sur l'indicatrice de conducteur novice, alors même que ces derniers n'apprennent pas sur cette variable. Cela appuie le choix fait en amont de ne pas conserver cette variable dans la modélisation.

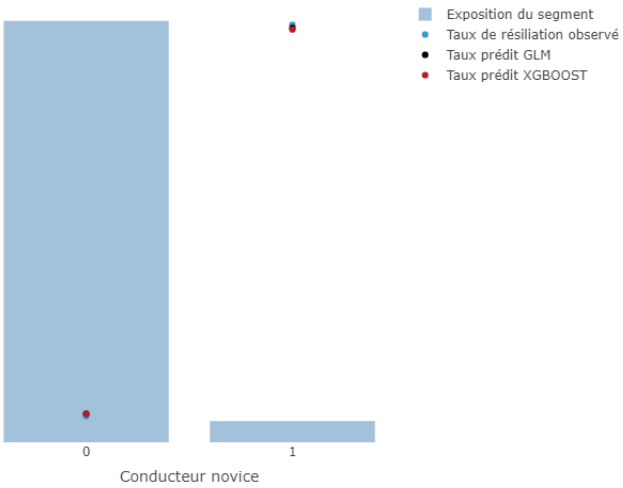


FIGURE 3.19 – Prédictions des probabilités de résiliation sur l'indicatrice de noviciat

A propos du véhicule assuré, seule la variable renseignant l'âge de la voiture a servi à entraîner les modèles. D'autres variables, très informatives des caractéristiques du véhicule, telles que le groupe SRA ou la classe de prix, auraient pu être intégrées aux modèles. Cependant, ces dernières n'ont pas été retenues suite à la phase de sélection de variables. Il peut être remarqué Figure 3.20 que les modèles restent néanmoins en capacité de fournir de bonnes prédictions sur les modalités de ces variables. Les prédictions sur l'âge du véhicule, disponibles Figure 3.21, indiquent que les deux modèles captent assurément la courbe observée.

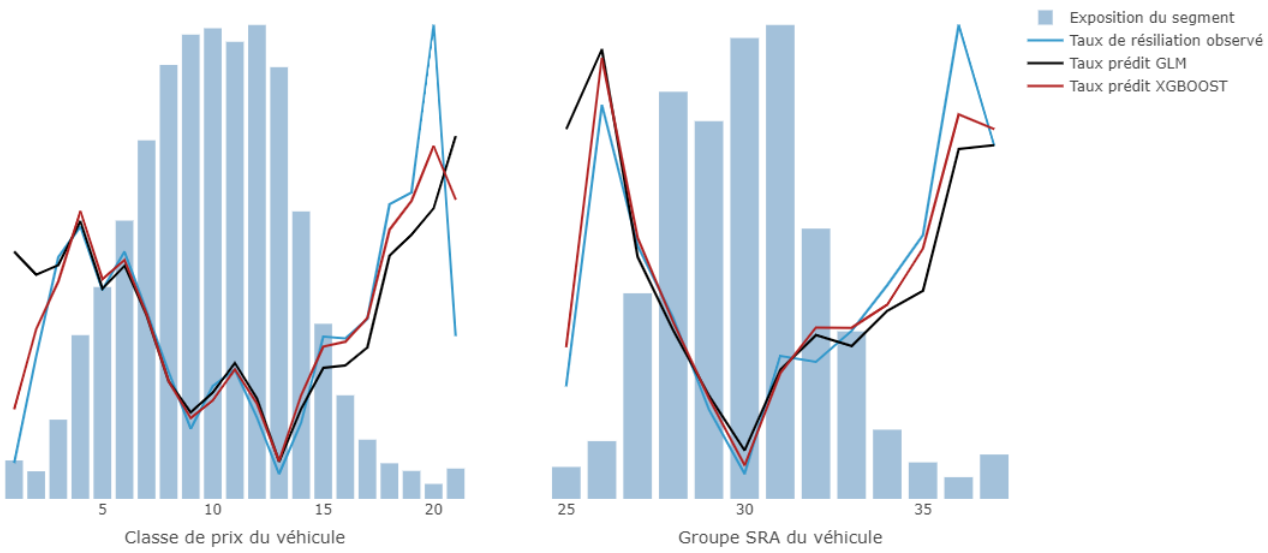


FIGURE 3.20 – Prédictions des probabilités de résiliation sur la classe de prix et le groupe SRA du véhicule

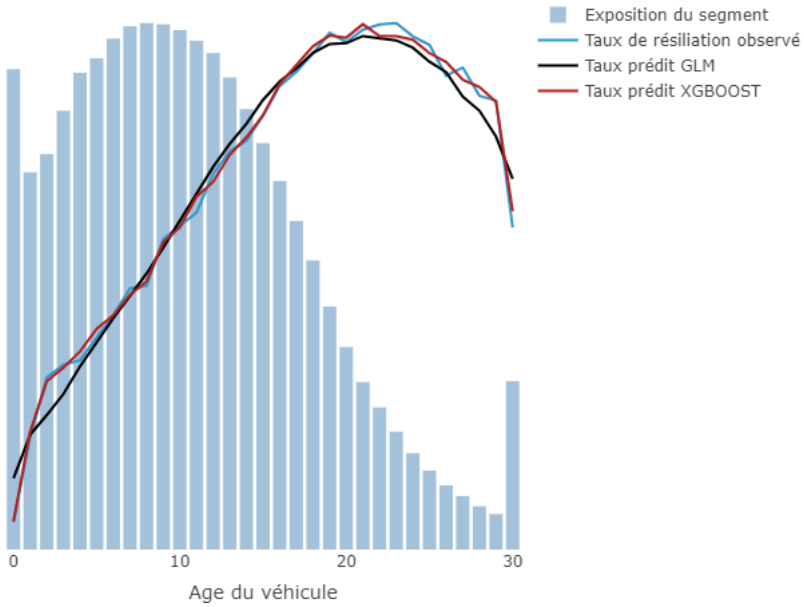


FIGURE 3.21 – Prédictions des probabilités de résiliation sur l'âge du véhicule

Le comportement responsable de l'assuré, bien que majoritairement capté par le coefficient de réduction majoration, n'a pas été retenu à la suite de la sélection par régression forward. Cela est en partie dû au fait que le CRM est très corrélé à l'âge de l'assuré et que d'autres variables, comme la présence d'un bonus supplémentaire ou d'un taux de dégressivité de la franchise, informent déjà sur le comportement de l'assuré. La Figure 3.22 confirme que cette variable était en redondance avec d'autres sur lesquelles les modèles ont pu s'appuyer pour prédire correctement les taux de résiliation par tranche de coefficient bonus malus.

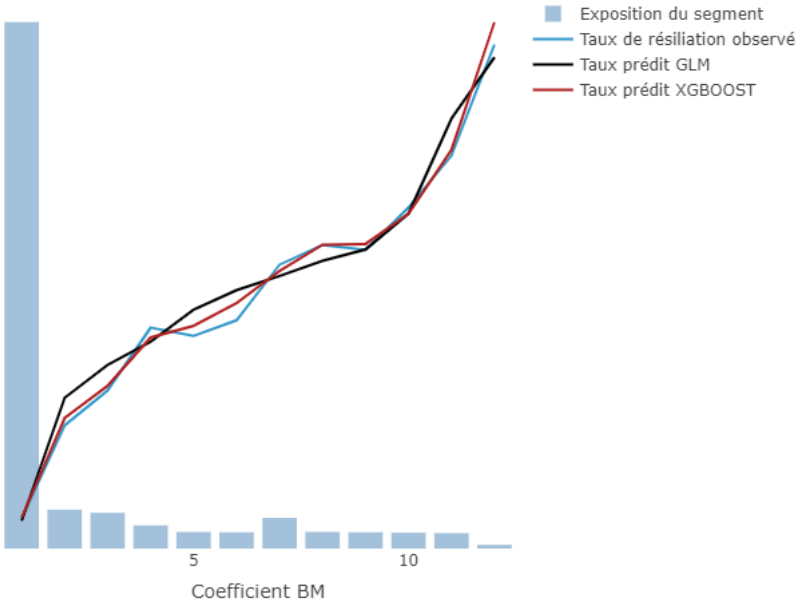


FIGURE 3.22 – Prédictions des probabilités de résiliation sur le coefficient de réduction majoration

Il en va de même avec le nombre de kilomètres réalisés par l'assuré : malgré le fait que cette information n'ait pas été prise en compte dans l'apprentissage des modèles, ces derniers fournissent des prédictions très satisfaisantes.

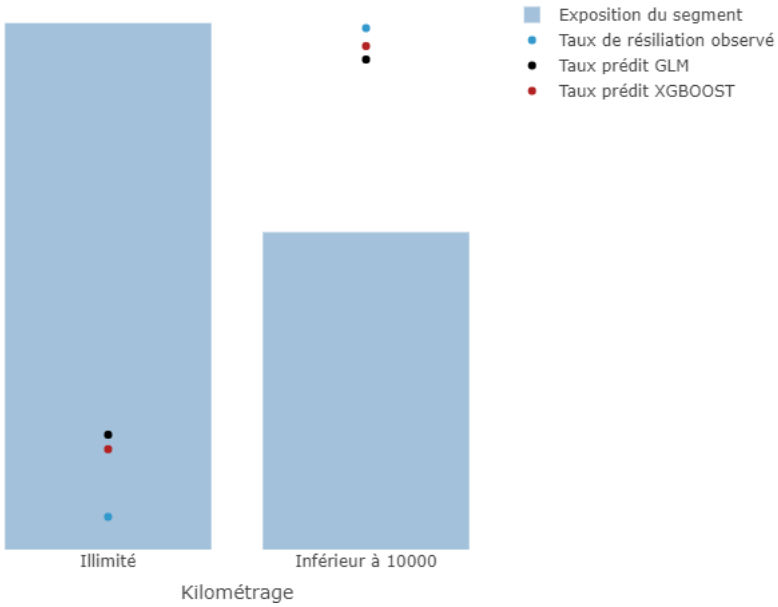


FIGURE 3.23 – Prédictions des probabilités de résiliation sur le kilométrage

Le type de réseau et le fractionnement des paiements sont des informations sur lesquelles les modèles ont été entraînés. Les prédictions sur les modalités de ces variables sont quasiment parfaites comme observé Figure 3.24. Un assuré ayant souscrit son contrat auprès d'un courtier tendra à plus résilier qu'un assuré étant passé par un agent. De plus, les assurés dont les paiements ne sont pas annuels résilient plus. Au-delà de la visualisation des prédictions, les coefficients de la régression logistique pour ces deux variables sont présentés Tableau 3.5.

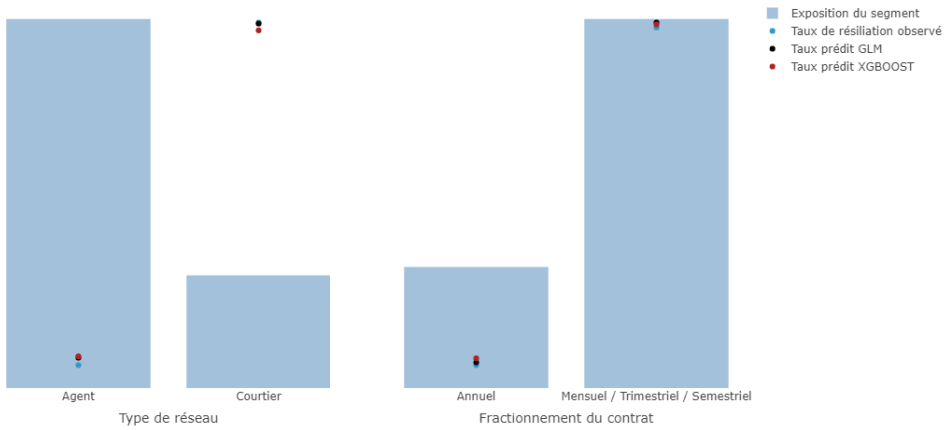


FIGURE 3.24 – Prédictions des probabilités de résiliation sur le type de réseau et le fractionnement des paiements

Modalité	Coefficient	Ecart-type	p-valeur
Réseau - Agent	-	-	-
Réseau - Courtier	1.2e-01	5.2e-03	$<$ 2e-16
Fractionnement - Annuel	-	-	-
Fractionnement - Autre	1.0e-01	5.7e-03	$<$ 2e-16

TABLE 3.5 – Coefficients régression logistique - Réseau et fractionnement

Le type de couverture pour lequel l'assuré opte est un indicateur essentiel de son comportement face à la résiliation, c'est d'ailleurs la variable retenue comme étant la plus significative par le XGBoost. Comme déjà remarqué dans l'analyse exploratoire, plus un assuré est couvert et moins il tendra à résilier en moyenne. Cela est capté par les modèles dont les prédictions sont observables Figure 3.25. La régression logistique reste moins performante que le XGBoost sur les segments les moins représentés. Le Tableau 3.6 présente le coefficient associé par le modèle linéaire à la variable nombre de garanties. Le modèle capte bien la décroissance du risque de résiliation avec l'augmentation du nombre de garanties.

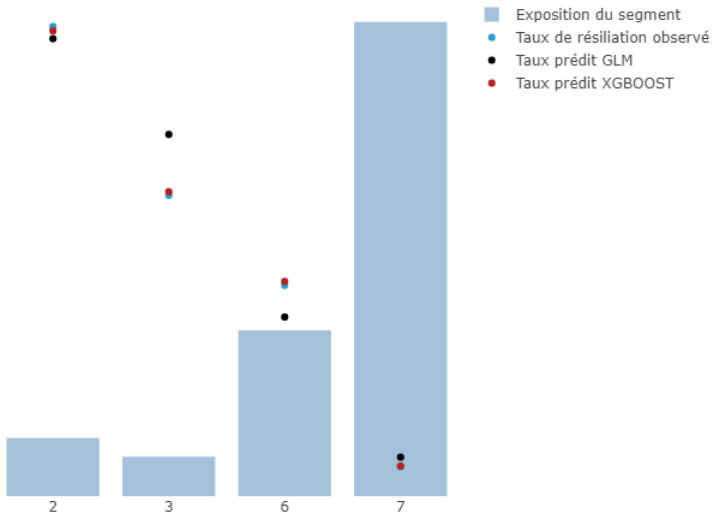


FIGURE 3.25 – Prédictions des probabilités de résiliation selon le nombre de garanties

Finalement, les prédictions sur les variables d'ordre tarifaire sont cohérentes avec ce qui est observé.

Modalité	Coefficient	Ecart-type	p-valeur
Nombre de garanties	-1.3e-01	1.6e-03	$<$ 2e-16

TABLE 3.6 – Coefficients régression logistique - Nombre de garanties

Le montant de la cotisation, approché par une fonction polynomiale, est représenté Figure 3.26. La marge et l'écart relatif au tarif médian, uniquement utilisés par les modèles sur leurs parties positives, sont représentés Figure 3.27. Il peut être noté que le modèle linéaire généralisé est moins performant sur les queues des distributions, où les segments présentent une exposition moindre.

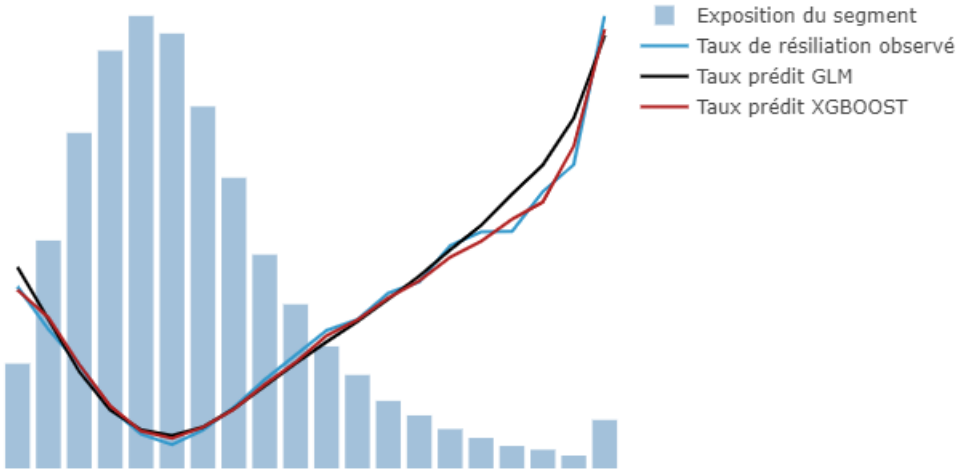


FIGURE 3.26 – Prédictions des probabilités de résiliation selon le montant de cotisation

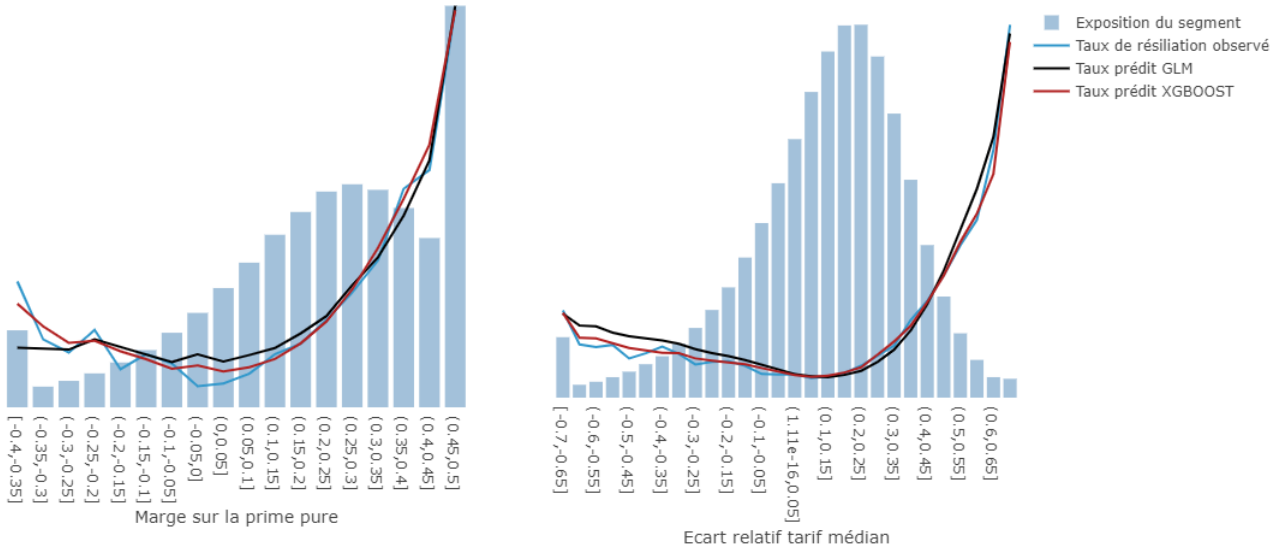


FIGURE 3.27 – Prédictions des probabilités de résiliation selon la marge et l'écart relatif au tarif médian

3.4.5 Interprétation des modèles

Cette dernière partie s'attache à l'interprétation des deux modèles construits. L'interprétation du modèle linéaire généralisé peut se faire par l'étude des odds ratios, celle du XGBoost par l'analyse des shap values.

Odds ratios

Les coefficients estimés de la régression logistique, passés à l'exponentielle, sont représentés graphiquement Figure 3.28. Les variables ayant subi une transformation de type polynomiale en amont de la régression ne sont pas représentées ici. La représentation des odds ratios permet une visualisation claire et rapide de l'impact de chacune des modalités. Certains des résultats obtenus sont détaillés :

— Les premières lignes renseignent sur les coefficients estimés pour chacune des modalités de la catégorie socioprofessionnelle, en prenant pour référence le groupe des salariés. Il est possible d'observer que les artisans, les commerçants et les personnes sans profession présentent un risque plus fort de résiliation relativement aux salariés. Inversement, les retraités représentent la catégorie la moins risquée.

— Le bonus supplémentaire est à comprendre dans le même sens que le CRM : plus ce dernier est élevé, et moins la réduction appliquée est grande. Il est observé qu'un bonus élevé, et donc qu'une réduction moindre, accru largement risque de résiliation. Cependant, l'effet du taux de dégressivité de la franchise sur la résiliation est peu marqué.

— La survenance d'avenants est un facteur important dans la prédiction de la résiliation. Observer un changement de véhicule réduit considérablement le risque de résiliation : l'assuré qui change de véhicule mais conserve le même assureur est fidélisé. Un changement de garantie impacte également la résiliation. Les effets d'une diminution et d'une augmentation de la garantie sont quasiment symétriques. Un assuré qui baisse sa couverture tendra à plus résilier toutes choses égales par ailleurs, et inversement pour un assuré qui l'augmente. Finalement, subir un sinistre au cours de l'année $n-1$ impacte négativement la rétention l'année n.

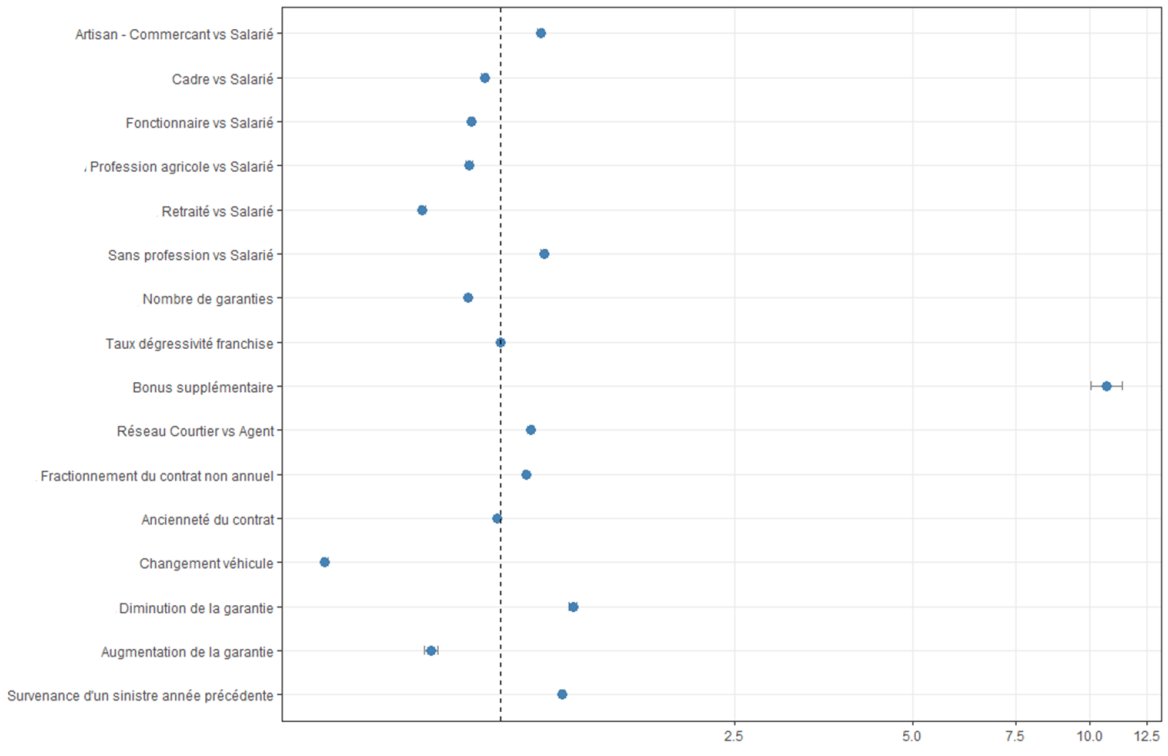


FIGURE 3.28 – Odds ratios - Prédiction de l'acte de résiliation

Shap values

Dans le cadre du XGBoost, l'étude des shap values, mesurant l'impact de chaque variable sur les performances du modèle, permet une meilleure appréhension des prédictions. Elles sont appréciées pour la stabilité apportée par leur consistance, et pour l'effet pur qu'elles expriment au travers du calcul des contributions marginales. La Figure 3.29 présente les shap values du XGBoost calibré. Les variables sont ordonnées par ordre décroissant d'importance avec en ordonnée l'impact sur la prédiction. Chaque point correspond à une observation et sa couleur indique la valeur de la variable. Plus un point est violet, et plus la valeur prise par l'observation est élevée (pour la variable considérée et relativement à ses modalités). Inversement pour un point qui tend vers le jaune. Une observation dont le point se trouve sur la partie positive des abscisses augmentera le risque de résiliation. Il est possible, entre autres, de réaliser les analyses suivantes :

— Plus le nombre de garanties détenues est faible, en jaune sur le graphique, et plus la contribution marginale de l'observation augmente le risque de résiliation.
— Pour l'âge du véhicule, les contributions sont moins nettes. Il se dégage tout de même qu'un assuré disposant d'un véhicule récent tendra à moins résilier.
— Un montant de cotisation, ou un écart relatif au tarif médian élevé augmentent le risque de résiliation.
— L'impact d'un changement de véhicule sur la résiliation est très marqué : deux groupes d'observations se distinguent de par leur couleur et de par leur signe. Un assuré dont l'indicatrice de changement de véhicule est positive résiliera moins qu'un autre assuré.

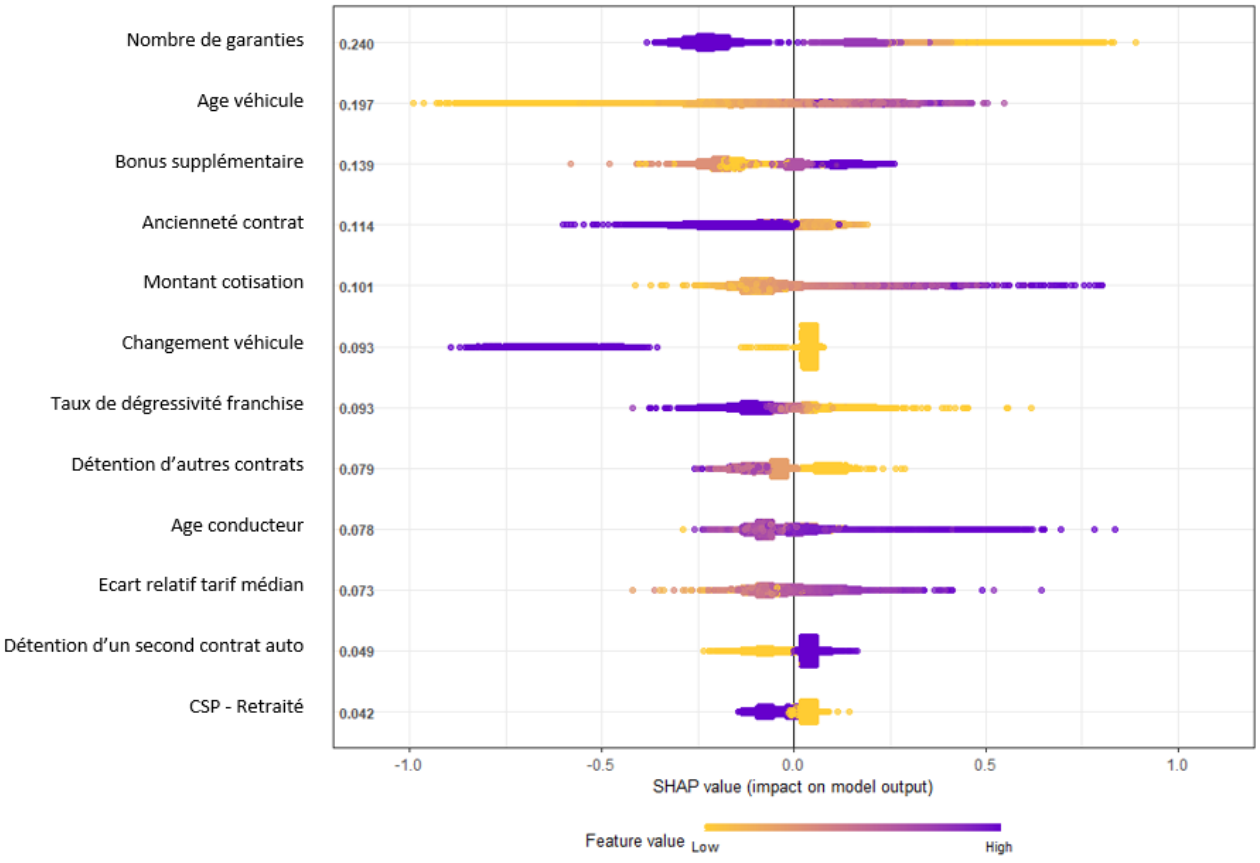


FIGURE 3.29 – Shap values (variables les plus importantes) - Prédiction de l'acte de résiliation

Chapitre 4

Durée de vie a priori

Sommaire

Ce chapitre présente la deuxième modélisation relative à la fidélité des assurés. Bien que peu utilisés en assurance dommage, les modèles de durée permettent de déterminer pour un continuum d'instants t, la probabilité que l'assuré soit toujours en portefeuille à cet instant. Ainsi, l'implémentation de tels modèles, à partir des caractéristiques à la souscription de l'assuré, fournit le recul nécessaire à l'appréhension des comportements de fidélisation des individus en portefeuille à moyen terme. Les trois premières sections de ce chapitre introduisent les notions théoriques des modèles de durée de vie, en présentant notamment les modèles de Kaplan-Meier et de Cox. Ensuite, les parties 4 et 5 s'attachent à appliquer sur le portefeuille la théorie introduite, au travers respectivement, d'une analyse descriptive et d'une modélisation.

4.1 Une théorie spécifique

L'analyse de durées consiste en l'étude d'une variable aléatoire T, qui représente le temps passé dans un certain état. T peut correspondre à une durée de vie humaine, à celle d'un composant électronique, d'un véhicule ou encore au temps qui s'écoule entre la survenance de deux sinistres. Dans le cadre du travail mené, T correspond au temps passé par l'assuré en portefeuille, entre sa souscription et sa résiliation. De telles variables, par leurs caractéristiques, nécessitent une approche mathématique spécifique. Premièrement, la variable aléatoire T est positive : la loi normale ne fait plus référence. De plus, l'estimation de quantités interprétables, dans le cadre de l'étude de survie, requiert de définir des objets, certes connexes aux probabilités classiques, mais néanmoins particuliers aux modèles de durée. Ensuite, comme introduit Section 2.5.1, l'analyse de durée est sujette à la censure : il arrive que la valeur prise par T ne soit pas observée, dans ce cas, seulement un minorant de cette dernière est connu. Ainsi, cette première section définit les outils essentiels à l'analyse et à la modélisation de variables de durées.

4.1.1 Fonctions de bases à l'analyse de durée

Soit T une variable aléatoire continue à support dans $\mathbb{R}^+$. Deux fonctions essentielles à l'étude de durées sont définies : la fonction de survie et le taux de risque instantané.

Fonction de survie

La distribution de T est définie de manière unique par sa fonction de répartition $F(.) : \mathbb{R}^+ \to [0; 1]$, telle que :

$$F(t) = \mathbb{P}(T \leq t)$$

Pour l'étude de durées, la fonction de survie $S(.)$, permettant une interprétation plus directe, est préférée à la fonction de répartition $F(.)$. Elle est définie par :

$$\begin{aligned} S(t) &= 1 - F(t^-) \\ &= \mathbb{P}(T \geq t) \end{aligned}$$

Dans le cadre de l'analyse de la durée de vie des contrats, pour t fixé, $S(t)$ représente la probabilité que le contrat ne soit pas encore résilié à la date t.

Taux de risque instantané

Aussi appelé taux de hasard, le taux de risque instantané permet de capter l'imminence du risque porté par la résiliation de l'assuré. Dans le cas continu, le taux de risque instantané correspond à la fonction $h(.) : \mathbb{R}^+ \to \mathbb{R}^+$, définie par :

$$h(t) = \lim_{dt \to 0^+} \frac{\mathbb{P}(T \in [t, t + dt] | T \geq t)}{dt}$$

Lors de l'étude de la durée de vie des contrats, $h(t)$ se comprend comme le risque que l'assuré résilie son contrat à l'instant t sachant qu'il ne l'a pas résilié jusqu'à lors. Le taux de hasard cumulé peut également être introduit, il correspond à l'intégrale du taux de risque instantané :

$$H(t) = \int_0^t h(u)du$$

En posant $f(.)$ la fonction de densité de T, le taux de risque instantané et la fonction de survie sont liés par les relations suivantes :

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ S(t) &= \exp(-\int_0^t h(u)du) = \exp(-H(t)) \end{aligned}$$

Au même titre que la densité $f(.)$ et la fonction de répartition $F(.)$, les fonctions $S(.)$ et $h(.)$ définissent de manière unique la loi observée par T.

4.1.2 La censure à droite

La Section 2.5.1 a permis de présenter les différents types de censures, et il a été identifié que la censure à droite de type I, aussi appelée censure déterministe, est celle présente dans les données. Pour rappel, la censure à droite perturbe la vision de la partie droite de la distribution. Formellement, soient $(T_1, ..., T_n)$, n variables indépendantes et identiquement distribuées (iid) de même loi que T, la variable d'intérêt. Soient $(C_1, ..., C_n)$, iid également, de même loi qu'une variable de censure C. En présence de censure, les variables $(T_1, ..., T_n)$ ne sont pas constatées. En revanche, le vecteur de variables $(Y_1, \delta_1, ..., Y_n, \delta_n)$ est observé. Ce dernier se définit $\forall i = 1, ..., n$:

$$\begin{cases} Y_i = \inf(T_i, C_i) \\ \delta_i = \mathbb{1}_{T_i \leq C_i} \end{cases}$$

où :

- Y_i : durée pendant laquelle le contrat i a été observé ;
- T_i : durée de vie du contrat i ;
- C_i : durée avant la censure du contrat i, correspond à l'âge du contrat lorsqu'il quitte l'observation pour une autre cause que la résiliation, ici pour cause d'extraction de la base ;
- δ_i : indicatrice de censure, si $\delta_i = 1$, le contrat i a été résilié avant la date d'extraction des données et sa durée de vie est connue, inversement si $\delta_i = 0$, l'observation est censurée et seul un minorant de sa durée de vie est observé.

4.1.3 Les modèles de durée

L'enjeu de tout modèle de durée est d'estimer la fonction de survie $S(.)$, adéquate au phénomène observé dans les données, tout en prenant en compte la censure. Les modèles de durée, au même titre que l'ensemble des modèles statistiques, se regroupent en trois familles : les modèles non paramétriques, semi-paramétriques et paramétriques.

Les modèles non paramétriques fournissent une estimation de la survie observée par T, sans faire d'hypothèse sur sa loi, et sans prendre en compte de covariable. Les estimateurs de Kaplan-Meier et de Nelson-Aalen en sont référents. Les modèles paramétriques supposent à l'inverse que la durée de vie appartient à une certaine une certaine loi, dont les paramètres sont alors à estimer, généralement par méthode du maximum de vraisemblance. L'enjeu est de déterminer la loi adaptée, notamment celle dont le taux de hasard $h(.)$ représente le mieux le risque considéré. Lors de l'étude de la durée de vie humaine, le taux de risque est croissant : à partir d'un certain âge, plus l'individu est âgé et plus ce dernier est susceptible de décéder. A l'inverse, il existe des phénomènes dits avec rajeunissement : plus le temps passé dans un état est long, et moins l'individu est sujet au risque. La loi exponentielle, ou sa généralisation la loi de Weibull, sont couramment utilisées dans la modélisation paramétrique de durées. Finalement, les modèles semi-paramétriques permettent d'expliquer la survie du contrat en fonction de covariables. Ces derniers disposent d'une partie non paramétrique, et d'une partie paramétrique dont les coefficients sont estimés à partir des caractéristiques des assurés. Ils sont appréciés pour leur flexibilité et la facilité d'interprétation de leur partie paramétrique.

Dans ce mémoire, deux approches sont étudiées. L'estimation non paramétrique par Kaplan-Meier permet une première appréhension des comportements, une analyse des durées de vie des contrats en fonction des caractéristiques des assurés et la validation ou l'invalidation des hypothèses nécessaires à la mise en œuvre d'un modèle semi-paramétrique. Dans un deuxième temps, un modèle semi-paramétrique de Cox, utilisé à des fins prédictives, permettra l'estimation de la fonction de survie des contrats en fonction des caractéristiques observées à la souscription.

4.2 Estimateur non paramétrique de Kaplan-Meier

Dans cette section, l'estimateur non paramétrique de Kaplan-Meier, est défini puis ses principales propriétés sont énoncées. Finalement, un des tests statistique, basé sur l'estimateur de Kaplan-Meier et permettant la comparaison des comportements de survie de différents groupes est proposé.

4.2.1 Définition

L'estimateur de Kaplan-Meier est un estimateur de la fonction de répartition, présentant des comportements asymptotiques solides, même en cas de censure. Une intuition sur sa construction est proposée.

Fonction de répartition empirique

Soit le vecteur de variables aléatoires indépendantes et identiquement distribuées $(T_1, ..., T_n)$. Dans le cas où la censure n'intervient pas, la fonction de répartition empirique s'exprime :

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{T_i \leq t}$$

$\hat{F}_n(.)$ est un estimateur non biaisé de la fonction de répartition $F(.)$, aux propriétés asymptotiques satisfaisantes :

- par la loi forte des grands nombres (Kolmogorov, 1929) : $\hat{F}_n(t) \xrightarrow[n \to +\infty]{p.s.} F(t)$;
- par le théorème de Glivenko-Cantelli : $\hat{F}_n(t) \xrightarrow[n \to +\infty]{unif} F(t)$ presque sûrement ;
- par le théorème central limite : $\sqrt{n}(\hat{F}_n(t) - F(t)) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, F(t)(1 - F(t)))$.

Construction de l'estimateur de Kaplan-Meier

En présence de censure, l'information portée par $\mathbb{1}_{T_i \leq t}$, le nombre exact d'observations dont la durée de vie est inférieure à un certain temps t, n'est plus disponible. L'estimateur de Kaplan-Meier, proposé en 1958 [19], permet d'obtenir, dans le cadre de données censurées, un estimateur de la fonction de répartition aux propriétés analogues à celles observées par $\hat{F}_n(t)$. Soit un échantillon de variables aléatoires sujettes à la censure. En reprenant les notation utilisées lors de la formalisation de la censure à droite Section 4.1.2, le couple de variables aléatoires (Y_i, δ_i) est observé pour tout $i = 1, ..., n$ et se définit par :

$$\begin{cases} Y_i = \inf(T_i, C_i) \\ \delta_i = \mathbb{1}_{T_i \leq C_i} \end{cases}$$

Les hypothèses suivantes sont nécessaires à la construction de l'estimateur empirique de Kaplan-Meier :

- $(T_1, ..., T_n)$ *iid* de même loi que T la variable d'intérêt ;
- $(C_1, ..., C_n)$ *iid* de même loi que C la variable de censure ;
- hypothèse d'identification : les variables d'intérêt T_i sont indépendantes des variables de censure C_i pour tout $i = 1, ..., n$.

Premièrement, il est possible, dans le cas discret, de montrer par récurrence que la fonction de survie peut s'exprimer à partir du taux de risque instantané :

$$S(t) = \prod_{t_i < t} (1 - h(t_i)) \tag{4.1}$$

Ensuite, les éléments de base à la construction de l'estimateur de la fonction de répartition empirique de Kaplan-Meier sont les fonctions $H(.)$ et $H_1(.)$, dont les estimateurs empiriques respectifs peuvent

être calculés à partir des données censurées :

$$\begin{aligned} H(t) &= \mathbb{P}(Y \leq t), & \hat{H}(t) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \leq t} \\ H_1(t) &= \mathbb{P}(Y \leq t | \delta = 1), & \hat{H}_1(t) &= \frac{1}{n}\sum_{i=1}^{n} \delta_i \mathbb{1}_{Y_i \leq t} \end{aligned} \tag{4.2}$$

Il est possible d'exprimer le taux de hasard en fonction des fonctions $H(.)$ et $H_1(.)$:

$$h(t) = \frac{dF(t)}{1 - F(t^-)} = \frac{dH_1(t)}{1 - H(t^-)} \tag{4.3}$$

Finalement, à partir des Équations 4.1, 4.2 et 4.3, l'estimateur empirique de Kaplan-Meier est obtenu :

$$\hat{F}_n^{KM}(t) = 1 - \prod_{i; Y_i \leq t} (1 - \frac{\delta_i}{\sum_{j=1}^{n} \mathbb{1}_{Y_i \leq Y_j}})$$

Cette expression est valable lorsque les Y_i sont tous distincts et que les T_i et C_i sont indépendants $\forall i$. Ces hypothèses assurent la convergence de l'estimateur.

4.2.2 Propriétés

L'estimateur ainsi construit présente plusieurs propriétés dont les principales sont présentées ici.

Si l'échantillon ne présente pas de censure, l'estimateur de Kaplan-Meier $\hat{F}_n^{KM}(.)$ coïncide avec la fonction de répartition empirique $\hat{F}_n(.)$. De plus, $\hat{F}_n^{KM}(.)$ est une fonction constante par morceau et peut se réécrire comme la somme :

$$\hat{F}_n^{KM}(t) = \sum_{i=1}^{n} W_{i,n} \cdot \mathbb{1}_{Y_{(i)} \leq t}$$

où $W_{i,n}$ représente la masse attribuée par l'estimateur à $Y_{(i)}$, la $i^{\text{ème}}$ observation. Afin de compenser la sous-représentation des grandes durées dans l'échantillon observé $(Y_1, ..., Y_n)$ par rapport à l'échantillon complet $(T_1, ..., T_n)$, les poids $W_{i,n}$ sont croissants avec le rang de l'observation dans l'échantillon. Plus une observation non censurée est grande, et plus le poids qui lui est attribué est conséquent, ce qui vient, en partie, équilibrer le déficit d'observations à droite de la distribution. Cependant, cette particularité est également vecteur d'une des limites de l'estimateur, qui tend à mal se comporter sur la queue droite de la distribution.

D'un point de vue théorique, l'estimateur de Kaplan-Meier dispose de bonnes propriétés. Premièrement, Droesbeke et al. montrent en 1989 [9] que l'estimateur $\hat{S}_n^{KM}$, obtenu à partir de $\hat{F}_n^{KM}$, est le seul estimateur cohérent de la survie. Formellement, l'estimateur $\hat{S}_n^{KM}$ vérifie l'équation implicite :

$$\hat{S}_n^{KM}(t) = \frac{1}{n}\left[\sum_{i=1}^{n} \mathbb{1}_{Y_i > t} + \sum_{i=1}^{n} \mathbb{1}_{Y_i \leq t, \delta_i = 0} \cdot \frac{\hat{S}_n^{KM}(t)}{\hat{S}_n^{KM}(Y_i)}\right]$$

Cela signifie que la probabilité d'être encore en vie à l'instant t est la somme de la probabilité de n'être ni décédé, ni censuré à t et celle d'être toujours en vie, mais censuré. Ensuite, un résultat voisin à celui de Glivenko-Cantelli, démontré par Stute [31] assure, sous certaines contraintes, la convergence uniforme de l'estimateur $\hat{F}_n^{KM}$ vers la fonction de répartition F. De plus, Gill [13] atteste, à conditions d'hypothèses supplémentaires, de la normalité asymptotique de l'estimateur en présence de censure. Enfin, en considérant la fonction S comme un paramètre dans l'espace des fonctions de survie, et $\hat{S}_n^{KM}$ comme un estimateur de S, il est possible de montrer que $\hat{S}_n^{KM}$ correspond à l'estimateur non paramétrique du maximum de vraisemblance dans l'espace des fonctions de survie. Les propriétés satisfaites par l'estimateur de Kaplan-Meier conduisent à le préférer, dans le cadre de l'étude, à d'autres estimateurs non paramétriques tels que celui de Nelson-Aalen.

4.2.3 Comparaison de courbes de survie

Dans le cadre de données censurées, des tests statistiques permettent d'éprouver l'hypothèse d'égalité des fonctions de survie de deux échantillons distincts. Le test du log-rank, présenté ici, est communément utilisé. Soient deux groupes : le groupe 1 dont la fonction de survie est S_1, le groupe 2 dont la fonction de survie est S_2. Une fois leurs estimations réalisées par Kaplan-Meier, il est possible de réaliser le test suivant :

$$H_0 : S_1 = S_2 \quad \text{contre} \quad H_1 : S_1 \neq S_2$$

L'idée du test du log-rank est de comparer le nombre d'évènements observés au sein de chacun des groupes à ceux attendus sous l'hypothèse nulle. Formellement, soit la suite ordonnée des décès observés dans l'ensemble de l'échantillon : $t_1 < ... < t_n$. Pour chaque instant t_i, r_{ij} et d_{ij} désignent respectivement le nombre d'individus et le nombre de décès observés dans le groupe j, r_i et d_i correspondent à ces mêmes valeurs mais sur l'ensemble de l'échantillon. Sous l'hypothèse nulle H_0, les proportions de décès dans les deux groupes doivent être égales à chaque instant. Ainsi, d_{ij} suit une loi hypergéométrique $\mathcal{H}(r_i, d_i, \frac{r_{ij}}{r_i})$, d'espérance notée $\mathbb{E}_{ij}$ et de variance $\mathbb{V}_{ij}$. La statistique de test prend la forme :

$$\phi = \sum_{j=1}^{2} \frac{(\sum_{i=1}^{n} d_{ij} - \mathbb{E}_{ij})^2}{\sum_{i=1}^{n} \mathbb{V}_{ij}}$$

et suit asymptotiquement une loi du χ^2 à 1 degré de liberté. Le test, de niveau α, rejette alors l'hypothèse nulle dans le cas où :

$$\phi \geq q_{\chi^2(1)}(1-\alpha)$$

avec $q_{\chi^2(1)}(1-\alpha)$ le quantile d'ordre $(1-\alpha)$ de la loi $\chi^2(1)$.

4.3 Modèle à hasard proportionnel de Cox

Les modèles semi-paramétriques permettent d'une part, de ne pas spécifier entièrement la famille de loi à laquelle la variable d'intérêt T appartient, et d'autre part, de prendre en compte et de mesurer les effets relatifs des différentes covariables. Il est alors possible d'expliquer T, la durée de vie du contrat, en fonction des caractéristiques portées par l'assuré. Les principales hypothèses de ce type de modèles reposent sur la forme de l'influence des covariables sur la variable d'intérêt. Deux classes de modèles sont communément différenciées, les modèles à hasards proportionnels et les modèles à temps accéléré, qui se distinguent par l'effet des covariables sur la fonction de hasard. Le modèle à hasards proportionnels de Cox, introduit par ce dernier en 1972 [7], est l'un des modèles les plus couramment utilisés en analyse de durées. Par sa flexibilité et la transparence de sa partie paramétrique, la régression de Cox est le modèle préféré dans le cadre de l'étude. Cette section s'attache à sa définition, à l'estimation et à l'interprétation de ses paramètres, à la présentation de ses hypothèses et aux métriques de validation du modèle.

4.3.1 Forme du modèle

Soit X le vecteur de covariables, correspondant aux caractéristiques de l'assuré à sa souscription, le modèle de Cox [7] s'écrit alors :

$$h(t|X) = h_0(t) \exp(\beta' X)$$

avec β, le vecteur de paramètres à estimer. La fonction $h_0(.)$, appelée taux de risque instantané de base, est inconnue, dans le sens où elle n'est pas supposée répondre à une certaine forme, d'où le terme de modèle semi-paramétrique. Le taux de hasard est alors basé sur deux quantités. Le taux de hasard de base, $h_0(t)$, indépendant des covariables, peut être interprété comme le taux de hasard d'un individu référent, dont toutes les covariables seraient nulles. La partie paramétrique, $\exp(\beta' X)$, indépendante du temps t, est comprise comme le risque relatif associé à l'individu aux caractéristiques X.

En posant $S_0(t) = \exp(-\int_0^t h_0(u)du)$, la fonction de survie, dans le cadre de la régression de Cox, peut être exprimée en fonction de la partie paramétrique du modèle, et donc des caractéristiques de l'assuré :

$$S(t|X) = S_0(t)^{\exp(\beta' X)}$$

4.3.2 Estimation des paramètres

L'estimation des paramètres du modèle de Cox se fait en deux temps. Premièrement, les coefficients liés aux covariables, sont obtenus par une méthode similaire à celle du maximum de vraisemblance. Ensuite, le risque de base est estimé.

Vraisemblance partielle de Cox

L'estimation du vecteur de paramètres β peut se faire par maximisation de la vraisemblance partielle, proposée par Cox en 1975 [8] :

$$\mathcal{L}(x;\beta) = \prod_{Y_i;\delta_i=1} \frac{\exp(\beta' x_i)}{\sum_{Y_i \leq Y_j} \exp(\beta' x_i)}$$

La vraisemblance partielle ne dépend pas du taux de risque de base $h_0(.)$, ce qui permet l'estimation de la partie paramétrique du modèle sans définir le hasard de base. De plus, bien que la vraisemblance partielle ne corresponde pas, dans le sens statistique, à une vraisemblance, elle se comporte de la même manière, ce qui lui confère des propriétés asymptotiques similaires. Le vecteur β est estimé par maximisation de la log-vraisemblance partielle, ce qui donne lieu à un système d'équations différentielles, résolu par méthode de Newton-Raphson.

Risque de base

Le risque de base, $h_0(t)$ correspond aux taux de hasard instantané $h(t)$ lorsque toutes les covariables du modèle sont nulles. En pratique, le risque de base cumulé, défini comme $H_0(t) = \int_0^t h_0(u)du$, est estimé à l'aide de l'estimateur de Breslow [3] :

$$\hat{H}_0(t) = \sum_{Y_i \leq t} \frac{\delta_i}{\sum_{Y_i \leq Y_j} \exp(\hat{\beta}' x_j)}$$

Survie estimée

Finalement, la forme finale de la fonction de survie estimée par la régression de Cox est obtenue :

$$\hat{S}^{COX}(t|X) = \exp(-\hat{H}_0(t)\exp(\hat{\beta}' X))$$

4.3.3 Interprétation des coefficients

Le modèle de Cox à risques proportionnels est apprécié, entre autres, pour l'interprétabilité de ses coefficients. L'étude des ratios de hasard, mesurant le risque de survenance de l'évènement chez un individu par rapport à un autre, permet une appréhension directe de l'effet d'une modalité sur la durée de vie. Soient deux individus, dont une des covariables prend pour valeur x_1 chez l'individu 1, et x_2 pour l'individu 2. Le ratio de hasard entre ces deux assurés se définit comme le rapport des deux taux de hasard :

$$\begin{aligned} HR(x_1, x_2) &= \frac{h(t|x_1)}{h(t|x_2)} \\ &= \frac{h_0(t)\exp(\beta' x_1)}{h_0(t)\exp(\beta' x_2)} \\ &= \exp(\beta'(x_1 - x_2)) \end{aligned} \tag{4.4}$$

Au même titre que lors de l'étude des odds ratios dans le cadre de la régression logistique, trois cas peuvent se présenter dans l'analyse des risques relatifs :

— $HR(x_1, x_2) = 1$: les risques de survenance de l'évènement étudié, ici la résiliation, portés par les deux individus sont identiques ;
— $HR(x_1, x_2) > 1$: les caractéristiques présentées par l'individu X_1 tendent à augmenter le risque relativement à celles de l'individu X_2 ;
— $HR(x_1, x_2) < 1$: situation inverse à celle du point précédent.

Ainsi, les coefficients estimés $\hat{\beta}$ peuvent être interprétés directement en sortie du modèle, de la même façon que ceux d'une régression logistique. Un coefficient positif indiquera que la variable, en croissant, si elle est continue, ou sa modalité, si elle est catégorielle, a un impact négatif sur la rétention.

4.3.4 Hypothèses

Le modèle de Cox repose sur deux hypothèses fondamentales : les covariables observent une relation linéaire avec le logarithme de la fonction de hasard et les risques sont proportionnels. Cette section s'attache à détailler ces hypothèses et à proposer des méthodes de validation de ces dernières.

Hypothèse de log-linéarité

L'expression du taux de hasard proposée par le modèle de Cox induit une relation linéaire entre les covariables continue et le log du ratio entre le risque instantané et le risque de base :

$$\ln(\frac{h(t|X)}{h_0(t|X)}) = \beta' X$$

Cela induit une linéarité de l'effet des variables explicatives continues sur le taux de hasard. En prenant l'exemple de l'âge du véhicule, l'hypothèse de log-linéarité implique que le risque relatif entre un assuré disposant d'un véhicule âgé de 5 ans et un assuré disposant d'un véhicule de 10 ans est le même que le risque relatif entre un véhicule de 15 ans et un véhicule de 20 ans. La linéarité du risque relatif s'observe clairement au travers de son expression Équation 4.4.

La méthode reposant sur l'analyse des résidus de martingale est proposée pour vérifier la validité de l'hypothèse, et dans le cas contraire, déterminer la fonctionnelle la plus adaptée pour satisfaire la log-linéarité. Cette approche repose sur la théorie des processus de comptage, dont le mémoire ne fait pas l'objet, la théorie de ces derniers ne sera abordée que succinctement. Dans le cadre du modèle linéaire généralisé, explicité Section 3.1, les résidus généralement notés ϵ, font partie de l'équation de régression et correspondent à la distance observée entre le réel et le prédit. Du fait de la censure, une expression directe des résidus dans le contexte des modèles de survie n'est pas disponible. En 1988, Barlow et Prentice [1] exposent une approche martingale pour l'obtention de tels résidus dans des modèles avec censure. A l'appui de ces travaux, Therneau et al. [33] présentent des approches graphiques, basées sur les résidus de martingale, pour la vérification des hypothèses du modèle de Cox. En se basant sur les notations des articles susmentionnés, une appréhension concise des résidus de martingale, et de leur étude graphique pour la validation de l'hypothèse de log-linéarité et la recherche de fonctionnelles adaptées est proposée.

Soient :

— $N_i(t) = \mathbb{1}_{T_i \leq t}$, le processus de comptage, qui vaut 1 une fois que l'individu i a observé l'évènement. Dans le cadre de l'étude réalisée, $N_i(t) = 1$ si l'assuré i est sorti du portefeuille. ;
— $\Lambda_i(t)$ le processus croissant, appelé compensateur, issu de la décomposition de Doob-Meyer de N_i en la somme d'un processus croissant et d'une martingale continue à droite ;
— $\lambda_i(t)$ la fonction d'intensité du processus de comptage N, telle que $\Lambda_i(t) = \int_0^t \lambda_i(u)du$;
— $Y_i(t) = 1 - N_i(t) = \mathbb{1}_{T_i > t}$, le processus qui indique si l'individu i est encore sujet au risque au temps t.

Dans le cadre du modèle de Cox, le processus d'intensité cumulée $\Lambda(t)$ s'exprime à partir du processus $Y(t)$ et de la fonction de hasard $h(t)$:

$$\begin{aligned}\Lambda_i(t) &= \int_0^t Y_i(u)h_i(u)du \\ &= \int_0^t Y_i(u)\exp(\beta' X_i)dH_0(u) \\ &= \int_0^{t\wedge T_i} \exp(\beta' X_i)dH_0(u)\end{aligned}$$

D'après la décomposition de Doob-Meyer, la différence entre le processus de comptage N et son intensité cumulée Λ est une martingale, appelée martingale compensée, dont l'expression est la suivante :

$$M_i(t) = N_i(t) - \int_0^{t\wedge T_i} \exp(\beta' X_i)dH_0(u)$$

La quantité d'intérêt est alors l'estimateur de la martingale $\hat{M}_i := \hat{M}_i(+\infty)$ pour chacune des observations. Elle correspond à la différence entre le nombre d'évènements observés et le nombre d'évènements attendus selon le modèle :

$$\hat{M}_i = \delta_i - \hat{H}_0(T_i)\exp(\hat{\beta}' X_i)$$

Therneau et al. [33] montrent que si le modèle vérifie effectivement $h(t|x) = h_0(t)\exp(f(x)\beta)$, les résidus de martingale du modèle nul, c'est-à-dire du modèle sans covariable, vérifient :

$$\mathbb{E}[M|x] \approx cf(x)$$

où c ne dépend que du nombre d'individus censurés. Ainsi, l'étude de la log-linéarité d'une covariable continue, ainsi que la détermination de sa transformation adaptée, peut se faire par représentation graphique des résidus de martingale lissés du modèle nul, en fonction des valeurs de la covariable. L'estimation lissée des résidus présente alors la forme de la fonctionnelle adaptée f à la covariable X de sorte que $f(X)$ respecte l'hypothèse de log-linéarité du modèle de Cox.

Hypothèse de proportionnalité

La forme du taux de hasard sur laquelle repose le modèle de Cox suppose une constance des ratios de risques. Autrement dit, l'effet d'un facteur de risque ne dépend pas du temps. Deux approches permettent de valider ou d'invalider cette hypothèse centrale. La première approche, s'appuyant sur des représentations graphiques, est utile dans le cadre de variables qualitatives ou discrètes, quand la seconde, basée sur l'analyse des résidus de Schoenfeld standardisés, permet l'étude de la validité de l'hypothèse pour les variables continues.

Pour rappel, à partir du modèle de Cox, la fonction de survie peut être exprimée comme $S(t|X) = S_0(t)^{\exp(X\beta)}$. En appliquant la fonction $\ln(-\ln(.))$ aux deux membres de cette égalité, la forme linéaire portée par la partie paramétrique apparaît :

$$\ln[-\ln(S(t|X))] = \ln[-\ln(S_0(t))] + X\beta$$

Soient deux individus différents, aux covariables X_1 et X_2. En soustrayant l'expression précédente appliquée à l'individu 1 à celle de l'individu 2, la relation suivante est obtenue :

$$\ln[-\ln(S(t|X_1))] = \ln[-\ln(S(t|X_2))] + (X_1 - X_2)\beta$$

Par conséquent, l'hypothèse de proportionnalité est vérifiée si les courbes $\ln[-\ln(S(t|X_1))]$ et $\ln[-\ln(S(t|X_2))]$, représentées dans le même plan, en fonction du temps t, sont parallèles. Cette méthode graphique ne peut être appliquée aux variables continues. Dans ce cas, il est possible de segmenter

la variable en plusieurs classes, avant d'inspecter graphiquement son respect de l'hypothèse de proportionnalité.

Bien qu'efficace dans le cadre de la validation graphique de l'hypothèse de proportionnalité, discrétiser une variable continue conduit nécessairement à une perte d'information. Une analyse basée sur les résidus de Schoenfeld, définis en 1982 [30], sera alors préférée pour les variables continues. Intuitivement, les résidus de Schoenfeld, calculés pour toutes les covariables de chaque individu non censuré, comparent au moment de la survenance de l'évènement, la valeur des covariables de l'individu et la moyenne pondérée par le risque des covariables des autres sujets exposés au risque. Autrement dit, ils correspondent à la différence entre la covariable observée et la covariable attendue, pondérée par le risque, lors de la réalisation de l'évènement. Les résidus sont ensuite standardisés à l'aide de la covariance. En posant $s^*_{k,j}$ le résidu de Schoenfeld standardisé au temps t_k de la covariable associée au coefficient estimé $\hat{\beta}_j$, Grambsch et Therneau [14] mettent en évidence que la valeur espérée de $s^*_{k,j}$ correspond à la déviation $\hat{\beta}_j$ au temps t_k :

$$\mathbb{E}[s^*_{k,j}] + \hat{\beta}_j \approx \beta_j(t_k)$$

Ainsi, en représentant graphiquement $\mathbb{E}[s^*_{k,j}] + \hat{\beta}_j$ en fonction du temps, il est possible d'estimer la dépendance temporelle $\beta_j(t_k)$ du coefficient. Une covariable respectant l'hypothèse de proportionnalité présentera une courbe $\mathbb{E}[s^*_{k,j}] + \hat{\beta}_j$ horizontale. Le lecteur intéressé par les subtilités calculatoires de cette méthode peut se référer à l'ouvrage proposé par Therneau et Grambsch [32].

4.3.5 La concordance : métrique d'évaluation d'un modèle de durée

La concordance, dont l'utilisation pour l'évaluation des modèles de durée a été popularisée par Harrell [16], est la métrique de performance la plus largement utilisée pour mesurer la qualité d'une régression de Cox. Évaluant l'aptitude du modèle à ordonner le risque attribué à chaque individu, la concordance peut être comprise comme une généralisation de l'AUC (métrique introduite formellement lors de l'étude de la résiliation à un an, Section 3.3), prenant en compte le phénomène de censure. La notion de concordance, ancienne dans le domaine de la statistique, repose sur le principe simple de comparaison de variables. Un couple d'observations (x_1, x_2) et (y_1, y_2) est dit concordant si $\{x_1 > x_2$ et $y_1 > y_2\}$ ou si $\{x_1 < x_2$ et $y_1 < y_2\}$. Soient deux observations, dont les durées de vie observées sont respectivement T_1 et T_2, et dont les prédicteurs linéaires, correspondant au $\hat{\beta}'X$, avec $\hat{\beta}$ estimé au travers du modèle de Cox, valent chacun η_1 et η_2. Dans le cadre du modèle de Cox, deux observations seront dites concordantes si :

$$\{T_1 > T_2 \text{ et } \eta_1 < \eta_2\} \text{ ou } \{T_1 < T_2 \text{ et } \eta_1 > \eta_2\}$$

En effet, le modèle est cohérent dans le cas où si un individu dispose d'une durée de vie plus longue qu'un second ($T_1 > T_2$), alors le risque qu'il supporte, représenté par le prédicteur linéaire η, est inférieur à celui du deuxième individu ($\eta_1 < \eta_2$). Ainsi, la métrique de concordance, notée C vise à mesurer la proportion d'observations concordantes relativement à l'ensemble des observations. Seules les observations non censurées, dont la durée de vie réelle est connue, peuvent être utilisées dans le calcul de la métrique. En notant n, le nombre d'individus en portefeuille et δ_i l'indicatrice de non censure de l'individu i, la mesure de concordance, notée C, se définit :

$$C = \frac{\sum_{i,j=1}^{n} \mathbb{1}_{T_i > T_j} \mathbb{1}_{\eta_i < \eta_j} \delta_j}{\sum_{i,j=1}^{n} \mathbb{1}_{T_i > T_j} \delta_j}$$

De la même manière que l'AUC, la métrique C est comprise entre 0 et 1. Un modèle dont les prédictions sont purement aléatoires disposera d'une concordance $C = 0,5$, et plus la mesure C approche de 1, meilleure est la qualité du modèle.

4.4 Analyses exploratoires

Les trois sections précédentes ont permis de définir le cadre théorique dans lequel s'inscrit l'étude de durées de vie. Les sections à venir s'attachent à l'application des modèles ainsi définis. Dans un premier temps, un travail d'analyse exploratoire est mené, par l'étude des différentes statistiques de durée et de censure sur le portefeuille à disposition. Ensuite, l'examen des courbes de survie empirique par l'estimateur de Kaplan-Meier propose une appréhension des comportements de survie des contrats en fonction des modalités des différentes covariables.

4.4.1 Statistiques préliminaires

La base de données utilisée dans cette section d'application est issue des travaux de préparation présentés Chapitre 2 et est décrite en fin de ce dernier, Section 2.6.2. Elle contient les informations à la souscription, la durée de vie ainsi que l'indicatrice de censure de près de 450 000 contrats souscrits entre 2009 et 2020. La répartition des durées de vie et les principales statistiques d'ordre et de position sont proposées Figure 4.1.

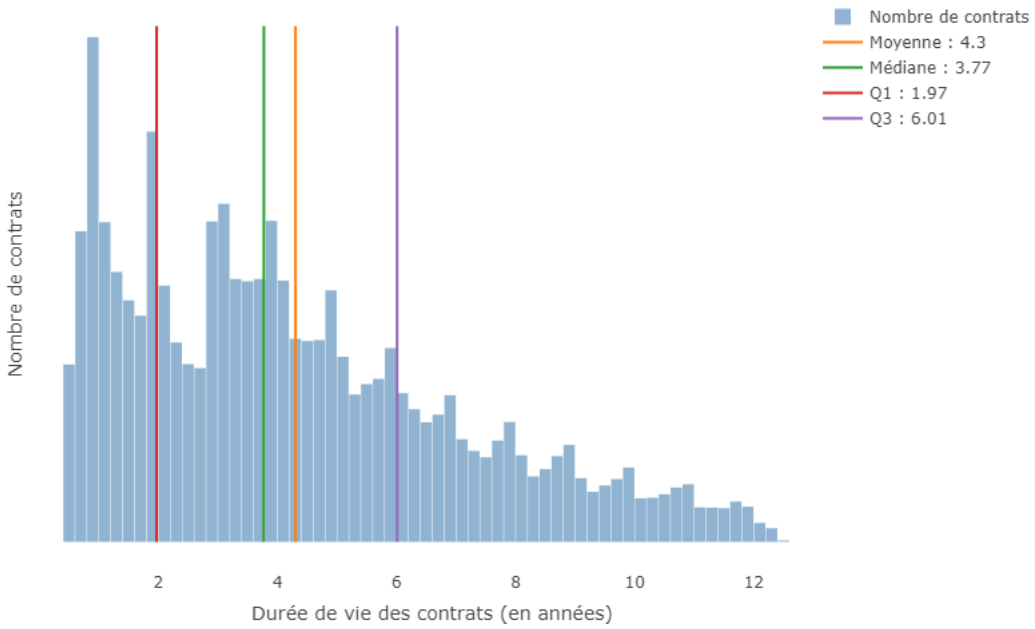


FIGURE 4.1 – Répartition et statistiques principales de la durée de vie des contrats

Les statistiques associées à la durée de vie sont inférieures à celles attendues. Précisément, à partir des taux de résiliation avoisinant 14% par an, une durée de vie moyenne cohérente, donnée par le temps de retour $1/p$, où p correspond à la probabilité de résiliation à 1 an, se situe autour des 7 ans. La moyenne observée, de 4.3 ans, s'explique en partie par la profondeur d'historique à disposition. En effet, bien que la période d'étude d'un peu moins de 13 ans semble large, cette dernière ne permet pas d'obtenir une répartition complètement réaliste des durées de vie. Seuls les contrats souscrits en 2009 ou ultérieurement sont étudiés. Cela implique que la durée de vie des contrats ne peut excéder les 13 ans, quand des assurés satisfaits et peu enclins à entreprendre des recherches de contrats qu'ils pourraient juger plus attractifs, sont susceptibles rester jusqu'à plusieurs décennies chez le même assureur. Certes, ces assurés très fidèles sont minoritaires, mais leur absence dans le portefeuille d'étude biaise négativement les statistiques de durée de vie des contrats. De plus, le phénomène de censure n'est pas à négliger : pour plus de 40% des individus en portefeuille, seul un minorant de la durée de vie est connu, ce qui fausse nécessairement, à la baisse, la durée de vie moyenne du portefeuille.

Outre cette première analyse, l'étude graphique d'objets spécifiques aux modèles de durée est

essentielle à la bonne appréhension des comportements de résiliation au sein du portefeuille. La Figure 4.2 propose la fonction de survie empirique obtenue par l'estimateur de Kaplan-Meier. En abscisses se trouvent les temps t, en années, en ordonnées est représentée la fonction de survie $S(t)$. A chaque instant t est estimée la probabilité $\mathbb{P}(T \geq t)$, que les assurées soient encore en portefeuille. Lorsque $t = 0$, la fonction de survie vaut 1 puis observe une décroissance au fil des années. Il est pertinent de remarquer que la fonction de survie décroît d'autant moins vite que le temps t augmente : le risque de résiliation s'avère plus prononcé sur les premières années de la vie du contrat. De plus, les estimateurs de la fonction de survie ne peuvent proposer des prédictions de $\mathbb{P}(T \geq t)$ pour un t supérieur à celui présent dans les données. Ainsi, la fonction de survie n'est estimée que jusqu'à $t \approx 12$ ans. Hormis au-delà de 12 ans, la quantité de données s'avère suffisante pour disposer d'intervalles de confiance très fins, ces derniers ne sont donc pas représentés.

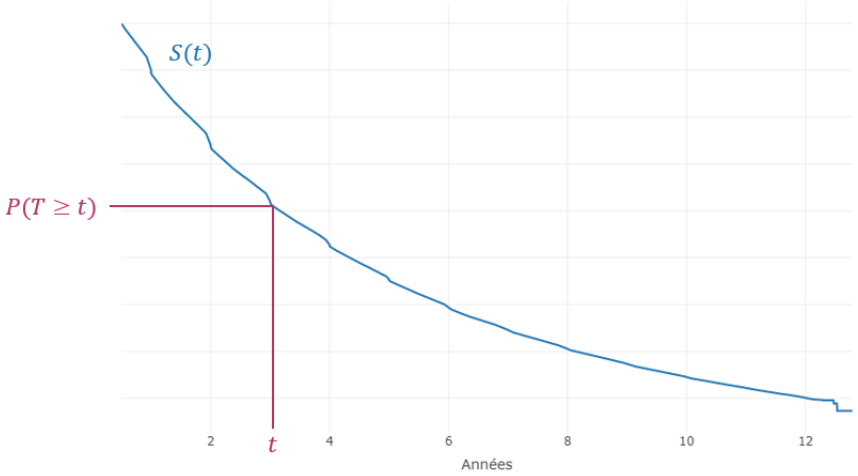


FIGURE 4.2 – Estimateur empirique de la fonction de survie par Kaplan-Meier

4.4.2 Analyse segment par segment

Une analyse segment par segment permet, avant l'étude des courbes de survie empirique et la modélisation, d'identifier les modalités sous représentées afin de réaliser les regroupements adaptés, de distinguer des comportements atypiques, en termes de durée de vie moyenne et de censure, et de dégager les tendances générales, qui seront à confirmer par des analyses plus fines. Dans l'intention de ne pas alourdir le mémoire par des statistiques redondantes à ce qui a été observé lors de l'analyse exploratoire réalisée dans le cadre de la probabilité de rétention Section 3.4.1, et à ce qui sera amené ultérieurement Section 4.4.3, seuls quelques exemples des analyses segment par segment effectuées sont proposés dans cette partie.

Les évolutions de la durée de vie moyenne, et du taux de censure, en fonction de l'âge de l'assuré à sa souscription sont disponibles Figure 4.3. Excepté le comportement atypique des 18-20 ans, la durée de vie est croissante avec l'âge, en cohérence avec l'allure des taux de résiliation étudiés lors du chapitre précédent. De plus, les taux de censure suivent la tendance observée par la durée de vie moyenne : plus un segment dispose d'une durée de vie élevée et plus un nombre important des individus le composant n'ont pas encore résilié au moment de l'étude, et sont donc censurés.

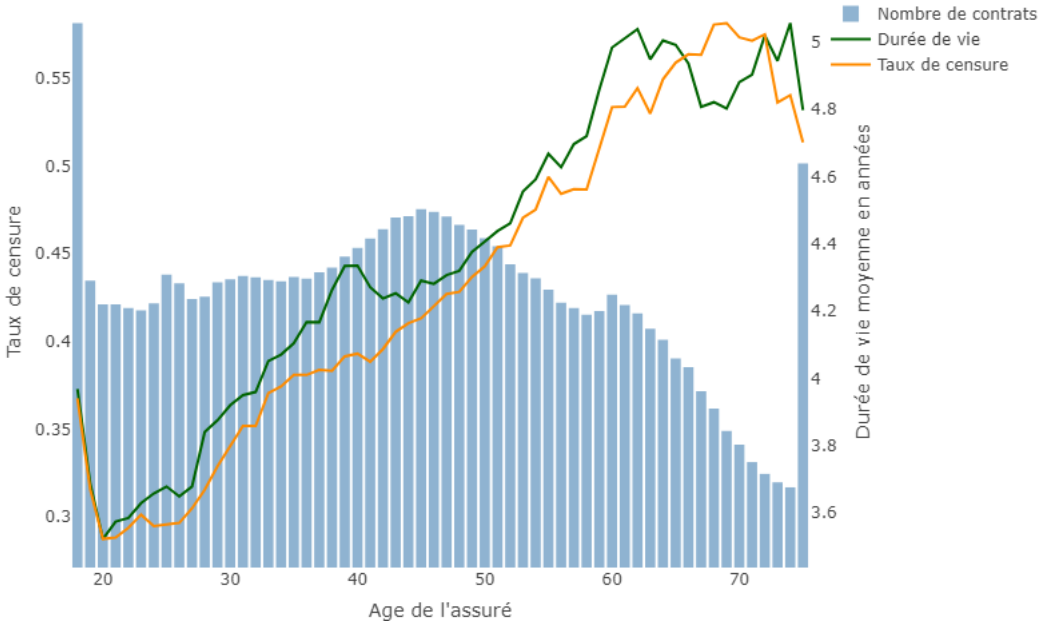


FIGURE 4.3 – Âge du conducteur - Durée de vie moyenne et taux de censure

Sur l'âge du véhicule également, les comportements des assurés Figure 4.4 sont consistants. Plus un assuré souscrit avec une voiture récente et plus ce dernier est susceptible en moyenne de rester longtemps en portefeuille. En effet, un assuré disposant d'une voiture récente, donc plus fiable, sera moins amené à devoir remplacer son véhicule, souvent synonyme de changement d'assureur. De plus, la possession d'un véhicule récent, souvent plus coûteux, témoigne du besoin de l'assuré, sur le long terme, de disposer d'un véhicule, et donc d'une assurance. Il est à relever que les expositions de chacun des segments étant moins élevées, des regroupements plus stricts que ceux opérés lors de l'étude de la résiliation sont nécessaires à la qualité des estimations. Les assurés de plus de 75 ans et les véhicules de plus de 25 ans sont ainsi rabattus à la plus grande modalité.

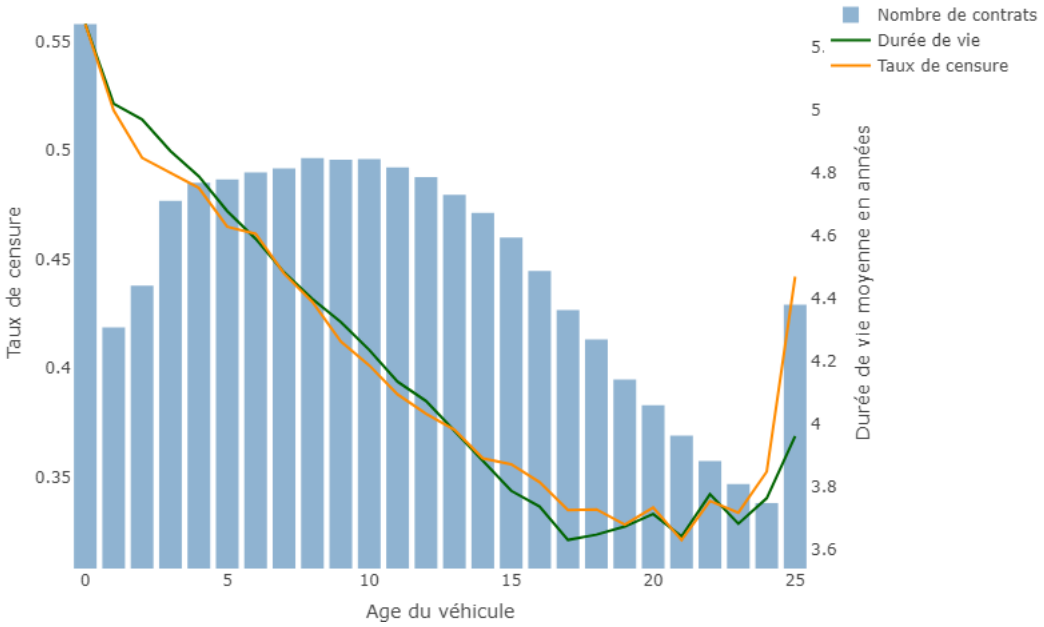


FIGURE 4.4 – Âge du véhicule - Durée de vie moyenne et taux de censure

Le dernier exemple proposé dans cette section est celui du bonus supplémentaire. Pour rappel, cette variable catégorielle présente 4 modalités. La modalité 4 correspond à une absence de bonus supplémentaire et plus la valeur prise est faible et plus le bonus est conséquent. Comme observé Figure 4.5, les durées de vie et les taux de censure sur les deux premières modalités semblent peu cohérentes. Les expositions de ces deux segments sont faibles ce qui induit une fiabilité limitée dans les informations portées par cette variable. De plus, le taux de censure sur la première modalité avoisine les 80%, ce qui est excessif pour le bon fonctionnement des différents modèles de durée.

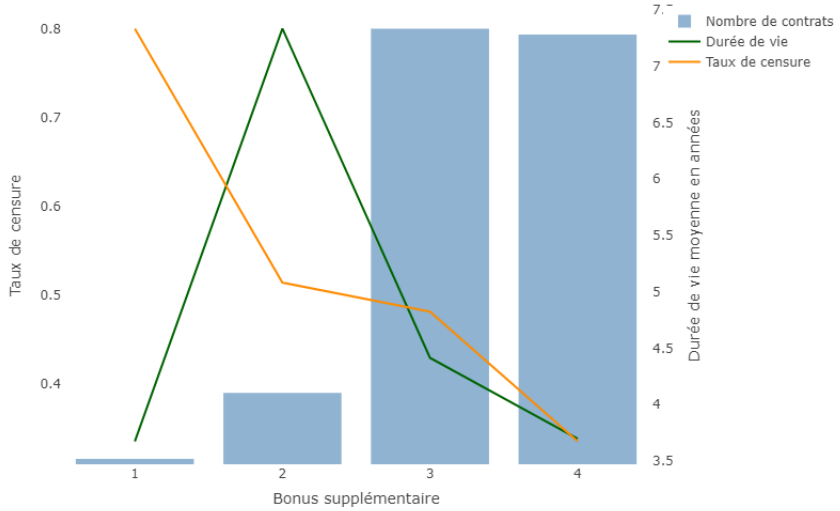


FIGURE 4.5 – Bonus supplémentaire - Durée de vie moyenne et taux de censure

Ainsi, la variable bonus supplémentaire est transformée en indicatrice, 0 pour les assurés ne bénéficiant pas de ce bonus et 1 pour ceux en bénéficiant, quel qu'en soit le niveau. La Figure 4.6 permet d'affirmer que les expositions de ces deux segments sont maintenant suffisantes et que les taux de censure restent raisonnables et cohérents avec la durée de vie moyenne.

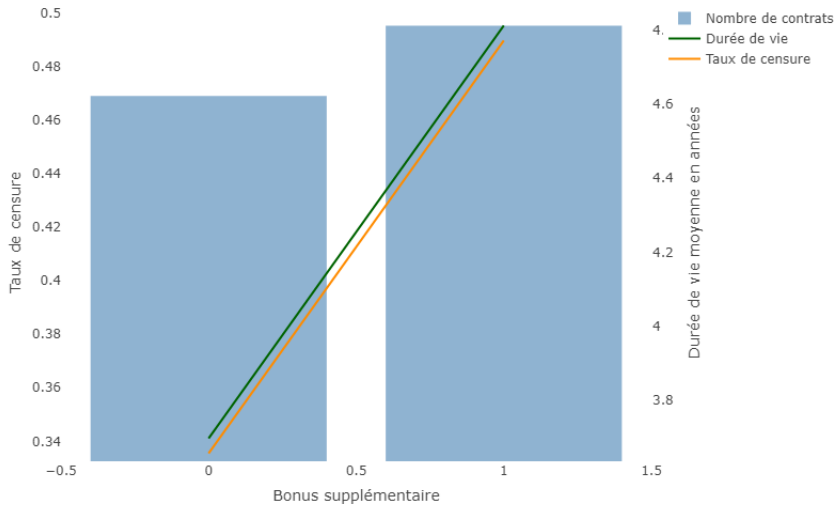


FIGURE 4.6 – Indicatrice de bonus supplémentaire - Durée de vie moyenne et taux de censure

4.4.3 Courbes de survie empirique

Il a été abordé Section 4.1 la nécessité de l'étude d'objets spécifiques pour l'appréhension de la durée de vie. Bien que l'analyse des segments, de leurs expositions, taux de censure et durées moyennes permette une première compréhension des comportements moyens, et soit essentielle aux transformations garantissant la fiabilité des données, l'analyse des courbes de survie empirique reste l'outil de visualisation statistique le plus informatif dans le cadre de l'étude de durées. Ainsi, les survies empiriques, en fonction des différentes modalités des covariables, estimées à partir de Kaplan-Meier, sont proposées dans cette partie. Les enjeux de l'étude de ces courbes sont multiples. Elle permet premièrement d'observer la forme des survies et de s'assurer de leur cohérence. Ensuite, les modalités aux fonctions de survies empiriques proches pourront être regroupées. Ce travail permet également d'identifier les variables dont certaines modalités sont récentes. Finalement, l'hypothèse de proportionnalité peut être validée ou invalidée par l'étude d'une transformation des fonctions de survie comme explicité Section 4.3.4. A moins d'avoir recours à des regroupements, l'étude des survies empiriques peut être menée uniquement sur les variables catégorielles.

Caractéristiques des assurés

Dans un premier temps, en regroupant la variable continue de l'âge des assurés en quelques modalités, il est possible de visualiser Figure 4.7 les fonctions de survies empiriques des individus selon leur classe d'âge. Les assurés jeunes observent une décroissance de leur fonction de survie d'autant plus rapide, quand les assurés les plus âgés disposent d'une fonction de survie quasiment linéaire. Au bout de 3 ans, un assuré de plus de 60 ans est une fois et demie plus susceptible qu'un assuré de moins de 30 ans d'être encore en portefeuille. Ce rapport est au-delà de deux après 8 ans.

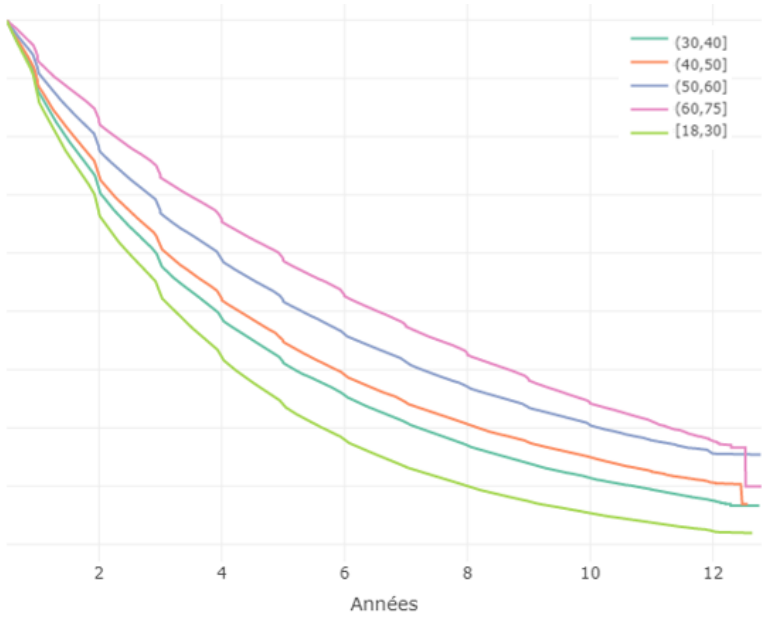
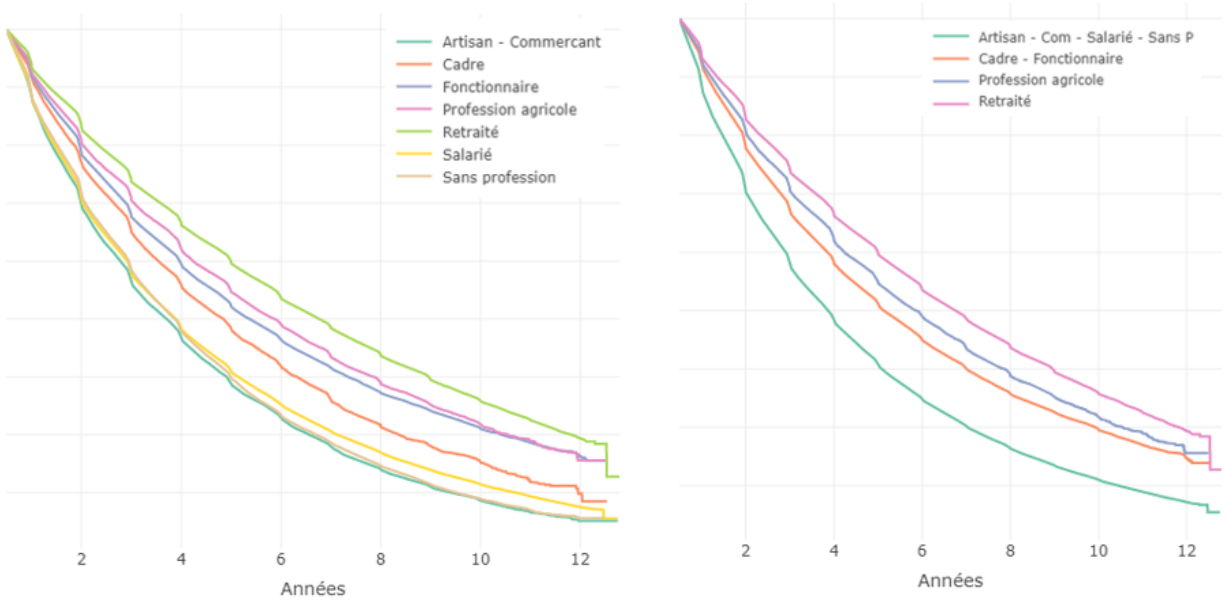


FIGURE 4.7 – Age du conducteur - Survie empirique Kaplan-Meier

Ensuite, les fonctions de survie selon les différentes catégories socioprofessionnelles sont représentées. Figure 4.8a, il est possible de constater que les catégories Artisan - Commercant, Sans profession et Salarié d'une part, et Fonctionnaire et Cadre d'autre part observent des comportements de survie proches. Des tests du log-rank sont réalisés pour éprouver l'hypothèse nulle affirmant que les courbes de survie sont identiques. Bien que l'hypothèse nulle soit rejetée pour les deux groupes, les modalités susmentionnées sont regroupées. Les survies des quatre modalités ainsi présentées sont disponibles Figure 4.8b.



(a) Avant regroupement

(b) Après regroupement

FIGURE 4.8 – Catégorie socioprofessionnelle - Survie empirique Kaplan-Meier

Caractéristiques du véhicule

En regroupant les modalités prises par l'âge du véhicule en quatre catégories, Figure 4.9, l'impact de l'ancienneté du véhicule lors de la souscription est clair : un véhicule récent induit en moyenne une fidélisation de l'assuré. Cependant, une fois le véhicule âgé de plus de 10 ans, que le véhicule soit plus ou moins ancien ne semble pas jouer considérablement sur l'allure de la survie. Quand les survies des véhicules de moins de 5 ans, de 5 à 10 ans et de 10 à 15 ans sont clairement distinctes, la courbe des véhicules âgés de 15 à 25 ans, bien que inférieure à celle des 10 à 15 ans, en est très proche. Cela remet en cause l'hypothèse de log-linéarité pour l'âge du véhicule, qui suppose que la variation du risque est constante.

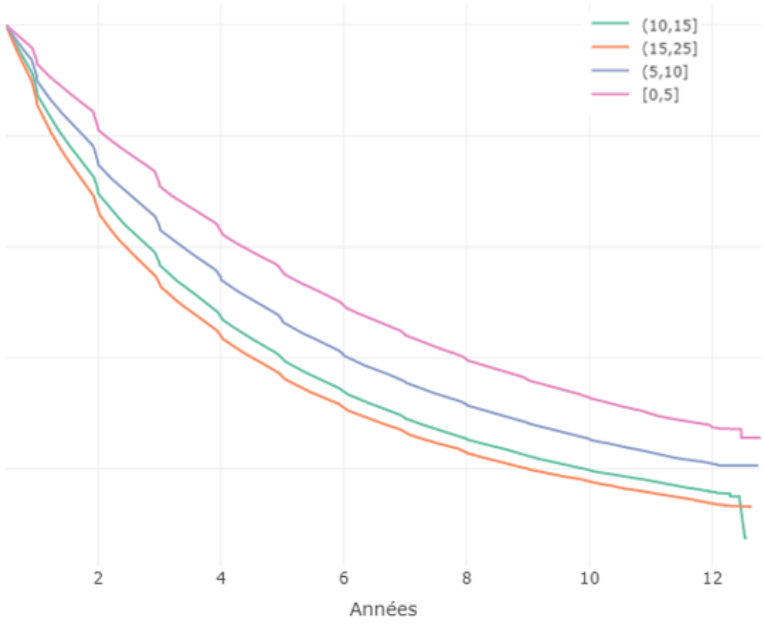
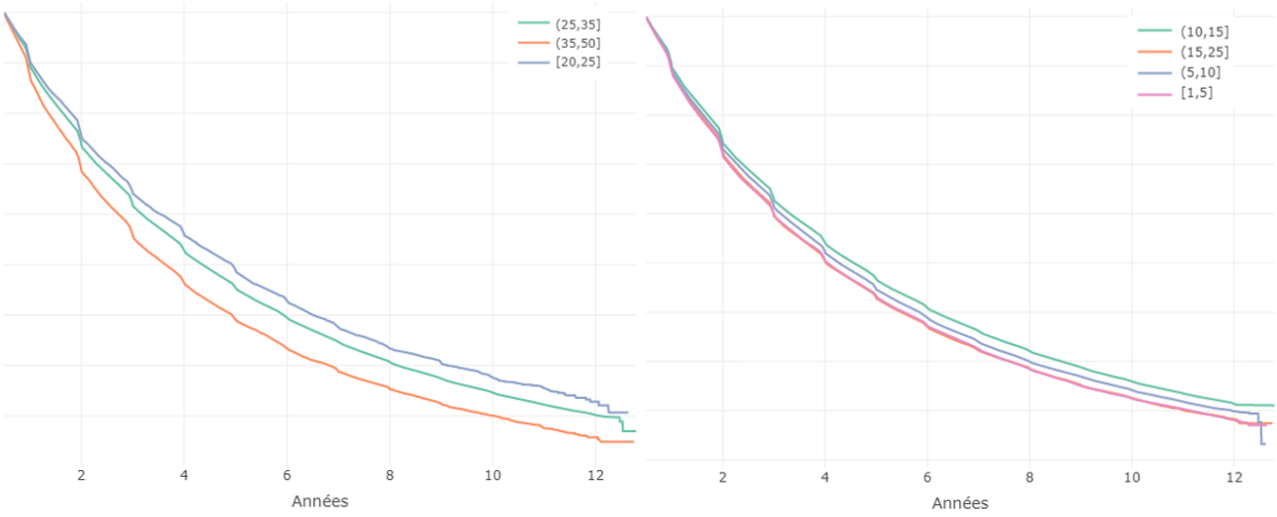


FIGURE 4.9 – Âge du véhicule - Survie empirique Kaplan-Meier

Également par regroupement de modalités, il est possible d'observer les différents survies en fonction

du groupe SRA Figure 4.10a, et de la classe de prix Figure 4.10b du véhicule. Les courbes des différents groupes sont toutes proches et aucun comportement clair ne semble s'en dégager. En moyenne, ces deux variables semblent être peu explicatives en termes de fidélisation des assurés. Cette conclusion avait déjà été posée dans le cadre de l'étude de la résiliation à un an.

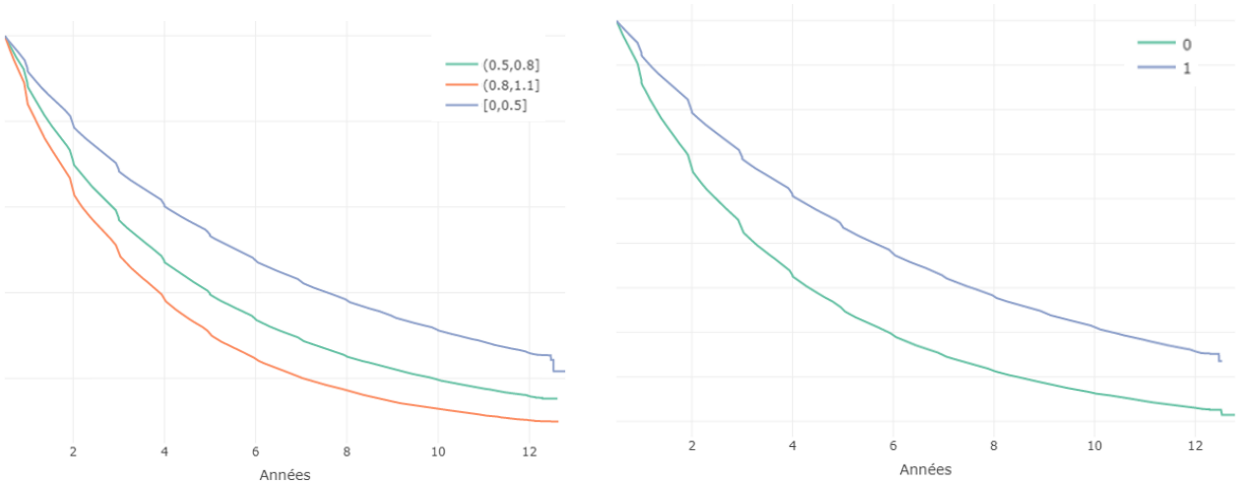


(a) Groupe SRA (b) Classe de prix

FIGURE 4.10 – Caractéristiques véhicules - Survie empirique Kaplan-Meier

Comportement de sinistralité de l'assuré

Les survies diffèrent en moyenne selon le comportement responsable de l'assuré. Le coefficient de réduction majoration est représenté Figure 4.11a après regroupement en trois classes. Comme cela avait été observé lors de l'étude des comportements de résiliation à un an, plus un assuré dispose d'un coefficient de réduction majoration conséquent et plus ce dernier est fidèle en moyenne. Cela s'explique en partie par l'âge de l'assuré, la réduction augmentant avec le nombre d'années sans sinistre responsable, et donc mécaniquement avec l'ancienneté du permis de conduire. L'indicatrice de bonus supplémentaire permet également une discrimination claire des profils plus ou moins fidèle, un assuré disposant d'un bonus supplémentaire, quel qu'il soit, présente une probabilité de rétention plus élevée tout au long de la vie de son contrat.



(a) Coefficient Bonus Malus (b) Indicatrice de bonus supplémentaire

FIGURE 4.11 – Comportement de sinistralité - Survie empirique Kaplan-Meier

Caractéristiques du contrat

Les fonctions de survie en fonction des différentes caractéristiques du contrat, à la souscription, sont également étudiées. Figure 4.12 sont représentées les survies empiriques de Kaplan-Meier en fonction de la garantie choisie par l'assuré lors de la souscription de son contrat. Les différences de survie sont nettes : plus un assuré choisit une couverture complète et plus ce dernier tend, en moyenne, à être fidèle. Au bout de 6 ans, un assuré ayant souscrit un contrat avec 7 garanties est deux fois plus susceptible d'être encore présent en portefeuille qu'un assuré ayant préféré une couverture au tiers.

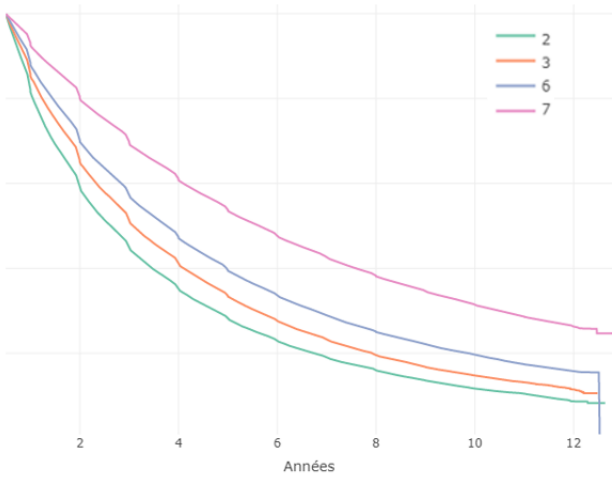
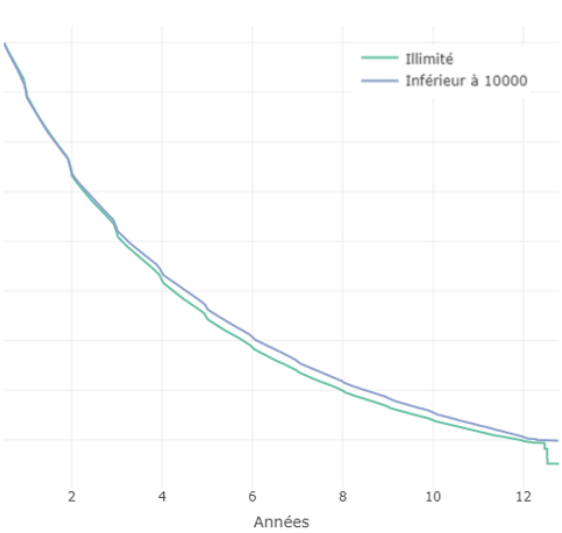
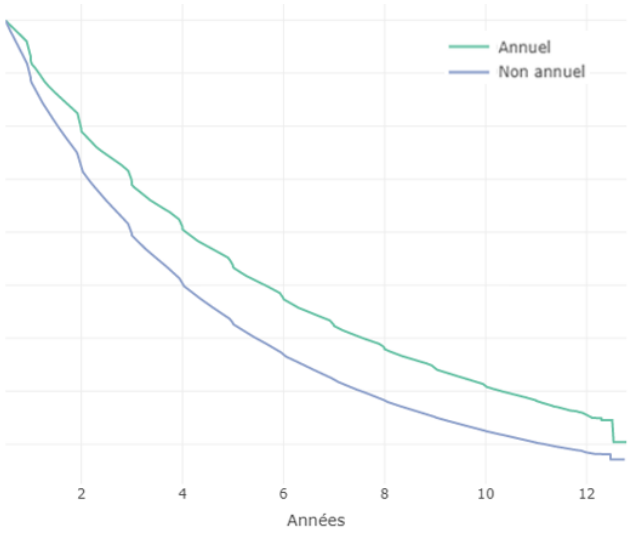


FIGURE 4.12 – Typologie de couverture - Survie empirique Kaplan-Meier

Outre le degré de couverture, le kilométrage et le fractionnement des paiements permettent de discriminer les profils en fonction de leur survie comme observé Figures 4.13a et 4.13b. Cependant, les écarts constatés entre les fonctions de survie pour les conducteurs au kilométrage limité et illimité sont faibles. Un assuré au kilométrage illimité observe une rétention sur le long terme légèrement inférieure à celle d'un assuré dont le kilométrage est limité à 10 000 kilomètres. La distinction de survie entre les assurés dont le paiement est annuel et ceux dont le paiement est fractionné est bien plus marquée. Les assurés au paiement annuel sont plus fidèles en moyenne : ils sont 15% plus susceptibles d'être encore en portefeuille au bout de 2 ans et 35% au bout de 8 ans.



(a) Kilométrage

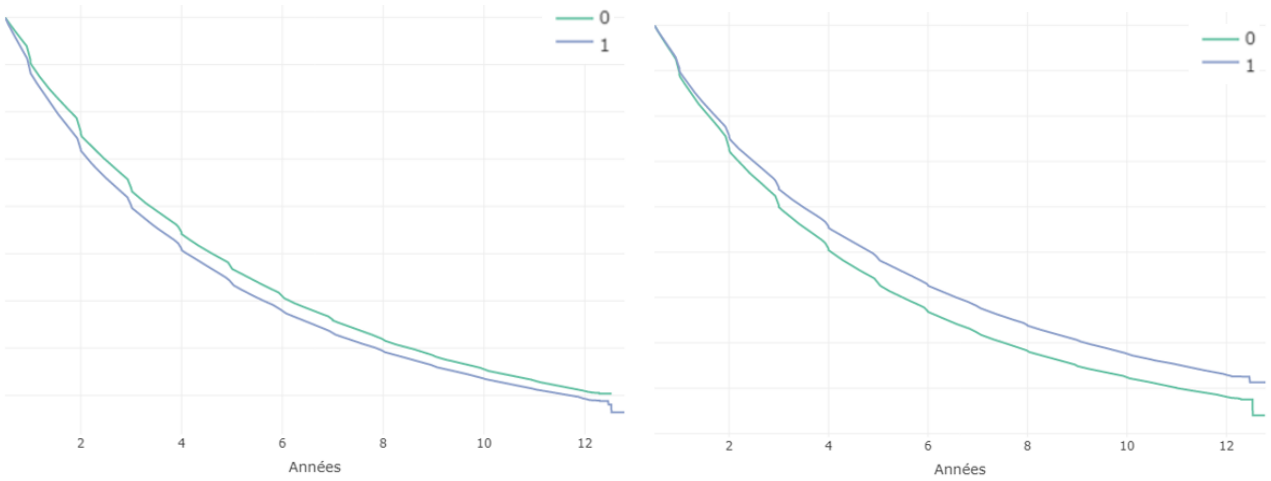


(b) Fractionnement des paiements

FIGURE 4.13 – Caractéristiques du contrat - Survie empirique Kaplan-Meier

Multi produit

Finalement, la détention de plusieurs contrats d'assurance est également vecteur de différents comportements de fidélisation. Figure 4.14a, un assuré détenant un autre contrat auto chez le même assureur semble très légèrement moins fidèle en moyenne. En revanche, Figure 4.14b il apparaît que la possession de produits d'assurance, autres que les contrats auto, augmente la fidélisation de l'assuré sur le long terme. En effet, les individus qui disposant d'au moins un contrat supplémentaire présentent une courbe de survie supérieure à celle des clients mono détenteurs.



(a) Présence d'un deuxième contrat auto

(b) Détention d'autres produits d'assurance

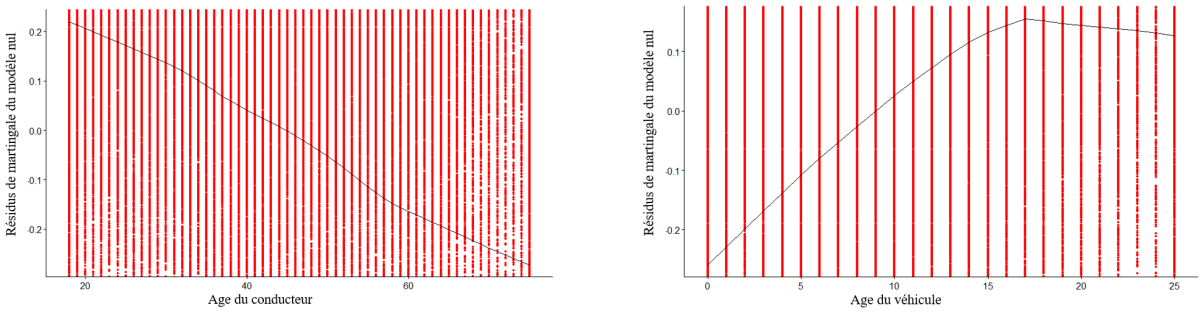
FIGURE 4.14 – Multi produit - Survie empirique Kaplan-Meier

4.5 Modélisation de la durée de vie sur le portefeuille

Les aspects théoriques nécessaires à l'étude de durées ont été développés dans les trois premières parties de ce chapitre. La Section 4.3 définit, entre autres, la forme du modèle de Cox, l'estimation de ses paramètres, ses hypothèses et les méthodes de validation associées. La Section 4.4 d'analyses exploratoires a permis d'appréhender la durée de vie des contrats, au global et selon les différentes caractéristiques des assurés. Au travers de ces analyses statistiques, les variables potentiellement discriminantes en termes de durée de vie des contrats ont pu être identifiées. Cette dernière partie s'attache à la mise en œuvre d'un modèle de Cox sur le portefeuille. Dans un premier temps, les hypothèses relatives au modèle sont brièvement redéfinies et vérifiées. Ensuite, un modèle est entraîné, puis les coefficients estimés sont analysés. Finalement, des tests de validité et de performance du modèle sont présentés.

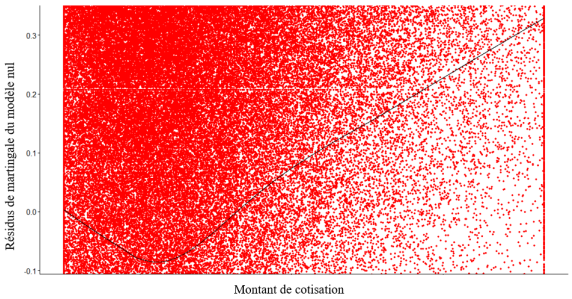
4.5.1 Hypothèse de log-linéarité

Le modèle de Cox repose comme énoncé Section 4.3.4 sur l'hypothèse que les covariables continues observent une relation linéaire avec le log du risque instantané. Cette hypothèse peut être validée, ou invalidée, au travers de l'étude des résidus de martingale du modèle nul, représentés en fonction des valeurs prises par la covariable. Les résidus de martingale lissés du modèle nul sont tracés pour les trois covariables continues exploitées. Les Figures 4.15a, 4.15b et 4.15c proposent respectivement les résidus lissés obtenus pour l'âge du conducteur, l'âge du véhicule et le montant de cotisation. En rouge sont représentés les résidus de martingale et en noir leur courbe lissée, ajustée par une méthode de régression non paramétrique. Pour l'âge, la courbe lissée des résidus est globalement linéaire ce qui valide, pour cette covariable, l'hypothèse de log-linéarité. Au niveau de l'âge du véhicule, Figure 4.15b, la linéarité est compromise au-delà de 17 ans. Les résidus lissés du modèle nul, représentés en fonction du montant cotisation Figure, 4.15c apparaissent clairement comme non linéaires. Dans les deux cas mentionnés, une fonctionnelle spécifique doit être appliquée aux covariables. Le recours aux fonctions splines cubiques, comme proposé par Heinlz et Kaider [17], permet d'assouplir l'hypothèse de log-linéarité.



(a) Âge du conducteur

(b) Âge du véhicule



(c) Montant de cotisation

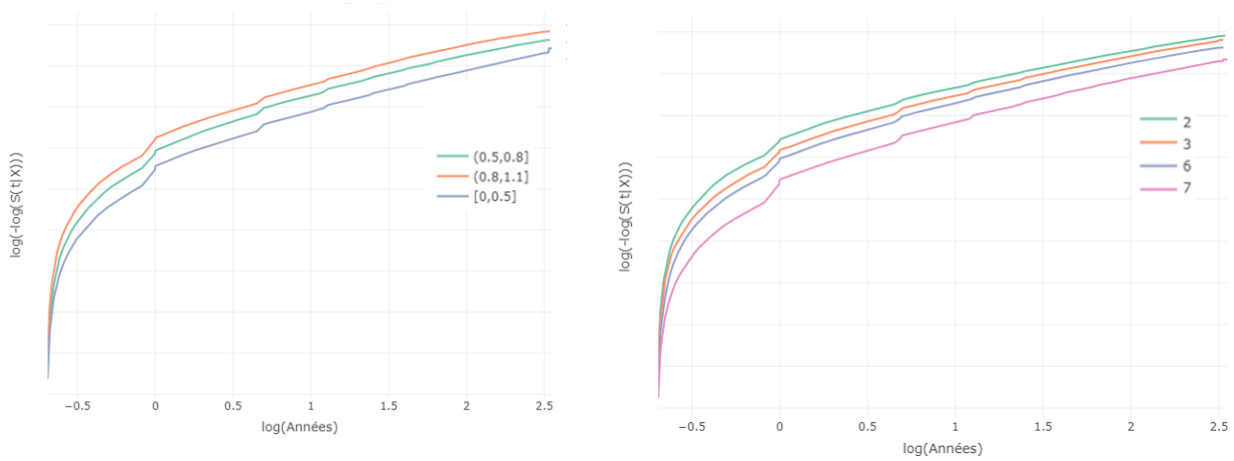
FIGURE 4.15 – Étude de la log-linéarité des covariables continues

4.5.2 Hypothèse de risques proportionnels

Le modèle de régression de Cox, aussi appelé modèle à hasards proportionnels, repose sur l'hypothèse forte que l'effet d'un facteur de risque ne dépend pas du temps. La théorie relative à cette hypothèse fondamentale est proposée Section 4.3.4. La vérification de cette dernière diffère pour les variables catégorielles et les variables continues.

Variables catégorielles

Pour les variables catégorielles uniquement, l'hypothèse peut être validée ou invalidée par l'observation des courbes $\ln[-\ln(S(t|X))]$, pour les différentes valeurs prises par la variable X et en fonction du temps. Ces courbes doivent être parallèles, ou à minima ne pas se croiser. Deux exemples sont proposés Figure 4.16, où les courbes $\ln[-\ln(S(t|.))]$ sont représentées, en fonction des modalités prises par le coefficient de réduction majoration d'une part, et par celles prises par le nombre de garanties d'autre part. Bien que les courbes soient confondues dans la partie extrême gauche de l'axe des abscisses, ces dernières sont rapidement disjointes puis parallèles. Ainsi, l'hypothèse de proportionnalité est validée pour ces deux covariables.



(a) Coefficient de réduction majoration

(b) Nombre de garanties

FIGURE 4.16 – Vérification de l'hypothèse de hasard proportionnel - Variables catégorielles

Un exemple de variable ne validant pas l'hypothèse de proportionnalité est maintenant proposé Figure 4.17, qui représente l'indicatrice de sinistralité précédente de l'assuré.

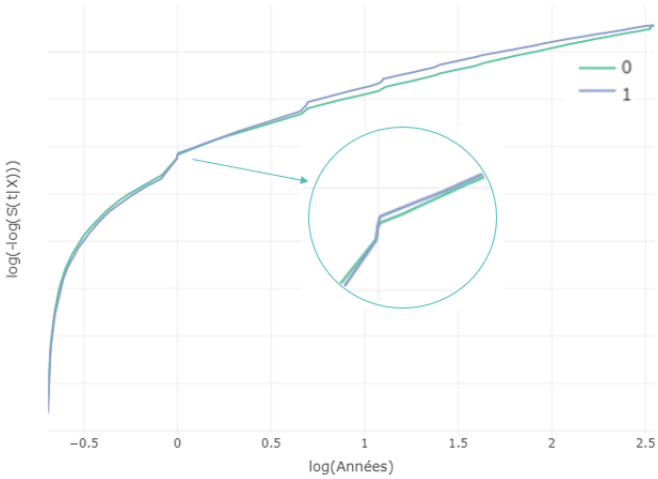


FIGURE 4.17 – Indicatrice de sinistralité précédente

Les courbes $\ln[-\ln(S(t|\text{pas de sinistre}))]$ et $\ln[-\ln(S(t|\text{sinistre}))]$, représentées en fonction du logarithme du temps $\ln(t)$, ne sont pas parallèles et se croisent même aux alentours de $\ln(t) = 0$. Ainsi, pour cette variable, l'hypothèse de hasard proportionnel n'est pas validée, et cette dernière ne pourra être exploitée dans le cadre du modèle de Cox.

Variables continues

Pour les variables continues, les résidus de Schoenfeld, dont le concept a été explicité Section 4.3.4, sont utilisés. Il est à relever que ces résidus se calculent une fois le modèle de Cox implémenté, néanmoins, dans un souci de fluidité de lecture, la validation de l'hypothèse de proportionnalité des risques est présentée en amont. Trois variables continues seront exploitées dans le modèle de Cox : l'âge du conducteur, de son véhicule et le montant annuel de sa première cotisation. Ainsi, l'hypothèse de proportionnalité du risque doit être vérifiée pour ces trois covariables. La Figure 4.18 propose d'observer, pour le montant de cotisation, une approximation du paramètre $\beta(t)$, représentée par la courbe noire. L'hypothèse de proportionnalité est validée lorsque le paramètre $\beta(t)$ approximé est constant en fonction du temps, ce qui confirme la supposition que les paramètres du modèle de régression de Cox ne dépendent pas de la durée. Pour le montant de cotisation, il est possible d'observer que le paramètre $\beta(t)$ est légèrement dépendant du temps sur les premières centaines de jours. Néanmoins, le paramètre se stabilise rapidement, puis reste constant, ce qui permet de valider l'hypothèse de proportionnalité du risque au cours du temps.

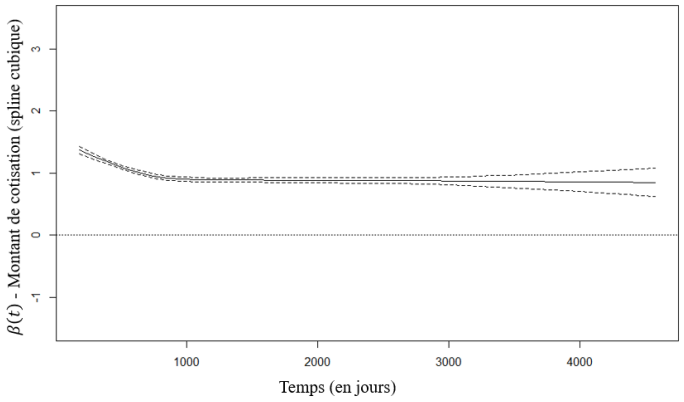
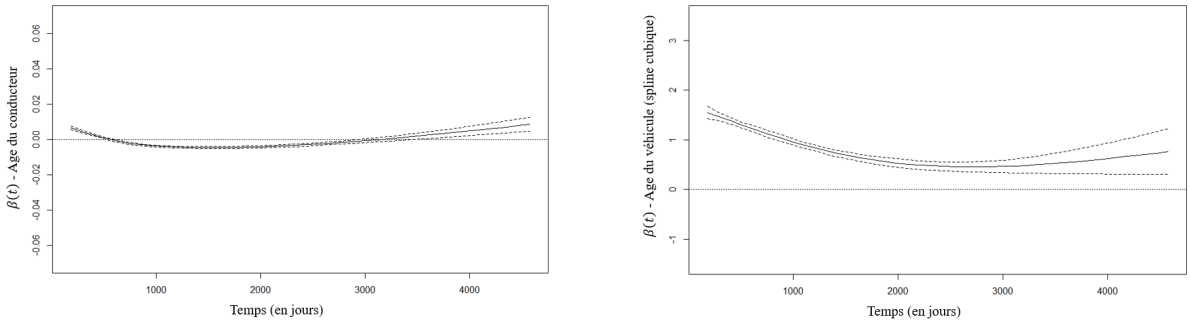


FIGURE 4.18 – Montant de cotisation - Résidus de Schoenfeld

Les résidus de Schoenfeld sont également représentés pour l'âge du conducteur et de son véhicule Figure 4.19. Bien que les deux courbes des coefficients $\beta(t)$ ne soient pas complètement horizontales, les variations des paramètres observés au cours du temps sont jugées suffisamment faibles pour ne pas rejeter l'hypothèse de proportionnalité du risque.



(a) Âge du conducteur

(b) Âge du véhicule

FIGURE 4.19 – Résidus de Schoenfeld

4.5.3 Interprétation et évaluation

Sélection des variables

Les travaux menés en amont de la calibration du modèle ont permis de dégager un premier ensemble de variables explicatives. Lors de l'analyse exploratoire, la représentation des courbes de survie notamment, Section 4.4.3, fournit un jeu de covariables au sein desquelles les modalités influent sur la durée de vie des contrats. Ensuite, les variables sélectionnées doivent satisfaire les hypothèses inhérentes à la validité du modèle de Cox. Ainsi, les covariables ne respectant pas l'hypothèse de proportionnalité ou de log-linéarité sont exclues du jeu de variables explicatives. Finalement, les variables retenues sont intégrées une à une au modèle. Les coefficients estimés se doivent d'être cohérents et statistiquement significatifs, sans quoi la covariable ne peut être conservée. De plus, l'ajout d'une covariable se fait en fonction de l'amélioration significative de la la concordance, principale métrique d'évaluation des modèles de durée. A l'issue de ce travail de sélection de variables et de calibration, un modèle de Cox est obtenu. Ce dernier repose sur sept covariables, dont trois sont continues. La suite de cette partie vise à interpréter et à évaluer le modèle ainsi construit.

Interprétation des coefficients

Comme exposé dans la partie théorique de ce chapitre, l'une des particularités qui rendent intéressante la régression de Cox consiste en la lisibilité de ses sorties. En effet, les coefficients estimés par le modèle sont naturellement interprétables, au même titre que ceux de la régression logistique par exemple. Pour rappel, le ratio de hasard entre deux modalités x_1 et x_2 d'une même covariable s'exprime $HR(x_1, x_2) = \exp(\beta(x_1 - x_2))$, et définit le risque relatif entre l'individu à la modalité x_1 et celui à la modalité x_2. Les taux de hasard du modèle obtenus sont représentés Figure 4.20.

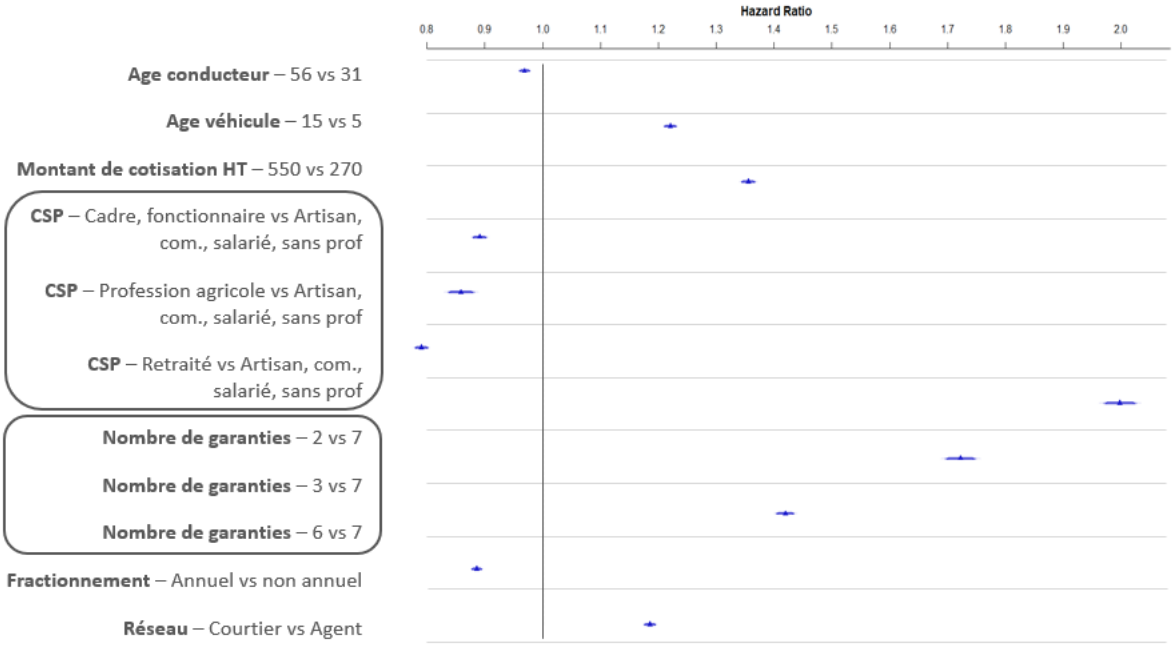


FIGURE 4.20 – Ratios de hasard - Modèle de Cox

Plusieurs analyses peuvent être menées à partir des taux de hasard :

— La catégorie socioprofessionnelle de l'assuré est un élément qui permet de différencier les comportements de fidélisation des individus. La catégorie de référence est celle regroupant les artisans, commerçants, les salariés et les sans profession qui, comme détaillé dans les analyses des courbes de survie, observent des comportements similaires. Relativement à cette référence, l'ensemble

des autres catégories socioprofessionnelles présentent des risques de sortie de portefeuille inférieurs. Les cadres et les fonctionnaires, avec un ratio de hasard de 0.9 environ, présentent un risque 10% inférieur à celui porté par le groupe de référence. Les assurés de profession agricole, puis les retraités, observent un risque d'autant plus faible avec des ratios de hasard de 0.85 et de 0.8 respectivement.

— Le degré de couverture choisi par l'assuré est également explicatif des comportements de survie des contrats. Moins un assuré est couvert, et plus ce dernier supporte un risque de sortie élevé. Un client souscrivant un contrat au tiers présente un risque de résiliation deux fois plus élevé à celui d'un assuré dont la couverture est complète.
— Un assuré dont le paiement est annuel résilie moins qu'un assuré dont le paiement est fractionné : le ratio de hasard entre ces deux modalités est de 0.87, ainsi le risque de sortie de portefeuille est diminué de 13% pour les clients s'acquittant de leur cotisation annuellement.
— Le type de réseau par lequel l'assuré a souscrit son contrat est vecteur de différents comportements : un client passé par un courtier présente un risque de sortie de portefeuille près de 20% supérieur à celui d'un assuré ayant souscrit auprès d'un agent.
— L'interprétation des ratios est moins directe pour les variables continues, d'autant plus pour celles introduites dans le modèle au travers d'une fonctionnelle spline. Le ratio pour la covariable de l'âge du conducteur est inférieur à 1 : plus un assuré souscrit âgé et moins il son risque de sortie de portefeuille est conséquent. L'âge du véhicule et le montant de cotisation ont un effet opposé : plus ces derniers sont élevés et plus l'assuré est risqué. Pour l'âge du véhicule, cet effet croissant du risque s'explique en partie par le fait qu'un véhicule ancien tendra à être remplacé, or le changement de véhicule est communément associé à un changement d'assureur. La hausse du risque de sortie de portefeuille portée par des assurés aux cotisations plus élevées se justifie par le fait que les primes élevées concernent les jeunes conducteurs, qui de par leurs changements de situation et de besoins plus fréquents, tendent à résilier plus régulièrement.

Pour les variables continues, et notamment pour celles dont une fonctionnelle spline a été appliquée, l'étude des variations du hasard relatif s'avère plus éloquente. Les modalités des autres variables sont fixées, et ainsi seul l'effet de la covariable d'intérêt sur le hasard relatif est représenté. Autrement dit, le ratio $\ln(\frac{h(t)}{h_0(t)})$ est représenté, toutes choses égales par ailleurs, en fonction des modalités de la variable en question. Les Figures 4.21 et 4.22 proposent les évolutions du hasard relatif en fonction de l'âge du véhicule, du conducteur et du montant de cotisation. Le hasard relatif est croissant en fonction de l'ancienneté du véhicule : plus un contrat est souscrit avec un véhicule âgé et plus ce dernier supporte un risque de sortie de portefeuille élevé. Néanmoins, cet effet se tasse au-delà de 15 ans, où l'augmentation du hasard relatif est plus modérée.

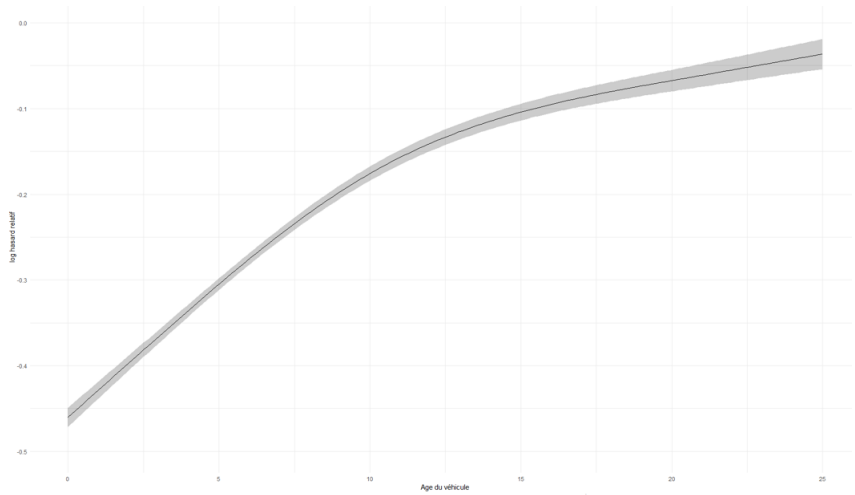
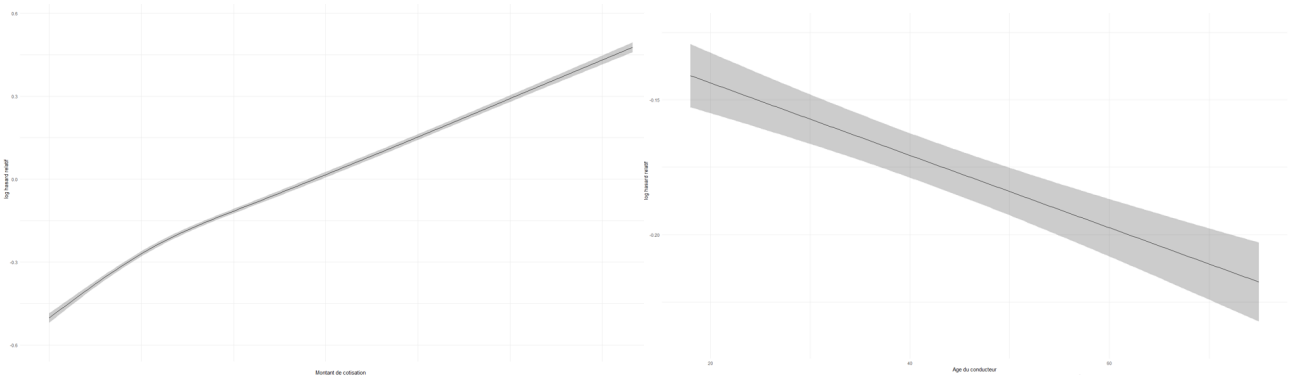


FIGURE 4.21 – Évolution du hasard relatif - Âge du véhicule

Cet effet positif sur le hasard relatif est également observé pour le niveau de cotisation Figure 4.22a. En revanche, plus un individu souscrit son contrat étant âgé, et moins ce dernier est risqué pour l'assureur en termes de résiliation. L'effet décroissant observé Figure 4.22b est purement linéaire, en effet, la covariable de l'âge du conducteur a été introduite au modèle sans transformation préalable.



(a) Montant de cotisation (b) Âge du conducteur

FIGURE 4.22 – Évolution du hasard relatif

Validité

La validité d'un modèle de Cox se détermine à partir de différents tests statistiques, tels que celui du rapport de vraisemblance ou de Wald. Ces deux tests, asymptotiquement équivalents, présentent des p-valeurs inférieures à 2.10^{-16}, ce qui appuie la significativité statistique du modèle.

Performance

Finalement, la concordance du modèle peut être obtenue par validation croisée, ce qui assure la stabilité de la métrique. La régression de Cox ainsi construite présente une concordance de 62,3%, performance très correcte dans le cadre de la modélisation de durée.

Chapitre 5

Classification des assurés et scénarios tarifaires

Sommaire

Ce chapitre propose la mise en place de divers outils d'aide à la décision, dans le cadre du pilotage d'un portefeuille d'assurance auto. Les assurés sont segmentés en classes selon d'une part, leur durée de vie a priori, et d'autre part, leur sensibilité de résiliation au prix. Cette classification donne lieu à la cartographie du portefeuille selon plusieurs critères, qui permet la mise en exergue de profils au comportement spécifique. Finalement, les prémices d'une optimisation tarifaire sont proposés au travers d'un scénario visant à augmenter le profit de l'assureur, tout en conservant son volume d'assurés sur le court et moyen termes. La dernière section de ce chapitre est consacrée aux ouvertures des travaux.

5.1 Classes de durée de vie

5.1.1 La classification ascendante hiérarchique

Principe de la classification ascendante hiérarchique

La classification ascendante hiérarchique (CAH) est une méthode de classification non supervisée. Cette dernière vise à regrouper les individus à partir d'un critère de similarité, de sorte que les classes construites soient chacune composées d'observations homogènes et qu'elles soient aussi distinctes les unes des autres que possible. Lors de l'initialisation, chaque individu constitue un groupe. A chaque itération, les observations sont regroupées selon le critère de ressemblance choisi au préalable, et ce, jusqu'à ce que l'ensemble des individus appartiennent à une unique classe.

La mise en place d'une CAH nécessite le choix du critère de similarité. La distance de Ward, qui permet de ne pas isoler les observations atypiques, est préférée. Notée W, la distance de Ward entre deux groupes G_i et G_j se définit :

$$W(G_i, G_j) = \frac{n_i n_j}{n_i + n_j} d_2(g_i, g_j)^2, \quad \forall \quad 1 \leq i, j \leq n \tag{5.1}$$

où :

— n, le nombre d'individus à classer ;
— n_i, le nombre d'individus présents dans le groupe G_i ;
— g_i, le centre de gravité, aussi appelé point moyen, du groupe G_i ;
— $d_2(g_i, g_j) = (\sqrt{g_i - g_j})^2$ la distance euclidienne.

De plus, la décomposition de Huygens est essentielle à la compréhension de l'algorithme de classification ascendante hiérarchique par critère de Ward. Soient $G = \{x_i, 1 \leq i \leq n\}$, le nuage de points comprenant l'ensemble des n observations, g son centre de gravité, et $(G_j)_{j=1}^K$, les K classes deux à deux disjointes telles que $G = \bigcup_{i=1}^K G_j$. La décomposition de Huygens permet d'exprimer l'inertie de G, dite totale, comme la somme de l'inertie inter-classe et intra-classe :

$$\begin{gathered} \mathcal{I}_{\text{totale}} = \mathcal{I}_{\text{inter}} + \mathcal{I}_{\text{intra}} \\ \frac{1}{n} \sum_{i=1}^{n} d_2(x_i, g)^2 = \frac{1}{n} \sum_{j=1}^{K} n_j d_2(g_j, g)^2 + \frac{1}{n} \sum_{j=1}^{K} \sum_{x \in G_j} d_2(x, g)^2 \end{gathered} \tag{5.2}$$

Quand l'inertie inter-classe, mesure la dispersion entre les classes, est à maximiser, l'inertie intra-classe, mesure de la variabilité des observations au sein d'une même classe, est à minimiser. Lors de l'initialisation, chaque observation constituant un groupe, l'inertie inter-classe, égale à l'inertie totale, est maximale, et l'inertie intra-classe est nulle. L'algorithme cherche ensuite à regrouper les classes selon les deux critères de minimisation du gain de l'inertie intra-classe et de la perte de l'inertie inter-classe. La décomposition de Huygens affirme l'équivalence des deux critères : minimiser la perte de l'inertie inter-classe et minimiser le gain de l'inertie intra-classe sont deux actions semblables.

L'initialisation de l'algorithme de Ward consiste en la définition de n groupes, comprenant chacun une observation. Ensuite, l'algorithme se décompose en les étapes suivantes :

1. Calcul de la matrice des distances de Ward, à partir de l'équation 5.1 : $\mathbb{W} = (W(G_i, G_j))_{1 \leq i,j, \leq n}$.
2. Réunion les deux groupes les plus proches, au sens de la distance de Ward, en une seule classe. Cela garantit la minimisation du gain de l'inertie intra-classe, et donc celle de la perte de l'inertie inter-classe, par la décomposition 5.2.
3. Remplacement des individus réunis par leur classe. La classe ainsi créée est représentée par le centre de gravité des observations qui la composent.

Les trois étapes sont répétées jusqu'à ce que l'ensemble des individus appartiennent à la même classe. Finalement, l'analyse du dendrogramme, représentation graphique des agrégations successives, et de la courbe de l'inertie représentée en fonction du nombre de classes, sont deux outils d'aide à la décision du nombre de groupes à fixer.

Principe du bagging

La classification ascendante hiérarchique repose sur le calcul de la matrice de distance $\mathbb{W} \in \mathcal{M}_n(\mathbb{R})$, dont la dimension dépend du nombre d'observations n. Dans le cadre de la classification d'un nombre conséquent d'individus, la matrice $\mathbb{W}$ peut s'avérer trop coûteuse à calculer. Dans ce cas, deux possibilités s'offrent. Une première solution serait de tirer aléatoirement un sous-ensemble d'observations, s'assurer que la répartition de l'ensemble tiré est semblable à celle du jeu de données complet, et effectuer la classification sur ce sous-ensemble. La deuxième option, dite du *bagging*, est choisie. Celle-ci consiste en le tirage aléatoire de k sous-échantillons de taille $N < n$, puis en la réalisation de la classification sur chacun k échantillons tirés. Les résultats des k classifications ascendantes hiérarchiques sont ensuite agrégés, afin d'obtenir un unique modèle. Le principal avantage de cette technique réside dans la réduction de variance du modèle construit.

5.1.2 Différents comportements de fidélisation

Le modèle de durée de vie construit Chapitre 4 a permis, au travers des travaux d'interprétation, de mieux appréhender le comportement de fidélisation à moyen et long terme des assurés. Pour aller au-delà des analyses réalisées au travers des fonctions de survies empiriques et des coefficients du modèle, une classification des assurés en fonction de leur durée de vie est effectuée.

Classification des assurés

Le modèle de Cox, élaboré Chapitre 4, repose sur une forme spécifique du taux de hasard : $h(t) = h_0(t)\exp(X\beta)$. Le risque porté par l'assuré et ses caractéristiques est transcrit uniquement au travers de la partie $X\beta$, appelée prédicteur linéaire. Plus ce dernier est élevé, et plus le risque d'une résiliation précoce de l'assuré est fort. A l'inverse, un individu au prédicteur linéaire plus faible tendra à rester longtemps en portefeuille. Ainsi, la segmentation des assurés en fonction de leur degré de fidélité est mise en place à partir des prédicteurs linéaires estimés $X\hat{\beta} \in \mathbb{R}^n$. La Figure 5.1 propose d'observer la répartition du prédicteur linéaire estimé sur les plus de 450 000 assurés.

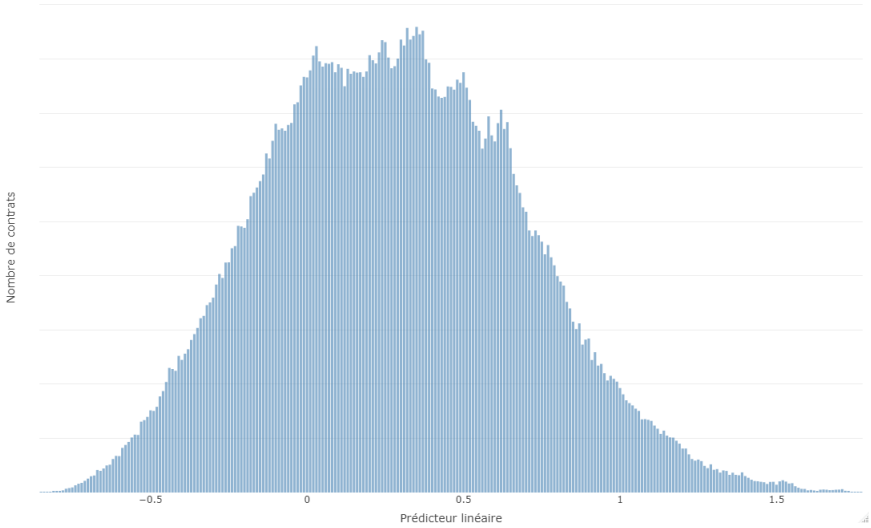
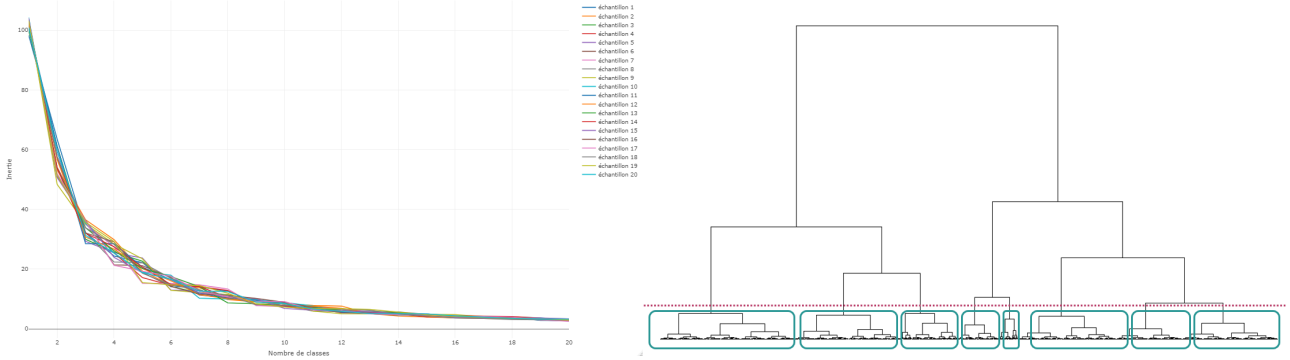


FIGURE 5.1 – Répartition du prédicteur linéaire $X\hat{\beta}$

La classification des assurés, réalisée sur le prédicteur linéaire, est mise en place comme expliqué dans la Section 5.1.1. Sous le principe du *bagging*, 50 000 observations sont tirées aléatoirement du jeu de données et la CAH est réalisée sur ce sous-échantillon. Ce processus est répété 20 fois, ce qui permet de s'assurer de la stabilité de l'agrégation à venir. Avant de poursuivre, le nombre de classes est à définir. Cela peut se faire par l'étude de l'inertie en fonction du nombre de groupes. Ce dernier est

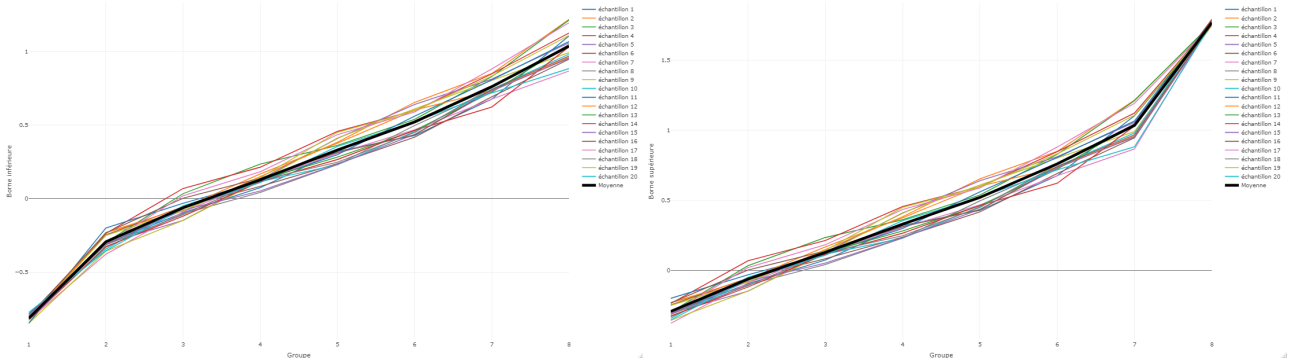
choisi après le dernier saut significatif de la courbe. Figure 5.2a sont représentées les courbes de l'inertie pour chacun des 20 échantillons tirés. L'étude de ces courbes suggère de définir huit groupes. La Figure 5.2 représente le dendrogramme, et les huit classes ainsi découpées, pour un des sous-échantillons considérés.



(a) Évolution de l'inertie (b) Dendrogramme

FIGURE 5.2 – Choix du nombre de classes

Une fois le nombre de classes défini, la classification est réalisée sur les vingt échantillons tirés aléatoirement. Pour chacune des classifications sont obtenues les bornes des huit groupes. Autrement dit, la CAH détermine que les individus du groupe 1 sont ceux ayant un prédicteur linéaire compris entre une valeur et une autre, et de même pour les sept autres groupes. Les Figures 5.3a et 5.3b proposent les bornes inférieures et supérieures de chacun des huit groupes, et ce, pour les vingt échantillons. De plus, ces figures illustrent le principe de l'agrégation : les bornes sont moyennées et apparaissent en noir. Ce sont ces dernières qui sont retenues à la suite du *bagging*.



(a) Bornes inférieures des classes (b) Bornes supérieures des classes

FIGURE 5.3 – Bornes des classes pour les 20 sous-échantillons

Il est intéressant de relever que les bornes inférieures et supérieures ainsi obtenues sont cohérentes, dans le sens où la borne inférieure du groupe j coïncide avec la borne supérieure du groupe $j-1$. Cela s'observe Figure 5.4 où les bornes agrégées sont représentées sur le même plan.

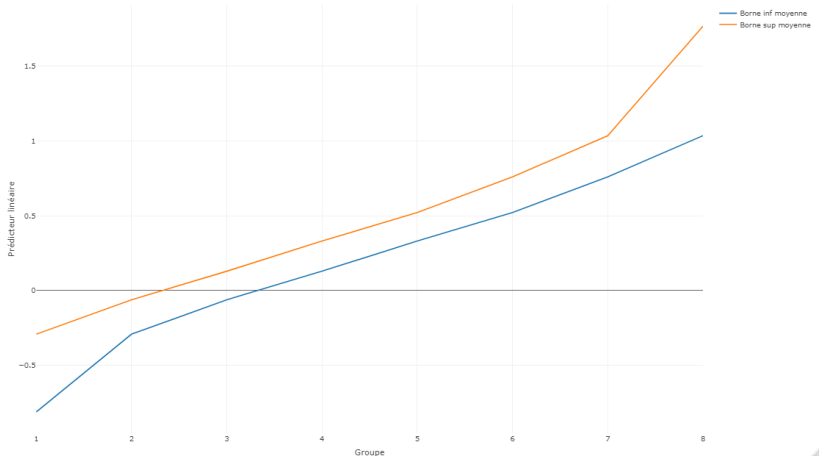
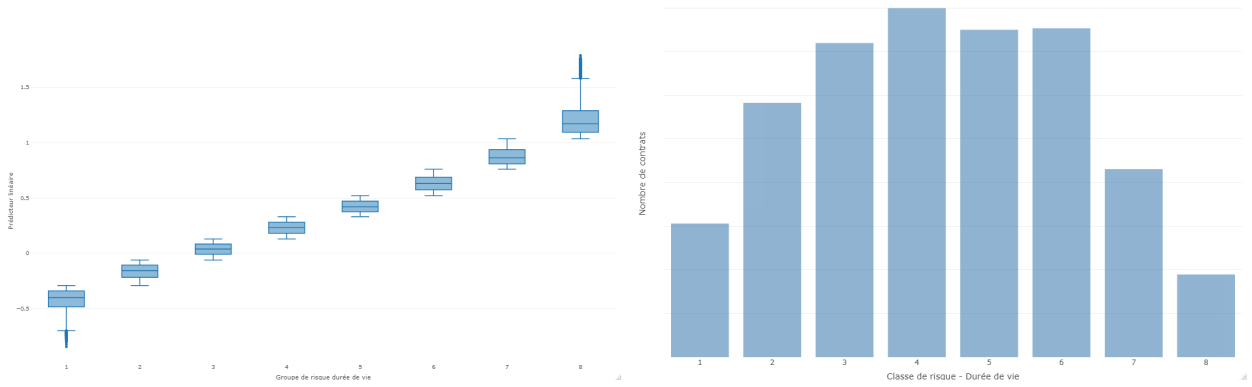


FIGURE 5.4 – Bornes des classes obtenues par *bagging* et classification ascendante hiérarchique

Finalement, la classification réalisée a permis de construire huit groupes disjoints, comprenant chacun des assurés au comportement de fidélité homogène. La Figure 5.5 donne une image d'ensemble des classes établies, avec d'une part les boxplots des différents groupes en fonction de leur prédicteur linéaire Figure 5.5a, et d'autre part la répartition des assurés au sein des huit classes Figure 5.5b. Les assurés du groupe 1 sont les plus fidèles quand ceux du groupe 8 sont les plus à risque de résilier leur contrat d'assurance rapidement.



(a) Prédicteur linéaire en fonction des classes (b) Histogramme des classes

FIGURE 5.5 – Répartition des classes

En outre, les survies empiriques des différents groupes, représentées Figure 5.6, sont clairement distinctes : les comportements de fidélisation sont très bien captés par la classification mise en œuvre. Par exemple, un assuré du groupe 1 est trois fois plus susceptible d'être encore au sein du portefeuille au bout de quatre ans qu'un assuré du groupe 8.

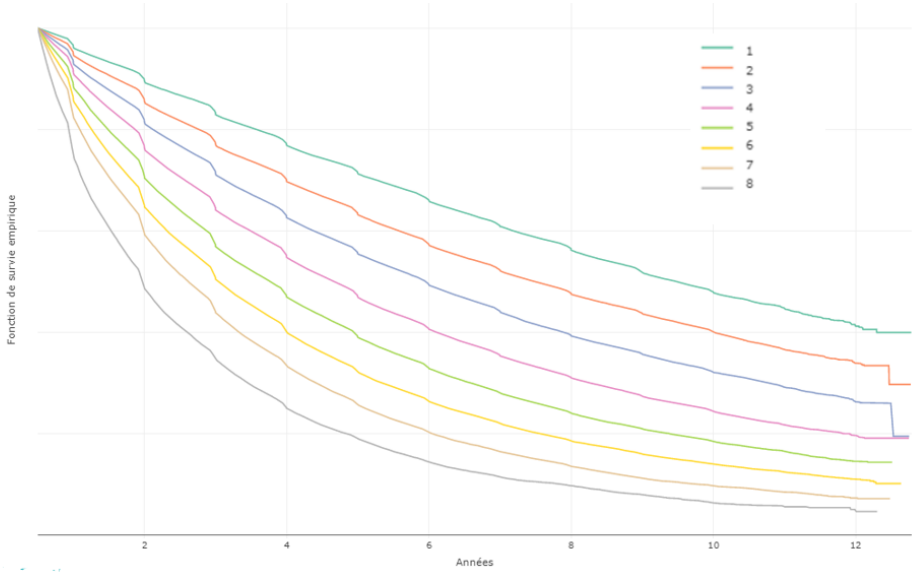
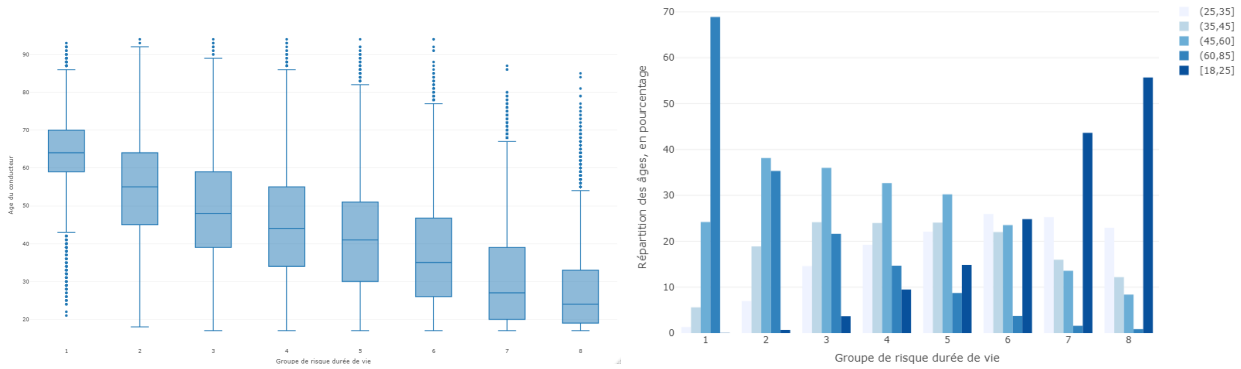


FIGURE 5.6 – Survies empiriques des classes construites

Analyse des classes

La Figure 5.7 permet l'appréhension des comportements de fidélité des assurés en fonction de leur âge lors de la souscription. Le boxplot à gauche illustre clairement le fait que plus l'assuré est jeune, et moins ce dernier sera fidèle sur les moyen et long termes. Quand l'âge médian de la classe 8, composée des individus les plus volatiles du portefeuille, est de 24 ans, celui de la classe 4 est de 44 ans, et celui de la classe 8, la plus fidèle, est de 64 ans. Cela peut également s'observer à l'aide de la répartition, en pourcentages, des différentes classes d'âge, à droite Figure 5.7b. Les assurés de moins de 25 ans, représentés en bleu foncé, sont majoritaires dans les classes 7 et 8 et ne sont que marginalement présents dans les classes les plus fidèles. Le schéma inverse s'observe chez les assurés les plus âgés, qui constituent près de 80% des individus du groupe 1 et moins de 1% du groupe 8. Les comportements sont plus nuancés chez les assurés d'âges moyens dont les répartitions sont relativement équilibrées au sein des classes de risque médianes.

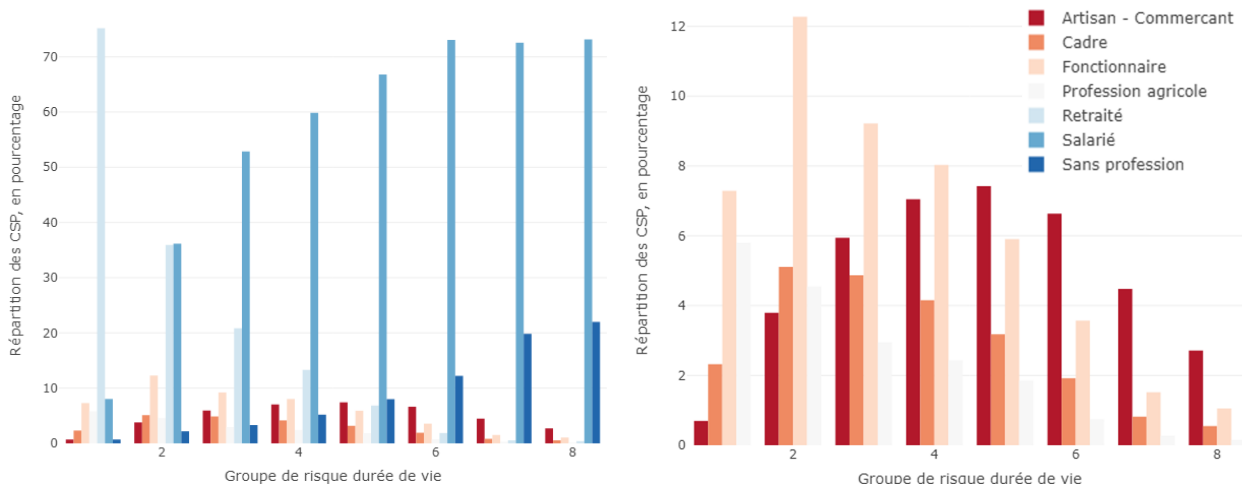


(a) (b)

FIGURE 5.7 – Analyse des classes de fidélité - Âge du conducteur

Au-delà de l'âge de l'assuré, la catégorie socioprofessionnelle constitue une caractéristique discriminante en termes de degré de fidélisation. L'analyse des catégories socioprofessionnelles au global, Figure 5.8a, permet notamment de confirmer les conclusions menées à la suite de l'étude des âges. Les retraités, représentant 75% des individus de la classe la moins risquée, sont de loin les assurés les plus fidèles. En revanche, les individus sans profession, catégorie en partie portée par les étudiants, sont très

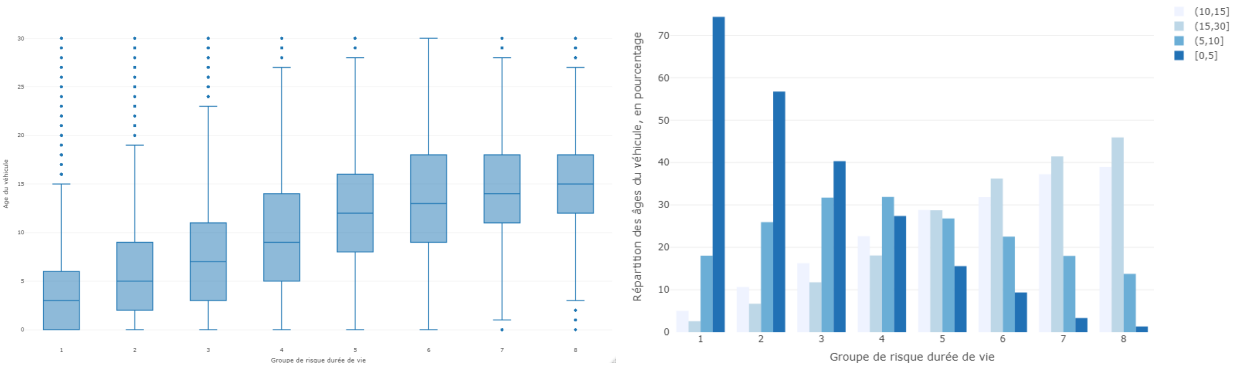
présents dans les classes les plus volatiles. Ces deux analyses sont cohérentes avec ce qui a été observé concernant les âges des assurés. Les salariés, majoritaires dans l'ensemble du portefeuille, sont moins représentés au sein des classes les plus stables. Finalement, en omettant les trois catégories susmentionnées, il est possible d'analyser avec plus de finesse les comportements des artisans, commerçants, cadres et fonctionnaires Figure 5.8b. Dans l'ensemble, la répartition de ces assurés au sein des classes de risque prend la forme d'une gaussienne : ces derniers sont majoritaires dans les classes médianes. Cette observation est à nuancer pour les fonctionnaires et les cadres, dont la distribution tend à se centrer vers les classes les moins risquées.



(a) Toutes CSP (b) Artisans, cadres et fonctionnaires

FIGURE 5.8 – Analyse des classes de fidélité - Catégorie socioprofessionnelle

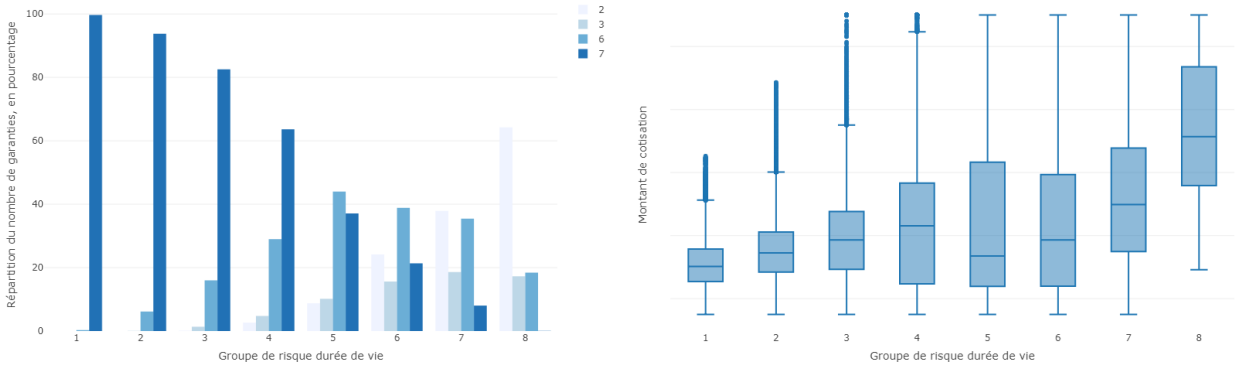
Une analyse similaire peut être menée sur l'âge du véhicule assuré Figure 5.9. Les contrats souscrits avec des véhicules âgés sont d'autant plus susceptibles d'être résiliés rapidement, phénomène en partie dû au fait que les véhicules anciens sont propices à être remplacés par leur propriétaire, or, il est d'usage qu'un changement de véhicule induise, en outre, un changement d'assureur. Les écarts des âges médians des classes adjacentes sont moins prononcés pour les classes les plus risquées, et la similarité de ces classes en termes d'âge des véhicules se confirme par la Figure 5.9b. En effet, les groupes de risque 5, 6, 7 et 8 présentent des répartitions de classes d'âges des véhicules foncièrement semblables. En revanche, les véhicules de moins de 5 ans sont nettement surreprésentés dans les classes les moins risquées, et inversement pour les véhicules les plus âgés, même si la tendance est moins marquée pour ces derniers.



(a) (b)

FIGURE 5.9 – Analyse des classes de fidélité - Âge du véhicule

Finalement, les caractéristiques du contrat peuvent être analysées, Figure 5.10a et Figure 5.10b, au travers du nombre de garanties et du montant de cotisation. Au sujet du nombre de garanties, les classes les plus stables sont constituées quasiment exclusivement d'assurés ayant choisi une couverture complète. Inversement, les couvertures au tiers sont surreprésentées dans les classes les plus volatiles. Concernant le niveau de cotisation, une tendance nette ne semble pas se dégager, sauf éventuellement pour les dernières classes dont les cotisations sont plus élevées. Cela s'explique en partie par le fait que les cotisations élevées sont supportées par les jeunes conducteurs, dont la prime pure est la plus conséquente et le coefficient de réduction majoration le moins avantageux.



(a) Nombre de garanties

(b) Montant de cotisation

FIGURE 5.10 – Analyse des classes de fidélité

5.2 Comportements d'élasticité au prix

Dans une démarche de meilleure compréhension des comportements de résiliation des assurés, la notion d'élasticité du taux de résiliation au prix est introduite. De la même façon que précédemment, les assurés sont segmentés en différentes classes de sensibilité, ce qui permet l'étude de leur composition.

5.2.1 Élasticité du taux de résiliation au prix

Soient P, le montant de cotisation de l'assuré, et $r(.)$, la fonction de $\mathbb{R}^+$ dans $[0;1]$ telle que $r(P)$ corresponde au taux de résiliation de l'individu lorsque son montant de cotisation est de P. Alors, l'élasticité du taux de résiliation au prix se définit comme :

$$E_P = \frac{\partial r(P)}{\partial P}.\frac{P}{r(P)} \tag{5.3}$$

La quantité E_P permet de mesurer la sensibilité de l'assuré lorsque ce dernier subit un choc sur le montant de sa cotisation. Elle correspond au pourcentage de la variation du taux de résiliation lorsque le montant de cotisation augmente de 1%, toutes choses égales par ailleurs. Ainsi, en notant P_i le montant de cotisation appliqué à l'assuré i, l'élasticité du taux de résiliation au prix peut se calculer pour l'ensemble des individus en portefeuille :

$$E_{P_i} = \frac{r(P_i(1+1\%)) - r(P_i)}{P_i(1+1\%) - P_i}.\frac{P_i}{r(P_i)}, \quad \forall i = 1, ..., n$$

L'élasticité est calculée sur les plus de 300 000 contrats en cours lors de l'extraction des données. Le modèle de résiliation, construit Chapitre 3 et formalisé dans la formule 5.3 par la fonction $r(.)$, est entraîné sur l'ensemble des états clos. La répartition des élasticités est proposée Figure 5.11.

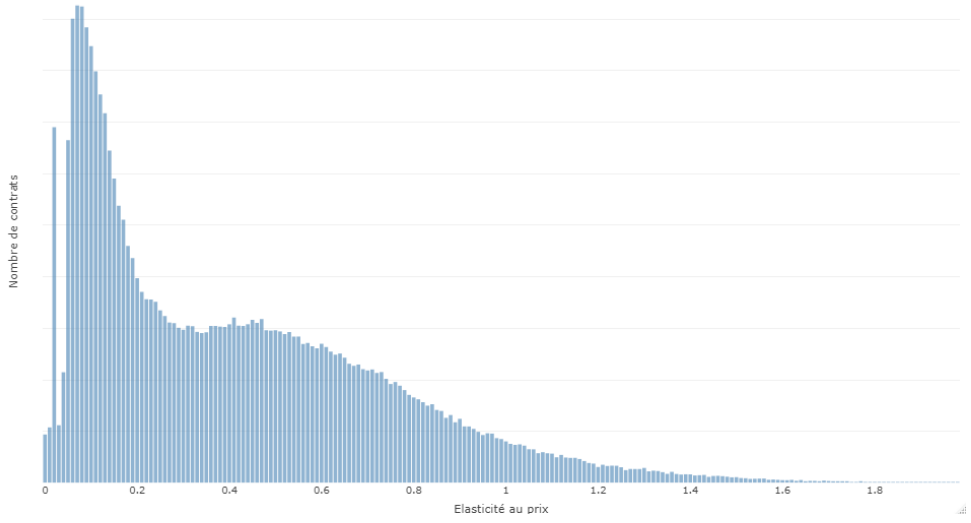


FIGURE 5.11 – Répartition de l'élasticité du taux de résiliation au prix

En appliquant la même méthodologie de classification ascendante hiérarchique combinée à du *bagging*, présentée Section 5.1.1 et détaillée en pratique Section 5.1.2, les assurés sont segmentés en 8 classes, au sein desquelles les comportements de sensibilité à une augmentation du tarif sont homogènes. La répartition des classes est disponible Figure 5.12, la classe 1 est composée des assurés les moins sensibles au prix, quand la classe 8 se constitue des individus dont la résiliation est la plus impactée par un changement de tarif.

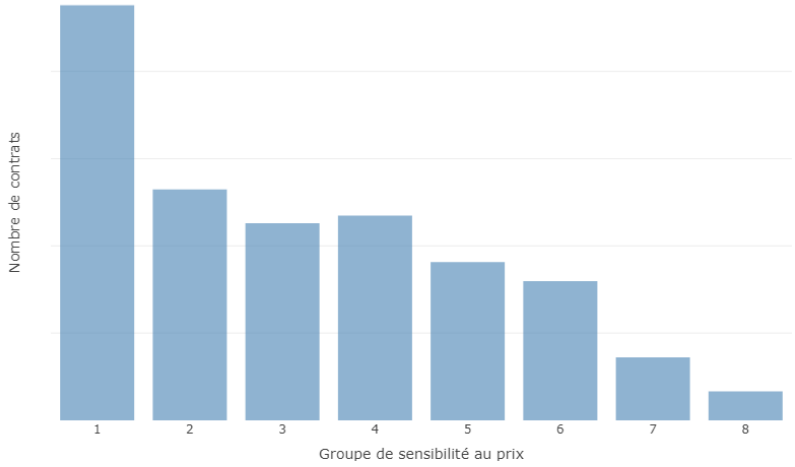
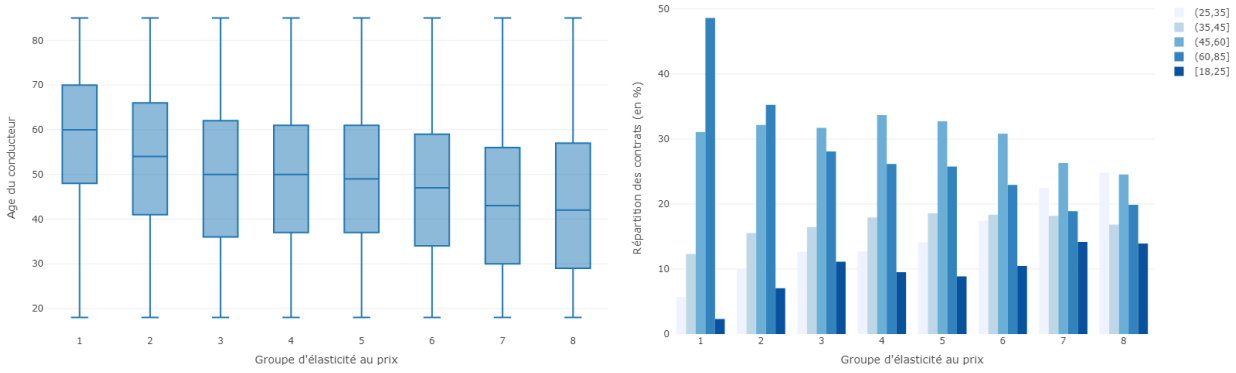


FIGURE 5.12 – Classes d'élasticité au prix

5.2.2 Analyse des classes

Une fois les classes d'élasticité au prix construites, une analyse des caractéristiques des assurés les composant accorde une appréciation des profils plus ou moins sensibles à une augmentation de leur cotisation.

La Figure 5.13 propose l'étude des différentes classes de sensibilité au prix en fonction de l'âge des assurés. Figure 5.13a, les différences d'âges entre les groupes sont peu marquées, bien que dans l'ensemble, l'âge médian des assurés décroît lorsque l'élasticité au prix augmente, pour passer de 60 ans dans le groupe 1 à 42 ans dans le groupe 8. La Figure 5.13b permet d'observer que les différences d'âges inter-classes sont principalement portées par les assurés les plus et les moins âgés. Quand globalement la répartition des âges des assurés de 35 à 60 ans est stable au travers des différentes classes, celles des assurés de moins de 35 ans et de plus de 60 ans varie. Les assurés les plus jeunes sont surreprésentés dans les classes où l'élasticité au prix est la plus élevée, et à l'inverse, les assurés les plus âgés sont en majorité dans les classes les moins sensibles.

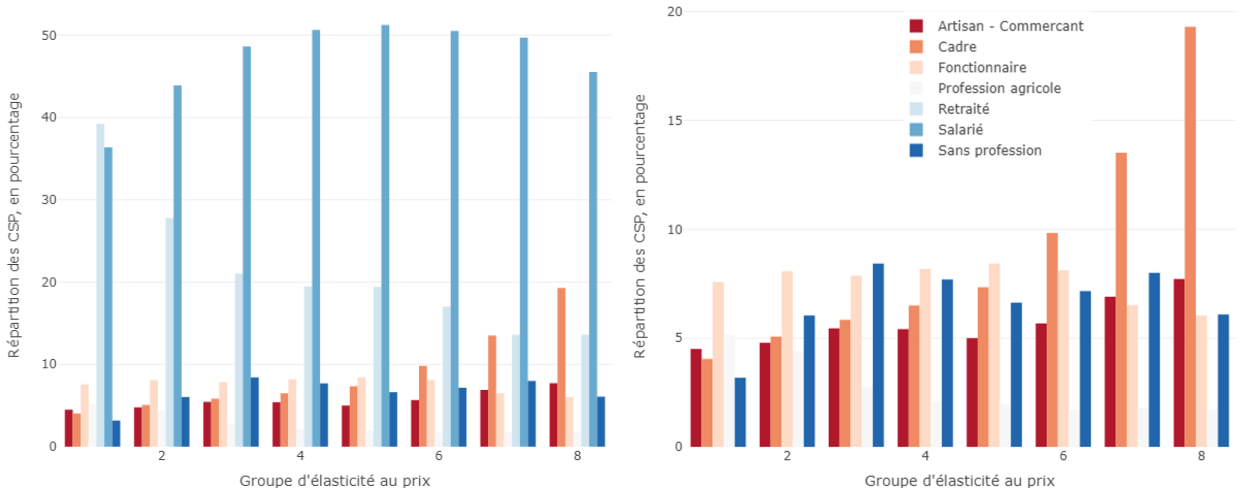


(a) (b)

FIGURE 5.13 – Analyse des classes d'élasticité au prix - Âge du conducteur

De même, les catégories socioprofessionnelles des assurés sont inégalement réparties au sein des différentes classes d'élasticité au prix. Premièrement sur la Figure 5.14a, où l'ensemble des CSP sont représentées, il est possible de notifier la surreprésentation des retraités dans les premières classes, composées des assurés les moins sensibles aux variations de leur tarif. Les salariés sont globalement équirépartis sur l'ensemble des groupes, quoique légèrement en sous nombre au sein des deux premières

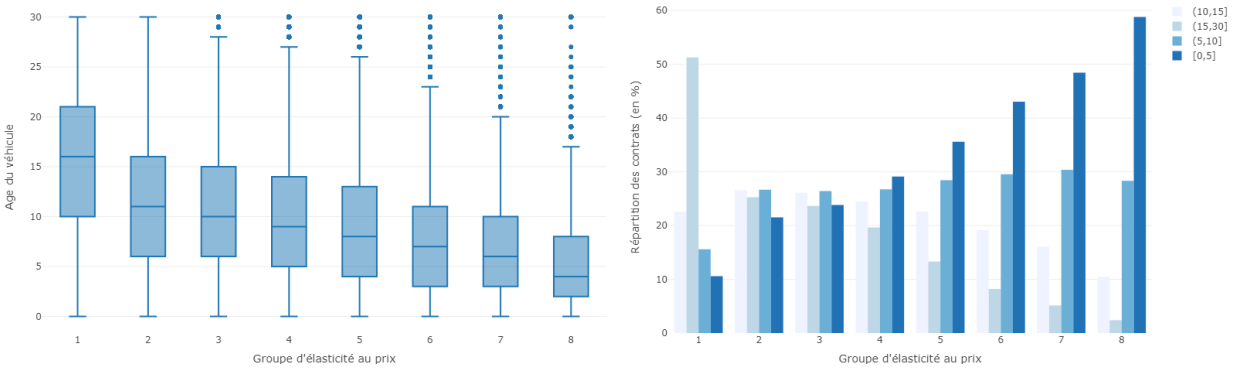
classes. Ensuite, en excluant les salariés et les retraités, très exposés dans le portefeuille, la Figure 5.14b permet l'analyse plus fine des autres assurés. Les cadres, et légèrement les artisans et commerçants, sont plus représentés dans les classes les plus à risque de résiliation lors d'une revalorisation tarifaire. A l'inverse, les professions agricoles sont sous-représentés dans ces classes très sensibles. A l'exception de la première classe, la moins à risque, où les assurés sans profession sont peu présents, il ne semble pas se dégager de comportement spécifique pour les fonctionnaires et les assurés sans profession.



(a) Toutes CSP (b) Artisans, cadres, fonctionnaires et sans prof.

FIGURE 5.14 – Analyse des classes d'élasticité au prix - Catégorie socioprofessionnelle

L'ancienneté du véhicule est également un facteur discriminant dans le cadre de la sensibilité. Figure 5.15a, il est clair que plus un véhicule est récent, et plus son conducteur est sensible au tarif appliqué. Le même constat se fait au travers de la Figure 5.15b. Encore une fois, les variations de répartition sont principalement portées par les extrêmes. Les véhicules les plus récents constituent 10% du premier groupe et leur part va croissante jusqu'à 60% dans le dernier groupe, le plus sensible à une augmentation du prix. Réciproquement, les véhicules de plus de 15 ans sont de loin majoritaires dans la première classe, composée d'assurés moins sensibles aux variations tarifaires.

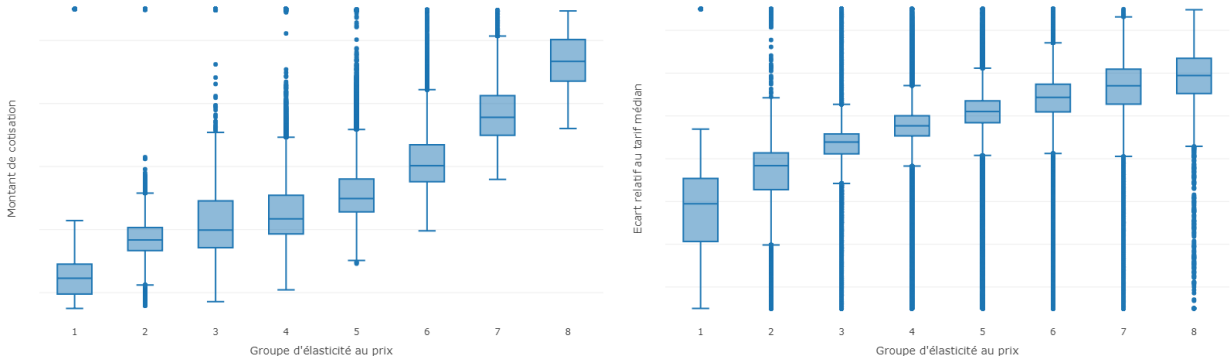


(a) (b)

FIGURE 5.15 – Analyse des classes d'élasticité au prix - Âge du véhicule

Les variables tarifaires, telles que le montant de cotisation et l'écart relatif au tarif médian, varient d'une classe de sensibilité à l'autre comme observé Figure 5.16. Figure 5.16a, il est clair que les assurés les plus susceptibles de résilier suite à une augmentation de tarif sont ceux dont la cotisation est déjà élevée, la cotisation médiane passant du simple au quadruple entre le premier groupe et le dernier. Cet

effet peut, en partie, être porté par le fait que les jeunes conducteurs, dont les cotisations sont élevées, sont à la recherche de tarifs préférentiels et donc plus sensibles aux prix. L'écart relatif de la cotisation au tarif médian progresse également d'autant plus que les assurés sont sensibles au prix. En effet, plus un assuré dispose d'un contrat dont le tarif est bien supérieur à celui du marché, et plus ce dernier sera susceptible de se tourner vers un assureur plus compétitif lors d'une revalorisation tarifaire.



(a) Montant de cotisation (b) Écart relatif au tarif médian

FIGURE 5.16 – Analyse des classes d'élasticité au prix - Variables tarifaires

Au niveau du degré de couverture choisi par l'assuré, la Figure 5.17 permet de conclure que les assurés couverts au tiers, c'est-à-dire présentant un contrat à trois garanties ou moins, sont moins sensibles aux variations de prix que les assurés dont la couverture est complète. Précisément, ils sont représentés significativement dans les trois à quatre premières classes uniquement. Les assurés au nombre de garanties élevé, et principalement ceux disposant de la couverture maximale proposée, représentent plus de 80% des assurés des classes 5 à 8. Ces individus dont le montant de cotisation est élevé, de part la complétude de leur contrat, sont concernés par une augmentation du tarif qui pourrait leur être appliquée et, le cas échéant, tendraient à se tourner vers un assureur dont la prime proposée aux nouveaux prospects est attractive.

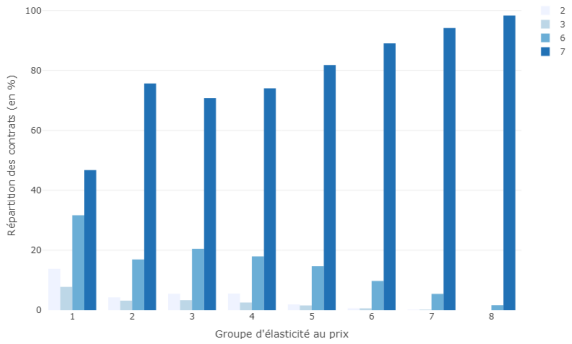
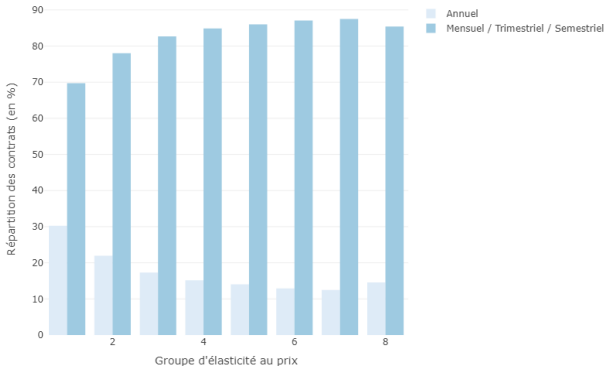
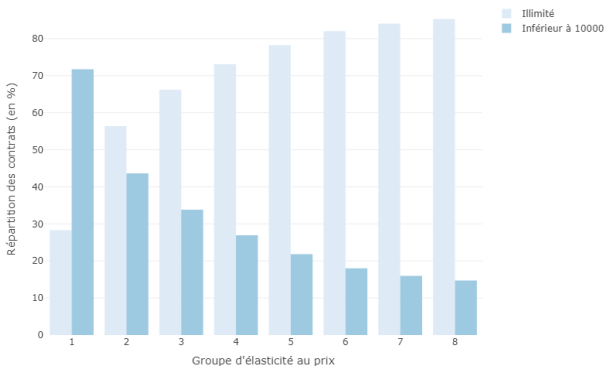


FIGURE 5.17 – Analyse des classes d'élasticité au prix - Type de couverture

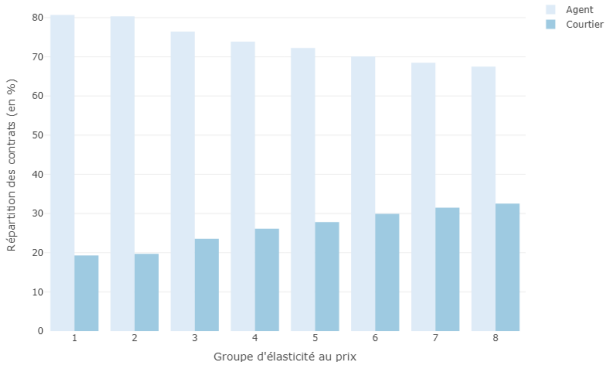
Finalement, les caractéristiques inhérentes au contrat de l'assuré peuvent être étudiées. Figure 5.18a, il apparaît que les individus dont le paiement est annuel sont légèrement moins sensibles aux variations de tarif. Figure 5.18b se manifeste clairement que les assurés au kilométrage limité, dont la cotisation est de fait moindre à celle des grands rouleurs, sont moins sensibles à une augmentation de leur prime. Enfin, Figure 5.18c, les assurés dont le contrat a été souscrit au travers d'un courtier s'avèrent plus représentés dans les classes à forte élasticité.



(a) Fractionnement des paiements



(b) Kilométrage



(c) Type de réseau

FIGURE 5.18 – Analyse des classes d'élasticité au prix

5.3 Cartographie des assurés

A l'issue des classes créées, dans le cadre de la durée de vie et de la sensibilité au prix, une cartographie des assurés peut être réalisée. Cette dernière se construit au travers de la marge, premier indicateur de la rentabilité d'un assuré.

Une première analyse peut être réalisée par l'étude de la Figure 5.19, cartographiant les assurés relativement à leur durée de vie a priori, en abscisses, et à leur marge, en ordonnées. Un assuré est d'autant plus fidèle qu'il se trouve à droite de la figure, et d'autant plus rentable qu'il se situe en hauteur. Un exposé des quatre segments entourés peut être conduit :

- Groupe A : les assurés de ce groupe sont a priori très fidèles et sont, de plus, rentables pour l'assureur. Selon le niveau de rétention, la sensibilité au prix, et le positionnement marché de l'assureur sur ces segments, certains de ces profils pourraient bénéficier d'une réduction commerciale visant à récompenser et asseoir leur fidélité.
- Groupe B : ces assurés, peu enclins toutes choses égales par ailleurs, à rester longtemps au sein du portefeuille, présentent un niveau de marge intéressant pour l'assureur. Bien que classés comme peu fidèles, une analyse plus fine de ces segments, au travers de la sensibilité au prix, permettrait de jauger de la pertinence, en termes de rétention, d'une stratégie visant à diminuer le tarif appliqué à ces assurés.
- Groupe C : les individus composant ce groupe sont fidèles, néanmoins, l'assureur est en déficit sur ces clients dont la marge est négative. Hormis pour les contrats très récents, dont la stratégie de fidélisation à long terme par une légère perte à court terme est pertinente, l'assureur pourrait gagner à augmenter tout ou partie de ces profils.
- Groupe D : sur ces profils très peu fidèles a priori, l'assureur bénéficierait à augmenter son tarif, les individus étant disposés, de par leurs caractéristiques intrinsèques à quitter rapidement le portefeuille.

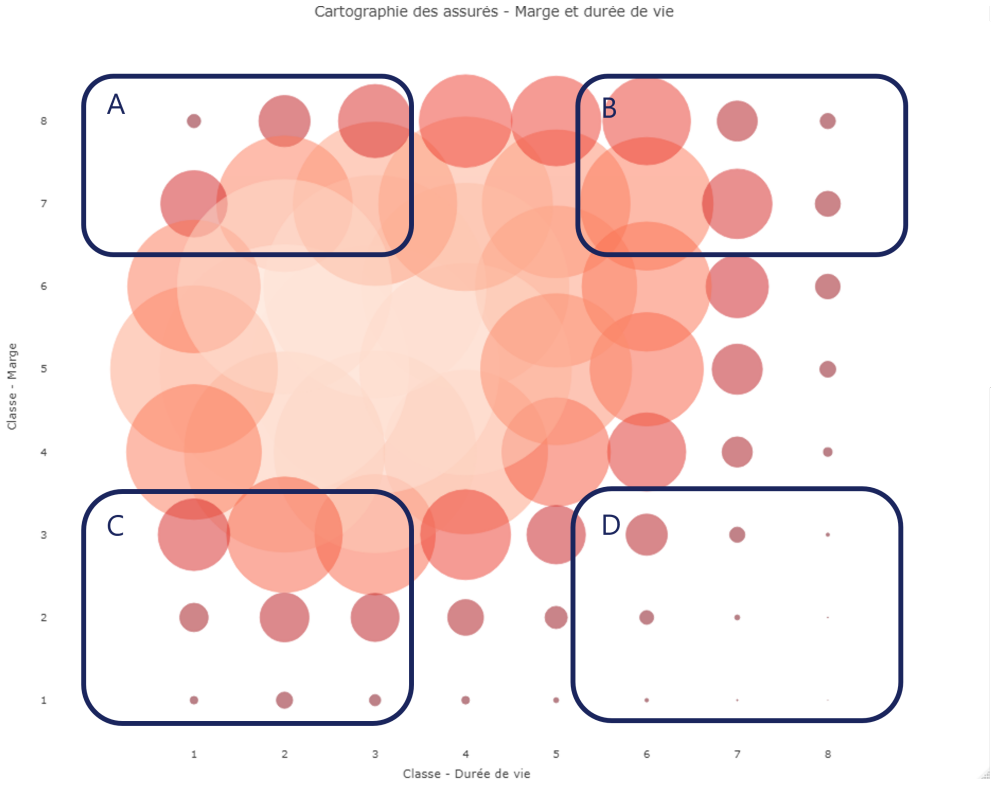


FIGURE 5.19 – Cartographie - Marge et groupe de durée de vie a priori

De la même manière, le portefeuille est cartographié Figure 5.20 à partir du niveau de marge et de la sensibilité au prix des assurés. Des analyses similaires à celles précédemment amenées peuvent être proposées. Par exemple, les assurés du groupe C présentent une marge négative, mais leur classe de sensibilité au prix indiquent que leur niveau de rétention ne serait que légèrement diminué dans la cadre d'une augmentation de tarif. Ainsi, après examen de son positionnement sur le marché, l'assureur pourrait considérer une légère augmentation de la cotisation de ces profils. A l'inverse, le groupe B est composé d'individus très sensibles aux variations tarifaires et dont la marge est élevée. De par leur niveau de sensibilité, une stratégie visant à diminuer légèrement leur cotisation pourrait engendrer une meilleure rétention, et donc rentabilité, sur le moyen terme.

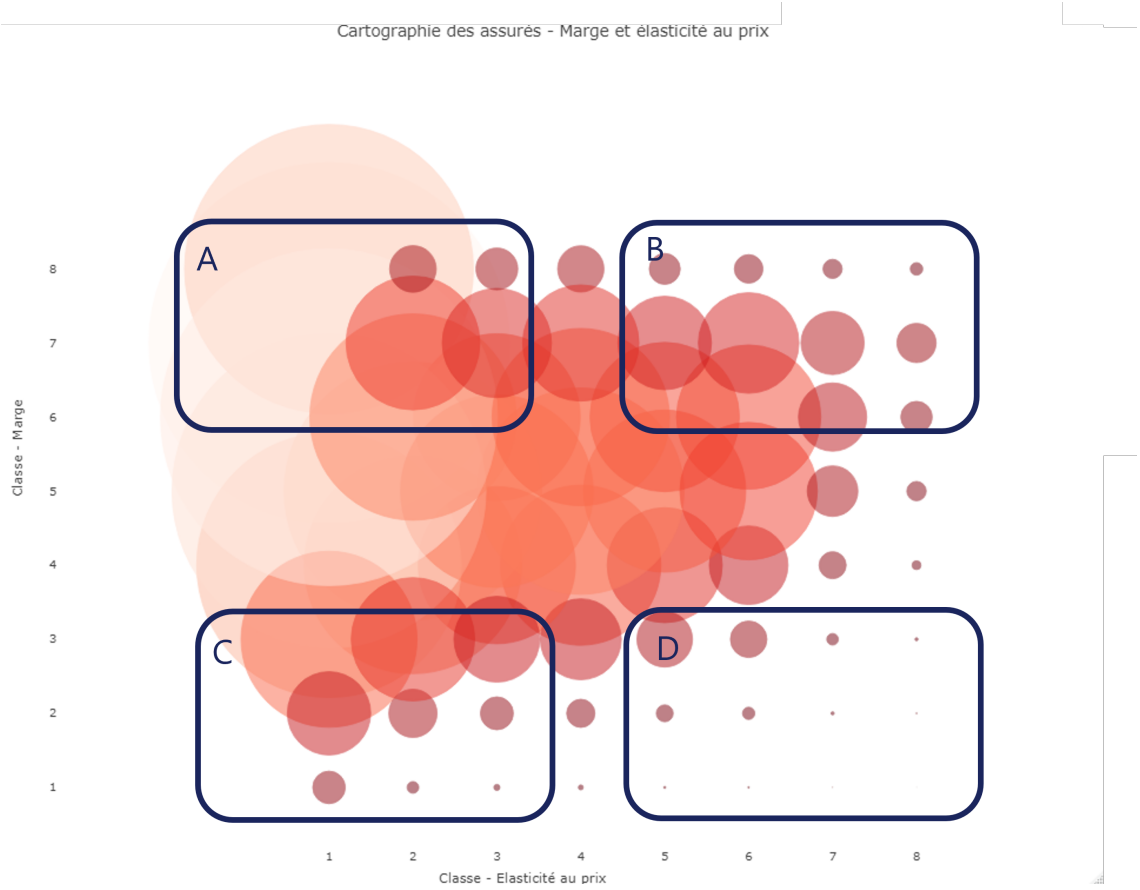


FIGURE 5.20 – Cartographie - Marge et groupe de sensibilité du taux de résiliation au prix

5.4 Amélioration des indicateurs clefs de la rentabilité de l'assureur : prémices d'une optimisation tarifaire

Les travaux réalisés en amont permettent la mise en œuvre d'un travail d'identification et de revalorisation spécifique de certains profils, dans une démarche d'optimisation de la stratégie tarifaire de l'assureur. Les principaux indicateurs de rentabilité sont définis puis la mise en œuvre de scénarios sur divers segments est évaluée.

5.4.1 Indicateurs clefs de rentabilité

Lors de la mise en place d'une optimisation tarifaire, l'assureur s'appuie sur différents indicateurs de santé financière. Entre autres, il cherche à améliorer son profit, son chiffre d'affaires ou encore son ratio de perte, tout en maintenant son volume d'assurés. Premièrement, la probabilité de résiliation de l'assuré i, dépendant du montant de cotisation $P_{i,HT}$ appliqué, est notée $\hat{p}_i$:

$$\hat{p}_i(P_{i,HT}) = \hat{f}_i(P_{i,HT}, P_{i,\text{pure}}, X_i)$$

Ensuite, l'ensemble des indicateurs mentionnés peuvent être définis, à un an, à partir de $\hat{p}_i$:

$$\begin{aligned}
\text{Volume du portefeuille}(P_{HT}) &= \sum_{i=1}^{n}(1-\hat{p}_i(P_{i,HT})) \\
\text{Chiffre d'affaires}(P_{HT}) &= \sum_{i=1}^{n} P_{i,HT}\cdot(1-\hat{p}_i(P_{i,HT})) \\
\text{Profit}(P_{HT}) &= \sum_{i=1}^{n}(P_{i,HT}-S_i)\cdot(1-\hat{p}_i(P_{i,HT})) \\
\text{Ratio de perte}(P_{HT}) &= \frac{\sum_{i=1}^{n} S_i\cdot(1-\hat{p}_i(P_{i,HT}))}{\sum_{i=1}^{n} P_{i,HT}\cdot(1-\hat{p}_i(P_{i,HT}))}
\end{aligned} \tag{5.4}$$

avec :

- n : nombre d'assurés en portefeuille ;
- P_{HT} : vecteur des $P_{i,HT}$, $i = 1, ..., n$;
- $P_{i,HT}$: tarif commercial hors taxes appliqué à la police d'assurance du client i ;
- $P_{i,\text{pure}}$: prime pure de l'assuré i ;
- S_i : montant des sinistres, additionné des frais, de l'assuré i, obtenu à partir de la prime pure. Le calcul des indicateurs se fait à un an, ainsi ce montant est déterministe ;
- X_i : caractéristiques de l'assuré i.

5.4.2 Mise en place de scénarios tarifaires sur des segments spécifiques

Les travaux d'analyses exploratoires, de modélisation de la résiliation, de la durée de vie et de la sensibilité au prix ont permis de dégager un certain nombre de profils aux comportements spécifiques. Certains de ces segments, se démarquant par leurs conduites de fidélité, leurs sensibilités au prix et par les positionnements tarifaires de l'assureur, méritent d'être étudiés. Les profils des jeunes conducteurs, des cadres, et des retraités, sont étudiés dans cette partie. Des scénarios de revalorisation tarifaire sont proposés dans une démarche d'accroissement du profit de l'assureur, tout en maintenant à leurs niveaux initiaux le volume, le chiffre d'affaires et le ratio de perte.

Jeunes conducteurs

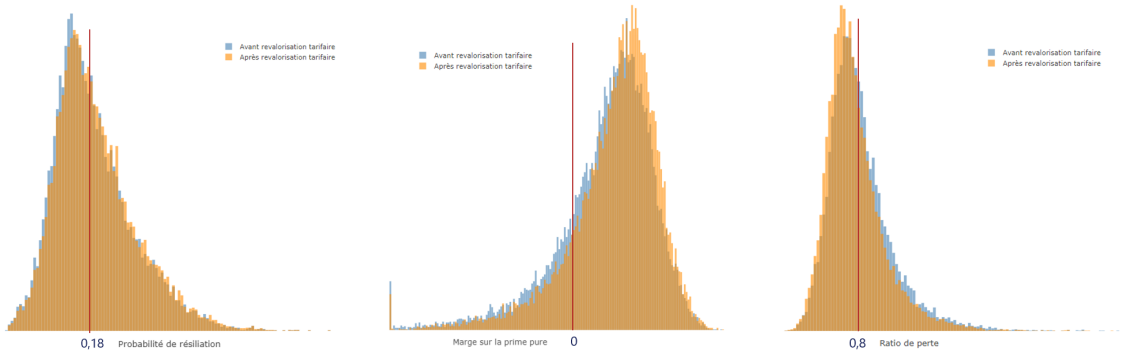
Le segment des jeunes conducteurs, se caractérisant par un niveau de risque élevé, présente un enjeu tout particulier pour l'assureur. Comme appuyé par les résultats des diverses analyses menées, les jeunes assurés se distinguent notamment par une sensibilité au prix accrue. Leur niveau de sensibilité au prix, s'observant également dans le cadre de la souscription, en fait des prospects difficiles à capter. Une stratégie pour l'assureur vise à proposer et à maintenir des tarifs bas, en dessous du tarif médian du marché, pour attirer et fidéliser ce segment de conducteurs. Néanmoins, les conclusions des modèles de résiliation et de durée sont sans équivoque : les jeunes assurés, toutes choses égales par ailleurs, présentent des taux de résiliation très élevés et constituent le segment le moins fidèle. Ainsi, augmenter légèrement le tarif de ces assurés permet d'améliorer les indicateurs de profit à court terme, sans prendre de risques inconsidérés sur le moyen terme.

Au sein du portefeuille étudié, les jeunes de moins de 23 ans constituent environ 5% des contrats. Pour cette sous population, la marge et l'écart relatif au tarif médian sont négatifs deux fois plus fréquemment que sur l'ensemble du portefeuille. Le ratio de perte y est également plus élevé. Ainsi, en moyenne, le bénéfice réalisé sur ce segment est moindre relativement au reste des contrats. Au regard du déficit porté par l'assureur sur les jeunes, qui pour autant n'observe pas d'amélioration de la rétention sur le long terme, une stratégie de revalorisation de ces contrats est proposée. Pour les conducteurs de moins de 23 ans, dont l'écart au tarif médian est négatif, les majorations suivantes sont appliquées :

— augmentation de 5% les assurés présentant un deuxième contrat auto ;
— augmentation de 10% les assurés ne présentant pas de deuxième contrat auto.

La distinction faite sur l'indicatrice de présence de deuxième contrat auto permet de se prémunir en partie du risque de multiples résiliations. En effet, l'assuré détenant le deuxième contrat auto, souvent un parent dans le cadre des jeunes conducteurs, pourrait être poussé à résilier l'ensemble de ses contrats suite à une hausse qu'il jugerait trop importante du montant de cotisation de l'un d'entre eux.

La Figure 5.21 permet une appréciation quantitative des effets du scénario tarifaire ainsi mis en place. Concernant la probabilité de résiliation à un an des assurés de moins de 23 ans, la Figure 5.21a indique un léger décalage vers la droite de la répartition. Naturellement, l'augmentation d'une partie des contrats conduit à une hausse du risque de résiliation. Précisément, une hausse de 2,6% du taux de résiliation moyen est constatée. Ensuite, la Figure 5.21b propose de comparer la répartition de la marge avant et après la mise en place de la revalorisation tarifaire. Celle-ci augmente de cinq points de pourcentage, ce qui permet de combler en partie les écarts de marge constatés entre les jeunes conducteurs et le reste du portefeuille. Finalement, le ratio de perte est également étudié Figure 5.21c, ce dernier décroît de manière significative, indiquant une meilleure rentabilité une fois le scénario établi.



(a) Probabilité de résiliation (b) Marge (c) Ratio de perte

FIGURE 5.21 – Étude du scénario tarifaire d'augmentation des jeunes assurés

De plus, malgré l'augmentation de certains contrats détenus par les jeunes conducteurs, l'assureur reste compétitif sur ce segment en présentant un écart relatif moyen au tarif médian toujours négatif.

Retraités

Les retraités se démarquent par une durée de vie a priori élevée, des taux de résiliation en deçà de la moyenne du portefeuille et une faible élasticité au prix. Cependant, l'assureur se voit en déficit, tant en termes de marge que d'écart au tarif médian, sur certains de ces profils. Une étude approfondie des assurés retraités permet de déterminer que ce comportement tarifaire concerne majoritairement les retraités au véhicule récent. Une stratégie visant à augmenter légèrement ce segment peut alors être menée. Néanmoins, plusieurs précautions sont prises : les assurés récemment acquis en portefeuille et multi équipés, plus sensibles, se voient attribuer une majoration moindre.

La stratégie suivante est adoptée sur les retraités au véhicule de moins de 10 ans :

1. Contrat présent depuis plus de 4 ans :
   — multi-équipé : augmentation de 8% ;
   — mono-équipé : augmentation de 10%.
2. Contrat présent depuis moins de 4 ans :
   — multi-équipé : augmentation de 4% ;
   — mono-équipé : augmentation de 6%.

La mise en place de ce scénario tarifaire induit, sur le segment des retraités :
— une augmentation de la probabilité de résiliation moyenne de moins de 1% ;
— une augmentation de 10% du profit ;
— une diminution du ratio de perte de 3%.
Ainsi, malgré une légère perte du volume d'assurés, le scénario mis en place est concluant en termes de rentabilité pour l'assureur.

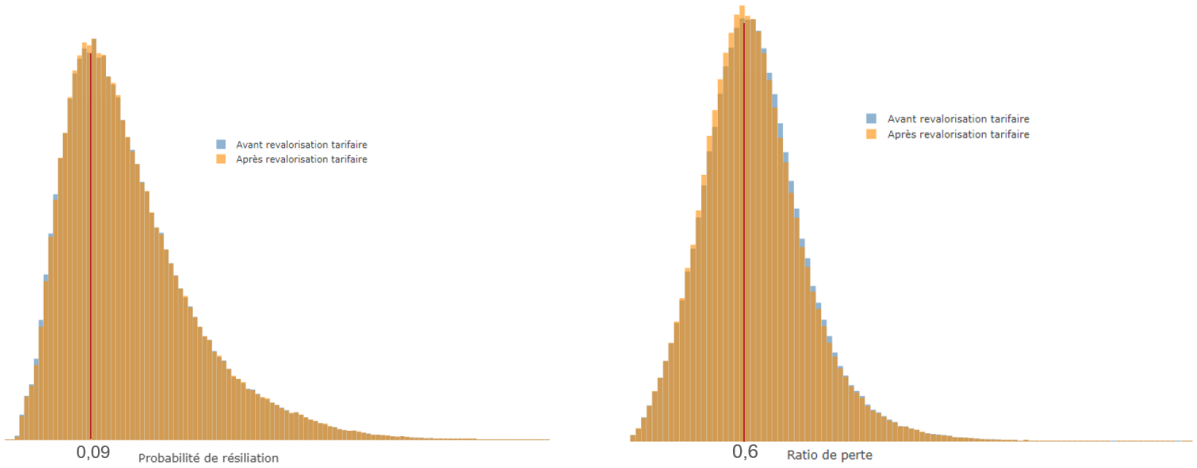
Cadres

La population des cadres est caractérisée par une rétention et une courbe de survie légèrement supérieures à la moyenne. Cependant, ils se démarquent par leur sensibilité au prix élevée. Le niveau de marge réalisé par l'assureur sur ce segment étant correct, une stratégie de minoration de ses cotisations est proposée. Cette dernière vise à accroître la rétention à court et moyen termes des cadres, en s'appuyant sur leur niveau de sensibilité face aux variations tarifaires. Sur les cadres, représentant près de 7% du portefeuille en cours, un scénario de diminution est mis en place. Les cadres dont le montant de cotisation est supérieur au tarif médian du marché et n'étant pas sinistrés se voient bénéficier d'une réduction de 10%. Les résultats de cette stratégie sont les suivants :
— augmentation de 5% de la rétention ;
— diminution de 30% du profit ;
— augmentation de 6% du ratio de perte.
Bien que l'assureur se défasse d'une partie du profit réalisé à court terme sur ce segment, l'augmentation de la rétention permet de garantir une compensation sur le plus long terme.

Résultats du scénario tarifaire

Les trois scénarios présentés en amont, sur les jeunes conducteurs, les retraités et les cadres, permettent la mise en place des prémices d'une optimisation tarifaire sur l'ensemble du portefeuille en cours. Les indicateurs de rentabilité de l'assureur évoluent de la manière suivante :
— augmentation de 3,5% du profit ;
— augmentation de 0,8% du chiffre d'affaires ;
— diminution de moins de 0,03% du volume du portefeuille ;
— diminution de 1% du ratio de perte.
Le scénario mis en place répond aux critères de l'optimisation tarifaire fixés en amont : le profit a augmenté de manière significative, les contraintes du programme liées au chiffre d'affaires et au ratio de perte sont remplies et le volume du portefeuille a diminué certes, mais de façon marginale seulement. La répartition des indicateurs et leurs évolutions respectives suite à l'optimisation tarifaire

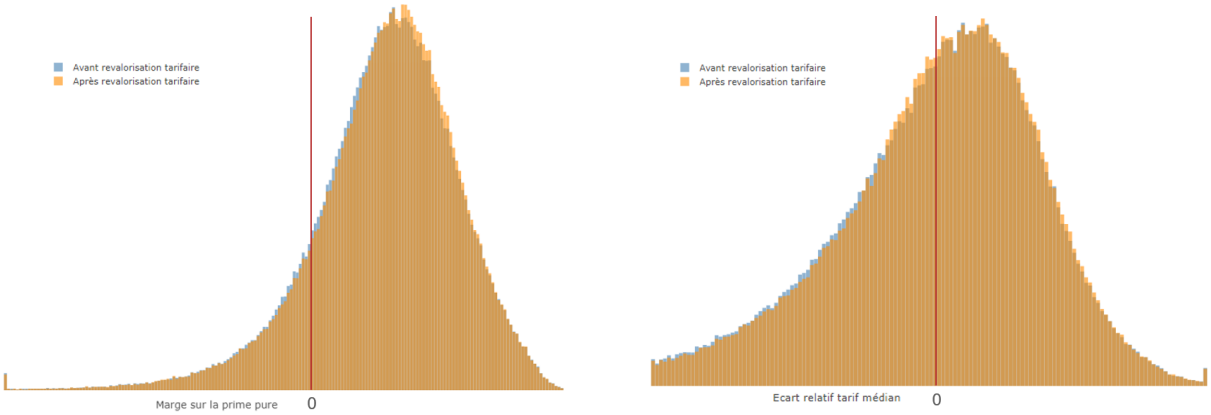
sont étudiées. Figure 5.22 sont représentées les répartitions de la probabilité de résiliation et du ratio de perte, sur l'ensemble du portefeuille, avant et après l'exécution de la revalorisation tarifaire. Figure 5.22, il est possible d'observer qu'une partie des individus qui présentaient une probabilité de résilier très faible se voient légèrement décalés vers des probabilités plus médianes, comprises entre 6 et 12%. Néanmoins, la revalorisation tarifaire menée n'induit pas une augmentation du nombre d'individus dans des zones très à risque, au-delà de la résiliation moyenne. Le ratio de perte, Figure 5.22b, observe un décalage vers la gauche, signe d'une meilleure santé financière.



(a) Probabilité de résiliation (b) Ratio de perte

FIGURE 5.22 – Étude du scénario tarifaire - Ensemble du portefeuille

Ensuite, la marge et l'écart relatif au tarif médian, indicateurs du positionnement tarifaire de l'assureur, peuvent être observés Figure 5.23. Les deux répartitions se décalent légèrement vers la droite, sans présenter de comportement extrême de sur tarification. L'assureur augmente son profit tout en gardant un positionnement sur le marché cohérent et compétitif.



(a) Marge (b) Écart relatif au tarif médian

FIGURE 5.23 – Étude du scénario tarifaire - Ensemble du portefeuille

5.5 Perspectives : optimisation tarifaire et valeur client

Les travaux réalisés au cours de ce mémoire proposent une base à l'optimisation tarifaire d'une part, et au calcul de la valeur client d'autre part. Ces deux aspects sont brièvement présentés dans cette dernière section.

5.5.1 Optimisation tarifaire

A partir des métriques de rentabilité présentées Équation 5.4, l'assureur dispose comme seul levier P_{HT}, le montant de cotisation qu'il propose à ses clients. Hormis cette variable, les autres quantités telles que la prime pure $P_{i,\text{pure}}$, le montant des sinistres agrémenté des frais S_i, ainsi que le vecteur de caractéristiques X_i sont fixes et inhérentes à l'assuré et son contrat. Sur une vision un an, une stratégie peut être de maximiser le profit tout en maintenant les autres quantités à leurs niveaux initiaux. Soit P^0_{HT}, le vecteur des tarifs appliqués en amont de l'optimisation tarifaire. Le programme d'optimisation est le suivant :

$$\begin{aligned} \max_{P_{i,HT},\quad i=1,\dots,n} \quad & \left\{ \text{Profit}(P_{HT}) = \sum_{i=1}^{n} (P_{i,HT} - S_i) \cdot (1 - \hat{p}_i(P_{i,HT})) \right\} \\ \text{s.c.} \quad & P_{i,HT} \in [P^0_{i,HT} \cdot (1 - 15\%), P^0_{i,HT} \cdot (1 + 15\%)] \quad \forall i = 1, \dots, n \\ & \text{Volume du portefeuille}(P_{HT}) \geq \text{Volume du portefeuille}(P^0_{HT}) \\ & \text{Chiffre d'affaires}(P_{HT}) \geq \text{Chiffre d'affaires}(P^0_{HT}) \\ & \text{Ratio de perte}(P_{HT}) \leq \text{Ratio de perte}(P^0_{HT}) \end{aligned} \tag{5.5}$$

avec, pour rappel :

- $\hat{p}_i(P_{i,HT}) = \hat{f}_i(P_{i,HT}, P_{i,\text{pure}}, X_i)$ la probabilité de résiliation de l'assuré i ;
- n : nombre d'assurés en portefeuille ;
- P_{HT} : vecteur des $P_{i,HT}$, $i = 1, ..., n$;
- $P_{i,HT}$: tarif commercial hors taxes appliqué à la police d'assurance du client i ;
- $P_{i,\text{pure}}$: prime pure de l'assuré i ;
- S_i : montant des sinistres, additionné des frais, de l'assuré i, obtenu à partir de la prime pure. Le calcul se fait à un an, ainsi ce montant est déterministe ;
- X_i : caractéristiques de l'assuré i.

Le programme d'optimisation tarifaire proposé permet une augmentation du profit tout en maintenant les autres indicateurs de rentabilité. De plus, la revalorisation des montants de cotisations ne peut excéder un certain seuil. Au-delà des indicateurs de rentabilité à un an, une optimisation tarifaire lors du renouvellement anniversaire des contrats doit également considérer une vision du portefeuille à plus long terme. C'est dans ce contexte que la durée de vie a priori des contrats intervient. L'assureur se doit d'être soucieux des segments les plus ou moins fidèles sur le moyen terme et d'ajuster sa stratégie en conséquence.

Selon le positionnement de la compagnie d'assurance, la fonction objectif et les contraintes du programme 5.5 seront différentes. Par exemple, une mutuelle ne cherchera pas à maximiser son profit mais à équilibrer de manière plus juste sa politique tarifaire.

La résolution d'une telle équation passe par une phase algorithmique complexe, faisant intervenir, entre autres, la méthode de Lagrange et le gradient projeté.

5.5.2 Valeur client

La valeur client se définit comme la valeur nette actualisée des profits passés et futurs. Elle correspond à la somme des profits générés par l'assuré au cours de sa relation avec l'assureur et se décompose en une valeur client passée, et une valeur client future. La valeur client future est particulièrement intéressante dans une démarche d'appréhension des comportements de fidélité et de rentabilité des assurés déjà en portefeuille. Cette dernière permet de capter, par un seul indicateur, la marge et la rétention potentielles d'un assuré sur un horizon T. Formellement, elle peut se définir comme la valeur actualisée de la marge, multipliée par un taux de chute :

$$\text{Valeur client future de l'assuré } i = \frac{\sum_{t=1}^{T} M_i(t) \cdot (1 - \hat{f}_i(t))}{1 + r(t)}$$

Puis, la valeur client de l'ensemble du portefeuille est obtenue :

$$\text{Valeur client future} = \sum_{i=1}^{n} \text{Valeur client future de l'assuré } i$$

avec :

— $M_i(t)$: marge de l'assuré i au temps t ;
— $\hat{f}_i(t)$: probabilité de résiliation de l'assuré i au cours de l'année t.
— $r(t)$: taux d'intérêt au temps t.

La probabilité $\hat{f}_i(t)$ de résiliation l'année t est obtenue au travers des modèles de durée de vie. Cependant, la quantité $M_i(t)$ demande la mise en place de modèles de projection de marge. Différentes approches peuvent être considérées, telles que la modélisation de la marge par des modèles actuariels ou économétriques [18]. Également, les évolutions de marge pourraient être captées par des modèles *black box* puissants. Des algorithmes de boosting ou des réseaux de neurones seraient en mesure d'estimer la pente des revalorisations annuelles et de la marge en fonction de différents paramètres.

De plus, la formule de la valeur client peut, à terme, être complexifiée afin de mesurer avec plus de précision le profit généré par les assurés au cours de leur relation avec l'assureur. L'intégration des coûts d'acquisitions, pour les affaires nouvelles, à amortir dans les premières années du contrat, est à considérer. Également, la notion de ventes croisées peut intervenir. Prendre en compte cet élément revient à passer de la valeur contrat à la valeur client, en examinant les revenus générés par un client au travers de ses différents contrats d'assurance. Cela demande notamment l'estimation des probabilités de souscription à de nouveaux contrats.

Conclusion

Dans un contexte de concurrence accrue, l'appréhension des comportements des assurés face à la résiliation est un enjeu majeur pour garantir la pérennité économique des organismes d'assurance. Cela se révèle d'autant plus incontestable face au fait que le turn-over d'un portefeuille représente un poids financier conséquent pour l'assureur. Ainsi, ce dernier bénéficie de la fidélisation des clients déjà acquis. L'objet de ce mémoire a été de proposer différents outils permettant de mieux capter les comportements des assurés et les facteurs conduisant à une rétention, ou au contraire à une résiliation.

La modélisation de la probabilité de résiliation repose sur une analyse fine des décisions prises par l'assuré dans un laps de temps d'un an. Des éléments, intrinsèques à l'assuré, son véhicule, sa couverture ou ses avenants ont pu être mis en évidence comme étant explicatifs d'une fidélisation plus ou moins prononcée de certains segments de clients. Dans une volonté de disposer d'un indicateur prenant plus de recul temporel, la durée de vie a priori des contrats d'assurance automobile a également été modélisée. Les deux modèles ainsi construits se complètent et proposent deux éclairages sur le sujet de la fidélisation des assurés. L'étude de la résiliation à un an capte l'imminence du risque porté par l'assuré et son comportement, quand l'analyse de la durée de vie du contrat a priori permet d'identifier les clients plus ou moins fidèles avec plus de recul.

A partir des modèles construits, un ensemble d'outils d'aide à la décision est proposé. Il permet à l'assureur de maîtriser son portefeuille, et de mener à bien ses politiques tarifaires et marketing. La cartographie du portefeuille en fonction de la rentabilité, de la durée de vie a priori, et de la sensibilité des assurés met en évidence certains profils de clients spécifiques, sur lesquels l'assureur peut entamer des actions. En outre, les modèles de résiliation et de durée de vie permettent l'évaluation de différentes stratégies sur la fidélisation des clients et donc, sur la rentabilité. Un scénario tarifaire permettant d'augmenter le profit de l'assureur tout en conservant son niveau d'assurés à court et moyen termes est proposé. Quand des indicateurs tels que le chiffre d'affaires ou le profit à un an peuvent être obtenus à partir des travaux réalisés, d'autres métriques, qui nécessitent la projection de la prime pure sur plusieurs années n'ont pas fait l'objet de l'étude. Par les deux modèles obtenus, ce mémoire propose une base pour le calcul de tels indicateurs, notamment celui de la valeur client.

Afin de disposer d'une part, de résultats interprétables, et d'autre part, d'équations dérivables en vue d'une optimisation tarifaire, les modèles employés présentent un niveau de sophistication limité. Bien qu'un modèle de machine learning puissant, le XGBoost, ait été réalisé dans le cadre de la modélisation de la résiliation, ce dernier est resté simple et a été construit à titre de comparaison de performance pour la régression logistique. Des modèles complexes et puissants pourraient être envisagés pour déceler un niveau d'interaction entre les variables, ou pour dégager des segments plus fins, qui échappent en partie aux régressions mises en place.

Bibliographie

Articles

[1] W. E. Barlow et R. L. Prentice. « Residuals for Relative Risk Regression ». In : *Biometrika* 75.1 (1988), p. 65-74.

[3] N. E. Breslow. « Analysis of Survival Data under the Proportional Hazards Model ». In : *International Statistical Review* 43.1 (1975), p. 45-57.

[6] T. Chen et C. Guestrin. « XGBoost : A Scalable Tree Boosting System ». In : *KDD '16 : Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), p. 785-794.

[7] D. R. Cox. « Regression Models and Life-Tables ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), p. 187-220.

[8] D. R. Cox. « Partial likelihood ». In : *Biometrika* 62.2 (1975), p. 269-276.

[12] Gibbons et Prat. « P-values : Interpretation and Methodology ». In : *The American Statistician, February 1975, Vol. 29, No. 1* (1975), p. 20-25.

[13] R. Gill. « Large Sample Behaviour of the Product-Limit Estimator on the Whole Line ». In : *The Annals of Statistics* 11.1 (1983), p. 49-58.

[14] P. M. Grambsch et T. M. Therneau. « Proportional Hazards Tests and Diagnostics Based on Weighted Residuals ». In : *Biometrika* 81.3 (1994), p. 515-526.

[15] P. J. Green. « Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternative ». In : *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 46, No. 2* (1984), p. 149-192.

[16] Harrell et al. « Evaluating the Yield of Medical Tests ». In : *JAMA* 247.18 (1982), p. 2543-2546.

[17] H. Heinzl et A. Kaider. « Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions ». In : *Computer Methods and Programs in Biomedicine* 54.3 (1997), p. 201-208.

[19] E. L. Kaplan et P. Meier. « Nonparametric Estimation from Incomplete Observations ». In : *Journal of the American Statistical Association, Vol. 53, No. 282* (1958), p. 457-481.

[26] Laurenz. « Methods of measuring the concentration of wealth ». In : *American Statistical Association, Vol. 9* (1905), p. 209-219.

[27] S. M. Lundberg et S.-I. Lee. « A Unified Approach to Interpreting Model Predictions ». In : *31st Conference on Neural Information Processing Systems (NIPS 2017)* (2017).

[28] J. A. Nelder et R. W. M. Wedderburn. « Generalized Linear Models ». In : *Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3* (1972), p. 370-384.

[29] F.F. Reichheld et P.M. Dawkins. « Customer retention as a competitive weapon ». In : *Directors Broads, 14* (1990), p. 42-47.

[30] D. Schoenfeld. « Partial Residuals for The Proportional Hazards Regression Model ». In : *Biometrika* 69.1 (1982), p. 239-241.

[31] W. Stute. « Consistent Estimation Under Random Censorship When Covariables Are Present ». In : *Journal of Multivariate Analysis* 45.1 (1993), p. 89-103.

[33] T. M. Therneau, P. M. Grambsch et Thomas R. Fleming. « Martingale-based residuals for survival models ». In : *Biometrika* 77.1 (1990), p. 147-160.

Articles en ligne

[4] A. Charpentier. *La loi des grands nombres et le théorème central limite comme base de l'assurabilité ?* 2011.

[5] A. Charpentier, M. Denuit et R. Elie. *Segmentation et mutualisation, les deux faces d'une même pièce ?* 2015.

[23] Fédération française de l'assurance. *Le marché de l'assurance automobile des particuliers en 2019*. 2019.

[24] Fédération française de l'assurance. *Les données clés de l'assurance française en 2021*. 2021.

Livres

[2] M. W. Berry, A. Mohamed et B. W. Yap. *Supervised and Unsupervised Learning for Data Science*. 2019.

[9] J.J. Droesbeke, B. Fichet et P. Tassi. *Analyse statistique des durées de vie*. 1989.

[11] R. A. Fisher. *Statistical Methods for Research Workers*. Springer, 1925.

[20] J. P. Klein et M. L. Moeschberger. *Survival Analysis Techniques for Censored and Truncated Data*. Springer, 2ème édition, 2003, p. 63-74.

[32] T. M. Therneau et P. M. Grambsch. *Modeling Survival Data : Extending the Cox Model*. Springer, 2000.

[34] P. M. Woodward. *Probability and Information Theory, with Applications to Radar*. 1953.

Mémoires d'actuariat

[10] Gauthier Eldin. « Construction d'un indicateur de Valeur Client et optimisation tarifaire en assurance non-vie ». EURIA, 2018.

[18] Damien Hennom. « Création d'un indicateur de valeur client en assurance non vie ». ENSAE IP Paris, 2016.

[21] Linda Krolikowski. « Modélisation de l'élasticité au prix à la souscription en assurance automobile dans le cadre d'une optimisation tarifaire ». Ensae IP Paris, 2021.

[22] Markéta Krúpová. « Construction d'un modèle de Machine Learning interprétable pour la tarification en assurance non-vie ». Université Paris-Dauphine, 2023.

[25] Claire Lamon. « Modélisation et analyse de comportements clients en assurance dommage - application au changement de véhicules et à la résiliation de contrats ». Université Paris-Dauphine, 2019.

Note de synthèse

Contexte et problématique

Le marché de l'assurance automobile est caractérisé, du fait notamment de la nature du bien assuré, par un dynamisme constant. L'entrée de nouveaux acteurs, le développement de comparateurs de tarifs en ligne et les évolutions réglementaires contribuent à accroître la concurrence du secteur. Face à des clients d'autant plus informés, les assureurs redoublent d'efforts tarifaires et marketing pour capter de nouveaux prospects, induisant une augmentation substantielle des coûts d'acquisition. La rétention des clients déjà acquis s'avère alors essentielle à la pérennité économique des assureurs.

Dans ce contexte concurrentiel, la fidélisation des assurés constitue un enjeu majeur et ainsi, l'appréhension des comportements s'impose comme un outil nécessaire au pilotage efficace d'un portefeuille. Pour répondre à cette problématique de la rétention client, deux modèles, aux portées complémentaires, sont proposés. Le premier modèle est celui de la probabilité de résiliation à un an. Il fournit une appréciation fine du risque imminent de sortie du portefeuille, en s'appuyant sur des données précises concernant l'assuré, ses avenants, son contrat et le positionnement tarifaire de l'assureur le concernant. Le deuxième modèle, de la durée de vie du contrat a priori, conduit à la compréhension des comportements de fidélisation sur le moyen terme, à partir des données à la souscription.

Ce mémoire s'attache à la mise en place de modèles prédictifs des comportements de fidélité, à court et moyen termes, afin de :

— comprendre le comportement des assurés face à la résiliation en vue d'améliorer la relation client ;
— segmenter les profils dépendamment de leur rentabilité et de leur fidélité, et ainsi entamer des plans d'action de nature marketing ou commerciale adaptés ;
— améliorer les indicateurs de profit clefs du portefeuille, et obtenir les résultats escomptés suite à une stratégie tarifaire lors de la reconduction annuelle des contrats.

1. Vers des données exploitables dans le cadre de l'étude de la fidélité

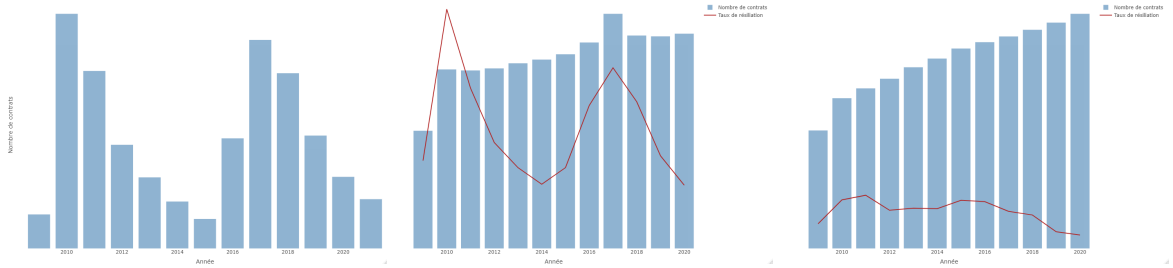
La base de données utilisée dans le cadre de cette étude comporte environ 1,4 million de contrats, souscrits entre 2009 et 2021, et répartis sur plus de 6 millions de lignes. Le traitement des données, inhérent à toute modélisation, permet d'asseoir la fiabilité des analyses et des modèles. Dans cette problématique de la fidélité, l'enjeu majeur est d'être en mesure d'identifier et de suivre les contrats de leur souscription à leur résiliation. De plus, nombre de variables explicatives des mouvements du portefeuilles sont créées.

La migration des contrats

Un frein à la qualité du suivi des contrats est celui de la migration. En effet, suite à des refontes tarifaires ou des lancements de nouveaux produits, les contrats peuvent être reconduits sous un autre numéro, compromettant l'identification des actes de souscription et de résiliation. Un travail conséquent de raccord des contrats, décrit dans le corps du mémoire, est mené afin de faire le lien entre les contrats migrés et les contrats initiaux. Ce dernier permet, avec fiabilité, d'identifier l'acte de résiliation et de calculer la durée de vie du contrat. La Figure 1 propose une appréciation de la correction apportée en représentant par année, de gauche à droite, le nombre de contrats migrés, l'évolution des taux de résiliation avant et après traitement. En bleu est présenté le nombre de contrats et en rouge les taux de résiliation. A l'issue de ce travail, les taux de résiliation Figure 1c ne présentent plus de pics et se situent autour de 14% ce qui est cohérent avec le marché de l'assurance automobile des particuliers.

Gestion des expositions

Dans le cadre d'une modélisation rigoureuse de la résiliation à un an, les états ne peuvent pas observer des expositions hétérogènes. En effet, supposons la présence de deux contrats observés sur



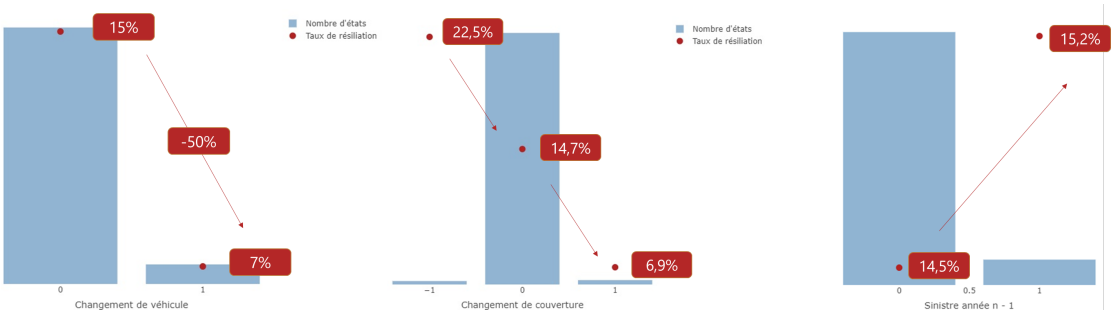
(a) Nombre de contrats migrés (b) Résiliation avant correction (c) Résiliation après correction

FIGURE 1 – Correction de l'erreur dans l'identification de la résiliation due aux contrats migrés

une unité de temps, l'un résilié et l'autre non. Alors le taux de résiliation est de 0,5. Cependant, si un des contrats est divisé sur l'unité de temps en trois états, alors le taux de résiliation calculé sera de 0,25. Le découpage des contrats en diverses états d'expositions inégales est courant dans les données. En effet, dès lors que le contrat subit un avenant, tel que la survenance d'un sinistre, le changement de véhicule ou de domicile, une nouvelle ligne est créée. Construire un modèle sur ces données brutes, dont les lignes observent des expositions de temps disparates, conduirait en des prédictions faussées. Un traitement est mené afin que chacune des lignes de la base corresponde à une exposition d'un an.

Identification des avenants

Les changements de situation, appelés avenants, sont des éléments importants de la vie des contrats. En effet, certains avenants, tels que le changement de véhicule, constituent un motif de résiliation. En revanche, les causes de la résiliation n'étant pas présentes dans les données, l'observation d'un avenant dans le portefeuille conduit, en moyenne, à une fidélisation accrue. Par exemple, un assuré ayant changé de véhicule, mais pas d'assureur aura tendance à moins résilier par la suite : cet assuré est fidélisé.



(a) Changement de véhicule (b) Changement de couverture (c) Survenance d'un sinistre

FIGURE 2 – Avenants et taux de résiliation

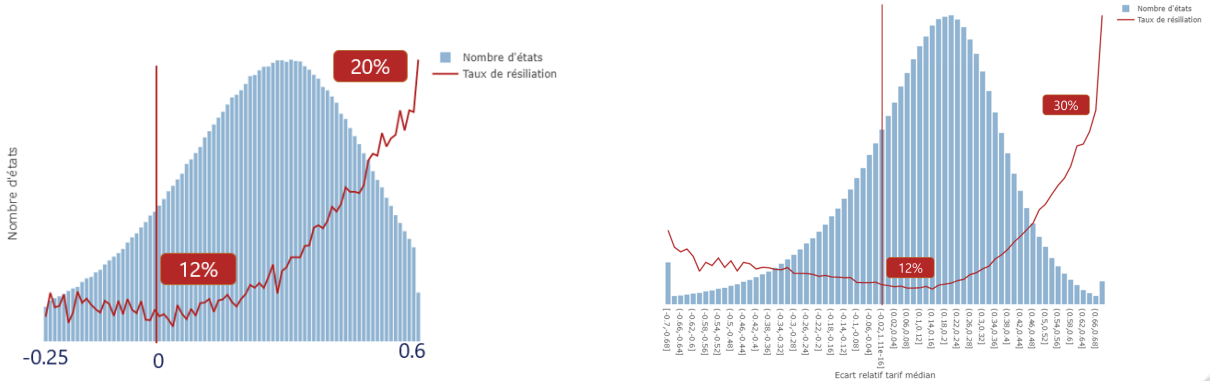
La Figure 2 présente les principaux avenants et leurs impacts sur la résiliation. En moyenne, un assuré venant de changer de véhicule résilie deux fois moins qu'un autre. Au niveau du changement de couverture, les effets sont symétriques. Un assuré diminuant son contrat, signe d'un véhicule déprécié et d'une recherche d'un tarif plus attractif, résilie 50% plus qu'un client ne modifiant pas sa couverture. Inversement, un individu qui augmente son contrat résilie deux fois moins en moyenne. De plus, la survenance d'un sinistre l'année $n-1$ impacte négativement la rétention de l'année n.

Variables tarifaires

La stratégie et le positionnement tarifaire de l'assureur impactent directement la rétention client. Ainsi, disposer de ces informations est essentiel à la qualité des modèles prédictifs. Au-delà du montant de cotisation, présent dans les données, les variables suivante sont intégrées :

— la prime pure, estimation du montant attendu du risque, est obtenue à partir de travaux de tarification réalisés en amont par le cabinet ;
— la marge correspond à l'écart relatif entre le montant de cotisation hors taxes proposé à l'assuré et son niveau de risque, quantifié par la prime pure ;
— le tarif médian du marché permet l'intégration de données concurrentielles [1] ;
— l'écart relatif au tarif médian retranscrit le positionnement de l'assureur sur le marché.

Les comportements de résiliation moyens en fonction du positionnement tarifaire de l'assuré sont proposés Figure 3. Une fois la marge et l'écart relatif au tarif médian positifs, les taux de résiliation moyens vont croissant.



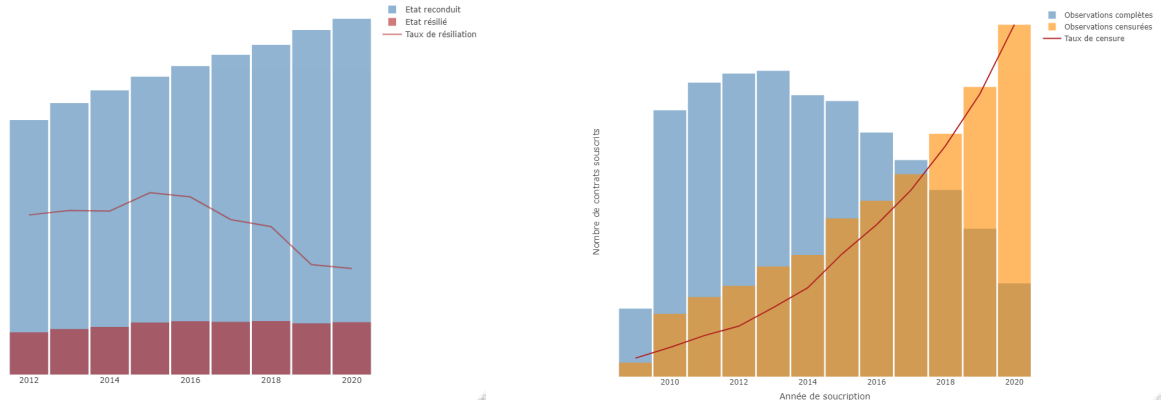
(a) Marge (b) Écart relatif au tarif médian

FIGURE 3 – Positionnement tarifaire et résiliation

Synthèse des bases de données

A l'issue des travaux réalisés sur le portefeuille d'assurés, deux bases de données sont obtenues, l'une pour l'étude de la résiliation à un an et l'autre pour la modélisation de la durée de vie.

1. Résiliation à un an : 2,7 millions d'états répartis sur 700 mille contrats, entre 2012 et 2021. La Figure 4a représente par an, les taux de résiliation, les états résiliés et ceux reconduits.
2. Durée de vie a priori : 650 mille contrats, souscrits entre 2009 et 2021. La Figure 4b représente les contrats souscrits par année, en bleu ceux résiliés, dont la durée de vie est connue, en orange les contrats censurés, encore en cours.



(a) Répartition des états reconduits et résiliés (b) Répartition des contrats souscrits, taux de censure

FIGURE 4 – Synthèse des bases de données

1. La collecte et l'intégration de données concurrentielles ont été exhaustivement réalisées par Mme. Linda Krolikowski, directrice de ce mémoire [21].

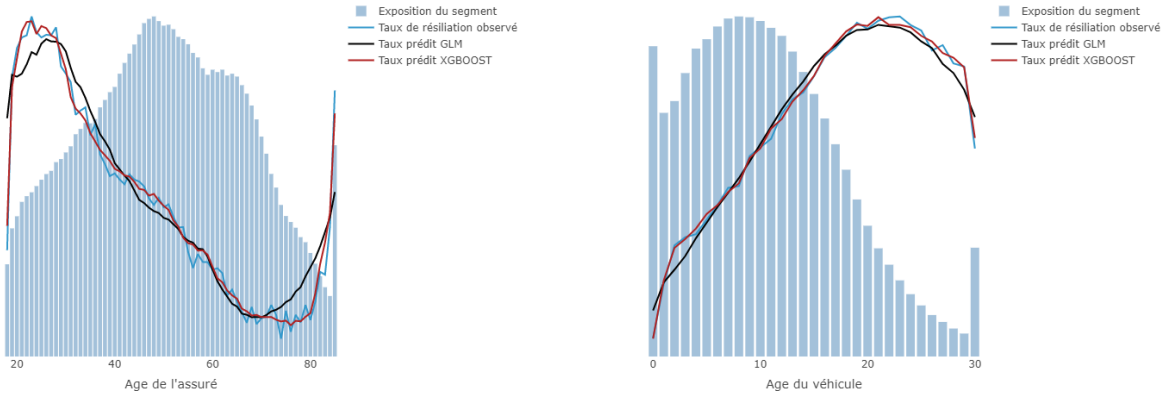
2. Probabilité de résiliation à un an

La prédiction de la résiliation à un an est réalisée au travers de deux outils :

— Régression logistique : appréciée pour sa transparence et son interprétabilité. Elle sera préférée dans le cadre d'une optimisation tarifaire pour ses propriétés de dérivabilité.

— XGBoost : appartenant à la famille des modèles dits *black box*, compte parmi les modèles de classification les plus performants, permet de fixer un élément de comparaison pour la régression.

Plusieurs métriques d'évaluation adaptées sont proposées dans le mémoire. Bien que le XGBoost dispose naturellement de meilleures performances, le gain n'est pas suffisant pour justifier de l'utilisation d'un modèle opaque. Une analyse segment par segment permet de s'assurer de la justesse de l'apprentissage. Les taux de résiliation en fonction de l'âge de l'assuré et de son véhicule sont pris en exemples Figure 5. Le modèle XGBoost, en rouge, s'ajuste avec plus de facilité aux taux observés, en bleu. Néanmoins, le modèle linéaire capte de manière satisfaisante les tendances principales.



(a) Âge de l'assuré (b) Âge du véhicule

FIGURE 5 – Étude des prédictions segment par segment - Échantillon de test

Finalement, les modèles peuvent être interprétés, au travers des valeurs estimées des coefficients pour la régression, et des shap values pour le XGBoost. Les shap values, Figure 6, mesurent l'impact de

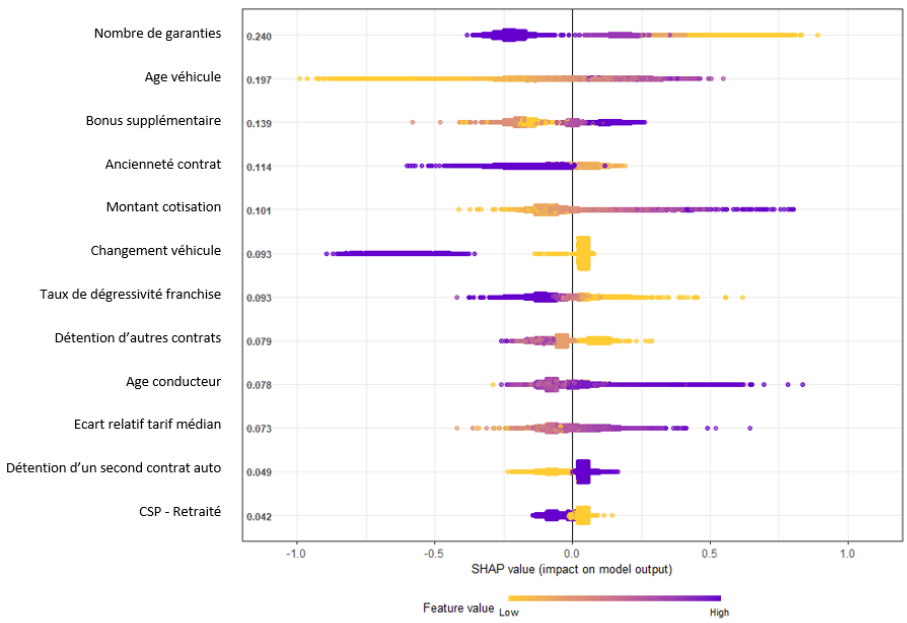


FIGURE 6 – Shap values (variables les plus importantes) - Prédiction de l'acte de résiliation

chaque variable sur les performances du modèle, et sont appréciées pour l'effet pur qu'elles expriment. Il est possible de réaliser diverses analyses au travers de leur étude. Par exemple, plus le nombre de garanties détenues est faible, en jaune sur le graphique, et plus la contribution marginale de l'observation augmente le risque de résiliation.

3. Durée de vie a priori

L'étude et la modélisation de durées consistent en l'estimation, pour un continuum d'instants $t \in [0; T]$, de la probabilité que l'individu soit toujours en portefeuille. L'estimateur non paramétrique de Kaplan-Meier permet l'analyse statistique des courbes de survie des individus. Le modèle de Cox, semi-paramétrique, explique la durée de vie des assurés à partir de leurs caractéristiques à la souscription. Après validation des hypothèses sur les données et sélection d'un jeu restreint de covariables, le modèle peut être interprété au travers des ratios de hasard Figure 7. Par exemple, un assuré dont le paiement est annuel résilie moins qu'un assuré dont la paiement est fractionné : le ratio de hasard entre ces deux modalités est de 0.87, ainsi le risque de sortie de portefeuille est diminué de 13% pour les clients s'acquittant de leur cotisation annuellement.

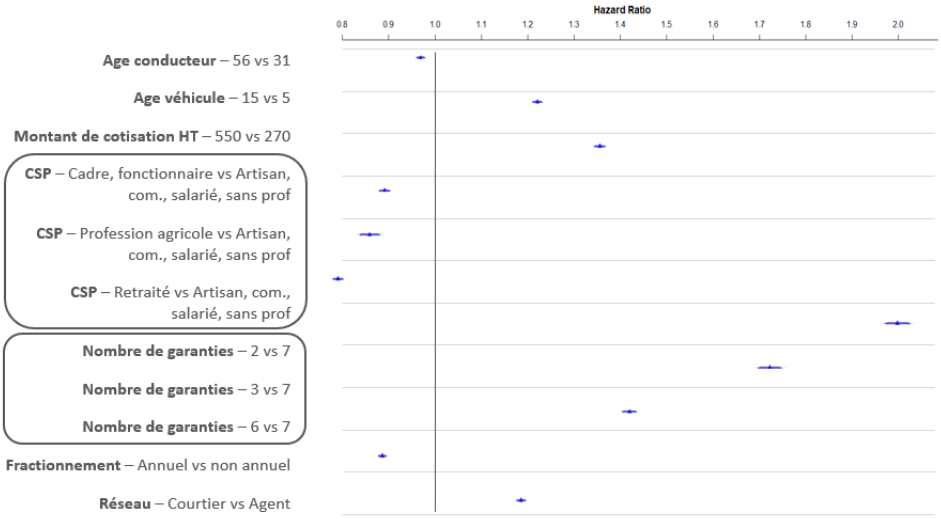
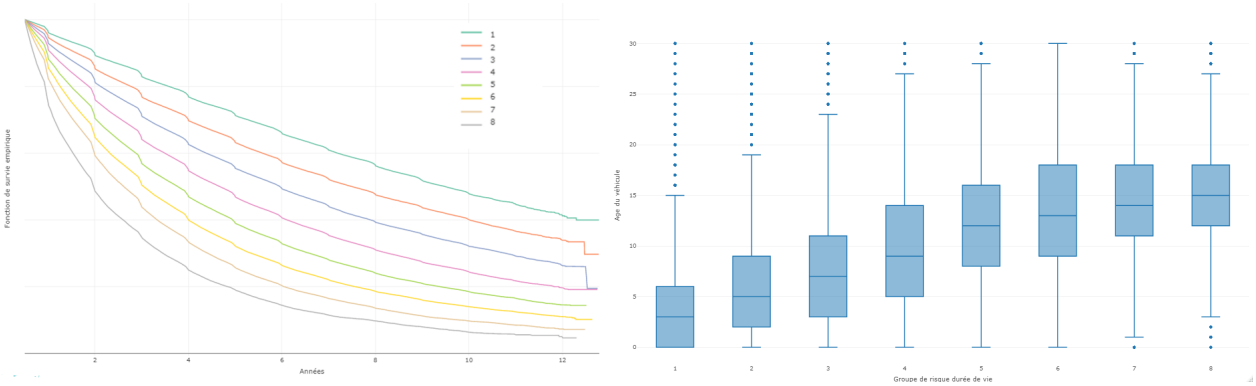


FIGURE 7 – Ratios de hasard - Modèle de Cox

4. Cartographie du portefeuille et scénario tarifaire

A partir des modèles construits, plusieurs outils, proposés dans une démarche d'aide à la décision, sont mis en place.

Premièrement le portefeuille est segmenté à l'aide de classifications ascendantes hiérarchiques. Ces classifications se font à partir des degrés de rentabilité, des comportements de fidélisation a priori et des élasticités du taux de résiliation au prix. Est proposée la classification réalisée en fonction du degré de fidélisation a priori de l'assuré. Figure 8a, les courbes de survie empirique des huit groupes construits sont présentées, le groupe 1 étant le plus fidèle. Les courbes, clairement distinctes, mettent en exergue la qualité de la segmentation. L'analyse de la composition des classes, comme soumis Figure 8b avec l'âge du véhicule assuré, donne matière à une compréhension d'autant plus juste des comportements. Il y est observé que plus un véhicule est âgé lors de la souscription du contrat, et plus ce dernier risque de quitter rapidement le portefeuille.



(a) Survies empiriques des classes (b) Âge du véhicule

FIGURE 8 – Segmentation en fonction du degré de fidélité a priori

Ensuite, des cartographies du portefeuille, toujours en fonction de la rentabilité, de la durée de vie a priori et de l'élasticité de la résiliation au prix sont mises en place. La Figure 9 propose de visualiser le portefeuille d'assurés en fonction de leur groupe de durée de vie en abscisse, et de leur marge en ordonnée. Des plans d'actions d'ordres commerciaux et marketing peuvent être entrepris par l'assureur à l'issue de l'analyse des différentes cartographies. Par exemple, pour les assurés du groupe A, fidèles et rentables, l'assureur qui bénéficie de la relation avec ces clients, peut cibler des actions pour récompenser leur fidélité. A l'inverse, les assurés du groupe D, ne sont pas rentables, mais très à risque de quitter rapidement le portefeuille. Les coûts de gestion et les actions menées pour ces clients doivent être minimisés.

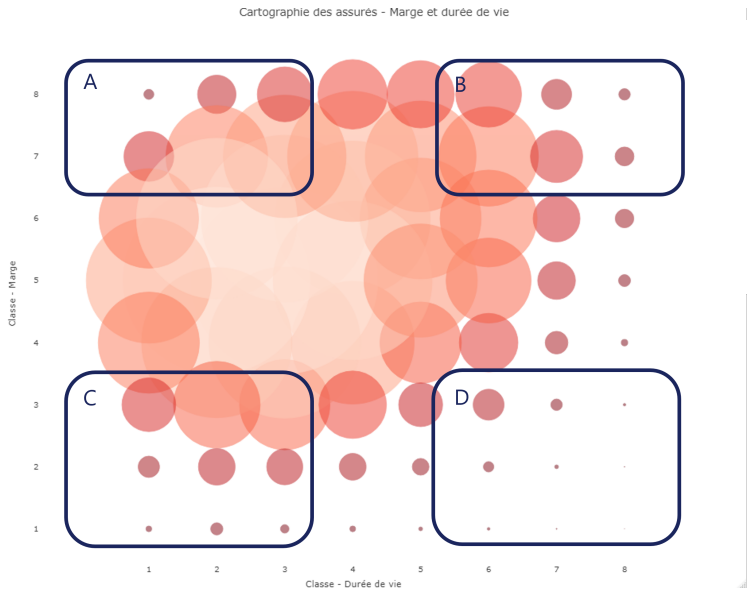
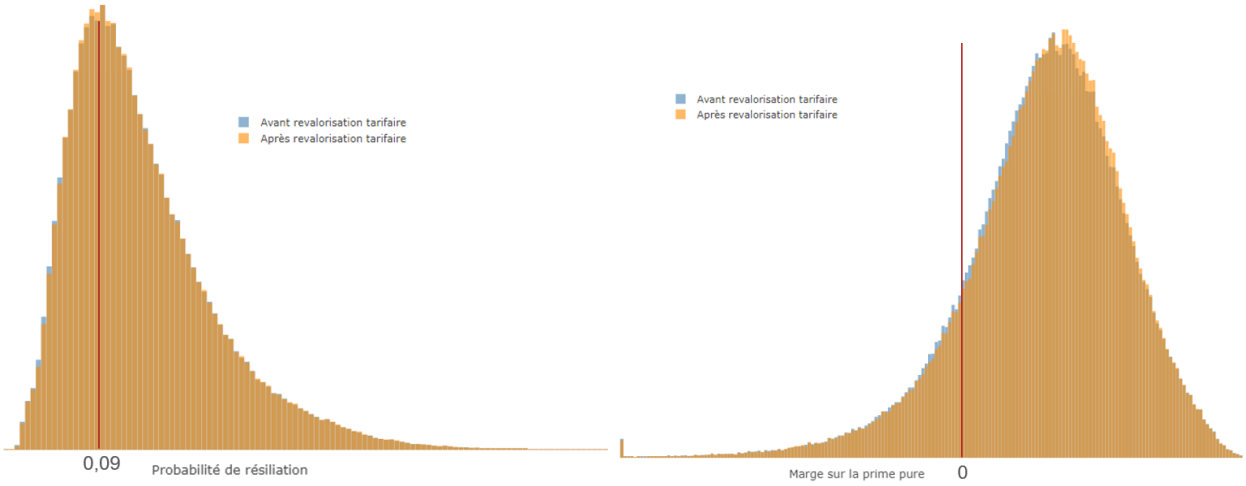


FIGURE 9 – Cartographie - Marge et groupe de durée de vie a priori

Finalement, les prémices d'une optimisation tarifaire sont proposés. Un scénario de revalorisation annuelle des contrats est étudié puis mis en œuvre sur le portefeuille des affaires en cours. A l'aide des divers modèles et outils proposés en amont, ce scénario rend compte des contraintes auxquelles sont soumises l'assureur. Il prend notamment soin de considérer les volumes d'assurés à moyen et long termes. En maintenant à leurs niveaux initiaux différents indicateurs, le scénario d'optimisation tarifaire permet une augmentation de 3% du profit. L'impact du scénario sur la probabilité de résiliation et la marge sont présentés Figure 10 : les taux de résiliation augmentent très marginalement quand la marge se voit sensiblement améliorée, sans décalage dans des niveaux extrêmes.



(a) Probabilité de résiliation (b) Marge

FIGURE 10 – Étude des impact du scénario tarifaire mis en place

Conclusion et perspectives

Au travers des modèles construits, de leurs interprétations et applications, le mémoire répond aux diverses problématiques posées. Les travaux menés permettent une compréhension fine des comportements de résiliation à court terme mais également de fidélité à moyen et long termes. L'assureur dispose de toutes les informations nécessaires pour entamer plans d'action visant à améliorer, d'une part, la relation qu'il entretient avec ses différents profils d'assurés, et d'autre part, obtenir les résultats attendus lors de ses démarches d'optimisation tarifaire. Ainsi, ce mémoire propose un premier niveau d'outils et de modèles, permettant à l'assureur une analyse de son portefeuille. Ce dernier est alors en mesure de mieux comprendre les mouvements qui s'y opère et de développer des actions pour cibler et fidéliser les clients à valeur.

Néanmoins, les outils d'aide à la prise de décision construits ont vocation à être améliorés. Premièrement, disposer des raisons de la résiliation permettrait de proposer des modèles plus fins, et l'assureur pourrait adapter sa politique sur certains axes de rétention. Ensuite, les modèles mis en place sont restés simples en vue d'être interprétables. Ces modèles étant maintenant stabilisés, faire appel à des modèles plus complexes, mais également plus opaques, permettrait d'affiner les outils implémentés. Finalement, ce mémoire propose une base à l'élaboration d'un indicateur de valeur client, captant en une seule mesure les degrés de rentabilité, sensibilité au prix et durée de vie des contrats.

Executive summary

Context and problem

The automobile insurance market is characterized by constant dynamism, particularly due to the nature of the insured property. The entry of new players, the development of online rate comparators and regulatory changes all contribute to increasing competition in the sector. Faced with more informed customers, insurers are redoubling their pricing and marketing efforts to attract new prospects, resulting in a substantial increase in acquisition costs. Retention of customers already acquired is therefore essential to the economic sustainability of insurers.

In this competitive context, the retention of policyholders is a major issue and thus, the apprehension of behaviors is a necessary tool for the efficient management of a portfolio. To address this issue of customer retention, two models with complementary scopes are proposed. The first model is the probability of termination within one year. It provides a detailed assessment of the imminent risk of portfolio exit, based on precise data concerning the insured, his changes in situation, his contract and the insurer's pricing position. The second model, of the a priori life of the contract, leads to an understanding of loyalty behavior over the medium term, based on data at the time of subscription.

This thesis focuses on the implementation of predictive models of loyalty behaviors, in the short and medium term, in order to :

— understand the behavior of the insureds in front of the termination in order to improve the customer relationship ;
— segment the profiles according to their profitability and loyalty, and thus initiate appropriate marketing or commercial action plans ;
— improve the key profit indicators of the portfolio, and obtain the expected results following a pricing strategy during the annual renewal of contracts.

1. Towards usable data for the study of customer loyalty

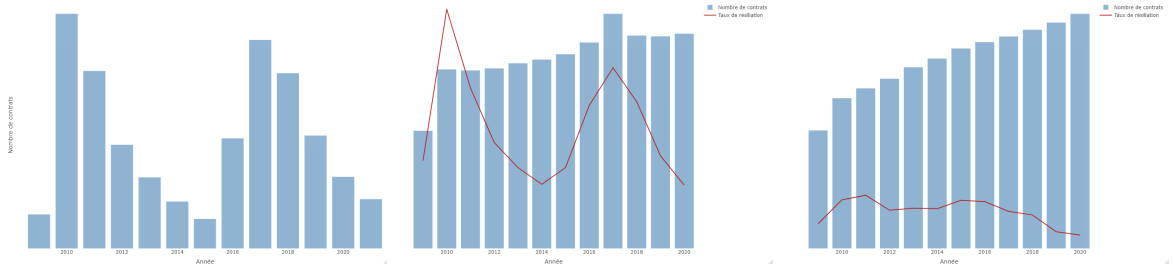
The database used in this study includes approximately 1.4 million contracts, signed between 2009 and 2021, and spread over more than 6 million lines. Data processing, which is inherent to all modeling, makes it possible to ensure the reliability of the analyses and models. In this loyalty issue, the major challenge is to be able to identify and track contracts from their subscription to their termination. In addition, a number of explanatory variables of portfolio movements are created.

Contract migration

One obstacle to the quality of contract monitoring is migration. Indeed, due to rate changes or new product launches, contracts may be renewed under a different number, compromising the identification of subscription and termination acts. A significant work of linking contracts, described in the main report, is carried out in order to make the link between the migrated contracts and the initial contracts. The latter makes it possible to reliably identify the act of termination and to calculate the life of the contract. Figure 1 proposes an appreciation of the correction made by representing by year, from left to right, the number of migrated contracts, the evolution of the cancellation rates before and after treatment. The number of contracts is shown in blue and the cancellation rates in red. At the end of this work, the cancellation rates Figure 1c no longer show peaks and are around 14

Exposures management

Under rigorous modeling of one-year termination, states cannot observe heterogeneous exposures. Indeed, assume the presence of two contracts observed over a unit of time, one terminated and the other not. Then the termination rate is 0.5. However, if one of the contracts is divided into three states on the time unit, then the calculated cancellation rate will be 0.25. The division of contracts into



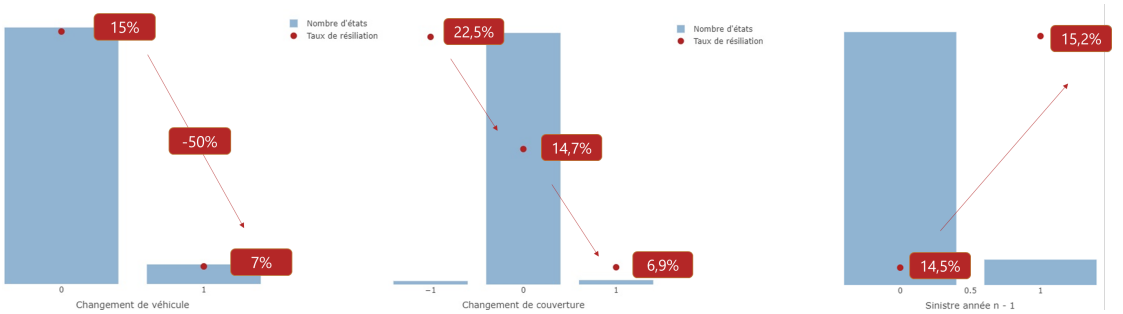
(a) Number of migrated policies (b) Termination before correction (c) Termination after correction

FIGURE 1 – Correction of the error in the identification of the termination due to migrated contracts

various states of unequal exposure is common in the data. Indeed, as soon as the contract undergoes an amendment, such as the occurrence of a claim, the change of vehicle or of residence, a new line is created. Building a model on this raw data, whose lines observe disparate time exposures, would lead to distorted predictions. A treatment is carried out so that each of the lines in the database corresponds to an exposure of one year.

Identifying changes in status

Changes in status are important elements in the life of a policy. Indeed, some changes, such as a change of vehicle, constitute a reason for termination. On the other hand, since the causes of termination are not present in the data, the observation of a rider in the portfolio leads, on average, to increased retention. For example, a policyholder who has changed vehicle but not insurer will tend to cancel less afterwards : this policyholder is more loyal.



(a) Vehicle change (b) Warranty change (c) Occurrence of a claim

FIGURE 2 – Change of status and termination rates

Figure 2 shows the main avenants and their impact on termination. On average, a policyholder who has just changed vehicle cancels half as much as another. The effects are symmetrical when it comes to changes in protection. A policyholder who reduces his contract, a sign of a depreciated vehicle and a search for a more attractive rate, cancels 50% more than a client who does not change his coverage. Conversely, an individual who increases his policy cancels half as much on average. In addition, the occurrence of a claim in year $n-1$ negatively impacts retention in year n.

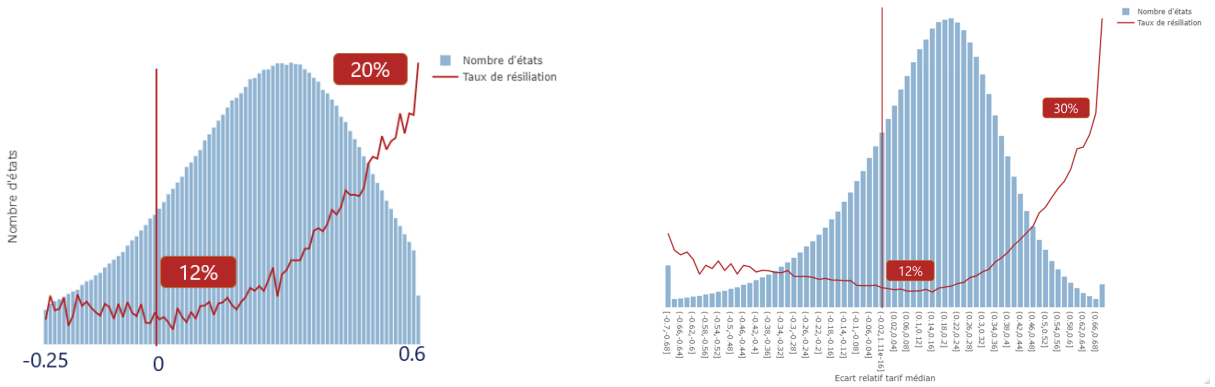
Pricing variables

The insurer's strategy and pricing position have a direct impact on customer retention. Thus, having this information is essential to the quality of predictive models. In addition to the amount of the premium, present in the data, the following variables are integrated :

— the pure premium, an estimate of the expected amount of the risk, is obtained from the pricing work carried out upstream by the cabinet ;

— the margin corresponds to the relative difference between the amount of the premium (excluding taxes) proposed to the insured and his level of risk, quantified by the pure premium ;
— the median market price allows for the integration of competitive data [2] ;
— the median rate spread reflects the insurer's market positioning.

The average cancellation behavior according to the insured's rate positioning is shown in Figure 3. Once the margin and the gap relative to the median rate are positive, the average cancellation rates increase.



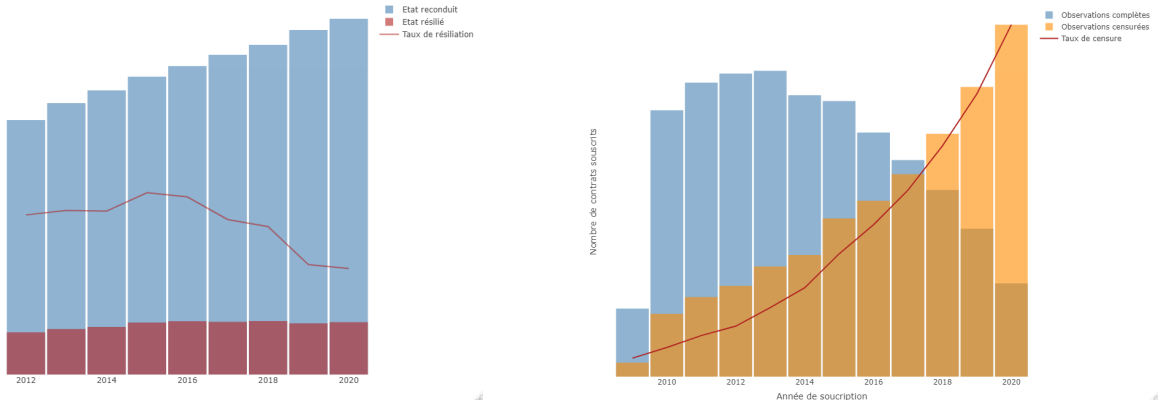
(a) Margin (b) Deviation from median price

FIGURE 3 – Price positioning and termination

Database synthesis

At the end of the work done on the portfolio of insureds, two databases are obtained, one for the study of termination at one year and the other for the modeling of the lifetime.

1. One-year termination : 2.7 million states spread over 700 thousand contracts, between 2012 and 2021. Figure 4a represents by year, the termination rates, the terminated states and the renewed ones.
2. A priori lifetime : 650 thousand contracts, signed between 2009 and 2021. Figure 4b represents the contracts subscribed by year, in blue those terminated, whose lifetime is known, in orange the censored contracts, still in effect.



(a) Distribution of renewed and terminated states (b) Distribution of contracts subscribed, censoring rate

FIGURE 4 – Database synthesis

2. The collection and integration of competitive data was done by Linda Krolikowski, the supervisor of this paper [21].

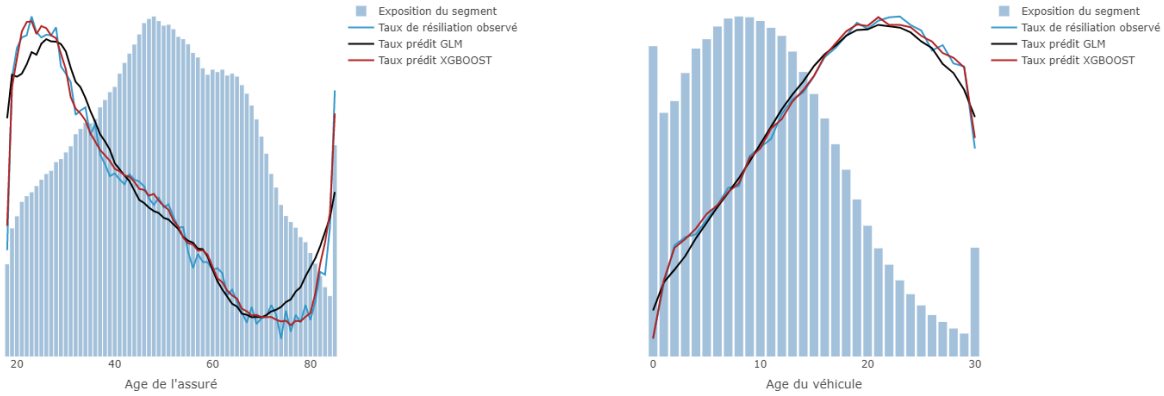
2. Probability of termination at one year

The prediction of one-year termination is performed through two models :

— Logistic regression : appreciated for its transparency and interpretability. It will be preferred in the context of price optimization for its derivability properties.

— XGBoost : belonging to the family of models known as black box, it is one of the most efficient classification models, it allows to fix a comparison element for the regression.

Several suitable evaluation metrics are proposed in the thesis. Although XGBoost naturally has better performance, the gain is not sufficient to justify the use of an opaque model. A segment-by-segment analysis ensures the correctness of the learning. The termination rates according to the age of the insured and his vehicle are taken as examples in Figure 5. The XGBoost model, in red, fits the observed rates, in blue, with greater ease. Nevertheless, the linear model captures the main trends well.



(a) Age of the insured person (b) Vehicle Age

FIGURE 5 – Segment by Segment Prediction Study - Test Sample

Finally, the models can be interpreted, through the estimated values of the coefficients for the regression, and the shap values for the XGBoost. The shap values, Figure 6, measure the impact of

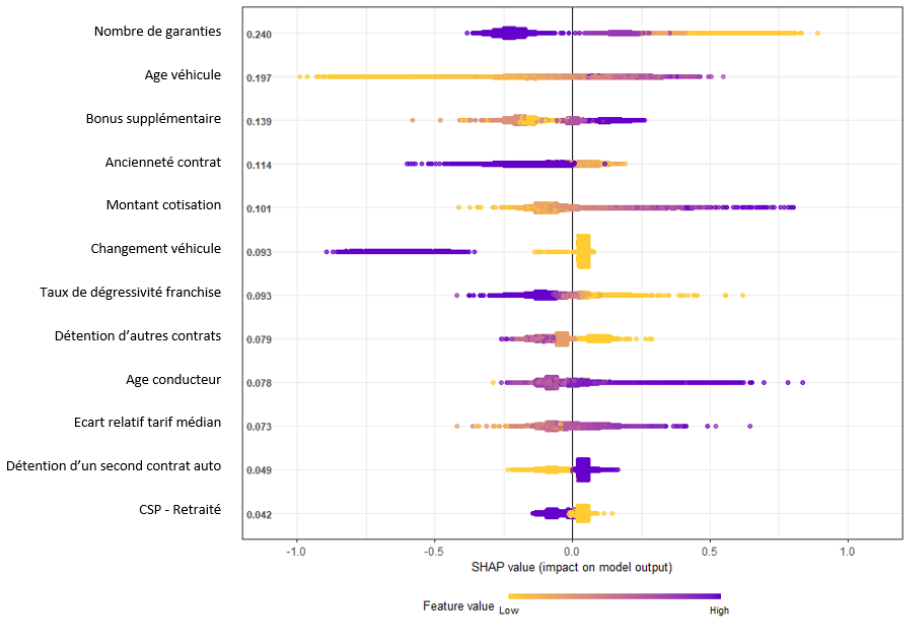


FIGURE 6 – Shap values (most important variables) - Prediction of the termination act

each variable on the performance of the model, and are appreciated for the pure effect they express.

It is possible to perform various analyses through their study. For example, the lower the number of guarantees held, shown in yellow on the graph, the higher the risk of termination.

3. Lifetime

The study and modeling of durations consist in estimating, for a continuum of instants $t \in [0; T]$, the probability that the individual is still in the portfolio. The non-parametric Kaplan-Meier estimator allows the statistical analysis of the survival curves of individuals. The semi-parametric Cox model explains the life expectancy of the insureds based on their characteristics at the time of subscription. After validation of the hypotheses on the data and selection of a restricted set of covariates, the model can be interpreted through hazard ratios Figure 7. For example, a policyholder with an annual payment terminates less than a policyholder with a split payment : the hazard ratio between these two terms is 0.87, so the risk of portfolio exit is reduced by 13% for clients who pay their premiums annually.

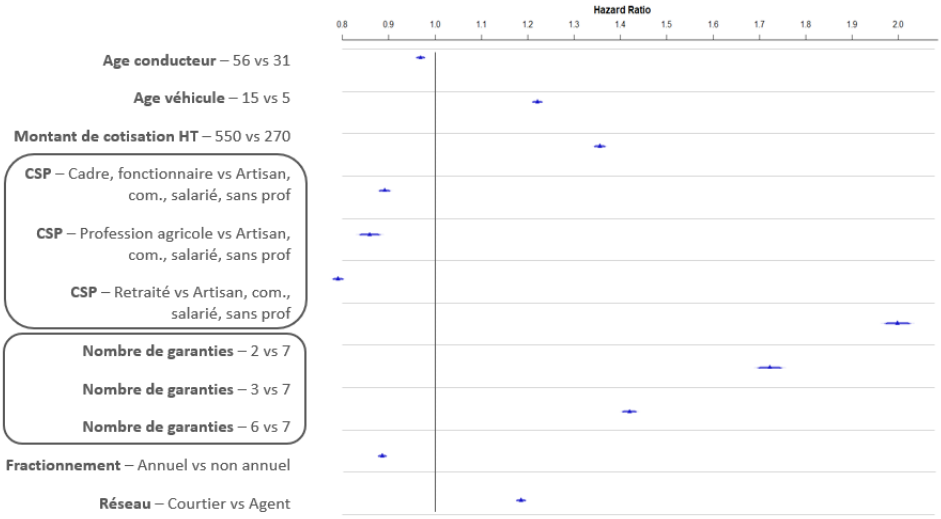
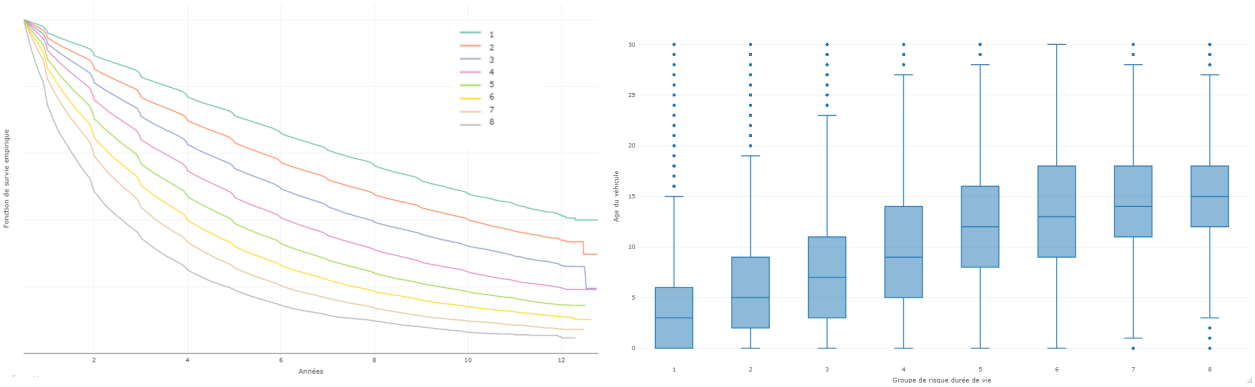


FIGURE 7 – Hazard ratios - Cox model

4. Portfolio mapping and pricing scenario

From the models built, several tools, proposed in a decision support approach, are implemented.

Firstly, the portfolio is segmented using ascending hierarchical classifications. These classifications are based on degrees of profitability, a priori loyalty behaviors and the elasticities of the termination rate to price. The classification carried out according to the degree of a priori loyalty of the insured is proposed. In Figure 8a, the empirical survival curves of the eight groups constructed are presented, group 1 being the most loyal. The curves are clearly distinct and highlight the quality of the segmentation. The analysis of the composition of the classes, as shown in Figure 8b, with the age of the insured vehicle, gives a more accurate understanding of the behavior. It is observed that the older a vehicle is at the time of contracting, the more likely it is to leave the portfolio quickly.



(a) Empirical class survivals (b) Vehicle age

FIGURE 8 – Segmentation according to the degree of loyalty

Next, portfolio mappings, again as a function of profitability, a priori lifetime, and price elasticity of termination are implemented. Figure 9 proposes to visualize the portfolio of insureds as a function of their lifetime group on the abscissa, and their margin on the ordinate. After analyzing the different maps, the insurer can undertake sales and marketing action plans. For example, for group A policyholders, who are loyal and profitable, the insurer who benefits from the relationship with these customers can target actions to reward their loyalty. On the other hand, group D policyholders are not profitable, but are very likely to leave the portfolio quickly. Management costs and actions for these clients must be minimized.

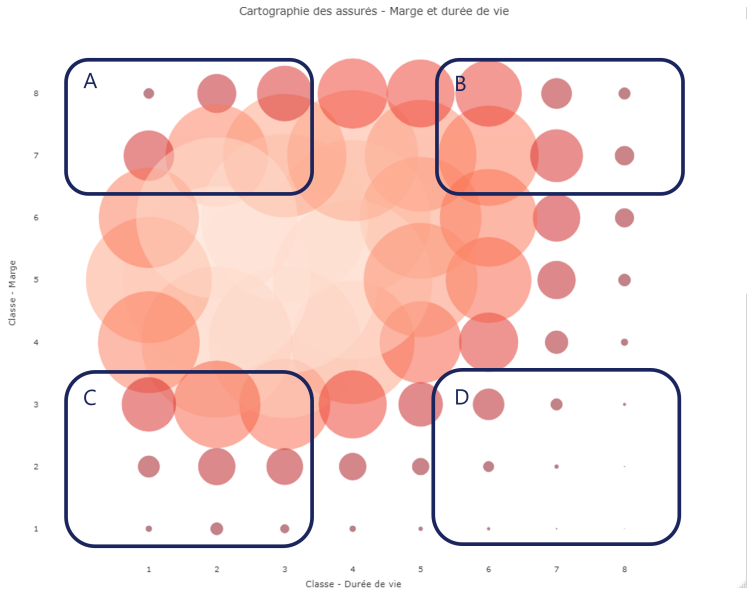
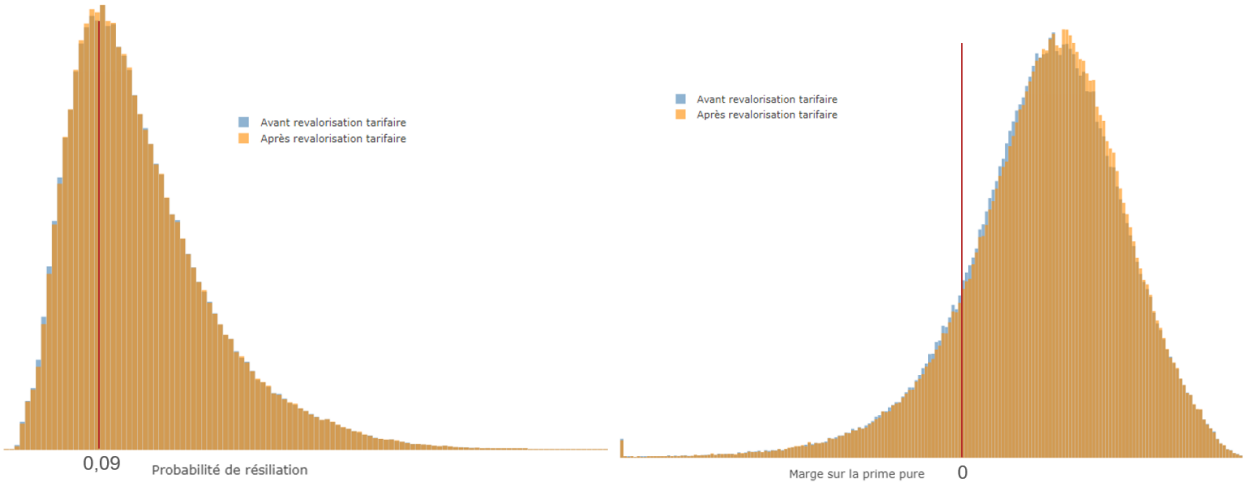


FIGURE 9 – Mapping - Margin and lifetime group

Finally, the beginnings of a tariff optimization are proposed. A scenario for the annual revaluation of contracts is studied and then implemented on the current portfolio. Using the various models and tools proposed beforehand, this scenario takes into account the constraints to which the insurer is subject. In particular, it takes care to consider the volumes of insureds in the medium and long term. By maintaining various indicators at their initial levels, the rate optimization scenario allows for a 3% increase in profit. The impact of the scenario on the probability of cancellation and the margin are presented in Figure 10 : the cancellation rates increase very marginally while the margin is significantly improved, without shifting to extreme levels.



(a) Probabilité de résiliation (b) Marge

FIGURE 10 – Impact study of the implemented tariff scenario

Conclusion and openings

Through the models built, their interpretations and applications, the thesis answers the various problems raised. The work carried out provides a detailed understanding of short-term cancellation behavior as well as medium and long-term loyalty. The insurer has all the necessary information to start action plans aiming at improving, on the one hand, the relationship it has with its different policyholder profiles, and on the other hand, to obtain the expected results during its price optimization process. Thus, this thesis proposes a first level of tools and models, allowing the insurer to analyze its portfolio. The latter is then in a position to better understand the movements that take place and to develop actions to target and retain valuable customers.

Nevertheless, the decision-making tools that have been built can be improved. Firstly, having the reasons for cancellation would make it possible to propose more refined models, and the insurer could adapt its policy on certain retention axes. Secondly, the models set up have remained simple in order to be interpretable. Now that these models have been stabilized, using more complex, but also more intransparent, models would allow us to refine the tools implemented. Finally, this thesis proposes a basis for the development of a customer value indicator, capturing in a single measure the degrees of profitability, price sensitivity and contract life.