IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Management
pour la validation du cursus à la Formation d'Actuaire
de l'Institut du Risk Management
et l'admission à l'Institut des Actuaires
le

Par : **Iuliana-Andreea OLARU**

Titre : **Identification des mesures tarifaires à l'affaire nouvelle en assurance moto**

Confidentialité : ☐ NON [x] OUI (Durée : ☐ 1 an [x] 2 ans)

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

Membres présents du jury de l'Institut des actuaires :

Entreprise : AXA France
Nom : Alix RAMBAUD
Signature :

Membres présents du jury de l'Institut du Risk Management :

Directeur du mémoire en entreprise :
Nom : Alix RAMBAUD
Signature :

Invité :
Nom :
Signature :

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

Secrétariat :

Signature du responsable entreprise

Bibliothèque :

Signature(s) du candidat(s)

Remerciements

Je tiens tout d'abord à adresser mes remerciements à Guillaume GORGE et Frédéric ARTRU pour m'avoir permis de réaliser ce mémoire au sein de la Direction de l'Offre IARD Particuliers Professionnels d'Axa France. Je les remercie également pour la confiance et le soutien qu'ils m'ont apporté tout au long de la formation du Centre d'Etudes Actuarielles.

Je souhaite exprimer toute ma gratitude envers ma responsable et tutrice de mémoire, Alix RAMBAUD, pour son encadrement régulier et avisé, qui m'a guidée avec intérêt et optimisme tout au long de ces travaux, et a largement contribué à l'achèvement de ce projet.

J'adresse notamment ma reconnaissance à mes anciens managers Gérald LUCAS et Sophie PASQUINELLI pour leurs conseils, leur disponibilité et l'aide qu'ils m'ont apportées pour réaliser la formation d'actuaire dans les meilleurs conditions possibles.

Enfin, je souhaite remercier mes proches pour leur soutien et leurs encouragements.

Résumé

Historiquement l'assurance deux-roues était un marché de niche qui attirait uniquement des acteurs ultra-spécialisés. Aujourd'hui le dynamisme des ventes de motos neuves entraîne l'émergence de nouveaux assureurs sur ce marché désormais confronté à une concurrence accrue. Par conséquent, les assureurs doivent continuellement revoir leur tarification et segmentation pour pouvoir proposer à la fois un tarif qui est techniquement cohérent et qui reste attractif par rapport à la concurrence.

L'objectif de ce mémoire est de proposer une stratégie de mise à jour du tarif à l'affaire nouvelle permettant de concilier la tarification technique du risque sous-jacent et le positionnement sur le marché de l'assurance. Cette modification se fera sous la contrainte d'une structure tarifaire équivalente. Nous ne nous intéressons donc pas à l'intégration de nouvelles variables dans le tarif commercial ou à une refonte intégrale de celui-ci.

A travers cette étude nous souhaitons répondre à deux questions :

- Quels sont les segments dont un ajustement du tarif est nécessaire ?
- De combien doit-on ajuster la prime commerciale afin d'améliorer la compétitivité du produit et/ou le mix du portefeuille sans dégrader les résultats techniques ?

Ce mémoire se décompose en cinq chapitres. Les deux premières chapitres permettent de présenter le contexte de l'étude et d'exposer la méthodologie utilisée pour la modélisation basée sur des techniques d'apprentissage automatique. Le troisième chapitre est consacré à l'analyse et à la modélisation du comportement des prospects à la souscription tandis que le quatrième chapitre vise à estimer la rentabilité sous-jacente attendue. Pour ce faire, nous parlerons de l'indicateur de rentabilité technique utilisé et nous allons détailler son calcul qui repose sur la refonte intégrale des modèles de prime pure par garantie. Pour finir, le dernier chapitre de ce mémoire présentera la stratégie d'ajustement de la prime commerciale basée sur l'analyse conjointe des deux indicateurs calculés précédemment ainsi que les résultats obtenus lors de l'application de cette stratégie.

Mots clés

Assurance moto, taux de conversion, prime pure, prime commerciale, garantie responsabilité civile matérielle, Expected Loss Ratio, gradient boosting, LASSO, sélection des variables

Abstract

Until recently, motorcycle insurance was a niche market that attracted only very specialized companies. Today, the dynamic of new motorcycle sales is leading to the emergence of new insurers on the market, which is now facing increased competition. As a result, insurers must continually review their pricing strategy and segmentation in order to be able to propose rates that reflect both the insured risk and remain attractive compared to the competition.

The objective of this dissertation is to propose a strategy for updating insurance rates while reconciling two concepts of the commercial premium : the technical pricing of the underlying risk and the global insurance market positioning. This update will be made taking into consideration the constraint of an equivalent rating structure. We are therefore not interested in integrating new variables into the actual ratebook nor in a complete re-do.

Through this study we wish to answer two questions :

- Which segments require rate adjustment ?
- How much should the commercial premium be adjusted in order to improve the competitiveness of the product and / or the mix of the portfolio without degrading the loss ratio ?

This thesis is mainly divided into five chapters. The first two chapters present the context of the study and expose the methodology used for modeling based on machine learning techniques. The third chapter is devoted to the analysis and the modeling of the customer demand while the fourth aims to estimate their expected profitability. To do this, we will introduce the concept of Expected Loss Ratio, one of the KPIs used for insurance portfolio management. We will go into details about its calculation which is based on the complete overhaul of pure premium models by warranties.Finally, the last chapter will present the strategy for adjusting the commercial premium based on the joint analysis of the two indicators calculated above as well as the results obtained while applying this strategy.

Key words

motorcycle insurance, conversion rate, pure premium, commercial pricing, third party liability, Expected Loss Ratio, gradient boosting, LASSO, variable selection

Sommaire

Remerciements **3**

Résumé **5**

Abstract **7**

Introduction **11**

1 Contexte et démarche de l'étude **13**
1.1 Présentation de l'assurance deux-roues 13
1.2 Problématique 17
1.3 Démarche suivie 22

2 Méthodologie de modélisation **27**
2.1 L'apport du machine learning 27
2.2 Modélisation avec des méthodes de linéarisation du gradient boosting 30

3 Modélisation du taux de transformation des devis **45**
3.1 Collecte et préparation des données 45
3.2 Analyse exploratoire des variables 48
3.3 Mise en œuvre de la méthodologie et résultats 53

4 Calcul de l'ELR **59**
4.1 Préliminaire : indicateurs de rentabilité 59
4.2 Modélisation de la prime pure 61
4.3 Calcul de l'ELR 86

5 Identification des mesures tarifaires **91**

Conclusion **101**

Bibliographie **103**

Annexes **105**

A Les catégories de permis moto **107**

B Le type de permis nécessaire par catégorie de deux-roues **108**

Liste des acronymes **109**

Table des figures **111**

Liste des tableaux **113**

Introduction

Les travaux présentés dans ce mémoire s'inscrivent dans le cadre du projet de mise à jour du tarif à l'affaire nouvelle du produit d'assurance moto d'AXA France.

En effet plusieurs mutations constatées sur le marché de l'assurance deux-roues incitent les assureurs non seulement à réviser de manière plus régulière leur tarification mais aussi à innover dans le domaine technique en tirant profit des avantages offerts par les outils d'apprentissage statistique.

Le marché de l'assurance deux-roues en France est actuellement confronté à une concurrence accrue. Cette concurrence est dûe à la multiplication des acteurs sur ce secteur, au développement fulgurant des assurances en ligne et des bancassureurs, ainsi qu'aux changements de comportement des assurés, moins fidèles à la marque et plus centrés sur le prix.

Dans ce contexte ultra-concurrentiel, les assureurs ont besoin de réinventer leur stratégie commerciale. Les techniques de tarification traditionnelles basées principalement sur la maîtrise du risque ne suffisent plus à relever les défis de croissance et de rentabilité demandés par l'entreprise. En effet, une approche basée uniquement sur la segmentation technique du portefeuille en fonction de la sinistralité estimée des assurés est considérée aujourd'hui comme obsolète et ne suffit plus aux compagnies pour se démarquer de la concurrence. Cette segmentation technique, aussi fine soit-elle, ne sera pas viable sans tenir compte du positionnement tarifaire sur le marché.

Pour tirer un maximum de valeur de son portefeuille, l'assureur doit mieux appréhender la demande d'assurance sur son marché et intégrer cette connaissance dans sa stratégie de prix. A travers ce mémoire nous proposons une stratégie de mise à jour du tarif à l'affaire nouvelle permettant de concilier la tarification technique du risque sous-jacent et le positionnement sur le marché de l'assurance, traduit par l'analyse du comportement des prospects au moment de la souscription.

L'approche utilisée peut se décomposer en trois grandes étapes :

1. L'analyse du comportement des prospects à la souscription : modélisation du taux de transformation de devis
2. L'estimation de la rentabilité future des devis : calcul de l'ELR (Expected Loss Ratio) comme rapport entre la prime pure prédite et la prime commerciale
3. L'analyse conjointe du taux de transformation et de la rentabilité estimée pour identifier les éventuelles évolutions tarifaires à envisager.

Dans le cadre de la modélisation du taux de transformation de devis et de la refonte des modèles de primes pures par garantie (modèles nécessaires au calcul de l'ELR) nous allons étudier la mise en œuvre d'une méthode de modélisation semi-automatisée basée sur les techniques d'apprentissage automatique. Nous allons exploiter le gradient boosting de bout en bout et transformer ensuite l'agrégat d'arbres décisionnels du modèle GBM en structure linéaire afin de proposer in fine une solution explicite et implémentable. Cette méthode a été développé et intégrée dans un outil de modélisation par Axa Global P&C et son but est de fournir une solution de production automatisée de modèles techniques, répondant aux critères de robustesse

et d'interprétabilité. Cette approche innovante a été choisie car, outre les critères de performance, elle offre un gain de temps conséquent grâce une procédure quasi-automatisée. De plus, les modèles ainsi construits sont facilement réutilisables et mis à jour par les autres actuaires de l'équipe. L'actuaire doit néanmoins porter un regard critique sur le modèle obtenu : son expertise métier et son opinion restent cruciales pour répondre aux critères d'interprétabilité, de robustesse, d'applicabilité opérationnelle tout en s'inscrivant dans la stratégie globale de développement de la branche.

Le premier chapitre du mémoire présentera le contexte économique et stratégique de l'assurance moto en France, ce qui permettra de bien situer la problématique de la mise à jour régulière du tarif commercial. On rappelle, entre autres, le cadre général de la tarification en assurance non-vie, les risques de modifications du tarif et on présente succinctement la démarche suivie.

Le second chapitre sera consacré à la méthodologie de modélisation. Après avoir synthétisé les particularités d'un modèle efficace, nous allons décrire la démarche globale de construction des modèles statistiques à la base de ce mémoire et mettre en évidence les différences par rapport à l'approche classique de modélisation. Nous allons introduire les outils permettant de valider et mesurer la performance des modèles d'apprentissage statistique et ensuite détailler étape par étape la méthodologie utilisée.

Le troisième chapitre vise à analyser le comportement des prospects à la souscription. Nous allons d'abord déterminer les critères expliquant l'attitude des clients potentiels en terme de transformation et ensuite construire, avec des méthodes de linéarisation du gradient boosting, un modèle de conversion de devis à partir de ces critères.

Le quatrième chapitre du mémoire est consacré au calcul de la rentabilité attendue des devis. Après avoir présenté les divers indicateurs de rentabilité utilisés pour le pilotage d'un portefeuille d'assurance, nous allons expliciter la notion d'ELR. Son calcul repose sur la refonte des modèles de primes pures par garantie. Nous allons modéliser, à travers une approche fréquence-coût moyen les quatre garanties principales du produit moto (RC matérielle, RC corporelle, incendie-vol et dommages) et illustrer comment cette approche de modélisation de la prime pure est mise en œuvre sur la garantie responsabilité civile matérielle. On calibrera ensuite les primes pures modélisées avec les charges finales prévisibles et on calculera l'ELR sur la base de prospects.

Le dernier chapitre présentera une stratégie d'ajustement de la prime commerciale basée sur une méthode graphique d'analyse conjointe des deux indicateurs calculés aux étapes précédentes : le taux de transformation et l'ELR. Cette partie permet de répondre à deux questions : quels sont les segments dont un ajustement du tarif est nécessaire et de combien doit-on ajuster la prime commerciale afin d'améliorer la compétitivité du produit et/ou le mix du portefeuille sans dégrader les résultats techniques ?

Chapitre 1

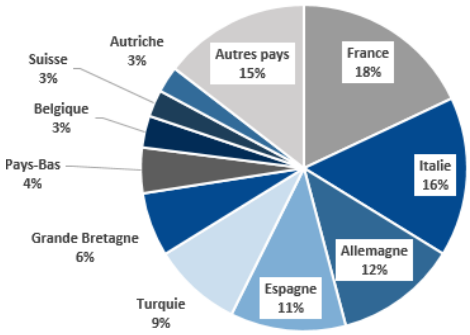
Contexte et démarche de l'étude

1.1 Présentation de l'assurance deux-roues

1.1.1 Le marché des deux-trois roues motorisés

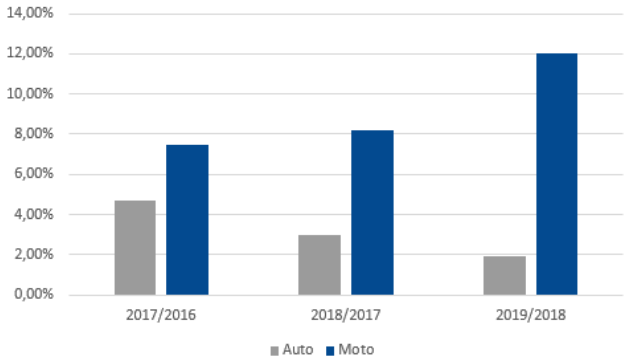
Un marché en forte croissance

Avec 274.000 ventes de deux et trois roues motorisés à fin octobre 2019 (source : MotorCycles Data, 2019), la France est le vingtième acteur du marché mondial et le leader sur le marché européen, devant l'Italie et l'Allemagne.



Graphique 1.1 – Distribution des ventes de deux et trois-roues motorisés en octobre 2019.

A fin 2019, l'évolution des immatriculations des deux et trois roues motorisés neuves était de +12% (source : Chambre Syndicale Internationale de l'Automobile et du Motocycle, 2019) par rapport à l'année 2018, tandis que le marché automobile français a enregistré une hausse de seulement 1,9% (source : Comité des Constructeurs Français d'Automobiles, 2019). En forte progression depuis plusieurs années, la dynamique du marché des 2/3 roues motorisés neuves dépasse chaque année celle enregistrée sur le marché de l'automobile, comme le montre le graphique ci-dessous :



Graphique 1.2 – Évolution annuelle des immatriculations neuves entre 2016 et 2019.

D'après la Fédération Française des Sociétés d'Assurances, en 2018 les deux roues représentaient seulement 7.5% du parc total de véhicules assurés contre 92.5% pour le parc automobile (source : Fédération Française de l'Asssurance, 2019). Suite à l'augmentation exponentielle du nombre des immatriculations neuves constatée sur le marché deux roues, la distribution du parc de véhicules motorisés va changer. Le produit moto présente donc un fort potentiel de développement des solutions d'assurances.

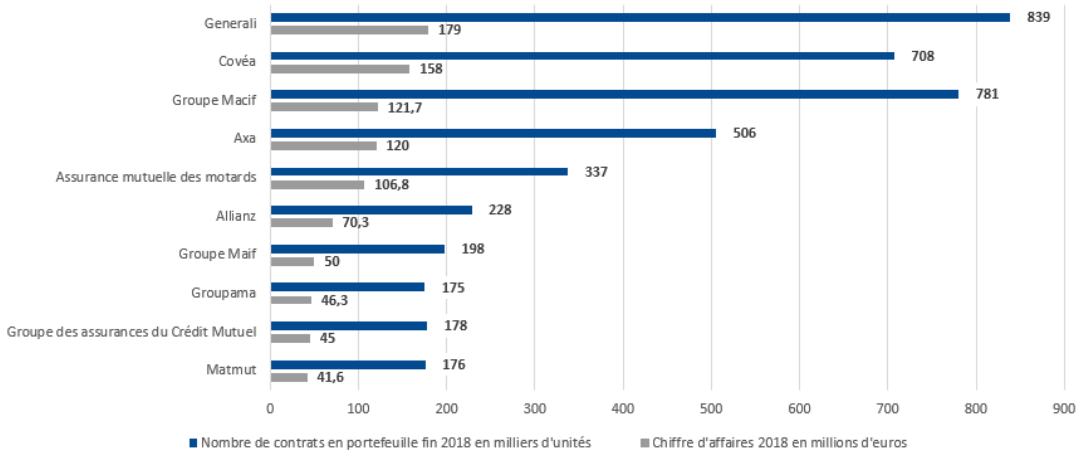
1.1.2 Le marché de l'assurance des deux-trois roues motorisés

Un marché qui évolue avec la société

Historiquement, les deux-roues représentaient un marché de niche attirant des acteurs ultra-spécialisés, qui partageaient avec leur client la même passion. Aujourd'hui, le profil des motards a changé. Les rues de plus en plus encombrées, les voitures de plus en plus chères et la montée en puissance des deux-roues électriques incitent les consommateurs à s'intéresser aux solutions de mobilité proposées par les scooters et les motos. Cette émergence de nouveaux utilisateurs, plus urbains, incite les assureurs généralistes et les bancassureurs à renforcer leurs solutions d'assurances pour le produit moto.

Un marché très concurrentiel

Le marché de l'assurance deux-roues est devenu depuis quelques années un marché hyperconcurrentiel, d'autant plus depuis la publication de la loi Hamon (2014), qui facilite la résiliation de l'assuré puisqu'il peut désormais résilier son contrat à tout moment dans l'année, après un an de souscription. Les nombreux acteurs qui se disputent aujourd'hui les parts de marché doivent rivaliser d'une part en termes de prix, et d'autre part en termes d'inventivité.



Graphique 1.3 – Classement des assureurs deux-roues en 2018 issu de l'Argus de l'assurance.

Dans ce marché hyperconcurrentiel Axa est en quatrième position et souhaite renforcer son positionnement sur ce marché de niche.

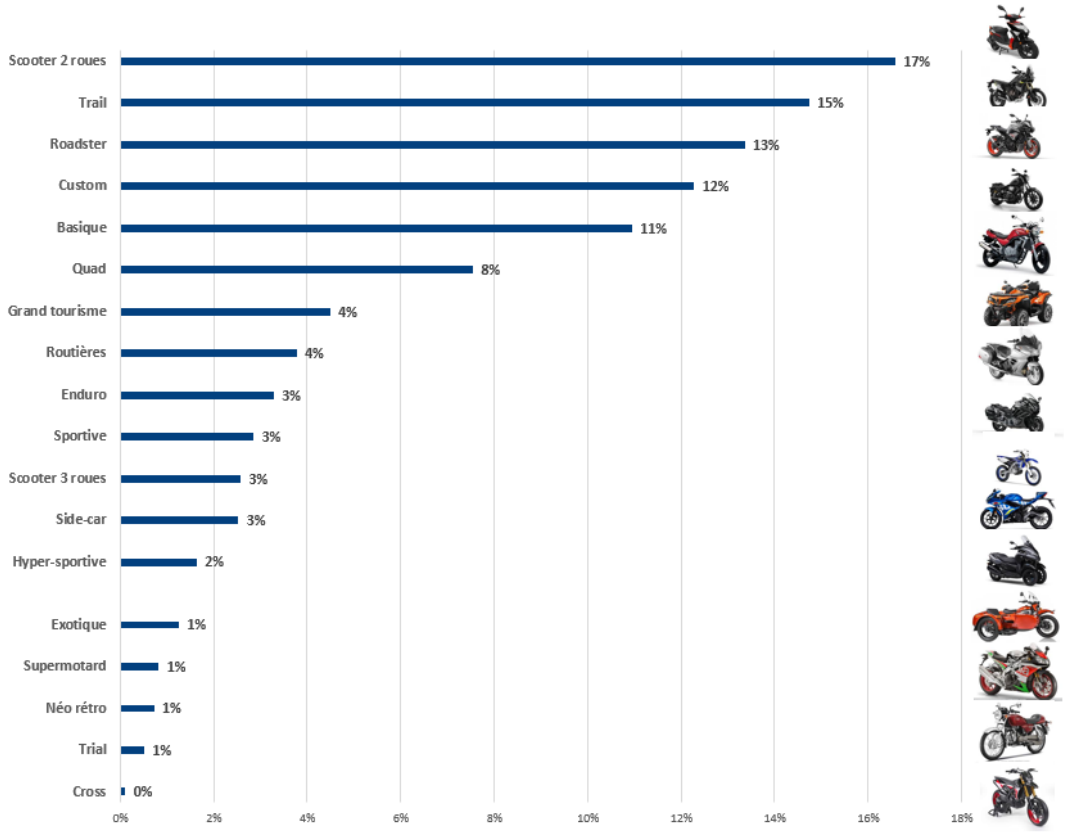
Un marché qui se transforme avec le digital

Le marché de l'assurance est en pleine mutation. Confronté à la concurrence accrue du secteur, dûe à la montée en puissance des comparateurs en ligne et à l'évolution grandissante de la part de marché des bancassureurs, les assureurs ont besoin de réinventer leur stratégie commerciale. Outre la digitalisation du parcours d'achat, nous observons l'essor d'une nouvelle génération hyperconnectée, moins fidèle à la marque et plus centrée sur le prix. Pour rester compétitifs sur ce marché, les assureurs doivent être en mesure d'offrir une meilleure expérience client grâce à la personnalisation des offres et ils doivent disposer d'outils leur permettant d'être flexibles et proactifs pour faire face à la concurrence.

1.1.3 L'offre AXA

Cette étude est basée sur le produit co-proposé par AXA France et le courtier Run Services qui ont noué un partenariat avec Club14 (C14), une association de passionnés de moto.

Contrairement à ce que son appellation pourrait laisser supposer, le produit moto ne concerne pas seulement les véhicules à deux roues, mais également les trois roues (tels que les scooters à trois roues et les trikes) et certaines catégories de quatre roues à moteurs (de type quad, ssv et buggy). Les différentes catégories de véhicules couvertes dans le produit moto sont présentées dans la figure ci-dessous.



Graphique 1.4 – Distribution du portefeuille d'assurances AXA par catégorie de deux et trois roues motorisés.

1.1.3.1 Les garanties

Le produit d'assurance moto AXA s'articule autour de plusieurs garanties.

Les garanties principales

- la garantie **Responsabilité Civile** (RC) ou assurance dite « au tiers » : elle constitue la seule garantie obligatoire qui couvre les dommages matériels et corporels causés à autrui.
- la garantie **Sécurité du conducteur** (SDC) : elle permet de couvrir les dommages corporels subis par le conducteur du véhicule assuré.
- la garantie **Incendie Vol** (IV) : elle s'applique en cas de vol, d'incendie ou d'explosion de véhicule.
- la garantie **Dommages collision** (DC) : elle permet de couvrir les dégâts matériels causés au véhicule assuré lors d'une collision avec un tiers identifié.
- la garantie **Dommages tous accidents** (DTA) : elle assure la même couverture que la garantie dommages collision mais sans condition sur l'identification du tiers impliqué.

Les garanties complémentaires

- la garantie **Défense pénale et recours suite à accident** (DPRSA) : elle permet le remboursement des honoraires d'avocats et d'experts engagés suite à un accident où la responsabilité civile ou pénale de l'assuré est impliquée.

- la garantie **Protection juridique** (PJ) : elle permet de bénéficier d'une assistance juridique en cas de litige lié au véhicule assuré, et pas seulement dans le cadre d'un accident de la route.

- la garantie **Assistance aux personnes** : elle couvre les frais de rapatriement au domicile des personnes blessées ou ne disposant plus de moyens de se déplacer ainsi que de leurs effets personnels, les frais médicaux liés à la panne ou à l'accident, l'hébergement sur place si cela s'avère nécessaire (longs trajets de vacances par exemple).

- la garantie **Assistance au véhicule** : elle permet de couvrir les frais de dépannage et de remorquage en cas d'accident.

- la garantie **Accessoires et vêtements** : elle prend en charge les dommages causés aux vêtements et aux accessoires moto lors d'un accident.

Selon le type de contrat choisi, les différentes garanties offrent une couverture limitée au tiers (RC) ou une couverture plus vaste, étendue au véhicule et à son conducteur. L'assuré peut choisir de compléter les garanties du socle avec d'autres couvertures facultatives, proposées en option, telles que la protection juridique confort, la sécurité du conducteur étendue, les accessoires avec un capital élargi, la garantie véhicule de remplacement et la garantie valeur à neuf.

1.1.3.2 Les formules

Quatre formules sont mises à disposition des assurés choisissant AXA. Celles-ci proposent différents niveaux de protection aux assurés, allant de l'assurance « au tiers » (protection minimale, F1) à l'assurance « tous risques » (protection maximale, F4) :

Garanties	F1	F2	F3	F4
Responsabilité Civile	Inclus	Inclus	Inclus	Inclus
D. P. R. S. A.	Inclus	Inclus	Inclus	Inclus
Protection Juridique	Inclus	Inclus	Inclus	Inclus
Décès du Conducteur	Inclus	Inclus	Inclus	Inclus
Assistance aux Personnes	Inclus	Inclus	Inclus	Inclus
Assistance au Véhicule sans franchise	Inclus	Inclus	Inclus	Inclus
Sécurité Du Conducteur (jusqu'à 15000 € - Franchise AIPP 15%)	Inclus	Inclus	Inclus	Inclus
Casque	Inclus	Inclus	Inclus	Inclus
Gilet Airbag	Inclus	Inclus	Inclus	Inclus
Incendie/Vol, Evénements Climatiques, Catastrophes Naturelles et Technologiques		Inclus	Inclus	Inclus
Dommages Collision			Inclus	Inclus
Dommages Tous Accidents				Inclus

Tableau 1.1 – Formules de garanties proposées dans le produit moto AXA.

Les conditions d'acceptation du risque reposent principalement sur des critères d'âge du conducteur, d'ancienneté du permis et sur l'adéquation entre le profil du pilote et le véhicule qu'il souhaite assurer.

A RETENIR :

- Le marché français des deux et trois roues motorisés est en pleine expansion avec une évolution des immatriculations neuves de +12% en 2019, dépassant largement la tendance constatée sur le marché automobile (de +2% seulement). Le marché moto offre donc un vrai potentiel de croissance à travailler pour les assureurs.
- Le marché de l'assurance deux-roues est devenu hyperconcurrentiel depuis quelques années. Il faut donc se démarquer des autres, en proposant un tarif attractif et une offre personnalisée. Pour ce faire, les assureurs doivent mettre à jour de manière régulière leur tarif commercial afin de proposer un tarif compétitif et ainsi gagner des parts de marché tout en assurant la pérennité de la branche dans le temps.
- Les travaux présentés dans ce mémoire s'inscrivent dans le cadre de la mise à jour annuelle du tarif à l'affaire nouvelle du produit d'assurance moto d'AXA France. L'offre s'articule autour de 4 formules d'assurances, couvrant la sinistralité survenue sur environ 12 garanties.

1.2 Problématique

L'objet de cette étude porte sur l'arbitrage entre le tarif technique et commercial, l'enjeu étant de trouver le bon curseur entre ces deux indicateurs pour assurer le développement commercial tout en garantissant la rentabilité de la branche. Cet exercice complexe soulève également la problématique du compromis à trouver entre la finesse dans la segmentation/personnalisation du tarif versus la mutualisation des risques. Avec la montée en puissance d'internet les assurés sont devenus de plus en plus sensibles au prix. Pour pouvoir faire face à la concurrence extrême, les assureurs doivent être en mesure de prendre en compte dans le tarif commercial :

- le risque sous-jacent
- l'évolution du comportement des clients face à la digitalisation

L'objectif de ce mémoire est donc de concilier dans la tarification le double enjeu de la prime commerciale : la tarification technique du risque sous-jacent et le positionnement sur le marché de l'assurance, traduit par l'analyse du comportement des prospects au moment de la souscription.

1.2.1 Principes de tarification et nécessité de segmenter

Afin de comprendre comment ces deux notions sont prises en compte dans la tarification d'un contrat d'assurance, il est nécessaire de comprendre comment cette prime est calculée.

Les composantes de la Prime Commerciale

La prime d'assurance, appelée également Prime Commerciale (PC), est la prime réellement versée par l'assuré. Elle représente le prix que l'assuré doit payer pour bénéficier de la couverture du risque en cas de sinistre. Suite à l'inversement du cycle de production, phénomène spécifique au secteur de l'assurance, la prime commerciale doit être calculée sans connaître le montant de l'indemnité qui sera effectivement versée par l'assureur en cas de sinistre.

La prime commerciale est composée de cinq éléments :

$$PC = PP + CH + CS + T + M, \tag{1.1}$$

avec

- **la Prime Pure (PP)** : elle correspond à la tarification du risque, à savoir au coût probable des sinistres sur la période de couverture (l'espérance de la charge de sinistres).

- **les Chargements pour Frais (CH)** : ils permettent de financer les différents coûts supportés par l'assureur pour l'exercice de son activité tels que les coûts d'acquisition des contrats, les frais de gestion des contrats, les frais d'administration des contrats.

- **les Chargements de Sécurité (CS)** : ils permettent à l'assureur de se protéger contre une faillite éventuelle en cas de sur-sinistralité imprévue.

- **les Taxes (T)** : les produits d'assurance sont assujettis à une taxe fiscale plus ou moins importante selon les garanties, réglementée par le Code des Assurances [1]. Les montants collectés sont alors versés par les compagnies d'assurance au Trésor Public. En plus de la taxe fiscale, la prime d'assurance comprend également une contribution pour financer des fonds ou organismes de solidarité nationale, tels que la Sécurité Sociale.

- **la Marge (M)** : elle comprend la rémunération des fonds propres demandée par les actionnaires et la marge bénéficiaire de l'assureur sur le produit.

La prime commerciale, réellement payée par l'assuré, peut grandement différer de la prime pure, issue du modèle de tarification. Cette dernière constitue la base de la prime commerciale. L'éloignement entre la prime commerciale et la prime pure résulte de la stratégie commerciale de l'entreprise d'assurance et de la prise en compte de la concurrence. Ces éléments amènent les assureurs à revoir leur tarification en apportant des correctifs, parfois individualisés, à la prime commerciale.

La nécessité de segmenter et le phénomène d'anti-sélection

La tarification en assurance représente la clef de la croissance rentable de l'entreprise. Un aspect important dont dépend la réussite d'un tarif performant est la bonne connaissance et compréhension de ses prospects et de ses clients afin de proposer à chacun le tarif le plus adéquat à ses besoins, à son profil et à son comportement. Pour atteindre cet objectif, les assureurs font appel aux techniques de segmentation de la clientèle, opération essentielle dans un marché si fortement concurrentiel.

La segmentation en assurance a été définie comme l'opération par laquelle l'assureur distingue les risques qu'il prend en charge, afin de les classer dans des catégories de risques homogènes et de leur appliquer un traitement adéquat en ce qui concerne le tarif et la garantie.

Une segmentation fine permet, outre une amélioration de la connaissance du coût des sinistres, une meilleure compréhension du comportement des clients, une amélioration de sa position concurrentielle et la rétention des clients les plus rentables du portefeuille. Prenons un exemple simple afin de comprendre la nécessité de segmenter son portefeuille, tout particulièrement en assurance deux-roues. Soient deux assureurs ayant les caractéristiques suivantes :

- Un assureur A, qui décide de ne pas segmenter son portefeuille.

- Un assureur B, qui segmente son portefeuille en fonction de la « qualité du conducteur ».

Chaque assureur propose un contrat proposant les mêmes garanties. Le tableau ci-dessous présente les différentes primes proposées par chacun des assureurs :

	Mauvais conducteurs	Bons conducteurs
Assureur A	300 €	300 €
Assureur B	600 €	200 €

Graphique 1.5 – Primes proposées par les assureurs A et B pour deux profils de conducteurs

1. Sauf pour les mutuelles, régies par le Code de la Mutualité.

L'assureur A propose la même prime d'assurance pour les bons et mauvais risques. Ainsi, il fait un profit avec les clients à faible probabilité d'avoir un sinistre et une perte avec les assurés avec forte probabilité d'avoir un sinistre. La prime perçue par les « bons conducteurs » sert, en partie, à couvrir les sinistres des « mauvais conducteurs ». L'assureur B propose, quant à lui, deux tarifs différents en fonction du profil de risque, déterminés à partir de variables telles que l'âge ou les antécédents de sinistres. Ainsi, les « mauvais condcuteurs » paient une prime supérieure du fait de leur exposition au risque.

Les conséquences de la présence de segmentation dans le portefeuille sont les suivantes :

- Les assurés prudents, ayant une faible probabilité d'avoir un sinistre et que l'on qualifiera de « **bons risques** » décideront de s'assurer auprès de l'**assureur B**, proposant des primes plus faibles.

- L'**assureur A**, proposant un tarif unique, verra venir s'assurer chez lui la plupart des personnes ayant un comportement à risque avec une probabilité élevée d'avoir un sinistre, qualifiés de « **mauvais risques** ».

En conclusion, la mise en concurrence de ces deux assureurs entraîne des migrations d'assurés d'un portefeuille vers l'autre. L'assureur choisissant de ne pas segmenter son portefeuille se retrouve dans un position délicate. En effet, il récupère tous les « mauvais risques », d'où la nécessité de segmenter son portefeuille d'assurés. Ce phénomène est connu sous le nom **d'antisélection**.

Les limites de la segmentation

L'exemple précédent, certes très simpliste, démontre bien la nécessité pour l'assureur de segmenter. Néanmoins, celle-ci présente des limites et ne peut être poussée à l'extrême. En effet, lorsqu'elle est trop fine, c'est-à-dire lorsque l'on tend vers un tarif individualisé, l'estimation de la prime pure est basée sur une classe tarifaire d'effectif trop faible et la validité du théorème des grands nombres est remise en question. L'estimation de la charge moyenne de sinistre peut être dans ce cas biaisée conduisant à de mauvais résultats pour l'assureur.

Il vaut donc mieux rester prudent et ne pas segmenter à l'extrême. Par conséquent, il faut savoir trouver le bon niveau de segmentation : les individus regroupés dans une même classe tarifaire présentent une certaine hétérogénéité. Le principe de mutualisation s'applique au sein de chacune des classes.

Segmentation en assurances : un compromis à trouver entre le tarif technique et commercial

Chaque assureur possède des techniques de tarification et de segmentation qui lui sont propres, en adéquation avec les profils de risque de son portefeuille clients. Au début des premières segmentations tarifaires, les assureurs utilisaient principalement des techniques statistiques d'analyse univariée, alors qu'aujourd'hui les assureurs ont migré vers des techniques multivariées, les plus répandues étant l'utilisation des modèles linéaires généralisés.

Dans un environnement concurrentiel, le choix des critères de segmentation, le degré de mutualisation plus ou moins important, le choix de faire des catégories de risques plus ou moins nombreuses, sont au cœur du métier d'actuariat.

Dans ce contexte, l'utilisation d'une approche basée entièrement sur le risque s'avère insuffisante, d'où la nécessité de prendre en considération la position sur le marché et de modifier la tarification en conséquence. Cette nouvelle approche de tarification permet de croiser deux enjeux qui s'opposent : la volonté de se développer commercialement (passant par une sous-tarification volontaire des affaires nouvelles) et la nécessité de garantir la rentabilité de la branche (passant par un tarif techniquement viable). La seconde notion est primordiale, car elle permet à l'assureur de faire face aux charges de sinistres futures. Il faudrait donc trouver un compromis entre ces deux notions, un seuil de rentabilité sous lequel on ne veut pas descendre, en dépit d'un mauvais positionnement sur le marché.

1.2.2 Le taux de transformation d'un contrat

Les étapes de transformation

Avant de détailler le parcours de souscription d'un contrat moto, nous allons d'abord caractériser le réseau de distribution d'AXA. L'entreprise distribue ses produits principalement à travers des intermédiaires traditionnels : un réseau d'agents classique qui vend les produits dans les agences, un réseau de salariés (plutôt orienté vie) et un réseau de courtiers proposant à leur clients les produits AXA mais également les produits d'autres assureurs. Le produit moto est également ouvert sous certaines conditions à la vente sur internet (suite au partenariat noué entre AXA et l'association Club 14).

Cependant, nous allons limiter notre étude aux devis réalisés à travers le réseau de distribution classique (agents, courtiers, salariés) car les interrogations tarifaires effectuées sur le site internet de Club 14 ne représentent pas un effectif assez significatif. En outre, le profil du client souhaitant souscrire son assurance deux roues sur internet diffère de manière significative de celui qui souhaite s'assurer à travers un intermédiaire traditionnel. Pour s'assurer donc de l'homogénéité de la base de données qui servira à notre analyse, les devis réalisés sur internet ont été écartés de l'étude.

Pour mieux appréhender l'acte de conversion des devis, nous allons par la suite illustrer le processus conduisant à la création d'une affaire nouvelle. Une fois arrivé en agence, le client potentiel doit répondre à un certain nombre de questions concernant :

- le souscripteur
- le conducteur principal
- les caractéristiques du véhicule à assurer
- les antécédents de sinistralité
- les formules de garanties
- la détention d'autres contrats AXA etc.

Ces éléments sont renseignés par l'agent/courtier dans le système d'informations et permettent d'obtenir un tarif. Après avoir vu le tarif, l'agent à la possibilité d'effectuer ou non l'enregistrement de la simulation tarifaire. Si l'enregistrement n'est pas réalisé, l'interrogation tarifaire n'est pas sauvegardée dans les bases des données et nous n'avons donc aucune trace des informations du prospect.

Si le prospect est intéressé par cette simulation de tarif, l'agent peut réaliser cinq actions :

1) Réaliser un **contrat** (appelé également affaire nouvelle directe) : dans ce cas, le prospect doit fournir à l'agent tous les documents dont il a besoin pour souscrire le contrat (carte grise, permis de conduire, attestation de l'ancien assureur etc.)

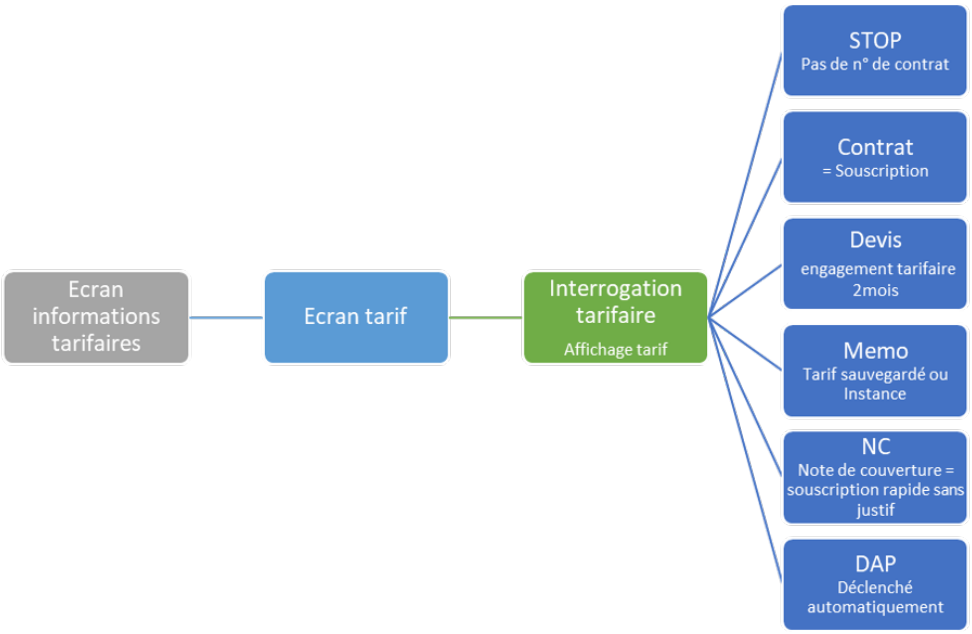
2) Réaliser un **devis** : l'agent s'engage à garder le tarif présenté inchangé pour une période limitée de temps (2 mois)

3) Réaliser un **mémo** : l'interrogation tarifaire est sauvegardée sans engagement sur le tarif proposé

4) Réaliser une **note de couverture** (NC) : souscription rapide d'un contrat temporaire, en l'absence des documents justificatifs. Si au bout de deux mois l'agent n'a pas encore reçu les documents justificatifs de la part du client, le contrat est resilié.

5) Réaliser une **demande d'autorisation préalable** (DAP) : si l'interrogation tarifaire concerne un profil interdit à la souscription, une DAP est déclenchée automatiquement. Dans ce cas, le contrat est mis en attente le temps que le siège décide la suite à donner concernant la souscription du risque.

Pour résumer ces différents éléments, le schéma représentant le parcours d'interrogation tarifaire est résumé ci-dessous :



Graphique 1.6 – Parcours de souscription en agence.

Le taux de transformation : indicateur du positionnement tarifaire sur le marché

Dans le cadre de cette étude nous souhaitons capter le positionnement d'Axa sur le marché moto. Pour ce faire nous avons fait le choix de modéliser la probabilité de transformation d'une interrogation tarifaire sauvegardée en affaire nouvelle. Deux indicateurs complémentaires permettent de mesurer la compétitivité du produit : le taux de pénétration et le taux de transformation. Le taux de pénétration retranscrit la part de marché prise par l'entreprise sur l'ensemble des immatriculations du marché. Il adresse donc d'avantage le sujet de répartition du réseau de distribution face à la demande. C'est un sujet de maillage indépendant du tarif. C'est pourquoi nous avons fait le choix de ne pas en tenir compte dans le cadre de ce projet. En revanche, nous allons nous concentrer sur l'indicateur du taux de transformation, indicateur qui permet de mesurer l'impact du tarif sur la concrétisation, et de ce fait, du positionnement tarifaire face à la concurrence.

Le but de l'étude est donc de savoir comment modifier la prime commerciale de manière segmentée afin d'améliorer le chiffre d'affaires et d'éviter en même temps le phénomène d'antisélection. Pour connaître le positionnement tarifaire sur le marché on se base donc sur le taux de transformation. Il est important de noter que dans le domaine de l'assurance classique (avec un réseau de distribution d'agents principalement), les agents ont un levier d'action sur le tarif final proposé aux prospects, à travers les rabais commerciaux. Ceux-ci sont accordés principalement lorsqu'il y a une mise en concurrence du tarif (présence d'un devis avec un tarif plus attractif de la concurrence), ou dans le cas d'un contexte client très favorable (client multi détenteur).

Nous souhaitons identifier d'éventuels segments sur lesquels les clients ne transforment pas bien, afin de baisser le prix dans la limite de la prime pure espérée. Ainsi, sans dégrader les résultats techniques, nous allons augmenter l'apport net (améliorer le chiffre d'affaires) et gagner des parts de marché. Dans la même logique, nous souhaitons identifier les segments sur lesquels les clients ont un très bon taux de transformation et des résultats techniques moins bons. En ajustant leur tarif à la hausse, nous allons conserver une charge commerciale globale suffisante en améliorant le mix des affaires nouvelles, sans trop détériorer l'apport net.

La première étape de l'analyse, qui est donc essentielle, est de modéliser le comportement des prospects face à la transformation.

1.2.3 Risques des modifications de tarif

Les modifications de tarif en assurance non-vie ne sont pas sans risques. En premièr lieu, une segmentation insuffisante peut générer un effet d'antisélection, et deuxièmement, une inadaptation du tarif par rapport au risque sous-jacent pour favoriser la compétitivité peut entraîner des conséquences négatives sur la solvabilité de l'entreprise.

Les ajustements tarifaires décrits plus loin feront en sorte de maximiser le nombre d'affaires nouvelles sous contrainte de la rentabilité. En effet, pour maîtriser le risque d'antisélection la diminution de la prime commerciale pour les segments dont la conversion est moindre doit se limiter au seuil de la prime pure simulée. On sera alors assuré de couvrir au minimum la charge de sinistres moyenne du prospect.

Les changements réalisés sur la prime commerciale de l'affaire nouvelle ont un fort impact sur le mix des affaires nouvelles et donc également sur la structure du portefeuille. En modifiant la prime sur un segment on attire de nouveaux profils clients et donc des risques potentiellement différents ce qui aura une incidence sur la prime pure. Si la modélisation initiale de la prime pure se base sur un grand nombre d'observations (la sinistralité du portefeuille deux-roues constatée sur trois années d'observations), cette dernière ne devrait pas changer de manière significative après modification du tarif. Cependant si la prime pure n'est pas suffisamment segmentée et/ou basée sur une faible part du marché, il est possible de voir la modélisation de la prime pure évoluer. Il faut alors prévoir une modélisation de la prime pure de manière régulière, ce qui est généralement le cas dans les compagnies d'assurance. De plus, un suivi régulier de la rentabilité estimée des affaires nouvelles devrait être mis en place pour s'assurer que les ajustements tarifaires n'ont pas comme effet une détérioration de cet indicateur.

En raison de ces risques inhérents à la modification de la prime commerciale, il est préférable de limiter les ajustements tarifaires. A ce titre, on se contente donc de limiter les ajustements de la prime commerciale sur le principe suivant : l'évolution de chaque coefficient ne pouvant pas dépasser 5%, à la hausse ou à la baisse. Non seulement cette contrainte va minimiser les deux risques que l'on vient de voir (en évitant une transformation brutale du portefeuille), mais il permettra aussi de ne pas donner une importance exagérée à la probabilité de transformation.

A RETENIR :

- L'objectif de ce mémoire est de concilier dans la tarification deux notions de la prime commerciale : la tarification technique du risque sous-jacent et le positionnement sur le marché de l'assurance.
- L'indicateur qui sera utilisé pour estimer l'attractivité du produit d'assurance deux-roues d'AXA est le taux de transformation des devis. Nous allons chercher à identifier les segments avec des bons résultats techniques mais qui ne transforment pas bien afin de baisser leur prime commerciale. Inversement, nous allons augmenter le tarif sur des segments présentant des mauvais résultats techniques et un bon taux de conversion.
- Les ajustements tarifaires en assurance peuvent avoir des conséquences négatives tant sur la structure du portefeuille que sur la solvabilité de l'entreprise. Nous préconisons donc de rester prudent et de limiter le niveau de chaque ajustement tarifaire afin d'éviter une transformation brutale du portefeuille.

1.3 Démarche suivie

Les travaux présentés dans ce mémoire s'inscrivent dans le cadre de la mise à jour annuelle du tarif à l'affaire nouvelle du produit d'assurance moto d'AXA France.

Dans un environnement fortement concurrentiel de l'assurance deux-roues, lors de la modification du tarif commercial, le rôle de l'actuaire est de proposer non seulement un tarif qui est à l'image du risque garanti mais aussi de s'assurer que le tarif reste attractif par rapport au marché afin de convertir le plus de prospects possibles.

A travers ce mémoire on propose une méthode de mise à jour du tarif à l'affaire nouvelle qui permet de concilier les deux notions de la prime commerciale : la tarification technique du risque et le positionnement sur le marché de l'assurance.

De plus, cette méthode doit prendre en considération une contrainte relative à l'ajustement de la prime d'assurance : la modification tarifaire doit être réalisée à structure tarifaire équivalente. Nous ne nous intéressons donc pas à l'intégration de nouvelles variables dans le tarif commercial ou à une refonte intégrale de celui-ci. Notre objectif est de faire évoluer le tarif commercial actuellement en place :

- à la hausse : pour les prospects avec un bon taux de conversion et une faible rentabilité estimée
- à la baisse : pour les prospects dont la transformation est moindre si leur rentabilité estimée la permet

L'approche utilisée peut se décomposer en trois grandes étapes :

- Analyse du comportement des prospects à la souscription
- Estimation de la rentabilité future des devis
- Analyse conjointe du taux de transformation et de la rentabilité estimée pour identifier les mesures tarifaires.

Décrivons plus en détail chaque étape.

1. Analyse du comportement des prospects à la souscription

La première étape consiste à mieux appréhender le comportement des prospects au moment de la souscription. Nous allons d'abord déterminer les critères expliquant l'attitude des clients potentiels par rapport à la transformation et ensuite construire un modèle de conversion de devis à partir de ces critères. Il est important de préciser que, suite à la contrainte imposée d'ajustement du tarif à structure tarifaire équivalente, le modèle de transformation final ainsi construit prendra en compte uniquement des critères présents dans la structure tarifaire actuellement en place.

L'objectif de cette première étape est, d'une part, d'identifier les critères pour lesquels un ajustement du tarif commercial est nécessaire (concrètement, sont concernées par une modification du tarif commercial les variables choisies dans le modèle de conversion ayant le pouvoir explicatif le plus élevé [2]) et d'une autre part de déterminer l'impact toutes choses égales par ailleurs de chacun de ces critères sur la transformation. L'utilisation des valeurs toutes choses égales par ailleurs pour le taux de transformation de devis lors de l'identification des mesures tarifaires permet d'assurer la prise en compte de l'effet propre de chacune des variables explicatives sur la conversion, indépendamment des autres.

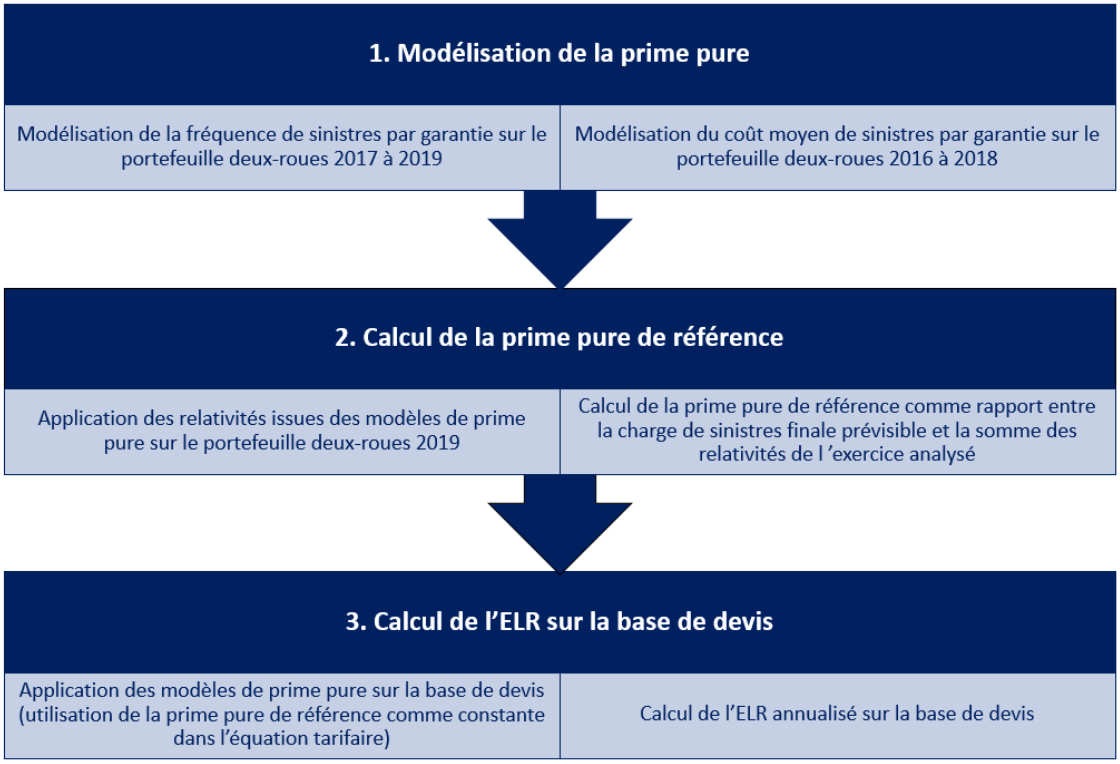
2. Estimation de la rentabilité future des devis

La deuxième étape de l'étude vise à estimer la rentabilité théorique annuelle des prospects à travers l'indicateur d'**Expected Loss Ratio** ou **ELR**. Calculé comme rapport entre la prime pure modélisée et la prime commerciale, l'ELR reflète la rentabilité technique annuelle estimée de façon individuelle au niveau de chaque contrat/devis. Puisque la prime pure est une estimation de la charge que devra payer l'assureur en cas de sinistre, l'ELR présente l'avantage de pouvoir être calculé au niveau des devis également; il sera donc utilisé dans le cadre de ce mémoire comme indicateur de la rentabilité théorique des prospects [3].

2. Dans le cas d'une variable avec des taux de conversion similaires pour toutes ses modalités l'ajustement du tarif est jugé superflu, le tarif commercial actuellement en place étant conforme à la pratique du marché.

3. Pour de plus amples détails concernant cet indicateur le lecteur intéressé est invité à se référer à la section 4.1.

L'estimation de la rentabilité théorique des prospects se décompose en trois étapes représentées dans le schéma 1.7 ci-dessous :



Graphique 1.7 – Schéma estimation de la rentabilité théorique des prospects.

Décrivons plus en détails chaque étape pour une meilleure compréhension.

2.1. Modélisation de la prime pure

La première étape consiste à bâtir des modèles de prime pure par garantie à l'aide des caractéristiques du contrat AXA, encore appelées «variables internes». On peut citer les caractéristiques de la moto assurée (la catégorie de la moto, la marque, la puissance, la cylindrée etc.), l'âge du conducteur, l'ancienneté de permis, les antécédents de sinistres et d'assurances etc.

Les modèles de primes pures seront construits à travers une approche classique de fréquence-coût moyen, sous condition que le volume de données soit suffisant.

Nous allons illustrer la manière dont cette approche de modélisation de la prime pure est mise en œuvre sur la garantie responsabilité civile matérielle. Cette garantie a été privilégiée car, d'une part, elle est la seule garantie obligatoire lors de la souscription d'un contrat d'assurance deux-roues (son taux de souscription est de 100%) et d'autre part elle représente le risque avec le plus grand volume de sinistres.

Cette démarche de modélisation est appliquée sur les trois autres garanties principales : la responsabilité civile corporelle, l'incendie et le vol (les deux garanties modélisées ensemble) et la garantie dommages tous accidents. Pour éviter trop de répétitions dans cette présentation on se limite à illustrer la méthodologie utilisée et les résultats obtenus uniquement sur la garantie RC matérielle.

A l'égard des autres garanties proposées à la souscription (telles que la protection juridique, la garantie assistance, le bris d'optique etc.), le faible volume de données disponibles a appuyé notre choix de les exclure de la modélisation. En réalisant une modélisation du risque qui se base sur un échantillon de taille insuffisante nous risquerions d'obtenir des résultats qui ne sont pas statistiquement significatifs et de tirer des conclusions qui ne sont pas fiables. C'est pour cette raison que nous avons décidé de ne pas réaliser de segmentation de la prime pure

en fonction des variables de risque pour ces garanties secondaires. Cependant, leurs charges sinistres seront prises en compte dans le calcul de la prime pure totale par l'intégration d'un montant forfaitaire de prime pure pour chaque contrat ayant souscrit la garantie.

2.2. Calcul de la prime pure de référence

Après avoir modélisé la fréquence et le coût moyen des sinistres pour les garanties principales, la deuxième étape dans le calcul de l'ELR est la calibration du montant de la prime pure de référence ultime. La prime pure de référence doit permettre d'obtenir en moyenne un ratio de S/P pur ultime de 100% sur l'exercice étudie (elle doit prendre en compte également des provisions pour les sinistres en cours et les sinistres tardifs). Cette quantité sera alors fixée garantie par garantie et calculée de la manière suivante pour un exercice donné :

- Application des modèles de fréquence et coût moyen par garantie et calcul de la prime pure (comme produit de ces deux composantes) sur le portefeuille deux-roues
- Estimation des charges ultimes de sinistres garantie par garantie (mise à disposition des données annuellement par les équipes de provisionnement) sur l'année étudiée
- Calcul de la prime pure de référence comme rapport entre les charges ultimes estimées et la prime pure modélisée

2.3. Calcul de l'ELR sur la base de devis

La dernière étape de cette partie vise à calculer l'ELR sur la base de devis. Pour ce faire, nous allons appliquer sur cette base les relativités obtenues dans les modèles de fréquence et coût moyen en utilisant comme constante pour chaque équation tarifaire les primes pures de référence obtenues à l'étape précédente. Nous obtenons ainsi une estimation de la charge annuelle théorique de sinistres au niveau de chaque devis. On calcule ensuite l'ELR au niveau de chaque prospect en divisant la charge théorique de sinistres par la cotisation affichée au moment de la réalisation du devis. Nous obtenons ainsi l'indicateur (nommé ELR) qui permet d'évaluer la rentabilité théorique des prospects et qui sera utilisé lors de l'identification des mesures tarifaires à l'affaire nouvelle.

3. Identification des mesures tarifaires

La dernière étape de ce mémoire a pour objectif de répondre à deux questions :

- Quels sont les segments dont un ajustement du tarif est nécessaire (à la hausse ou à la baisse) ?
- De combien doit-on ajuster la prime commerciale soit pour redresser la compétitivité du produit sans dégrader les résultats techniques, soit pour améliorer les résultats techniques sans dégrader la compétitivité ?

Pour répondre à ces deux questions, nous allons mettre au point une stratégie d'ajustement de la prime commerciale en se basant sur une méthode graphique d'analyse conjointe des deux indicateurs calculés aux étapes précédentes (dans une approche de type rendement/risque) : le taux de transformation et l'ELR.

L'analyse sera réalisée uniquement sur les critères tarifaires dont le pouvoir explicatif dans le modèle de transformation des prospects est le plus important.

A travers cette méthode nous allons identifier les prospects qui ne transforment pas bien, afin de baisser leur prix dans la limite de leur prime pure espérée. Ainsi, sans dégrader les résultats techniques, nous allons augmenter l'apport net (améliorer le chiffre d'affaires) et gagner des parts de marché. Dans la même logique, nous allons trouver les prospects avec un très bon taux de transformation et qui ont de mauvais résultats techniques. En ajustant leur tarif à la hausse, nous allons conserver une charge commerciale globale suffisante en améliorant le mix des affaires nouvelles, sans trop détériorer l'apport net.

Après avoir identifié les ajustements nécessaires du tarif commercial, nous allons mesurer l'impact engendré par ces modifications sur la base des affaires nouvelles en termes de volume, de rentabilité estimée et de changement du mix.

Pour les deux premières étapes de notre approche, la démarche de modélisation utilisée se distingue des autres travaux de modélisation par l'objectif de coupler les modèles linéaires généralisés avec les méthodes d'apprentissage statistique ce qui présente l'avantage d'éviter l'aspect « boîte noire » de ces outils et de conserver la lisibilité du modèle. La lisibilité et l'interprétabilité des modèles assurées par la structure des GLM sont des éléments essentiels pour l'actuaire tarificateur. Celui-ci doit pouvoir échanger de façon transparente sur le contenu des modèles avec différents interlocuteurs comme sa hiérarchie, les agents généraux, les équipes marketing etc. ; ce souci de la transparence est aussi crucial vis-à-vis du régulateur, ainsi que pour l'audit des modèles. L'utilisation des outils de machine learning est justifiée pour leur vélocité, leur capacité à traiter des données plus volumineuses et leur performance accrue.

En conclusion, cette étude nous permet d'analyser l'appétence à la souscription ainsi que la rentabilité estimée des prospects deux-roues tout en bénéficiant des avantages liés aux techniques de machine learning. De cette manière nous allons d'abord construire des modèles robustes et facilement interprétables et ensuite nous allons analyser conjointement ces deux KPI. L'approche choisie permet, d'une part, d'identifier les segments pour lesquels une modification du tarif commercial est nécessaire et d'une autre part de trouver le juste niveau d'ajustement du tarif commercial actuellement en place, sans pour autant dégrader la rentabilité.

A RETENIR :

- Cette analyse s'inscrit dans le cadre de la mise à jour annuelle du tarif à l'affaire nouvelle du produit d'assurance moto d'AXA France.
- A travers cette étude nous souhaitons répondre à deux questions : quels sont les segments dont le tarif doit être ajusté et de combien doit-on ajuster le tarif pour qu'il soit plus compétitif sans pour autant dégrader la rentabilité (ou inversement).
- Pour répondre à ces questions nous proposons une méthode d'ajustement du tarif commercial prenant en compte d'une part l'attractivité du produit sur le marché (mesurée à travers le taux de transformation) ainsi que la rentabilité estimée des devis (mesurée à travers l'Expected Loss Ratio).
- L'approche utilisée se décompose en trois étapes : d'abord nous allons analyser et modéliser le comportement des prospects à la souscription. Ensuite nous allons estimer la rentabilité future des devis à travers l'ELR et, en dernière étape, nous allons confronter ces deux indicateurs dans une approche d'optimisation de type rendement-risque pour identifier les mesures tarifaires à appliquer.

Chapitre 2

Méthodologie de modélisation

2.1 L'apport du machine learning

L'approche proposée pour l'identification des mesures tarifaires à l'affaire nouvelle se base sur des méthodes de modélisation statistiques. Celles-ci interviennent lors des deux premières étapes de la démarche suivie. Elles sont d'abord utilisées pour analyser et modéliser le taux de transformation des interrogations tarifaires en affaire nouvelle. Ensuite, elles interviennent lors du calcul de l'ELR [1], dont l'estimation repose sur la modélisation des primes pures par garantie.

En pratique, plusieurs techniques statistiques sont mises à disposition des actuaires pour réaliser ces types de modélisation. Nous allons d'abord déterminer quelles sont les caractéristiques d'un modèle efficace. Ensuite nous allons présenter succinctement deux méthodes de construction de tels modèles : la méthode de modélisation dite "classique" et la méthode choisie dans ce mémoire, renforcée par des techniques de machine learning.

L'approche utilisée à travers ce mémoire fait intervenir ces techniques lors des deux premières étapes de construction des modèles : la **sélection** des variables ainsi que leur **simplification**. Deux méthodes seront mises en avant : **la régression pénalisée LASSO** ainsi que le **Gradient Boosting Model (GBM)**. Malgré leur vélocité et leur capacité à traiter des données volumineuses, ces techniques présentent une certaine opacité. Pour éviter donc l'aspect « boîte noire » de ces outils, la sortie du GBM est ensuite transformée en structure linéaire, permettant d'aboutir, in fine, à un modèle de type **GLM**, facile à interpréter et analyser.

2.1.1 Les caractéristiques d'un modèle efficace

Un modèle efficace est un compromis délicat entre simplicité et pouvoir prédictif.

Un modèle d'une très grande complexité, où on reprendrait quasi-exhaustivement les variables, leurs interactions et des découpages très fins, pourrait mal capturer le phénomène sous-jacent et devenir trop dépendant aux données d'entraînement et aux petites fluctuations aléatoires, non représentatives du phénomène. Il y aurait donc un nombre très important de profils possibles, très peu d'effectifs par profils, et donc les coefficients auraient un très grand écart-type, ce qui entraîne deux choses : d'une part les intervalles de confiance des vrais paramètres et de la vraie probabilité seraient très grands, d'autre part, une grande partie d'entre eux ne serait pas significativement différent de zéro. En outre, on risquerait d'arriver à une séparation complète ou quasi-complète des données, ce qui n'aurait, conjugué à des faibles effectifs par profils, aucun sens.

Il ne faut pas non-plus négliger la question des interactions entre les variables. Deux variables fortement corrélées présentes toutes les deux dans un modèle risquent d'entraîner une non-significativité des coefficients sur l'une des deux variables. Il vaut mieux alors simplifier le modèle en retenant une seule des deux, ou tenter de les résumer en une seule et unique variable.

1. Pour de plus amples détails concernant cet indicateur nous invitons le lecteur à se référer à la section 4.1.

A contrario, un modèle trop simple ne parviendrait pas à capturer toute la complexité du phénomène analysé. Il ne refléterait pas suffisamment la réalité et il prédirait mal.

En bref, le but est d'arriver à des modèles pas trop complexes, avec pas trop de modalités, en limitant les interactions, avec des coefficients de chaque modalité significativement différents de zéro et significativement différents entre eux.

2.1.2 La méthode classique de modélisation

Dans une approche classique de modélisation, la première étape à suivre est la sélection de variables ayant le pouvoir discriminant le plus élevé sur le phénomène analysé. Cette sélection est réalisée grâce à des algorithmes de sélection pas à pas (sélection ascendante, descendante ou mixte) qui permettent de trouver simultanément le meilleur modèle parcimonieux, en optimisant les critères de performance, tels que l'AIC ou le BIC. Ces critères, de manière générale, décroissent lors de l'ajout de variables pertinentes, mais augmentent lors de l'ajout de variables non pertinentes. Le but étant donc d'essayer de les minimiser.

Ces méthodes de sélection des variables sont très coûteuses en terme de temps de calcul, car celui-ci augmente de manière exponentielle avec le nombre de variables p, puisqu'on dispose de 2^{p-1} modèles. Pour réduire le temps de calcul, une pré-sélection des variables en amont est donc nécessaire. Celle-ci est réalisée, habituellement, par des techniques d'apprentissage statistique comme le GBM ou les forêts aléatoires. L'avantage de ces méthodes est qu'elles permettent de traiter un grand nombre de variables explicatives dans un délai de temps acceptable.

Une fois les variables explicatives sélectionnées, l'étape suivante est de construire le modèle final, un GLM prenant comme régresseurs la liste des variables avec la plus forte influence relative sur le phénomène analysé.

Ensuite, le modèle ainsi réalisé est simplifié, pour garder une certaine cohérence. Cette simplification est réalisée manuellement, grâce à des regroupements de modalités, afin de réduire le nombre de paramètres du modèle et d'accroître son pouvoir généralisant.

Une fois les regroupements réalisés, il est important d'analyser les corrélations entre les variables explicatives pour pouvoir mettre en place un modèle issu de la théorie des modèles linéaires généralisés. Le risque d'introduire des variables fortement corrélées dans le modèle final est une estimation biaisée de ses paramètres et une possible non-convergence du modèle. Si, après avoir réalisé cette analyse, on observe des variables fortement corrélées, nous avons le choix soit de garder dans le modèle finale une seule des deux variables corrélées (celle qui permet de diminuer le plus les critères de performance du modèle : AIC ou BIC), soit d'introduire dans le modèle final l'interaction de ces deux variables.

2.1.3 La modélisation basée sur les techniques de machine learning

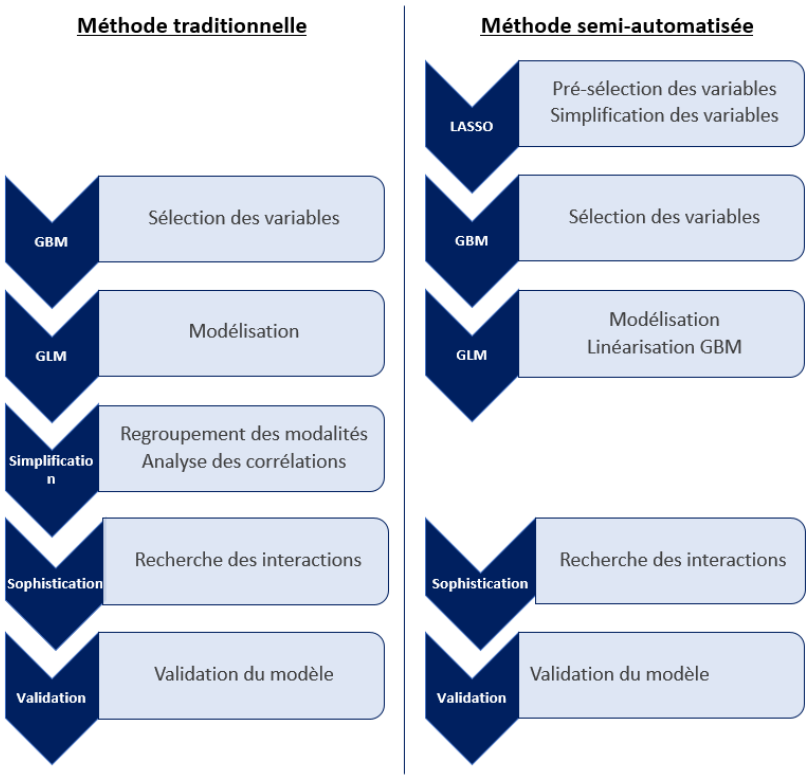
La construction des modèles bâtis lors de ce mémoire (que ce soit la modélisation du taux de transformation des devis ou les différents modèles de fréquence et de coût moyen des diverses garanties) a été réalisée à travers un outil de modélisation quasi-automatisée développé en interne par AXA Global P&C. Cette approche présente plusieurs avantages :

- un gain de temps considérable (étant donné le nombre conséquent des modèles nécessaires à la réalisation de cette étude)
- des performances comparables ou même supérieures aux modèles de tarification traditionnels (le logiciel a été testé en interne par les équipes de tarification non-vie conjointement avec les équipes de Data Science de AXA Global P&C les trois dernières années)
- les modèles peuvent être facilement réutilisables et mis à jour par les autres actuaires de l'équipe

En dépit de ces avantages, l'actuaire doit néanmoins porter un regard critique sur le modèle obtenu et contraindre le nombre de variables qu'il contient, car sa parcimonie et son interprétabilité sont décevantes sans son intervention. L'expertise métier et l'opinion de l'actuaire

restent donc cruciales afin de répondre aux critères de transparence, robustesse et applicabilité opérationnelle du modèle construit.

Pour mieux comprendre le processus de modélisation utilisé dans ce mémoire, la figure ci-dessous présente schématiquement les étapes de modélisation de la méthode traditionnelle comparées avec la méthode semi-automatisée [2] :



Graphique 2.1 – Schéma méthodes de modélisation.

Pour synthétiser l'approche de modélisation à la base de cet outil, plusieurs étapes s'enchaînent : d'abord une pré-sélection et un regroupement automatique des variables catégorielles à travers la **régression pénalisée LASSO** suivie par une sélection des variables avec un **GBM**. Ensuite, une fois les variables simplifiées et sélectionnées, le modèle final est construit, en transformant directement le **GBM en structure linéaire** afin de garder la lisibilité et l'interpretabilité du modèle construit.

L'outil permet également un calcul automatisé des coefficients de corrélation entre les variables explicatives, pour que les actuaires puissent exclure de la modélisation les variables fortement corrélées.

Comme dans la méthode traditionnelle, une des dernières étapes de la modélisation est la recherche des croisements entre les variables explicatives (les interactions) à intégrer aux modèles. L'identification ainsi que la mesure de l'importance relative de chaque croisement sur le phénomène analysé sont réalisées de manière automatisée par l'outil de modélisation. L'intégration des interactions dans le modèle final est, quant à elle, laissé à la main de l'utilisateur (les modèles initiaux construits avec la méthode innovante ne contiennent des interactions que si elles ont été définies en amont de la modélisation).

De plus, cet outil présente l'avantage de pouvoir bâtir et intégrer dans le modèle ainsi construit des nouveaux zoniers (à la maille géographique souhaitée).

Cette approche de modélisation, dite "semi-automatisée", repose également sur l'intervention de l'actuaire à plusieurs niveaux :

2. La présentation comparative entre la méthode classique de tarification et la méthode semi-automatisée fait l'objet d'un autre mémoire d'actuariat, les deux techniques étaient appliquées lors de la refonte tarifaire du produit d'assurances multirisque professionnelle d'Axa France (source : Sultan, 2018)

1. Sélectionner le meilleur modèle (parmi l'ensemble des modèles construits avec le logiciel) en termes de parcimonie, robustesse et interprétabilité

2. S'assurer de la validité des regroupements proposés et faire des ajustements s'ils avèrent nécessaires

3. S'assurer de l'exclusion du modèle final choisi des variables fortement corrélées ou des variables a posteriori

4. Sélectionner et inclure des croisements des variables si ceux-ci permettent d'augmenter les performances du modèle de manière significative

5. Sélectionner le meilleur zonier (parmi ceux construits par le logiciel) dans le cas d'une refonte intégrale de celui-ci.

En conclusion, l'approche de modélisation utilisée dans ce mémoire permet de tirer profit des outils d'apprentissage statistique pour la construction de modèles tarifaires techniques qui soient à la fois robustes et interprétables.

A RETENIR :

- L'approche proposée pour l'identification des mesures tarifaires repose d'une part sur la compréhension et la modélisation du taux de transformation des devis et d'autre part sur la modélisation des primes pures par garantie (pour le calcul de l'ELR).
- Les différents modèles bâtis lors de ce mémoire sont réalisés dans une nouvelle approche pour le périmètre moto en se basant sur les techniques de machine learning : notamment la régression pénalisée LASSO et le GBM.
- Cette nouvelle approche présente divers avantages : un gain de temps considérable, des performances optimisés par rapport aux techniques de modélisation classiques, des modèles faciles à interpréter et à mettre à jour.

2.2 Modélisation avec des méthodes de linéarisation du gradient boosting

Après avoir présenté les avantages de la nouvelle approche de modélisation intégrant des techniques des machine learning, nous allons maintenant rentrer plus dans les détails dans la description de la méthodologie utilisée et des fondements statistiques sous-jacents.

2.2.1 Préliminaire : outils de validation et mesures de performance des modèles

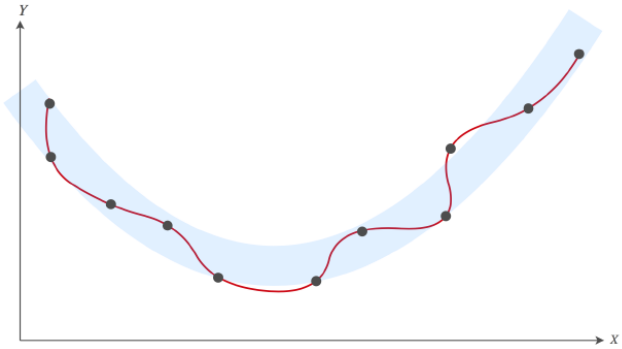
L'apprentissage statistique désigne un ensemble de techniques d'analyse prédictive réalisées à travers un algorithme d'apprentissage itératif.

Le phénomène de sur-apprentissage et le dilemme biais-variance

Contrairement à la statistique classique qui exige de se fixer des hypothèses concernant la structure et la distribution des données analysées, les méthodes d'apprentissage automatique se basent sur une seule hypothèse : la variable à expliquer, que nous noterons Y, est générée de façon i.i.d par un processus P à partir du vecteur des variables explicatives X. Ainsi, l'objectif est de déterminer la fonction $f_\lambda(x)$, fonction qui dépend des variables d'entrée X et du paramètre de complexité λ (qui désigne par exemple le nombre d'arbres dans un modèle GBM ou le nombre de nœuds dans un arbre de décision).

Pour déterminer cette fonction, nous allons construire un modèle à partir d'un échantillon de données appelé échantillon d'apprentissage. Pour évaluer sa qualité de prédiction, il ne suffit pas de comparer la prédiction des individus de l'échantillon d'apprentissage avec la réalisation de la variable réponse. En effet, l'objectif est de créer un modèle généralisable, c'est à dire un modèle qui peut être appliqué sur une base de données qui n'a pas été utilisée à la construction du modèle pour éviter le phénomène de sur-apprentissage.

Le but d'un algorithme d'apprentissage est donc de capter la tendance observée sur un échantillon entre la variable réponse et ses covariables, et de la généraliser sur des individus qui ne sont pas présents dans la base d'apprentissage. Cependant, l'algorithme peut également modéliser du bruit propre à l'échantillon sur lequel le modèle a été construit.



Graphique 2.2 – Exemple de sur-apprentissage.

Dans cet exemple simple, la tendance observée est tracée en bleu, tandis que la courbe rouge correspond à un modèle qui a sur-appris. Même si le modèle représenté par la courbe rouge donne des meilleurs résultats sur la base d'apprentissage, il n'est sûrement pas généralisable sur une autre base de données. Pour éviter donc ce phénomène, il faudrait utiliser le modèle tracé en bleu, qui permet d'être utilisé sur des observations qui ne sont pas présentes dans le jeu de données initial.

Dans un modèle de régression, l'estimation d'un modèle repose sur la minimisation de l'erreur quadratique moyenne (MSE). La fonction à modéliser, que nous allons noter y, peut s'écrire comme

$$y = f(x) + \varepsilon, \tag{2.1}$$

où ε est un bruit d'espérance nulle et de variance σ.

Lors de la création d'un modèle de régression, nous allons créer un estimateur $\hat{f}(.)$ de telle sorte que pour chaque point de l'échantillon y_i, son estimation soit $\hat{f}(x_i)$.

Mathématiquement, on peut donc écrire :

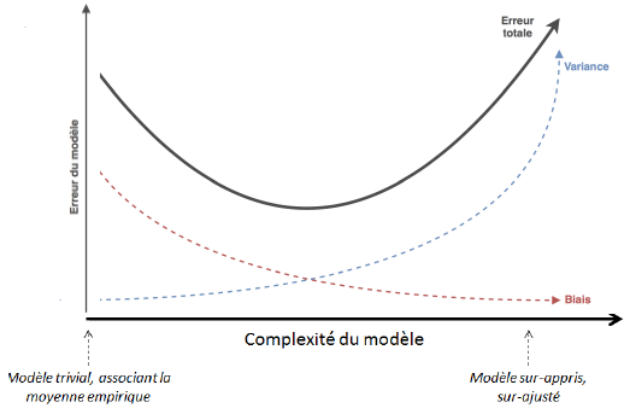
$$\begin{aligned} MSE &= E[(y - \hat{f}(x))^2] = E[y^2 + \hat{f}^2(x) - 2y\hat{f}(x)] \\ &= Var[y] + Var[\hat{f}(x)] + E[f(x) - \hat{f}(x)]^2 \\ &= Var[\hat{f}(x)] + Biais^2[\hat{f}(x)] + Var[y] \\ &= Variance + Biais^2 + \text{Erreur irréductible.} \end{aligned} \tag{2.2}$$

La variance correspond à la sensibilité du modèle à la suite des fluctuations de l'échantillon d'apprentissage. Un modèle avec une grande variance est un modèle complexe, qui modélise du bruit. On est alors dans le cas du sur-apprentissage.

Un modèle avec un biais important est un modèle trop simple qui ne capte pas correctement les relations entre la variable à expliquer et les covariables. On est alors dans le cas du sous-apprentissage.

L'objectif est donc de systématiquement trouver le bon compromis entre le biais et la variance du modèle. Ce phénomène, connu sous le nom de dilemme biais-variance, est illustré dans la

figure 2.3.



Graphique 2.3 – Dilemme biais-variance. Source : Dreyfus, 2008

Si le modèle est très simple, alors la variance sera petite mais le biais sera grand. C'est le cas contraire si le modèle est trop paramétré. Le meilleur modèle est donc situé au croisement entre ces deux courbes et il permet de minimiser l'erreur totale.

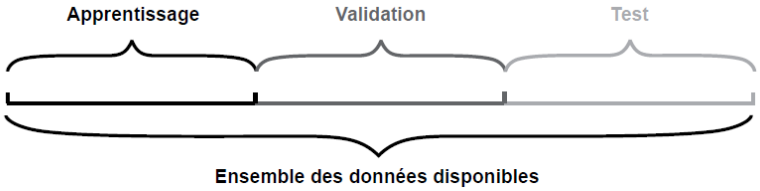
La validation croisée

Comme indiqué plus haut, la méthode de minimisation du risque (mesuré à travers l'erreur quadratique moyenne) conduit à concevoir des modèles de complexités différentes et à choisir celui qui est susceptible d'avoir les meilleures propriétés de généralisation.

Nous avons vu qu'il est impossible, en général, d'estimer la capacité de généralisation d'un modèle à partir des résultats de l'apprentissage ; une telle procédure conduirait systématiquement à sélectionner un modèle de biais faible et de variance élevée, donc sur-ajusté. Pour sélectionner le meilleur modèle parmi des modèles de complexités différentes, il convient donc de les comparer sur la base des prédictions qu'ils effectuent sur des données qui n'ont pas servi à l'apprentissage.

La première étape à suivre, donc, une fois la base de données construite, est le découpage de celle-ci en trois échantillons :

- l'échantillon d'apprentissage : à partir duquel on va construire le modèle
- l'échantillon de validation : qui permet de calibrer et de tester la performance du modèle construit sur la base d'apprentissage
- l'échantillon de test : qui permet d'évaluer la performance du modèle sélectionné

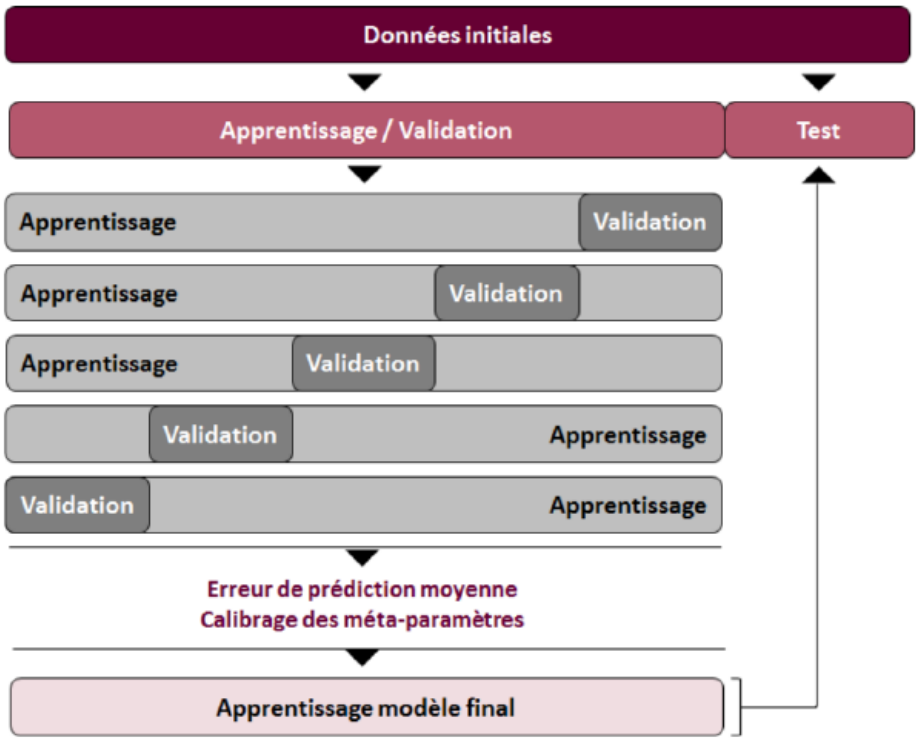


Graphique 2.4 – Validation simple.

Concernant les proportions à garder dans chaque échantillon il n'existe pas de règle précise dans la littérature statistique. L'idée principale est d'avoir le maximum de données dans la base d'apprentissage tout en en gardant suffisamment pour les bases de validation et de test. L'inconvénient de ce découpage est donc la perte de données pour l'apprentissage. Néanmoins, il existe des techniques pour découper une base en limitant la perte de données, la plus connue étant la technique de validation croisée, utilisée dans ce mémoire.

Le principe de la validation croisée (ou k-Fold Cross Validation) consiste d'abord à diviser la base initiale en 2 parties : la base d'apprentissage/validation et la base de test (en proportion de

80-20 par exemple). Ensuite la base d'apprentissage/validation est découpée en k échantillons de même taille. Un modèle est construit en utilisant tous les échantillons sauf le dernier qui sert d'échantillon de validation. La procédure est répétée en utilisant l'avant-dernier échantillon comme échantillon de validation et le reste comme échantillon d'apprentissage, etc. Les modèles ainsi construits disposent chacun d'un score de performance. La moyenne des scores permet d'obtenir un score de performance moyen. Le modèle déterminé en regroupant les résultats obtenus sur l'ensemble des k-folds (en moyennisant les valeurs obtenues) est appliqué ensuite sur la base de test. Cette technique est illustrée dans la figure 2.5 :



Graphique 2.5 – Validation croisée par 4-fold. Source : Guillot, 2015

La validation croisée est levier utilisé dans les techniques de machine learning pour optimiser la robustesse des modèles et éviter le sur-apprentissage. Dans le cadre de ce mémoire, cette technique sera employée pour le calibrage du paramètre de pénalisation λ du LASSO (cf. partie 2.2.2.3) et pour valider la sélection des variables issues du gradient boosting (cf. partie 2.2.3).

La courbe de lift

En plus de la validation croisée, un autre outil de validation des modèles utilisés dans ce mémoire est la courbe de lift. Le principe est simple : après avoir trié la base de données par ordre croissant des valeurs prédites, elle est divisée en k segments d'exposition égale. Ensuite, nous allons calculer pour chaque segment la prédiction moyenne et l'observation moyenne et tracer graphiquement la courbe des valeurs prédites et observées. Pour s'assurer que le modèle s'ajuste bien aux données, les deux courbes doivent présenter globalement les mêmes tendances et ne pas être trop éloignées l'une de l'autre.

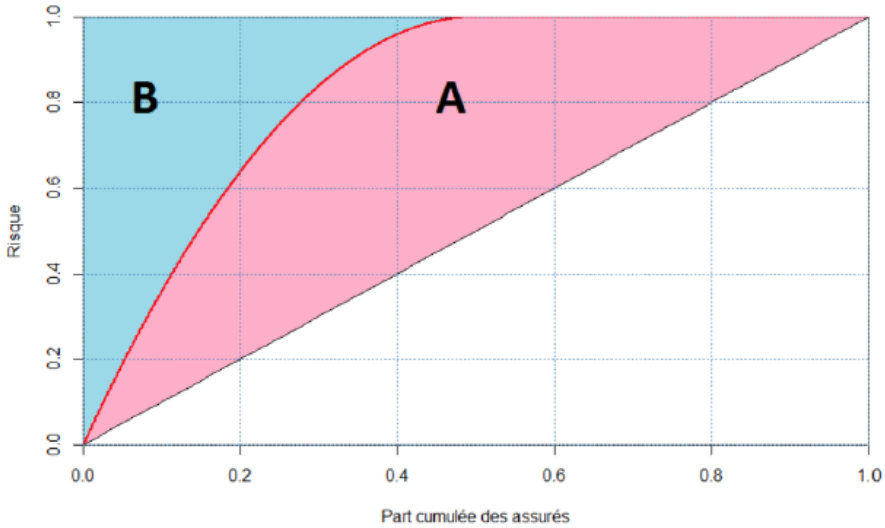
Mesure de la performance des modèles

Pour évaluer la performance d'un modèle, il est important d'analyser à la fois une mesure de précision et une mesure du pouvoir discriminant car l'analyse d'un seul indicateur ne suffit pas. Ici, on utilise principalement deux indicateurs pour évaluer la performance des modèles bâtis : l'indice de Gini et l'EDR .

L'indice de Gini

L'indice de Gini est un indicateur mesurant la capacité de segmentation et de tri d'un modèle. Il est calculé à partir de la fonction représentée par la courbe de Lorenz. Aussi appelée courbe de gain, cette courbe, développée par Lorenz (source : Lorenz, 1905) en 1905, permettait à la base de représenter graphiquement les inégalités de revenus au sein d'une population. On lui associe la fonction qui calcule la part des revenus par rapport à la part des détenteurs.

Si le phénomène analysé est la fréquence des sinistres, la courbe de gain représente en abscisse la part cumulée des expositions, et en ordonnée, la part cumulée du nombre de sinistres. L'indice de Gini permet de comparer deux modèles. Il est calculé à partir de l'aire entre la bissectrice et la courbe de gain notée A, et correspond à deux fois l'aire A de la figure 2.6.



Graphique 2.6 – Calcul de l'indice de Gini comme étant deux fois l'aire A entre la courbe de gain et la bissectrice. Source : Pariente, 2016.

La bissectrice représentée dans la figure 2.6 correspond à un modèle aléatoire ou à une égalité parfaite lorsqu'on observe une fréquence des sinistres identique pour l'ensemble des assurés : "x % des assurés ont x % du nombre de sinistres total". Si l'aire entre la courbe de gain et la bissectrice vaut A, et que celle au-dessus de la courbe de Lorenz vaut B alors l'indice de Gini est défini comme :

$$G = \frac{A}{A+B}. \tag{2.3}$$

Puisque A + B = 0,5, l'indice de Gini est égal à G = 2A ou encore G = 1 - 2B. Si l'on suppose que la courbe de gain est représentée par la fonction y = L(x), alors la valeur de l'aire B peut s'exprimer avec l'intégrale :

$$B = 1 - \int_0^1 L(x)\,dx. \tag{2.4}$$

L'indice de Gini peut donc s'écrire :

$$G = 2\int_0^1 L(x)\,dx - 1. \tag{2.5}$$

D'un point de vue pratique on cherche à maximiser l'indice de Gini, c'est a dire avoir une courbe de Lorenz la plus éloignée possible de la bissectrice. C'est un indicateur que nous allons utiliser pour tester l'ajout des nouvelles variables dans un modèle.

L'EDR , indicateur du pouvoir de prédiction

L'expected deviance ratio (EDR) est utilisé pour mesurer la performance d'un modèle. Il permet de déterminer la part du signal capturé par le modèle et il est calculé comme rapport entre les écarts de vraisemblances par rapport au modèle sans variables explicatives :

$$EDR = \frac{LLH(\text{modèle analysé}) - LLH(\text{modèle sans variables explicatives})}{LLH(\text{modèle saturé}) - LLH(\text{modèle sans variables explicatives})}. \tag{2.6}$$

Sous la dénomination de modèle saturé nous entendons un modèle de régression tel que le nombre des paramètres à estimer est égal au nombre total d'observations. Par conséquent, un tel modèle peut être considéré comme fournissant une description complète des données et peut ainsi servir de référence.

l'EDR est un bon indicateur pour comparer différents modèles : entre deux modèles avec un indicateur de Gini équivalent, nous allons choisir le modèle avec l'EDR le plus élevé, car son pouvoir prédictif est plus proche de celui du modèle saturé.

En conclusion, l'indice de Gini est un indicateur qui reflète "le bon classement" des risques. Il ne peut pas être utilisé seul car il ne tient pas compte des écarts absolus par rapport au modèle saturé ou par rapport au modèle sans variables explicatives (seul l'ordre compte) : il constitue une mesure du pouvoir discriminant du modèle. Il est donc conjointement utilisé avec l'EDR (expected deviance ratio).

2.2.2 Présélection et simplification des variables catégorielles avec LASSO

Lors de la construction d'un modèle linéaire généralisé, la réduction du nombre des coefficients non-nuls du prédicteur linéaire est importante à plusieurs égards.

L'ajout d'une contrainte sur les coefficients accélère la convergence de l'algorithme de maximisation de la vraisemblance et permet un gain du temps de calcul considérable. Limiter le nombre des coefficients à estimer permet de diminuer la variance des valeurs prédites et ainsi augmenter le pouvoir prédictif du modèle. Un autre avantage est une meilleure interprétabilité du modèle : en général, lors de la construction d'un modèle statistique nous cherchons a identifier un petit nombre de variables ayant un fort pouvoir prédictif.

En pratique, dans le secteur de l'actuariat IARD, les méthodes de sélection des variables les plus utilisées sont les méthodes pas à pas (forward, backward et stepwise). Ces méthodes permettent de trouver simultanément le meilleur modèle parcimonieux, en optimisant les critères de performance, tels que l'AIC ou le BIC.

Malgré leur utilisation régulière, ces méthodes présentent l'énorme inconvénient d'être très coûteuses en termes de temps de calcul, car celui-ci augmente de manière exponentielle avec le nombre de variables. De plus, Frank Harrell, dans son livre "Regression Modeling Strategies" (source : Harrell, 2001), met en évidence plusieurs défauts dans cette méthode :

- Le coefficient de détermination R^2 est biaisé à la hausse.
- Les hypothèses de distribution sur lesquels s'appuient les tests de significativité (test de Fisher et du Chi-deux) entre deux itérations ne sont pas valides.
- Les intervalles de confiance obtenus pour les coefficients et les valeurs prédites sont biaisés à la baisse.

- Les problèmes de colinéarité sont exacerbés.

Pour pallier ces différents biais, la méthode de régularisation LASSO a été privilégiée lors de la construction de l'outil de modélisation.

2.2.2.1 Méthodes de régularisation

Développée par Robert Tibshirani en 1996, le LASSO (Least Absolute Shrinkage and Selection Operator, source : Tibshirani, 1996) est la plus célèbre des méthodes de pénalisation ou régularisation. Les méthodes de régularisation consistent à imposer une contrainte à la complexité du modèle afin de favoriser la parcimonie. L'objectif principal de ces méthodes est d'améliorer la robustesse du modèle (en limitant le sur-apprentissage et en contrôlant l'erreur de type variance) par rapport aux données de la base d'apprentissage. Cette contrainte se traduit par l'ajout d'une norme sur les paramètres à estimer, norme intégrée à la fonction objectif à minimiser.

La norme L1 correspond à la régression LASSO, tandis que la norme L2 correspond à la régression ridge (pénalisation de Tikhonov). Ce terme supplémentaire va donc contraindre le programme d'optimisation à favoriser le choix des coefficients plus faibles, limitant également la volatilité des coefficients les plus élevés et contribuant donc à la robustesse du modèle. Il s'agit donc d'augmenter volontairement le biais du modèle pour favoriser la réduction de la variance.

La description du cadre général des méthodes de régularisation requiert de présenter la méthode Elastic Net, développée par Zou et Hastie en 2005 (source : Zou, H. & Hastie, T., 2005). Celle-ci permet de combiner les deux méthodes de régularisation précédentes, en introduisant les deux normes simultanément. L'estimateur associé dans le cas d'une régression aux moindres carrés est alors défini par la minimisation de la fonction de coût R suivante :

$$\hat{\beta} = \underset{\beta}{argmin}\left[\|Y - X^t\beta\|_{L_2}^2 + \lambda\left((1-\alpha)\frac{1}{2}\|\beta\|_{L_2}^2 + \alpha\|\beta\|_{L_1}\right)\right], \tag{2.7}$$

où le paramètre α, compris entre 0 et 1, permet de définir l'équilibre entre les deux pénalités et λ est le coefficient de pénalisation. Pour $\alpha = 1$ la fonction de coût correspond à la régression LASSO, pour $\alpha = 0$, à la régression ridge et enfin pour $\lambda = 0$ à celle de la régression linéaire classique. Le coefficient de pénalisation λ est très important puisqu'il définit le poids attribué à la pénalité. Plus il est élevé, plus la pénalisation est forte et plus les coefficients estimés sont proches de zéro. On doit donc accorder une importance particulière au calibrage de ce méta-paramètre : il est généralement effectué à l'aide d'une technique de validation croisée, dont l'application au LASSO est décrite plus loin. De même, on pourrait calibrer le paramètre α, représentant l'équilibre entre les deux types de pénalités, par une procédure similaire. Cette procédure est complexe et elle s'avère généralement excessive. Il suffit donc de se restreindre à l'une des deux méthodes élémentaires la composant.

En grande dimension, lorsqu'il y a un risque élevé que les predicteurs soient fortement corrélés entre eux, les deux paramètres de pénalisation présentent des effets relativement opposés par rapport à la colinéarité. La régression ridge a tendance à réduire simultanément les estimations des coefficients des variables corrélées, tandis que le LASSO est connu pour privilégier une des variables et d'écarter les autres, en étant plutôt indifférent aux phénomènes de colinéarité. Même si la méthode elastic net offre un bon compromis entre ces deux approches, étant donné le grand nombre de variables potentiellement explicatives que les actuaires doivent tester lors de la construction des modèles de statistiques, une vraie technique de réduction de la dimension a été favorisée lors de la construction de l'outil. Seule la méthode LASSO permet d'écarter véritablement une partie des facteurs de risques utilisés.

2.2.2.2 Résolution numérique

Concrètement, pour pouvoir implémenter cette méthode de régularisation, on doit estimer les deux paramètres du modèle. On doit résoudre donc le programme de minimisation décrit par l'équation 2.7. Cette équation n'admet pas de solution analytique par formule fermée comme en régression linaire classique, en raison de la présence de la norme L1, non inversible [3].

La résolution numérique est est donc réalisée à travers un algorithme de descente de gradient. Ainsi, le programme d'optimisation se décline selon chaque direction β_j, par l'équation impliquant la dérivée partielle de la fonction de coût R :

$$\frac{\partial R}{\partial \beta_j}|_{\beta=\hat{\beta}} = -\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - x_i^t\hat{\beta}) + \lambda(1-\alpha)\beta_j + \lambda\alpha = 0, \tag{2.8}$$

où le paramètre $\hat{\beta}$ est l'éstimateur de l'itération précedente, fixé dans la dérivation précédente pour toutes les coordonnées $k \neq j$.

Dans le cas d'une régression GLM, le terme de pénalité s'ajoute à la log-vraisemblance du modèle. Pour la modélisation de la fréquence, par exemple, le programme de maximisation de la vraisemblance devient alors :

$$\begin{aligned} &\max_{\beta\in\mathbb{R}^p}\left[\sum_{i=1}^{n}\{y_i ln(\lambda_i) - ln(y_i!) - \lambda_i\} - \lambda\left\{(1-\alpha)\frac{1}{2}\|\beta\|_{L_2}^2 + \alpha\|\beta\|_{L_1}\right\}\right] \\ &\Leftrightarrow \max_{\beta\in\mathbb{R}^p}\left[\sum_{i=1}^{n}\left\{y_i x_i^t\beta - ln(y_i!) - e^{x_i^t\beta}\right\} - \lambda\left\{(1-\alpha)\frac{1}{2}\|\beta\|_{L_2}^2 + \alpha\|\beta\|_{L_1}\right\}\right]. \end{aligned} \tag{2.9}$$

Dans ce cas, la méthode de résolution numérique de ce problème d'optimisation correspond à une variante de l'algorithme de Newton-Raphson, homologue de la descente de gradient pour les vraisemblances. Pour une explication plus détaillée de cet algorithme, le lecteur intéressé pourra se référer aux articles de Friedman et collab. (source : Friedman, J., Hastie, T. & Tibshirani, R., 2010).

Il est important de constater que la vraisemblance pénalisée et la vraisemblance maximisée dans le cadre d'un GLM classique ne sont pas comparables. Les coefficients estimés avec la méthode pénalisée sont de facto biaisés, car nous avons volontairement introduit dans la fonction de perte une source de biais. De même, les intervalles de confiance ainsi que les erreurs standards usuellement calculés dans un modèle linéaire généralisé n'ont plus aucun sens dans cette situation. La condition d'interprétabilité de ces indicateurs est l'hypothèse de distribution asymptotiquement gaussienne des estimateurs. Cette hypothèse n'est pas respectée dans le modèle pénalisé. Ainsi, si l'on souhaite comparer les résultats entre les modèles classiques et les modèles pénalisées on doit faire preuve d'une grande précaution, en gardant bien à l'esprit les divergences de signification précédentes.

2.2.2.3 Calibrage du méta-paramètre

Comme vu dans la partie précédente, le choix du méta-paramètre λ, qui représente le poids de la pénalisation, est crucial. Son estimation doit être la plus robuste possible.

Par la suite, on propose de détailler comment ce choix est réalisé. La valeur optimale de λ est généralement déterminée par validation croisée (conforme sous-section 2.2.1).

Dans cette partie de l'analyse, l'approche suivie examine une sélection de cinquante valeurs candidates de λ, comparées par la méthode de validation croisée de type k-fold, à l'aide de la déviance comme mesure d'erreur. Le choix de ces cinquante valeurs candidates est déterminé par l'algorithme de FRIEDMAN et collab. (source : Friedman, J., Hastie, T. & Tibshirani, R., 2010), qui est computationnellement efficient.

3. Contrairement à la régression ridge, qui admet une solution analytique, grâce au caractère quadratique de la norme L2.

Concrètement, l'algorithme propose de partir d'une valeur initiale λ_{max} qui est la plus petite valeur pour laquelle tous les coefficients estimés sont nuls. Ensuite l'algorithme détermine une valeur finale λ_{min} =ϵ x λ_{max} (typiquement $\epsilon = 10^{-3}$), puis construit une suite de valeurs décroissantes entres ces deux bornes sur l'échelle logarithmique :

$$\lambda_{max} > \lambda_2 > \lambda_3 > ... > \lambda_{199} > \lambda_{min}. \tag{2.10}$$

Pour chacune de ces valeurs on estime ensuite l'erreur de validation croisée CV(λ), calculée comme la moyenne des déviances dev_i sur les k partitions du k-fold, soit :

$$CV(\lambda) = \frac{1}{k}\sum_{i=1}^{k} dev_i(\lambda). \tag{2.11}$$

L'algorithme propose deux alternatives d'estimation pour le méta-paramètre :

- $\lambda = \lambda_{Dev}^{min}$: on retient donc le paramètre pour lequel l'erreur de validation croisée est minimale,

- $\lambda = \lambda_{Dev}^{lse}$: on retient comme valeur proposée pour le méta-paramètre la plus grande valeur de λ telle que son erreur type est inférieure à l'erreur type de la déviance pour $\lambda = \lambda_{Dev}^{min}$.

Mathématiquement, ces deux expressions sont définies comme suit :

$$\begin{aligned} \lambda_{Dev}^{min} &= \text{argmin } CV(\lambda) \\ \lambda_{Dev}^{lse} &= \max(\lambda) \text{ tel que } CV(\lambda) \leq CV(\lambda_{Dev}^{min}) + \sigma(\lambda_{Dev}^{min}). \end{aligned} \tag{2.12}$$

La première approche, même en étant la plus utilisée pour les modèles de prédiction, a tendance a sur-apprendre et peut amener a choisir un modèle trop complexe. C'est donc λ_{Dev}^{lse} qui est retenu dans cette approche de modélisation car il permet d'obtenir un modèle plus parcimonieux car il permet d'obtenir un modèle plus simple (pénalisé d'avantage) et qui, en même temps, ne peut pas être distingué du meilleur modèle en termes d'erreur (étant donné l'incertitude d'estimation de cette erreur).

2.2.3 Sélection des variables avec la méthode du gradient boosting

Dans la partie précédente nous avons détaillé la méthode LASSO utilisée pour réaliser à la fois la sélection préalable des variables catégorielles ainsi que leur simplification. Mais, dans la base de données utilisée pour la modélisation, nous avons à disposition non seulement des variables catégorielles, mais également des variables numériques. Une deuxième sélection des variables en amont de la modélisation est donc nécessaire.

Dans cette partie du mémoire, nous allons exposer la méthode de sélection des variables utilisée dans notre nouvelle approche : le gradient boosting. D'abord nous allons exposer les principes théoriques à la base de cette méthode et ensuite nous allons présenter la manière dont sa mise en place est articulée lors de la construction d'un modèle statistique.

Préliminaire : principe des arbres de décision et théorie du gradient boosting

Introduit par Friedman (source : Friedman, 2001) en 2001, l'algorithme du **gradient boosting** est une méthode ensembliste d'apprentissage automatique, utilisée pour des problèmes de régression ou de classification.

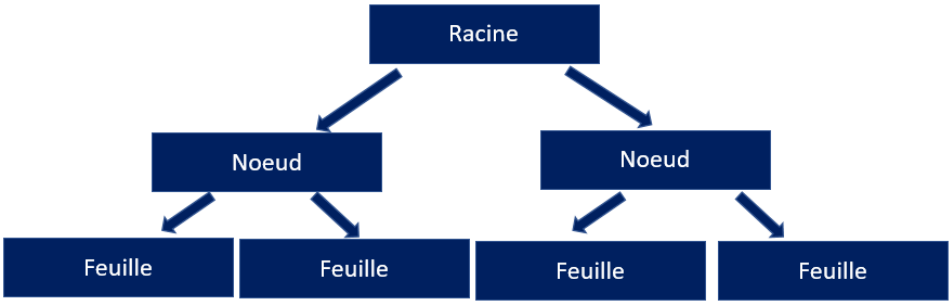
En reprenant le principe général du boosting, il optimise les performances d'un ensemble de modèles de prédiction dits "faibles", en les assemblant dans un modèle final. Concrètement, il consiste à réaliser une succession d'arbres de décision (de type CART dans notre cas) où chaque modèle est construit sur l'erreur résiduelle de l'itération précédente.

Le but d'un arbre de décision est d'expliquer une valeur à partir d'une série de variables discrètes ou continues. L'idée principale est de construire une partition de l'ensemble des variables explicatives de manière itérative, en tenant compte de la capacité prédictive des variables explicatives et en essayant à chaque étape d'obtenir des parties les plus homogènes possible. Cette hiérarchie des variables explicatives permet de visualiser les résultats dans un arbre et d'identifier des règles explicatives explicites, orientées métier. Chaque division est obtenue en sélectionnant la variable qui produira la meilleure séparation des individus (elle maximisera le critère de séparation). Cette division définit des sous-populations, représentées par les nœuds de l'arbre. L'opération est répétée pour chaque sous-population, jusqu'à ce que plus aucune séparation ne soit possible ou qu'un critère d'arrêt soit rencontré. On obtient alors des nœuds terminaux, appelés feuilles de l'arbre. Chaque feuille est caractérisée par un chemin spécifique à travers l'arbre qu'on appelle une règle. L'ensemble des règles pour toutes les feuilles constitue le modèle. Puis, une fois le modèle bâti, chaque individu se voit attribuér la moyenne de la feuille (ou classe) à laquelle il appartient.

La construction d'un arbre de décision peut donc se décomposer en trois étapes :

- **Initialisation** : la partition contient une seule prédiction, constituant le nœud initial (également appelé racine).
- **Expansion** : on agrandit l'arbre jusqu'à ce qu'un ou plusieurs critères d'arrêt soit satisfaits, comme par exemple :
  – Le nombre d'éléments dans chaque feuille doit être supérieur à une valeur prédéfinie.
  – Le nombre total de feuilles ne doit pas dépasser une valeur donnée.
- **Elagage** : les branches de l'arbre qui n'améliorent pas la qualité de prédiction sont supprimées.

La construction d'un arbre de décision est représenté schématiquement dans la figure 2.7 :



Graphique 2.7 – Schéma simplifié d'un arbre de décision.

Il existe plusieurs types d'arbres de décision. Dans cette étude, nous nous concentrons sur l'algorithme CART (Classification And Regression Tree), favori pour le gradient boosting. Cet algorithme présente de très bonnes performances en classement et ne nécessite pas de paramétrage. Pour de plus amples détails concernant la méthode d'apprentissage automatique CART, le lecteur peut se référer à l'ouvrage de Stéphane Tufféry, *Data mining and statistics for decision making* (source : Tufféry, 2005). La principale qualité des arbres de décision est qu'ils sont très facilement lisibles et interprétables en un ensemble de règles simples. En revanche, les arbres CART sont en général très instables. Par conséquent, les prédictions calculées varient beaucoup.

Pour pallier ces inconvenants des arbres de décision et afin de gagner en efficacité et en précision, le partitionnement est parfois répété un grand nombre de fois. La solution générale s'écrit alors comme la combinaison des réponses de chacun. Ces approches sont appelées « méthode d'agrégation » ou « méthodes ensemblistes ». L'idée essentielle des méthodes d'agrégation est donc d'offrir une meilleure stabilité et fiabilité des résultats, en palliant le sur-apprentissage,

principal défaut des arbres de décision. Le fait de construire un grand nombre d'arbres permet de réduire considérablement la variance de l'estimation et d'obtenir des estimateurs plus consistants.

Il y a deux grands types des « méthodes ensemblistes » : les méthodes d'arbres générés en parallèle (Random Forest) et les méthodes d'arbres générés de façon itérative (gradient boosting machine (GBM) utilisés à travers cette étude. Le principe général du GBM consiste à construire un ensemble d'arbres de manière récursive. Chaque arbre est une version adaptative du précédent en donnant plus de poids aux observations mal ajustées ou mal prédites. A l'issue de chaque arbre, on obtient une prédiction et on peut alors calculer les résidus de l'estimation. L'arbre suivant tient alors compte de ces résidus, sous la forme du gradient d'une fonction de perte de ces derniers. Chaque arbre apprend donc des erreurs du précédent, et on retient les estimations du dernier arbre.

Le principal défaut de ce type de méthodes ensemblistes basées sur les arbres est leur opacité ou, en tout cas, leur interprétabilité assez limitée. Friedman (source : Friedman, 2001) a développé une extension de l'influence relative pour les arbres boostés. Pour les méthodes arborescentes simples, l'influence relative d'une variable x_i est approximée par

$$\hat{J}_i^2 = \sum_{\text{splits sur } x_i} I_t^2, \tag{2.13}$$

où ${I_t}^2$ est l'amélioration empirique (égale à la diminution de l'erreur au carré) obtenue en divisant sur x_i sur ce point. L'extension de Friedman aux modèles GBM consiste à moyenner l'influence relative de la variable x_i à travers tous les arbres générés par l'algorithme de boosting. Ainsi, le métier peut avoir un début d'interprétation du modèle GBM en prenant connaissance des variables ordonnées de la plus influente à la moins influente.

Dans le but de limiter le sur-apprentissage et d'accroître le pouvoir généralisant du modèle, le gradient boosting machine possède plusieurs paramètres qui permettent d'optimiser son utilisation :

- le shrinkage ou le rétrécissement : il pondère l'apport de chaque nouvel arbre lors de son ajout à la prédiction.
- la fonction de perte, choisie en fonction du type de données à expliquer.
- le nombre d'arbres réalisés, qui correspond au nombre d'itérations effectuées par l'algorithme
- le nombre de variables choisies aléatoirement à chaque split
- le nombre d'observations minimal par feuille
- le nombre maximal de feuilles (qui est un critère d'arrêt)
- la profondeur des arbres : le nombre maximal d'interactions pour obtenir une feuille

Pour de plus amples détails concernant le paramétrage de la méthode du gradient boosting, le lecteur peut se référer à l'ouvrage d'Eric Biernat et Michael Lutz, *Data science : fondamentaux et études de cas* (source : Biernat, E. & Lutz, M., 2015).

Mise en place de la théorie du gradient boosting

Après avoir présélectionné les variables catégorielles avec la méthode de LASSO, nous allons appliquer la théorie du gradient boosting sur l'ensemble des variables (les variables catégorielles sélectionnées et les variables numériques).

La régression par GBM est paramétrée comme suit :

- des arbres de profondeur 1 (dits stumps)
- une distribution de Bernoulli pour le modèle de transformation des devis, une distribution de Poisson pour les modèles de fréquence des sinistres et une distribution Gamma pour les modèles de coût moyen des sinistres

- un nombre d'arbres minimum égal à 500 et un maximum à 2000
- un nombre de partitions égal à 5 pour la validation croisée
- la valeur initiale du paramètre de rétrécissement (shrinkage) à 0,25
- 50% des variables choisies aléatoirement à chaque split (bag.fraction égal à 0,5)
- pas de taille minimale des feuilles

Le modèle est optimisé de façon à ce que le nombre d'arbres optimal soit entre 500 et 2000 : s'il est inférieur à 500, on divise par deux le paramètre de rétrécissement pour apprendre moins vite ; s'il est supérieur à 2000, on double ce méta-paramètre pour apprendre plus vite et réduire le nombre d'arbres nécessaires pour atteindre l'optimalité. Le nombre d'itérations optimal est celui qui minimise la moyenne des erreurs de généralisation sur les cinq partitions.

En sortie du GBM deux éléments principaux sont à disposition. Le premier est l'ensemble constitué de l'objet de type GBM ainsi que les influences relatives (standard et issues de la validation croisée) calculées pour chaque variable explicative. Le second est la liste finale des variables sélectionnées et l'influence relative issue de la validation croisée calculée pour chacune d'entre elles. L'influence relative issue de la validation croisée de chaque régresseur est calculée comme la moyenne des influences relatives obtenues sur chacune des cinq partitions. On trie par importance décroissante les facteurs de risque et on les ajoute successivement à la liste finale de variables jusqu'à obtenir une importance cumulée (ou information) égale à 99,9%.

2.2.4 Analyse des corrélations entre les variables retenues

Une fois que la sélection des variables est finalisée, une analyse des variables retenues s'impose, avant que celles-ci soient intégrées dans le modèle de prédiction.

En effet, bien qu'un maximum d'information aide toujours à mieux expliquer le risque à modéliser, il faut tout de même éviter la redondance d'information afin d'aider le modèle à mieux performer. Deux variables fortement corrélées, présentes toutes les deux dans un modèle, risquent d'entraîner une non-significativité des coefficients sur l'une des deux variables. Il vaut mieux alors en retenir une seule des deux, ou tenter de les résumer en une seule et unique variable, sans prendre en compte toutes les interactions, mais en retenant uniquement celles les plus significatives. Une analyse des corrélations est donc faite en amont avant de réaliser le modèle de prédiction.

Pour l'étude des corrélations, les indicateurs adoptés sont exposés dans le tableau suivant :

Type de corrélation	Indicateur
Corrélation entre deux variables qualitatives	V de Cramer
Corrélation entre deux variables quantitatives	Coefficient de Pearson

Tableau 2.1 – Indicateurs utilisés pour étudier les types de corrélation.

La corrélation des variables qualitatives

L'intensité de liaison entre deux variables nominales peut être mesurée soit à travers le test de Chi-Deux, soit à l'aide du coefficient V de Cramer.

Le test de Chi-Deux est obtenu en comparant les effectifs théoriques et les effectifs observés de chaque classe d'un tableau de contingence. Sa formule est la suivante :

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - T_{ij})^2}{T_{ij}}. \tag{2.14}$$

Le coefficient V de Cramer représente une alternative à ce test qui ne fait pas recours à une table de χ^2. Sa formule est la suivante :

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}}, \tag{2.15}$$

où χ^2_{max} = effectif x (minimum (nombre de lignes, nombre de colonnes) - 1).

Le V de Cramer est compris entre 0 (dans le cas d'une absence de liaison entre les variables) et 1 (liaison parfaite).

La probabilité associée au test de Chi-Deux manque de lisibilité quand elle est extrêmement sensible à la taille de la population. Des lors, nous avons préféré l'utilisation du V de Cramer au test classique de Chi-Deux car il fournit une mesure absolue de l'intensité de la liaison entre deux variables qualitatives, indépendamment du nombre de modalités et de l'effectif de la population. Il intègre l'effectif de la population par l'intermédiaire de χ^2_{max}.

Si l'on se fie à la probabilité du Chi-Deux pour la sélection des variables, les conclusions peuvent être opposées selon que l'on travaille sur la population entière ou sur un échantillon (car ce test est sensible à l'effectif) et de surcroît, on risque de ne pas sélectionner des variables qui apparaîtront pourtant discriminantes après un regroupement éventuel de certaines de leurs modalités.

Vu les avantages du V de Cramer par rapport au simple test de Chi-Deux, c'est cet indicateur qui a été choisi dans cette étude pour l'analyse des corrélations entre deux variables explicatives catégorielles.

La corrélation des variables quantitatives

L'analyse des corrélations entre deux variables ordinales peut se réaliser au travers de divers indicateurs tels que le coefficient de Pearson ou le coefficient de Spearman. C'est le coefficient de Pearson qui a été utilisé dans cette méthode. Etant une mesure canonique de la dépendance en univers gaussien, le coefficient de corrélation linéaire de Pearson entre deux variables aléatoires X_1 et X_2 est défini par :

$$r(X_1, X_2) = \frac{\mathbb{E}[X_1, X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]}{\sqrt{\mathbb{V}[X_1]\mathbb{V}[X_2]}}. \tag{2.16}$$

Compris entre -1 et 1 cet indicateur traduit une forte corrélation positive dans le cas où il est compris entre 0,5 et 1 et une forte corrélation négative dans le cas où est compris entre -1 et -0,5. Le cas ou le coefficient de Pearson r (X1,X2) = 0 apparaît dans les cas où les deux variables quantitatives X1 et X2 sont indépendantes. Dans le cadre de la modélisation nous ne cherchons pas à déterminer si les corrélations sont positives ou négatives, mais seulement si elles sont importantes. C'est donc la valeur absolue de cet indicateur qui sera analysée.

La corrélation entre une variable quantitative et une variable qualitative

Même si plusieurs indicateurs existent pour mesurer les corrélations entre une variable qualitative et une variable quantitative (tels que le test de Kruskal-Wallis), ceux-ci n'ont pas été jugés assez performants et robustes pour pouvoir être adaptés et automatisés lors de la construction du logiciel de modélisation. Nous n'allons donc pas réaliser d'analyse préalable entre ces deux types de variables explicatives.

Le seuil de corrélation maximal choisi est de 0,5. Les paires de variables qui dépassent le seuil de corrélation choisi ont été analysées à part. Dans ce cas, afin d'éviter la redondance de l'information et d'aider le modèle à mieux performer il vaut mieux retenir dans le modèle final une seule des deux variables (celle avec l'influence relative par rapport à la variable à expliquer la plus élevée d'après les résultats du modèle GBM), ou tenter de les résumer en une seule et unique variable, en gardant uniquement les interactions les plus significatives.

2.2.5 Linéarisation du gradient boosting

Dans les approches de modélisation traditionnelles, la méthode de gradient boosting est utilisée seulement pour des problèmes de classification. Le logiciel de modélisation utilisé dans ce mémoire applique cette technique d'apprentissage automatique non seulement pour réaliser la sélection des variables explicatives, mais il sert également en tant que modèle de régression pour prédire la variable réponse.

Les modèles de Gradient Boosting ne s'interprètent pas de manière aussi évidente qu'un modèle linéaire généralisé. Il était donc nécessaire de transformer la sortie du GBM en structure linéaire pour faciliter la lisibilité du modèle de prédiction construit. Nous allons, donc, par la suite décrire cette méthode de linéarisation du GBM.

La première étape est la construction d'un nouveau GBM à partir uniquement de la liste de régresseurs sélectionnés tel que le nombre de stumps optimal soit compris entre 2000 et 6000 (en ajustant le méta-paramètre de rétrécissement ou shrinkage).

En deuxième étape, nous allons parcourir les arbres où chacun de ces régresseurs interviennent et nous allons ajouter de manière successive la prédiction de ces arbres sur l'ensemble des modalités du régresseur concerné. Et celle-ci jusqu'à l'obtention de l'arbre optimal.

Concrètement, le coefficient associé à la modalité A de la variable *var* est égal à la somme des prédictions des arbres où la variable de scission est *var* pour la modalité A. Finalement, on obtient bien un coefficient pour chaque modalité de chaque régresseur. L'intercept est égal à la valeur prédite initiale sur laquelle les arbres font des ajustements. On obtient donc la suivante formule de calcul pour les prédictions :

$$prédiction = exp(\beta_{var1} + \beta_{var2} + intercept). \tag{2.17}$$

On applique ensuite l'offset en multipliant cette prédiction avec le poids considéré dans le problème de modélisation (le nombre de devis réalisés si l'on souhaite modéliser le taux de transformation des devis en contrats, le nombre d'année polices pour la modélisation de la fréquence des sinistres ou le nombre de sinistres si l'on souhaite modéliser le coût moyen des sinistres).

La différence entre la formule du GLM et celle du GBM linéarisé est subtile. Elle réside dans le fait que pour une variable numérique *var*, dans un modèle GLM, un seul et unique β_{var} est obtenu, qui sera ensuite multiplié avec la valeur de la modalité *var*, tandis que, dans le cadre de linéarisation du GBM, la variable numérique *var* est discrétisée. On n'utilise pas le produit entre β_{var} et *var*, mais simplement le coefficient $\beta_{modalité}$ correspondant à la modalité de la variable explicative numérique.

2.2.6 Recherche d'interactions avec GBM

L'approche mise en place consiste à utiliser le modèle construit à l'étape précédente en offset [4] et à entraîner un nouveau modèle qui cherche à prédire les quasi-résidus avec des variables croisées. Précisons que le caractère confidentiel de la technique utilisée par AXA Global P&C nous interdit de détailler cette étape.

2.2.7 Construction d'un zonier avec GBM

Lors de la construction d'un modèle, nous avons le choix soit d'intégrer dans le modèle le zonier actuellement utilisé dans le tarif commercial, soit de bâtir un nouveau zonier en partant de la maille géographique souhaitée (code postal, code iris etc.). Le logiciel de modélisation développé par AXA Global P&C permet de construire un nouveau zonier de manière automatisée. La méthode mise en place est similaire à celle utilisée pour la recherche des interactions.

4. Mettre le modèle en offset signifie que l'on fixe tous les coefficients.

Concrètement, pour construire un nouveau zonier, le modèle initial (sans variables géographiques) est d'abord mis en offset. Puis un nouveau modèle GBM est entraîné qui cherche à prédire les quasi-résidus avec la variable géographique choisie pour construire le zonier. Le caractère confidentiel de la technique développée nous interdit de détailler cette étape mais nous pourrons voir plus loin (dans la section : 4.2.2) les résultats obtenus dans le cas d'application.

A RETENIR :

- Le modèle de conversion de devis ainsi que les 8 modèles de fréquence et de coût moyen bâtis lors de ce mémoire sont réalisés à travers une nouvelle approche de modélisation intégrant les techniques de machine learning.
- Pour répondre à la contrainte de temps de calcul et de traitement d'importantes bases de données, les méthodes non paramétriques, telles que la régression pénalisée ainsi que les arbres de décision et le Gradient Boosting, sont de bonnes alternatives aux modèles linéaires généralisés. La méthode retenue ici pour la présélection et la simplification des variables catégorielles est la régression pénalisée LASSO. La théorie du gradient boosting est ensuite appliqué sur les variables catégorielles choisies en amont et sur l'ensemble des variables numériques. Après suppression des variables fortement corrélés un dernier modèle GBM est réalisé et sa sortie est transformée en structure linéaire afin de faciliter la lisibilité du modèle de prédiction construit. De plus, cette méthode semi-automatisée présente l'avantage de permettre l'identification des interactions ainsi que de bâtir des nouveaux zoniers à la maille géographique souhaitée.
- Pour éviter le sur-apprentissage et ainsi d'assurer la capacité de généralisation des modèles construits deux outils de validation sont utilisés : d'une part la technique de validation croisée et d'une autre part la courbe de lift qui permet de valider graphiquement la qualité d'ajustement des modèles. La performance des modèles est, quant à elle, mesurée à l'aide de l'indice de Gini ainsi que de l'EDR .

Chapitre 3

Modélisation du taux de transformation des devis

3.1 Collecte et préparation des données

Les données disponibles pour cette étude sont issues de sources différentes et ne sont donc pas construites de manière identique. Une bonne partie du travail préparatoire a d'abord consisté à assembler et uniformiser l'intégralité des données. Ce travail représente une part de temps non négligeable dans l'étude, mais permet de constituer une base de données robuste, ce qui est primordial pour la construction du modèle.

3.1.1 Choix de la période d'observation

L'étude se base sur les interrogations tarifaires deux-roues réalisées par des agents et courtiers pendant la période 01.01.2019 - 30.06.2019. Comme présenté dans la section 1.2, les interrogations tarifaires contiennent l'ensemble des contrats (interrogations tarifaires transformées en affaires nouvelles), devis, instances, note de couverture (NC) et demandes d'autorisation préalable (DAP).

Ce choix du périmètre provient de deux contraintes. D'une part, l'analyse a servi comme base pour le choix des mesures tarifaires appliqués sur les affaires nouvelles à partir de l'année 2020. Ainsi, il a fallu choisir la période de devis la plus récente au moment de réalisation de l'étude. La deuxième contrainte est représentée par le souhait d'avoir une base de données homogène. On aurait pu choisir une période d'analyse plus longue, mais, pour ne pas biaiser l'étude en prenant en compte des interrogations tarifaires réalisés avec des versions tarifaires trop anciennes, nous avons jugé suffisant de prendre en considération une période d'observation étendue sur 6 mois seulement.

3.1.2 Construction de la base de données

Une fois la période d'observation définie, nous avons procédé à la collecte des données. La base de départ de l'analyse est la base d'interrogations tarifaires transversale d'AXA, contenant toutes les interrogations tarifaires des produits IARD : devis automobiles, multirisque habitation, multirisque professionnels, deux-roues, collection. Cette base de données contient des informations d'ordre général, telles que : le numéro de devis, le produit concerné, la région de vente, et des informations concernant le réseau de distribution (le numéro de l'agent/courtier, le numéro du portefeuille). Dans cette base, les informations tarifantes des devis (quelque soit le type de produit) ne sont pas présentes.

Pour pallier ce manque d'informations, une partie importante dans la constitution de la base d'analyse est la récupération des informations tarifantes pour chaque devis.

Pour reconstituer cette base, il a fallu constituer un outil sous VBA permettant de récupérer les informations tarifaires à partir de l'outil de gestion des contrats interne à AXA. L'outil ainsi

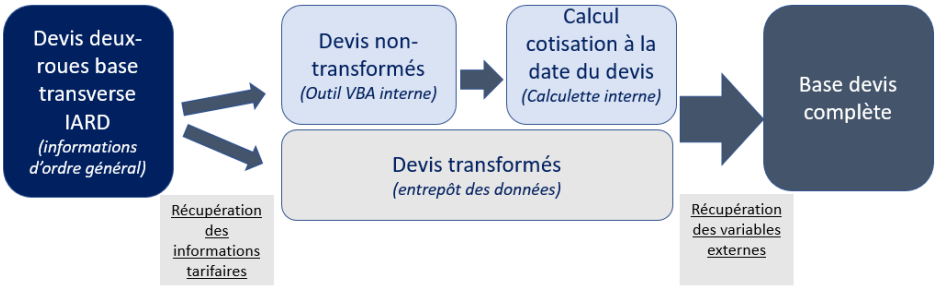
constitué contient plusieurs informations nécessaires à la récupération des variables demandées à la souscription, telles que l'emplacement de chaque information tarifaire dans l'écran de gestion des contrats, la longueur de chaque variable et le numéro de l'écran où chaque information est disponible.

De plus, il a fallu prendre en considération au moment de la construction de l'outil certains aspects qui peuvent bloquer le processus et empêcher ainsi la récupération de l'intégralité des informations tarifantes. Ces éléments concernent en général des changements intervenus dans la situation du devis entre le moment de réalisation de celui-ci et le moment de l'extraction des informations de l'outil de gestion. Au titre d'exemple, un devis deux roues peut bénéficier d'une réduction commerciale si le client est propriétaire d'un contrat d'assurance automobile chez AXA. Dans ce cas, au moment de réalisation du devis, le distributeur doit renseigner une clause de réduction dans l'écran de gestion et il doit remplir également le numéro du contrat automobile en cours. Si, entre le moment de réalisation du devis et le moment d'extraction des informations avec l'outil, le contrat référencé dans la clause de réduction est résilié, un motif de blocage s'affiche à l'écran. Dans ce cas l'outil ne peut pas récupérer l'intégralité des informations du devis.

Ainsi, une fois que tous ces éléments ont été analysés et intégrés dans l'outil, nous avons réussi, de manière automatisée, à reproduire le procès de reprise des devis réalisé par le réseau de distribution dans l'outil de gestion des contrats. Afin de diminuer le temps nécessaire à la construction de la base de devis, nous avons fait le choix d'extraire avec cet outil seulement les interrogations tarifaires non transformées, en sachant que pour les devis transformés en affaire nouvelle, les informations tarifaires sont présentes dans l'entrepôt de données SAS.

Une fois que les devis non-transformés ont été extraits, un travail de formatage des données a été nécessaire afin d'uniformiser les informations extraites des plusieurs sources des données. Une contrainte de plus posé par l'absence des informations tarifaires dans la base de devis de l'entrepôt est le calcul de la cotisation proposée au moment de réalisation du devis. L'outil d'extraction construit permet de récupérer la cotisation du devis, mais calculée au moment de l'extraction des informations. Celle-ci pose un problème dans le cas où la version tarifaire a changé entre le moment de réalisation du devis et le moment de l'extraction. Il était donc nécessaire de recalculer la cotisation que le prospect a vu au moment de la réalisation de son devis. A partir des informations tarifantes récupérées, nous avons pu recalculer, à l'aide d'une calculette interne, le tarif vu par le client au moment de réalisation du devis. La troisième étape de la construction de la base de données est l'intégration des données externes, non présentes dans le formulaire d'interrogation tarifaire.

Le schéma ci-dessous permet de résumer les différentes étapes parcourues :



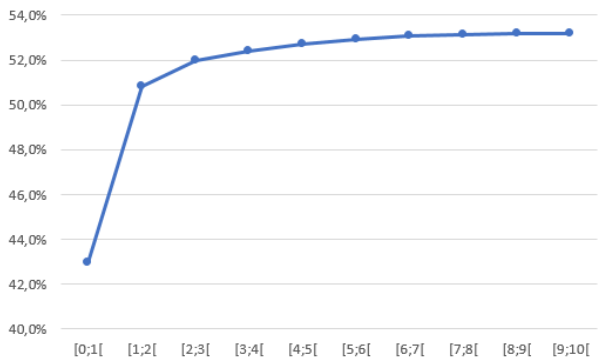
Graphique 3.1 – Schéma de construction de la base de devis deux-roues.

3.1.3 Dédoublonnement de la base et choix du recul de 4 mois

Le premier travail sur la base ainsi obtenue est l'identification des doublons : si plusieurs devis ont été réalisés par le même client ou prospect pendant le même mois, nous avons gardé le devis le plus récent. Si on n'avait pas réalisé ce travail préliminaire, on risquait de construire un modèle biaisé, dont le même profil serait pris en compte plusieurs fois (à titre d'exemple, si l'on suppose que la plupart des conducteurs des deux-roues puissantes décident de créer le même devis plusieurs fois, le coefficient associé à une modalité dont les conducteurs des deux-roues puissantes sont sur-représentées est superficiellement augmenté, ce qui ne reflétera pas la réalité).

En ayant une période d'étude fixée de 6 mois, nous remarquons que le taux de transformation des premiers mois est plus important que le taux de transformation des derniers mois puisque la durée d'observation post devis est décroissante avec le mois de transformation. Par exemple, pour une étude réalisée en octobre 2019, les interrogations tarifaires réalisées en janvier 2019 ont eu 10 mois pour réaliser leur passage en affaire nouvelle, alors que les devis réalisés au mois de juin 2019 n'auront eu que 4 mois pour réaliser cette transformation. Il est donc nécessaire d'homogénéiser la base d'étude suivant le mois d'émission de l'interrogation tarifaire, en déterminant le recul nécessaire pour stabiliser le taux de transformation et ensuite l'appliquer à chaque interrogation tarifaire.

Le choix de la période de recul à prendre en compte pour considérer le passage d'un devis en affaire nouvelle est réalisé en se basant sur le graphique ci-dessus, représentant le taux de transformation en fonction du délai entre la date de visualisation du tarif et celui de la transformation brute :



Graphique 3.2 – Taux de transformation cumulé par délai (exprimé en mois).

A travers ce graphique, nous remarquons qu'au bout de 4 mois près de 99% des affaires nouvelles sont réalisées. A partir ce délai la courbe est stabilisée et par conséquent, un recul de quatre mois pour chaque tarif vu stabilise le taux de transformation.

Ainsi, on considère qu'une interrogation tarifaire ne peut passer en affaire nouvelle que sous les quatre mois après visualisation du tarif. Si la date de transformation n'est pas comprise entre la date de réalisation du devis + 4 mois, alors on considère que le prospect n'a pas transformé son projet. On supprime ainsi certaines affaires nouvelles de la base de l'étude, mais l'impact est marginal.

A RETENIR :

- L'étude se base sur les interrogations tarifaires deux-roues réalisées par des agents et courtiers pendant la période 01.01.2019 - 30.06.2019.
- La construction et la mise en forme de la base de données représente une part de temps non négligeable dans l'étude. La récupération des critères tarifaires absents initialement de la base de devis transverse a demandé le développement d'un outil automatisé sous VBA capable de reproduire le processus de reprise des devis réalisé par le réseau de distribution dans l'outil de gestion des contrats.
- La cotisation aperçue par le client au moment de la réalisation du devis a été calculée à partir des informations tarifaires.
- Une période de recul de 4 mois a été choisie pour considérer le passage d'un devis en affaire nouvelle.

3.2 Analyse exploratoire des variables

L'analyse descriptive des variables est une étape indispensable avant d'initier les travaux de modélisation. Elle permet notamment de s'approprier les données, de détecter d'éventuelles sources futures d'instabilité des résultats (faible exposition) mais également d'éliminer les critères non liés à la variable réponse étudiée. Une variable pour laquelle le taux de transformation observé est similaire pour toutes ses modalités ne présente pas d'intérêt a priori pour la modélisation.

Après quelques spécifications concernant les variables présentes dans la base de données, nous allons préciser la méthode utilisée pour la discrétisation des variables continues. Ensuite nous réalisons quelques analyses descriptives permettant de mieux appréhender la conversion des interrogations tarifaires en affaires nouvelles, en amont de la modélisation.

3.2.1 Les variables initiales

Afin d'identifier les facteurs discriminants qui expliquent le mieux la transformation des interrogations tarifaires en affaire nouvelle, le point de départ de l'étude comprend une analyse exhaustive des devis, à partir d'un grand nombre d'informations (données demandées dans le formulaire de souscription, données clients etc.). Ainsi, nous avons fait un tri au préalable pour nous focaliser sur les variables qui nous semblent les plus intéressantes notamment des variables tarifaires sur lesquels nous pourrons agir suite à une révision tarifaire. A la première étape nous avons choisi 39 variables explicatives, dont 24 variables qualitatives nominales, 3 variables qualitatives ordinales, 8 variables quantitatives continues et 4 variables quantitatives discrétisées :

Liste de variables	
Variables concernant le conducteur	1. Le statut marital 2. Le type de permis 3. L'âge du conducteur 4. L'ancienneté du permis du conducteur 5. L'antécédents d'assurance croisés avec l'expérience de conduite 6. La durée de détention du véhicule déclarative 7. L'expérience de conduite 8. La profession du conducteur 9. Les antécédents d'assurance
Variables concernant le véhicule à assurer	1. La clause de réduction moto 2. L'ancienneté du véhicule 3. La classe croisé avec la carrosserie 4. La cylindrée 5. La durée de détention du véhicule actuel 6. Le genre (la carrosserie) 7. Le genre croisé avec la cylindrée 8. Le groupe 9. La marque 10. Le type de roues 11. L'usage du véhicule
La sinistralité	1. Présence de sinistres dans les 24 dernières mois 2. Nombre de sinistres dans les 24 dernières mois 3. Type de sinistralité dans les 24 dernières mois
Variables géographiques	1. La région Axa 2. La zone de risque
Le produit	1. La formule des garanties 2. Le fractionnement 3. L'option sécurité du conducteur étendue 4. La version tarifaire
La cotisation	1. La cotisation TTC 2. Le coefficient technique du contrat 3. Le crédit commercial appliqué
Le distributeur	1. Coefficient technique personnalisé du distributeur 2. Le type de distributeur 3. Agent VIP
Le client	1. La clause de multi détention 2. Le type de client
La cotation	1. Ecarts en jours entre la date d'émission du devis et la date d'effet souhaitée de l'AN 2. Type de l'interrogation tarifaire

Tableau 3.1 – Taux de transformation de devis : les variables analysées.

Dans cette liste finale de variables retenues on retrouve des variables classiques de description du risque de type auto ou moto (âge du conducteur, ancienneté de permis) mais aussi des informations spécifiques au produit moto (cylindrée, le genre), ou des variables spécifiques AXA (région AXA, agent vip, clause multi-détention). Dans la section 3.2.3 nous allons revenir sur certaines de ces variables pour donner plus d'explications et faciliter leur compréhension.

3.2.2 Discrétisation des variables quantitatives continues

Comme précisé dans la section 3.2.1, nous avons discrétisé en amont quatre variables quantitatives afin qu'elles soient utilisées conjointement aux autres.

La raison principale qui justifie ce traitement est une contrainte imposée par le logiciel de modélisation interne d'AXA. En effet, le logiciel impose un nombre maximal de 255 modalités différentes pour chaque variable continue. Si ce seuil est dépassé, la variable est automatiquement rejetée par le logiciel, et elle ne peut pas être intégrée au modèle de régression. Pour les autres variables quantitatives (de moins de 255 modalités), le regroupement en classes est réalisé a posteriori par le logiciel de modélisation avec des ajustements à la main de l'actuaire s'ils se sont avérés nécessaires.

Outre la contrainte informatique, ce traitement des variables quantitatives apportera plus de simplicité et de lisibilité pour notre modèle de régression.Après avoir découpé en classes ces

variables, le modèle construit va appliquer un coefficient pour chaque modalité, et non pas un coefficient par variable. Pour que l'on puisse utiliser une variable continue sans la discrétiser dans un modèle de régression, il est nécessaire qu'entre cette variable et la variable à expliquer il existe une liaison monotone. Il y a certaines variables, comme l'âge du conducteur par exemple, qui ne se prêtent pas forcement bien à être considérées comme variables quantitatives. La relation entre cette variable et le taux de conversion des devis n'est pas monotone ; il est préférable donc de considérer cette variable comme une variable qualitative ordinale.

En plus, la discrétisation offre l'avantage de gérer les valeurs extrêmes (qui seront regroupées dans la première et la dernière tranche) et les valeurs manquantes qui seront rassemblées dans une tranche supplémentaire spécifique (si c'est le cas).

En conséquence, quatre variables ont été discrétisées a priori : **la cylindrée de la moto**, **la cotisation TTC**, **le crédit commercial appliqué** et **l'écart en jours entre la date d'émission du devis et la date d'effet souhaitée de l'affaire nouvelle**.

Même si plusieurs méthodes statistiques de discrétisation existent, nous avons choisi de créer le nombre de classes pour ces quatre variables de façon empiriques, en faisant attention d'avoir suffisamment d'individus dans chaque classe, pour que le taux de transformation reste interprétable sur chaque tranche.

Pour discrétiser **la cylindrée de la moto**, nous nous sommes appuyés sur l'avis des experts Club 14. Cette caractéristique d'une moto, qui traduit également la puissance de celle-ci, n'est pas seulement une des variables prises en compte dans la tarification du produit d'assurances, mais elle est étroitement liée à la règlementation française concernant le produit deux-roues. Par exemple, les motos légères, de moins de 125 cm3, peuvent être conduites que si le conducteur est détenteur d'un permis de type A1 ou B. Il est donc nécessaire d'avoir une tranche a part pour ces types de motos, et de ne pas les inclure dans une autre classe avec des véhicules de cylindrée différente. Pour une explication plus détaillée concernant le lien entre la cylindrée de la moto et le type de permis nécessaire à sa conduite, le lecteur intéressé pourra se référer à l'annexe B.

Par souci de lisibilité, **le montant TTC** du tarif proposé (qui prend en compte également la réduction appliquée par le distributeur) a été discrétisé par tranche de 50€. Les interrogations tarifaires dont le prix proposé était supérieur à 1.300 € ont été regroupés dans la même tranche, ce seuil étant considéré comme un seuil pécunier psychologique pour le client. Une analyse plus approfondie de cette dernière classe s'est avérée nécessaire, ayant comme objectif d'identifier ces profils et en conséquence soit de les exclure de la souscription, soit de plafonner la cotisation pour ces profils d'assurés. Cette analyse fait l'objet d'une étude à part et elle ne sera pas traitée dans le cadre de ce mémoire.

Le crédit commercial accordé par le distributeur au moment de réalisation du devis est aussi une des variables qu'on pense, a priori, très explicative du taux de transformation. Si un contexte de mise en concurrence apparaît, le distributeur peut choisir d'accorder un rabais commercial. De manière similaire, cette variable a été séparée par tranche de 10€, limitée à 150€.

La dernière variable à discrétiser est **l'écart entre la date d'émission du devis et la date d'effet souhaitée de l'affaire nouvelle**. Le choix des classes a été fait en fonction du nombre de devis de chaque classe et en fonction d'une logique chronologique (les devis ayant plus d'un mois d'écart entre la date de l'interrogation tarifaire et la date souhaitée de l'affaire nouvelle, qui représentent environ 10% de la base d'analyse, ont été regroupés dans la même classe).

3.2.3 Statistiques descriptives univariées

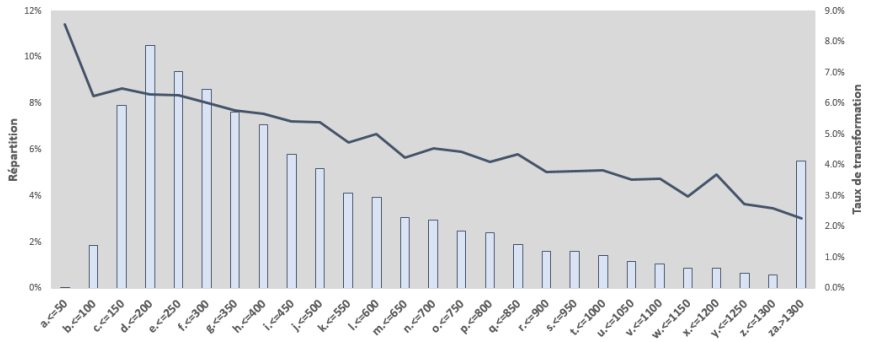
La première partie de l'étude consiste à s'approprier les variables afin de connaître le contenu des bases, de faire des retraitements selon la qualité des données (gestion des valeurs manquantes, valeurs extrêmes, valeurs inconnues ou mal renseignés) et aussi de comprendre les caractéristiques qui impactent le plus le taux de transformation.

Au global, le taux de transformation observé est de 52,4% sur environ 60.000 interrogations sauvegardées. Ce taux, que nous souhaitons modéliser à travers ce mémoire, correspond au rapport entre le nombre d'affaires nouvelles et le nombre de tarifs sauvegardés.

Pour mieux appréhender le phénomène analysé on peut représenter graphiquement la répartition des interrogations tarifaires sauvegardées et le taux de transformation observé sur chaque modalité de chaque variable. Ce travail préliminaire a été réalisé sur l'ensemble des 39 variables, en amont de la modélisation. Pour ne pas alourdir l'étude nous allons représenter graphiquement quatre variables parmi les plus explicatives du taux de transformation (ordre d'importance issue du modèle GBM).

La cotisation TTC

La variable qui différencie le plus le taux de transformation est la cotisation effective qui doit être payée par le prospect s'il décide de s'assurer chez AXA. Celle-ci est une cotisation toutes taxes comprises et elle prend également en compte les éventuels rabais appliqués par les distributeurs au moment de l'interrogation tarifaire.

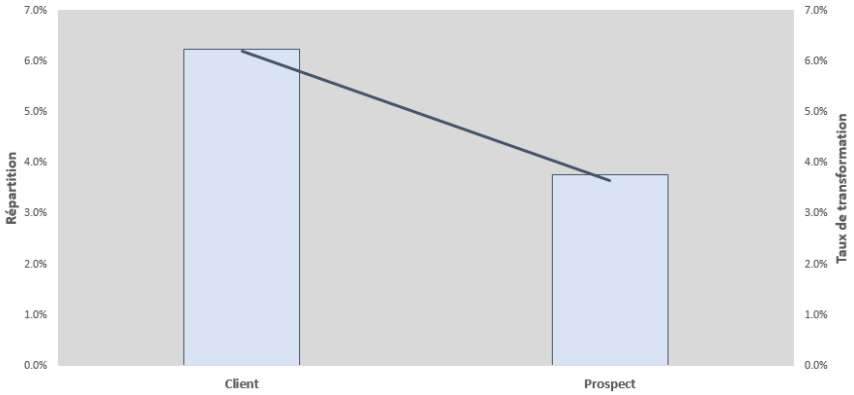


Graphique 3.3 – Le taux de transformation par tranche de cotisation TTC.

Sans grande surprise, le taux de transformation est une fonction décroissante du prix. On constate cependant quelques exceptions sur certaines tranches de cotisation (par exemple la tranche 1150€-1200€) qui ne respectent pas la tendance décroissante du taux de conversion en fonction de la prime à payer. Ces exceptions sont généralement liées à la faible exposition de certains segments. Lors de la construction du modèle multivarié, elles seront regroupées avec des tranches de prix inferieures ou supérieures, pour respecter la tendance constatée.

Le type de prospect

Une autre variable très explicative de la transformation des devis en contrats est le type de prospect.



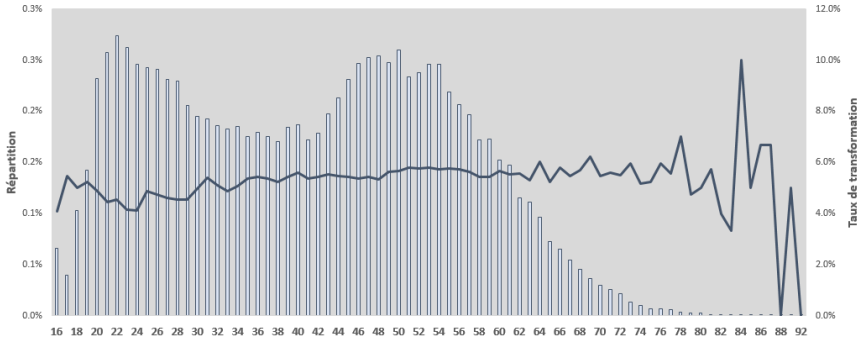
Graphique 3.4 – Le taux de transformation par type de client.

Pendant la période de janvier à juin 2019, 60% des interrogations tarifaires ont été réalisées par des clients ayant au moins un autre produit d'assurance AXA. Leur taux de transformation moyen est de 62%, tandis que le taux de transformation des devis réalisés par des nouveaux clients est de seulement 37%. On constate donc qu'il est plus facile de transformer une affaire pour des clients déjà assurés chez Axa, ce qui s'explique facilement pour différentes raisons :

un lien qui existe déjà avec l'agent général, une volonté de regroupement de l'ensemble des contrats pour faciliter leur gestion, des avantages tarifaires afin d'encourager la multidétention.

L'âge du conducteur principal

Une autre variable fortement liée à la conversion des devis en affaires nouvelles est l'âge du conducteur principal.

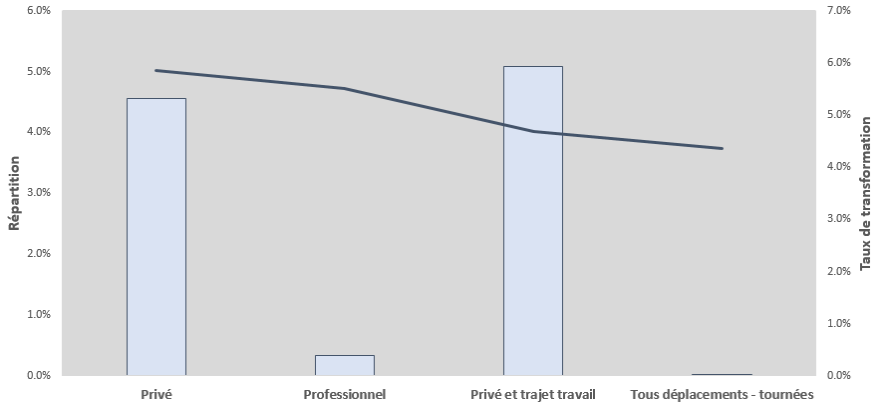


Graphique 3.5 – Le taux de transformation en fonction de l'âge du conducteur principal.

La représentation graphique montre une tendance croissante de la transformation en fonction de l'âge du conducteur principal, à partir de l'âge de 23 ans. A partir du seuil de 60 ans, les effectifs par âge sont très faibles. En conséquence, pour consolider les résultats, les modalités de cette variable seront regroupées en tranche, en même temps que la modélisation du taux de transformation. On constate aussi une très bonne conversion pour les conducteurs très jeunes, de moins de 20 ans, ce qui peut signifier que les concurrents proposent des prix majorés pour ce segment par rapport au produit moto AXA/Club 14.

L'usage de la moto

La dernière variable présentée est celle de l'usage de la moto. Cette variable est tarifante et elle fait aussi partie des critères de refus à la souscription en association avec le groupe du véhicule assuré. Par exemple, un conducteur souhaitant assurer une moto de groupe élevée en usage professionnel se verra interdit à la souscription.



Graphique 3.6 – Le taux de transformation en fonction de l'usage de la moto.

Malgré un tarif plus élevé pour les deux-roues en usage professionnel, le taux de transformation pour cette catégorie de motos est plus important que la conversion observée sur les motos en usage privé-trajet travail (qui représentent la majorité du périmètre d'analyse).

A RETENIR :

- L'analyse exploratoire des variables permet de s'approprier les données, de traiter les éventuelles problèmes de qualité des données usuelles (valeurs manquantes) et de détecter d'éventuelles sources futures d'instabilité des résultats (faible exposition) mais également d'éliminer les critères non liés à la variable réponse étudiée.
- Pour analyser le taux de transformation des devis, nous nous sommes focalisés sur une sélection de 39 variables qui nous semblent les plus intéressantes, notamment des variables tarifaires sur lesquels nous pourrons agir suite à une révision tarifaire. Parmi les facteurs analysés on compte des variables classiques du secteur IARD (telles que des variables concernant le conducteur, les caractéristiques du véhicule à assurer, des informations concernant la sinistralité antérieure etc.) ainsi que des variables spécifiques au produit d'assurance deux-roues AXA.
- Une discrétisation en amont de la modélisation de 4 facteurs de risque a été nécessaire : la cotisation TTC du devis, le crédit commercial accordé par l'agent, la cylindrée de la moto ainsi que l'écart en jours entre la date d'émission et la date d'effet du devis.
- L'analyse univariée du taux de transformation observé permet d'identifier empiriquement les facteurs de risque ayant le pouvoir discriminant le plus important, tels que : la cotisation du devis, le type de prospect, l'âge du conducteur principal ainsi que l'usage de la moto.

3.3 Mise en œuvre de la méthodologie et résultats

Après avoir préparé les données pour qu'elles soient exploitables, puis les avoir explorées à travers des analyses descriptives univariées, nous avons mis en œuvre le processus de modélisation.

Nous souhaitons construire un modèle qui permette de restituer la probabilité de transformation d'une interrogation tarifaire en contrat. La variable à expliquer est binaire, déclinée en deux modalités :

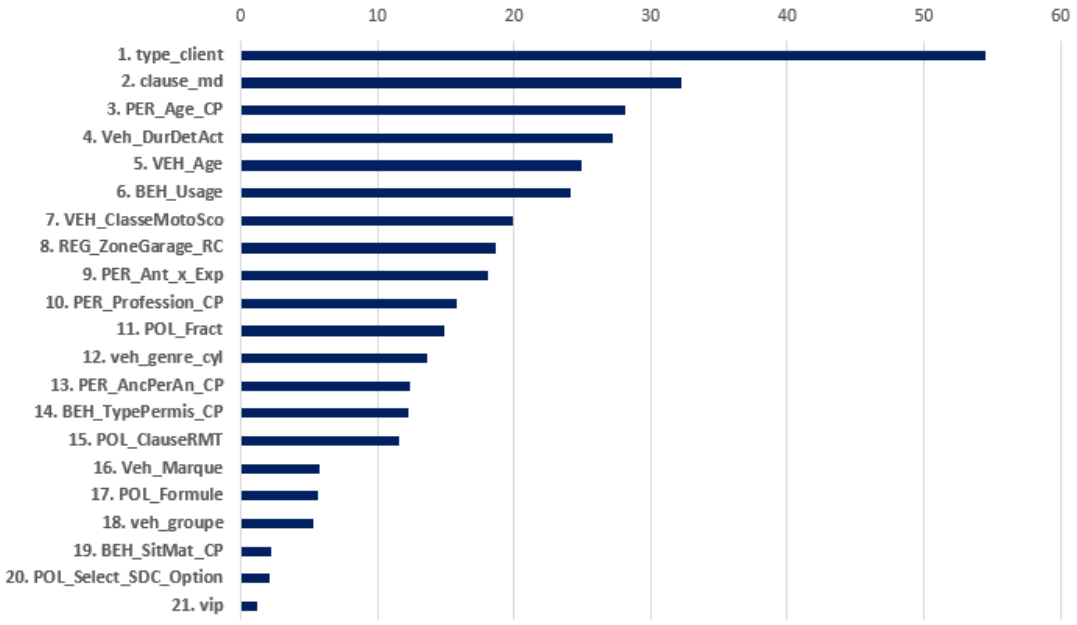
- le devis est transformé en contrat : Y=1
- le devis n'est pas transformé en contrat : Y=0.

D'abord, la base initiale de devis a été divisée en deux partitions : la base d'apprentissage et la base de validation, en proportion de 80-20. Ensuite, la première base a été divisée de façon aléatoire en 4 échantillons indépendants de taille égale. Le modèle est alors bâti sur l'ensemble des 3 premières bases (folds), puis sa performance est évaluée sur la 4-ème base, la base de validation. Ainsi, pour la validation du modèle, on utilise la technique de validation croisée (4-fold) dont les avantages ont été exposés dans la section 2.2.1 du mémoire.

Dès lors que la variable à modéliser est une variable binaire, la distribution utilisée pour la modélisation est la distribution de Bernoulli.

Variables sélectionnées dans le modèle final

Parmi les 39 variables présentes dans la base de modélisation, seulement 21 ont été choisies dans le modèle final : 4 variables numériques et 17 variables catégorielles (dont 2 qui sont des croisements des variables).



Graphique 3.7 – Influence relative des variables sur le taux de transformation (exprimé en %).

Le graphique ci-dessous illustre l'influence relative de chacune des 21 variables sur la variable à expliquer, le taux de transformation des interrogations tarifaires en contrat. Cette influence est mesurée à travers le spread 95/5.

Permettant de mesurer le pouvoir discriminant des variables sélectionnées dans le modèle, le spread peut être calculé sur chaque niveau de variables et/ou en supprimant les niveaux avec une exposition inférieure à 5%. Sa formule de calcul est la suivante :

$$Spread = \frac{1+\beta_{max}}{1+\beta_{min}} - 1.$$

Une variable est d'autant plus discriminante dans un modèle que son spread est élevé. Si une variable a un spread nul, cela signifie que ses betas (les coefficients régresseurs) sont proches de 0 et donc, par principe de parcimonie, la variable n'est pas incluse dans le modèle (son ajout dans le modèle ne permet pas d'augmenter le pouvoir explicatif du modèle).

Le spread 95/5 a été préféré au spread 100/0. En choisissant le spread 100/0 comme mesure du pouvoir discriminant d'une variable, le risque est de surestimer l'importance des variables très segmentées, pour lesquelles les segments de moindre exposition (inférieure à 5%) ont des coefficients estimés extrêmes. Le spread 95/5, permet de palier ce risque, en ne prenant en compte que les coefficients associés à des modalités d'exposition supérieures à 5% pour chaque régresseur.

Il est important de préciser que dans le cadre de cette analyse, deux modèles de transformation des devis ont été testés : un premier modèle prenant en compte la cotisation du devis et un deuxième sans cette variable. Nous avons fait le choix d'utiliser le deuxième modèle, car, des lorsqu'on inclue la cotisation de l'interrogation tarifaire dans le modèle de transformation, le pouvoir explicatif des autres variables intégrées dans le modèle est très faible. De plus, l'objectif final de l'étude est l'identification des segments sur ou sous-tarifés par rapport à la concurrence, en prenant en compte leur rentabilité attendue et d'ainsi identifier les mesures tarifaires à réaliser. Les ajustements tarifaires vont être réalisés pour chaque segment identifié, et non pas au niveau global, en fonction de leur cotisation finale.

Coefficients et interprétation

Une fois les variables sélectionnées, un nouveau GBM est entraîné sur cette sélection de vingt-et-un facteurs de risque, puis la transformation du GBM en GLM est opérée : on obtient alors un jeu de coefficients pour chacun des régresseurs que nous allons analyser et interpréter avec un regard “métier” par la suite. Le tableau suivant résume les variables retenues dans le modèle final, leur granularité initiale et finale, ainsi que leur pouvoir discriminant.

Liste de variables	Descriptif variable	Nom de la variable	Type variable	Granularité initiale	Granularité finale	Importance dans le modèle
Variables concernant le conducteur	1. Le statut marital	BEH_SitMat_CP	Qualitative nominale	6	4	19
	2. Le type de permis	BEH_TypePermis_CP	Qualitative nominale	6	4	14
	3. L'âge du conducteur	PER_Age_CP	Quantitative continue	75	7	3
	4. L'ancienneté du permis du conducteur	PER_AncPerAn_CP	Quantitative continue	68	6	13
	5. Les antécédents d'assurance croisés avec l'expérience de conduite	PER_Ant_X_Exp	Qualitative nominale	3	2	9
	6. La profession du conducteur	PER_Profession_CP	Qualitative nominale	8	7	10
Variables concernant le véhicule à assurer	1. La clause de réduction moto	POL_ClauseRMT	Qualitative nominale	2	2	15
	2. L'ancienneté du véhicule	VEH_Age	Quantitative continue	79	6	5
	3. La classe croisé avec la	VEH_ClasseMotoSco	Qualitative ordinale	17	12	7
	4. La durée de détention du	VEH_DurDetAct	Quantitative continue	43	5	4
	5. Le genre croisé avec la	veh_genre_cyl	Qualitative nominale	44	20	12
	6. Le groupe	veh_groupe	Qualitative ordinale	16	5	18
	7. La marque	veh_marque2	Qualitative nominale	17	11	16
	8. L'usage du véhicule	BEH_Usage	Qualitative nominale	4	3	6
Variables géographiques	1. La zone de risque	REG_ZoneGarage_RC	Qualitative ordinale	13	9	8
Le produit	1. La formule des garanties	POL_Formule	Qualitative nominale	8	4	17
	2. Le fractionnement	POL_Fract	Qualitative nominale	2	2	11
	3. L'option sécurité du conducteur étendue	POL_Select_SDC_option	Qualitative nominale	2	2	20
Le distributeur	1. Agent VIP	vip	Qualitative nominale	2	2	21
Le client	1. La clause de multi détention	pol_clausemd_vf	Qualitative nominale	2	2	2
	2. Le type de client	type_client	Qualitative nominale	2	2	1

Tableau 3.2 – Modèle de transformation de devis : résumé des variables retenues dans le modèle final.

Le modèle contient 14 variables qualitatives nominales, 3 variables qualitatives ordinales et 4 variables quantitatives continues. Après simplification, la granularité initiale de ces 21 variables explicatives a été divisée par 4, le nombre de coefficients régresseurs passant alors à environ 120.

On se propose d’analyser les coefficients obtenus par linéarisation du GBM pour une sélection

des variables explicatives, en gardant à l'esprit les contraintes métier et opérationnelles. Cette interprétation des tendances observées et prédites à travers les coefficients régresseurs n'a de sens que toutes choses égales par ailleurs.

Graphique 3.8 – Modélisation du taux de transformation de devis : exemples des paramètres β estimés.

Dans le graphique 3.8.1 on constate que les devis réalisés par des clients multi-détenteurs ont plus de chances de se transformer en affaires nouvelles que les devis réalisés par des clients mono-détenteurs ou des prospects.

Le taux de transformation des devis a une tendance décroissante en fonction du groupe de la moto (figure 3.8.2). On constate également qu'une des modalités se détache nettement en termes de transformation par rapport aux autres : le groupe de moto 39. Dans cette catégorie on retrouve les motos les plus puissantes. Puisque le risque associé à cette catégorie de motos est très élevé, le tarif commercial a été fixé en conséquence, d'où le faible taux de transformation constaté sur ce segment.

Dans la figure 3.8.3 nous avons représenté les valeurs observées et prédites, ainsi que les coefficients toutes choses égales par ailleurs en fonction de la classe du véhicule. Cette variable, issue du véhiculier interne d'AXA, classe les motos et les scooters en fonction de leur risque associé (allant de J à R, par ordre croissante du risque). Les motos des classes extrêmes (les classes J, P, Q et R) ont des coefficients toutes choses égales par ailleurs nettement inférieurs à la moyenne. Pour augmenter le taux de transformation global une des solutions serait d'ajuster le tarif commercial de ces segments sous réserve que l'analyse du risque associé le permette.

Concernant la profession du conducteur principal, les artisans, les professions libérales et les commerciaux transforment mieux que la moyenne, tandis que pour les retraités la transformation est moindre.

Comme attendu, le taux de transformation est décroissant en fonction de la zone de risque. Une attention particulière doit être accordée à la zone de risque la plus élevée, car son taux de transformation est nettement inférieur à la moyenne constatée.

Enfin, on constate que le tarif commercial est peut-être trop élevé par rapport à la concurrence sur le segment des motos en usage privé et trajet-travail. De la même manière, un ajustement du tarif commercial serait à étudier de manière croisée avec le risque associé.

La performance et la validation du modèle

La performance du modèle est évaluée à l'aide de deux indicateurs : le Gini et l'EDR , définis dans la section 2.2.1.

Indicateur	échantillon d'apprentissage	échantillon de test	total
Taux de transformation moyenne observée	52,41%	52,46%	52,42%
Taux de transformation moyenne prédite	52,68%	52,63%	52,67%
Indice de Gini	20,18%	19,76%	20,02%
EDR	10,33%	9,88%	10,13%

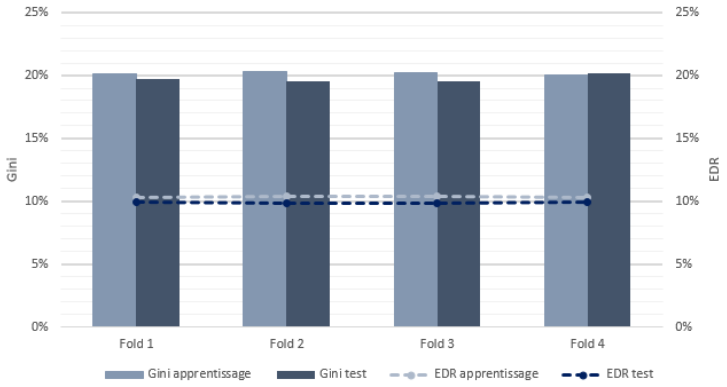
Tableau 3.3 – Principaux indicateurs de performance du modèle de transformation de devis.

Dans une première étape, ces indicateurs ont été calculés sur les échantillons d'apprentissage et de test comme moyennes obtenues sur les k-folds. Étant donné que le modèle a été bâti sur la base d'apprentissage, il est normal d'obtenir des résultats légèrement meilleurs sur cette base, par rapport à l'échantillon de validation.

Dans une deuxième étape nous avons appliqué le modèle sur la base totale des devis et nous avons calculé les indicateurs de performance sur cette base également. On constate que le modèle de régression a une légère tendance à surestimer le taux de transformation de devis.

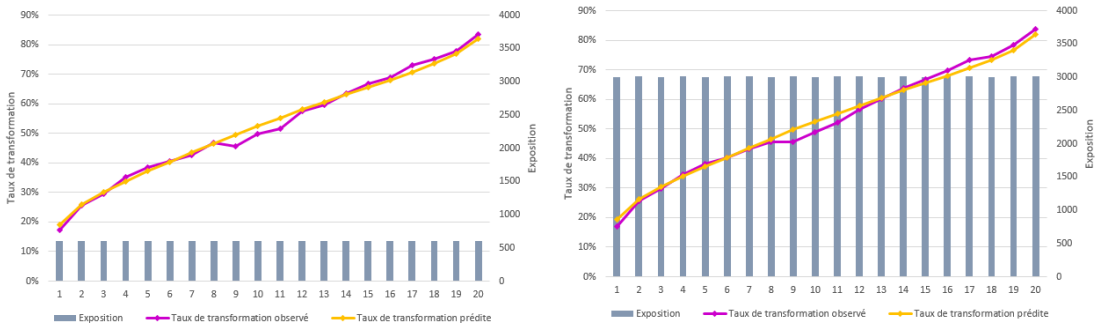
La validation du modèle a été réalisée à travers la méthode de validation croisée (k-fold) et à travers la courbe de lift, définis dans la section 2.2.1.

La k-fold cross validation donne des résultats concluants puisque les indicateurs de Gini et les EDR sur la base d'apprentissage et de test sont stables au cours des 4 cross validations, comme le montre le graphique 3.9. Le modèle final retenu est obtenu en calculant la moyenne des prédicteurs sur les 4 folds d'apprentissage.



Graphique 3.9 – Validation du modèle de transformation de devis par la méthode de validation croisée.

La deuxième méthode de validation utilisée est la validation par quantiles, à travers la courbe de lift. Comme précisé dans la section 2.2.1, cette méthode consiste à construire des quantiles sur les taux de transformation prédits issus du modèle et de tracer les taux de transformation observés et prédits par quantile. De par la construction du graphique, le taux de transformation prédit est croissant. Si le modèle est performant, l'observé doit également être croissant suivant les quantiles et doit se rapprocher le plus de la valeur prédite.



Graphique 3.10 – Courbes de lift construites sur la base de validation (à gauche) et sur la base totale de devis (à droite).

Nous avons choisi de réaliser la courbe de lift sur 20 quantiles. Malgré une légère surestimation sur les segments n° 9, 10 et 11, la validation est satisfaisante puisque le taux transformation prédit est globalement proche du taux de transformation observé tant sur la base de validation que sur la totalité de la base de devis.

Globalement, le modèle de transformation des interrogations tarifaires en affaires nouvelles est satisfaisant.

A RETENIR :

- La modélisation du taux de transformation des devis a été réalisée par linéarisation du GBM avec, pour fonction de lien, la loi de Bernoulli. Une pré-sélection et simplification des variables catégorielles est réalisée à travers la méthode de LASSO. Ensuite le choix des variables intervenant dans le modèle est réalisé à l'aide de la méthode d'agrégation GBM.
- Deux méthodes de validation du modèle sont mises en œuvre : la validation croisée ainsi que la validation par quantiles. Parmi les différentes modèles valides construits le modèle final retenu est celui qui présente l'indice de Gini le plus élevé et l'EDR le plus bas (pour le même niveau de Gini).
- Le modèle final utilise 21 variables comme facteur de risque, la simplification des variables ayant permis de diviser par 4 leur granularité initiale.

Chapitre 4

Calcul de l'ELR

4.1 Préliminaire : indicateurs de rentabilité

L'identification des mesures tarifaires à l'affaire nouvelle ne repose pas seulement sur l'analyse de compétitivité mesurée à travers le taux de transformation. En nous basant uniquement sur le marché nous risquerions d'avoir des prix attractifs mais des contrats non rentables. Il est donc primordial de prendre en compte également la rentabilité théorique attendue des devis, mesurée par le biais de l'indicateur ELR.

Dans cette partie nous allons présenter succinctement les principaux indicateurs de rentabilité utilisés pour le pilotage d'un portefeuille d'assurance, puis nous allons détailler la méthodologie de calcul de l'ELR, en commençant par le calcul de la prime pure par garanties, puis en réalisant les ajustements aux charges finales prévisibles sur le portefeuille d'analyse. En dernière étape nous allons présenter les résultats obtenus sur la base de devis de janvier à juin 2019.

4.1.1 Le S/C et ses limites

Le S/C est l'indicateur de rentabilité le plus utilisé pour le pilotage des portefeuilles d'assurances. Calculé comme rapport entre le montant de sinistres et celui de cotisations acquises sur la même période, il permet de savoir si le tarif est suffisant pour couvrir le risque. Notons que la cotisation payée par un assuré correspond à la prime commerciale définie dans la section 1.2.1.

$$\text{S/C} = \frac{\text{Montant de sinistres}}{\text{Cotisations acquises}}.$$

Un contrat rentable pour l'assurance est donc un contrat qui a un S/C peu élevé.

Malgré le fait que le S/C soit un indicateur indispensable dans le suivi et le pilotage de notre rentabilité, il comporte certaines limites. En effet, cet indicateur se base uniquement sur la sinistralité observée. Il reflète uniquement ce qui s'est réellement passé.

De ce fait, nous ne pouvons pas, à travers le S/C, estimer la rentabilité future d'un contrat. Il est impossible de calculer un S/C en affaire nouvelle. En effet, lorsque l'assuré souscrit, nous ne connaissons pas encore sa sinistralité à venir, de ce fait nous ne pouvons pas utiliser le S/C comme indicateur de la rentabilité en affaire nouvelle.

En outre, le montant de sinistres pris en compte dans le calcul de ce ratio ne permet pas d'avoir une vision finale de la sinistralité ; pour avoir une vision ultime de la charge de sinistres il faudrait également prendre en compte les sinistres non encore déclarés ou survenus (les IBNR).

Le S/C est un bon indicateur pour suivre la rentabilité du portefeuille dans le temps mais il est inadapté pour analyser la rentabilité des affaires nouvelles.

4.1.2 L'ELR et ses avantages

l'Expected Loss Ratio (ELR), ou le ratio de sinistres espérés à primes, est un autre indicateur très utilisé dans le pilotage du portefeuille d'assurances. Il est défini de la manière suivante :

$$\text{ELR} = \frac{\text{PP}}{\text{PC}}.$$

Calculé comme rapport entre la prime pure estimée (comme défini dans la section 1.2.1) et la prime commerciale, l'ELR reflète la rentabilité technique annuelle estimée de façon individuelle au niveau de chaque police d'assurance. Il représente une prédiction du S/C, parce que la prime pure est une estimation de la charge que devra payer l'assuré en cas de sinistres.

Cet indicateur permet d'analyser la rentabilité de certains contrats pour lesquels la sinistralité n'est pas renseignée. Ainsi, il permet :

- D'analyser la rentabilité des affaires nouvelles et des devis : lors de la souscription d'une affaire nouvelle et au moment de la réalisation d'un devis, nous ne connaissons pas la sinistralité à venir des devis; nous pouvons donc réaliser une estimation par le biais de la prime pure.

- De mieux estimer la rentabilité d'un contrat récent : de la même façon que pour les affaires nouvelles et les devis, lorsque le contrat n'a pas beaucoup d'ancienneté dans le portefeuille, nous n'avons pas assez de recul pour avoir une vision robuste de sa sinistralité. C'est pourquoi nous utilisons la prime pure comme indicateur de la sinistralité attendue.

- De tenir compte du vieillissement des sinistres dans l'estimation de la charge finale : les futurs sinistres d'un assuré ne sont pas connus en intégralité au moment de l'analyse, mais l'estimation des IBNR est prise en compte par le biais des modèles de prime pure. Cet aspect est d'autant plus important car le poids de sinistres corporels avec des longues déroulés représente une part significative sur la branche d'assurance moto .

Ainsi, l'ELR permet de pallier les inconvenants du S/C et il est devenu un indicateur de suivi et de pilotage de la rentabilité indispensable au secteur de l'assurance [1]. De plus, il est le seul indicateur qui permet d'estimer la rentabilité attendue des devis. C'est donc cet indicateur qui a été choisi pour être analysé conjointement avec le taux de transformation des devis dans le but d'identifier les mesures tarifaires à l'affaire nouvelle.

A RETENIR :

- ► Les principaux indicateurs de rentabilité utilisés pour le pilotage des portefeuilles d'assurance sont le loss ratio (S/C) et l'Expected loss ratio (ELR)
- ► Calculé comme rapport entre la sinistralité observée et le montant des cotisations acquises, le S/C présente le désavantage de ne pas pouvoir être utilisé pour estimer la rentabilité des affaires nouvelles et des devis.
- ► Pour estimer la rentabilité attendue des devis nous allons utiliser l'ELR, indicateur calculé comme rapport entre la prime pure modélisée et la prime commerciale.

1. De plus il permet de contrôler l'adéquation entre les modèles de prime pure et de prime commerciale. Pour de plus amples détails le lecteur intéressé peut se référer au mémoire IA de M. Vandekerchove (source : Vandekerchove, 2019)

4.2 Modélisation de la prime pure

Comme défini dans la partie précédente, les deux composantes qui rentrent dans le calcul de l'ELR sont la prime pure et la prime commerciale. La prime commerciale des devis est présente dans la base de données, car elle est calculée au moment de la réalisation de l'interrogation tarifaire. Pour pouvoir calculer les ELR des devis, la composante manquante est donc la prime pure individuelle.

Pour pouvoir la calculer, nous devons d'abord réaliser le calcul de la prime pure globale sur les contrats en portefeuille. Celle-ci s'obtient en additionnant la prime pure de chacune des garanties :

- Responsabilité civile matérielle
- Responsabilité civile corporelle
- Sécurité du conducteur
- Incendie et vol (chez AXA France, il est coutume de regrouper les deux garanties ensemble)
- Dommages (regroupant les garanties dommages collision et dommages tous accidents)
- Les autres garanties (protection juridique, casque, assistance etc.)

L'objectif principal de cette section est de présenter les notions théoriques relatives au calcul de la prime pure, et notamment l'approche fréquence-coût moyen et d'ensuite illustrer la mise en œuvre de cette approche sur la garantie responsabilité civile (RC) matérielle.

4.2.1 Préliminaire : le modèle collectif de l'assurance et le calcul de la prime pure

L'assurance non-vie se distingue notamment de l'assurance vie tant par la survenance du sinistre que par son coût. En effet, en assurance vie, la survenance du sinistre est certaine, seule sa date est inconnue ; tandis qu'en assurance non-vie, sa survenance est probable (avec une probabilité strictement comprise entre 0 et 1). De plus, en assurance non-vie le coût du sinistre est rarement connu à l'avance (on parle de remboursement indemnitaire), alors que ce n'est pas le cas en assurances de personnes, le coût du sinistre étant un coût forfaitaire.

Par conséquent, pour modéliser la sinistralité d'un portefeuille de contrats d'assurance non-vie, on se place dans le cadre traditionnel du modèle collectif qui permet de distinguer à la fois la fréquence des sinistres sur le portefeuille (modèle de comptage) et aussi leur coût associé (ou sévérité) quel que soit l'assuré (modèle de montant).

Dans le cadre du modèle collectif, la charge totale de sinistres S_i d'un assuré i peut s'écrire de cette forme :

$$S_i = \sum_{k=1}^{N_i} Y_{i,k}, \tag{4.1}$$

où N_i est le nombre aléatoire de sinistres de l'assuré i et $Y_{i,k}$ est le coût de son $k^{\text{è}me}$ sinistre, avec les hypothèses suivantes :

- les $Y_{i,k}$ sont indépendants et identiquement distribués, pour tout i et pour tout k
- $\forall k \geq 1, \mathrm{Y}_{i,k}$ est indépendante de N_i pour tout i.

Dans le cadre de la modélisation de la prime pure, le modèle collectif distingue donc la modélisation de la fréquence des sinistres par assuré et celle de la sévérité des sinistres quel que soit l'assuré. Les arguments en faveur de cette distinction sont multiples. Tout d'abord, il n'y a pas de raison particulière de penser que les déterminants de la fréquence et du coût des sinistres soient les mêmes. Une modélisation spécifique à chaque quantité parait donc a priori plus pertinente.

Ensuite, la fréquence est connue pour présenter un comportement beaucoup plus stable que le coût. Il est donc recommandé de valoriser cette robustesse naturelle du modèle de fréquence, et de considérer la problématique de coût indépendamment. Dans le cadre de ce mémoire, nous emploierons systématiquement cette distinction lors de l'élaboration des différents modèles de prime pure.

La modélisation de la prime pure consiste donc à prédire la charge future espérée d'un contrat. En passant l'équation précédente à l'espérance on obtient le résultat suivant :

$$\begin{aligned}
\mathbb{E}(S_i) &= \mathbb{E}\Big[\sum_{k=1}^{N_i} Y_{i,k}\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big(\sum_{k=1}^{N_i} Y_{i,k}|N_i\Big)\Big] \\
&= \mathbb{E}\big[\mathbb{E}\big(N_i Y_i|N_i\big)\big] \text{ car } Y_{i,k} \text{ i.i.d} \\
&= \mathbb{E}\big[N_i \mathbb{E}\big(Y_i|N_i\big)\big] \\
&= \mathbb{E}\big[N_i \mathbb{E}\big(Y_i\big)\big] \text{ par indépendance de } Y_i \text{ et } N_i \\
&= \mathbb{E}\big(N_i\big)\mathbb{E}\big(Y_i\big).
\end{aligned} \tag{4.2}$$

La charge espérée de l'assuré i peut donc être calculée comme produit des espérances du nombre de sinistres et du coût moyen d'un sinistre. Si on rapporte cette valeur à l'exposition t_i du contrat, on obtient la prime pure :

$$\text{prime pure} = \text{fréquence x coût moyen} = \frac{\mathbb{E}(S_i)}{t_i}, \tag{4.3}$$

$$\text{fréquence} = \frac{\text{nombre de sinistres}}{\text{exposition}} = \frac{\mathbb{E}(N_i)}{t_i}, \tag{4.4}$$

$$\text{coût moyen} = \frac{\text{charge de sinistres totale}}{\text{nombre de sinistres}} = \frac{\mathbb{E}(S_i)}{\mathbb{E}(N_i)}. \tag{4.5}$$

L'exposition du contrat, également connue sous le terme de nombre d'années-police est noté t_i et désigne la durée de couverture du risque. Au titre d'exemple, un contrat souscrit le 1er juillet de l'année N et toujours en cours au 31 décembre de l'année, a une exposition égale à 0,5 années-police. Si un sinistre survenait sur ce contrat au cours de cette demi-année, on aurait donc schématiquement une fréquence observée de 200% car la fréquence est une moyenne annualisée : 1 sinistre sur une demi-année équivaut, en termes de risque, à en moyenne 2 sinistres par an.

On a vu que la prime pure était classiquement le produit de deux composantes : la fréquence et la sévérité qui sont modélisées indépendamment par la méthode de linéarisation du gradient boosting machine.

Il est important de préciser que dans le cadre de ce mémoire, nous avons modélisé à travers l'approche fréquence-coût moyen les quatre garanties principales du produit moto (RC matérielle, RC corporelle, incendie-vol et dommages), obtenant ainsi, en première étape huit modèles (quatre modèles de fréquence et quatre modèles de sévérité). Concernant les autres garanties (les garanties optionnelles et les garanties dont le poids dans la charge de sinistres est moins important) nous avons fait le choix de ne pas réaliser une segmentation de la prime pure mais d'intégrer un montant forfaitaire de prime pure, proportionnel avec leur exposition.

Nous allons par la suite illustrer comment cette approche de modélisation de la prime pure est mise en œuvre sur la garantie responsabilité civile matérielle.

4.2.2 Modélisation de la fréquence des sinistres RC matériels

4.2.2.1 Collecte et préparation des données

Périmètre d'étude et vision de sinistralité

On s'intéresse à modéliser la fréquence des sinistres RC matériels sur les contrats deux-roues en portefeuille sur la période 2016 à 2018. En couvrant les dommages matériels causés à autrui, la garantie RC (responsabilité civile) constitue la seule garantie obligatoire sur le produit d'assurance deux-roues. Nous conservons donc tous les contrats qui ont été en cours au moins un jour pendant la période analysée, quelle que soit la formule de garanties souscrite. Nous travaillons sur les contrats gérés par les agents généraux et les courtiers, pour les cinq régions de souscription d'AXA France.

Pour ne pas omettre les sinistres tardifs dans le modèle de fréquence, pour l'analyse de la sinistralité nous avons pris en compte la vision la plus récente au moment de l'étude. Nous avons donc enrichi la base de contrats en portefeuille avec les sinistres survenus entre le 1er janvier 2016 et le 31 décembre 2018, tous les sinistres étant vus à fin octobre 2019. De cette manière, un sinistre survenu en 2018 mais déclaré tardivement, pendant l'année 2019, est bien pris en compte dans la modélisation.

Construction de la base de données

Les données chiffrées précisées ci-après concernent uniquement les contrats deux-roues incluant la garantie responsabilité civile matérielle ainsi que les sinistres de cette nature mais la constitution de la base pour chacune des autres garanties est similaire.

La base de données relative à la garantie RC matérielle, représente environ 2,3 millions de lignes, chaque ligne correspondant à une image. Un contrat conserve la même image tant qu'il ne connaît pas de modifications, c'est à dire tant qu'il conserve les mêmes caractéristiques. Dès lors que le contrat subi un mouvement, une nouvelle ligne est générée. Nous pouvons donc avoir plusieurs images par contrat, l'exposition de chaque image étant calculée comme l'écart annualisé entre la date de fin d'image et la date de début d'image.

Les mouvements qui justifient la création d'une nouvelle image, et donc d'une nouvelle ligne sont :

- Les affaires nouvelles
- Les avenants
- Les anniversaires de contrats

Chaque image est considérée comme une observation à part entière dans nos modélisations. Ainsi les 2,3 millions d'images (lignes) représentent un peu plus de 1 million d'années-polices. La base contient les caractéristiques d'environ 22.000 sinistres RC matérielle, survenus pour environ 21.700 assurés. La base de données ainsi construite contient plus de 180 variables, que l'on peut classer comme suit :

- Variables concernant le conducteur principal
- Variables concernant le véhicule à assurer
- Variables géographiques
- Variables concernant la sinistralité déclarée à la souscription, ou la sinistralité survenue du contrat AXA (par nature de sinistre)
- Variables concernant le produit
- Variables concernant le distributeur

Conducteur	La moto	Géographiques	Sinistres	Le produit	Le distributeur
PER_ : conducteur BEH_ : comportamental CLI_ : client	VEH_ : véhicule à assurer	REG_	CLA_	POL_	POL_

Graphique 4.1 – Types de variables à disposition.

Nettoyage de données

Pour la construction de la base finale de modélisation, un nettoyage de la base initiale constitué est réalisé en amont en dimension verticale et horizontale.

Suppression des colonnes sans intérêt préalable

Etant donné la multitude de facteurs de risque inclus dans la base de données, on exclut d'abord les nombreuses variables qui nous paraissent a priori n'avoir aucun pouvoir explicatif tels que des variables de type identificateur (le numéro de sinistre, le numéro de client, le numéro de contrat etc.), les variables présentant trop de valeurs manquantes, les variables concernant les cotisations par type de garantie etc. On exclut également les charges de sinistres par garanties pour la modélisation de la fréquence des sinistre, mais on la conserve pour la modélisation du coût moyen. Pour la deuxième partie de la modélisation (la modélisation du coût moyen des sinistres) la charge de sinistres pour la garantie considérée est bien-sûr gardée dans la base de données. Après ce retraitement, nous obtenons une base de 62 variables qui servira à la modélisation de la fréquence des sinistres.

Suppression des lignes

Les lignes dont l'exposition est nulle ont été écartées de l'analyse. De même, avant de réaliser la jointure entre la base de contrats et la base de sinistres, les sinistres annulés ont été ôtés de la base. Les sinistres sans suite ont cependant été conservés car ils génèrent usuellement des frais d'expertise médicale et d'autres frais de gestion de sinistres. Ils sont en conséquence comptés dans le nombre de sinistres final. On écarte également les contrats temporaires de l'analyse, car ils constituent un risque à part, et leur présence dans le périmètre d'analyse pourrait biaiser les résultats.

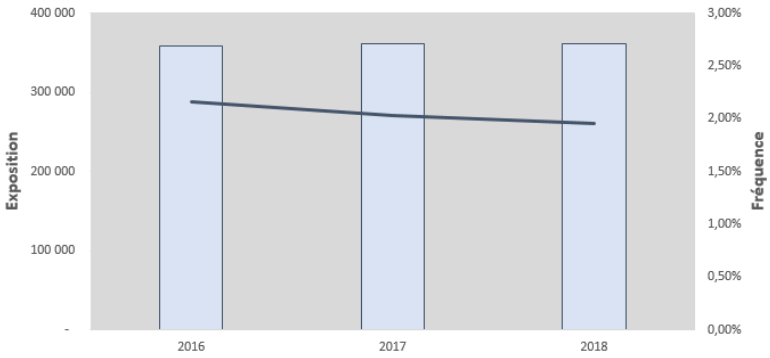
4.2.2.2 Analyse exploratoire des variables

Cette étape est indispensable avant d'initier les travaux de modélisation. Elle permet notamment de s'approprier les données, de détecter d'éventuelles sources futures d'instabilité des résultats (faible exposition, absence de sinistres pour une modalité) mais également d'éliminer les critères non liés à la variable réponse étudiée. Une variable pour laquelle le risque associé est similaire pour toutes ses modalités ne présente pas d'intérêt a priori pour la modélisation.

Après quelques spécifications sur la variable que l'on souhaite modéliser (le nombre de sinistres), nous réalisons quelques analyses descriptives permettant de mieux appréhender la sinistralité de la garantie RC matérielle, en amont de la modélisation.

Le nombre de sinistres

La base utilisée, obtenue par jointure entre les différentes sources mentionnées précédemment, est constituée d'autant de lignes que d'images. Pour chaque image le nombre de sinistres correspond au nombre de sinistres survenus au cours de la période intra-annuelle de chaque année étudiée où le contrat était en vigueur.



Graphique 4.2 – Fréquence RC matérielle observée par année d'étude.

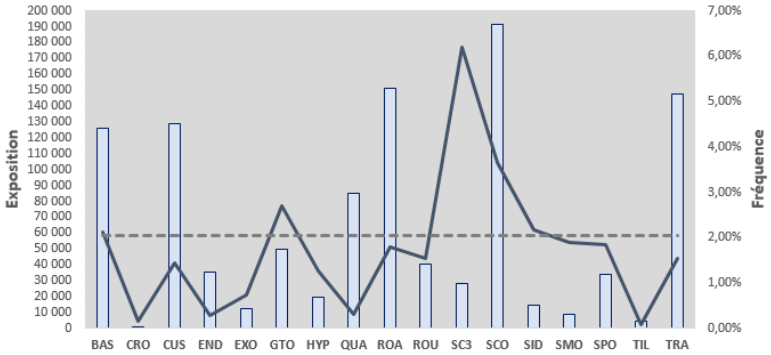
On constate, d'après la figure 4.2, que la fréquence des sinistres RC matériels est relativement stable sur les trois années d'étude, avec une tendance légèrement décroissante, elle se situant à un niveau d'environ 2%.

Analyse univariée des variables classiques

Nous allons par la suite représenter graphiquement l'évolution de la sinistralité moyenne en fonction de quelques facteurs de risque. Dans chaque graphique, la courbe de la fréquence moyenne observée est accompagnée d'un histogramme représentant l'exposition de chaque modalité sur la période analysée. Pour faciliter l'analyse nous avons également ajouté la fréquence moyenne constatée sur le portefeuille analysé (en pointillé). Les principaux intérêts de cette étape sont :

- Observer l'effet a priori de chaque critère sur la variable réponse étudiée, en l'occurrence la fréquence des sinistres RC matériels
- Éliminer les critères dont une seule modalité concentre la quasi-totalité de l'exposition car cet aspect traduit un faible pouvoir discriminant de ces critères
- Détecter d'éventuelles sources pouvant perturber les modélisations (des modalités avec une faible exposition)

Le genre de la moto



Graphique 4.3 – Fréquence RC matérielle observée selon le genre de moto.

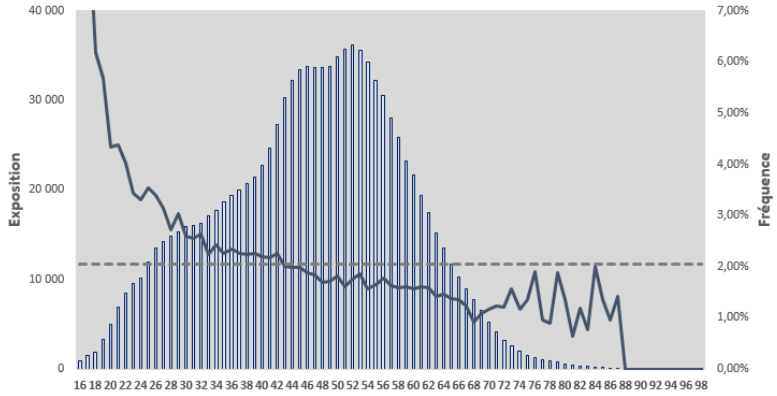
La figure 4.3 fait apparaître des disparités concernant la fréquence des sinistres RC matériels en fonction de la catégorie de moto. Ainsi, on constate une nette sur-sinistralité pour les scooters trois-roues et les scooters deux-roues (mais à moindre mesure). Ce phénomène se traduit notamment par l'usage urbain de ce type de motos. En effet, les scooters sont devenus depuis quelques années le moyen de transport le plus souvent utilisé dans les grandes villes ou les voies complètement saturées ont poussé les habitants à opter plutôt pour un scooter que pour une automobile. De plus, les conducteurs de scooters trois roues et des scooters deux-roues de moins de 125 cm^3 ne sont pas obligés de passer le permis moto afin de pouvoir les conduire. Un

permis auto suivi d'un stage de 7h sont jugés suffisants. Ces deux éléments peuvent expliquer la sur-sinistralité de ce segment.

On constate également une sur-sinistralité pour les motos de type grand tourisme (GTO). Ce sont des motos lourdes, peu maniables, utilisées généralement par des conducteurs qui parcourent beaucoup de kilomètres annuellement, ce qui peut expliquer le phénomène constaté.

A l'inverse, les quads sont faiblement sinistrés. Ce phénomène nous semble cohérent, étant donné que ce type de moto n'est pas utilisé pour parcourir beaucoup de kilomètres, mais plutôt pour un usage de loisir. La même situation est constatée sur les motos de type custom, qui sont des motos "plaisir", peu puissantes, utilisées également pour un usage de loisir, donc de manière plus occasionnelle.

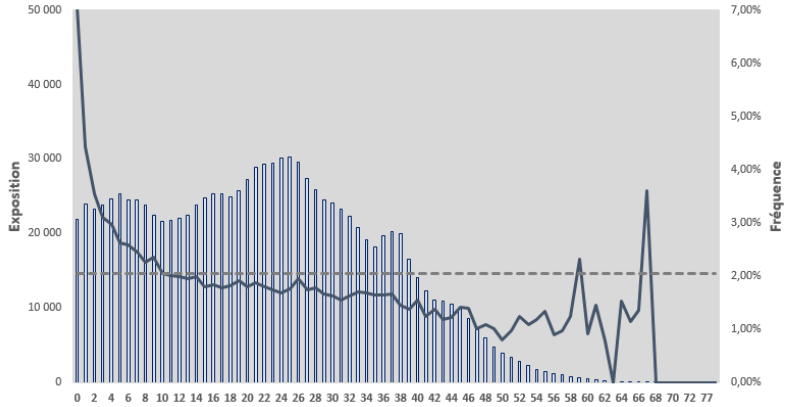
L'âge du conducteur



Graphique 4.4 – Fréquence RC matérielle observée en fonction de l'âge du conducteur.

La figure 4.4 fait apparaître la relation négative entre l'âge du conducteur et la sinistralité. On constate une tendance décroissante de la fréquence des sinistres RC matériels en fonction de l'âge du conducteur. Les valeurs observées à partir de l'âge de 70 ans sont assez volatiles. Cela fait suite à la faible exposition constatée dans notre portefeuille d'étude des conducteurs de plus de 70 ans.

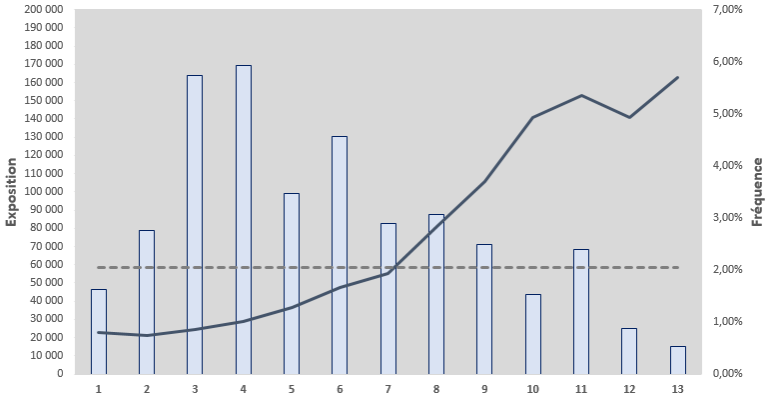
L'ancienneté de permis



Graphique 4.5 – Fréquence RC matérielle observée en fonction du nombre d'années d'ancienneté de permis.

Dans la figure 4.5 on constate la tendance décroissante de la fréquence des sinistres RC matériels en fonction de l'ancienneté de permis. A partir de 50 ans de permis, les valeurs observés ne sont pas très consistantes, suite à la faible exposition de ces profils.

La zone de risque



Graphique 4.6 – Fréquence RC matérielle observée par zone de risque.

La zone géographique du modèle actuel, quant à elle, présente une relation positive linéaire avec le nombre de sinistres RC matériels (à l'exception de la zone de risque 12). Le sujet de zonier est un élément important pour l'appréhension du risque. Malgré cette première analyse qui semble cohérente il sera peut-être nécessaire de retravailler ce sujet en profondeur. Pour s'assurer de la bonne adéquation du modèle de prime pure à la composante zonier du risque actuel il nous semble important de faire une mise à jour du zonier pour avoir la vision la plus récente possible. Nous allons revenir sur ce sujet dans la partie 4.2.2.3.

4.2.2.3 Modèle final

Après avoir traité et nettoyé les données pour qu'elles soient exploitables, puis les avoir explorées à travers des analyses descriptives univariées, le processus de modélisation par linéarisation du gradient boosting est mis en place.

Nous souhaitons par la suite construire un modèle qui permette de prédire le nombre de sinistres. La distribution du portefeuille en fonction du nombre de sinistres RC matériels survenus sur les trois années d'études est donnée dans le tableau suivant :

Nombre de sinistres	Exposition	Exposition (%)
0	1.068.764	98,88 %
1	11.822	1,09 %
2	265	0,02 %
3	8	0,01 %

Tableau 4.1 – Distribution du nombre de sinistres RC matériels

Sur la période considérée, le nombre de sinistres déclarés sur une image varie entre 0 et 3. Si la grande majorité des contrats n'a pas déclaré de sinistres RC matériels pendant la période 2016 à 2018 (98,88%), il est à noter toutefois que 8 contrats ont totalisé 3 sinistres déclarés. La fréquence annuelle empirique moyenne est de 2,05% pour une variance de 3,14% ce qui indique une sur dispersion caractéristique d'un portefeuille hétérogène (justifiant de cette manière la segmentation a priori).

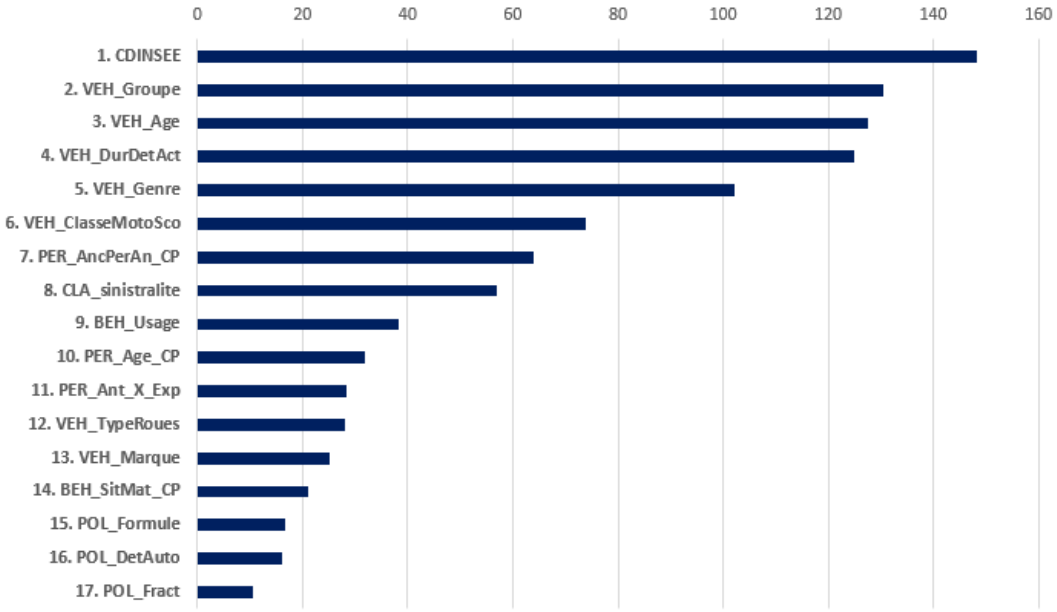
Pour la modélisation du nombre de sinistres, la loi de Poisson apparaît comme la plus légitime, le nombre de sinistres déclarés sur un exercice ne pouvant prendre que des valeurs entières positives ou nulles. C'est donc cette distribution qui a été utilisé comme fonction de lien pour la modélisation de la fréquence des sinistres.

Pour la validation du modèle nous avons utilisé la technique de validation croisée (4-fold). La base initiale a été divisée en deux partitions : la base d'apprentissage et la base de validation,

en proportion de 80-20. Ensuite, la première base a été divisée de façon aléatoire en 4 échantillons indépendants de taille égale. Le modèle est alors bâti sur l'ensemble de 3 premières bases (folds) , puis sa performance est évaluée sur la 4-ème base, la base de validation.

Variables sélectionnées dans le modèle final

Parmi les 62 variables présentes dans la base de modélisation, seulement 17 ont été choisies dans le modèle final : 4 variables numériques et 13 variables catégorielles (dont 1 qui représente un croisement des deux autres variables).



Graphique 4.7 – Influence relative des variables sur la fréquence des sinistres RC matériels (exprimé en %).

Le graphique ci-dessus illustre l'influence relative de chacun des régresseurs sur la variable à expliquer, c'est à dire le nombre de sinistres RC matériels. Cette influence est mesurée par le biais du spread 95/5, indicateur défini dans la section 3.3.

On constate que la variable la plus discriminante du modèle de prédiction de la fréquence des sinistres est le code INSEE. Les trois autres variables les plus discriminantes sont le groupe de la moto, l'ancienneté du véhicule et la durée de détention du véhicule actuel, qui ont des influences relativement équivalentes sur le phénomène analysé.

Il est important de préciser que dans le cadre de cette analyse, pour chaque type de garantie modélisée, deux modèles de fréquence des sinistres ont été réalisés : un premier modèle avec la zone de risque utilisée actuellement dans le tarif commercial et un deuxième modèle dans lequel nous avons construit un nouveau zonier de fréquence en fonction du code INSEE du lieu de garage.

Le zonier est la variable catégorielle qui a pour but de capter les effets purement géographiques de la sinistralité. L'introduction d'un tel critère dans la tarification est courante en assurance IARD. Elle est justifiée par le fait que la zone géographique impacte la sinistralité, par sa structure (nombre de voies rapides, de routes nationales), par la qualité de son réseau routier (les routes mal entretenues augmentent le risque d'accidents) ainsi que par les conditions climatiques propres à chaque région. La densité de la population, le degré d'urbanisation sont également des critères qui jouent sur la sinistralité.

Comme précisé dans la partie 2.2.7, le logiciel développé par AXA Global P&C permet de construire un nouveau zonier de manière automatisée en partant de la maille de zonage souhaitée (code postal, code iris etc.). Nous avons donc testé les deux zoniers (le zonier actuel et le nouveau

zonier à la maille INSEE) et nous avons choisi le modèle qui permet d'obtenir de meilleurs résultats en termes de performance (concrètement, entre les deux modèles construits, nous avons choisi comme modèle final celui avec le meilleur indice de Gini), c'est à dire le zonier actualisé. Ce qui confirme la pertinence d'avoir remis en question le zonier actuellement en place.

Coefficients et interprétation

Une fois les variables sélectionnées, un nouveau GBM est entraîné sur cette sélection de dix-sept facteurs de risque, puis la transformation du GBM en GLM est opérée : on obtient alors un jeu de coefficients pour chacun des régresseurs que nous allons analyser et interpréter avec un regard "métier" par la suite.

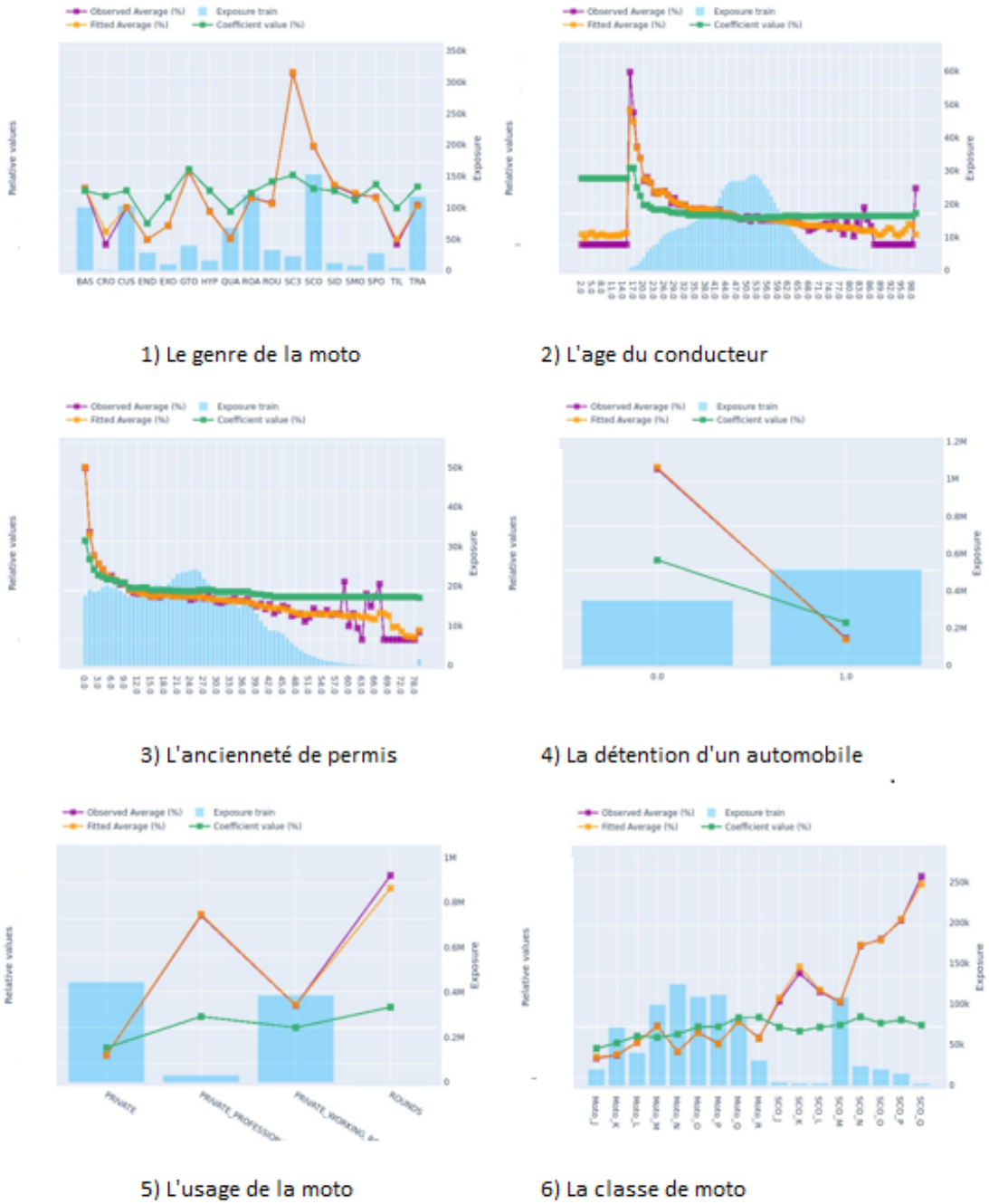
Le tableau suivant résume les variables retenues dans le modèle final, leur granularité initiale et finale, ainsi que leur pouvoir discriminant.

Liste de variables	Descriptif variable	Nom de la variable	Type variable	Granularité initiale	Granularité finale	Importance dans le modèle final
Variables concernant le conducteur	1. L'ancienneté du permis du conducteur	PER_AncPerAn_CP	Quantitative continue	77	21	7
	2. L'âge du conducteur	PER_Age_CP	Quantitative continue	98	26	10
	3. Les antécédents d'assurance	PER_Ant_X_Exp	Qualitative nominale	3	3	11
	4. Le statut marital	BEH_SitMat_CP	Qualitative nominale	7	5	14
Variables concernant le véhicule assuré	1. Le groupe	VEH_Groupe	Qualitative ordinale	17	11	2
	2. L'ancienneté du véhicule	VEH_Age	Quantitative continue	92	25	3
	3. La durée de détention du véhicule actuel	VEH_DurDetAct	Quantitative continue	58	10	4
	4. La carrosserie	VEH_Genre	Qualitative nominale	17	15	5
	5. La classe croisé avec la	VEH_ClasseMotoSco	Qualitative ordinale	17	16	6
	6. L'usage du véhicule	BEH_Usage	Qualitative nominale	4	4	9
	7. Le type de roues	VEH_TypeRoues	Qualitative nominale	3	3	12
	8. La marque	VEH_Marque	Qualitative nominale	17	17	13
Variables géographiques	1. La zone de risque	CDINSEE	Qualitative nominale	33561	33103	1
Variables sinistralité	1. Le score de sinistralité	CLA_Sinistralite	Qualitative nominale	5	5	8
Le produit	1. La formule des garanties	POL_Formule	Qualitative nominale	6	4	15
	2. Le fractionnement	POL_Fract	Qualitative nominale	4	3	17
Le client	1. La clause de multi détention	POL_DetAuto	Qualitative nominale	2	2	16

Tableau 4.2 – Résumé des variables retenues dans le modèle de fréquence des sinistres RC matériels.

Le modèle contient 11 variables qualitatives nominales, 2 variables qualitatives ordinales et 4 variables quantitatives continues. Après simplification avec la méthode de LASSO, la granularité initiale des variables explicatives incluses dans le modèle (hors zonier) a été divisée par 2. Le nombre de coefficients régresseurs final est d'environ 170.

On se propose d'analyser les coefficients obtenus par linéarisation du GBM pour une sélection de variables explicatives, en gardant à l'esprit les contraintes métier et opérationnelles. Cette interprétation des tendances observées et prédites à travers les coefficients régresseurs n'a de sens que toutes choses égales par ailleurs.



Graphique 4.8 – Modélisation de la fréquence des sinistres RC matériels : exemples des paramètres β estimés.

Dans chaque graphique, nous avons représenté les coefficients toutes choses égales par ailleurs (courbe verte), la fréquence des sinistres observée (courbe violette) et la fréquence des sinistres modélisée à l'aide de tous les coefficients régresseurs du modèle (courbe orange). On constate un bon ajustement entre les valeurs observées et modélisées pour toutes les variables analy-

sées.

Le graphique 4.8.1 met en évidence la sur-sinistralité des motos de type grand tourisme et des scooters trois roues. Malgré une forte sinistralité observée sur les scooters deux-roues, le coefficient toutes choses égales par ailleurs pour cette catégorie de deux roues est nul. Le scooter est un moyen de locomotion utilisé notamment dans le milieu urbain. Le zonier segmenté à la maille INSEE étant la première variable la plus discriminante du modèle, la sur-sinistralité du milieu urbain et implicitement des scooters deux-roues est capté donc par le biais de cette variable (d'où le coefficient nul en toutes choses égales par ailleurs).

Comme nous l'avons également constaté lors de l'analyse univarié par variable, la fréquence des sinistres RC matériels est décroissante avec l'âge du conducteur (cf. figure 4.8.2). Un léger écart est constaté entre le risque observé et prédit à partir de l'âge de 70 ans dû à la faible exposition sur cette tranche d'âge.

L'occurrence du risque est aussi décroissante avec l'ancienneté de permis, comme le montre la figure 4.8.3. Les profils ayant une ancienneté de permis supérieure à 43 ans ont été regroupés dans la même catégorie suite à la faible présence de ce type de risque dans le portefeuille d'étude.

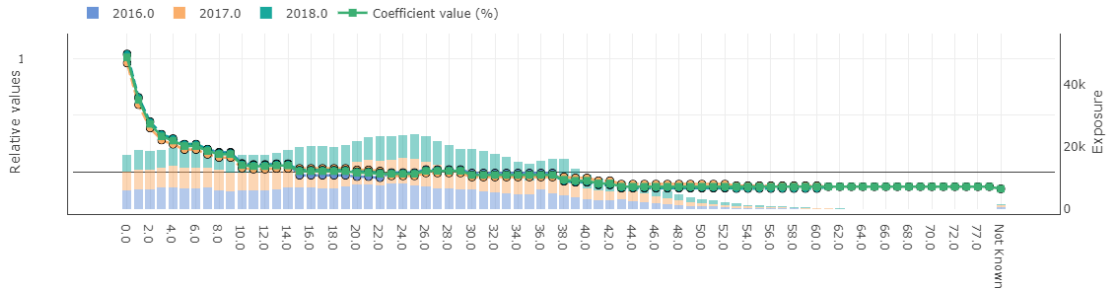
La figure 4.8.4 révèle une sous-sinistralité pour les conducteurs de moto qui possèdent également une automobile. Ce phénomène s'explique par le fait qu'en tant que propriétaires de deux véhicules à usage courant (auto, moto), les assurés utilisent à moindre mesure leur moto en parcourant moins de kilomètres annuellement. En conséquence ils ont une probabilité plus faible d'avoir un sinistre par rapport aux assurés qui ne possèdent qu'un seul véhicule.

De la même manière, les motards qui utilisent leur véhicule pour un usage professionnel ou pour réaliser des trajets quotidiens vers leurs lieu de travail sont susceptibles de parcourir plus de kilomètres par rapport à ceux qui utilisent leur deux-roues pour un usage privé. En conséquence, l'occurrence du risque est plus élevée pour ces deux types d'usage.

Enfin, on constate une tendance croissante de la fréquence des sinistres en fonction de la classe de moto. Le véhiculier moto étant différent du véhiculier scooter, nous avons donc croisé la classe avec la catégorie de moto, mais la tendance croissante de la sinistralité en fonction de la classe est respectée pour les deux types de véhicules analysés. Cependant, la pente des coefficients toutes choses égales par ailleurs est plus élevée pour les motos standards que pour les scooters (ce qui indique un plus fort pouvoir discriminant du véhiculier moto standard).

Consistance temporelle

Pour chaque variable intégrée dans le modèle nous avons vérifié si la tendance moyenne observée et modélisée est consistante dans le temps. A titre d'exemple, nous avons représenté l'outil graphique permettant de vérifier la consistance temporelle pour la variable ancienneté de permis (cf. fig. 4.9). Les histogrammes représentent l'exposition pour chaque année d'analyse pour une modalité de la variable étudiée, les fréquences de sinistres observées par année d'analyse et le risque modélisé étant représentés par des courbes. On ne constate pas d'écart systématique entre les quatre courbes donc la tendance modélisée pour la variable ancienneté de permis est consistante dans le temps.



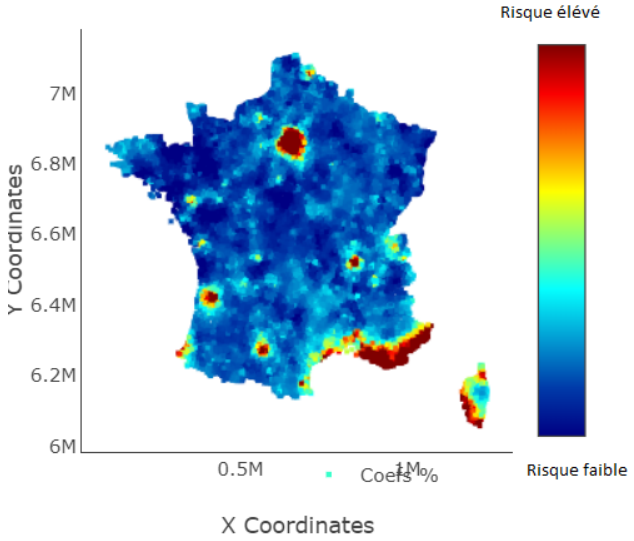
Graphique 4.9 – Modélisation de la fréquence des sinistres RC matériels : consistance temporelle pour la variable ancienneté de permis.

Nous réalisons la même analyse graphique sur les autres régresseurs du modèle. Les tendances des fréquences selon chaque variable sont bien consistantes dans le temps sur les trois années d'étude 2016 à 2018.

Construction d'un zonier avec GBM

Une attention particulière doit être accordé au nouveau zonier construit lors de cette modélisation de la fréquence des sinistres RC matériels. Le logiciel de modélisation développé par AXA Global P&C permet de construire un nouveau zonier de manière automatisée. La méthode mise en place est similaire à celle utilisée pour la recherche des interactions.

Concrètement, pour construire ce nouveau zonier, le modèle initial (sans variables de zonage) est d'abord mis en offset. Puis un nouveau modèle GBM est entrainé qui cherche à prédire les quasi-résidus avec la variable de zonage choisie pour construire le zonier, dans notre cas, à la maille INSEE.



Graphique 4.10 – Zonier de fréquence pour la garantie RC matérielle.

Comme précisé plus haut, pour chaque garantie nous avons testé deux modèles : un avec le zonier actuel et un deuxième avec le nouveau zonier, en choisissant comme modèle final celui avec le meilleur indice de Gini. Dans la figure 4.10 nous avons représenté les coefficients toutes choses égales par ailleurs en fonction de la zone de risque. Ainsi, on constate une surfréquence des sinistres responsabilité civile en région Ile-de-France, dans le sud-ouest de la France (Alpes-Maritimes, le Var et les Bouches du Rhône), ainsi que dans la Corse du Sud. On observe également que l'occurrence du risque est plus élevée autour des grandes villes (Bordeaux, Toulouse, Lyon). Afin de vérifier la consistance du nouveau zonier bâti, nous avons représenté la sinistralité pour chacune des trois années d'exposition. Le nouveau zonier est bien consistant dans le temps, les tendances étant respectées pour les trois années d'observation.

La performance et la validation du modèle

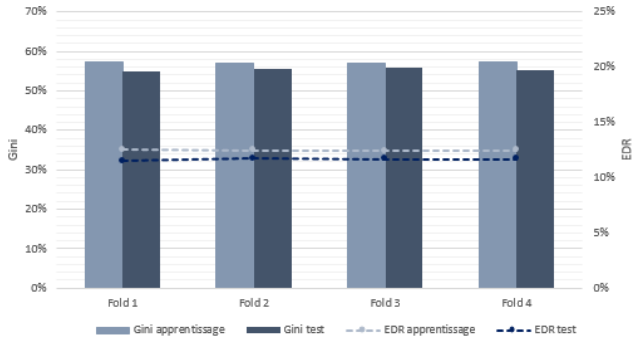
La performance du modèle est évaluée à l'aide de deux indicateurs : le Gini et l'EDR , définis dans la section 2.2.1.

Indicateur	échantillon d'apprentissage	échantillon de test
Fréquence RC MAT moyenne observée	2,06%	2,05%
Fréquence RC MAT moyenne prédite	2,06%	2,05%
Indice de Gini	57,18%	55,44%
EDR	12,46%	11,63%

Tableau 4.3 – Principaux indicateurs de performance du modèle.

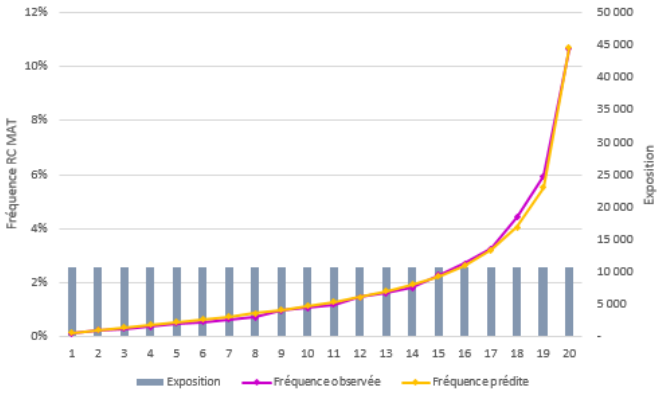
Les indicateurs ont été calculés sur les échantillons d'apprentissage et de test comme moyennes obtenues sur les k-folds. Etant donné que le modèle a été bâti sur la base d'apprentissage, il est normal d'obtenir des résultats légèrement meilleurs sur cette base, par rapport à l'échantillon de validation.

La validation du modèle a été réalisée à travers la méthode de validation croisée (k-fold) et à travers la courbe de lift, définis dans la section 2.2.1. La k-fold cross validation donne des résultats concluants puisque les indicateurs de Gini et les EDR sur la base d'apprentissage et de test sont stables au cours des 4 cross validations, comme le montre le graphique 4.11. Le modèle final retenu est obtenu en calculant la moyenne des prédicteurs sur les 4 folds d'apprentissage.



Graphique 4.11 – Validation du modèle par la méthode de validation croisée.

La deuxième méthode de validation utilisée est la validation par quantiles, à travers la courbe de lift. Comme précisé dans la section 2.2.1, cette méthode consiste à trier la base de données par ordre croissant de la fréquence des sinistres prédite et de diviser ensuite la base de données en tranches égales d'exposition. En dernière instance on va tracer les fréquences de sinistres observées et prédites pour chaque quantile. Par construction la fréquence prédite est croissante. Si le modèle est performant, l'observé doit également être croissant suivant les quantiles et doit se rapprocher le plus de la valeur prédite.



Graphique 4.12 – Courbe de lift construite sur la base de validation.

Nous avons choisi de réaliser 20 quantiles. Malgré une légère sous-estimation sur les segments n° 18 et 19, la validation est satisfaisante puisque la fréquence des sinistres RC matériels prédite est globalement proche du risque observé sur la base de validation.

4.2.3 Modélisation du coût moyen de la garantie RC matérielle

4.2.3.1 Collecte et préparation des données

Périmètre d'étude et vision de sinistralité

Si dans la partie précédente nous avons exposé toutes les étapes de construction du modèle de fréquence des sinistres RC matériels, on s'intéresse maintenant à la modélisation du coût

moyen des sinistres de même nature sur le portefeuille deux-roues.

Contrairement à la modélisation de la fréquence des sinistres dont l'analyse était basée sur la sinistralité survenue sur le portefeuille pendant la période 2016 à 2018, pour la modélisation du coût moyen nous avons fait le choix de modifier le périmètre d'étude. Ainsi, cette partie de l'étude porte sur la sinistralité survenue entre les années 2015 et 2017, les sinistres étant vus à fin octobre 2019, comme dans le modèle de fréquence. Ce choix est justifié par la nécessité d'avoir une période suffisante de développement pour les charges de sinistres, afin qu'elles soient le plus rapprochés de la charge ultime. Ce choix est d'autant plus important sur les sinistres corporels car ce sont des sinistres d'intensité avec une période de liquidation plus longue que les sinistres de fréquence (tels que les sinistres dommages, incendie, vol etc.).

Nous avons donc pris en compte la vision de sinistralité la plus récente au moment de l'étude afin de ne pas omettre les sinistres tardifs et d'avoir la vision la plus proche de l'ultime.

Nous conservons tous les contrats qui ont été en cours au moins un jour pendant la période analysée et qui ont eu au moins un sinistre RC matériel survenu entre les années 2015 et 2017. La responsabilité civile étant la seule garantie obligatoire, nous gardons dans le périmètre tous les contrats, quel que soit la formule de garanties souscrite. Nous travaillons sur les contrats gérés par les agents généraux et les courtiers, pour les cinq régions de souscription d'AXA France.

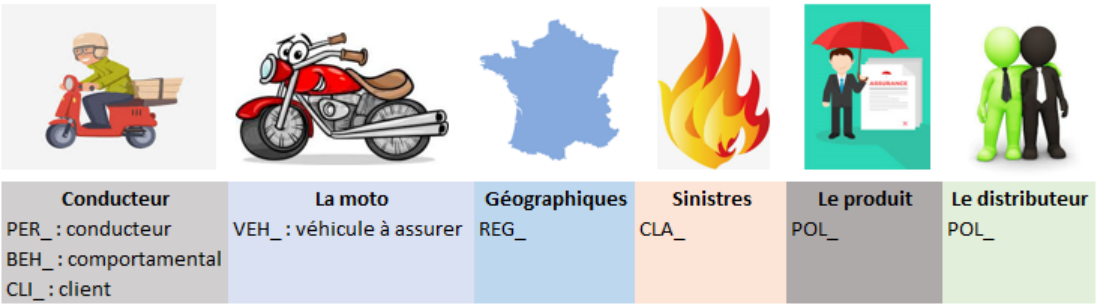
Construction de la base de données

Les données chiffrées précisées ci-après concernent uniquement les contrats deux-roues ayant subi un sinistre RC matériel pendant la période 2015 à 2017 mais la constitution de la base de modélisation du coût moyen pour toutes les autres garanties est similaire.

La base de données relative au coût moyen de la garantie RC matérielle, représente environ 11.400 lignes, chaque ligne correspondant à une image. Concrètement, afin d'obtenir la base de données utilisée pour la modélisation du coût moyen, le point de départ est représenté par la base totale par image sur la période considérée. Ensuite les lignes (les images) ayant un nombre de sinistres nul ont été écartées.

Chaque image est considérée comme une observation à part entière dans nos modélisations. Pour chaque ligné gardée dans la base de données, le coût moyen est obtenu en divisant la charge de sinistres par le nombre de sinistres, et celle-ci garantie par garantie. Ainsi le 11.400 images (lignes) représentent environ 11.700 sinistres. De la même manière que l'analyse de la fréquence des sinistres, la base de données utilisée pour la modélisation du coût moyen contient plus de 180 variables, que l'on peut classer comme suit :

- Variables concernant le conducteur principal
- Variables concernant le véhicule à assurer
- Variables géographiques
- Variables concernant la sinistralité déclarée à la souscription, ou la sinistralité survenue du contrat AXA (par nature de sinistre)
- Variables concernant le produit
- Variables concernant le distributeur



Graphique 4.13 – Types de variables à disposition.

Nettoyage de données

Pour la construction de la base finale de modélisation, un nettoyage de la base initiale constitué est réalisé en amont en dimension verticale et horizontale.

Suppression des colonnes sans intérêt préalable

Etant donné la multitude de facteurs de risque inclus dans la base de données, on exclut d'abord des nombreuses variables qui nous paraissent a priori n'avoir aucun pouvoir explicatif tels que des variables de type identificateur (le numéro de sinistre, le numéro de client, le numéro de contrat etc.), des variables présentant trop de valeurs manquantes, des variables concernant les cotisations par type de garantie etc. On exclut également le nombre de sinistres pour les garanties non modélisés, car lors de la modélisation du coût moyen de la garantie RC matérielle ces variables présentent peu d'intérêt (par exemple, le nombre de sinistres dommages n'a pas d'influence a priori sur le coût moyen RC matérielle). Après ces exclusions, nous obtenons une base de 60 variables qui servira à la modélisation du coût moyen.

Suppression des lignes

Les lignes dont l'exposition est nulle ont été écartées de l'analyse.

On écarte également les contrats temporaires de l'analyse, car ils constituent un risque à part, et leur présence dans le périmètre d'analyse pourrait biaiser les résultats.

De même, avant de réaliser la jointure entre la base de contrats et la base de sinistres, les sinistres annulés [2] et les sans suite [3] ont été ôtés de la base. Parmi ces sinistres, certaines catégories ont également été exclues de l'analyse :

- Les sinistres de montant négatif ou nul : ils correspondent aux situations où la responsabilité de notre assuré n'a pas été engagée lors d'un accident de la circulation avec un tiers identifié. En l'absence de garanties dommages souscrites (assurance au tiers), le gestionnaire enregistre alors le sinistre au titre de la garantie RC, un recours étant ensuite exercé auprès de la compagnie d'assurance du tiers identifié.

- Les sinistres RC matériels dont le coût correspond au forfait d'ouverture ont été exclus de l'analyse, car pour ces sinistres nous n'avons aucune visibilité sur leur vrai coût ultime. Ils représentent moins de 1% du nombre total de sinistres.

- Les sinistres au coût conventionnel : ils ont été écartés de l'analyse car leur coût final est forfaitaire. Il s'agit de sinistres pour lesquels la responsabilité de notre assuré est engagée (totalement ou partiellement) et qui vérifient les conditions d'application de la convention IRSA.

2. Il s'agit de sinistres clos dont toutes les composantes du coût sont nulles
3. Il s'agit de sinistres de coût non nul mais qui n'entrainent pas de garantie. Principalement les charges entrainées représentent des frais d'expertise médicale. Afin de ne pas biaiser l'analyse, ils ont été supprimés de la modélisation de coût moyen, mais ils sont pris en compte dans le modèle de fréquence

- Les sinistres dits sinistres graves : pour cette étude nous avons fixé un seuil de 10.000€ d'après une analyse préalable de la distribution de sinistres RC matériels. Les sinistres supérieurs à ce seuil sont considérés comme des sinistres atypiques. Pour ne pas biaiser la modélisation, ces sinistres plus volatiles ont été écartés de l'analyse. Ils représentent 0,5% du nombre total de sinistres et 6,5% de la charge totale.

Il est rappelé que pour mener à bien les travaux qui visent à déterminer les critères tarifaires et les échelles de risque associées, il n'est pas indispensable de disposer d'une vision à l'ultime des sinistres, ni même de considérer tous les sinistres. C'est au cours des travaux relatifs au calibrage de la prime pure de référence que ces dimensions devront être prises en compte afin d'évaluer au plus juste l'ensemble des engagements purs de l'assureur sur l'exercice. Nous allons revenir sur ces aspects dans la section 4.3.

Sur ce périmètre, on dénombre environ 11.700 sinistres RC matériels représentant une charge de sinistres totale de plus de 18 M€, le coût moyen observé étant de 1.546 €.

Point d'attention concernant la modélisation du coût moyen des sinistres RC

Il est important de préciser que, compte tenu du caractère singulier des sinistres RC, une modélisation segmentée du coût moyen apparaît, au premier abord, difficile. En effet, contrairement aux autres garanties modélisées dont le coût de sinistres est quasi-exclusivement lié aux caractéristiques du véhicule assuré, le coût d'un sinistre RC dépend également des caractéristiques du véhicule adverse (dans le cas d'un sinistre RC matériel) ou des caractéristiques des tiers personnes impliquées dans l'accident (dans le cas d'un sinistre RC corporel). Ainsi, on peut penser, au premier abord, que les critères impactant le plus le coût moyen d'un sinistre RC sont :

- Les caractéristiques du véhicule adverse : par exemple, le coût sera d'autant plus important que la valeur du véhicule adverse est élevée
- Le nombre de véhicules impliqués dans l'accident : le coût sera, a priori, d'autant plus important que le nombre de véhicules impliqués est élevé
- La nature du choc et la vitesse d'impact : un sinistre sera d'autant plus important que la vitesse d'impact est grande
- La présence des dommages corporels et le nombre de victimes impactées
- La gravité des dommages corporels, l'âge des victimes et leur typologie (motocycliste, piéton, conducteur d'une automobile etc.)

Du fait de l'absence des informations concernant la partie adverse dans l'entrepôt de données, nous allons segmenter le coût moyen selon les variables disponibles (relatives au conducteur et au véhicule assuré).

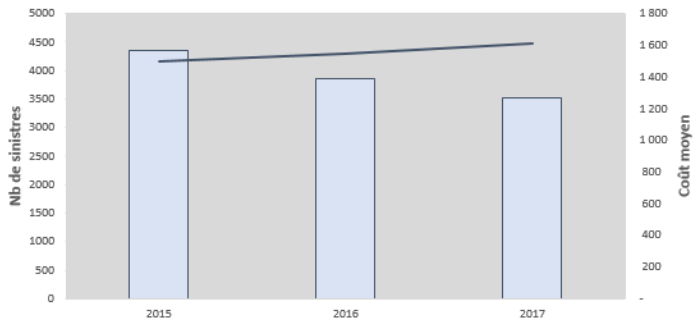
4.2.3.2 Analyse exploratoire des variables

Avant de commencer la modélisation du coût moyen des sinistres RC matériels une analyse exploratoire des facteurs de risque s'impose. Cette étape permet notamment de s'approprier les données, de détecter d'éventuelles sources futures d'instabilité des résultats mais également d'identifier en amont les critères non liés à la variable réponse étudiée.

Nous allons d'abord apporter quelques spécifications sur la variable que l'on souhaite modéliser (le coût moyen des sinistres), et ensuite présenter les analyses descriptives des certains facteurs de risque.

Le coût moyen des sinistres RC matériels

La base utilisée, obtenue par jointure entre les différentes sources mentionnées précédemment, est constituée d'autant de lignes que d'images ayant subi au moins un sinistre pendant la période considérée. Pour chaque image le nombre de sinistres correspond au nombre de sinistres survenus au cours de la période intra-annuelle de chaque année étudiée où le contrat était en vigueur.



Graphique 4.14 – Coût moyen sinistre RC matériel observé par année d'étude.

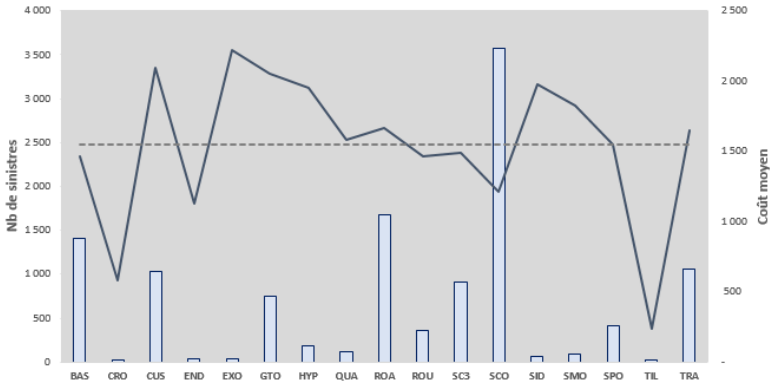
On constate, d'après la figure 4.14, que le coût moyen empirique de la garantie RC matérielle est relativement stable sur les trois années d'étude, avec une tendance légèrement croissante, se situant à un niveau d'environ 1.550€. On remarque également la tendance à la baisse du nombre de sinistres RC matériels d'une année à l'autre.

Analyse univariée des variables classiques

Nous allons par la suite représenter graphiquement l'évolution du coût moyen en fonction de quelques facteurs de risque. Dans chaque graphique, la courbe du coût moyen observé est accompagnée d'un histogramme représentant le nombre de sinistres de chaque modalité sur la période analysée. Pour faciliter l'analyse nous avons rajouté également le coût moyen empirique de la sinistralité analysée (en pointillé). Cette étape nous permettra de :

- Observer l'effet a priori de chaque facteur de risque sur la variable réponse étudiée, en l'occurrence le coût moyen des sinistres RC matériels
- Identifier les critères dont une seule modalité concentre la quasi-totalité de sinistres survenus. Cet aspect traduit un faible pouvoir discriminant de ces critères dans la modélisation du coût moyen.
- Détecter d'éventuels sources pouvant perturber la modélisation (des modalités dont le nombre de sinistres survenus est moindre)

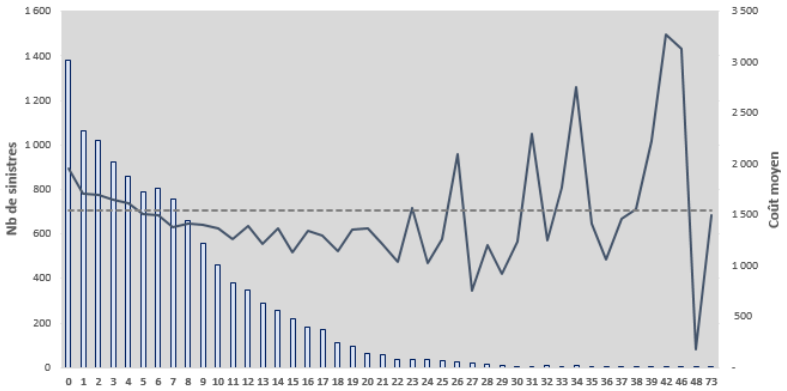
Le genre de la moto



Graphique 4.15 – Coût moyen observé des sinistres RC matériels par genre de moto.

La figure 4.15 fait apparaître des disparités concernant le coût moyen des sinistres RC matériels en fonction de la catégorie de moto. Ainsi, on constate que le coût moyen est plus important pour les exotiques (EXO), les super motards (SMO) et les hyper sportives (HYP). Dérivées des modèles de compétition, ces catégories de motos peuvent facilement attendre des niveaux de vitesse très élevés générant, dans le cas d'un sinistre, un fort impact avec le véhicule adverse, ce qui explique le coût plus conséquent du sinistre. Les motos de type grand tourisme et les motos customisés (CUS) génèrent également des sinistres dont le coût moyen est nettement supérieur à la moyenne. Ce phénomène est lié notamment au type de conduite de cette catégorie de motos. Etant des motos lourdes, peu maniables, les GTO et Customs génèrent plutôt des chocs latéraux qui sont plus ravageurs que les chocs arrière. A l'inverse, le coût d'un sinistre RC matériel généré par un scooter (SCO) est moins important que la moyenne empirique. Celle-ci est a priori due au type de conduite. Ainsi, les scooters sont utilisés notamment dans le milieu urbain, sur des voies complétement saturées, et donc avec une vitesse limitée (certes ce phénomène de conduite en ville génère un nombre important de sinistre, mais de coût moindre). La même situation est constatée sur les motos de type basique. Elles sont généralement des motos de puissance faible et leur faible protection n'incite pas les motards à rouler vite, d'où le faible impact en termes de coût moyen des sinistres RC matériels.

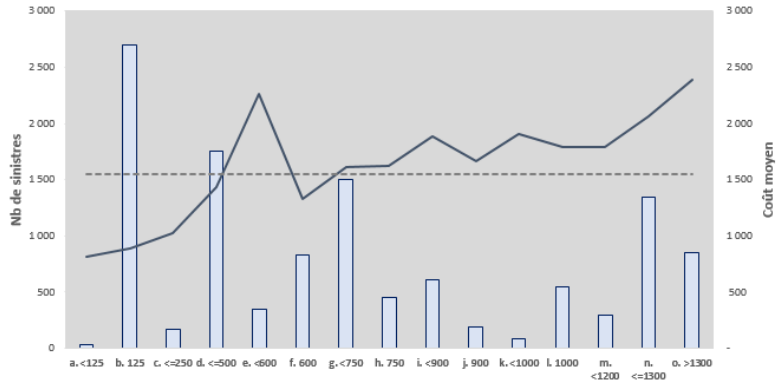
L'ancienneté du véhicule



Graphique 4.16 – Coût moyen observé des sinistres RC matériels par ancienneté du véhicule.

La figure 4.16 fait apparaître la relation négative entre l'ancienneté de la moto et le coût moyen de la garantie RC matérielle. Bien que, au premier abord, on peut penser que l'ancienneté du véhicule assuré n'a aucun impact sur le montant total d'un sinistré causé à autrui, empiriquement on constate que l'impact de ce facteur de risque est non négligeable. Celle-ci peut s'expliquer par le fait que, avec les nouvelles technologies, les motos sont devenues de plus en plus puissantes avec le temps, les deux-roues neuves générant des chocs plus importants que leurs homologues plus anciennes.

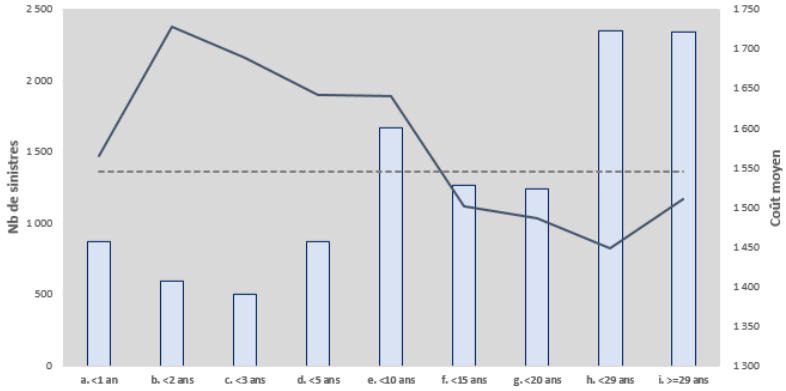
La cylindrée de la moto



Graphique 4.17 – Coût moyen observé des sinistres RC matériels par cylindrée.

Dans la figure 4.17 on constate la tendance croissante du coût moyen des sinistres RC matériels en fonction de la cylindrée du véhicule. Traduisant la puissance d'un deux-roues, le choc avec un tiers est d'autant plus important que la cylindrée est grande. Une exception est cependant constatée sur les véhicules de cylindrée comprise entre 500 et 600 cm^3. Il s'avère que sur ce segment 91% de sinistres constatés sont générés par des scooters et les scooters de cylindrée élevé ont nativement des coûts moyens des sinistres RC matériels plus élevés que les motos, ce qui explique donc le pic observé sur ce segment un peu atypique.

L'ancienneté de permis du conducteur



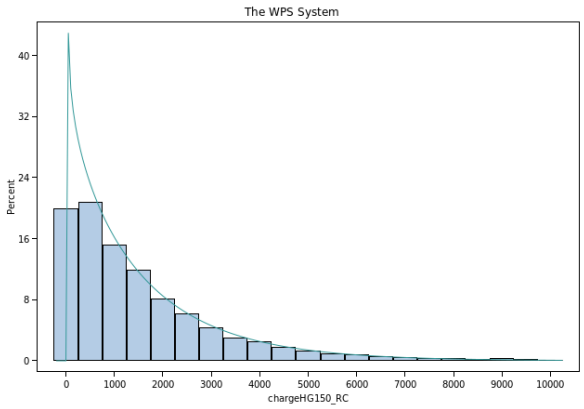
Graphique 4.18 – Coût moyen observé des sinistres RC matériels par ancienneté de permis.

Dans la figure 4.18 on constate la tendance décroissante du coût moyen des sinistres RC matériels en fonction de l'ancienneté de permis. Avec l'expérience les conducteurs sont d'une part moins incités à faire des excès de vitesse et d'autre part ils ont une meilleure maîtrise de leur conduite et de leur véhicule et, en cas d'accident, ils arrivent à éviter les chocs importants ce qui explique la relation négative entre l'ancienneté de permis et le coût d'un sinistre RC. Si la tendance n'est pas respectée pour les jeunes permis de moins d'un an d'ancienneté, celle-ci est dûe au bridage des motos pour les conducteurs sans expérience de conduite (la puissance ne pouvant pas excéder 35 kW).

4.2.3.3 Modèle final

Après avoir traité et nettoyé les données pour qu'elles soient exploitables, puis les avoir explorés à travers des analyses descriptives univariées, le processus de modélisation par linéarisation du gradient boosting est mis en place.

Nous souhaitons par la suite construire un modèle qui permette de prédire le coût moyen des sinistres RC matériels. La distribution de sinistres en fonction de leur coût étant représentée dans le graphique suivant :



Graphique 4.19 – Distribution du coût moyen des sinistres RC matériels.

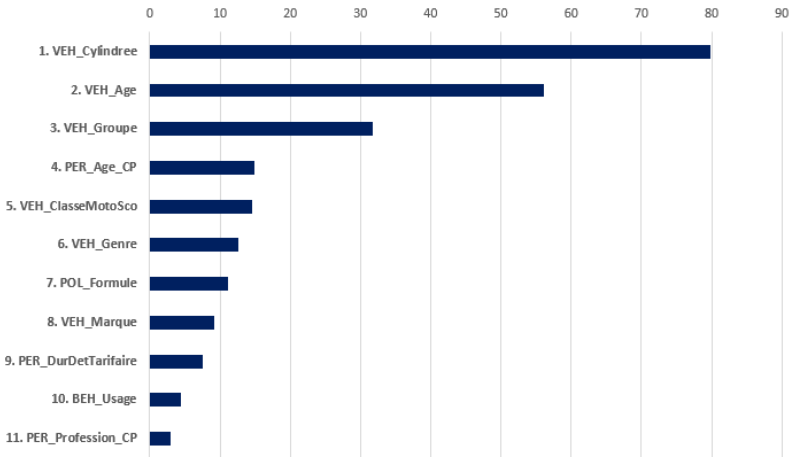
Sur la période considérée, la majorité de sinistres ont un coût inférieur à 2.000€ (49% de sinistres ayant un coût inférieur à 1.000€ et 23% de sinistres de montant situé entre 1.000 et 2.000€). Cette distribution présente une asymétrie à droite modérée (skewness égale à 1.9) dûe à la présence de sinistres avec un coût supérieur à la moyenne empirique.

La coût empirique moyen est de 1.546€ pour un écart type de 1.682€ ce qui indique une sur dispersion caractéristique d'un portefeuille hétérogène (justifiant de cette manière la segmentation a priori). Pour la modélisation du coût moyen, la loi de Gamma apparaît comme la plus légitime. Pour s'en convaincre, un test d'adéquation des données à une loi de Gamma de paramètre de forme $\alpha = 0.87$ et de paramètre d'échelle $\sigma = 0.0005$ a conduit, au seuil de 5%, à l'acceptation de l'hypothèse nulle d'adéquation. C'est donc cette distribution qui a été utilisée comme fonction de lien pour la modélisation du coût moyen des sinistres.

Pour la validation du modèle, nous avons utilisé la technique de validation croisée (4-fold). La base initiale a été divisé en deux partitions : la base d'apprentissage et la base de validation, en proportion de 80-20. Ensuite, la première base a été divisée de façon aléatoire en 4 échantillons indépendants de taille égale. Le modèle est alors bâti sur l'ensemble de 3 premières bases (folds), puis sa performance est évaluée sur la 4-ème base, la base de validation, et ainsi de suite.

Variables sélectionnées dans le modèle final

Parmi les 60 variables présentes dans la base de modélisation, seulement 11 ont été choisies dans le modèle final : quatre variables numériques et sept variables catégorielles.



Graphique 4.20 – Influence relative des variables sur le coût moyen des sinistres RC matériels (exprimé en %).

Le graphique ci-dessous illustre l'influence relative de chacun des régresseurs sur la variable à expliquer, le coût moyen des sinistres RC matériels. Cette influence est mesurée par le biais du spread 95/5, indicateur défini dans la section 3.3.

On constate que la variable la plus discriminante du modèle de prédiction du coût moyen des sinistres est la cylindrée. Elle est suivie par l'ancienneté du véhicule en deuxième position et le groupe de la moto en troisième position. Les 8 facteurs de risque suivant ont un spread 95/5 proche de 10%, leur influence relative par rapport à la variable réponse étant moins forte. Les critères qui impactent le plus le phénomène analysé sont des facteurs de risque comportementaux, qui traduisent notamment soit la puissance du véhicule (la vitesse de conduite) soit l'expérience de conduite (dans le cas d'un sinistre, un conducteur expérimenté arrive à mieux gérer le choc causé à la partie adverse).

Il est important de préciser que dans le cadre de cette analyse, pour chaque type de garantie modélisée, deux modèles de coût moyen des sinistres ont été réalisés : un premier modèle incluant la zone de risque utilisée actuellement dans le tarif et un deuxième modèle dans lequel nous avons bâti un nouveau zonier de coût moyen en fonction du code INSEE du lieu de garage.

Naturellement, lors de la construction d'un nouveau zonier, le choix est porté par la construction d'un zonier de fréquence plutôt que sur la construction d'un zonier de coût moyen. Les coûts moyens sont plus volatiles que la fréquence et sont sujets à l'inflation, ce qui nécessite une actualisation plus fréquente du zonier. De plus, pour que le nouveau zonier soit robuste et stable dans le temps, il est primordial d'avoir un nombre de sinistre suffisant pour toutes les modalités de la maille de zonage souhaitée.

Construction d'un nouveau zonier de coût moyen

Puisque le logiciel développé par AXA Global P&C permet de bâtir un nouveau zonier de manière automatisée et assez rapide, nous avons testé l'élaboration d'un zonier de coût moyen pour la garantie RC matérielle. Le zonier ainsi obtenu n'est pas stable dans le temps et son inclusion dans le modèle final ne permet pas d'augmenter les performances du modèle.

Nous avons également testé l'inclusion du zonier actuel dans le modèle de coût moyen mais, une fois de plus, les résultats obtenus ne sont pas satisfaisants. Les variables géographiques ne feront pas partie des critères retenus lors de la modélisation du coût moyen.

Coefficients et interprétation

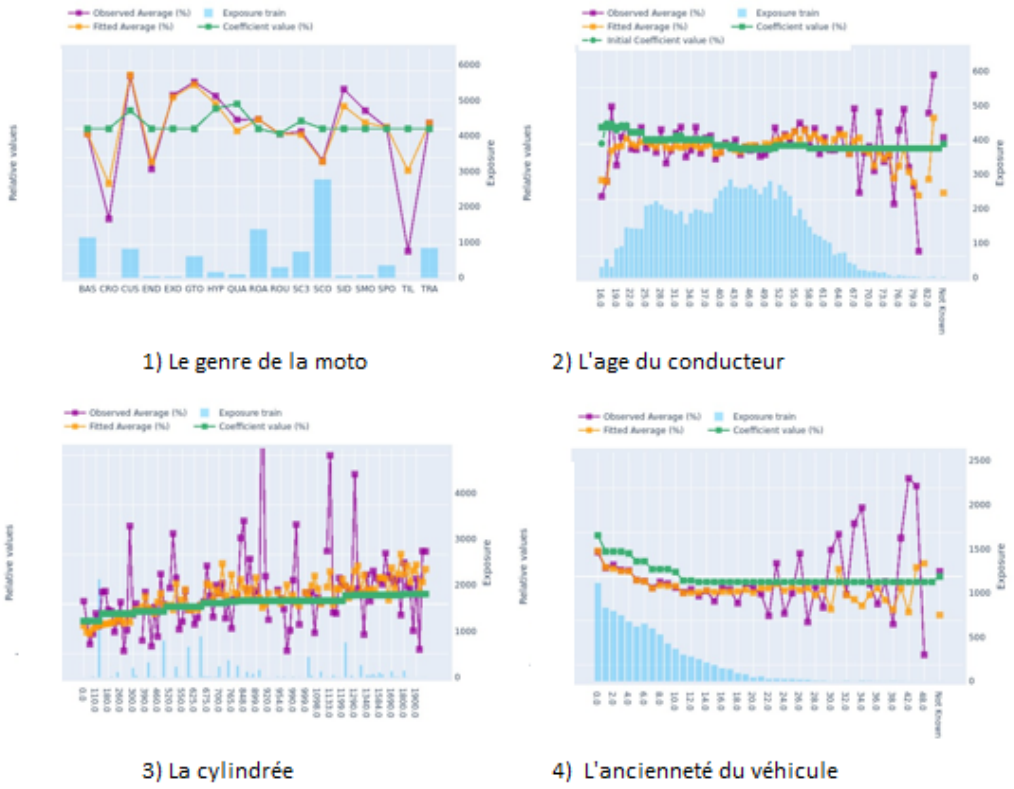
Une fois les variables sélectionnées, un nouveau GBM est entraîné sur cette sélection de onze facteurs de risque, puis la transformation du GBM en GLM est opérée : on obtient alors un jeu de coefficients pour chacun des régresseurs que nous allons analyser et interpréter avec un regard "métier" par la suite. Le tableau suivant résume les variables retenues dans le modèle final, leur granularité initiale et finale, ainsi que leur pouvoir discriminant.

Liste de variables	Descriptif variable	Nom de la variable	Type variable	Granularité initiale	Granularité finale	Importance dans le modèle final
Variables concernant l'assuré	1. L'âge du conducteur	PER_Age_CP	Quantitative continue	70	8	4
	2. La profession du conducteur	PER_Profession_CP	Qualitative nominale	8	4	11
Variables concernant le véhicule assuré	1. La cylindrée	VEH_Cylindrée	Quantitative continue	112	12	1
	2. L'ancienneté du véhicule	VEH_Age	Quantitative continue	45	10	2
	3. La groupe	VEH_Groupe	Qualitative ordinale	16	6	3
	4. La classe de moto/sco	VEH_ClasseMotoSco	Qualitative ordinale	17	10	5
	5. La carrosserie	VEH_Genre	Qualitative nominale	17	6	6
	6. La marque	VEH_Marque	Qualitative nominale	17	12	8
	7. La durée de détention	VEH_DurDetTarifaire	Quantitative continue	33	3	9
	8. L'usage de la moto	BEH_Usage	Qualitative nominale	4	4	10
Le produit	1. La formule des garanties	POL_Formule	Qualitative nominale	6	4	7

Tableau 4.4 – Résumé des variables retenues dans le modèle final.

Le modèle contient 5 variables qualitatives nominales, 2 variables qualitatives ordinales et 4 variables quantitatives continues. Après simplification, la granularité initiale des variables explicatives incluses dans le modèle a été divisée par 4, le nombre de coefficients régresseurs étant à environ 80.

On se propose d'analyser les coefficients obtenus par linéarisation du GBM pour une sélection des variables explicatives, en gardant à l'esprit les contraintes métier et opérationnelles. Cette interprétation des tendances observées et prédites à travers les coefficients régresseurs n'a de sens que toutes choses égales par ailleurs.



Graphique 4.21 – Modélisation du coût moyen des sinistres RC matériels : exemples des paramètres β estimés.

Dans chaque graphique, nous avons représenté les coefficients toutes choses égales par ailleurs (courbe verte), le coût moyen des sinistres observé (courbe violette) et le coût moyen des sinistres modélisé (courbe orange). Globalement on constate un bon ajustement entre les valeurs observées et modélisées pour toutes les variables analysées (sauf pour les modalités dont le nombre de sinistres est trop faible).

Le graphique 4.21.1 met en évidence les écarts de coût moyen en fonction de la catégorie de la moto. On constate que, toutes choses égales par ailleurs, les motos customisées (CUS) et les hyper-sportives (HYP) sont responsables des sinistres matériels de montant le plus élevé. Comme vu dans la partie d'analyse exploratoire des variables, cet aspect est lié au rapport poids/puissance de ce type de motos. Comme nous l'avons également remarqué lors de l'analyse univariée de cette variable, les motos EXO tiques (EXO) et les grand tourisme (GTO) ont également des coûts de sinistres empiriques importants. Cependant, lors de l'analyse multivariée, on constate que le coefficient estimé pour ces deux catégories de moto est nul. L'effet est capté sûrement par la cylindrée, la variable la plus discriminante du modèle (car, malgré un paramètre estimé nul pour ces deux modalités, on constate un bon ajustement entre l'observé et le modélisé).

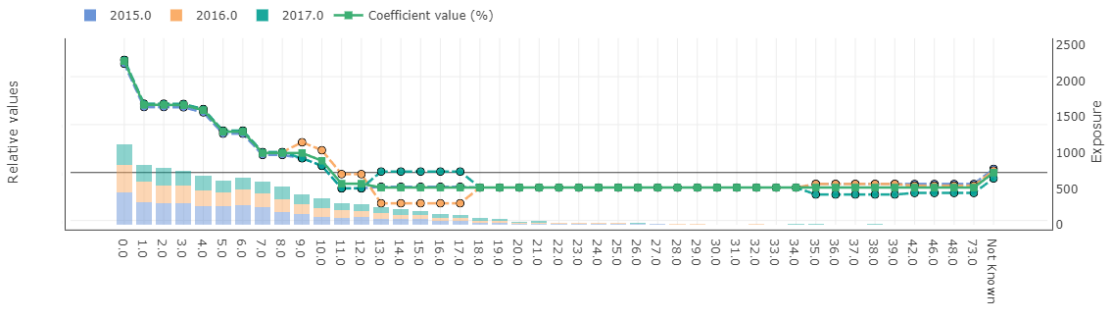
Le coût moyen des sinistres RC matériels est décroissant avec l'ancienneté du conducteur (cf. fig. 4.21.2). D'abord parce que les jeunes conducteurs ont tendance à rouler vite, générant des chocs plus importants mais aussi parce que leur comportement en cas de sinistre est différent par rapport à celui d'un motard expérimenté. Dans le cas d'un sinistre, un conducteur expérimenté arrive à mieux gérer le choc causé à la partie adverse, générant moins de dégâts.

Dans la figure 4.21.3 on remarque la relation positive, croissante, entre la cylindrée de la moto et le coût moyen d'un sinistre RC matériel. Directement lié à la puissance de la moto, le choc avec un tiers est d'autant plus important que la cylindrée de la moto impliquée dans l'accident est grande.

La figure 4.21.4 confirme la tendance remarquée lors de l'analyse univariée : le coût moyen de la garantie RC matérielle est décroissant avec l'ancienneté du véhicule. Avec les nouvelles technologies, les motos sont devenues de plus en plus puissantes avec le temps, les deux-roues neuves générant des chocs plus importants que leurs homologues plus anciennes.

Consistance temporelle

De la même manière que dans le modèle de fréquence, pour chaque variable intégrée dans le modèle de coût moyen nous avons vérifié si la tendance moyenne observée et modélisée est consistante dans le temps. A titre d'exemple, nous avons représenté l'outil graphique permettant de vérifier la consistance temporelle pour la variable ancienneté du véhicule (cf. fig. 4.22). Les histogrammes représentent le nombre de sinistres pour chaque année d'analyse pour une modalité de la variable étudiée, le coût moyen empirique par année d'analyse et le risque modélisé étant représentés par des courbes. On ne constate pas d'écart systématique entre les quatre courbes. Les seuls écarts constatés sont pour les motos d'ancienneté comprise entre 13 et 17 ans, pour les années de survenance 2016 et 2017. Cependant, lorsque ces écarts ne sont pas observés de manière systématique, on peut considérer que la tendance modélisée pour la variable ancienneté du véhicule est consistante dans le temps.



Graphique 4.22 – Modélisation du coût moyen de la garantie RC matérielle : consistence temporelle pour la variable ancienneté du véhicule.

Nous réalisons la même analyse graphique sur les autres régresseurs du modèle. Les tendances du coût moyen selon chaque variable sont bien consistantes dans le temps sur les trois années d'étude 2015 à 2017.

La performance et la validation du modèle

La performance du modèle est évaluée à l'aide de deux indicateurs : le Gini et l'EDR , définis dans la section 2.2.1.

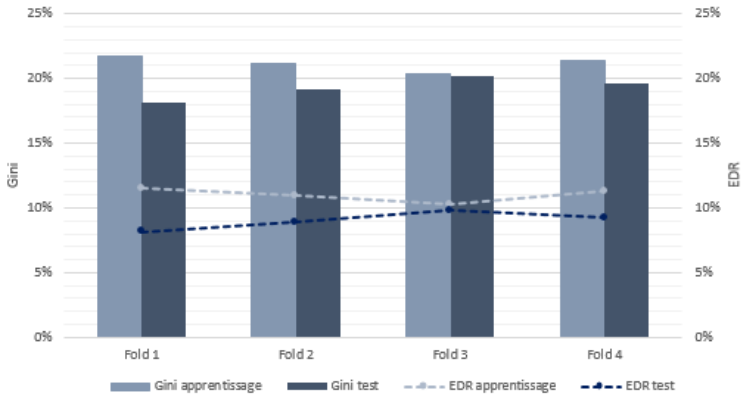
Indicateur	échantillon d'apprentissage	échantillon de test	Base totale
Coût moyen RC MAT observé	1 549 €	1 534 €	1 546 €
Coût moyen RC MAT modélisé	1 538 €	1 537 €	1 538 €
Indice de Gini	20,86%	18,46%	20,38%
EDR	10,67%	19,28%	10,06%

Tableau 4.5 – Modélisation du coût moyen de la garantie RC matérielle : principaux indicateurs de performance du modèle.

Les indicateurs ont été calculés sur les échantillons d'apprentissage et de test comme moyennes obtenus sur les k-folds. Etant donné que le modèle a été bâti sur la base d'apprentissage, il est normal d'obtenir des résultats légèrement meilleurs sur cette base, par rapport à l'échantillon de validation. En appliquant le modèle sur la base totale les résultats sont aussi satisfaisants.

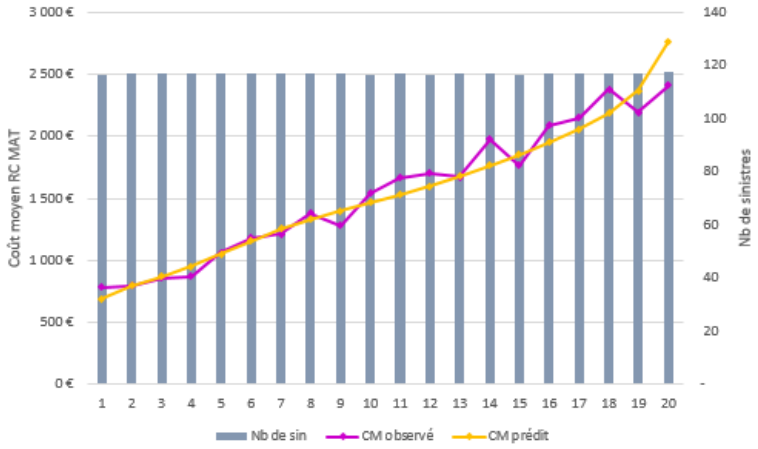
La validation du modèle a été réalisée à travers la méthode de validation croisée (k-fold) et à travers la courbe de lift, définis dans la section 2.2.1.

La k-fold cross validation donne des résultats concluants puisque les indicateurs de Gini et les EDR sur la base d'apprentissage et de test sont stables au cours des 4 cross validations, comme le montre le graphique 4.23. Le modèle final retenu est obtenu en calculant la moyenne des prédicteurs sur les 4 folds d'apprentissage.



Graphique 4.23 – Modélisation du coût moyen de la garantie RC matérielle : validation du modèle par la méthode de validation croisée.

La deuxième méthode de validation utilisée est la validation par quantiles, à travers la courbe de lift. Comme précisé dans la section 2.2.1, cette méthode consiste à trier la base de données par ordre croissant du coût moyen prédit et de diviser ensuite la base de données en tranches égales d'exposition. En dernière instance on va tracer le coût moyen observé et prédit pour chaque quantile. Par construction du graphique, le coût moyen prédit est croissant. Si le modèle est performant, l'observé doit également être croissant suivant les quantiles et doit se rapprocher le plus de la valeur prédite.



Graphique 4.24 – Modélisation du coût moyen de la garantie RC matérielle : courbe de lift construite sur la base de validation.

Nous avons choisi de réaliser 20 quantiles. Malgré quelques écarts constatés, la validation est satisfaisante puisque le coût moyen prédit de sinistres RC matérielle est globalement en cohérence avec le risque observé sur la base de validation.

A RETENIR :

- La modélisation de la prime pure par garantie a été réalisée à travers une approche fréquence x coût moyen. Par souci de redondance on se contente de présenter en détail la méthodologie utilisée ainsi que les résultats obtenus uniquement sur la garantie RC matérielle.
- La fréquence sinistres et le coût moyen ont été modélisés à travers la méthode de linéarisation du GBM avec pour fonction de lien, la loi de Poisson dans le premier cas et la loi Gamma, dans le deuxième cas.
- 17 facteurs de risque ont été sélectionnés dans le modèle final de fréquence des sinistres RC matériels, dont les plus importants sont la zone de risque, le groupe de la moto, l'ancienneté et la durée de détention du véhicule.
- La modélisation de la fréquence des sinistres RC matériels met en évidence l'obsolescence du zonier deux-roues actuel. Le zonier actuellement en place n'est plus adapté au risque observé, une mise à jour de celui-ci est donc recommandée.
- 11 variables ont été sélectionnées dans le modèle final de coût moyen de la garantie RC matérielle, dont les plus importantes sont la cylindrée, l'ancienneté du véhicule, le groupe du véhicule et l'âge du conducteur.
- Si le modèle de fréquence des sinistres permet d'obtenir un indice de Gini proche de 60%, le modèle de coût moyen, quant à lui, présente des performances inférieures (mais toutes mêmes optimisés et validés). Cela fait suite à l'absence dans l'entrepôt de données des facteur concernant le véhicule adverse ainsi que les tiers impliqués.

4.3 Calcul de l'ELR

Après avoir identifié, sur la base des modèles de fréquence et de coût moyen, les critères tarifaires pertinents et les échelles de risque associées à intégrer dans l'équation tarifaire de chaque garantie, l'objectif de cette nouvelle étape est de définir le « juste niveau » de la prime pure de référence à mettre en place pour l'exercice de survenance 2019 [4], garantie par garantie.

Par « juste niveau » on entend un niveau de prime pure suffisant, permettant de couvrir l'ensemble des engagements purs de l'assureur sur l'exercice de survenance analysé. Compte tenu de la structure du risque du portefeuille, la prise en compte de l'ajustement entre les primes pures modélisées et les charges ultimes estimées permet d'obtenir en moyenne un ratio d'ELR au plus juste. Cette étape est donc fondamentale car un mauvais calibrage de ce montant peut s'avérer fortement dommageable pour l'assureur.

L'estimation des charges finales prévisibles (charge à l'ultime) par année de survenance est réalisée et revue fréquemment par les équipes de provisionnement d'AXA France, pour chaque branche d'assurances, de manière centralisée. Ce sont donc les valeurs fournies par les équipes de provisionnement qui ont été utilisées pour le calibrage des primes pures de référence.

La prime pure de référence de chaque garantie est obtenue en rapportant l'estimation de la charge ultime à la somme des relativités comptabilisées sur l'exercice 2019.

Concrètement, afin de déterminer les niveaux de référence de chaque garantie, la première étape est la construction de la base de données relative au portefeuille deux-roues de l'année 2019 qui contient l'intégralité des critères tarifaires. Cette base est construite de façon similaire à la base utilisée pour la modélisation de la fréquence des sinistres RC matériels. Les seules différences concernent :

- La période d'observation : lors de la modélisation de la fréquence des sinistres RC matériels, l'analyse a été basée sur le portefeuille 2016 à 2018, tandis que, pour le calibrage des primes pures on utilise les données relatives au portefeuille 2019 (le niveau de la prime pure de référence doit être le plus à jour possible).

- Les informations relatives à la sinistralité : si, lors de la construction des modèles de primes pures ces informations sont indispensables, dans cette partie de l'analyse l'intérêt porte uniquement sur la sinistralité modélisée (il s'agit de l'application des modèles de prime pure bâtis en amont sur le portefeuille deux-roues 2019) et non pas sur la sinistralité observée.

Une fois la base de données constituée, l'étape suivante est l'application des modèles de fréquence et de coût moyen bâtis lors de l'étape précédente, en faisant attention à prendre en compte uniquement les paramètres β estimés. Bien que les constantes (les β_0) de chaque modèle construit soient connues, lors de l'estimation de la prime pure de référence, ces valeurs sont remplacées par la valeur 1.

Il est important de rappeler que, lors de l'étape précédente, seulement les garanties RC matérielle, RC corporelle, IV (incendie et vol) et DO (dommages) ont été modélisées à travers une approche fréquence - coût moyen. Pour les garanties non modélisées, la prime pure de référence est calculée au global, comme rapport entre la charge de sinistres ultime fournie par les équipes de provisionnement et l'exposition de la garantie (le nombre d'années-polices ayant souscrit la garantie). De cette manière la prime pure des garanties non modélisées sera un montant forfaitaire pour tous les contrats ayant souscrit la garantie.

Mathématiquement le calcul de la prime pure de référence s'écrit :

4. Il s'agit de la dernière année de survenance de sinistres qui est finalisée au moment de réalisation du mémoire

- Si la garantie a été modélisée :

$$PP_0^{\text{garantie}} = \frac{\text{Charge ultime de sinistres de la garantie sur l'exercice de survenance 2019}}{\sum_{i=1}^{n} \lambda_i^{\text{garantie}} expo_i^{\text{garantie}}}, \tag{4.6}$$

  où λ_i - l'échelle de risque associé au contrat i et exp_i - l'exposition du contrat pour la garantie considérée sur l'exercice 2019. Pour chaque contrat, le risque associé λ_i est calculé comme produit entre les coefficients β issus des modèles de fréquence et coût moyen.

- Si la garantie n'a pas été modélisée :

$$PP_0^{\text{garantie}} = \frac{\text{Charge ultime de sinistres de la garantie sur l'exercice de survenance 2019}}{\sum_{i=1}^{n} expo_i^{\text{garantie}}}. \tag{4.7}$$

On en déduit alors l'équation tarifaire :

$$\pi(Z_i^{\text{garantie}}) = PP_0^{\text{garantie}} x \lambda_i^{\text{garantie}}, \tag{4.8}$$

où :

$\pi(Z_i^{\text{garantie}})$: la prime pure du profil de risque i, i=1, 2, ..., I,

$\text{PP}_0^{\text{garantie}}$: la prime pure de référence de la garantie,

$\lambda_i^{\text{garantie}}$: l'échelle de risque tarifaire associé au profil de risque i, i=1, 2, ..., I.

Si nous prenons l'exemple de la garantie RC matérielle, compte tenu des relativités comptabilisées sur l'exercice 2019, la prime pure de référence est égale à 7,75 euros. L'équation tarifaire pour la garantie RC matérielle s'écrit donc :

$$\pi(Z_i^{\text{RC Mat}}) = PP_0^{\text{RC Mat}} x \lambda_i^{\text{RC Mat}}, \tag{4.9}$$

où :

$\lambda_i^{\text{RC Mat}} = \prod_{j=1}^{17} Y_{ij}^{\text{fréq RC Mat}} X \prod_{j=1}^{11} Y_{ij}^{\text{coût RC Mat}}$,

$Y_{ij}^{\text{fréq RC Mat}}$: le coefficient issu du modèle de fréquence du jème critère (j=1, 2,..., 17) associé au profil i,

$Y_{ij}^{\text{coût RC Mat}}$: le coefficient issu du modèle de coût moyen du jème critère (j=1, 2,..., 11) associé au profil i.

<u>Le calcul de l'ELR</u>

Après avoir calibré les primes pures on peut calculer l'ELR 2019 global et unitaire pour les contrats en portefeuille. Comme défini dans la section 4.1.2, l'ELR représente le rapport entre l'espérance de sinistres et la prime payée. Deux notions distinctes d'ELR sont utilisés lors de l'analyse d'un portefeuille d'assurances :

- **L'ELR classique** : ELR = Exposition * Prime pure modélisée / Primes acquises

  La prime pure permet de prédire la sinistralité sur un an mais dans ce calcul nous étudions la sinistralité seulement sur la période pendant laquelle le contrat est resté dans le portefeuille. Les primes acquises sont également proratisées, elles représentent les primes réellement payées sur la même période. Suite au calibrage de la prime pure de référence, le niveau global de l'ELR classique coïncide avec le S/C, indicateur défini dans la partie 4.1.1. Cependant, des disparités peuvent être constatées entre ces deux indicateurs lors d'une analyse contrat par contrat. L'ELR est un indicateur qui reflète la rentabilité modélisée tandis que le S/C reflète la rentabilité observée sur le contrat pour la période considérée.

- **L'ELR annualisé** : ELR annualisé = Prime pure modélisée / Prime annualisée

  Les deux éléments qui font partie du calcul de cet indicateur ne sont pas proratisés. Les risques sont étudiés sur un an et non seulement sur le temps passé dans le portefeuille. De la même façon nous n'utilisons plus la prime acquise mais la prime annualisée, ce qui correspond à l'espérance de primes sur un an. Avec ce calcul, nous accordons moins d'importance au mix du portefeuille. Tous les contrats du portefeuille se voient attribuer le même poids.

Pour une meilleure compréhension du calcul de ces deux indicateurs voici un exemple. Soit un contrat C qui est resté six mois dans le portefeuille. Dans ce cas, l'exposition est alors de 0,5.

Prime pure proratisée	Prime acquise	Prime pure modélisée	Prime annualisée
75	100	150	200

Tableau 4.6 – ELR annualisé vs ELR classique

La prime pure modélisée permet de prédire la sinistralité sur un an, de ce fait, afin d'obtenir la prime pure proratisée nous multiplions par l'exposition (0,5). A partir de nos bases (fonctionnant par image), la prime annualisée est obtenue en multipliant la prime acquise sur la période avec l'inverse de l'exposition. Bien entendu, sur un contrat donné, les deux ELR sont identiques :

ELR classique = $\frac{75}{100} = 75\%$,
ELR annualisé = $\frac{150}{200} = 75\%$.

Si nous souhaitons étudier l'ELR de façon segmentée, prenons un deuxième exemple simple. On considère deux contrats C1 et C2, correspondant au risque X :

N°	Expo	Prime acquise	Prime annualisée	PP modélisée	PP proratisée
C1	0,5	100	200	150	75
C2	0,25	30	120	100	25

Tableau 4.7 – ELR annualisé vs ELR classique 2

On calcule ensuite les ELR correspondantes au risque X :

ELR classique = $\frac{75+25}{100+30} = 79,92\%$,
ELR annualisé = $\frac{150+100}{200+120} = 78,12\%$.

Des écarts sont alors constatés entre les deux ELR : le premier prend en compte l'exposition réelle du portefeuille, tandis que le deuxième considère les deux contrats comme ayant le même poids, comme s'ils étaient restés un an dans le portefeuille tous les deux.

Etant donné les différences entre ces deux indicateurs, leur utilisation est différente en fonction du phénomène que l'on souhaite analyser. Si l'objectif est d'analyser la rentabilité estimée d'un portefeuille, l'ELR classique nous semble un bon indicateur car il reflète la réalité d'un portefeuille d'assurances sur une année. Si l'objectif est d'analyser la rentabilité espérée des affaires nouvelles, puisqu'on ne peut pas connaître en amont la période de temps qu'un nouveau contrat restera dans le portefeuille, il est souhaitable d'accorder le même poids à tous les contrats, et donc, d'utiliser l'ELR annualisé comme indicateur de leur rentabilité espérée. C'est pourquoi nous retenons pour cette étude l'ELR annualisé.

Les deux indicateurs peuvent être calculés au niveau total, sur toutes les garanties acquises d'un contrat, mais également au niveau de chaque garantie.

Calcul de l'ELR sur la base des devis

Après avoir montré, à travers un exemple, la méthode de calcul de ces deux indicateurs, on calcule ensuite les ELR annualisés sur la base des devis. Comme dans le cas des affaires nouvelles, le calcul de l'ELR classique sur la base des devis est impossible car on ne peut pas connaitre en amont la période de temps qu'un devis est susceptible de rester dans le portefeuille, dans le cas où il serait concrétisé. On accorde donc le même poids à tous les devis, en calculant l'ELR annualisé. Pour réaliser ce calcul on applique les modèles de primes pures sur la base des devis, en faisant attention à prendre en compte les primes pures de référence calculées précédemment (calibrées avec les charges ultimes). L'ELR est ensuite obtenu en faisant le rapport entre la prime pure calibrée et la cotisation du devis.

On peut se poser la question de l'utilité du calcul de cet indicateur sur la base des devis. L'objectif du mémoire est l'identification des mesures tarifaires à l'affaire nouvelle en prenant en compte la compétitivité (par le biais du taux de transformation de devis) et la rentabilité estimée des devis (à travers l'ELR). L'ELR calculé sur la base de devis est le seul indicateur capable de refléter la rentabilité estimée des devis.

L'ELR calculé sur la base du portefeuille reflète la rentabilité estimée du portefeuille et au niveau global il coïncide avec le S/C sur la période considérée. Il est utilisé notamment pour piloter les mesures tarifaires passés lors de renouvellement des contrats, à échéance. Même si l'objet du mémoire n'est pas l'analyse de cet indicateur, nous avons calculé également l'ELR du portefeuille et l'ELR des affaires nouvelles. Il est intéressant de comparer, au niveau global ces trois indicateurs : l'ELR du portefeuille, l'ELR des affaires nouvelles et l'ELR des devis.

Si l'on compare l'ELR des affaires nouvelles et l'ELR des devis plusieurs cas sont possibles :

- ELR AN < ELR DEVIS : suite à la segmentation des risques, le tarif appliqué à l'affaire nouvelle est adapté au risque sous-jacent. Avec une politique tarifaire adéquate, les "bons" risques décident de souscrire auprès de la compagnie, tandis que les "mauvais risques", pour lesquels le tarif affiché est élevé, décideront de s'assurer à la concurrence. La sélection des risques par le biais du tarif est efficace. Au delà de cette analyse très technique il faut également prendre en considération la stratégie tarifaire et les latitudes commerciales accordés qui peuvent avoir un fort impact sur les ELR, et biaiser la comparaison de ces deux indicateurs.

- ELR AN = ELR DEVIS : la sélection de risques pratiqué par la compagnie est très similaire par rapport à celle réalisé par les autres compagnies d'assurance.

- ELR AN > ELR DEVIS : la segmentation du risque est insuffisante ce qui génère un effet d'antisélection. Le tarif est inadapté par rapport au risque et les "mauvais risques" s'assurent auprès de la compagnie.

Il est important de connaître également l'écart entre le niveau d'ELR du portefeuille et celui des affaires nouvelles. Dans un marché de l'assurance devenu de plus en plus compétitif, la plupart des assureurs proposent des tarifs très attractifs à la souscription. Ils acceptent de dégrader les résultats techniques à la souscription, en capitalisant sur la fidélité des clients, qui deviennent rentables au bout de plusieurs années passés en portefeuille. Mais, depuis l'arrivée de la loi Hamon, permettant la résiliation des contrats d'assurances à tout moment après un an d'engagement, les hypothèses initiales fixées par les assureurs peuvent devenir obsolètes. Il est important donc de limiter, dans la mesure du possible, l'écart entre ces deux indicateurs. Une baisse trop importante des tarifs à l'affaire nouvelle, afin de gagner en compétitivité, doit être suivie d'une augmentation de tarifs pour les contrats en portefeuille afin de compenser la perte générée à la souscription, et l'écart entre ces deux indicateurs se creuse encore plus. Dans le cadre de ce mémoire nous allons travailler sur le tarif à l'affaire nouvelle. Il est cependant important de garder à l'esprit les impacts sous-jacents sur la rentabilité du portefeuille dans le temps (on ne doit pas décorréler ces indicateurs). Accentuer trop l'écart entre les deux pourrait mettre en jeux la pérennité du portefeuille sur le long terme.

A RETENIR :

- Les modèles de prime pure bâtis à la section 4.2 ne permettent pas de calibrer la prime pure de référence mais uniquement d'identifier les critères tarifaires pertinents et les échelles de risque associées à intégrer dans l'équation tarifaire de chaque garantie. Malgré l'usage de la vision de sinistralité la plus à jour au moment de l'étude, la modélisation de ne permet pas d'intégrer une vision à l'ultime de la charge des sinistres.

- L'objectif de cette nouvelle étape est de définir le « juste niveau » de la prime pure de référence à mettre en place pour l'exercice de survenance 2019, garantie par garantie. Pour ce faire les modèles de prime pure construits seront appliqués sur le portefeuille deux-roues 2019. Ensuite la prime pure de référence de chaque garantie est calculée comme rapport entre la charge ultime de sinistres (charge finale prévisible) et la somme des relativités comptabilisées sur l'exercice analysé.

- Après que la calibration des primes pures ait été réalisée, l'ELR est calculé contrat par contrat comme rapport entre la prime pure modélisée et calibrée et la cotisation acquise. Parmi les deux visions d'ELR possibles, nous retenons l'ELR annualisé qui est le seul à avoir du sens sur les devis. Ce KPI est calculé au niveau global, sur toutes les garanties acquises d'un contrat, mais également au niveau de chaque garantie.

Chapitre 5

Identification des mesures tarifaires

L'objectif de ce mémoire est de proposer une méthode d'ajustement du tarif commercial permettant de concilier les enjeux techniques (tarifer le risque au plus juste) et commerciaux (proposer un tarif compétitif, en adéquation avec le marché).

Pour connaître le positionnement tarifaire sur le marché, on se base sur le taux de transformation des devis modélisé dans le troisième chapitre de l'étude. Et pour analyser la rentabilité estimée des devis, on se base sur l'expected loss ratio (ELR), dont la méthode de calcul a été exposée dans le quatrième chapitre de ce mémoire.

L'importance de l'analyse conjointe de l'ELR et du taux de transformation

Pour chaque compagnie d'assurances, l'objectif est d'attirer les contrats rentables tout en améliorant le chiffre d'affaires. Pour arriver à ces fins, une méthode serait d'augmenter le tarif commercial des mauvais risques à l'affaire nouvelle afin d'améliorer leur rentabilité et de diminuer le tarif des bons risques, afin d'avoir un ELR élevé. Ainsi, nous pouvons gagner des parts de marché pour les bons risques, tandis que les mauvais risques, pour lesquels nous avons augmenté le tarif, vont souscrire à la concurrence. En utilisant seulement l'ELR comme levier d'ajustement de la prime commerciale, le risque est une baisse de rentabilité des affaires nouvelles. Si nos tarifs en affaire nouvelle deviennent avantageux au point de ne plus être rentables, il faut absolument les équilibrer avec les mauvais risques, qui eux, auront un ELR très bas. Nous devons donc être sûr que le mix des affaires nouvelles permet d'assurer l'équilibre entre les mauvais et les bons risques. Cependant, si nous augmentons de manière trop importante le tarif commercial de mauvais risques, ils ne souscriront plus et se retourneront vers la concurrence ce qui engendre une baisse du chiffre d'affaires et l'équilibre nécessaire pour assurer la rentabilité des affaires nouvelles ne sera plus réalisé.

Pour une meilleure compréhension, nous allons exemplifier cette problématique avec un cas concret. Soit A un bon risque et B un mauvais risque. On souhaite modifier la prime commerciale afin d'augmenter l'ELR des bons risques et de baisser l'ELR pour les mauvais risques.

Risque	Prime pure	Objectif ELR	Tarif commercial avec objectif	Tarif marché
A	160	80%	200	250
B	400	50%	800	750

Tableau 5.1 – Prime commerciale basée uniquement sur l'ELR

En se basant uniquement sur l'ELR pour l'ajustement de la prime commerciale, nous aurions un tarif de 200 euros pour le risque A versus 250 euros sur le marché. Nous serions donc très attractifs sur les bons risques car le prix fixé est bien en dessous du marché. Toutefois, en choisissant un tarif commercial à peine en dessous du marché, de 245 euros par exemple, le tarif reste toujours attractif (car toujours en dessous du marché) et cela nous permet d'améliorer la rentabilité ainsi que le chiffre d'affaires. Concernant le risque B, pour atteindre l'objectif d'ELR fixé à 50%, la prime commerciale correspondante doit être au niveau de 800 euros. Ce tarif nous permet d'être très rentable mais il est au-dessus du marché, ce qui génère peu de souscriptions.

Si nous choisissons de nous aligner à la concurrence, en proposant le même tarif de 750 euros, le nouvel ELR sera de 53%, et le segment reste rentable.

En recapitulant, la prise en compte simultanée de l'ELR et de la concurrence lors de la fixation du tarif commercial permet d'obtenir ces indicateurs :

Risque	PP [1]	Objectif ELR	PC [2] avec objectif ELR	PC marché	PC choisie	ELR réel
A	160	80%	200	250	245	65%
B	400	50%	800	750	750	53%

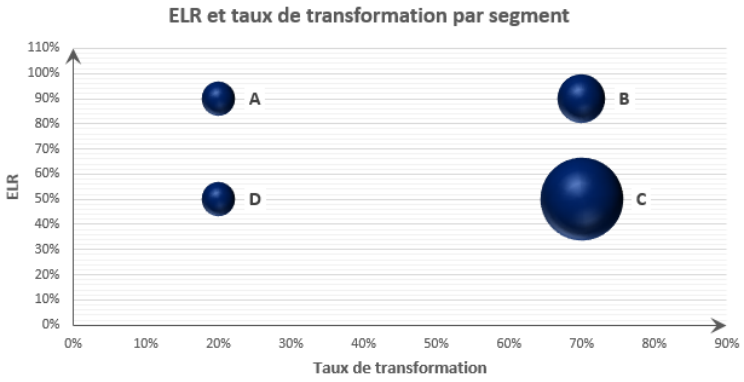
Tableau 5.2 – Prime commerciale basée sur l'ELR et la concurrence

Pour le risque A, choisir le tarif en fonction uniquement d'un objectif d'ELR n'aurait pas été la solution la plus pertinente. La prise en compte du tarif de la concurrence nous a permis de gagner en rentabilité, tous en proposant un prix compétitif. Pour le risque B, si nous nous étions basés uniquement sur l'objectif de l'ELR, le prix de 800 euros proposé initialement pour ce profil de risque aurait été hors de marché et aurait engendré une souscription très limitée des contrats. La prise en compte du tarif du marché nous a permis de proposer un tarif attractif, tout en restant rentables.

La méthode graphique d'identification des mesures tarifaires

Pour mettre au point une stratégie d'ajustement de la prime commerciale il est donc primordial de prendre en compte d'une part la concurrence (par le biais du taux de transformation) et d'autre part la rentabilité estimée des devis (par le biais de l'ELR). Nous cherchons à trouver un bon compromis entre compétitivité et rentabilité et réajuster le tarif des affaires nouvelles à la hausse ou à la baisse selon les besoins identifiés. Cet exercice d'optimisation tarifaire devrait permettre de redonner plus de compétitivité tout en améliorant la rentabilité des contrats transformés.

En ayant comme input les deux indicateurs (le taux de transformation et l'ELR) plusieurs cas de figure peuvent se présenter :



Graphique 5.1 – ELR versus taux de transformation par segment.

- Cas A : ELR élevé et taux de transformation bas

  Signification : Segment sous tarifé par rapport au risque et prix pas assez attractif par rapport à la concurrence

  Mesure : Le tarif commercial n'est pas assez attractif, de ce fait nous souscrivons peu et les contrats souscrits ne sont pas rentables. Si nous diminuons nos tarifs nous gagnerons en compétitivité, en contrepartie la rentabilité globale des affaires nouvelles va diminuer. On se propose de laisser le tarif commercial tel qu'il est actuellement. Ainsi nous allons continuer à souscrire une petite partie de ces contrats, sans trop dégrader les résultats techniques.

- Cas B : ELR élevé et taux de transformation élevé

  Signification : Segment sous tarifé par rapport au risque et prix attractif par rapport à la concurrence

  Mesure : Le tarif commercial est attractif mais les contrats ne sont pas rentables. Nous devons essayer d'augmenter la rentabilité des segments en se rapprochant de l'ELR moyen. Pour cela nous devons revoir à la hausse les coefficients de la prime commerciale. De cette façon le taux de transformation diminuera peut-être mais nous gagnerons en rentabilité sur les contrats souscrits.

- Cas C : ELR bas et taux de transformation élevé

  Signification : Segment sur tarifé par rapport au risque mais prix attractif par rapport à la concurrence

  Mesure : Le tarif commercial proposé est élevé mais il reste compétitif. Nous pouvons conserver le tarif tel quel qui est gagnant-gagnant pour la compagnie (très bonne compétitivité et très bonne rentabilité).

- Cas D : ELR bas et taux de transformation faible

  Signification : Segment sur tarifé par rapport au risque et prix pas assez attractif par rapport à la concurrence

  Mesure : Le tarif n'est pas assez compétitif, le risque étant surestimé. Nous pouvons diminuer le coefficient de la prime commerciale sur ce segment. Ainsi, nous allons gagner en compétitivité sans forcement trop dégrader la rentabilité des affaires nouvelles.

Dans le cadre de l'analyse conjointe de ces deux indicateurs il est également important de prendre en compte le volume de devis concernés par chaque segment. L'impact global des mesures tarifaires n'est pas le même si le changement de tarif concerne 5% de la base de devis, ou 60% de la base de devis. Une fois que les segments sur lesquels le tarif commercial doit être ajusté ont été identifiés, et que le niveau d'ajustement a été validé, nous devons mesurer l'impact généré sur la base des affaires nouvelles, en termes de volume (quel sera le nouveau taux de transformation) et en termes de rentabilité (quel sera le nouveau niveau d'ELR des affaires nouvelles).

Ainsi, à travers cette analyse nous souhaitons répondre à deux questions :

- Quels sont les segments sur lesquels nous devons ajuster le tarif commercial ?

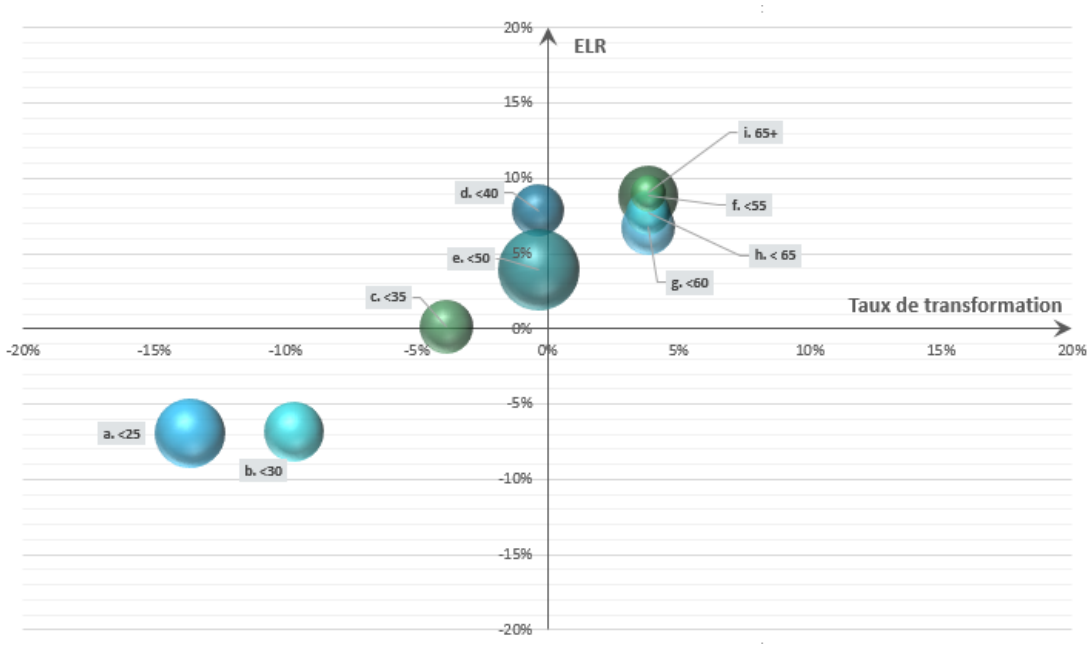
- De combien doit-on ajuster le tarif sans dégrader la compétitivité ni les résultats techniques ?

Nous proposons par la suite une méthode d'identification des profils dont la prime commerciale nécessite d'être revue. Pour chaque variable explicative et tarifaire sélectionnée dans le modèle de transformation de devis, nous avons représenté, à travers un nuage de points le lien entre le taux de conversion et l'ELR calculé sur la base de devis par segment. Nous avons représenté le meme type de nuage des points que celui présenté dans l'exemple théorique précèdent (cf graphique 5.1). Pour chacune des variables analysées, l'abscisse représente le taux de transformation standardisé en toutes choses égales par ailleurs et l'ordonnée représente l'ELR standardisé. Nous avons préféré la prise en compte des indicateurs standardisés aux valeurs non retraitées car ce qui nous intéresse ce sont les relativités par rapport à la moyenne plutôt que les niveaux globaux. En prenant les valeurs toutes choses égales par ailleurs pour le taux de transformation de devis nous nous assurons de la prise en compte de l'effet propre de chacune des variables explicatives sur la conversion, indépendamment des autres.

Pour rappel, lors de la construction du modèle de transformation de devis moto, 21 variables ont été sélectionnées dans le modèle final. Avant de commencer l'analyse conjointe de deux KPI (le taux de transformation et l'ELR) on doit s'assurer que les variables prises en compte dans l'analyse sont bien des variables actuellement présentes dans le moteur tarifaire et que les modalités de chaque facteur sont identiques entre l'analyse et le moteur de prime commerciale actuellement en place. Le cas contraire, l'ajout des nouvelles variables dans le moteur tarifaire, ou un changement concernant les modalités d'une des variables déjà présentes dans l'outil, demanderait un développement informatique assez lourd et coûteux. Pour cette raison, on se contente dans un premier temps de réaliser l'analyse graphique uniquement sur les variables déjà présentes dans le moteur tarifaire existant.

L'ensemble des variables a été analysé. On présente ici, à titre d'exemple, les résultats obtenus sur les 3 variables les plus intéressantes.

L'âge du conducteur



Graphique 5.2 – ELR et taux de transformation en fonction de l'âge du conducteur.

L'analyse conjointe du taux de transformation et des ELR en fonction de l'âge du conducteur fait apparaître un problème de tarification sur les jeunes conducteurs. Pour les conducteurs âgés de moins de 30 ans, tant l'ELR que la transformation toutes choses égales par ailleurs sont inférieurs à la moyenne. Un ajustement à la baisse du tarif commercial pour ce profil de conducteurs est nécessaire afin de proposer des tarifs plus compétitifs sans trop dégrader la rentabilité.

Les conducteurs d'âge compris entre 30 et 35 ans transforment moins bien que la moyenne, mais leur rentabilité estimée ne diverge pas de façon significative par rapport à la moyenne constatée sur la base des devis. Pour ce type de profils, deux choix s'offrent à nous :

- Favoriser la transformation : une diminution du tarif commercial peut être réalisée dans la limite du taux de conversion. La rentabilité estimée sera alors dégradée dans la même proportion, l'ELR ne sera plus à l'équilibre.
- Favoriser la rentabilité : le tarif commercial reste inchangé afin de ne pas dégrader l'ELR, malgré un tarif légèrement surestimé par rapport au marché.

La deuxième option a été préférée à la première car assurer l'équilibre en termes de rentabilité nous semble plus important. De plus, les assurés d'âge compris entre 30 et 35 ans représentent seulement 9% des devis.

En ce qui concerne les conducteurs d'âge compris entre 35 et 50 ans, l'effet toute choses égales par ailleurs sur le taux de transformation est nul tandis que leur ELR observé est supérieur à la moyenne, ce qui se traduit par une rentabilité estimée inférieure à la moyenne sur ces tranches d'âge. Deux stratégies sont alors possibles :

- Favoriser la rentabilité : une augmentation du tarif commercial peut être réalisée afin d'assurer l'équilibre de l'ELR. Le tarif sera dans ce cas surestimé par rapport à la concurrence et le taux de transformation va diminuer.

- Favoriser la transformation : le tarif commercial reste inchangé afin de ne pas dégrader la transformation.

Nous avons préféré laisser le tarif inchangé afin de ne pas dégrader la transformation, car ces profils de conducteurs représentent environ un tiers des clients ayant réalisé un devis pendant la période janvier à juin 2019. Une baisse de tarif pour favoriser la rentabilité aurait généré un impact négatif trop important sur le volume d'affaires nouvelles.

Enfin, la conversion de conducteurs de plus de 50 ans est meilleure que la moyenne, mais leurs ELR sont également plus élevés. Une augmentation du tarif commercial sera donc réalisée afin d'améliorer leur rentabilité estimée ce qui nous permettra également de compenser la réduction du tarif appliqué sur les jeunes motards.

La zone de risque

Graphique 5.3 – ELR et taux de transformation en fonction de la zone de risque.

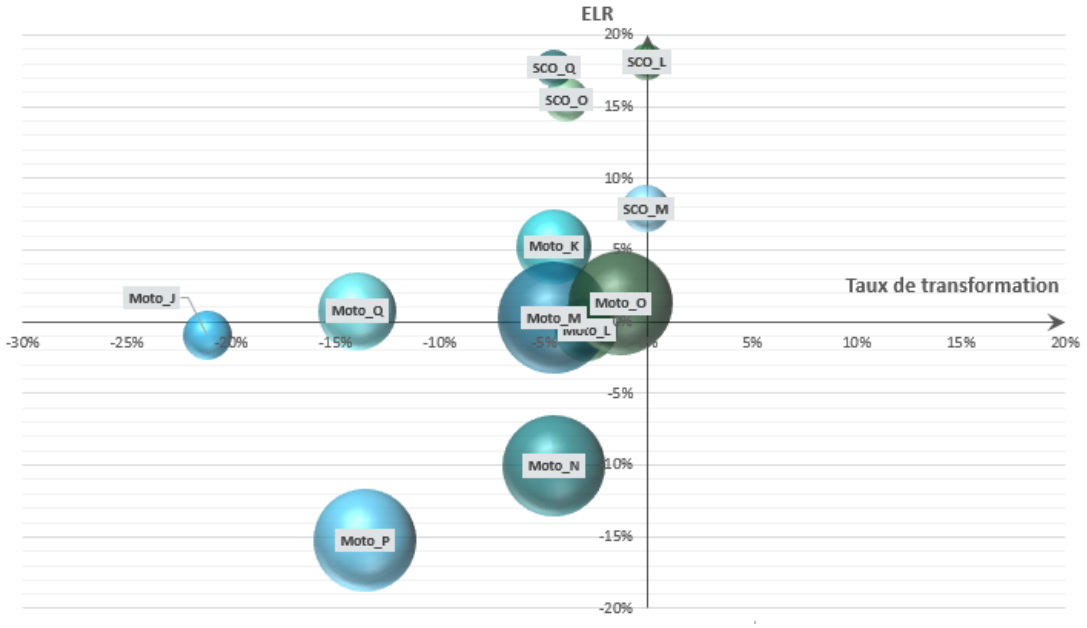
La figure 5.3 fait apparaître des disparités de conversion et de rentabilité en fonction de la zone de risque. Deux profils se démarquent de façon significative : les zones de risque 12 et 13.

En ce qui concerne la zone 12, tant la transformation que l'ELR sont nettement inférieurs à la moyenne. Un ajustement à la baisse du tarif commercial pour favoriser la compétitivité est nécessaire. L'ELR de ce profil va augmenter, mais pas au point de dégrader la rentabilité de façon trop significative.

Concernant la zone de risque 13, le tarif commercial actuellement en place est surestimé par rapport au marché, mais le tarif est justifié techniquement, car ce segment présente l'ELR le plus élevé. Etant donné que la zone de risque 13 cumule les villes les plus risquées en termes de résultats techniques, le tarif doit rester inchangé pour que la rentabilité ne soit pas dégradée.

Lors de la modélisation de la prime pure, détaillée dans la section 4.2, nous avons remarqué qu'une adaptation du zonier actuel par rapport à la fréquence de sinistres serait opportune. Une refonte intégrale du zonier nous semble nécessaire. Même si l'analyse que nous venons de présenter met en évidence que des mesures d'ajustement du tarif commercial en fonction de la zone de risque doivent être mis en place pour gagner en compétitivité, nous avons fait le choix de laisser inchangés les coefficients de la prime commerciale à ce stade. Nous préconisons d'initier un travail de fond pour élaborer un nouveau zonier ultérieurement.

La classe de moto/scooter



Graphique 5.4 – ELR et taux de transformation en fonction de la classe de moto/scooter.

L'analyse conjointe du taux de transformation et des ELR en fonction de la classe de moto ou de scooter fait apparaître un problème de tarification sur les motos de classe P. Pour cette catégorie de motos, tant l'ELR que la transformation toutes choses égales par ailleurs sont inférieurs à la moyenne. Un ajustement à la baisse du tarif commercial pour ce profil de motos est nécessaire afin de proposer des tarifs plus compétitifs sans pour autant dégrader de façon significative la rentabilité.

Concernant les scooters (toutes classes confondues), leur transformation ne diffère pas de façon significative par rapport à la moyenne de devis. Cependant ils ont les ELR les plus élevés. On propose de laisser le tarif actuellement en place inchangé car ils font partie des profils ayant le risque le plus élevé.

Ce type d'analyse graphique est réalisé sur l'intégralité des variables explicatives sélectionnées dans le modèle de transformation des devis et qui sont présentes actuellement dans le moteur tarifaire. L'analyse conjointe des ELR et du taux de transformation nous a permis d'identifier 8 segments sur lesquels le tarif commercial nécessite des ajustements à la hausse ou à la baisse.

Après avoir identifié les segments nécessitant des ajustements du tarif commercial, ainsi que le sens de l'ajustement (majoration ou minoration du tarif), on doit déterminer quel est le niveau optimal de majoration/minoration pour chacun de segments. Afin de préserver l'équilibre de l'ELR, l'ajustement du tarif est limité à l'ELR d'équilibre. Par exemple, si l'on souhaite ajuster à la baisse le tarif d'un segment dont l'ELR est de 10% inférieur à la moyenne, la limite maximale de diminution du tarif commercial est également de 10%.

Cependant, si les ajustements sont trop importants par rapport au tarif actuel, le mix des affaires nouvelles peut être fortement impacté également. On a donc fait le choix de limiter les

ajustements de la prime commerciale par une évolution maximale de 5% sur chaque coefficient, à la hausse ou à la baisse. Cependant nous avons volontairement levé cette limitation pour un segment que nous souhaitons tout particulièrement animer commercialement pour 2019. Sur ce segment nous avons baissé le coefficient de -20%. A contrario, le segment fortement animé de la même manière les années précédents a été remis au niveau tarifaire standard, et son coefficient a enregistré une augmentation de +20%.

Pour les segments identifiés, on regarde également les ELR par garantie, afin de déterminer si l'ajustement doit s'appliquer sur toutes les garanties ou sur une garantie en particulier (car les coefficients de la prime commerciale peuvent être différents en fonction des garanties).

Impacts générés sur la base des devis

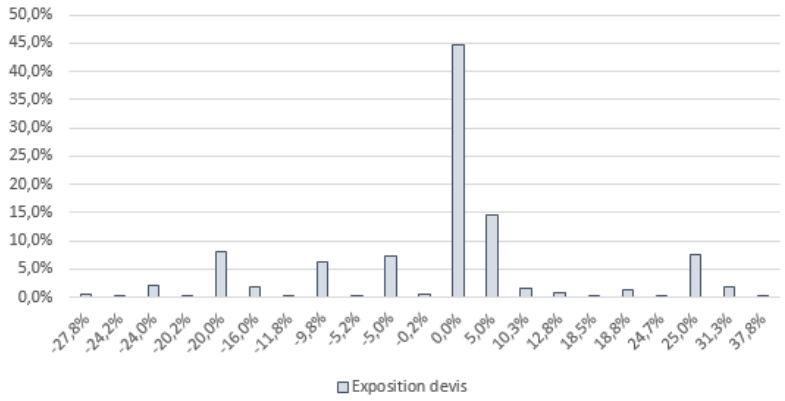
Après avoir fixé les niveaux d'ajustement tarifaire ainsi que les profils qui seront impactés par les changements, nous pouvons estimer l'impact généré sur la base des devis. Le tableau ci-dessous met en évidence pour chaque niveau d'ajustement tarifaire le volume de devis concernés, l'évolution estimée du taux de transformation ainsi que l'évolution de la rentabilité estimée des devis.

Evolution du tarif commercial	Nombre devis	% devis	Nombre AN initial	Nombre d'AN après mesures	Evol AN	Taux transfo initial	Taux transfo final	PP	PC initiale	PC ajusté	ELR initial	ELR ajusté
-27,8%	307	1%	160	204	44	52%	66%	174 €	243 €	176 €	71%	99%
-24,2%	4	0%	1	1	-	25%	25%	161 €	437 €	331 €	37%	49%
-24,0%	1 227	2%	480	595	115	39%	48%	363 €	611 €	464 €	59%	78%
-20,2%	74	0%	34	41	7	46%	55%	338 €	513 €	410 €	66%	82%
-20,0%	4 843	8%	2 159	2 591	432	45%	53%	324 €	513 €	410 €	63%	79%
-16,0%	1 133	2%	641	744	103	57%	66%	215 €	311 €	261 €	69%	83%
-11,8%	108	0%	69	77	8	64%	71%	220 €	307 €	271 €	71%	81%
-9,8%	3 818	6%	2 083	2 286	203	55%	60%	188 €	215 €	194 €	87%	97%
-5,2%	55	0%	31	33	2	56%	60%	221 €	260 €	247 €	85%	90%
-5,0%	4 480	7%	2 125	2 231	106	47%	50%	403 €	531 €	505 €	76%	80%
-0,2%	307	1%	171	171	-	56%	56%	414 €	459 €	458 €	90%	90%
0,0%	26 956	45%	13 888	13 888	-	52%	52%	323 €	403 €	403 €	80%	80%
5,0%	8 763	15%	5 154	4 896	- 258	59%	56%	261 €	293 €	308 €	89%	85%
10,3%	1 002	2%	665	597	- 68	66%	60%	250 €	269 €	297 €	93%	84%
12,8%	432	1%	234	204	- 30	54%	47%	232 €	232 €	261 €	100%	89%
18,5%	5	0%	3	2	- 1	60%	40%	235 €	172 €	203 €	137%	115%
18,8%	781	1%	400	325	- 75	51%	42%	523 €	526 €	625 €	99%	84%
24,7%	46	0%	18	14	- 4	39%	30%	646 €	573 €	714 €	113%	90%
25,0%	4 534	8%	2 423	1 817	- 606	53%	40%	490 €	481 €	601 €	102%	81%
31,3%	1 177	2%	678	466	- 212	58%	40%	333 €	310 €	407 €	107%	82%
37,8%	120	0%	74	46	- 28	62%	38%	311 €	294 €	405 €	106%	77%
TOTAL	**60 172**	**100%**	**31 491**	**31 229**	**- 262**	**52%**	**52%**	**323 €**	**397 €**	**397 €**	**81%**	**82%**

Tableau 5.3 – Impact de mesures tarifaires par tranche d'évolution de la prime commerciale.

Au niveau de chaque devis, l'évolution de la prime commerciale à la suite des changements tarifaires est comprise entre -27,8% et +37,8%. Le tarif commercial est un tarif multiplicatif et c'est pourquoi les évolutions de certaines primes commerciales peuvent être au delà de l'évolution maximale fixée de 5% lorsqu'elles cumulent plusieurs évolutions.

Cependant, un nombre limité de devis est impacté par des évolutions extrêmes. Ainsi, comme on peut le remarquer dans le graphique 5.5, la distribution des devis en fonction de chaque niveau de majorations est symétrique et centrée en 0. Pour 74% de devis, l'évolution de la prime commerciale est comprise entre (-10%, +10%), 13% de devis ayant une évolution inférieure à -10% et 13% de devis ayant une évolution supérieure à +10%.



Graphique 5.5 – Distribution de devis par niveau d'évolution de la prime commerciale.

Au niveau global, l'évolution de la prime commerciale est quasi nulle (-0,17%), la prime commerciale initiale étant au même niveau que la prime commerciale ajustée. C'est pour cette raison que le niveau d'ELR global sur la base des devis reste inchangé également. Pour estimer l'impact sur le taux de transformation nous avons mis l'hypothèse forte d'une élasticité unitaire du tarif, une réduction de 1% de la prime commerciale permettant une augmentation dans la même mesure du taux de transformation de devis. Habituellement, afin d'estimer l'élasticité au prix, les assureurs mettent en place une procédure connue sous le nom de AB testing. Cette procédure consiste à réaliser aléatoirement des ajustements tarifaires de niveau modéré pour un même profil de conducteur. Cette procédure permet, d'une part, de connaître le niveau global de l'élasticité au prix et d'autre part de réaliser des modèles multivariés de sensibilité au prix en fonction de variables tarifaires. Malgré la forte utilité de ce type de procédure dans le cadre de l'analyse des ajustements du tarif commercial, la procédure d'AB testing sur le produit deux-roues a été mise en place récemment. Nous ne disposons pas d'une période de recul suffisante afin de mesurer la sensibilité au prix de façon robuste. À la suite de cette contrainte, l'hypothèse d'une élasticité au prix unitaire nous semble la plus pertinente.

Suite à cette hypothèse d'élasticité au prix unitaire et à l'impact nul des évolutions tarifaires sur la prime commerciale, il est donc normal de constater un effet quasi nul sur le taux de transformation des devis également.

Malgré des impacts nuls en terme d'ELR et de taux de transformation des devis, les ajustements tarifaires identifiés permettent d'améliorer de façon significative le mix des affaires nouvelles. Les bons risques, pour lesquels nous avons réalisé des baisses de tarif souscriront d'avantage, tandis que les profils risqués verront leur tarif augmenter et donc, la conversion diminuera dans la même mesure. Les devis se situant dans la tranche la plus importante de réduction des tarif (de -27,8%) sont les profils les moins risqués, avec une prime pure modélisée de 174 €. On préfère donc diminuer leur tarif, en dépit d'une hausse d'ELR car, en tant que bons risques, on souhaite attirer ces profils en proposant un tarif très compétitif et ainsi souscrire d'avantage d'affaires nouvelles.

Dans le cas inverse, nous avons augmenté le tarif pour les segments les plus risqués et qui ont une conversion supérieure à la moyenne. Parmi les segments identifiés, 8% de devis vont voir leur tarif augmenter de 25%. Le niveau de leur prime pure est de 490€ tandis que leur prime commerciale actuelle est inférieure. Les devis ne sont donc pas à l'équilibre car la prime commerciale actuelle ne permet pas de couvrir l'intégralité des sinistres estimés pour ce segment. L'ajustement à la hausse du tarif commercial est justifié, d'autant plus que la transformation moyenne est supérieure à la moyenne (notre tarif est donc inférieur au prix pratiqué sur le marché).

Evolution du tarif commercial	Nombre devis	% devis	Nombre AN initial	Nombre d'AN après mesures	Evol AN	Taux transfo initial	Taux transfo final	PP	PC initiale	PC ajusté	ELR initial	ELR ajusté
Diminution	16 356	27%	7 954	8 974	1 020	49%	55%	305 €	433 €	373 €	70%	82%
Augmentation	16 860	28%	9 649	8 367	- 1 282	57%	50%	340 €	353 €	408 €	96%	83%
Tarif constant	26 956	45%	13 888	13 888	-	52%	52%	324 €	403 €	403 €	80%	80%
TOTAL	**60 172**	**100%**	**31 491**	**31 229**	**- 262**	**52%**	**52%**	**323 €**	**397 €**	**397 €**	**81%**	**82%**

Tableau 5.4 – Synthèse des impacts de mesures tarifaires.

Le tableau 5.4 permet de résumer les impacts des changements tarifaires sur la base de devis moto réalisés entre janvier et juin 2019. En moyenne, les 27% de devis concernés par une diminution de la prime commerciale, sont de meilleurs risques, avec une prime pure inférieure à la moyenne. Suite à la réduction du tarif et à l'hypothèse fixée de l'élasticité unitaire, on estime une augmentation de 6 points taux de transformation, pour attendre le niveau de 55%. Dans le cas contraire, 28% de devis sont concernés par une augmentation du tarif commercial. Ils ont une prime pure supérieure à la moyenne et leur ELR est de 15 points supérieur à la moyenne. L'augmentation de la prime commerciale nous a permis d'améliorer la rentabilité de ces devis et de diminuer la part de ces risques dans le portefeuille d'affaire nouvelles. Ainsi, l'effet généré par l'amélioration du mix des affaires nouvelles se traduit par une meilleure rentabilité estimée des contrats souscrits, leur ELR enregistrant un baisse d'environ 1 point.

Pour conclure, l'ajustement du tarif commercial nous a permis d'une part d'améliorer le mix des affaires nouvelles, en proposant des tarifs plus attractifs pour les bons risques et en augmentant le tarif pour les mauvais risques, et d'autre part d'avoir un tarif commercial qui s'ajuste mieux au risque, tout en prenant en compte la compétitivité du produit par rapport au marché.

Mesures tarifaires non segmentés

L'approche d'identification des mesures tarifaires décrite dans ce mémoire est une approche segmentée. Elle nous a permis de réaliser des ajustements des coefficients tarifaires au niveau de segments identifiés, sous contrainte d'une structure tarifaire équivalente.

Outre cet aspect, l'analyse conjointe du taux de transformation et de l'ELR a également mis en évidence une manque de rentabilité pour les faibles primes et qui présentent un taux de transformation élevé.

Une attention particulière a dû être accordée aux segments avec des primes commerciales très faibles dont l'ELR dépasse largement le seuil de 100%. Pour pouvoir redresser la rentabilité de ces segments, deux solutions sont envisageables : soit l'identification des profils avec des primes commerciales faibles (afin de réaliser des mesures segmentées), soit la définition de seuils minimaux de prime par garantie. La première solution ne permet pas de cibler uniquement les contrats concernés par cette problématique. Nous avons choisi la deuxième solution qui permet de remettre l'intégralité des devis non rentables à un seuil d'ELR acceptable.

Pour ce faire nous avons réalisé une analyse approfondie afin d'identifier, garantie par garantie, le niveau de prime minimale permettant d'améliorer l'ELR des contrats avec des primes très faibles. Pour ne pas alourdir l'étude mais aussi pour respecter la confidentialité, les résultats de cette analyse ne font pas partie de ce mémoire.

A RETENIR :

- L'ajustement du tarif à l'affaire nouvelle repose sur l'analyse conjointe du taux de transformation et de l'ELR.
- La stratégie mise en place repose sur la diminution du tarif pour les profils avec un ELR bas et un taux de transformation faible, respectivement, une augmentation de la prime commerciale pour les profils avec un ELR élevé et un taux de transformation élevé.
- Nous utilisons une méthode graphique d'analyse conjointe de ces deux KPI calculés sur la base des devis deux-roues. Cette méthode sera appliquée uniquement sur les variables les plus discriminantes du modèle de conversion des devis. 8 segments nécessitant des ajustements de la prime commerciale ont été ainsi identifiés.
- Afin de préserver l'équilibre de l'ELR et pour ne pas impacter de manière trop importante la structure des affaires nouvelles deux-roues, deux contraintes ont été imposées : d'une part l'ajustement du tarif est limité à l'ELR de l'équilibre, et d'une autre part les ajustements par segment seront limités à +/-5% par rapport au tarif commercial actuellement en place.
- Au global, sur l'ensemble des devis, l'évolution de la prime commerciale suite aux changements tarifaires est comprise entre -27,8% et +37,8 (caractère multiplicatif du tarif). La distribution de devis en fonction de chaque niveau de majoration est symétrique et centrée en 0, l'impact global sur la prime commerciale étant quasi nul.
- Les mesures appliquées permettent d'une part d'améliorer le mix des affaires nouvelles, et d'autre part de proposer des tarifs mieux ajustés par rapport au marché et par rapport au risque. Ainsi, sans dégrader le chiffre d'affaires, les changements tarifaires proposés nous ont permis d'améliorer la rentabilité estimée des affaires nouvelles, l'ELR enregistrant une baisse d'environ 1 point.
- Des mesures tarifaires non segmentés ont également été identifiées : l'inclusion des seuils minimaux de prime au niveaux de chaque garantie afin de redresser la rentabilité des contrats avec des niveaux de primes très faibles.

Conclusion

Dans un contexte assurantiel de plus en plus concurrentiel, la tarification des produits d'assurance deux-roues est de plus en plus étudiée et analysée. Pour faire face à la concurrence accrue, les assureurs doivent être en mesure d'ajuster régulièrement leur tarif et d'innover dans le domaine technique en tirant profit des avantages offerts par les outils d'apprentissage statistique. Le rôle de l'actuaire est de proposer un tarif qui est à la fois en adéquation avec le risque assuré et qui reste en même temps attractif par rapport au marché.

Le but du présent mémoire était d'exposer une technique de mise à jour du tarif à l'affaire nouvelle permettant de concilier dans la tarification deux notions de la prime commerciale : la tarification technique du risque sous-jacent et le positionnement sur le marché de l'assurance.

Si la notion de risque est considérée comme bien appréhendée par l'assureur, la position tarifaire par rapport à la concurrence est un élément moins utilisé dans la tarification. Dans ce contexte, la première étape de l'étude a été consacrée à la modélisation du comportement des prospects à la souscription par le biais du taux de transformation de devis. Il est important de préciser que, suite à la contrainte imposée d'ajustement du tarif à structure tarifaire équivalente, le modèle retenu se base uniquement sur des critères présents dans la structure tarifaire actuellement en place. Cette analyse nous a permis, d'une part, d'identifier les critères pour lesquels un ajustement du tarif commercial est nécessaire (concrètement, sont concernés par une modification du tarif commercial les variables choisies dans le modèle de conversion ayant le pouvoir explicatif le plus élevé) et d'autre part de déterminer l'impact toutes choses égales par ailleurs de chacun de ces critères sur la transformation.

Dans la seconde partie de l'étude notre but était d'estimer la rentabilité théorique annuelle des prospects par le biais de l'ELR. Calculé comme rapport entre la prime pure et la prime commerciale, cet indicateur repose sur la refonte intégrale des modèles de prime pure par garantie. Nous avons bâti à travers une approche classique de fréquence-coût moyen, des modèles de prime pure sur les 4 garanties principales proposées par le produit d'assurance moto d'Axa France (RC matérielle, RC corporelle, incendie-vol et dommages). Après avoir calibré les primes pures modélisées avec la charge ultime des sinistres, nous avons calculé l'ELR au niveau de chaque devis en divisant la charge théorique de sinistres (la prime pure modélisée et calibrée) par la cotisation affichée au moment de la réalisation du devis.

Au vu du nombre important de modèles nécessaires à la réalisation de ce mémoire, nous avons tiré profit des avantages offerts par les outils d'apprentissage statistique, ce qui n'a jamais été fait auparavant sur le périmètre moto. Cette approche permet de construire des modèles robustes et performants et elle offre un gain de temps conséquent.

Dans la dernière étape de l'étude, l'objectif était de proposer une stratégie d'ajustement de la prime commerciale par le biais des deux indicateurs calculés aux étapes précédentes : le taux de transformation et l'ELR. La prise en compte simultanée de ces indicateurs visait l'identification des prospects qui ne transforment pas bien, afin de baisser leur prix dans la limite de leur prime pure théorique et, réciproquement, l'identification des prospects avec un bon taux de conversion et une faible rentabilité estimée afin d'ajuster leur tarif commercial à la hausse.

Pour ce faire nous avons proposé une méthode graphique d'analyse conjointe des deux KPI qui a été appliquée uniquement sur les critères les plus discriminants sur la transformation

des devis. Ainsi nous avons pu identifier 8 segments nécessitant des ajustements de la prime commerciale. Pour préserver l'équilibre de l'ELR et pour ne pas impacter de manière trop importante la structure des affaires nouvelles deux-roues, deux contraintes ont été imposées : d'une part l'ajustement du tarif a été limité à l'ELR de l'équilibre, et d'une autre part les ajustements par segment ont été limités à +/-5% par rapport au tarif commercial actuellement en place. Les modification tarifaires réalisées pour ces 8 segments nous ont permis de diminuer d'environ un point la rentabilité estimée des affaires nouvelles, d'améliorer leur mix et de proposer des tarifs mieux ajustés par rapport au marché et par rapport au risque.

En conclusion, cette étude nous a permis d'analyser le comportement des motards à la souscription ainsi que la rentabilité estimée des prospects deux-roues tout en bénéficiant des avantages liés aux techniques de machine learning. D'une part nous avons identifié les segments pour lesquels une modification du tarif commercial était nécessaire et d'une autre part nous avons trouvé le niveau d'ajustement de la prime commerciale permettant de proposer un tarif compétitif, sans pour autant dégrader la rentabilité.

De plus, la mise à jour des modèles de fréquence de sinistres par garantie a soulevé la problématique de la bonne adéquation du zonier actuel par rapport au risque. Les modèles construits permettent d'obtenir des résultats significativement améliorés avec la prise en compte d'un zonier actualisé ; un travail en profondeur de ce sujet est à envisager.

L'analyse réalisée a mis en évidence également certains problèmes de rentabilité sur les contrats ayant des faibles primes. Suite à une analyse plus approfondie de ces segments, des niveaux minimaux de prime commerciale par garantie ont été fixés.

Le stratégie d'ajustement du tarif à l'affaire nouvelle proposée dans ce mémoire peut être envisagée comme un point de départ pour la construction d'un modèle d'optimisation tarifaire. Dans la suite de l'analyse, il conviendrait d'enrichir l'étude avec la modélisation de l'élasticité au prix (étude AB testing) et de disposer d'un benchmark significatif de la concurrence pour pouvoir prendre en compte les écarts de prix constatés par rapport au marché.

Bibliographie

Biernat, E. & Lutz, M. (2015). *Data science : fondamentaux et études de cas : Machine learning avec Python et R*. Eyrolles.

Chambre Syndicale Internationale de l'Automobile et du Motocycle (2019). *Communiqué de presse*. http ://www.csiam-fr.org/.

Comité des Constructeurs Français d'Automobiles (2019). *Communiqué de presse*. https ://ccfa.fr/.

Dreyfus, G. (2008). *Apprentissage statistique*. Eyrolles.

Friedman, J. (2001). "Greedy function approximation : a gradient boosting machine". In : *Annals of statistics*, p. 1189-1232.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). "Regularization paths for generalized linear models via coordinate descent". In : *Journal of statistical software* 33.1, p. 1.

Fédération Française de l'Asssurance (2019). *L'assurance française. Données clés 2018*. https ://www.ffa-assurance.fr/infos-assures/assurance-des-motos-scooters-et-autres-deux-roues-moteur.

Guillot, A. (2015). *Apprentissage statistique en tarification non-vie : quel avantage opérationnel ?* mémoire d'actuariat,ENSAE.

Harrell, F. (2001). *Regression Modeling Strategies*. Springer.

Lorenz, M. (1905). "Methods of measuring the concentration of wealth". In : *Publications of the American statistical association* 9.70, p. 209-219.

MotorCycles Data (2019). *European Motorcycles sales*. https ://motorcyclesdata.com/2019/11/24/european-motorcycles-sales/.

Pariente, J. (2016). *Modélisation du risque géographique en assurance habitation*. mémoire d'actuariat, Université Paris Dauphine.

Sultan, D. (2018). *Tarif multirisque professionnelle : construction de modèles robustes et interprétables*. mémoire d'actuariat, Centre d'études actuarielles.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". In : *Journal of the Royal Statistical Society : Series B (Methodological)* 58.1, p. 267-288.

Tufféry, S. (2005). *Data mining et statistique décisionnelle : l'intelligence dans les bases de données*. Editions Technip.

Vandekerchove, M. (2019). *Contrôle des modèles de prime pure et de prime commerciale de la Responsabilité Civile par le biais de l'expected loss ratio*. mémoire d'actuariat,ISFA.

Zou, H. & Hastie, T. (2005). "Regularization and variable selection via the elastic net". In : *Journal of the royal statistical society : series B (statistical methodology)* 67.2, p. 301-320.

Annexes

Annexe A

Les catégories de permis moto

Les permis « moto »

AM	•Remplace le BSR (Brevet de Sécurité Routière) •Peut être délivré à partir de 14 ans après 7 heures de formation •Permet de conduire : •Des cyclomoteurs dont la cylindrée n'excède pas 50cm3, dont la vitesse maximale est comprise entre 6 km/h et 45km/h – limité à 4kW •Des quadricycles légers à moteur (L6e) dont la vitesse maximale est comprise entre 6 km/h et 45 km/h - limité à 4kW •N'entre pas dans le régime du permis à points
A1 (Ex AL)	•Peut être délivré à partir de 16 ans après 20 heures de formation minimum •Permet de conduire des motocyclettes dont la cylindrée n'excède pas 125 cm3 et dont la puissance n'excède pas 11 kW (15cv) •Permet de conduire des tricycles (L5e) limités à 15 kW (20cv)
A2	•Peut être délivré à partir de 18 ans après 20 heures de formation minimum •Permet de conduire des motocyclettes de toutes cylindrées d'une puissance maxi de 35 kW (47.6cv) et d'un rapport puissance/poids de 0,2 kW/Kg •Les véhicules dépassant 70 kW (95cv) d'origine ne peuvent être ramenées à 35 kW, elles sont donc interdites au permis A2
A	•Peut être délivré à partir de 20 ans après 2 ans minimum de permis A2 et une formation post-permis de 7h. •Permet de conduire des motocyclettes de toutes cylindrées sans limitation de puissance •Permet de conduire des tricycles sans limitation de puissance

Annexe B

Le type de permis nécessaire par catégorie de deux-roues

Quel permis dois-je obtenir pour conduire un…?

Genres	Catégorie CE (Zone J de la CG)	Genre National (Zone J1 de la CG)	Puissance max (Zone P2 de la CG)	Puissance / Poids (Zone Q de la CG)	Permis requis et possibles
Cyclomoteur	L1e L2e	CL CYCL	≤ 4 kW	Tous	**AM(ex BSR)**conducteur âgé d'au moins 14 ans **A1, A2, A** **B**
Motocyclette Légère	L3e L4e	MTL	≤ 11 kW	Tous	**A1, A2, A** **B**(1) (2) obtenu depuis au moins de 2 ans + formation obligatoire de 7h
Motocyclette	L3e L4e	MTT1	≤ 25 kW	≤ 0,2	**A2, A**
				> 0,2	**A**
	L3e L4e	MTT2	> 25 kW et ≤ 35 kW	≤ 0,2	**A2, A**
				> 0,2	**A**
	L3e L4e	MTT2	> 35 kW	Tous	**A**
Tricycle à Moteur	L5e	TM Carrosserie TMP1	≤ 15 kW	Tous	**A1, A2, A** **B**(2) obtenu depuis au moins de 2 ans + conducteur âgé d'au moins 21 ans + formation obligatoire de 7h
	L5e	TM Carrosserie TMP2	> 15 kW	Tous	**A** **B**(2) obtenu depuis au moins de 2 ans + conducteur âgé d'au moins 21 ans + formation obligatoire de 7h
Quadricycle Léger à Moteur	L6e	QM Carrosserie QLEM	≤ 4 kW	Tous	**AM (BSR)** conducteur âgé d'au moins 14 ans **A1, A2, A** **B1** **B**
Quadricycle Lourd à Moteur	L7e	QM Carrosserie QLOMP	≤ 15 kW	Tous	**B1** **B**

- **Sont exemptés de la formation de 7 heures :**
  - **(1)** Les titulaires d'une catégorie quelconque du permis obtenue avant le 1er mars 1980 **pour les Motocyclettes Légères uniquement**
  - **(2)** Les titulaires du permis B ayant assuré et utilisé une motocyclette légère ou un tricycle à moteur (L5e) au cours des cinq dernières années précédant le 01/01/2011

Liste des acronymes

CART Classification and regression tree. 38, 39

CH Chargements pour frais. 17, 18

CS Chargements de sécurité. 17, 18

CUS moto de type custom. 78, 83

EDR expected deviance ratio. 34, 35, 44, 57, 72, 73, 84

ELR Expected loss ratio. 9, 11, 12, 23, 25–27, 30, 59–61, 86–100, 112, 113

EXO moto de type exotique. 78, 83

GBM Gradient boosting machine. 11, 28–30, 40, 41, 43, 44, 55, 58, 69, 70, 72, 81, 82, 85

GLM generalized linearmodel ou modèle linéaire généralisé. 26, 28, 37, 43, 55, 69, 81

GTO moto de type grand tourisme. 66, 78, 83

HYP moto de type hypersportive. 78, 83

i.i.d indépendants et identiquement distribués. 30, 62

KPI Key Performance Indicator. 26, 90, 94, 100

LASSO Least Absolute Shrinkage and Selection Operator. 29, 30, 33, 35, 36, 38, 40, 44, 58

LLH likelihood. 35

PC Prime commerciale. 17, 60, 92

PP Prime pure. 17, 18, 60

RC Garantie responsabilité civile. 12, 24, 62–67, 69–87, 111–113

SCO scooter. 78

SMO moto de type super-motard. 78

Table des figures

1.1 Distribution des ventes de deux et trois-roues motorisés en octobre 2019. 13
1.2 Évolution annuelle des immatriculations neuves entre 2016 et 2019.. 13
1.3 Classement des assureurs deux-roues en 2018 issu de l'Argus de l'assurance. . . 14
1.4 Distribution du portefeuille d'assurances AXA par catégorie de deux et trois roues motorisés. 15
1.5 Primes proposées par les assureurs A et B pour deux profils de conducteurs . . . 18
1.6 Parcours de souscription en agence. 21
1.7 Schéma estimation de la rentabilité théorique des prospects. 24

2.1 Schéma méthodes de modélisation. 29
2.2 Exemple de sur-apprentissage. 31
2.3 Dilemme biais-variance. Source : Dreyfus, 2008 32
2.4 Validation simple. 32
2.5 Validation croisée par 4-fold. Source : Guillot, 2015 33
2.6 Calcul de l'indice de Gini comme étant deux fois l'aire A entre la courbe de gain et la bissectrice. Source : Pariente, 2016. 34
2.7 Schéma simplifié d'un arbre de décision. 39

3.1 Schéma de construction de la base de devis deux-roues. 46
3.2 Taux de transformation cumulé par délai (exprimé en mois). 47
3.3 Le taux de transformation par tranche de cotisation TTC. 51
3.4 Le taux de transformation par type de client. 51
3.5 Le taux de transformation en fonction de l'âge du conducteur principal. 52
3.6 Le taux de transformation en fonction de l'usage de la moto. 52
3.7 Influence relative des variables sur le taux de transformation (exprimé en %). . . 54
3.8 Modélisation du taux de transformation de devis : exemples des paramètres β estimés. 56
3.9 Validation du modèle de transformation de devis par la méthode de validation croisée. 58
3.10 Courbes de lift construites sur la base de validation (à gauche) et sur la base totale de devis (à droite). 58

4.1 Types de variables à disposition. 64
4.2 Fréquence RC matérielle observée par année d'étude.. 65
4.3 Fréquence RC matérielle observée selon le genre de moto. 65
4.4 Fréquence RC matérielle observée en fonction de l'âge du conducteur. 66
4.5 Fréquence RC matérielle observée en fonction du nombre d'années d'ancienneté de permis. 66
4.6 Fréquence RC matérielle observée par zone de risque. 67
4.7 Influence relative des variables sur la fréquence des sinistres RC matériels (exprimé en %). 68

4.8 Modélisation de la fréquence des sinistres RC matériels : exemples des paramètres β estimés. 70
4.9 Modélisation de la fréquence des sinistres RC matériels : consistance temporelle pour la variable ancienneté de permis. 71
4.10 Zonier de fréquence pour la garantie RC matérielle. 72
4.11 Validation du modèle par la méthode de validation croisée. 73
4.12 Courbe de lift construite sur la base de validation. 73
4.13 Types de variables à disposition. 75
4.14 Coût moyen sinistre RC matériel observé par année d'étude. 77
4.15 Coût moyen observé des sinistres RC matériels par genre de moto. 77
4.16 Coût moyen observé des sinistres RC matériels par ancienneté du véhicule. . . . 78
4.17 Coût moyen observé des sinistres RC matériels par cylindrée. 79
4.18 Coût moyen observé des sinistres RC matériels par ancienneté de permis. 79
4.19 Distribution du coût moyen des sinistres RC matériels. 80
4.20 Influence relative des variables sur le coût moyen des sinistres RC matériels (exprimé en %). 80
4.21 Modélisation du coût moyen des sinistres RC matériels : exemples des paramètres β estimés. 82
4.22 Modélisation du coût moyen de la garantie RC matérielle : consistence temporelle pour la variable ancienneté du véhicule. 84
4.23 Modélisation du coût moyen de la garantie RC matérielle : validation du modèle par la méthode de validation croisée. 84
4.24 Modélisation du coût moyen de la garantie RC matérielle : courbe de lift construite sur la base de validation. 85

5.1 ELR versus taux de transformation par segment. 92
5.2 ELR et taux de transformation en fonction de l'âge du conducteur. 94
5.3 ELR et taux de transformation en fonction de la zone de risque. 95
5.4 ELR et taux de transformation en fonction de la classe de moto/scooter. 96
5.5 Distribution de devis par niveau d'évolution de la prime commerciale. 98

Liste des tableaux

1.1 Formules de garanties proposées dans le produit moto AXA. 16

2.1 Indicateurs utilisés pour étudier les types de corrélation. 41

3.1 Taux de transformation de devis : les variables analysées. 49
3.2 Modèle de transformation de devis : résumé des variables retenues dans le modèle final. 55
3.3 Principaux indicateurs de performance du modèle de transformation de devis. . 57

4.1 Distribution du nombre de sinistres RC matériels 67
4.2 Résumé des variables retenues dans le modèle de fréquence des sinistres RC matériels. 69
4.3 Principaux indicateurs de performance du modèle. 72
4.4 Résumé des variables retenues dans le modèle final. 82
4.5 Modélisation du coût moyen de la garantie RC matérielle : principaux indicateurs de performance du modèle.. 84
4.6 ELR annualisé vs ELR classique 88
4.7 ELR annualisé vs ELR classique 2 88

5.1 Prime commerciale basée uniquement sur l'ELR 91
5.2 Prime commerciale basée sur l'ELR et la concurrence 92
5.3 Impact de mesures tarifaires par tranche d'évolution de la prime commerciale. . 97
5.4 Synthèse des impacts de mesures tarifaires.. 99