Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 15/11/2021

Par : **Linda Krolikowski**

Titre : **Modélisation de l'élasticité au prix à la souscription en assurance automobile dans le cadre d'une optimisation tarifaire**

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

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

Membres présents du jury de la filière

Entreprise : ADDACTIS France
Nom du référent en entreprise : Quang Do
Signature :
Nom du référent pédagogique ENSAE : Caroline Hillairet
Signature :

Membres présents du jury de l'Institut des Actuaires

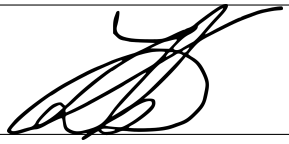
Directeur du mémoire en entreprise : Quang Do
Signature :

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

Secrétariat :

Signature du candidat

Bibliothèque :



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

Remerciements

Je tiens à remercier tous les membres du pôle P&C d'Addactis pour leur soutien et leur bienveillance. Je remercie tout particulièrement Guillaume ROSOLEK de m'avoir accueillie au sein de cette équipe.

Je remercie également Quang DO qui m'a guidée avec une grande expertise sur ce sujet riche par sa technicité et ses champs d'application.

Résumé

La digitalisation croissante accorde un poids plus impactant à la souscription en ligne qui rend la compétition entre acteurs de l'assurance automobile plus ardue. Capter de nouvelles cibles conduit à une course effrénée aux techniques les plus innovantes en termes de critères collectés et de méthodes employées. Poser le tarif adéquat ne se résume pourtant pas à la mesure du risque caractérisée par la prime pure, mais doit tenir compte du positionnement de l'assureur sur le marché, auquel cas atteindre un bon niveau de marge sur l'ensemble du portefeuille n'est pas garanti. Il s'agit ici d'une entrave au principe d'anti-sélection. L'objectif est de modéliser avec précision l'élasticité-prix individuelle pour répondre aux problématiques d'optimisation tarifaire.

La première partie se focalise sur la modélisation du taux de transformation : il s'agit de prendre connaissance du score de conversion pour chaque client. Les facteurs impliqués traditionnellement sont les caractéristiques du conducteur, du véhicule et du contrat auxquelles sont ajoutés le tarif leader du marché pour ces caractéristiques (obtenu par collecte de données automatisée), la prime commerciale TTC et la prime pure (estimée). Ce volet permet d'améliorer la connaissance des stratégies tarifaires mises en place et de comprendre ce qui favorise la réalisation d'un devis et sa transformation en affaire nouvelle. La modélisation du score de conversion individuel ne répond que partiellement à notre problématique qui interroge d'avantage le tarif initialement proposé par l'acteur.

L'élasticité-prix individuelle mise à disposition octroie la possibilité d'évaluer le niveau de marge obtenue pour plusieurs tarifs et constitue alors un outil d'aide à la décision pour l'élaboration de nouvelles politiques tarifaires. Traditionnellement, des price tests sont réalisés pour tenter de représenter le comportement client dans un contexte de prix modifié. Mais ces méthodes sont compliquées à mettre en place et l'évolution dans le temps du comportement d'un profil empêche de se servir du passé. Le modèle de conversion intègre la co-variable de prix pour tenter d'identifier les réactions qui sont propres à un segment. Or, l'estimateur en devient biaisé puisque l'attribution même d'une stratégie dépend des caractéristiques du profil de l'assuré. Cette section envisage d'emprunter une méthode employée massivement dans l'univers médical et social pour l'appliquer au secteur de l'assurance automobile : il s'agit du *propensity score matching*. Cette technique vise à estimer le score d'appartenance à un groupe de stratégie et de faire intervenir un algorithme de jumelage entre les individus des différents groupes à partir de ce score pour assigner un taux de conversion. Différents dérivés sont testés pour améliorer la qualité de l'appariement comme l'estimation du score de propension par *XGBoost* ou le *Genmatch* qui mêle le score de propension et la distance de Mahalanobis.

La grille des taux de transformation individuels effectuée pour chaque prospect, un modèle global de ces taux est construit pour obtenir l'élasticité-prix individuelle. Deux modèles interprétables sont sollicités : les GLM, et, dans le prolongement des GAM (*Generalized Additive Models*), les *Explainable Boosting Machine*.
Mots clés : assurance automobile, prime pure, marge, taux de transformation/conversion, élasticité-prix, optimisation tarifaire, machine learning.

Abstract

Increasing digitalization is giving more impact to online underwriting, which makes the competition between car insurance players more difficult. Capturing new targets leads to a frantic race for the most innovative techniques in terms of criteria collected and methods used. However, setting the right tariff is not just a matter of measuring the risk in terms of the pure premium, but must also take into account the insurer's market positioning, in which case achieving a good margin level on the entire portfolio is not guaranteed. This is an obstacle to the principle of anti-selection. The objective is to accurately model the individual price elasticity in order to address the issues of rate optimization.

The first part focuses on modeling the conversion rate : this involves taking the conversion score for each customer. The factors traditionally involved are the characteristics of the driver, the vehicle and the contract, to which are added the market-leading rate for these characteristics (obtained by automated data collection), the commercial premium including tax and the pure premium (estimated). This component allows us to improve our knowledge of the pricing strategies in place and to understand what encourages a quote and its transformation into new business. The modeling of the individual conversion score only partially responds to our problem, which questions the tariff initially proposed by the actor.

The individual price elasticity made available allows us to evaluate the level of margin obtained for several tariffs and thus constitutes a decision-making tool for the development of new tariff policies. Traditionally, price tests are performed to try to represent customer behavior in a modified price context. However, these methods are complicated to implement and the evolution of a profile's behavior over time makes it impossible to use the past. The conversion model integrates the price co-variate in an attempt to identify reactions that are specific to a segment. However, this biases the estimator since the very attribution of a strategy depends on the characteristics of the insured's profile. This section considers borrowing a method widely used in the medical and social world and applying it to the automobile insurance sector : it is called *propensity score matching*. This technique aims at estimating the score of belonging to a strategy group and at using a matching algorithm between individuals of different groups from this score to assign a conversion rate. Different derivatives are tested to improve the quality of the matching such as the estimation of the propensity score by *XGBoost* or the *Genmatch* which mixes the propensity score and the Mahalanobis distance.

Once the individual transformation rates have been calculated for each prospect, a global model for these scores is built in order to include it in the resolution of the rate optimization equation. Two interpretable models are used : GLMs and, in the extension of GAMs (Generalized Additive Models), Explainable Boosting Machines.

Key words : car insurance, pure premium, margin, transformation/conversion rate, price elasticity, rate optimization, machine learning.

Table des matières

1 Stratégie tarifaire en univers concurrentiel : enjeux et environnement d'étude

1.1 Contexte et objectifs

Addactis est un cabinet de conseil, intervenant sur tous les sujets de l'assurance (Vie – Prévoyance / Santé – IARD (Incendie, Accidents et Risques Divers) ; les travaux réalisés s'intégrant au sein du pôle IARD et plus précisément en assurance automobile.
L'assurance automobile comporte plusieurs garanties : la responsabilité civile matérielle, la responsabilité civile corporelle, le vol, l'incendie, le bris de glaces, les dommages tous accidents, le conducteur responsable. Seules les deux premières sont obligatoires pour tous les automobilistes. La Responsabilité Civile permet de rembourser un tiers suite aux dégâts matériels et corporels causés par l'assuré. Les garanties dommages et conducteur, qui ne sont pas obligatoires, couvrent les dégâts relatifs à l'assuré, causés par l'assuré responsable. La particularité du marché de l'assurance automobile est d'être très concurrentiel, avec des prix d'entrée agressifs permettant d'en faire un produit d'appel pour les assureurs multi-produits. La présence des sites comparateurs renforce la transparence et par cela même la compétition, la comparaison entre les tarifs à la souscription étant désormais aisée.

Le rôle d'un consultant est de résoudre l'ensemble des problématiques actuarielles sur la chaîne de valeur d'un organisme d'assurance.
Les études de ce mémoire ont été impulsées par la volonté d'un acteur d'augmenter son volume de portefeuille tout en améliorant ses indicateurs de performance financiers. Est-ce que le tarif qu'il propose pour chacun de ses segments d'intérêt, compte tenu des garanties, options et couvertures doit être modifié aux vues des offres concurrentes ? Est-il dans l'intérêt de l'assureur qu'il soit modifié ? Quels sont ces principaux leviers existants pour l'amélioration de la conversion ? Sur quel segment la conversion est-elle à améliorer ? L'estimation de l'élasticité au prix apporte des éléments de réponse à ces questions.
Une base devis sera à disposition avec les caractéristiques client, le prix affiché, les frais et la variable d'intérêt, c'est-à dire si oui ou non le devis s'est transformé en affaire nouvelle, donc si le client a finalement fait partie intégrante de la base portefeuille. Le périmètre opté est le marché des particuliers uniquement, sur la formule "Tous risques" du produit automobile qui comporte toutes les garanties (responsabilité civile, bris de glaces, vol/incendie, dommage etc.). L'année 2020 ayant été atypique, l'historique comportera les années 2018 et 2019.

L'élaboration d'une politique tarifaire fait écho à la notion de marge individuelle. Avant de poser un niveau de marge sur un profil de risque il est important d'en déterminer le coût réel. Les actuaires s'attellent à estimer ce coût réel, appelé prime pure, de la manière la plus exacte possible. Une fois le coût estimé, l'actuaire l'agrémente des frais puis détermine un niveau de marge à appliquer : c'est le tarif hors taxes. Une fois les taxes ajoutées, il s'agit du tarif TTC payé par l'assuré. Avoir une bonne vision du risque apparaît comme le moyen le plus certain d'obtenir une rentabilité sur son portefeuille. Cependant, il ne peut être l'unique outil permettant de définir une politique tarifaire sur un prospect.

1.2 Définitions

Sont posés les calculs des indicateurs de profit et de rentabilité dans le cadre d'une simulation de résultat d'une stratégie tarifaire future. Ces formules nécessitent l'estimation de l'élasticité au prix.

$$Profit = \sum_{i=1}^{n} (P_i - S_i) * \hat{f}(P_i, X_i) \tag{1}$$

$$Chiffre\ d'affaires = \sum_{i=1}^{n} P_i * \hat{f}(P_i, X_i) \tag{2}$$

- P_i : la variable endogène qui correspond au tarif Hors Taxes payé par l'assuré i
- S_i : correspond à tous les frais et la charge sinistre de l'assuré i
- $\hat{f}(P_i, X_i)$: la probabilité de conversion estimée de l'assuré i
- X_i : les caractéristiques de l'assuré i (âge, profession, bonus/malus...), les caractéristiques du contrat de i (présence de franchise, d'offres commerciales, de dérogation, marge appliquée...), les caractéristiques du marché pour le profil i (prix du concurrent, écart avec la médiane du marché...)

Le profit s'écrit comme la somme de la marge de chaque client potentiel $(P_i - S_i)$ multipliée par la probabilité $f(P_i, X_i)$ que le devis effectué par lui, au prix P_i, se concrétise en affaire nouvelle. Tandis que l'augmentation du prix accroît l'expression de la marge, elle affecte négativement la conversion.

Lorsque l'on cherche l'allocation optimale de tarifs $(P_1, ..., P_n)$ telle que le profit ou le chiffre d'affaires est maximisé, alors les équations 1 et 2 sont des équations d'optimisation tarifaire (sous des contraintes de volume de portefeuille souhaité par l'assureur) à résoudre par un lagrangien.

Définitions préliminaires :

Prime pure = évaluation moyenne du montant attendu du sinistre

Prime hors taxes = prime pure + frais

Tarif hors taxes = prime pure + frais + marge

Tarif TTC = prime pure + frais + marge + taxes

Rentabilité totale = marge individuelle * nombre de clients

Chiffre d'affaires = primes acquises * nombre de clients

Taux de conversion = $\frac{Nombre\ de\ prospects\ ayant\ souscrit\ \ un\ devis}{Nombre\ de\ devis\ effectués}$

Elasticité-prix = $\frac{\frac{Nombre\ de\ devis\ transformés\ (T+1)\ -\ Nombre\ de\ devis\ transformés\ (T)}{Nombre\ de\ devis\ transformés\ (T)}}{\frac{(Tarif\ (T+1)\ -\ Tarif\ (T)}{Tarif\ (T)}}$

Ratio de sinistralité = $\frac{coût\ des\ indemniations\ +\ charges\ estimées\ correspondant\ aux\ sinistres\ en\ cours}{primes\ acquises}$

1.3 Tarification : le compromis entre segmentation et mutualisation du risque

La manière d'assurer la solvabilité, la liquidité et la rentabilité du secteur de l'assurance s'avère ardue par le fait distinctif de cette industrie qu'est "l'inversion du cycle de la production" : la prime, prix de la prestation, est encaissée avant la fourniture du service et l'événement qui déclenche et justifie cette prestation. La prime commerciale proposée aux assurés englobe la prime pure qui caractérise le coût de l'aléa, les frais et les charges, les taxes et la marge de l'assureur. Chacun de ces éléments est à évaluer le plus justement possible pour que le tarif final s'inscrive au mieux dans la compétition. La précision dans l'estimation de la prime pure, qui est le produit du coût moyen et de la fréquence, traduit la connaissance du risque de l'assureur et s'avère importante à la fois pour la survie de l'entreprise et pour la conquête de nouvelles parts de marché. Les mécanismes d'antisélection et d'aléa moral, bien connu du monde l'assurance, pose les bases du compromis entre mutualisation et segmentation du risque. La compétition entre les acteurs les pousse à investir dans l'amélioration continue des modèles de prime pure afin de bénéficier de la meilleure vision du risque sur le marché et conduit à la pratique d'une quasi-"nanotarification". Les acteurs poursuivent leurs avancées techniques par le biais de l'enrichissement de leur écosystème de données internes -avec des variables innovantes qui vont au-delà des critères traditionnels- et externes -notamment avec les critères de géolocalisation pour le zonage- qui accompagne le phénomène du Big Data, et par la sophistication des outils de modélisation empruntés à la Data Science.

Au sein de cette section, les travaux de Charpentier, Denuit et Elie seront repris, ainsi que

leurs notations [1] (réf. (4)). Un concours (le *Pricing Game*) mettant en compétition des acteurs du monde de l'assurance (assureurs et cabinets de conseil) avait été organisé avec le classement final des modèles de tarification en assurance automobile selon différents critères : le total de primes perçues, le ratio de sinistralité (S/P), la Var(99.5%) et la part de marché. Un portefeuille de 31 000 contrats était mis à disposition et les clients choisissent parmi les tarifs les plus avantageux du marché.

Les conclusions de l'étude ont été apportés par Arthur Charpentier lors de l'évènement annuel "100% actuaires / 100% Data Science" le 5 novembre 2015.

1.3.1 Antisélection et aléa moral

Sans segmentation de la population d'assurés présente sur le marché, le prix actuariellement juste est l'espérance mathématique de la charge sinistre multipliée par la fréquence moyenne, auquel on ajoute des frais, proportionnels à la prime ou fixes. Sans segmentation, les individus n'ont alors aucune incitation à choisir un assureur plutôt qu'un autre. Le partage du marché est équitable et les portefeuilles d'assurés sont de taille égale. La somme des primes versées par l'ensemble des assurés couvre la sinistralité S durant la période couverte.

$$S = \sum_{i=1}^{N} C_i$$

Avec N le nombre de sinistres et C le coût moyen individuel de la sinistralité. L'hypothèse d'un coût unitaire équivalent et fixé à 1000 euros pour tous les types de profil est émise pour plus de simplification :

$$\text{S} = \text{N} * 1000$$

et donc

$$\text{Prime pure} = \text{E[S]} = \text{E[N]} * 1000$$

	Assurés	Assureur
Dépense	E[S]	S - E[S]
Dépense moyenne	E[S]	0
Variance	0	Var(S)

TABLE 1 – Répartition des risques entre l'assureur et les assurés

Sans segmentation, l'assureur porte à lui-seul tout l'aléa car la prime commerciale est fixée comme étant la même pour tous les risques.

Dans le cadre d'un environnement réaliste où les assureurs segmentent leur tarif, ces derniers imposent un tarif bas pour les profils à risque faible et un tarif plus élevé pour les profils à

1. Charpentier, Denuit, Elie, *Pricing Game*, 100% ACTUAIRES / 100% DATA SCIENCE, 2015

risque élevé . Segmenter permet donc à la fois d'attirer les "bons risques" (et de maîtriser les mauvais) et de transférer une partie du risque à l'assuré, l'incitant à être plus prudent.
Il existe en réalité une multitude de profils de risque sur le marché. La population d'assurés est divisée en sous-groupes. En fonction de la granularité de segmentation -la délimitation du risque selon les croyances des assureurs- un profil similaire peut avoir à choisir entre plusieurs tarifs différents. On part du principe qu'il choisira le tarif leader, le plus compétitif du marché. Si un assureur choisit de ne pas segmenter suffisamment, cela revient à dire que dans certains cas, des profils à faible risque se retrouvent avec une sur-tarification puisqu'ils se trouvent dans le même panier que des profils à haut risque : c'est l'antisélection [2] (réf. (1)). Les bas risques peuvent alors se réfugier auprès d'offres concurrentes et l'assureur qui n'a pas suffisamment segmenté se retrouve avec uniquement des profils à haut risques : son produit est donc sous-tarifé (son Loss Ratio risque de dépasser les 100% pour cette catégorie de profil).

	Assurés	Assureur
Dépense	E[S |Ω]	S - E[S |Ω]
Dépense moyenne	E[S]	0
Variance	Var(E[S |Ω])	E[Var(S |Ω)]

TABLE 2 – Répartition des risques entre l'assureur et les assurés en présence d'une segmentation parfaite

Dans le cas où l'assureur aurait une vision parfaite de l'aléa Ω porté par chaque profil d'assuré, un partage du risque est effectué :

$$\mathrm{Var}(S) = \mathrm{Var}(E[S\,|\Omega]) + E[\mathrm{Var}(S\,|\Omega)]$$

En pratique, on est en présence d'une asymétrie d'information car la segmentation n'est jamais parfaite : l'assureur ne dispose pas de la totalité de l'information et les probabilités de survenance peuvent évoluer au cours du temps. Chaque assureur dispose d'une base de caractéristiques X qui lui est propre. Ces croyances sur les classes de risque ne suffisent pas à les tarifer parfaitement. A titre d'exemple, le retrait de la variable *Genre* [3] des modèles due à une nouvelle réglementation contre certaines discriminations occulte une part explicative du comportement de l'assuré. Par ailleurs, ces comportements genrés pourraient converger dans le temps. De la même manière, des changements socio-économiques et politiques (éco-mobilité, taxes carbone, augmentation du niveau de vie...) influencent les comportements de certains groupes de population.

2. Akerlof, *The Market for "Lemons" : Quality Uncertainty and the Market Mechanism, Quarterly Journal of Economics*, p. 488–500, 1970

3. *Gender Directive*, 21 décembre 2012

	Assurés	Assureur
Dépense	E[S $\|X$]	S - E[S $\|X$]
Dépense moyenne	E[S]	0
Variance	Var(E[S $\|X$])	E[Var(S $\|X$)]

TABLE 3 – Répartition des risques entre l'assureur et les assurés en présence d'une segmentation parfaite

La variance de l'assureur est plus importante à cause du manque d'information précise concernant l'aléa, les classes de risque ne sont donc pas parfaitement homogènes.

$$\text{Var(S } |X) = \text{E}[S^2|X] \text{ - } E^2[S|X]$$

$$= \text{E}[\text{E}[S^2|\Omega]|X] \text{ - } E^2[E[S|\Omega]|X]$$

Par ailleurs,

$$\text{Var(E[S}|\Omega]|X) = \text{E}[E^2[S|\Omega]|X] \text{ - } {}^2[E[S|\Omega]|X]$$

et donc,

$$\text{Var(S } |X) = \text{E}[\text{E}[S^2|\Omega]|X] + \text{Var(E[S}|\Omega]|X) \text{ - E}[E^2[S|\Omega]|X]$$

$$\implies \text{Var(S } |X) = \text{E}[\text{Var(S } |\Omega)|X] + Var(E[S|\Omega]|X)$$

$$\implies \text{E[Var(S } |\Omega)] = \text{E[Var(S } |\Omega)] + \text{E[Var(E(S}|\Omega)|X)]$$

Ce résultat est l'élément déclencheur de la course frénétique aux facteurs de discrimination et aux meilleurs techniques de modélisation. Néanmoins, le pouvoir de la segmentation demeure limité au regard de ce qui se passe réellement sur le marché de l'assurance automobile.

1.3.2 Limites de l'antisélection

Les modèles présentés dans le cadre du *Pricing Game* vont des techniques d'économétrie à la Data Science. Malgré les divergences en termes de méthodes optées, les modélisations sont cohérentes aux vues des conjonctures de marché. Chaque modèle privilégie un type de population, selon les cibles jugées d'intérêt.

Certains ont une segmentation très fine et repèrent des niches leur permettant d'obtenir un loss ratio plus faible. En revanche, sa variance est plus importante due au fait qu'un petit portefeuille est plus volatil. Les modèles avec une segmentation grossière ont un loss ratio dégradé mais moins volatil et possède une part de marché plus conséquente. Un bon modèle de prime pure n'est pas suffisant pour atteindre un meilleur niveau de marge.

D'autres conclusions ont été établis à l'issu du concours. Pour les illustrer, considérons le scénario simplifié d'un marché où seuls trois acteurs (A, B et C) et deux facteurs de risque (l'âge du véhicule et de l'assuré) seraient présents :

— Assureur A : tarification en utilisant la moyenne sur l'ensemble des assurés (pas de modèle de prime pure)

— Assureur B : tarification basée sur un bon modèle de prime pure en utilisant une variable (l' âge du conducteur)

— Assureur C : tarification basée sur un modèle de prime pure parfait en utilisant deux variables (l' âge du conducteur et l' âge du véhicule)

Les trois assureurs visent une prime commerciale égale à leur vision de la prime pure ajoutée de 10% de marge. Dans cet exemple simple, la prime commerciale est calculée à partir de la vision de prime pure ajoutée de la marge souhaitée.

					prime proposée		
	# de clients	jeune conducteur	jeune voiture	prime pure réelle	Assureur A	Assureur B	Assureur C
tarif le plus bas du marché	500	1	0	300	1075	990	330
	500	1	1	1500	1075	990	1650
	700	0	0	1000	1075	1146	1100
	500	0	1	1100	1075	1146	1210

FIGURE 1 – Représentation du marché

1. Jeune conducteur/ jeune $voiture^C$:
   Les assureurs A et B ont une mauvaise vision de l'aléa : ce segment est le meilleur en termes de risque et les primes proposées sont beaucoup trop élevées. L'assureur C en revanche aun tarif beaucoup plus juste et attire ces "bons risques". En revanche, au regard du positionnement du marché, C pourrait augmenter sa prime et profiter des moins bonnes performances des modèles concurrents.
2. Jeune conducteur/ jeune voiture :
   L'assureur C est le seul à proposer un tarif cohérent avec le risque de ce profil. L'assureur B et C ont une vision imparfaite du risque : ils proposent une prime commerciale en-deça de la prime pure réelle. Or, comme l'assureur B se positionne en tant que leader, il est préféré aux autres acteurs pour cette population.
3. Jeune $conducteur^C$/ jeune $voiture^C$:
   Les trois assureurs proposent un tarif commercial cohérent compte tenu du risque de cette classe d'assurés. L'assureur A est le leader et rafle cette part de marché. C'est un exemple où la compétition peut être responsable d'un moins bon résultat et pénaliser un bon modèle de tarification.
4. Jeune $conducteur^C$/ jeune voiture :
   Les assureurs B et C ont une meilleure vision du risque que A. L'assureur B affiche cependant un tarif plus attractif que C qui n'est pas suffisamment compétitif. L'assureur A qui se situe en-dessous de la prime pure réelle gagne cette part de marché.

L'assureur C a une vision exacte du risque et affiche une prime commerciale constamment au-dessus de la prime pure réelle. Il capte les jeunes conducteurs mais n'est pas suffisamment attractif pour les deux dernières lignes de profils. L'assureur A qui ne segmente pas et offre

sur le marché un équilibre pooling est déficitaire pour le segment des conducteurs plus âgés possédant un véhicule plus ancien mais se rattrape avec une marge positive sur les conducteurs possédant un véhicule neuf, d'autant plus qu'il représente une part de marché conséquente. L'assureur B qui a une vision partielle du risque se retrouve avec un niveau de chiffre d'affaires négatif.

Assureur A	Assureur B	Assureur C	MARGE A	MARGE B	MARGE C
0	0	500	-	-	15 000
0	500	0	-	- 255 000	-
700	0	0	52 500	-	-
500	0	0	- 12 500	-	-
		total	40 000	- 255 000	15 000

FIGURE 2 – Compte de résultats (vision simplificatrice) des trois assureurs

La mobilisation de ressources suffisantes pour la collecte de données permettant de pallier à la présence d'asymétrie d'information. Complétée de la bonne performance technique d'un modèle de tarification, elle permet d'établir une prime pure précise qui octroie la possibilité de bien se positionner par rapport au marché et de bénéficier d'une marge positive. L'ensemble de caractéristiques n'est pas figé : l'environnement d'assurés est dynamique et évolue dans le temps. De nouveaux segments apparaissent (utilisateurs de voitures électriques, co-voiturage...) et il convient de se tenir informer de l'apparition de nouvelles parts de marché pour mieux répondre à leurs besoins lors de la conception du produit d'assurance et de l'établissement de son prix. La réalisation d'un benchmark sur les offres et les critères tarifaires concurrents esquisse un aperçu des innovations et maintient le contact avec l'actualité du secteur. Néanmoins, la recherche effrénée d'informations concernant l'assuré possède des limites de gain. D'autres critères entrent en jeu dans un univers concurrentiel.

En amont s'effectue une analyse de la cible visée qui doit être cernée au préalable. Dans l'exemple présenté, la catégorie "$jeune\ conducteur^C/jeune\ véhicule^C$" est certes plus risquée que la catégorie "$jeune\ conducteur/jeune\ véhicule^C$" mais représente une plus grande part du marché.
Une fois les modèles de prime pure et de frais établis et la détermination de la valeur de chaque profil effectuée, le travail repose sur l'optimisation de la marge individuelle. Deux facteurs principaux sont à prendre en compte :

1. Le positionnement des concurrents.
2. La sensibilité tarifaire propre à chaque assuré.

Le terme "concurrents" correspond aux acteurs visant les mêmes cibles ou proposant des offres de couverture similaire ou alors, possédant les mêmes caractéristiques de ressources financières et opérationnelles. Comme mentionné précédemment, si l'assureur C avait connu le positionnement de A et de B pour la catégorie $jeune\ conducteur/jeune\ voiture^C$, il aurait pu accroître son chiffre d'affaires et sa rentabilité sans affecter sa part de marché sur cette cible.

Le deuxième facteur se rapporte au comportement client face aux différents tarifs lors de la souscription. De quelle ampleur doit être la variation du tarif pour atteindre un certain volume dans le portefeuille d'un client cible ? A partir de quel écart avec le leader du marché, le produit d'assurance ne devient plus attractif ? Quel seuil le tarif ne peut franchir sur un profil au risque de perdre des parts de marché ?

1.4 Du taux de transformation à l'élasticité au prix

1.4.1 Manque de données et absence d'A/B testing

En premier lieu, il s'agit d'estimer le taux de conversion pour chacun des prospects i. Dans la base des devis, cela correspond à l'estimation de Y à partir des caractéristiques X et du prix P :

$$Y = \begin{cases} 1 & si\ l'individu\ souscrit, \\ 0 & sinon. \end{cases} \tag{3}$$

Cependant, la détermination d'une probabilité de transformation sur chacun des profils de la base devis n'est pas suffisant pour établir une nouvelle politique tarifaire : que se passe t-il si l'assureur souhaite appliquer un nouveau tarif à un type de profil ? Lorsque la question de savoir si oui ou non les tarifs proposés sont optimaux, le besoin de tester de nouveaux tarifs aux sujets est implicite. Or, si on prend l'exemple d'un jeune parisien ayant un tarif de 33 euros et à qui on veut appliquer un tarif de 39 euros et connaître sa réaction face à cette évolution de la prime, nous sommes face à un manque de données. Il s'avère impossible de remplacer dans l'équation du modèle de Y l'ancien prix par le nouveau au niveau des inputs puisque le coefficient b_{prix} relié au prix a été estimé avec des données d'apprentissage ne comportant pas de jeunes parisiens ayant un prix de 39 euros.

$$Y_i = b'X_i + b_{prix} * P_i$$

Substituer simplement 33 par 39 euros dans les inputs conduit directement à un biais de sélection puisque dans la base devis, les prospects ayant un tarif de 39 euros correspondent à un autre type de profil de risque. Ici, l'exemple a été donné avec un modèle glm pour expliquer la probabilité de conversion mais cette problématique s'étend à tous types d'algorithmes : il faut alors faire face à un manque de données dans la base d'apprentissage.

L'idéal serait de pouvoir réaliser un price test pour obtenir un bon modèle d'élasticité. Le price test est un exercice consistant à simuler des conditions commerciales plausibles afin d'en étudier les conséquences sur la demande. Par exemple, affecter un tarif de 39 euros sur les conducteurs parisiens durant un laps de temps court mais assez important pour collecter suffisamment de devis et déterminer ainsi le nombre de souscrits parmi eux. Le taux de transformation estimé sur ce laps de temps est proche de la réalité car il provient de données observées sur le marché. Cette technique est difficile à mettre en oeuvre car elle affecterait le bilan de par l'engagement de l'assureur auprès de ses assurés. La baisse de tarif d'une cible doit être compensée par une hausse sur un même nombre de prospects durant l'exercice. Aussi, rien ne peut justifier la différence de tarifs entre deux profils similaires ayant souscrits à la même période. Une difficulté opérationnelle est également présente puisque le processus nécessite la participation active des

réseaux de distribution de l'assureur.

Pour répondre à cette problématique et pouvoir proposer différentes fourchettes de tarifs en connaissant la conversion en résultant, sans passer par un quelconque price test , des techniques mathématiques seront employées notamment le jumelage par score de propension ou *propensity score matching*.

1.4.2 Principe du jumelage par score de propension

Cette technique est largement employée dans l'univers médical pour mesurer l'effet d'un traitement par rapport à un autre sur des malades. Dans le cadre de cette étude, l'effet d'un tarif par rapport à un autre sur la demande est recherché.
Pour chaque ligne de la base devis, la probabilité de transformation $\hat{Y}$ est supposée déjà estimée.
A des sujets de la base devis, appartenant à une stratégie tarifaire x (niveau de marge x), l'assureur souhaite connaître le taux de transformation s'ils appartenaient à une stratégie y. La première étape vise à constituer deux groupes : les prospects d'intérêt, ceux appartenant à la stratégie x, et les prospects de la base devis faisant déjà partis de la stratégie y mais ayant des profils différents de ceux de la stratégie x. Ensuite, pour les sujets des deux groupes, on estime leur probabilité d'appartenance au groupe x à partir de leurs caractéristiques X : le score de propension. Tour à tour, les assurés du groupe ayant un niveau de marge correspondant à la stratégie x seront jumelés à un assuré du groupe y, celui qui a le score de propension le plus proche du sien. Cette technique sous-entend que puisqu'ils ont la même probabilité d'appartenir au groupe x alors ils possèdent un support commun de caractéristiques essentiel pour la détermination d'une stratégie, et donc leur comportement face à une politique tarifaire est similaire. Le taux de transformation du jumeau du groupe y est attribué au sujet jumelé du groupe x. Les sujets du groupe x sont donc dédoublés : chacun possède deux lignes avec les mêmes variables de profil, seuls le prix et le taux de transformation sont modifiés. In fine, c'est l'ensemble de ces taux qui seront modélisés, avec en base d'apprentissage, la nouvelle base devis avec ce surcroît d'observations obtenues après jumelage.

Le protocole à effectuer est le suivant :

1. Estimer de manière la plus exacte possible la probabilité de transformation sur chaque ligne de la base de devis. Il s'agit de créer une série de taux quasiment observés qui seront utilisés lors du jumelage et estimés pour déterminer l'élasticité au prix.
2. Regrouper les sujets de la base devis en fonction de la politique tarifaire qui leur est appliquée.
3. Jumeler les sujets entre ces stratégies pour connaître l'élasticité-prix individuelle de chaque prospect si on lui appliquait un autre tarif
4. Modéliser le score de conversion avec ces nouvelles données pour obtenir un modèle d'élasticité-prix individuelle.

2 Estimation de la probabilité de conversion

Pour chacun de ces individus, un score de transformation, c'est-à dire une estimation de la probabilité de souscrire auprès de notre assureur est recherchée. In fine, chaque individu devra posséder un score de conversion pour différentes stratégies existantes, en s'appuyant sur l'appariement des sujets entre les groupes de stratégie. Attribuer un score de conversion à chaque sujet revient à la création de probabilités observables à partir de la valeur cible. Ces probabilités observées seront ensuite modélisées en vue de fournir une prédiction du score à n'importe quel assuré présent sur le marché et pour différents prix (modélisation de l'élasticité au prix). C'est à cette dernière étape que le pouvoir prédictif détient une importance capitale. En revanche, l'enjeu dans cette section est d'avoir la meilleure qualité de précision dans l'approche de la probabilité de transformation : seule la performance d'apprentissage est considérée.

2.1 Équation de la demande et premières analyses

2.1.1 Base devis

Le périmètre de l'étude concerne la formule "Tous risques" du produit automobile chez les particuliers. La base regroupe l'ensemble des devis réalisés sur les années 2018 et 2019 comprenant les caractéristiques du conducteur et de son véhicule, du contrat et si, oui ou non, le devis s'est transformé en affaire nouvelle.
Est notée Y la variable qui décrit si l'assuré transforme son devis en souscription ou non.

Donc

$$Y = \begin{cases} 1 & si\ l'individu\ souscrit, \\ 0 & sinon. \end{cases} \tag{4}$$

Comptage des devis transformés

Au sein de la base devis, chaque ligne représente un contrat. Chaque contrat possède un numéro d'identification unique. Plusieurs contrats peuvent en revanche correspondre à la même personne : un client a la possibilité d'effectuer plusieurs devis car il se voit proposer des offres promotionnelles différentes. Les seuls contrats considérés sont ceux de la garantie « Tous risques».

Un contrat est converti si : -le code du contrat figure dans la base portefeuille d'affaires nouvelles correspondant à l'année en cours

Une personne est convertie si : -le code de la personne figure dans la base portefeuille d'affaires nouvelles correspondant à l'année en cours

Une ligne peut alors être convertie dans le sens de la personne mais pas du contrat lorsqu'une personne effectue plusieurs devis mais ne souscrit qu'à l'un d'entre eux. Si deux lignes ont exactement le même code personne, le même code de véhicule et exactement le même contrat (tarif TTC, franchise, réductions) alors le contrat non converti parmi les deux est définit comme doublon et est supprimé de l'étude. En revanche, et il s'agit du cas le plus courant, une même personne peut se voir proposer deux contrats « Tous risques » aux franchises, aux promotions

distinctes qui impactent directement le tarif annuel. Environ 15% des personnes présentes dans les bases devis entrent dans cette configuration. Dans ce cas, il est utile de conserver les deux contrats : ce sont les variables relatant des offres promotionnelles et des tarifs qui vont départager les contrats convertis et les contrats non convertis.

L'intérêt de la modélisation du taux de conversion est la détermination d'un score de transformation, et non pas la création d'un classificateur. Ainsi, si un assuré a sous les yeux un devis à 3 euros (franchise à 3 euros) et un devis à 5 euros (franchise à 5 euros) et qu'il choisit la première option, alors le modèle est valide si la différence entre la probabilité de conversion estimée pour le premier devis et la probabilité de conversion estimée pour le second devis est suffisamment grande.

Le taux de transformation moyen de la base devis regroupant les années 2018 et 2019 est de 37%. Le taux de transformation est supérieure en 2019 avec 10000 devis supplémentaires pour cette année.

Pour évaluer le taux de conversion du client, il faut comprendre les facteurs qui appuient la conversion d'un devis en affaire nouvelle. Plusieurs types de données sont nécessaires :

1. Les caractéristiques liées au profil de l'assuré (âge, profession, bonus/malus...). Ces données sont présentes dans la base devis.
2. Les caractéristiques liées au véhicule (âge du véhicule, puissance, marque...). Ces données sont présentes dans la base devis.
3. Les caractéristiques liées au contrat (tarif TTC, garantie, franchise, réductions...). Ces données sont présentes dans la base devis.
4. Les caractéristiques liées au positionnement de l'assureur sur le marché (tarifs des concurrents pour chaque prospect, marge de l'assureur...). La stratégie de l'assureur est retranscrite par le niveau de marge sur chacun des devis. L'obtention de cette donnée nécessite l'évaluation de la prime pure et le montant des frais. La base devis sera également alimentée par des tarifs concurrentiels, en particulier le tarif le plus agressif du marché et le tarif médian du marché.

2.1.2 Intégration de données concurrentielles

La modélisation de la réaction des souscripteurs face à la concurrence nécessite l'intégration des prix, à la fois de la concurrence et de l'assureur dans l'équation de demande.

Une base marché, constituée de profils fictifs représentatifs du marché, a été créée dans le but de collecter des données tarifaires externes provenant de sites comparateurs. Le tarif leader (le tarif minimal rencontré pour chacun des profils fictifs) et le tarif médian (calculé à partir de l'ensemble des tarifs collectés) constituent des baromètres auxquels l'assureur doit se comparer. La méthodologie est la suivante :

1. Création de profils représentatifs du marché.
2. Collecte de données automatisées sur un site comparateur.

3. Mapping des variables entre la base marché et la base devis.
4. Modélisation du tarif leader et du tarif médian.

L'ensemble des travaux relatifs à l'intégration de données concurrentielles est présenté en annexe (**?**) afin de ne pas alourdir le rapport. Y est détaillée l'ensemble de la méthodologie évoquée.

La prime pure a été modélisée afin de s'assurer une vision du risque homogène entre les prospects, en appliquant un glm sur l'ensemble des garanties proposées. Les frais étant donnés, la marge de l'assureur a pu être obtenue avec le concours du montant hors taxes soustrait des frais et de la prime pure modélisée.

Analyse exploratoire préliminaire

La présentation des comportements évolutifs du tarif commercial de l'assureur, du tarif leader, du taux de conversion et de la marge en fonction des segments d'assurés offre un premier aperçu des politiques tarifaires de l'assureur et de la réaction associée de ses clients potentiels. Pour rappel :

— Le tarif leader correspond au prix commercial TTC le plus compétitif du marché proposé au client.
— La tarif commercial TTC est le prix proposé au client par l'assureur.
— Le tarif commercial HT correspond au prix commercial TTC retiré des taxes.
— La prime pure représente le risque exact du client vu par l'assureur.
— La prime HT est la prime pure augmentée des frais.
— La marge est la différence entre le tarif commercial HT et la prime HT rapportée à la prime commercial HT. Lorsque l'année n'est pas spécifiée, il s'agit de la moyenne entre 2018 et 2019.
— Le taux de transformation est la moyenne des sujets convertis sur un segment donné. Lorsque l'année n'est pas spécifiée, il s'agit de la moyenne entre 2018 et 2019.

Les chiffres ont été bruités ou ré-échelonnés dans un but de confidentialité.

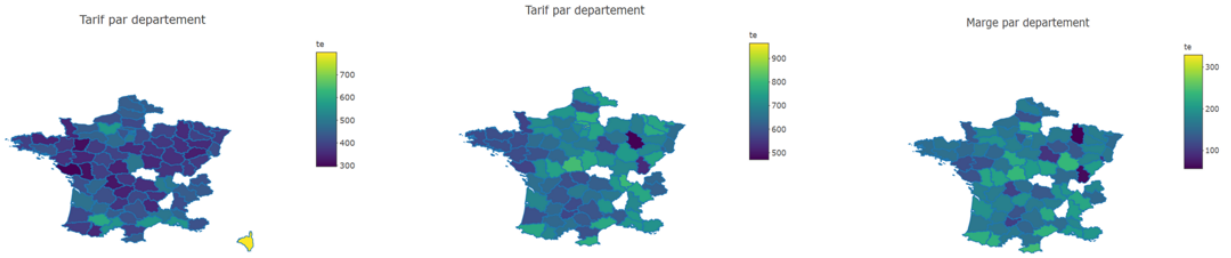


FIGURE 3 – Représentation du tarif leader (gauche), du tarif commercial TTC (centre) et de la marge (droite). Ici, la marge n'est pas en pourcentage, il s'agit de la simple différence entre le tarif commercial HT et la prime HT (différence entre la carte centrale et celle de gauche). Les marges sont les plus élevées pour les départements de la Côte-d'Or, de l'Indre et de l'Oise. En revanche, l'assureur est plus compétitif dans le Jura, la Meuse, le sud de la Bretagne et le Pays-de-la-Loire.

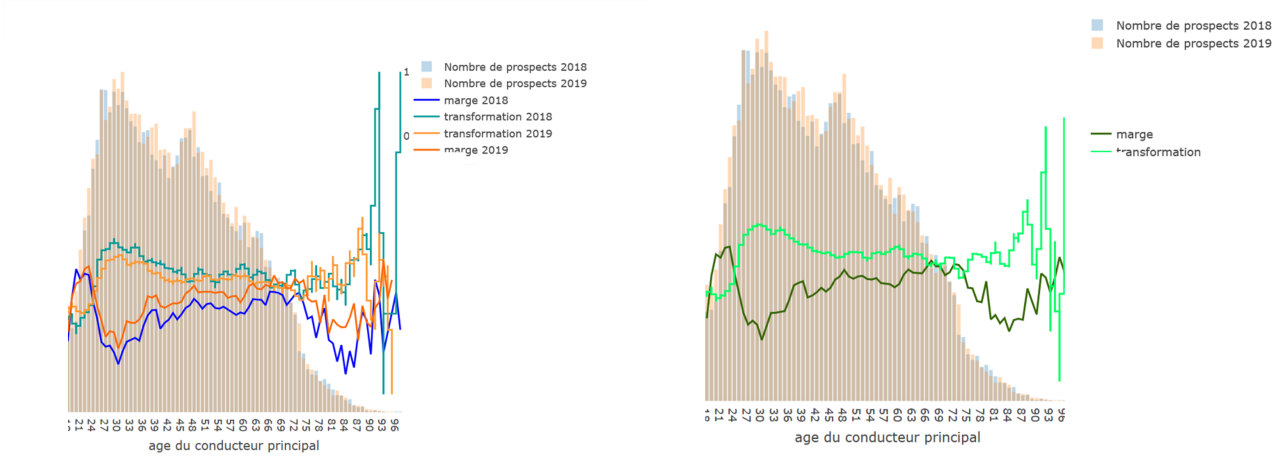


FIGURE 4 – Une demande de devis accrue est décelée entre 2018 et 2019, notamment dans les segments qui se convertissent davantage. Ceci explique l'amélioration du taux de conversion entre les deux années. La corrélation négative entre le taux de transformation et le niveau de marge est visible au sein des analyses des différents segments. De surcroît, ce lien entre les deux variables est vérifié au cours du temps : la marge augmente sur l'ensemble des profils, le taux de transformation est détérioré entre 2018 et 2019. La transformation est plus élevée pour les âges entre 24 et 45 ans, et c'est aussi pour ces âges que sont rencontrés les taux de marge les plus faibles : le pic a lieu pour les conducteurs de 30 ans avec un taux de transformation de 41,2% et une marge de 12,1% par rapport à la prime commerciale hors taxe. Le produit devient donc intéressant pour ces âges. Il est également avantageux pour les personnes entre 73 et 88 ans avec un deuxième pic à 84 ans qui regroupe un pourcentage de conversion de 32,5% et une marge de 14,2%. Les résultats sont néanmoins plus volatils car les grands âges sont sous-représentés dans l'échantillon. Ce sont les jeunes conducteurs qui enregistrent les marges les plus hautes qui dissuadent la souscription.

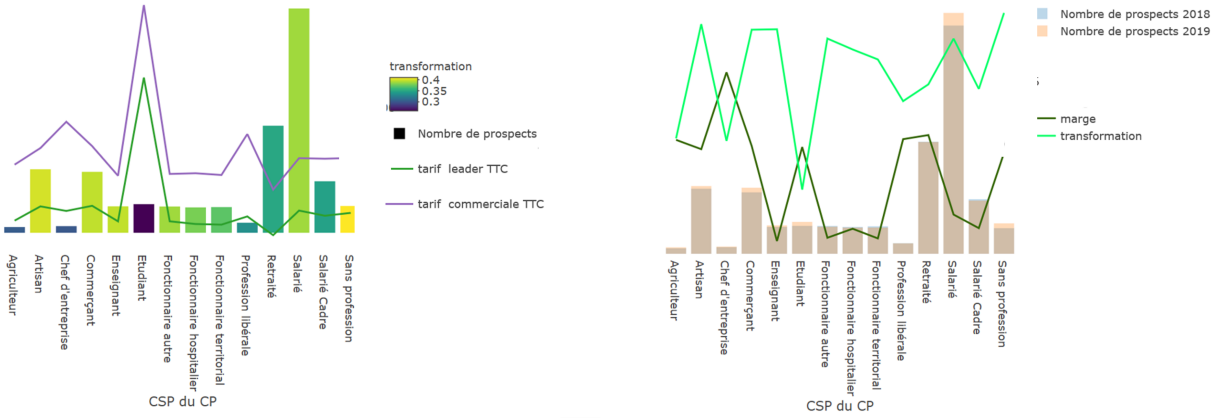


FIGURE 5 – Le produit est avantageux pour les fonctionnaires, les salariés et les salariés cadres. Les taux de conversion dépassent les 32% pour ces catégories. Les retraités et les personnes sans profession ont un bon taux de transformation qui avoisine les 32% en comparaison avec les autres professions. En revanche les étudiants, les chefs d'entreprises et les professions libérales sont les moins concernés par l'offre, possédant les montants de prime TTC et de marge les plus élevées. Le tarif leader du marché prévoit pourtant un tarif préférentiel pour les chefs d'entreprise par rapport aux artisans ou des commerçants. Même esprit pour les professions libérales dont le tarif demeure en-dessous de celui du salarié ou du salarié cadre. Les csp les plus présentes dans la base sont les salariés et les retraités (ils vont s'intéresser davantage au devis de l'assureur par rapport aux autres classes), ce qui justifie en partie la sélection tarifaire de l'assureur et le délaissement de certains segments.

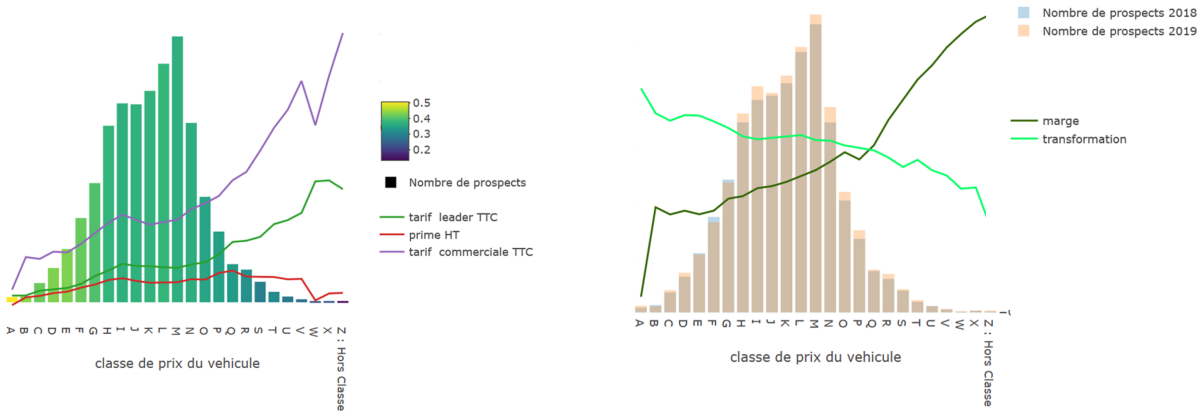


FIGURE 6 – D'avantage de voitures de moyenne gamme sont représentées parmi les devis de 2018 et 2019. Les voitures bas de gamme enregistrent un plus fort taux de transformation. C'est directement en lien avec la marge effectuée. La chute du taux de conversion de 47 à 39% entre les gammes A et B s'explique par une marge négative pour la gamme A. Il s'agit d'une anomalie : certes, le taux de souscription est fort mais les clients assurés de cette catégorie ne sont pas rentables. Le faible volume de cette classe de véhicules explique cette négligence. De même, la catégorie W présente un caractère spécifique puisque le risque est estimé moins important que pour la catégorie V alors que le coût moyen est certainement plus élevé pour un véhicule de meilleure gamme, étant plus onéreuse à la réparation. Cette catégorie doit être corrélée à une autre variable qui amoindrie la fréquence. Le tarif leader, lui, continue sa progression mais redescend brutalement une fois arrivé aux automobiles hors classes, variation souvent justifiée par un garage mort. Ici, l'analyse exploratoire ne tient pas compte des intéractions et ne mesure pas l'impacte d'un segment sur la prédiction du taux de transformation. L'effet montré ici n'est pas marginal, ce qui installe des mouvements inexpliqués dans la définition du tarif final.

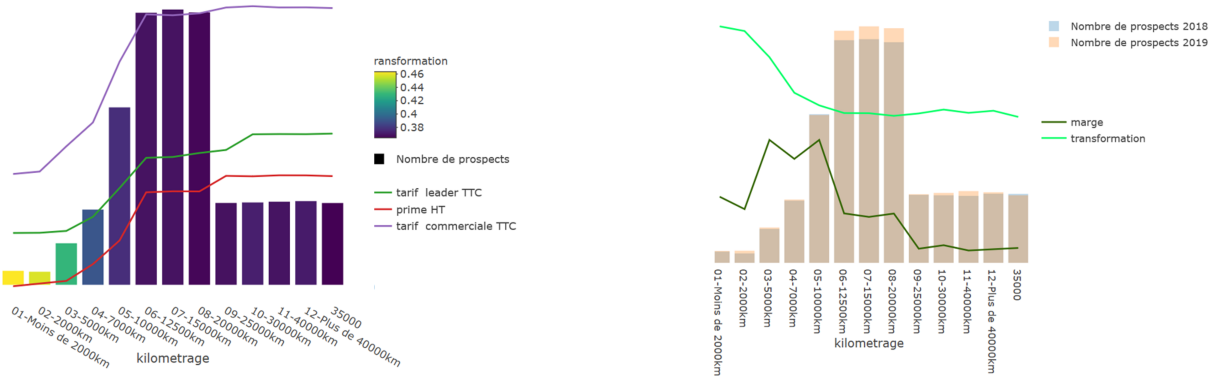


FIGURE 7 – Le produit semble correspondre aux petits rouleurs puisque leur taux de souscription est plus élevé. C'est aussi pour les petits rouleurs que l'offre est la plus avantageuse. Le tarif se stabilise à partir de 12500 km moyen annuel, tous les gros rouleurs connaissent le même tarif et bénéficient de la même marge. Opter pour un découpage plus précis peut être intéressant pour capter le risque d'une plus forte utilisation du véhicule qui intervient dans la dégradation de la fréquence et attirer davantage les rouleurs moins risqués. Le leader émet quant à lui une distinction entre le parcours de 25000 et 30000 km. De plus, l'estimation de la prime pure indique bien une distinction de risque entre les personnes parcourant 20000 et 25000 km par an.

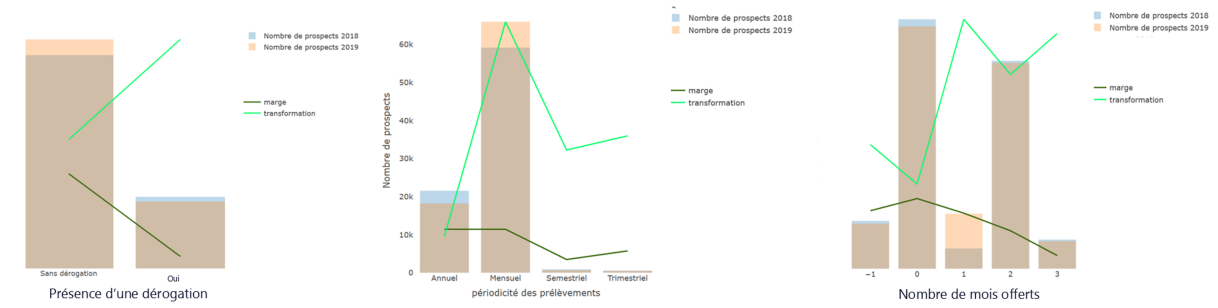


FIGURE 8 – La dérogation octroie une réduction ponctuelle de la première prime annuelle. Pour le nombre de mois offerts la première année, -1 est une donnée manquante (peut être assimilé à 0 mois de réduction ici). Une dérogation est accordée aux assurés optant pour un paiement annuel plutôt que mensuel, afin de garantir le règlement de la prime. Néanmoins, la transformation est la plus élevée lorsque le paiement de la prime est dilué sur plusieurs mois. L'assureur est tenté d'expliquer ce phénomène par un effet psychologique qui pousse l'assuré à penser que la somme de petits paiements est plus simple à gérer financièrement qu'un paiement unique, même si celui-ci est plus faible. De surcroît, tous les ménages ne sont habilités à avancer un montant important.

L'assureur est de caractère très prudent. Dès qu'un profil de conducteur possède un risque plus élevé, la marge augmente de manière considérable. C'est le cas pour les gammes de prix de véhicule à partir de la classe Q, les âges jeunes, les étudiants (associés à la notion de conducteur novice), les retraités et des grands rouleurs.

2.2 Critères de sélection d'un modèle

2.2.1 Performance

Les taux de transformation estimés pour chacun des devis dans cette section seront attribués à d'autres sujets ayant une politique tarifaire différente au moment du jumelage de sujets, puis estimés à nouveau par un modèle pour obtenir la sensibilité au prix individuelle de chaque prospect réalisant un devis particulier. Il convient alors d'avoir la meilleure estimation possible pour ne pas perdre de l'information lors des prochaines étapes.

Il est nécessaire de rappeler que la modélisation du score de transformation n'implique pas la performance de classification. En effet, un sujet converti peut avoir une probabilité de conversion en-deça du seuil de classification (et donc être prédit comme non converti), de par une faible conversion chez les personnes similaires, sans heurter la validité et le bon fonctionnement du modèle. L'important est d'obtenir une hiérarchie de probabilités entre les profils. Pour mesurer la qualité de la prédiction, une analyse segment par segment, sur lesquels figureront les scores estimés et observés, sera effectuée. En outre certains critères quantitatifs comme l'AUC, la comparaison des moyennes de prédiction et l'indice de gini seront abordés.

Lorsqu'une excellente adéquation du modèle aux données est requise, les méthodes "boîtes noires", comme le XGBoost, sont habituellement sollicitées, au détriment de certaines propriétés d'interprétabilité.

2.2.2 Interprétabilité

Lors de la modélisation d'un phénomène, l'un des critères les plus analysés concerne le pouvoir prédictif du système, représenté par la fonction de perte. L'un des points majeurs que doit cocher le classificateur est d'être capable d'approcher la probabilité se souscrire d'un assuré donné. Les résultats doivent être stables et s'adapter à une faible variation au sein de l'échantillon. La qualité de la classification fait figure d'élément de comparaison principal entre plusieurs modèles et méthodologies.

Dans un même temps, comprendre les raisons qui ont contribué à une certaine prédiction est tout aussi crucial que d'obtenir un score de prédiction exacte. L'assureur, dans l'estimation du score de conversion, attend des réponses à ses questions afin de confirmer ses intuitions ou apprendre de nouveaux liens de causalité.

— Quelles sont les caractéristiques client impactant davantage la souscription ? En d'autres termes, quels sont les attributs qui augmentent la probabilité de souscrire ?

— Quels sont les segments les plus récalcitrants ?

A partir de la causalité entre inputs et output instaurée par le modèle, l'assureur déduit des propriétés afin d'élaborer une stratégie tarifaire adéquate à partir de l'état des lieux de la situation. Le fonctionnement interne du classificateur constitue alors une aide à la décision et la précision descriptive arrive à hauteur de la précision prédictive dans l'échelle de nos besoins. Par ailleurs, l'intelligibilité du système de prédiction permet de détecter des problèmes de biais par

un regard expert et de les piloter, en collectant de nouvelles données (variables ou individus) ou en les retirant (étude de corrélations...)

Cependant, les modèles complexes qui permettent les meilleures performances sont également les moins intelligibles : c'est le cas des méthodes XGBoost ou des réseaux de neurones, considérées comme des "boîtes noires". Il s'agit du fameux compromis entre précision et interprétabilité. De nouveaux algorithmes sont disponibles pour apporter un éclaircissement quant au rôle des variables dans la prédiction de Y. En revanche, ils ne répondent pas nécessairement à tous les critères d'interprétabilité fixés par l'assureur, en réponse à ses besoins. Les méthodes "boîtes blanches" sont définis ainsi car l'utilisateur peut voir et raisonner sur le fonctionnement interne du système. Deux critères sont respectés au sein d'un modèle dit *glassbox* :

— Simulabilité : à partir des résultats obtenus par le modèle, l'humain doit pouvoir retrouver la prédiction.

— Modularité : une portion significative du processus peut être interprétée indépendemment.

Autrement, le modèle est interprétable si les inputs peuvent être interprétés indépendemment des autres et que les termes sont en unités interprétables. C'est le cas des modèles additifs, où les contributions de chaque variable sont additionnées pour aboutir à la prédiction.

Le modèle linéaire généralisé répond aux critères d'interprétabilité, ce qui justifie son utilisation répandue dans le monde de l'actuariat. Cependant il peut s'avérer ardu à calibrer : de la sélection des variables à leur segmentation en passant par la détection d'interactions ; il paraît difficile de concilier ce type de modèle avec la contrainte de temps imparti et la contrainte de performance. Un autre modèle interprétable, l'*Explainable Boosting Machine*, satisfaisant d'avantage ces contraintes opérationnelles, est présenté dans cette section.

2.3 Modèle linéaire généralisé classique

2.3.1 Cadre du modèle linéaire classique

Nous disposons d'un échantillon de données $(x_i, Y_i)_{i=1,...,n}$, chaque couple étant à valeurs dans $\mathbb{R}^p \times \mathbb{R}$ et identiquement distribué. Nous travaillons dans le cas du modèle linéaire gaussien : Y = Xβ + ϵ où $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ avec $\epsilon = (\epsilon_1, ..., \epsilon_n)$' , $Y = (Y_1, ..., Y_n)'$ et $X = (x'_1, ..., x'_p) \in$R^{n*p}. On écrit alors :

$$\hat{Y} = X\hat{\beta} \tag{5}$$

qui est un estimateur sans biais et consistent d'après les propriétés du terme d'erreur. Parmi les hypothèses du modèle linéaire classique, celles numérotées (1), (4), (5) ne sont alors pas vérifiées par notre output :

1. La relation entre l'espérance de la variable à expliquer et les variables explicatives est une relation linéaire.
2. Les observations sont indépendantes.
3. E[ϵ|X] = 0

4. La variance des variables aléatoires représentant les observations est constante (homocédasticité).
5. Les observations sont distribuées suivant une loi normale. Cette hypothèse est notamment essentielle pour réaliser les tests. Grâce au théorème central limite, le modèle linéaire est robuste aux écarts à la normalité.

Les modèles linéaires sont mal adaptés pour la problématique étudiée. En effet, si Y $\in \{0, 1\}$, on a

$$E(Y|X) = P(Y = 1|X) \in [0, 1] \tag{6}$$

Dans un modèle linéaire, sous l'hypothèse (3) d'exogénéité,on a E(Y |X) = Xβ. Mais rien n'assure que X$\beta \in$ [0, 1].
Dans notre cas, la variance varie en fonction de la moyenne et diffère par classe de risque. Nos observations sont issues d'une loi discrète et l'hypothèse de normalité semble intenable.
Pour que l'équation (4) soit satisfaite, on va supposer que

$$E(Y|X) = F(X\beta), \tag{7}$$

où F(.) est une fonction (connue) strictement croissante bijective de R dans]0, 1[, donc une fonction de répartition. C'est l'équation d'un modèle linéaire généralisé, c'est-à-dire un modèle de la forme :

$$g(E(Y_i|X_i)) = X_i\beta \tag{8}$$

La technique est de proposer une transformation des données pour rendre la relation linéaire avec $g(E(Y|X)) \in$. Cela permet d'envisager l'intégration d'observations de nature variée comme des données de conversion.

2.3.2 Famille exponentielle naturelle

En identifiant la famille exponentielle naturelle de notre cible Y, ses propriétés permettront de déterminer la fonction de lien g^{-1} définie plus haut.
Soit P_Y la loi de probabilité de la variable Y. Y appartient à la famille exponentielle naturelle si elle s'écrit sous la forme

$$P_Y(y) = exp(a(\phi)(y\theta - b(\theta)) + c(y, \phi)) \tag{9}$$

où c est une fonction dérivable, b est trois fois dérivable et sa dérivée première b'est inversible. Le paramètre θ est appelé paramètre naturel de la loi. ϕ est un paramètre appelé paramètre de nuisance ou de dispersion. Si la distribution P_Y appartient à la famille exponentielle naturelle, alors :

— E(Y) = μ = b(θ)
— V(Y) = a(ϕ)b"(θ)

Pour montrer qu'une loi de probabilité appartient à la famille exponentielle naturelle, il suffit de l'écrire sous la forme d'une exponentielle et d'identifier les termes. Dans le cas de la loi de Bernoulli :

$$P[Y = y; p] = p^y * (1-p)^{1-y} = e^{y*log(p)+(1-y)*log(1-p)} \tag{10}$$

ce qui donne :

— $\theta = log(\frac{p}{1-p})$
— b(θ) = $log(1-p)$ = log (1 + exp(θ))
— a(ϕ) = 1
— c = 0

D'après l'expression de l'espérance μ, cela revient à choisir g(μ) = $b'^{-1}(\mu)$.
Dans le cas des données binaires suivant une loi de Bernoulli, la fonction de lien naturel est la fonction logit, définie par logit(p) = $log(p/1-p)$.

Le modèle linéaire s'écrit alors comme suit :

$$\text{g(E[Y])} = log(\tfrac{p}{1-p}) = \beta_0 + \sum_{k=1}^{p} \beta_k X_k + \epsilon$$

2.3.3 Estimation des paramètres

Le modèle étant posé, il s'agit d'estimer le vecteur de paramètres $\beta = (\beta_1, \ldots, \beta_p)$ et le paramètre de dispersion ϕ. Notons que ce dernier paramètre n'est le plus souvent pas le paramètre d'intérêt, il n'apparaît pas en effet dans la partie explicative (i.e. l'espérance). Nous utilisons ici la méthode classique d'estimation du maximum de vraisemblance.
La vraisemblance en y s'écrit :

$$L(y;\beta,\phi) = \prod_{i=1}^{n} f(y_i;\theta_i,\phi) = \prod_{i=1}^{n} f(y_i;x_i\beta,\phi) \tag{11}$$

et la log-vraisemblance :

$$l(y;\beta,\phi) = \sum_{i=1}^{n} a(\phi) * (y_i x_i \beta - b(xi\beta)) + c(y_i,\phi) \tag{12}$$

Les valeurs deβ et de ϕ qui rendent maximale cette fonction de log-vraisemblance sont solutions du système d'équations aux dérivées partielles suivant :

— $\frac{\partial l(y;\beta,\phi)}{\partial \beta_j} = 0$ pour j = 1, . . . , p
— $\frac{\partial l(y;,)}{\partial \phi} = 0$

D'après l'expression de la log-vraisemblance donnée ci-dessus, on peut remarquer que la maximisation de la log-vraisemblance en β ne dépend pas de ϕ. Il n'y a pas d' expression explicite pour les estimateurs : pour obtenir les estimations du maximum de vraisemblance, on a recours à des algorithmes d'optimisation itératifs.

A partir d'une estimation $\hat{\beta}$ de β, on obtient la prédiction par le modèle au point x_k qui est tout simplement l'estimation de la moyenne :

$$\mu_k = \hat{p}_k = g^{-1}(\hat{\theta}_k) = x_k\hat{\beta} \tag{13}$$

On obtient alors finalement :

$$\hat{Y_k} = \begin{cases} 1 & si\ \hat{p_k} > q, \\ 0 & sinon. \end{cases} \tag{14}$$

Le seuil q serait égal au taux de souscription moyen du portefeuille qui est de 37% ou égal à la valeur du seuil de Bayes classique qui est de 50%. De manière générale, la classification n'est pas l'objet d'intérêt ici, estimer des probabilités de conversion μ_k cohérente par profil est privilégié.

2.3.4 Interprétabilité

Dans les modèles binaires, l'effet marginal de X_j n'est plus β_j, et il dépend de x :

$$\frac{\partial E(Y|X=x)}{\partial_j} = f(x'\beta)\beta_j \; avec \; f = F \tag{15}$$

Cependant, la comparaison des différents paramètres est licite :

$$\frac{\beta_i}{\beta_j} = \frac{\frac{\partial E(Y|X=x)}{\partial x_i}}{\frac{\partial E(Y|X=x)}{\partial x_j}} \tag{16}$$

Pour faciliter les comparaisons, un ré-échelonnage des données continues est souvent appliqué pour esquiver les conclusions hasardeuses quant aux écarts entre les coefficients.
Dans le cadre de la distribution de Bernoulli et l'approche du modèle linéaire via le lien logit, il est intéressant de définir l'odd ratio :

$$odd\ ratio = \frac{Probabilité\ de\ conversion}{Probabilité\ de\ non\ conversion} = \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \frac{\frac{1}{(1+e^{-\ x'\beta})}}{\frac{e^{-\ x'\beta}}{(1+e^{-\ x'\beta})}} = e^{x'\beta} \tag{17}$$

Il s'agit d'un rapport des risques et, appliqué à notre problématique, il permet de comparer les propensions à souscrire entre les classes homogènes. Plus l'odd ratio d'un sous-groupe du portefeuille est proche de 0, moins il sera représenté dans le portefeuille d'assurés. Dans notre modèle, l'augmentation d'une unité de x_i accroît de β_i le log-odd ratio de convertion. Pour plus de fluidité, l'expression de rapport de risque peut être plus directe en passant à l'exponentielle : l'augmentation d'une unité de x_i correspond à un accroissement de $e^{\beta i}$ *100 % de l'odd ratio, toute chose égale par ailleurs.

Chaque facteur peut s'interpréter indépendemment : les glm sont modulaires et simulables. La parcimonie est privilégiée pour renforcer le caractère interprétable du modèle. Le nombre de coefficients est restreint grâce à la sélection de variables et évite la mauvaise adéquation du modèle (sur-apprentissage), la prédiction erronée portée par la non pertinence de certaines données, l'apparition de biais due aux corrélations. Ces phénomènes poussent à la mauvaise interprétation d'un facteur et sont étroitement liés.

La structure additive des glm permettent aisément de constituer un ordre d'importance des variables selon leur contribution. Le classement des variables par p-valeur est communément employé. Les p-valeurs sont issues du test de Student :

$$H_0 : \beta = 0 \; vs \; \beta \neq 0$$

$$t = \frac{\hat{\beta}}{\sqrt{Var(\hat{\beta})}} \sim N(0,1)$$

Plus la p-valeur est faible et plus le facteur explicatif est important : l'idée est que plus la p-valeur augmente et plus la probabilité de rejeter l'hypothèse H_0 statistiquement alors qu'elle est vraie (erreur de première espèce) diminue. En termes très généraux, un résultat statistiquement significatif est un résultat qui advient très rarement lorsque l'hypothèse nulle est vraie (Sawyer et Peter, 1983).

De nombreuses critiques des tests de significativité paraissent dans la littérature. Entre autres, la confusion entre significativité statistique et substantielle ainsi que la validation des tests franchement déterminée par le volume de l'échantillon. Le lecteur intéressé pourra se référer au livre de Mbengue paru en 2010 cité en pied de page [4] (réf. (14)).

2.4 Explainable Boosting Machine

Le modèle linéaire généralisé est interprétable mais facilement challengé en termes d'exactitude. Les modèles additifs généralisés (GAM), proposés par Hastie et Tibshirani (1987,1990) conservent la structure additive qui favorise le caractère modulaire et simulable de la méthode mais procurent davantage de flexibilité et peuvent présenter des résultats plus performants dans le cadre de certains travaux de recherche (McCullagh et Nelder (1989), Schimek (2000), Shuman (2010) entre autres).

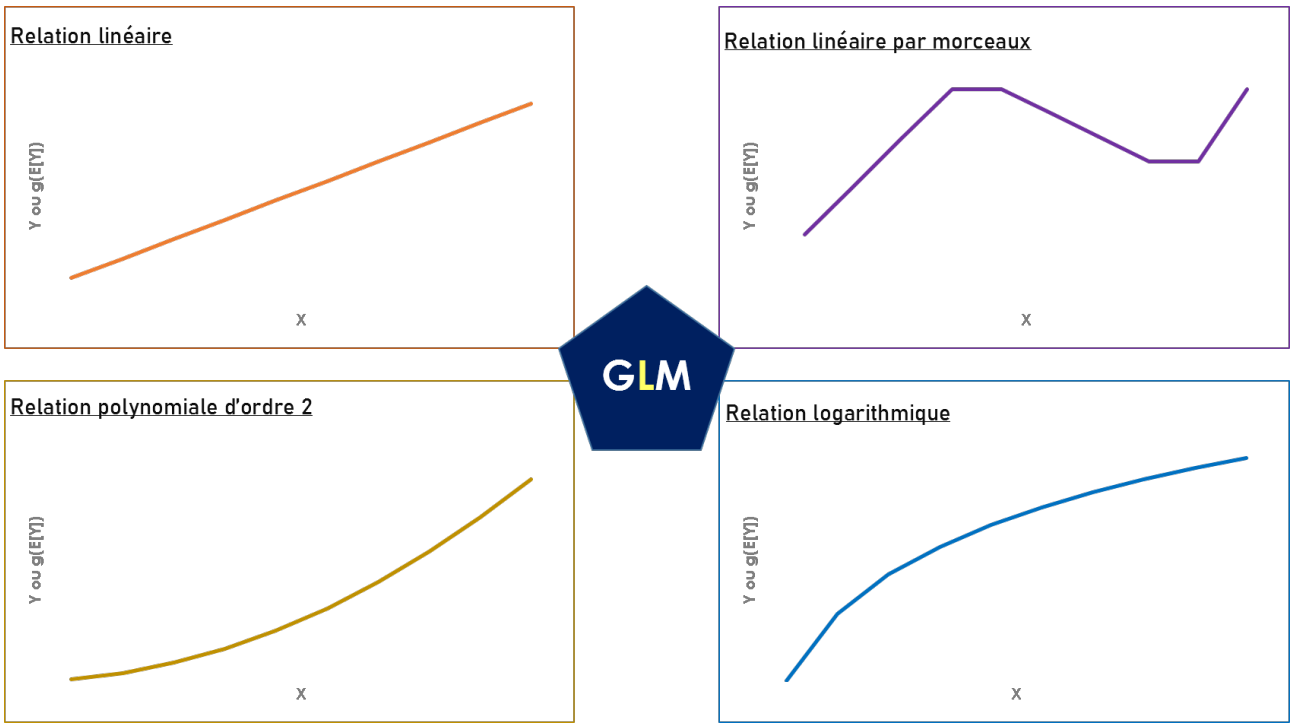
Les explainable boosting machine (EBM), ou GA2M, renforcent la performance des GAM en accompagnant les estimations par du boosting et du bagging et en permettant les intéractions entre deux variables. Les EBM sont une implémentation automatique et rapide des GA2M proposé par lou et coll. en 2013. L'additivité est conservée, l'effet marginal de chaque facteur ou de chaque duo de facteurs est facilement isolé (en cas d'hypothèse de non corrélation ou d'indépendance avec les autres variables). Il devient aisé pour les modèles additifs de comprendre l'importance de chaque variable et la prédiction du modèle. C'est en cela que les GLM, les GAM et les EBM sont considérés comme *glassbox* et font partis de notre pré-sélection de modèles.

Cette section détaille les algorithmes utilisés par les GAM et les EBM, étaye les distinctions avec les GLM classiques et offre des discussions sur leur application pour l'estimation du score de conversion au sein de notre portefeuille.

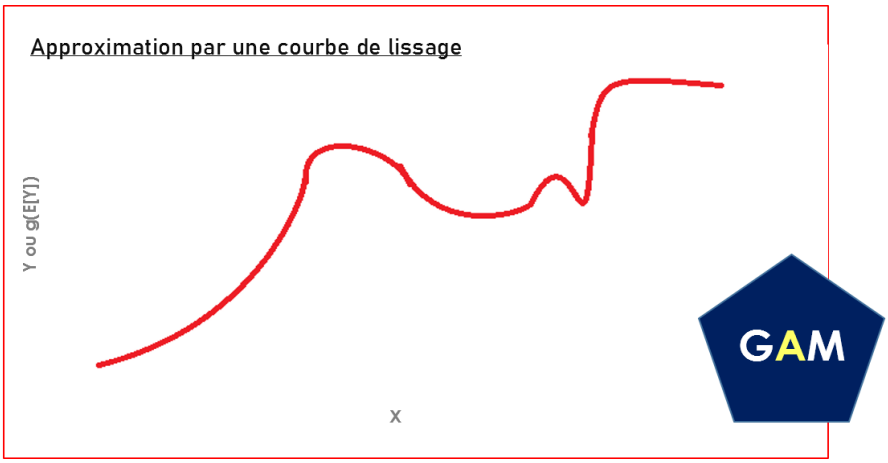
2.4.1 GAM : Modèle Additif Généralisé

Nous avons répondu au problème de manquement au principe de normalité des résidus du modèle par le modèle linéaire généralisé. Un autre point doit être soulevé et concerne la relation entre la variable d'intérêt et les variables explicatives. Jusqu'ici, on a supposé qu'elle était linéaire. Considérons les relations entre output et input suivantes :

4. Mbengue, *Faut-il brûler les tests de signification statistique ?, M@n@gement*, Vol. 13, p. 100-127, 2010



Dans les cas de figure ci-dessus, il est aisé de rendre linéaire la relation entre les deux variables. Dans le cas de linéarité par morceaux, on procède à la segmentation de la variable x. L'interprétation se fera pour chaque tranche qui bénéficiera de son propre coefficient estimé. Dans les exemples de non linéarité, il suffit de poser z égale à $\sqrt{x}$ ou égale à ln(x) pour conserver la linéarité. Dans ce cas, on a recours à une transformation fonctionnelle simple, qui est appliquée à l'ensemble du domaine de définition de x. Cependant, malgré un changement de variable aisé, l'interprétation du modèle devient plus complexe (pourcentage pour la transformation logarithmique et facteur multiplicatif de 2 pour la transformation polynomiale à prendre en compte lors de l'interprétation). Considérons maintenant l'exemple suivant :



La relation n'est pas linéaire, elle requiert une forme fonctionnelle plus complexe. Pour décrire au mieux le comportement de cette courbe, les techniques de régression locale sont adaptées. Elles consistent à déterminer, pour chaque point du jeu de données X, les coefficients d'un polynôme de faible degré (au maximum 3) pour effectuer la régression d'un sous-ensemble des données, les valeurs des variables aléatoires étant proches du point pour lequel on effectue

la régression, puis à calculer la valeur de ce polynôme pour le point considéré. Les coefficients du polynôme sont calculés à l'aide de la méthode des moindres carrés pondérés, qui donne plus de poids aux points proches du point dont la réponse est estimée, et moins de poids aux points plus éloignés. Si la fenêtre est égale à 5 points, alors pour chaque point x :

1. Les 4 points les plus proches de x sont considérés.
2. Une régression polynomiale est effectuée sur ces 4 points et x. Le degré du polynôme dépend du lissage souhaité, si le degré vaut 1 alors la relation est linéaire, et plus le degré s'élève plus le lissage est important.
3. Sur la régression estimée, le point correspondant à l'abscisse de x est fixé.

Les points estimés pour chaque x sont reliés et forme une approximation du lien entre le facteur et la variable d'intérêt. L'amplitude de la fenêtre décrit le degré d'exactitude souhaité. Une segmentation fine aboutit à une estimation juste de la relation mais le risque de sur-apprentissage ne peut être ignoré. L'intensité du lissage débouche sur la même conclusion. L'arbitrage entre la réduction du biais d'une part et le sur-apprentissage d'autre part doit être abordé. Les GAM emploie cette technique pour mieux approcher les correspondances entre inputs et output. La fonction de lien s'exprime ainsi :

$$g(E(Y|X)) = \mu = f(X) + \epsilon = \beta_0 + \sum_{k=1}^{p} f_k(x^k) + \epsilon \tag{18}$$

Pour déterminer les formes fonctionnelles, plusieures approches sont possibles comme l'estimation par noyau. Cependant, l'approche par spline sera privilégiée dans ce cadre, ce qui justifie l'introduction de la régression locale. Des bases fonctionnelles sont largement utilisées dans le lissage de données expérimentales qui sont, à titre d'exemple, les bases polynomiales et de Fourier ou encore la base de splines définie sous la forme :

$$(x^j - v)_+ \; j \in \mathbb{N}, v \in \mathbb{R} \tag{19}$$

En d'autres termes, une spline est une fonction définie par morceaux par des polynômes ce qui permet d'octroyer des contours complexes à l'interpolation entre μ et X. Une segmentation est réalisée : la variable d'état est découpée de telle sorte à affecter un score sur chaque branche (b) de chaque facteur (k) $f_k(x^{k,b})$.

$$f_k(x^k) = \sum_j \xi_j \phi_j(x^k) \tag{20}$$

avec ϕ les éléments de la base de splines.
Les splines font partis de ce qu'on appelle les *shape functions*. Les *shape functions* sont utilisées pour déterminer la valeur de la variable d'état en tout point de l'élément. Les arbres de décision en font également partis et sont largement employés dans le cadre des GAM et des EBM. Ils déterminent le degré de segmentation de la la variable explicative, sous contraintes des paramètres imposés.
On relève deux inconvénients majeurs qu'il faudra contourner dans notre recherche :

1. Le compromis précision/volatilité qui se solde par un sur-apprentissage pour les modèles GAM.

2. L'interprétabilité complexe du modèle par l'approximation de formes complexes.

Ces contraintes se résument à contrôler le(s) paramètre(s) de lissage, les splines étant une forme aisée à interpréter dans le cas d'une segmentation de variables contrôlée.
L'interprétabilité du modèle GAM est facilité dans le sens où la contribution individuelle de chaque variable d'état est compréhensible, et par les scores obtenus pour chaque segment de chacune d'entre elles, on peut quantifier son impact.

2.4.2 Algorithme EBM

Les Explainable Boosting Machines (EBM) possèdent des propriétés plus avantageuses que les GAM. Ils combinent l'exactitude des méthodes telles que le XGBoost ou les Forêts Aléatoires avec l'intelligibilité des méthodes glassbox comme les régressions logistiques ou les GAM [5] (réf (11)). Dans les modèles additifs généralisés, il faut manipuler les fonctionnalités f_k une à une et chacune d'elle modifiera la p-valeur, les critères d'Akaike. Les splines sont flexibles mais également complexes : il faut trouver le nombre de noeuds, la position de chaque noeuds, le degré de liberté...les EBM permettent d'obtenir des résultats similaires à des splines, mais sans saisie manuelle, le meilleur modèle est obtenu automatiquement.

Les intéractions entre variables sont désormais possible, ce qui redéfinit la relation entre Y et les facteurs explicatifs :

$$g(E(Y|X)) = \beta_0 + \sum_{k=1}^{p} f_k(x^k) + \sum_{k,l=1}^{p} f_{k,l}(x_k, x_l) + \epsilon \tag{21}$$

Par ailleurs, la prédiction d'un EBM est moins coûteuse qu'une prédiction par arbres de décision puisqu'il s'agit d'un modèle additif. Le procédé général de l'algorithme s'établit comme suit :

1. **Détermination des fonctionnalités** par bagging ou gradient boosting. Chaque f_k est déterminée tour à tour selon la méthode de round-Robin. De manière imagée, les facteurs explicatifs sont placés sur une liste d'attente en forme de tourniquet. Chacun leur tour, ils sont engagés dans une procédure de renforcement pour déterminer la meilleure fonction associée afin de montrer comment chaque caractéristique contribue à la prédiction, en atténuant les effets de colinéarité détectée par les premiers facteurs de la liste d'attente. Ce traitement est effectué en un temps limité (quantum de temps) par facteur, après cette unité de temps imparti, la variable explicative cédera sa place à celle se situant derrière elle sur le tourniquet. Le quantum est volontairement court pour ne pas donner d'importance à l'ordre des co-variables sur la liste d'attente, qui impacte la contribution de la variable, dégradée par son degré de colinéarité avec les facteurs précédents.
2. **Sélection et détermination des intéractions entre les variables**
3. **Interprétation** du modèle par la quantification de la contribution de chaque paramètre et des graphes des fonctions f_k contre μ.

5. Lou & al., *Intelligible Models for Classification and Regression, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 150-158, 2012

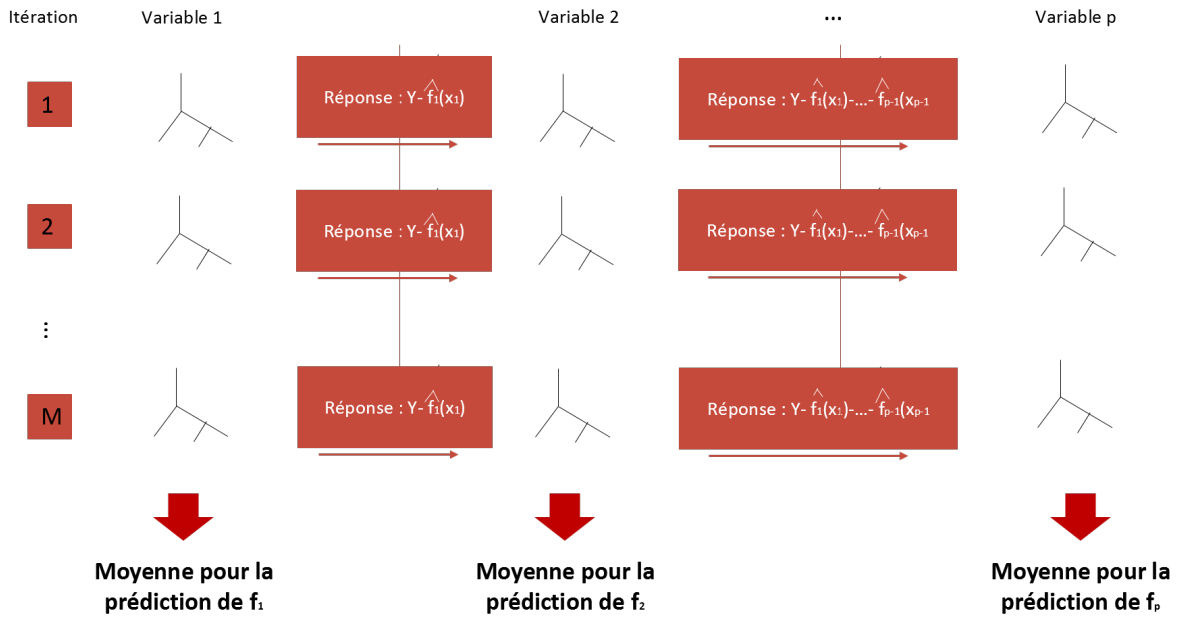


FIGURE 9 – Algorithme de construction d'un modèle EBM avec M itérations de bagging/boosting

A chaque itération, le modèle est amélioré en attribuant un poids plus important aux branches comportant d'avantage d'erreur. Il s'agit alors d'agréger les M arbres pour chaque variable afin de combiner tous les résultats et de diminuer le biais. C'est en ce sens où le modèle demeure interprétable : les arbres prédictifs ont une contrainte sur le nombre de variables possiblement présentes pour la construction des noeuds contrairement au XGBoost classique.

2.4.3 Paramétrisation du modèle

Paramètres	Résultats du tuning
interactions	5
binning	"quantile"
Taux d'apprentissage	0.1
Nombre maximale de feuilles	3
outer bags	25
max bins	75
max interactions bins	25
Nombre d'arbres utilisés pour le boosting	2000

TABLE 4 – Hyperparamètres sélectionnés

2.4.4 Interprétabilité des EBM

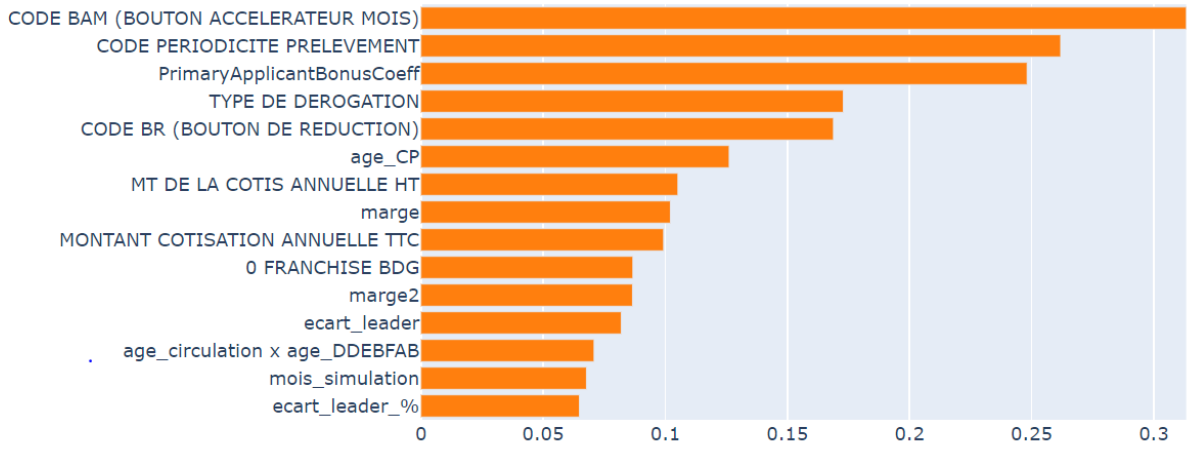


FIGURE 10 – Classement des variables selon leur importance

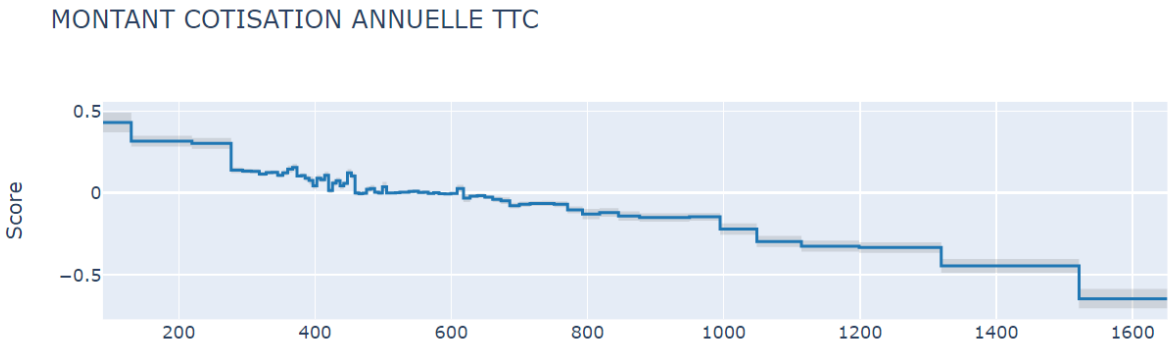


FIGURE 11 – Score d'importance du tarif TTC

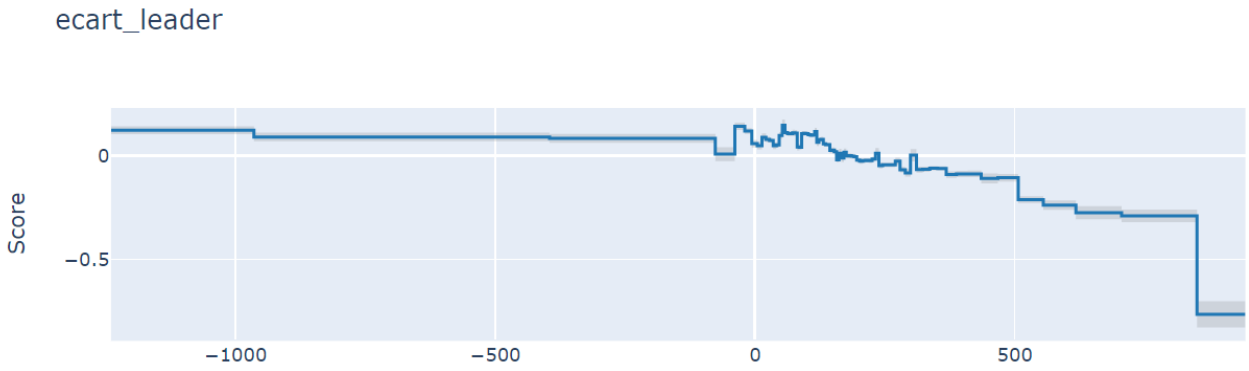


FIGURE 12 – Score d'importance de l'écart avec le tarif leader

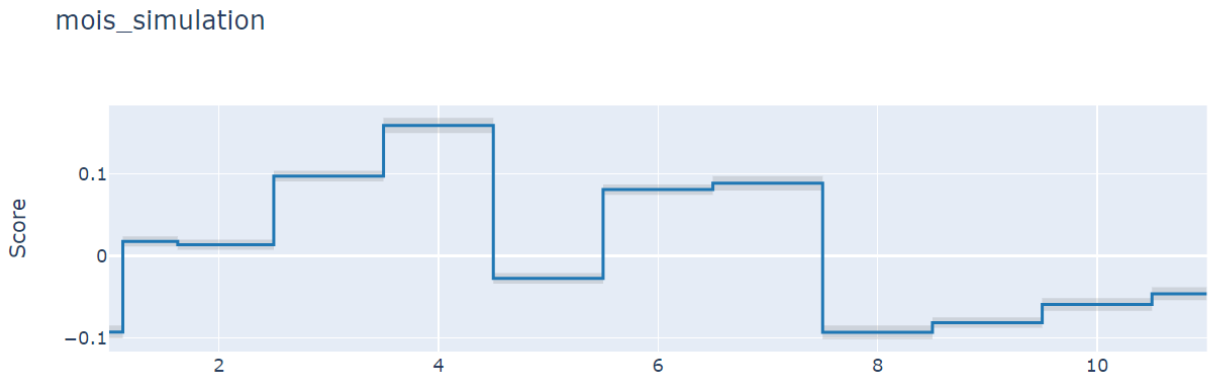


FIGURE 13 – Score d'importance du mois de simulation du devis

FIGURE 14 – Score d'importance du kilométrage moyen parcouru

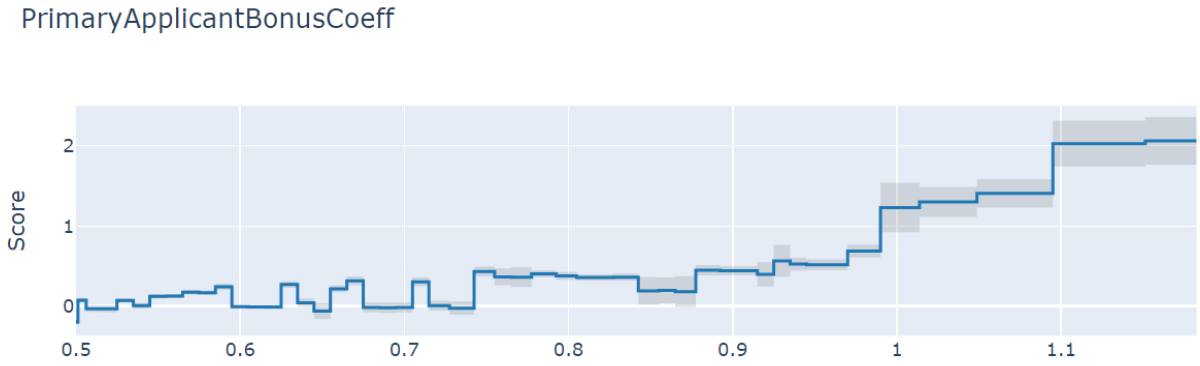


FIGURE 15 – Score d'importance du bonus/malus

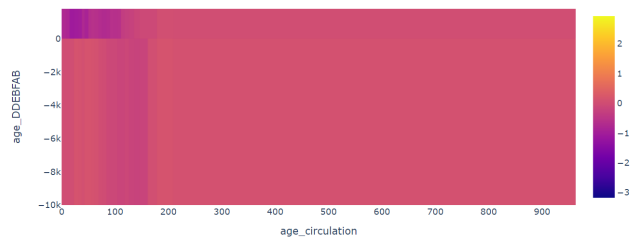


FIGURE 16 – Heatmap : âge du véhicule et âge de sa fabrication

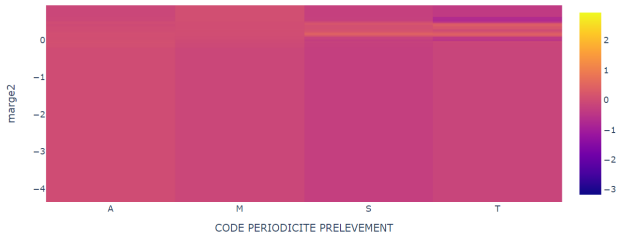


FIGURE 17 – Heatmap : périodicté des prélèvements et marge

2.5 Détermination du score par un modèle XGBoost

2.5.1 Principe

Parmi l'éventail des modèles existants, le XGBoost est opté pour ses bonnes propriétés prédictives, ayant déjà fait ses preuves dans de nombreux concours. Le XGBoost est en réalité issu du *Gradient Boosting*[6] auquel est ajouté des techniques de parallélisation afin d'améliorer les coûts algorithmiques :

1. Un premier modèle est construit pour prédire la valeur cible. Ce modèle est en sur-apprentissage.
2. A partir du modèle construit précédemment, des poids plus importants sont attribués aux observations mal prédites.
3. Un nouveau modèle est construit. Il renforce la performance de prédiction au niveau des observations ayant davantage de poids.
4. L'étape 2 et 3 sont répétées autant de fois que l'utilisateur l'aura suggéré : le *Gradient Boosting* est caractérisé comme séquentiel.

Le résultat final est obtenu par agrégation de l'ensemble des modèles. Afin de profiter de la robustesse de cet algorithme, la bonne paramétrisation est primordiale. Les hyperparamètres se décomposent en plusieures catégories :

— Les paramètres généraux qui incluent le modèle employé, le contrôle de l'apparition de messages d'erreur, les contraintes de parallélisation. Le modèle choisi est l'arbre de régression.

— Les paramètres de boosting qui incluent notamment le poids minimal qu'un noeud de l'arbre doit contenir, la profondeur maximal de l'arbre, le nombre maximal de noeuds terminaux, le seuil du taux d'apprentissage (seuil à partir duquel on considère qu'il y a eu une amélioration de la fonction de perte), la proportion du sous-échantillon utilisé pour la construction du modèle, le nombre de facteurs maximal intervenant dans le découpage d'un noeud et enfin, la pose de contraintes de type L1 ou L2. La calibration de ces paramètres permet un meilleur arbitrage entre biais et variance et visent à éviter le sur-apprentissage. Un ensemble de valeurs est proposée pour chacun d'entre eux et la convergence vers un unique jeu de paramètres sera obtenue par validation croisée sur la base d'apprentissage (procédé de *tuning*).

6. Friedman, *Greedy function approximation : a gradient boosting machine*, 1999

— Les paramètres d'apprentissage qui comprennent la définition de la fonction de perte et la métrique d'évaluation. La distribution appliquée est une loi de Bernoulli et le critère d'évaluation pour l'apprentissage est le logloss (approprié pour ce type de loi parmi l'éventail des métriques proposées), mesurant ainsi la distance entre l'échantillon observé et prédit.

2.5.2 Paramétrisation du modèle

Etant donné le domaine de définition de la variable cible, la loi à calibrer est celle d'une Bernoulli et l'estimation porte sur le paramètre de probabilité de la loi. Les relations suivantes sont retrouvées, avec g la fonction de lien :

$$\mathrm{g(E[Y])} = \mathrm{logit(p)} = log(\tfrac{p}{1-p}) = \beta_0 + \sum_{k=1}^{p} \beta_k X_k + \epsilon$$

La métrique d'erreur de prédilection pour les classifieurs restent le *logloss* contrairement aux fonctions de pertes telles que le MAE ou le RMSE qui sont plus adaptées pour les régressions classiques. La fonction de log-likelihood devant être maximisée, le log-loss est l'opposé de cette fonction. Le log-loss doit donc être minimale pour éviter les pertes d'information.
Un tuning est également effectué pour contrôler l'apprentissage.

Paramètres	Résultats du tuning
Proportion de facteurs utilisée pour chaque arbre	50%
Réduction minimale de la fonction de perte pour effectuer un découpage	1%
Taux d'apprentissage	0.01
Profondeur maximale d'un arbre	7
Poids minimal dans un noeud	1
Métrique d'erreur (apprentissage)	logloss
Proportion de l'échantillon utilisé pour chaque arbre	100%
Nombre d'arbres utilisés pour le boosting	500

TABLE 5 – Hyperparamètres sélectionnés

2.5.3 Interprétabilité du modèle

Pour comprendre la prédiction du modèle, des graphes d'importance sont réalisés. Selon la définition de l'importance, le classement des variables est modifié. Le principe *cover* caractérise l'importance d'une variable comme la couverture moyenne en termes de présence dans la détermination d'un découpage. Le principe *gain* quant à lui, comptabilise l'importance comme la somme totale des gains en termes de réduction de la fonction de pertes à chaque découpage où le facteur est impliqué. L'ordre des facteurs selon leur importance est instable suivant la méthode empruntée.

Les valeurs de Shapley, appréciées pour leur consistence (contrairement aux méthodes *gain*, *cover* ou LIME), stabilisent l'ordre d'importance des variables. Elles font parties des approches

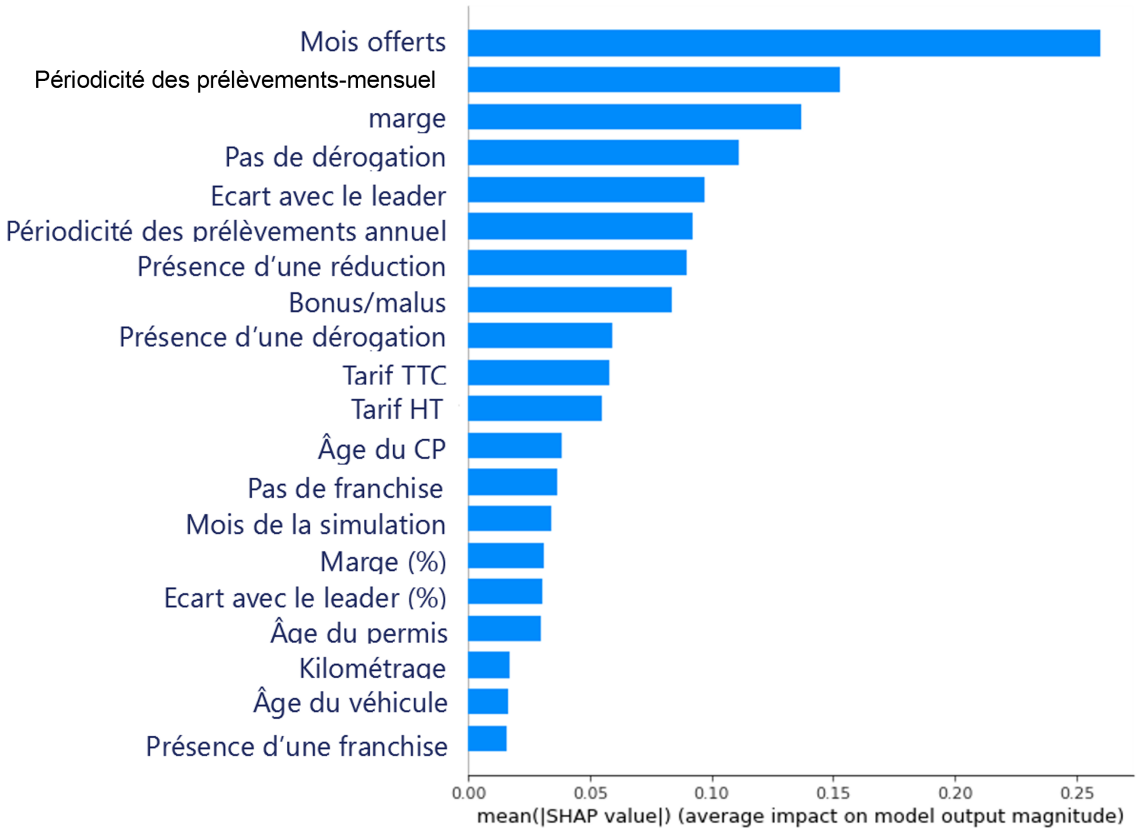
additives (comme LIME) : la somme des contributions de chaque facteur intervenant pour une prédiction en particulier est égal à la valeur de la prédiction. La contribution prédictive se calcule alors localement et pourra être agrégée. Soit F l'ensemble des facteurs présents dans le modèle et soit k un facteur dont on souhaite calculer la contribution. On nomme par S un sous-ensemble quelconque de F ne contenant pas k. On note par $f_K(x_K)$ la prédiction des observations par le modèle entraîné avec l'ensemble de facteurs K.

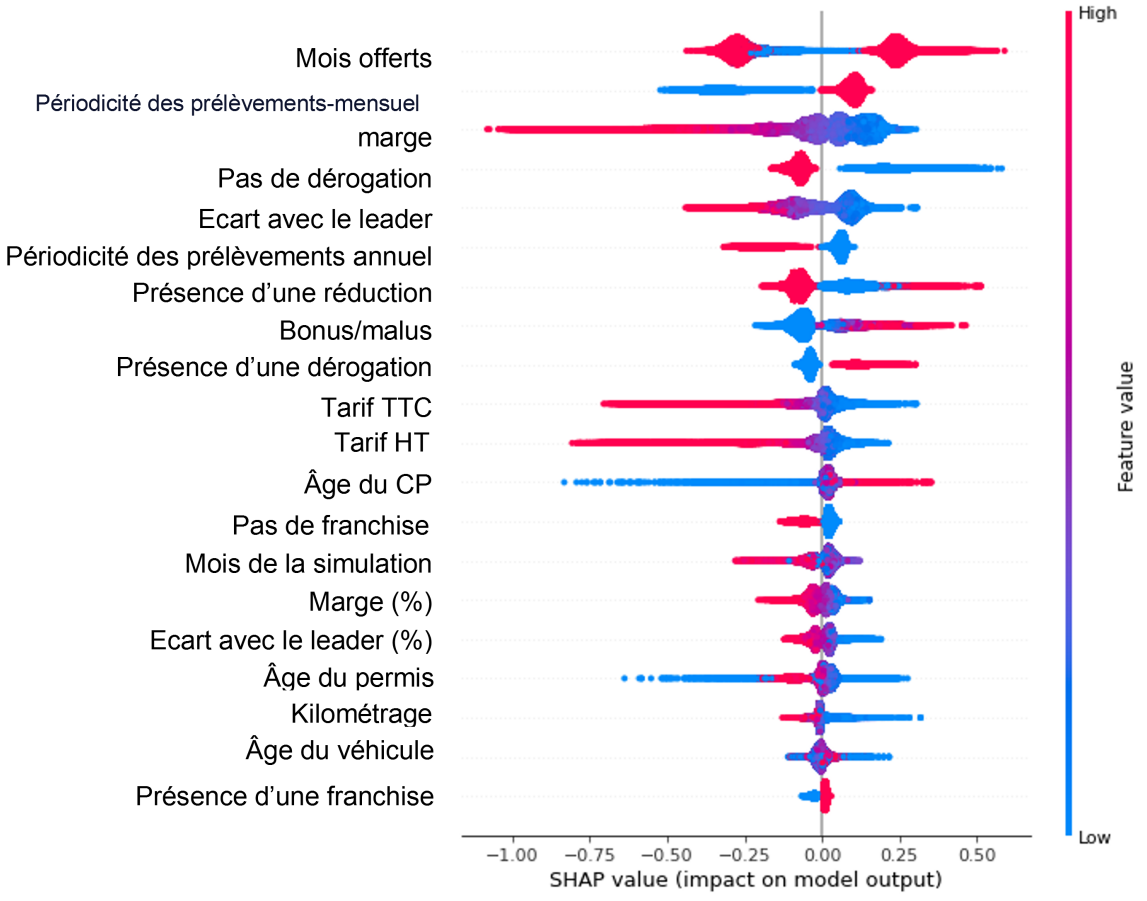
$$\phi_k = \sum_{S \in F} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup k}(x_{S \cup k}) - f_S(x_S)] \tag{22}$$

Autrement dit, la contribution du facteur k est égale à la somme pondérée des contributions marginales de k dans chaque coalition S. La valeur de Shapley offre donc un poids marginal à chaque variable, toute chose égale par ailleurs. Ce concept est issu de la théorie des jeux. Des approximations des valeurs de Shapley existent pour réduire le coût algorithmique sous certaines hypothèses sur les covariables. Le lecteur intéressé pourra se référer à l'article en bas de page pour des approfondissements autour de la théorie sur le calcul de ces contributions [7] (réf. (12)).

Parmi les variables qui impactent davantage le taux de transformation estimé, que ce soit à la hausse ou à la baisse, le nombre de mois de réduction semble convaincre les clients potentiels, le paiement mensuel de la prime, le niveau de marge de l'assureur qui traduit la compétitivité de l'offre, la réduction du montant de la première prime annuelle, l'écart avec le leader, le paiement annuel de la prime qui est généralement moins apprécié mais offre une dérogation, les montant de prime TTC et HT, le niveau de franchise de la garantie Bris de Glace. Les premières places du palmarès sont donc réservées aux facteurs liés au contrat et aux offres promotionnelles, à la position de l'assureur sur le marché et par rapport au leader. En d'autres termes, la première motivation de l'assuré vient de la stratégie tarifaire de l'assureur et ses politiques marketing. Viennent ensuite le bonus/malus, l'âge du conducteur principale, la date d'obtention de son permis, le nombre de kilomètre parcourus, la date de circulation du véhicule. Ces caractéristiques sont hautement significatives dans la tarification du contrat. Les contributions de Shapley délivrent l'effet marginal de chaque variable, toutes choses égales par ailleurs. Il ne faut donc pas interpréter le bon positionnement de ces caractéristiques liées au conducteur et à son véhicule comme une simple corrélation avec les données de contrat. Le comportement et le profil de l'assuré impactent la transformation par eux-mêmes. Le mois de simulation traduit le phénomène périodique de vague de souscription durant les mois de septembre et de janvier/février.

7. Lundberg et Lee, *A unified approach to interpreting model predictions, NIPS'17 : Proceedings of the 31st International Conference on Neural Information Processing Systems (USA)*, 2017





Pour chaque variable présente sur l'axe des ordonnées à gauche sont représentées les contributions de Shapley pour chaque sujet. Chaque point correspond à une observation et la couleur

traduit l'intensité de la valeur du facteur. Pour une variable continue, le rouge signifie que la valeur est élevée tandis que le bleu équivaut à une faible valeur. Pour les variables dichotomiques, l'intensité est binaire, le rouge le plus vif coïncide avec la valeur 1 et le bleu le plus clair coïncide avec la valeur 0 ("yes/no"). Une contribution de Shapley positive participe à faire augmenter le taux de transformation estimé, et ce, proportionnellement à sa valeur. A l'inverse, une contribution de Shapley négative participe à faire diminuer le taux de transformation estimé. Voici quelques exemples d'interprétation élaborés à partir de la lecture de ce graphe :

— Plus l'écart avec le leader est important et plus la valeur de shapley tend vers les négatifs, c'est-à dire que le facteur pousse à prédire un taux de transformation proche de 0. Les assurés se tournent vers des assureurs plus compétitifs.

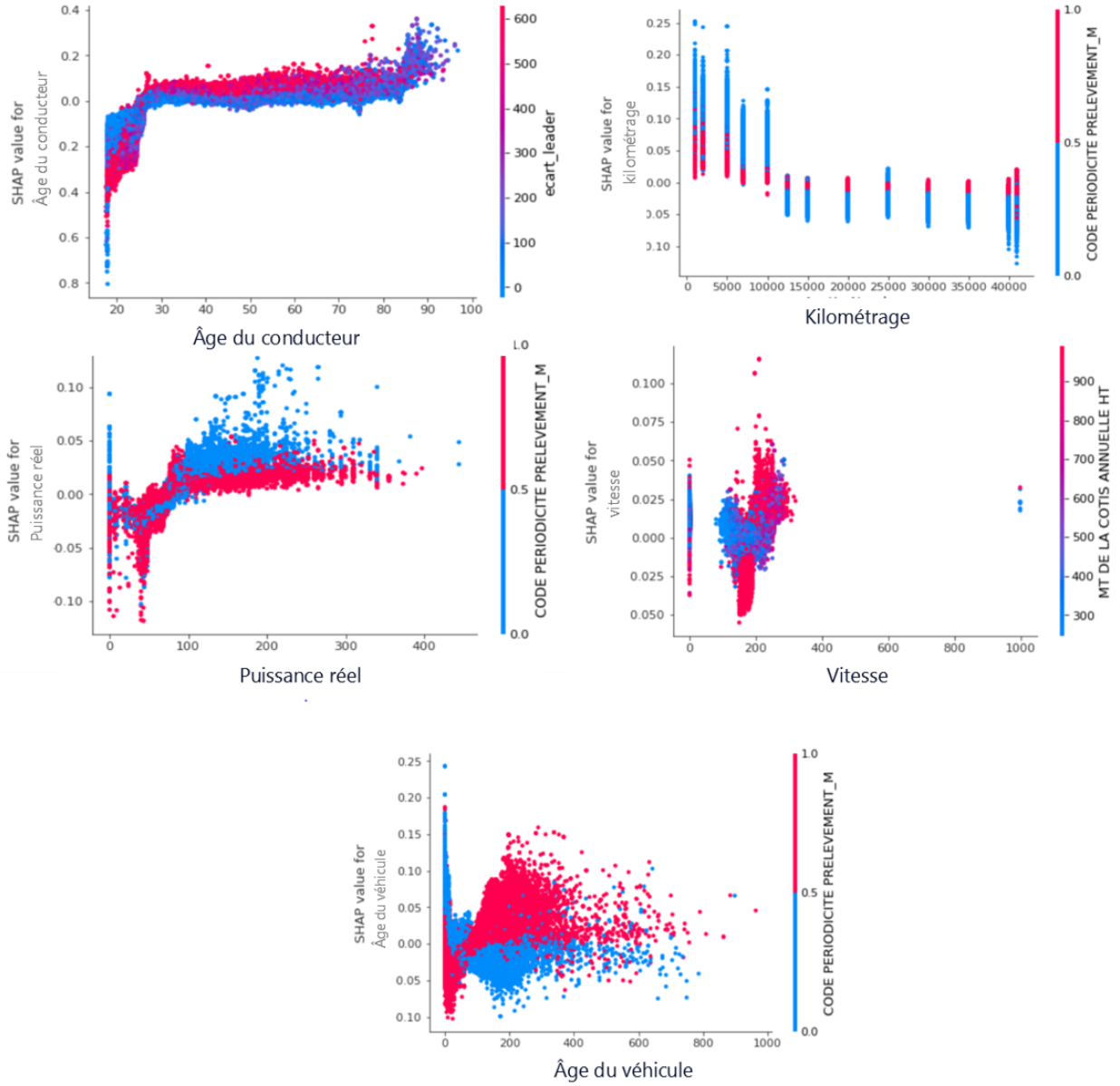
— Lorsque la marge est faible, la contribution est positive et donc le facteur contribue à faire tendre le score vers 1.

— Lorsque le paiement de la prime se fait annuellement, la contribution de la variable est négative (fait tendre le score vers 0). L'assureur a une perte de clients qui paient annuellement leur prime malgré la dérogation accordée. Ce constat fournit une information quant aux profils qui se dirigent vers cette entreprise.

— Les petits rouleurs ont une valeur de shapley positive pour cette variable, le modèle tend à prédire un taux vers 1 grâce à cette caractéristique.

— Le nombre de mois offert offre une analyse plus mitigée : augmenter le nombre de mois de gratuité ne permet pas toujours de favoriser un taux de transformation élevé.

— Avoir une franchise bris de glace de type "N" diminue le taux de transformation.



Les cinq graphes mettent en lien l'évolution des valeurs de Shapley de certaines variables continues avec l'intensité de la valeur d'une deuxième variable d'intérêt.

— Plus l'âge augmente et plus la contribution augmente positivement. C'est aussi à partir des grands âges que l'écart avec le leader est faible. Pour les âges jeunes, les contributions sont négatives et le sont d'autant plus que l'écart avec le leader est important. En revanche, parmi les sujets qui ont entre 30 et 70 ans, ceux qui ont une contribution positive élevée (et donc une probabilité de conversion plus grande) sont ceux qui ont un + grand écart avec le leader.

— Les contributions sont négatives à partir des 20k annuel parcourus. Les petits rouleurs ont une contribution positive accrue lorsque le paiement des primes n'est pas mensuel. L'inverse est vrai pour les grands rouleurs.

— Les contributions augmentent d'autant plus que la puissance réelle augmente et que le prélèvement n'est pas mensuel.

— Lorsque le véhicule a une vitesse maximale en-deça de 250, la contribution est positive

uniquement lorsque le montant de la prime est plus faible. Ce sont des véhicules moins risqués qui exigent une diminution de la prime. En revanche les véhicules à grande vitesse sont moins regardant sur la prime pour avoir une transformation élevée.

— Les véhicules neufs peuvent avoir une contribution plus élevée que les autres âges de véhicule si le prélèvement des primes n'est pas mensuel. La tendance s'inverse quand le véhicule est plus ancien.

2.6 Analyse de performance

Après l'interprétabilité, la deuxième partie de l'arbitrage concerne la qualité des systèmes de classification. Les modèles complexes autorisant les intéractions multiples et renforçant la performance -par *boosting* ou *bagging* par exemple- sont aptes à délivrer de meilleurs résultats dans la prédiction de l'échantillon de test. Ainsi le XGBoost, suivi de l'EBM, part favori : les indicateurs de performance présentés ci-dessous favorisent la comparaison entre les deux modèles et permettront de conclure quant aux a priori émis dans l'application pratique à notre jeu de données.

2.6.1 Analyse segment par segment

La méthode privilégiée pour s'assurer de la qualité d'approximation d'un modèle est la réalisation d'une étude sur chacune des modalités des variables. Elle permet :

1. de s'assurer que la moyenne des taux observés est atteinte par les modèles sur chaque sous-population d'assurés et pas uniquement d'un point de vue globale laissant libre place à la volatilité des erreurs ;
2. de détecter les observations où les erreurs sont les plus importantes et de s'assurer qu'il ne s'agisse pas d'un volume de données trop important ;
3. d'étudier la régularité dans la précision de l'estimation pour chacun des deux modèles ;
4. de garantir la cohérence des évolutions entre les segments et ainsi garantir une nouvelle fois que les prédictions aient un sens.

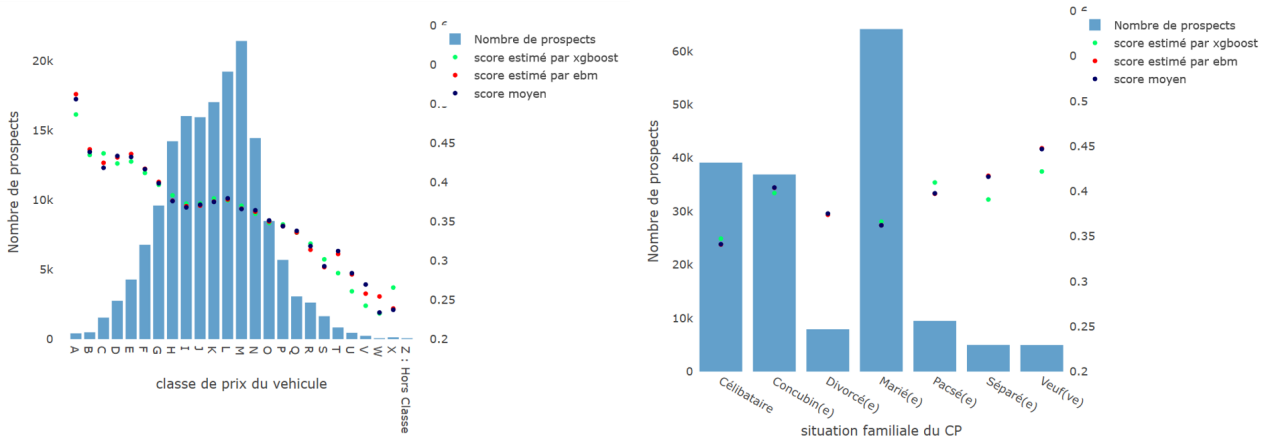


FIGURE 18 – Distributions de la classe de prix du véhicule et du statut marital avec, sur l'axe de droite, le score estimé par XGBoost, EBM et le score observé.

L'ajustement des modèles XGBoost et EBM au taux de transformation est bonne sur chacun des segments. Les résultats peuvent paraître moins justes sur les segments avec un faible volume d'observations, là où la volatilité est plus importante. Néanmoins, l'écart n'est pas suffisamment important pour s'en inquiéter au regard de ce qu'un modèle peut réaliser en termes de performance dans la pratique.

2.6.2 Métriques usuelles des classificateurs

Le premier outil utilisé pour mesurer la performance d'un classificateur est la matrice de confusion. Simple à interpréter, elle permet un premier coup d'oeil sur l'exactitude (ou l'inexactitude) des prédictions. Les lignes correspondent aux classes réelles et en colonnes, les classes estimées. Sur la diagonale principale figure alors le nombre d'individus de la base de test correctement prédits tandis que sur la diagonale secondaire, on compte le nombre de faux positifs et négatifs. Dans le cas de notre étude, la matrice de confusion ne reflète pas la performance attendue de la part de nos modèles. Parmi les profils ayant une faible probabilité de conversion, les contrats convertis auront également un score suffisamment faible pour être prédit comme non converti. En revanche la probabilité de ce devis devra être supérieure à celle octroyée aux profils similaires dont le devis n'a pas été converti. A l'identique, pour les profils ayant une forte probabilité de conversion, une différence sera notée entre les devis convertis et les non convertis. Des détails sur les matrices de confusion et les critères associés (score d'exactitude, taux d'erreur, sensibilité, spécificité, score F1 etc.), qui ne seront pas abordés dans cette section, sont disponibles en annexe.

	Devis converti	Devis non converti
Profil à forte probabilité de conversion	0.48	0.45
Profil à faible probabilité de conversion	0.28	0.25

TABLE 6 – Exemple de score de transformation estimé selon le profil et le devis. Cet exemple illustre l'inadéquation de la matrice de conversion en tant qu'indicateur de performance

Par ailleurs, cet exemple souligne la difficulté potentielle de déterminer un seuil de Bayes à proprement parlé.

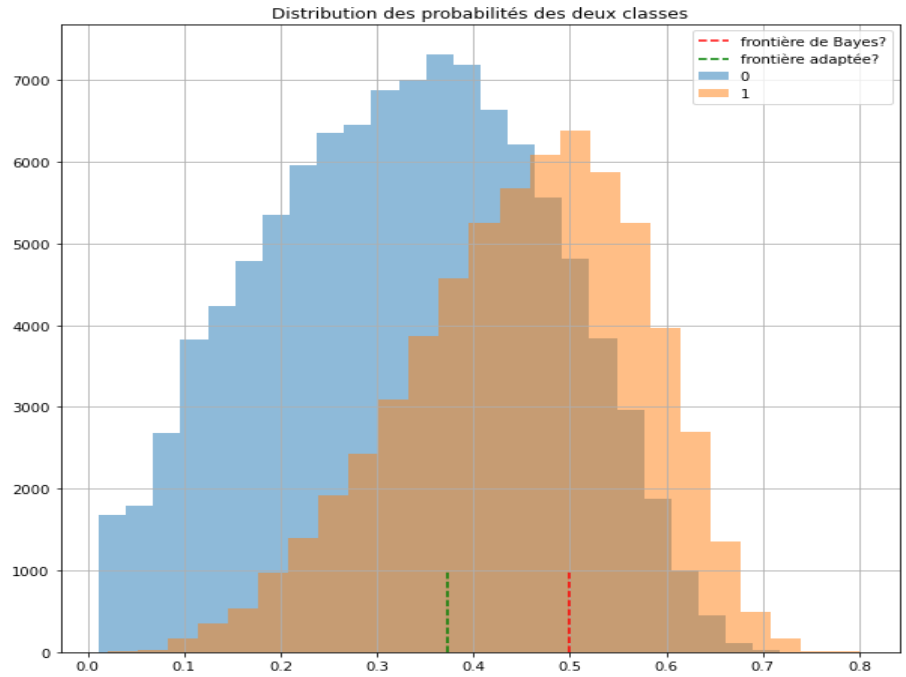


FIGURE 19 – Distributions du score estimé par XGBoost des deux groupes

Le seuil de bayes est à 0.5 et ne détermine pas une frontière claire pour les prédictions de notre output. Les conclusions pour un seuil égal au taux de transformation moyen du portefeuille sont les mêmes.
D'après les estimations du taux de conversion de chaque individu, la frontière entre les deux groupes n'est pas claire.

Néanmoins, le décalage entre le groupe des contrats convertis et non convertis est présent : les deux distributions ne sont pas les mêmes. L'objectif ici n'est pas de classifier les contrats mais d'estimer un score de conversion en fonction des caractéristiques fournies par l'assuré, les caractéristiques du contrat proposées par l'assureur et le tarif proposé par le concurrent.

Critères	Performances sur la base d'apprentissage	
	XGBoost	EBM
Gini normalisé entre score estimé et variable-cible	47.8%	41.3%
Score moyen estimé de la base devis	37.4%	37.3%

TABLE 7 – Comparaison entre les modèles XGBoost et EBM des critères d'évaluation

La métrique de gini (Corrado Gini) permet d'évaluer le niveau d'inégalité entre l'échantillon observé et prédit en tenant compte de la distribution des taux de conversion. Mathématiquement, la métrique de gini se définit par la différence de l'aire des courbes de Lorenz [8] (réf. (10)) :

8. Lorenz, *Methods of measuring the concentration of wealth, American Statistical Association*,vol. 9, p. 209-219, 1905

Courbe de $Lorenz_y \equiv$ (x,y') $\equiv$ (part d'observations cumulées (ajoutées par ordre croissant en fonction de y), part de y détenue par ce cumul d'observations)

$Gini_{Taux\ observés,Taux\ prédits} \equiv$ aire entre la Courbe de $Lorenz_{Taux\ observés}$ et la Courbe de $Lorenz_{Taux\ prédits}$ $\equiv$ aire entre la courbe des taux prédits en fonction des taux observés et la bissectrice

Les indicateurs de performance positionne le XGBoost comme favori, bien que l'EBM s'en rapproche. La probabilité de transformation estimée moyenne est quasiment égale au taux de transformation moyen de la base devis. Le léger décalage est expliqué par la pénalisation de la régression pour le modèle de *Gradient Boosting*. L'idéal serait de comparer le score de chaque profil avec la probabilité moyenne de ce sous-groupe de profils similaires.

2.6.3 Cas des devis proposés à la même personne

Ce qui est intéressant à présent est de vérifier que l'algorithme réussit à déterminer les contrats convertis parmi les personnes converties ayant rempli plusieurs devis. Pour cela on apparie le contrat converti avec le contrat non converti de chaque assuré converti. On a alors deux échantillons appariés et on effectue le test des rangs signés de Wilcoxon pour déterminer si les rangs des deux séries sont similaires (hypothèse nulle).
Ci-dessous, parmi les personnes converties, on a un écart de 9 points de pourcentage entre les contrats convertis et les contrats non convertis et le décalage entre les deux distributions est visible.

	XGBoost	EBM
Taux moyen estimé des devis convertis des personnes ayant souscrit	0.46	0.45
Taux moyen estimé des devis non convertis des personnes ayant souscrit	0.36	0.36
Statistique de Wilcoxon	991447334.0	949613467.0
p-valeur du test de Wilcoxon	0	0

Les p-valeurs sont quasiment nulles. Il y a assez d'évidence statistique pour rejeter l'hypothèse nulle à tous niveaux de confiance usuels. Les deux échantillons ne suivent statistiquement pas la même distribution, ce qui nous conforte sur la qualité de l'estimation du taux de conversion.

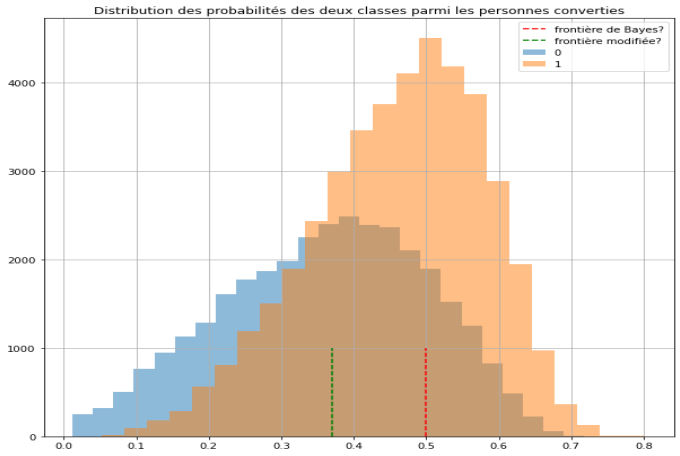


FIGURE 20 – Distribution des taux estimés par XGBoost pour les classes de devis convertis et non convertis parmi les personnes converties.

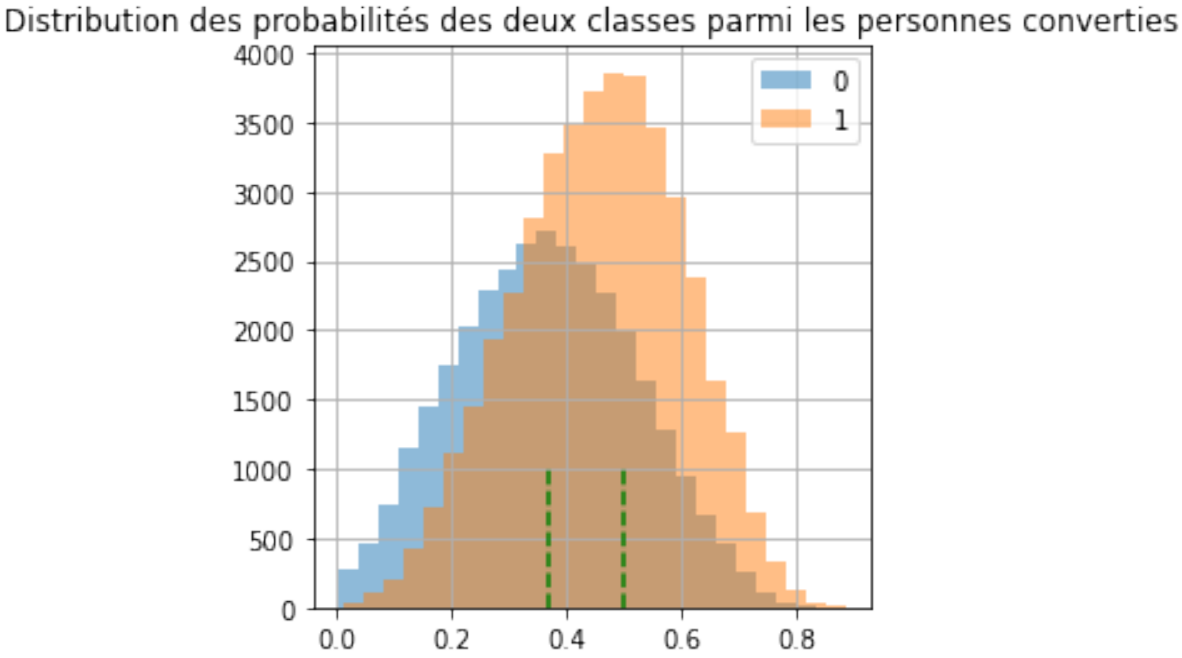


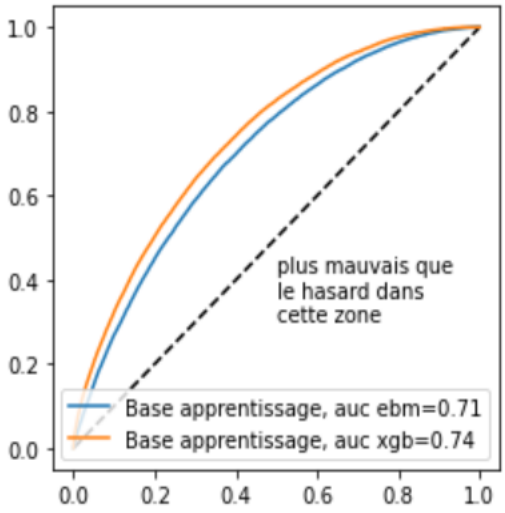
FIGURE 21 – Distribution des taux estimés par EBM pour les classes de devis convertis et non convertis parmi les personnes converties.

2.6.4 Courbe ROC et AUC

Il est intéressant de pouvoir visualiser la performance de l'ensemble de nos modèles statistiques. L'enjeu de cette section est de pouvoir sélectionner le meilleur modèle en termes de qualité de classification.
La courbe ROC (Receiver Operating Characteristic) est une mesure de la performance d'un classificateur binaire qui reprend les notions de sensibilité et de spécificité. Graphiquement, on représente souvent la mesure ROC sous la forme d'une courbe qui donne le taux de "vrais assurés convertis" (fraction des assurés convertis qui sont effectivement détectés : sensibilité) en fonction du taux de "faux assurés convertis" (fraction des assurés non convertis qui sont

incorrectement détectés : 1-spécificité) à chaque seuil de classification.

Diminuer la valeur du seuil de classification permet de classer plus d'assurés comme convertis, ce qui augmente le nombre de faux convertis et de vrais convertis.
Il faut savoir qu'un modèle qui classe les individus au hasard dans l'un ou l'autre classe, aura une courbe de ROC équivalente à celle en pointillés noirs sur la figure ci-dessOus. L'aire sous la courbe, ou AUC (Area Under the Curve), sera alors égale à 0.5 et correspondra au "pile ou face" : en effet, peu importe le seuil, la sensibilité et la spécificité demeurent équivalentes (leur somme égale à 1). La proportion des assurés qui n'ont pas souscrits prédits comme convertis est de 50% de même que la proportion des assurés qui ont souscrits prédits correctement est égale à 50%. De l'autre coté, un modèle qui prédit très bien, aura une aire égale à 1 et la courbe ROC formera un carré. Ainsi, plus l'air sous la courbe de ROC d'un modèle est grand, plus ce modèle est bon.



Les modèles les plus performants sont dans l'enveloppe convexe. Le XGBoost domine le modèle EBM pour tous les seuils de classification.

2.7 Conclusion

Le modèle retenu est le XGBoost.
La première phase est achevée : chacun des devis réalisés possèdent un score de transformation estimé. Les prochains paliers de la méthodologie sont rappelés :

1. Création des groupes de politique tarifaire.
2. Jumelage des sujets entre les groupes pour obtenir des données d'élasticité.
3. Modélisation de l'élasticité au prix.

Deux modèles ont été testés pour la modélisation du taux de transformation, l'EBM et le XGBoost, tous deux ayant l'avantage d'être simple à implémenter et précis. Etant donné que le critère prédominant demeure la robustesse des estimations, le XGBoost l'emporte même si l'EBM n'a pas réalisé de contre-performance. Les analyses des valeurs de Shapley ont permis

de vérifier la cohérence des résultats du XGBoost même si ce dernier n'a pas les qualités d'interprétabilité de l'EBM.

Dans le cas où l'assureur se servirait de la connaissance de l'élasticité-prix pour optimiser son profit en résolvant l'équation par la recherche d'une allocation de tarifs, le XGBoost ne pourrait pas être utilisé dans sa modélisation. Les avantages de l'EBM sont requis dans la phase d'estimation de l'élasticité-prix puisque l'expression de l'équation d'optimisation du profit ou du chiffre d'affaires doit être explicite au niveau du prix. La propriété de modularité deviendra intransigible. La performance et la rapidité d'implémentation du modèle étant très proches de celles du XGBoost sur nos données, l'EBM est un modèle privilégié pour l'estimation du taux de transformation après jumelage des sujets entre les différents groupes de stratégies.

3 Définition des politiques tarifaires et biais de sélection

3.1 Problématique des chocs tarifaires

Question : quel est l'effet causal d'une augmentation/diminution des prix de X% sur la demande d'un individu i de la classe de risque k ?

Cette question revient à chercher l'élasticité-prix :

$$\frac{\frac{Y_{i,n+1}-Y_{i,n}}{Y_{i,n}}}{\frac{P_{i,n+1}-P_{i,n}}{P_{i,n}}} \tag{23}$$

ce qui revient à écrire :

$$\frac{\frac{Y_{i,n+1}-Y_{i,n}}{Y_{i,n}}}{\frac{P_{i,n}*(1+X\%)-P_{i,n}}{P_{i,n}}} \tag{24}$$

Dans le cas où il existe un autre individu du segment k avec un prix P_i * (1 + X%) alors le modèle de taux de transformation réalisé est suffisant pour l'estimation de l'élasticité-prix. En revanche, les données ne sont généralement pas suffisamment exhaustives pour traiter tous les chocs sur tous les individus. Ce manque de données expérimentales peut occourrir dans le cas où une telle évolution n'a jamais eu lieu pour certains profils, ou si le segment est pauvre en données. Malgré cette absence de données comportementales, beaucoup d'acteurs ignorent le problème d'extrapolation qui en est lié et estiment l'élasticité-prix à partir du modèle de demande [9] (réf. (6)). Le biais engendré n'est alors pas pris en compte.

D'autres compagnies réalisent des price tests. Pour réaliser ce choc sur les tarifs, ils effectuent le procédé :

- Scinder une classe de risque en deux groupes de manière aléatoire.
- Appliquer une évolution de X% sur les tarifs d'un des deux sous-groupes.
- Évaluer l'évolution de la demande des deux sous-groupes.
- Comparer la demande avec et sans évolution pour obtenir l'élasticité-prix.

Cette manoeuvre pourrait entraîner des fluctuations dans le portefeuille d'assurés à l'origine de pertes économiques réelles pour la compagnie d'assurance. Pour contrer cet effet, des price tests opposés sont pratiqués : des diminutions de tarifs au sein de sous-groupes viennent compenser des augmentations pour d'autres. Ce procédé pose d'emblée un problème moral : il est difficile de justifier auprès des clients une modification tarifaire sachant que d'autres clients au profil similaire bénéficie d'un tarif plus avantageux. Cette méthode n'entre pas dans le cadre de la transparence vis-à-vis des assurés concernant les processus de tarification. Pour pouvoir tester un choc tarifaire il est alors nécessaire de comparer deux populations d'assurés différentes, l'individu i appartenant à la classe de risque k pour le prix originel P_i et un autre individu l appartenant à une autre classe j dont le prix serait égal à P_i * (1 + X%).

L'effet causal doit être mesuré sans A/B testing, au conditionnel : que se passerait-il si jamais on décide de faire évoluer les tarifs de X% ?

9. Alexandre de Larrard, *Commercial price optimization strategies in car insurance*, 2016

3.2 Discrétisation de la marge : classement des sujets en classe de stratégie

Dans un premier temps, il convient de regrouper les prospects qui ont la même politique tarifaire afin de mieux comprendre les stratégies appliquées par l'assureur (par l'étude des profils de chaque politique) et de mieux maîtriser la bascule d'un profil sur une autre politique tarifaire.
Afin d'établir des typologies de stratégies, la segmentation de la marge calculée se révèle indispensable pour restreindre le nombre de niveaux.

$$marge\ en\ \% = (Cotisations\ HT\ -\ (Prime\ pure\ +\ frais))/Cotisations\ HT \tag{25}$$

Le regroupement des marges individuelles par classe fait échos à deux interrogations :

1. Comment choisir le nombre d'intervalles (ie le nombre de stratégies recensées) ?
2. Comment choisir les bornes de découpage (ie quels sont les niveaux de marge appartenant à une même stratégie) ?

La connaissance experte est ce qui a plus fiable dans cet exercice mais sans a priori, il est difficile de procéder à la main au découpage de la marge. Les statistiques inhérentes à la série comme le min, le max, l'écart-type ou la moyenne sont des indicateurs utilisés pour ce type de procédé. Ils interviennent notamment pour le choix du nombre d'intervalles déterminé par les formules de Brooks-Carruthers, Huntsberger, Sturges, Scott et Freedman-Draconis [10] (réf. (18)). A partir du nombre de classes ainsi pré-calculé (qui varie selon la formule optée), il est possible de regrouper les marges en groupes de taille égale. Un inconvénient majeur est qu'il est également possible à travers cette technique de regrouper des marges trop éloignées en termes de valeurs et d'en séparer certaines qui sont similaires. Cette remarque est d'autant plus vraie pour les intervalles se situant aux deux extrémités qui contiennent des points sensiblement atypiques. Un exemple de ce type de découpage très connu repose sur les quantiles de la série et abouti aux mêmes conclusions. Le regroupement par valeurs pose également un problème, non pas pour l'homogénéité des classes comme précédemment, mais pour la simple raison que certains intervalles pourraient contenir trop peu d'observations et être instables.

Ce manque de connaissance au regard de la distribution des marges individuelles appelle à l'utilisation d'algorithmes de classification. La méthode optée est la classification par ascendance hiérarchique mais d'autres existent, comme par exemple la descendance hiérarchique ou les moyennes mobiles.
L'objectif est d'estimer le nombre de segments et d'obtenir une homogénéité des valeurs dans chaque intervalle.

A l'initialisation, chaque observation représente une classe (principe d'une classification ascendante). Il y a donc autant de groupes que d'observations.
La méthode suggère la détermination de critères de classification. Tout d'abord, une mesure de ressemblance : comment dire qu'une classe est homogène ? Pour ce faire, la distance de Ward notée W est naturellement choisie et calculée à partir de la distance euclidienne d :

10. Youness, *Contributions à une méthodologie de comparaison de partitions* ,2004

$$\mathrm{W} = (\frac{p_i p_j}{p_i + p_j} * d(M_i, M_j)^2 \ 1 \leq i \leq n, 1 \leq j \leq n.$$

avec n le nombre de sujets, M_i la marge de l'individu i, p_i le poids de la classe i (nombre de sujets). Cette mesure pondère la distance euclidienne en fonction du nombre d'individus appartenant au groupe. Cela permet d'éviter d'isoler les observations atypiques. Ensuite, une mesure de ressemblance entre les classes est définie. L'algorithme de Ward, communément utilisé s'appuie sur le théorème de Huygens :

$$\text{Inertie totale} = \text{inertie inter-classe} + \text{inertie intra-classe}$$

$$\Longrightarrow \sum_{i=1}^{n}(x_i - \overline{x})^2 = n_k \sum_{k=1}^{K}(\overline{x_k} - \overline{x})^2 + \sum_{k=1}^{K}\sum_{i=1}^{n}(x_{i,k} - \overline{x_k})^2$$

$$0 \leq v = \frac{inertie\ inter-classe}{inertie\ totale} \leq 1$$

L'inertie intra-classe décrit la variabilité à l'intérieur de la classe : notre objectif étant de regrouper des individus qui se ressemblent cette quantité est à minimiser. A contrario, l'inertie inter-classe décrit la variabilité entre les classes : pour justifier la segmentation, les groupes doivent être hétérogènes et cette quantité est à maximiser. D'après ce théorème, minimiser la variance intra-classe ou minimiser la variance inter-classe sont deux actions équivalentes.

Le découpage ne peut avoir lieu sur ce seul critère. En effet, à l'initialisation, v=1. Plus le nombre de classes est grand et plus les deux termes tendent à diminuer. A l'opposé, plus le nombre de classes est petit et plus les deux termes tendent à augmenter. A chaque itération, on risque alors de dégrader l'inertie inter-classe. En réalité, la méthode de Ward minimise la diminution de l'inertie intra-classe à chaque itération. La construction des groupes se fait étape par étape :

1. Calculer la distance de Ward entre chaque groupe.
2. Fusionner les 2 groupes les plus proches et on les relie dans le dendrogramme. Calculer la perte d'inertie interclasse (ou gain d'inertie intraclasse) dû au regroupement précédent : il s'agit exactement de l'écart de Ward des deux individus regroupés. Le trait qui nous permet de relier les 2 groupes dans le dendrogramme est d'autant plus long que le gain d'inertie intraclasse entre les groupes est élevé.
3. Remplacer les deux individus de distance minimale par une classe, qui sera représentée par le centre de gravité des individus et affectée de la somme des poids des individus.

Reproduire les trois étapes jusqu'à ce qu'il ne reste plus qu'un seul et unique groupe réunissant tous les individus.

Le niveau (hauteur) de chaque noeud de l'arbre est donc choisi proportionnel à la nouvelle d'inertie intraclasse ; en particulier, il correspond au rapport 1-v. Ce niveau est nul lorsque tous les individus sont séparés (en bas) et vaut 1 lorsqu'il sont tous réunis en une seule classe (en haut).

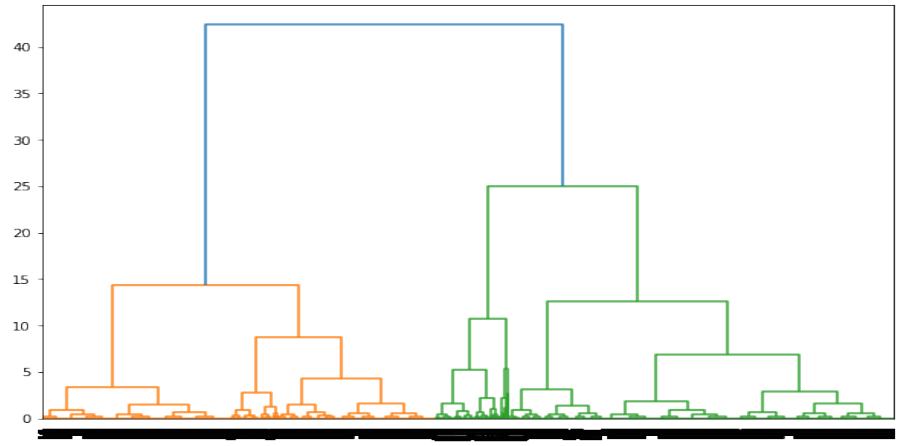


FIGURE 22 – Dendogramme obtenu à l'issu des trois étapes.

Le dendogramme tracé a pour but de visualiser le niveau optimal de coupure de cet arbre afin de réaliser la meilleure partition de l'ensemble initial. Il est judicieux de couper le dendrogramme à un niveau où le regroupement entre classes conduit à une perte d'inertie entre les classes maximale.

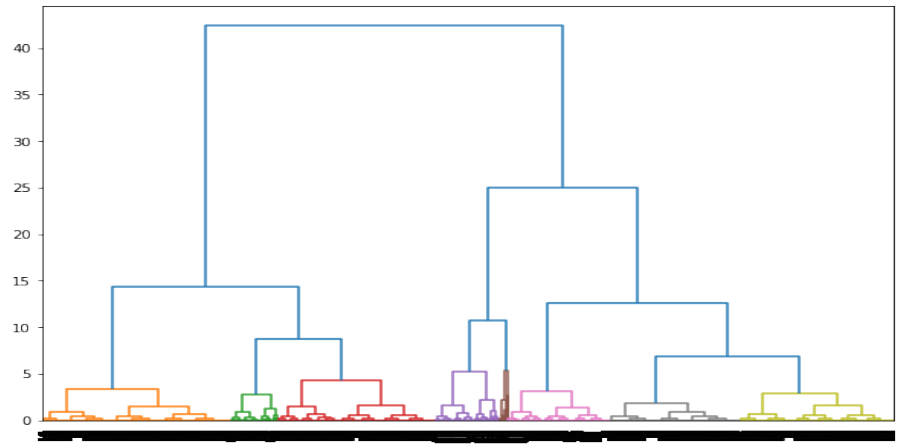


FIGURE 23 – Dendogramme avec la représentation des groupes correspondant au niveau 6

Couper à un niveau supérieur à 6 engendre un saut important de la variance intra-classe, tout d'abord entre les classes grise et jaune, verte et rouge, puis violette et marron. Le regroupement conduit à une perte encore plus importante au-delà du niveau égal à 13. Réduire de manière importante le nombre de groupes revient à confondre des niveaux de stratégies distinguables de par les profils auxquelles elles sont appliquées.
En-dessous de 6, le nombre de groupe apparaît très important (au-delà de 13) et la segmentation peu justifiable. En effet, si la proportion d'individus présente dans un groupe est faible avec le jumelage entre les individus de deux classes éloignées en termes de valeur de marge (et donc de caractéristiques bien différentes) peut s'avérer ardu de par un manque de données.

Les bornes d'intervalles finalement choisies sont :

	Minimum	Maximum
Groupe 1	-2.42	-0.42
Groupe 2	-0.41	-0.05
Groupe 3	-0.04	0.1
Groupe 4	0.11	0.22
Groupe 5	0.23	0.3
Groupe 6	0.31	0.42
Groupe 7	0.43	0.59
Groupe 8	0.6	0.9

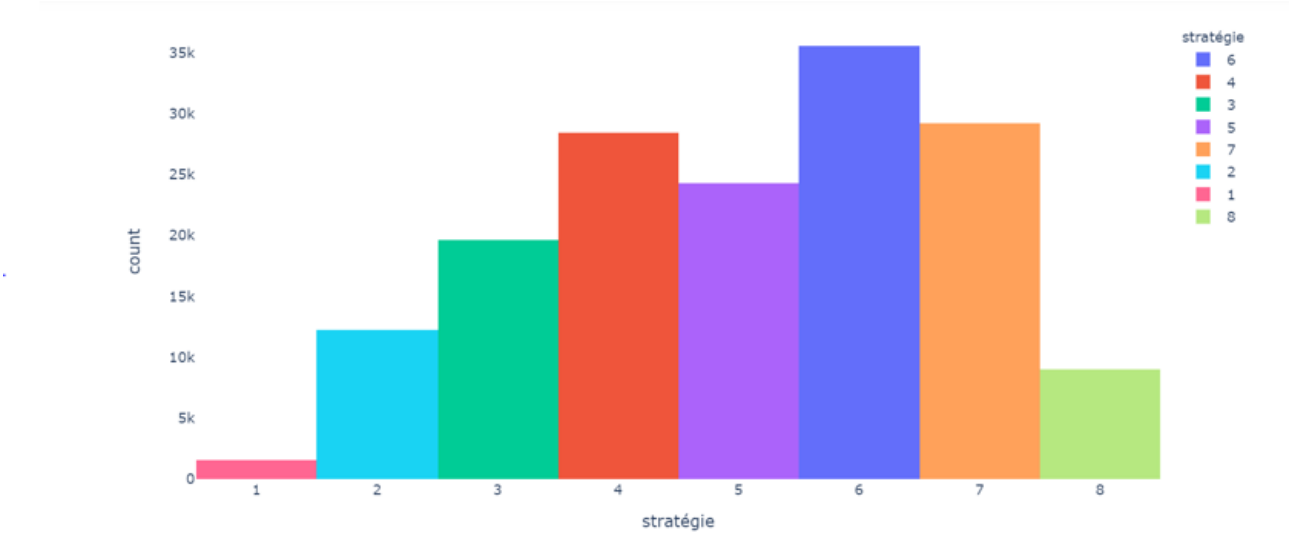


FIGURE 24 – Histogramme des stratégies

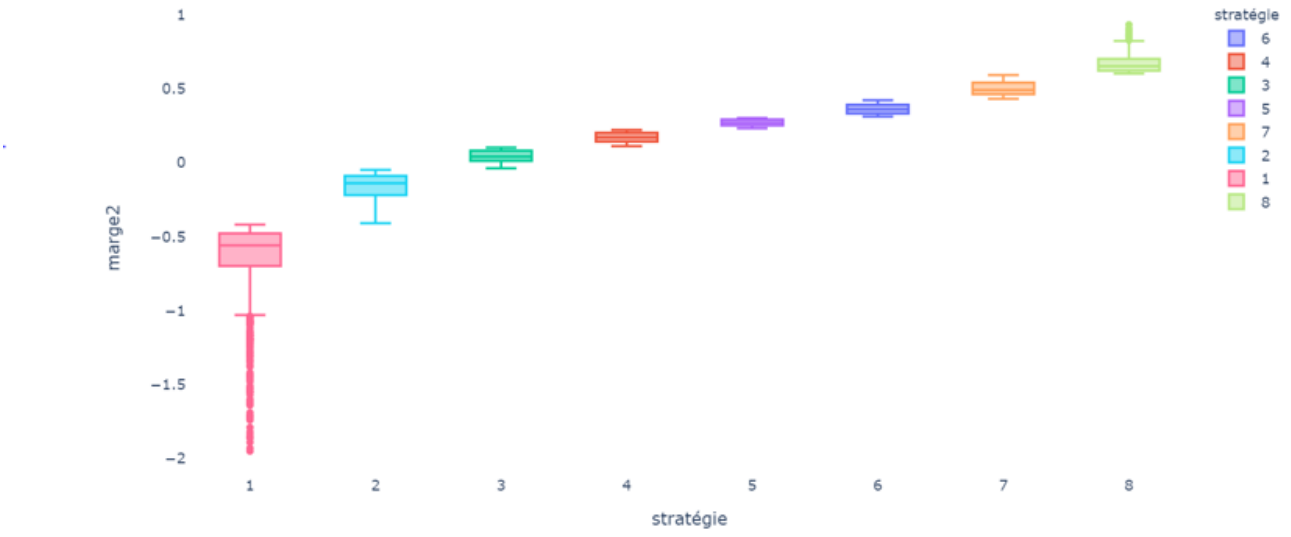


FIGURE 25 – Répartition par stratégie

Le groupe 1 détient un faible nombre de données. Sa proportion est 4 fois inférieure au deuxième groupe le moins volumineux, c'est-à dire le groupe 8 qui se situe également dans les

extrêmes. Pour homogénéiser le volume de données au sein de chaque classe et par cohérence compte tenu de la valeur des marges (marges négatives, contrats non rentables), la base devis sera segmentée autour de sept politiques tarifaires :

	Minimum	Maximum
Groupe 1-2	-2.42	-0.05
Groupe 3	-0.04	0.1
Groupe 4	0.11	0.22
Groupe 5	0.23	0.3
Groupe 6	0.31	0.42
Groupe 7	0.43	0.59
Groupe 8	0.6	0.9

L'inconvénient majeur de l'algorithme, souvent cité dans la littérature, est lié à son coût lorsque le nombre d'observations devient trop élevé. Pour les bases de données volumineuses la méthode des moyennes mobiles est souvent privilégiée. Cependant, le pilotage de l'initialisation peut s'avérer complexe et les différents tests pour trouver le nombre de segments initial sont à prendre en compte dans le coût général du découpage. Le lecteur intéressé pourra se référer aux travaux de la thèse référée précédemment. Pour ces travaux, seul 20% de la base devis a servi à stratifier les niveaux de marge. Un contrôle des statistiques élémentaires des séries a permis d'attester la robustesse du sous-échantillon.

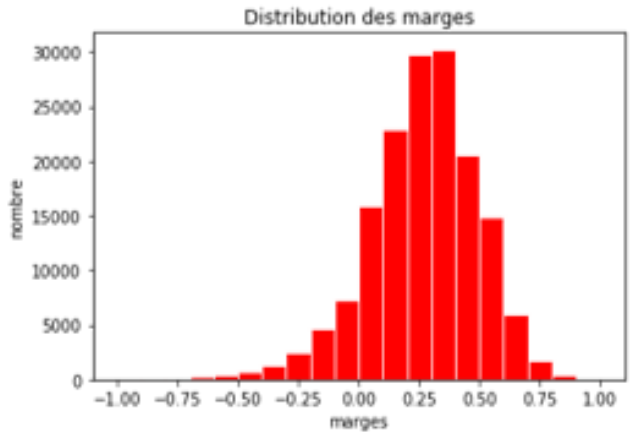


FIGURE 26 – Distribution de la série entière

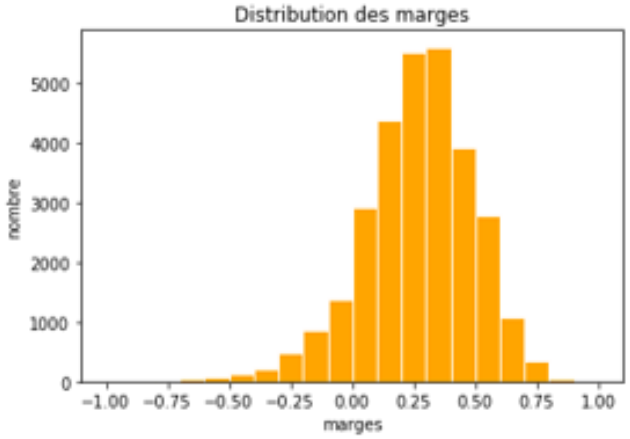


FIGURE 27 – Distribution du sous-échantillon

	Série entière	Sous-échantillon
Minimum	-4.36	-2.42
Moyenne	0.27	0.27
Maximum	0.93	0.9

3.3 Profils par stratégie : entre points communs et distinctions

Lors de l'étude des caractéristiques propres aux différents sous-groupes de stratégies, la première remarque doit porter sur l'existence de similitudes : sur chaque variable, tous les individus à l'exception de quelques extrêmes peuvent trouver un assuré d'un autre groupe tarifaire ayant la même valeur pour cette variable. En revanche, les moyennes et les proportions de sont pas les mêmes pour l'ensemble des caractéristiques.

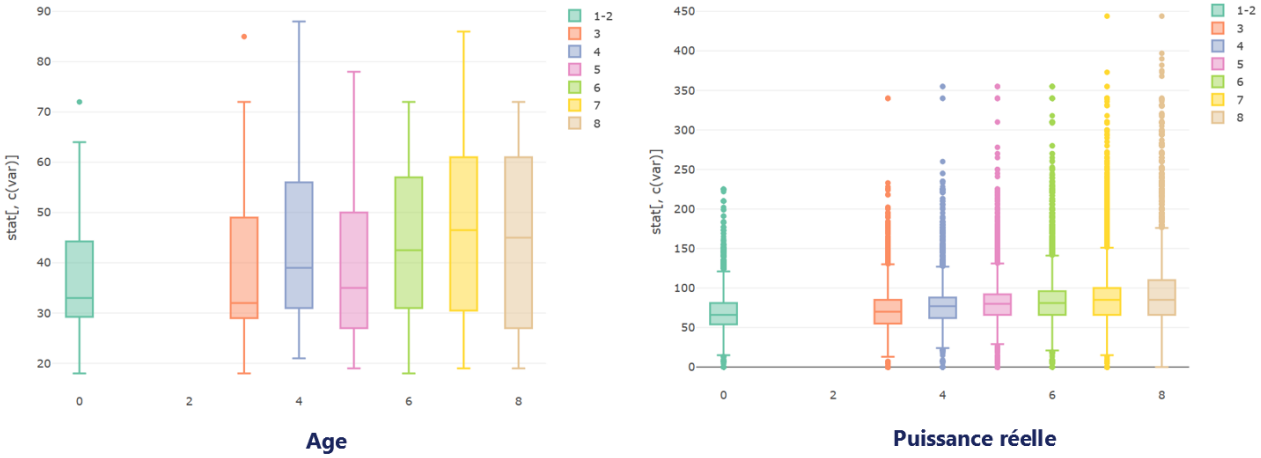


FIGURE 28 – Âge et puissance réelle du véhicule par groupe de stratégie. Les assurés des stratégies 1-2 et 3 sont plus jeunes tandis que ceux des groupes 7 et 8 sont plus âgés. La marge s'accroît avec la montée en gamme des véhicules : la puissance réelle augmente de manière croissante avec la stratégie.

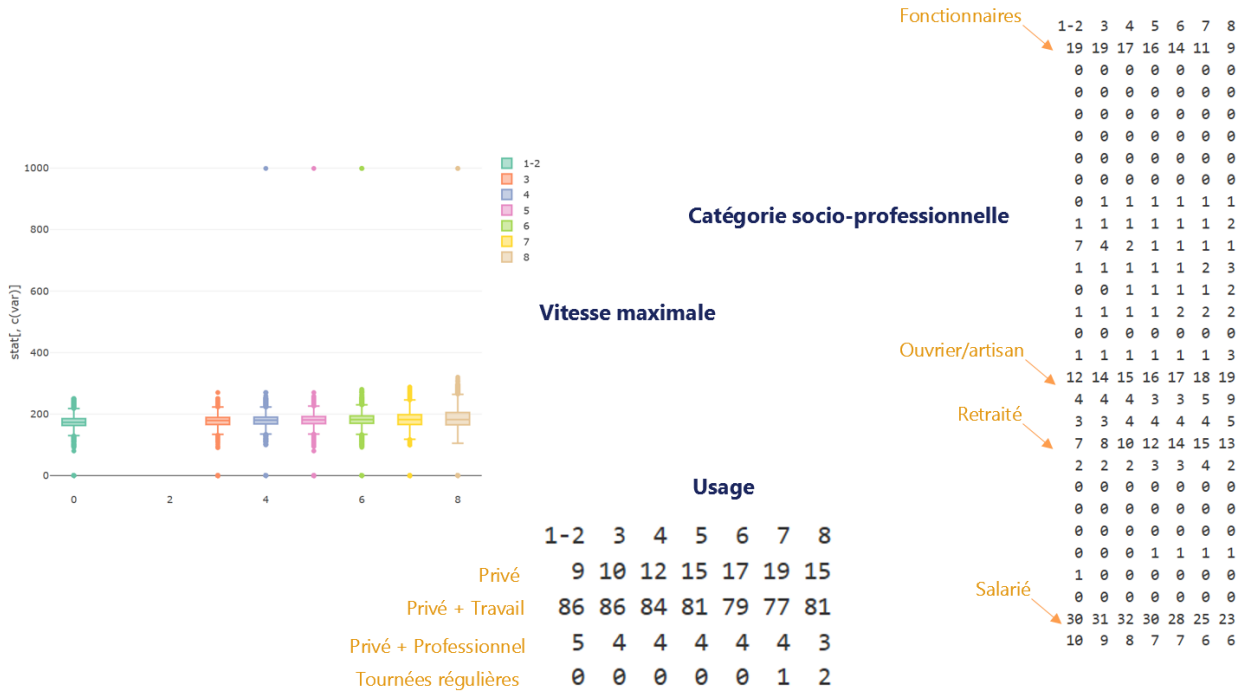


FIGURE 29 – Vitesse du véhicule, profession et usage du véhicule par stratégie. De même que pour la puissance avec qui elle est corrélée, la marge augmente avec la vitesse maximale. La proportion de fonctionnaires et de salariés diminuent avec la marge tandis que celle des retraités et des ouvriers/artisans augmentent.

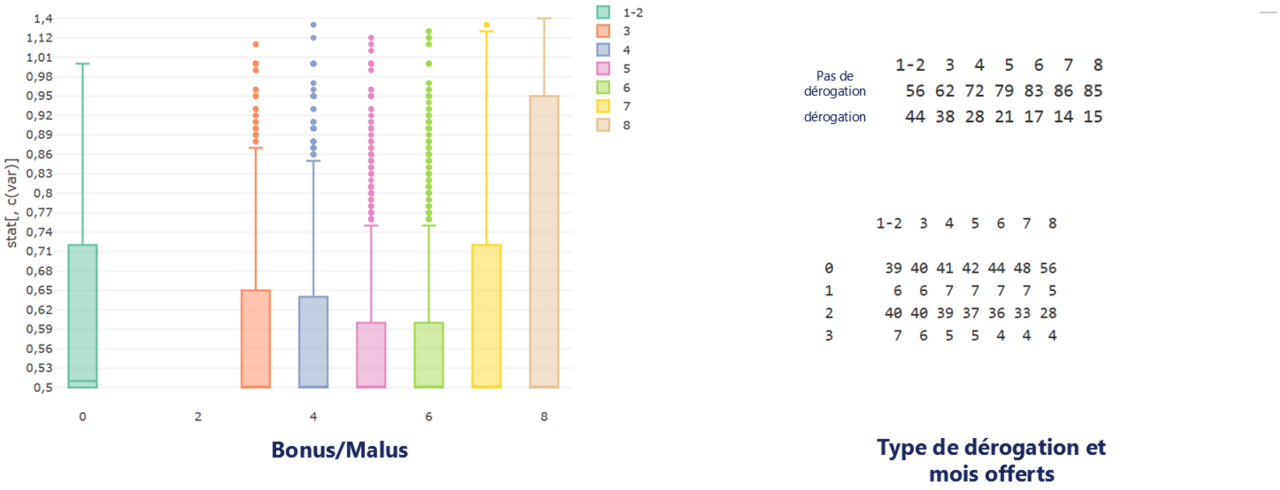


FIGURE 30 – Bonus/malus, dérogation, nombre de mois offerts par stratégie. Les deux premiers groupes sont d'avantages constitués de jeunes que les suivantes ce qui explique la proportion de prospects ayant un bonus entre 0.5 et 1 (jeunes conducteurs). En revanche la proportion d'assurés avec un bonus supérieur à 0.5 ne s'explique pas de la même manière pour les stratégies 7 et 8, leur typologie de conducteurs étant des malusés ayant déclarés un ou plusieurs sinistres. Les profils à marge faible bénéficient d'offres commerciales comme des dérogations ou des mois offerts (qui contribuent à l'abaissement des marges). Ils constituent alors de segments ciblés par l'assureur.

3.4 Modélisation : cas randomisé

La modélisation du taux de transformation a été effectuée pour chaque *prospect*. Il constitue une mesure du comportement d'un client donné face à un produit proposé doté d'une stratégie. Parmi les caractéristiques du produit automobile "Tous risques", le montant toutes taxes comprises, le niveau de franchise, les différentes réductions sont recensés. La stratégie a quant à elle été définie comme la différence entre la prime commerciale hors taxe et hors frais et la prime pure estimée. En d'autres termes, la marge appliquée sur un tarif correspond à une stratégie de la part de l'assureur qui peut décider de rogner cette marge en vue de capter de nouvelles cibles. Ces stratégies ont été regroupées en fonction de la valeur des marges pour aboutir à un dictionnaire composé de 7 stratégies.

Pour optimiser un tarif, il faut pouvoir évaluer le comportement client dans le cas hypothétique d'un changement de stratégie (variation de prix).

Le problème suivant est formulé :
Quel serait le taux de conversion du sujet i appartenant à la stratégie m $\in$ {1-2,3,..,8} si l'assureur lui attribuait un niveau de marge appartenant à la stratégie m' $\in$ {1-2,3,..,8} $\setminus$ m ?
Deux sous-groupes apparaissent avec d'une part les devis avec un niveau de marge correspondant au groupe m et d'autre part les devis avec un niveau de marge correspondant au groupe m'. Ils seront appelés respectivement groupe de traitement (car ce sont les individus dont on va appliquer un traitement au tarif) et groupe de contrôle (individus servant de référence).

Le numérateur de l'élasticité-prix est redéfinit dans ce cadre :

$$\frac{Y_{i,m'} - Y_{i,m}}{Y_{i,m}} | D = m \tag{26}$$

où $Y_{i,m'}$ est inconnu puisque i appartient au groupe m. Dans le cas où les individus sont indépendants et identiquement distribués, les méthodes de Monte Carlo sont utilisées pour estimer la sensibilité tarifaire moyenne d'un groupe à un autre. Plus précisément, ces méthodes permettent d'estimer les deux quantités d'intérêt suivantes :

$$ATE_{m,m'} = E[Y_{m'} - Y_m] \tag{27}$$

$$ATT_{m,m'} = E[Y_{m'} - Y_m | D = m] \tag{28}$$

La première est caractérisée comme l'*average treatment effect*, soit l'effet causal moyen du traitement (ie. de la stratégie S_m par rapport à la stratégie $S_{m'}$). La deuxième est nommée *average treatment effect on the treated* ou effet causal moyen sur les personnes traitées et traduit l'effet sur le taux de transformation de la stratégie $S_{m'}$ par rapport à la stratégie S_m des assurés ayant bénéficié de la stratégie S_m.

Dans le cas où l'assignation à un traitement serait aléatoire, les sujets seraient interchangeables, ce qui se traduit par la condition :

$$(Y_m, Y_{m'}) \perp D \tag{29}$$

et donc :

$$E[Y_{m'}|D = m'] - E[Y_m|D = m] = E[Y_{m'}] - E[Y_m] = ATE$$

et

$$ATT = E[Y_{m'} - Y_m|D = m] = E[Y_{m'} - Y_m] = ATE$$

Ces deux mesures des effets du traitement coïncident car, en raison de la randomisation, la population traitée ne sera pas, en moyenne, systématiquement différente de l'ensemble de la population.
Or, $E[Y_{m'}|D = m'] - E[Y_m|D = m]$ est estimable par l'estimateur de Monte Carlo de la moyenne :

$$\hat{ATE} = \frac{1}{\#j:D_j=1} \sum_{j:D_j=m} Y^j_{m'} - \frac{1}{\#j:D_j=m} \sum_{j:D_j=m} Y^j_0$$

Dans les expériences randomisées, les résultats des deux groupes de traitement peuvent souvent être directement comparés parce que leurs attributs sont quasiment similaires alors que, dans les expériences non randomisées, ces comparaisons directes peuvent être trompeuses parce que les sujets exposés à un traitement diffèrent généralement de façon systématique des sujets exposés au deuxième traitement.

En comparant le résultat moyen du traitement et la moyenne résultat de non-traités, nous estimons alors :

$$E[Y^i|D_i = m'] - E[Y^i|D_i = m] = E[Y^i_m|D = m] - E[Y^i_{m'}|D_i = m']$$

$$= E[Y^i_m|D_i = m] - E[Y^i_{m'}|D_i = m] + E[Y^i_{m'}|D_i = m] - E[Y^i_{m'}|D_i = m']$$

$$\implies ATT^i = E[Y^i_m|D_i = m] - E[Y^i_{m'}|D_i = m'] + E[Y^i_{m'}|D_i = m] - E[Y^i_{m'}|D_i = m']$$

1. $E[Y^i_m|D_i = m] - E[Y^l_{m'}|D_i = m]$ est le treatment effect for the treated ou effet du traitement pour les personnes traitées
2. $E[Y^i_m|D_i = m] - E[Y^i_{m'}|D_i = m']$ est la quantité estimable
3. $E[Y^i_{m'}|D_i = m] - E[Y^i_{m'}|D_i = m']$ est le biais de sélection

En d'autres termes, la différence de moyennes de l'échantillon n'identifie l'effet causal moyen (et l'effet causal moyen sur les personnes traitées) que s'il n'y a pas de biais de sélection :

— Dans le cas où les données expérimentales seraient exhaustives, c'est-à dire si chaque individu se serait vu octroyer les deux stratégies tarifaires dans le passé, le résultat Y serait observable pour les deux traitements (et donc le taux de transformation précédemment estimé par XGBoost devient suffisant pour obtenir la sensibilité au prix individuel pour tous les niveaux de marge relative présents dans la base devis).

— Dans le cas où S_m aurait été attribuée au sujet i et $S_{m'}$ aurait été attribuée au sujet j, et que les deux sujets seraient interchangeables (ie. posséderaient sensiblement les mêmes caractéristiques) alors la conclusion serait similaire : on pourrait interchanger les résultats sur le taux de transformation des deux assurés : $Y^i_m \sim Y^j_{m'}$ et $Y^j_m \sim Y^i_{m'}$.

3.5 Modélisation : cas non randomisé

Or, le taux de transformation n'est connu historiquement que pour les marges auparavant appliquées. La remontée vers un historique trop lointain pour accroître le nombre de stratégies observées s'avère erronée puisque le comportement client évolue avec le temps.

L'objectif est de connaître, pour chaque assuré potentiel présent dans la base devis, le taux de transformation associé à chaque stratégie m $\in$ {1-2,3,4,5,6,7}. Comme énoncé précédemment, le manque de données empêche la modélisation directe. Comme les stratégies ne sont pas distribuées aléatoirement aux segments d'assurés distincts, il est également impossible d'attribuer directement le taux de transformation d'un autre sujet ayant subi la stratégie d'intérêt, au risque de créer un biais lors de l'estimation de l'élasticité-prix.

Prenons un client i de notre base devis auquel la stratégie m_i du groupe de politique m a été appliquée avec un taux de transformation t_{i,m_i}. On cherche $t_{i,m'i}$ pour tout $m' \in$ M \m.$Onfixe$m'$\in$ M \m.$Deuxsous-groupesdelabasedevissontd'intrt$: $legroupedetraitementcomprenantiq$ et le groupe de contrôle qui bénéficie de la stratégie m'.

La quantité qui nous intéresse est l'effet du traitement sur le traité mais à l'échelle individuelle :

$$ATT_{i,m,m'} = E[Y^i_{m'}|D^i = m] - E[Y^i_m|D^i = m] \tag{30}$$

C'est la différence des taux de transformation dans chaque situation pour le traité. La procédure est donc la suivante :

Pour $m \in$ 1-2,3,4,5,6,7,8 :
  Pour $m' \in$ M = 1-2,3,4,5,6,7,8 $\setminus$ m :
    Pour chaque client i ayant la stratégie m :
      On estime $ATT_{i,m'(i)}$

3.6 Hypothèses de Rosenbaum et Rubin (1983)

τ la variation de taux de transformation moyen :

$$\tau|(D = m) = E(Y_{i,m'}|D = m) - E(Y_{i,m}|D = m) \tag{31}$$

Propriété de non-confusion (unconfoundedness propriety)

$$(Y_m, Y_{m'})\perp D|X \tag{32}$$

A caractéristiques équivalentes, les taux de transformation liés à une stratégie sont équivalents peu importe le groupe de stratégie d'appartenance initial.

Propriété de chevauchement

$$0 < P(D = m) < 1 \tag{33}$$

Tous les sujets du groupe m' ont une probabilité non nulle d'appartenir au groupe m c'est-à dire il n'existe pas de variable x telle que l'intersection des valeurs des deux groupes soit vide. Il n'existe donc pas de variable discriminante permettant de distinguer formellement deux stratégies.

Lorsque les propriétés de non-confusion et de chevauchement sont respectées alors les données vérifient l'ignorance forte du traitement. Ce qui implique :

$$\tau|(D = m) = E(Y_{i,m'}|D_i = m) - E(Y_{i,m}|D_i = m)$$

$$= E(E(Y_i|D_i = m, X_i) - E(Y_i|D_i = m', X_i)|D_i = m)$$

$$= E(E(Y_{i,m}|D_i = m, X_i) - E(Y_{i,m'}|D_i = m', X_i)|D_i = m)$$

avec $E(Y_{i,m'}|D_i = m', X_i)$ le taux de transformation d'un sujet du groupe m' ayant les mêmes caractéristiques que le sujet i du groupe m. L'idée est alors d'associer à chaque sujet de la stratégie m, un sujet de la stratégie cible m' qui lui "ressemble".

4 Jumelage des données

Des méthodes de jumelage sont proposées pour pallier à ces problématiques. Il convient de recentrer nos objectifs afin d'opter pour la méthode adéquate en fonction de nos besoins. L'estimation de la quantité $ATT_{i,m,m'}$ [11], et donc de la sensibilité au prix, s'effectue par jumelage : chaque sujet du groupe de traitement m doit être apparié à un sujet du groupe de contrôle m' et adopter son taux de transformation. Plusieurs algorithmes de matching sont disponibles et présentes leurs avantages et leurs inconvénients pour des cas bien spécifiques :

— **Le jumelage exact** Notre base de devis comprend plusieurs variables continues (âge du conducteur, du véhicule, du permis, vitesse du véhicule, valeur du véhicule, montant de prime...) et des variables regroupant de nombreuses catégories (catégorie socio-professionnelle, classe et groupe du véhicule...). Si nous jumelons des sujets qui ont exactement les mêmes valeurs d'attribut mais qui diffèrent seulement dans le traitement qu'ils ont reçu, nous pouvons atteindre un équilibre parfait. Dans le cas où il y a peu de caractéristiques client et aucune variable continue (qui pourrait occasionner des "valeurs vides" de sujets), les données peuvent se superposer. Les variables continues peuvent être discrétisées pour qu'il y ait suffisamment d'individus de traitement et de contrôle pour chaque modalité.
L'équilibre parfait n'est pas possible, même pour un nombre modéré d'attributs dans X (ou si X contient des attributs continus), nous avons donc besoin de méthodes alternatives.L'appariement exact n'est pas adapté à notre base. On a besoin de comprendre de plus près le mécanisme d'affectation des individus au traitement.

— **Le jumelage par score de propension estimé par un modèle logit**

— **Le jumelage par score de propension estimé par un modèle de machine learning robuste (XGBoost)**

— **Le genetic matching**

— **Le matching Frontier (Gary King)** le *matching frontier* essaie de concurrencer le nombre de co-variables impliquées dans l'appariement des individus. Cette méthode est utilisée dans le cas d'un nombre très important de co-variables. Ici, cette technique n'est pas nécessaire au regard des données : la dimension du modèle est correcte vis-à-vis du

11. Remarque : Dans le cas où les sujets qui composent chaque groupe de stratégie ne sont pas interchangeables, l'ATT et l'ATE ne sont plus équivalent. Les chercheurs doivent décider si l'ATE ou le ATT est le plus utile ou le plus intéressant dans leur contexte de recherche. Pour estimer l'efficacité d'un programme intensif de sevrage tabagique, l'ATT peut présenter un plus grand intérêt que l'ATE. En raison des obstacles potentiellement élevés à la participation et à l'achèvement du programme de sevrage tabagique, il peut être irréaliste d'estimer l'effet du programme s'il était appliqué à tous les fumeurs actuels. Au lieu de cela, un plus grand intérêt peut résider dans l'effet du programme sur les fumeurs actuels qui choisissent de participer au programme. En revanche, lors de l'estimation de l'effet sur le sevrage tabagique d'une brochure d'information donnée par les médecins de famille aux patients qui sont fumeurs, l'ATE peut présenter un plus grand intérêt que l'ATT. Le coût et l'effort de distribution d'une brochure d'information sont relativement faibles, et les obstacles à la réception de la brochure par un patient sont minimes (Austin, 2011).
Il est nécessaire de considérer pour chaque prospect, sa réaction face à une variation de prix, ce qui introduit la notion de conditionnement. Par ailleurs, c'est à l'échelle individuelle que l'élasticité-prix sera estimée car tous les devis ne subiront pas de modification de tarif ou tous les devis d'une politique m ne basculeraient pas vers la stratégie m' (pour des raisons de cohérence en termes de tarif à appliquer et de profil)

volume des observations ; la base comprend 168000 devis pour une trentaine de variables utilisées dans le modèle).

Les méthodes de jumelage par score de propension et le genetic matching seront traitées dans le cadre de cette étude.

4.1 Jumelage par score de propension *(Propensity score matching)*

Le propensity score matching emprunte son langage au corps médical. Les problématiques qui ont fait émerger ses contours théoriques appartenaient aux domaines de la médecine et de la pharmaceutique. Le premier article (Rosenbaum et Rubin, 1983) mentionnant cette méthode utilisait une estimation du score de propension pour sous-classer les patients dans une étude des thérapies pour les maladies des artères coronaires. Les traitements étaient le pontage coronarien, D = 1, et la pharmacothérapie, D = 0. Les co-variables x étaient des mesures cliniques, hémodynamiques et démographiques effectuées sur chaque patient avant l'attribution du traitement. L'idée était de mesurer l'effet d'un traitement par rapport à un autre sur les malades. Or, l'affectation à un traitement n'étant pas aléatoire, il est impossible de comparer directement les effets du groupe de traitement au groupe de contrôle.

Peu à peu,la pratique du propensity score matching s'est étendue à des questionnements socio-économiques. Parmi les exemples d'utilisation connue de ces méthodes en dehors du domaine médical figurent l'évaluation des effets de la maternelle sur le développement socio-affectif des enfants (Hong Yu, 2008), l'efficacité des Alcooliques Anonymes (Ye Kaskutas, 2009), les effets de la petite taille des écoles sur les résultats en mathématiques (Wyse, Keesler, Schneider, 2008), et l'effet de la consommation d'alcool des adolescents sur le niveau d'éducation (Staff, Patrick, Loken, Maggs, 2008).
En 2018, sont recensés 93000 articles publiés utilisant des variantes du jumelage par score de propension.

4.1.1 Définition

Les scores d'équilibrage, peuvent être utilisés pour regrouper les unités traitées et les unités de contrôle afin que les comparaisons directes soient plus significatives. Un score d'équilibrage, b(x), est une fonction des covariables observées x telle que la distribution conditionnelle de x donnée par b(x) est la même pour les unités traitées (D = m) et de contrôle (D = m') ; c'est-à-dire [12] (réf. (5)) :

$$X \perp D | b(X) \tag{34}$$

Il s'agit donc de ré-écrire la propriété de non-confusion de manière plus "faible" afin d'alléger les conditions d'application de la méthode d'appariement.
Afin de motiver l'ajustement pour un score d'équilibre, la distribution de l'échantillon pour l'affectation au traitement est considérée. Étant donné les co-variables, la probabilité conditionnelle

12. Dawid,*Conditional Independence in Statistical Theory, J. R. Statist. Soc. B 41*, p.1-31, 1979

d'affectation au traitement est désignée par (Rosenbaum et Rubin, 1983) :

$$\pi(X) = Prob(D = m|X) \tag{35}$$

La fonction $\pi(x)$ est appelée le score de propension, c'est-à-dire la propension à être assigné au traitement m compte tenu des co-variables x observées. Dans une assignation au traitement effectuée au hasard, le score de propension est une fonction connue, de sorte qu'il existe une spécification pour $\pi(x)$. Dans une expérience non randomisée, la fonction du score de propension est presque toujours inconnue, de sorte qu'il n'existe pas de spécification pour $\pi(x)$. Cependant, $\pi(x)$ peut être estimé à partir de données observées, en utilisant un modèle tel que le logit, communément employé. Lorsque deux individus ont le même score de propension alors ils ont la même probabilité de participer au traitement (bénéficier de la stratégie S_m dans notre environnement). En d'autres termes, ils possèdent un support de caractéristiques commun permettant de prédire une même probabilité d'assignation au traitement. Une hypothèse moins forte que l'indépendance peut être émise, sous condition d'existence de ce support commun de caractéristiques entre individus traités et non traités : il s'agit de l'ignorance de traitement.
Cette méthode réduit la dimension du jeu de données à 1, ce qui donne une solution au problème de grande dimension, en ne regardant que les variables qui interviennent dans l'estimation de ce score. La norme L1 standard intervient ensuite pour déterminer le plus proche voisin dans cet espace à une dimension. L'approche classique est de minimiser la distance L1 individuelle. L'algorithme fonctionne comme suit, sachant que l'on souhaite déterminer pour chaque sujet de la stratégie m $\in \{1-2,3,4,5,6,7,8\}$ son score de conversion s'il appartenait à la stratégie m' $\in \{1-2,3,4,5,6,7,8\} - \{m\}$.

1. Estimer le score de propension π pour chaque individu des groupes m et m'.
2. Choisir le premier sujet du groupe m et lui associer le sujet du groupe m' qui a le score de propension le plus proche en valeur absolue.
3. Affecter la marge et le score de conversion du jumeau à ce premier sujet.
4. Recommencer la deuxième étape pour le second sujet du groupe etc.

Il est important de mentionner pour la compréhension de l'algorithme que deux sujets du groupe de traitement peuvent avoir le même jumeau. Ainsi, les sujets du groupe de contrôle qui ressemblent davantage à ceux du groupe de traitement seront davantage sollicités pour le jumelage tandis que les individus éloignés en termes de caractéristiques seront écartés : c'est ainsi que le biais est corrigé, en associant deux individus interchangeables selon le score de propension (condition plus faible que la propriété de non-confusion qui exigeait un appariement selon toutes les caractéristiques).

4.1.2 Propriétés du score de propension–EPBR

Rosenbaum et Rubin ont énoncé et démontré les théorèmes suivants, dans le cas où les covariables seraient distribuées selon une loi Normale :

1. Le score de propension est un score d'équilibre.
2. Tout score "plus fin" que le score de propension est un score d'équilibre ; en outre, x est le score d'équilibre le plus fin et le score de propension le plus grossier.

3. Si l'hypothèse d'ignorance de traitement forte est vérifiée (propriété de non-confusion + chevauchement), alors l'ignorance de traitement ($(Y_m, Y_{m'}) \perp D|b(X)$ et $0 < P(D = 1|b(X)) < 1$) est respectée pour tout score d'équilibrage.
4. Pour toute valeur de score d'équilibre, la différence des moyennes entre les réponses du groupe de traitement et de contrôle est une estimation non biaisée de l'ATE à cette valeur du score d'équilibre, si l'ignorance au traitement est forte. Par conséquent, si l'affectation au traitement est fortement ignorable, le jumelage des paires sur un score d'équilibrage, la sous-classification sur un score d'équilibrage et l'ajustement de covariance sur un score d'équilibrage peuvent tous produire des estimations non biaisées des effets du traitement.
5. L'utilisation d'estimations des scores d'équilibre à partir de l'échantillon X peut conduire à un équilibre de l'échantillon X. $\hat{b}$ est donc un score d'équilibrage plus grossier que le score de propension

Autrement dit, si les hypothèses de non-confusion et de chevauchement sont vérifiées et que le score d'équilibrage suit une loi normale alors le jumelage par score de propension permet de corriger le biais présent dans l'estimation de l'ATT (ou de l'élasticité au prix). En outre, sous l'hypothèse de normalité, la méthode est *Equal Per cent Bias Reducing* (EPBR), c'est- à dire que le biais présent pour chaque co-variable est réduit dans les mêmes proportions pour chacune d'entre elles.
Le biais initial est définit par B : B = E(x|D = m) - E(x|D = m') Et le biais sur x entre l'échantillon de traitement et le sous-échantillon issu du groupe de contrôle obtenu par la méthode de jumelage est de : $B_j = E(x|D = m) - E_j(x|D = m')$ Sous la propriété de EPBR, $B = \gamma * B_j$ avec $0 < \gamma < 1$: le biais sur chaque coordonnées de x est réduit par $100(1 - \gamma)\%$. Si l'EPBR n'est pas vérifiée alors il existe un vecteur w tel que wB_j>wB et donc le jumelage augmente le biais et dégrade l'estimation de l'élasticité pour une fonction linéaire de x.

4.1.3 Propensity score matching pour l'effet d'une stratégie tarifaire sur le taux de renouvellement en assurance automobile

La méthode du jumelage par score de propension a déjà été mentionnée en tant que solution pour corriger le biais de sélection, proposée lors d'un séminaire à Dallas du 9-11 mars 2015 [13].

Le contexte d'étude du séminaire sera repris fidèlement dans cette section, quitte à reprendre les problématiques de biais déjà énoncées. Puis, les procédés présentés dans ce séminaire seront adaptés à nos objectifs.

L'environnement d'étude demeure limitrophe au nôtre : il s'agit d'estimer l'élasticité-prix en assurance automobile mais cette fois, pas au niveau de la première souscription mais lors du renouvellement.
Usuellement, pour prédire si un individu va renouveler son contrat, un ensemble d'attributs statiques, qui inclut notamment le tarif proposé par l'assureur, contribue au modèle de prédiction. Or, chercher à établir une nouvelle politique de tarifs afin d'améliorer sa marge ou son chiffre d'affaires implique la modification des tarifs au sein des caractéristiques assurés.

13. Kim & Guelman, *Propensity Scoring : Theory and Applications*, 2015

Comme la politique n'a pas été mise en place et qu'on essaie de prédire l'output Y (ici si le contrat de l'individu a été reconduit ou non), la même confrontation au problème d'extrapolation des données a lieu : le modèle d'apprentissage n'a pas été alimenté avec ces caractéristiques.

Chaque individu a reçu un traitement (une stratégie tarifaire ou combinaison de stratégies tarifaires) en date n. Le tarif de chaque individu a alors évolué entre n-1 et n. On dispose alors dans l'historique de l'élasticité-prix entre n-1 et n. Mais les stratégies tarifaires effectuées entre ces deux dates ne sont pas nécessairement optimales : l'idée dans le cadre d'une optimisation tarifaire est de trouver la stratégie optimale pour chaque individu, c'est-à dire une stratégie qui permet le renouvellement tout en maximisant la marge individuelle. De multiples stratégies ont été élaborées entre n-1 et n. C'est parmi cet ensemble de stratégies E que va être appliquée celle pour un individu en particulier en n+1. Comme expliqué à la section précédente, il est impossible d'attribuer à un premier individu en date n+1 l'élasticité-prix entre n-1 et n d'un deuxième individu qui a subi la stratégie tarifaire en n que l'on souhaiterait appliquer au premier individu en n+1. En effet, les deux individus n'ont pas le même profil car les traitements n'ont pas été attribués au hasard et n'ont, par conséquence, la même réaction face à une évolution des prix.

L'objectif est de déterminer grâce aux méthodes de propensity score matching, pour chaque stratégie de E, quelle aurait été la probabilité d'un renouvellement pour chaque individu. On aurait ainsi une estimation du volume de portefeuille et de primes pour chacune des stratégies tarifaires individuelles. On pourrait alors, à l'échelle du portefeuille, déterminer la frontière d'efficience composée de toutes les stratégies individuelles qui permettent d'atteindre les niveaux de rentabilité et de chiffre d'affaires cibles.

Supposons qu'entre n-1 et n, deux stratégies tarifaires ont été déployées :

1. Le tarif a augmenté de 10% (D = 1)
2. Le tarif a augmenté de 5% (D = 0)

Dans l'encadré ci-dessous, est proposé un exemple où les deux stratégies ont pu être appliquées à l'ensemble des individus :

Taux d'augmentation de la prime	+5%	+10%
Nombre de contrats	10000	10000
Contrats renouvelés	9200	8700
Taux de rétention	92%	87%

L'ATE (pour rappel, ATE = $E[Y_{(1)} - Y_{(0)}]$) correspond ici à l'élasticité-prix car les individus sont interchangeables et répond à la question : "comment se répercute une augmentation de tarif de 10% par rapport à une augmentation de 5% sur le taux de transformation ?". On a $\hat{ATE}$ = 87% - 92% = -5% : appliquer un accroissement des tarifs de 10% au lieu de 5% a pour conséquence une diminution de la demande de renouvellement à hauteur de 5%.

Dans la réalité, les stratégies tarifaires adoptées historiquement dépendent du profil et les individus ne sont alors plus interchangeables [14] (réf. (16)). Une action sur le tarif ne peut jamais avoir lieu pour une partie de la population d'assurés.

Taux d'augmentation de la prime	+5%	+10%
Age $<$ 25 ans	$\surd$	$\surd$
Age $>=$ 25 ans	$\surd$	?

Une idée courante est de mesurer l'ATE par régression linéaire, par estimation du coefficient μ :

$$Y = \beta_0 + \mu D + \beta X + \epsilon \tag{36}$$

Les méthodes basées sur la régression extrapolent les inférences dans les régions des prédicteurs où les traitements n'ont pas été observés [15] (réf. (3)). Le grand volume de caractéristiques client peut camoufler le problème car il est plus difficile de remarquer le manque de chevauchement de celles-ci qui pourrait justifier une analogie entre les profils.

L'absence de données expérimentales conduit donc à un biais appelé "biais de sélection" car il découle du choix du type d'assurés pour une stratégie donnée. Rosenbaun et Rubin (1983) proposent des algorithmes de matching pour corriger le bais de sélection dans l'estimation de l'effet de traitement. Ils supposent l'existence d'un support commun de caractéristiques, ce qui émet l'hypothèse que des assurés similaires pour cet ensemble de variables appelé support commun ont subi des traitements différents. Les estimations des effets de traitement ne sont fiables que dans la région de chevauchement. En dehors de cette région, l'estimation de l'effet causal implique une extrapolation. Une autre hypothèse est émise selon laquelle les stratégies historiques ont été fondées sur les attributs observés X de l'assuré : aucun autre attribut n'a d'influence sur l'action ou la réponse. Les sujets sont donc affectés aux traitements (stratégies) sur la base de leurs attributs X. Pour détecter cette zone de caractéristiques communes entre deux individus ayant reçu deux traitements différents, comprendre le mécanisme d'assignation au traitement est primordial. La probabilité conditionnelle d'assignation au traitement se définie comme :

$$\pi(X) = P(D = 1|X) \tag{37}$$

L'estimation de π donne le score de propension, qui s'interprète comme la probabilité d'appartenir à la stratégie telle que D = 1. Si l'on souhaite appliquée la stratégie s à un assuré i (qui ne s'est jamais vu appliquer s), les algorithmes de matching vont jumeler l'assuré i à un autre assuré parmi ceux qui ont historiquement bénéficié de la stratégie s. Pour effectuer cet appariement, on associe à l'individu i l'assuré le plus proche de lui en termes de caractéristiques présentes dans le support commun. En d'autres termes, la paire de sujets diffèrent dans le traitement qu'ils ont reçu mais bénéficient de la même probabilité d'affectation au traitement, ie

14. Rosenbaum & Rubin, *The central role of the propensity score in observational studies for causal effects, Biometrika*, p. 41-55, 1983

15. Berk, *Regression Analysis : A Constructive Critique*, Chap. 5, 2004

d'un score de propension similaire.

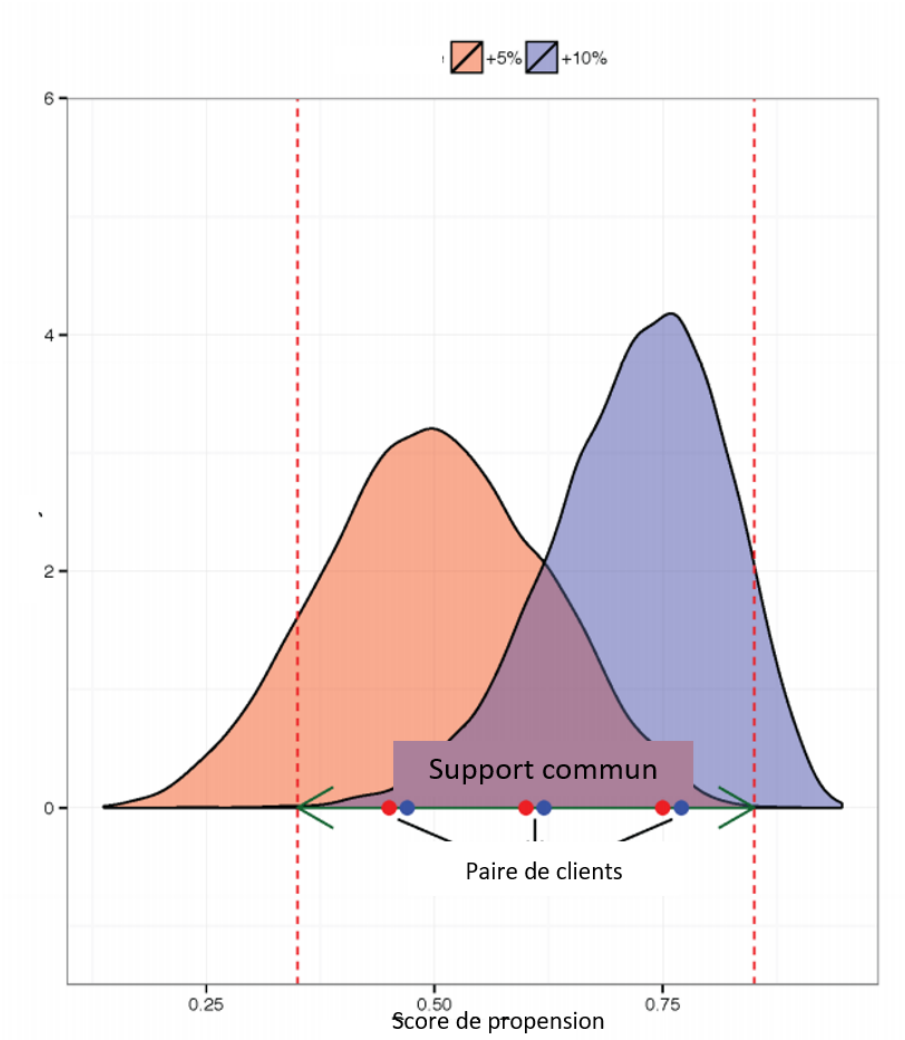


FIGURE 31 – Appariement dans la région de chevauchement selon le score de propension

Au sein de l'échantillon apparié, la mise en place d'une stratégie est indépendante des caractéristiques. Les jumeaux sont interchangeables car ils ont la même probabilité d'être affecté au traitement :

$$D \perp X|\pi(X)$$

La distribution de X est la même pour chaque paire. Sur la figure ci-dessus, si les deux densités étaient superposées, alors les individus des deux groupes seraient interchangeables et il n'y aurait eu nul besoin de procéder à l'estimation du score de propension. A l'opposé, si les deux courbes ne se juxtaposaient jamais (ou très peu) alors les deux sous-classes de sujets ne possèdent aucune propriété commune et l'appariement est impossible.

On rappelle les enjeux présentés au cours du séminaire :

1. Obtenir des estimations de l'élasticité des prix de l'assurance automobile niveau du portefeuille (Effet de traitement moyen ou ATE). L'action a-t-elle une incidence sur le taux

de renouvellement ?

2. Identifier les sous-groupes de clients ayant une sensibilité variable au prix. L'action affecte-t-elle le taux de renouvellement différemment selon le type de client ?

Les clients ont été historiquement exposés à des niveaux de changement de tarifs basés sur (i) un exercice de modélisation des prix, (ii) des contraintes réglementaires, (iii) une analyse de la concurrence, et (iv) des objectifs commerciaux généraux.

$r_{l,d} \in \{0,1\}$ indiquent le résultat (renouvellement ou non) observé chez l'assuré l = 1,....L lorsqu'il est exposé au niveau de changement de taux D=d . Ici, le portefeuille compte 5 politiques tarifaires différentes perpétrées dans le passé. Les points signifient que la stratégie n'est pas observée pour l'individu.

Client	Stratégie 1	Stratégie 2	Stratégie 3	Stratégie 4	Stratégie 5
1	.	$r_{1,2}$	.	.	.
2	.	.	$r_{2,3}$	.	.
3	$r_{3,1}$	.	.	.	.
4	.	.	.	$r_{4,4}$	.
5	.	$r_{5,2}$	.	.	.
6	.	.	.	.	$r_{6,5}$
...	...	...	...	...	...
L	.	.	.	.	$r_{L,5}$

TABLE 8 – Indicatrice de renouvellement pour les stratégies historiques adoptées

Le protocole suivant sera appliqué :

Étape 1 : Remplacer les outputs réels décrivant le renouvellement par des estimations de probabilité.

Client	Stratégie 1	Stratégie 2	Stratégie 3	Stratégie 4	Stratégie 5
1	.	$\hat{r}_{1,2}$	.	.	.
2	.	.	$\hat{r}_{2,3}$	.	.
3	$\hat{r}_{3,1}$	.	.	.	.
4	.	.	.	$\hat{r}_{4,4}$	.
5	.	$\hat{r}_{5,2}$	.	.	.
6	.	.	.	.	$\hat{r}_{6,5}$
...	...	...	...	...	...
L	.	.	.	.	$\hat{r}_{L,5}$

TABLE 9 – Estimation des taux de renouvellement par GLM

Étape 2 : déduire le score de renouvellement pour chaque stratégie et chaque individu à partir des appariements obtenus par algorithme de matching (dans la mesure où le chevauchement est possible).

Client	Stratégie 1	Stratégie 2	Stratégie 3	Stratégie 4	Stratégie 5
1	$\hat{r}_{1,1}$	$\hat{r}_{1,2}$	$\hat{r}_{1,3}$	$\hat{r}_{1,4}$	$\hat{r}_{1,5}$
2	$\hat{r}_{2,1}$	$\hat{r}_{2,2}$	$\hat{r}_{2,3}$	$\hat{r}_{2,4}$	$\hat{r}_{2,5}$
3	$\hat{r}_{3,1}$	$\hat{r}_{3,2}$	$\hat{r}_{3,3}$	$\hat{r}_{3,4}$	$\hat{r}_{3,5}$
4	$\hat{r}_{4,1}$	$\hat{r}_{4,2}$	$\hat{r}_{4,3}$	$\hat{r}_{4,4}$	$\hat{r}_{4,5}$
5	$\hat{r}_{5,1}$	$\hat{r}_{5,2}$	$\hat{r}_{5,3}$	$\hat{r}_{5,4}$	$\hat{r}_{5,5}$
6	$\hat{r}_{6,1}$	$\hat{r}_{6,2}$	$\hat{r}_{6,3}$	$\hat{r}_{6,4}$	$\hat{r}_{6,5}$
...	...	...	...	...	...
L	$\hat{r}_{L,1}$	$\hat{r}_{L,2}$	$\hat{r}_{L,3}$	$\hat{r}_{L,4}$	$\hat{r}_{L,5}$

TABLE 10 – Estimation des taux de renouvellement par GLM

Étape 3 : Élaboration d'un "modèle global" de la réponse correspondant aux estimations des taux de renouvellement observés et associés.

Client	Stratégie 1	Stratégie 2	Stratégie 3	Stratégie 4	Stratégie 5
1	$\hat{\hat{r}}_{1,1}$	$\hat{\hat{r}}_{1,2}$	$\hat{\hat{r}}_{1,3}$	$\hat{\hat{r}}_{1,4}$	$\hat{\hat{r}}_{1,5}$
2	$\hat{\hat{r}}_{2,1}$	$\hat{\hat{r}}_{2,2}$	$\hat{\hat{r}}_{2,3}$	$\hat{\hat{r}}_{2,4}$	$\hat{\hat{r}}_{2,5}$
3	$\hat{\hat{r}}_{3,1}$	$\hat{\hat{r}}_{3,2}$	$\hat{\hat{r}}_{3,3}$	$\hat{\hat{r}}_{3,4}$	$\hat{\hat{r}}_{3,5}$
4	$\hat{\hat{r}}_{4,1}$	$\hat{\hat{r}}_{4,2}$	$\hat{\hat{r}}_{4,3}$	$\hat{\hat{r}}_{4,4}$	$\hat{\hat{r}}_{4,5}$
5	$\hat{\hat{r}}_{5,1}$	$\hat{\hat{r}}_{5,2}$	$\hat{\hat{r}}_{5,3}$	$\hat{\hat{r}}_{5,4}$	$\hat{\hat{r}}_{5,5}$
6	$\hat{\hat{r}}_{6,1}$	$\hat{\hat{r}}_{6,2}$	$\hat{\hat{r}}_{6,3}$	$\hat{\hat{r}}_{6,4}$	$\hat{\hat{r}}_{6,5}$
...	...	...	...	...	...
L	$\hat{\hat{r}}_{L,1}$	$\hat{\hat{r}}_{L,2}$	$\hat{\hat{r}}_{L,3}$	$\hat{\hat{r}}_{L,4}$	$\hat{\hat{r}}_{L,5}$

4.1.4 Propensity score matching pour l'effet d'une stratégie tarifaire sur le taux de conversion en assurance automobile, exemple fictif

Dans cette section, un exemple fictif est introduit afin de comprendre facilement le fonctionnement du propensity score matching.
La stratégie 3 est considérée (au hasard), elle constitue donc le groupe de traitement. L'assureur peut vouloir accroître sa marge au sein de cette classe et ainsi, faire basculer chacun des assurés de la stratégie 3 sur la stratégie 4 (groupe de contrôle, choisi au hasard) qui est délimitée par des niveaux de marge plus élevés. Cependant, une augmentation de la marge induirait une augmentation du tarif TTC, ce qui, avec la dégradation de son positionnement sur le marché sur ce segment, entraînerait une baisse de la conversion qui serait à contrôler.

Dans cet exemple fictif, les sujets de la stratégie 3 sont au nombre de 3, ils sont agriculteurs (sujet 1 et 2) et ouvrier (sujet 3) et ont 22,24 et 29 ans respectivement. Leurs taux de transformation sont de 34,36 et 41%. Les sujets de la stratégie 4 sont également au nombre de 3, sont majoritairement ouvriers (sujet 5 et 6) et agriculteur (sujet 4). Ils ont 26, 33 et 45 ans

(ils sont donc en moyenne plus âgés que les assurés du groupe 3). Leurs taux de transformation sont de 30,38 et 40%.

Sujet	stratégie	âge	csp	taux de transformation	Score de propension	jumeau	taux de transformation_jumeau
1	3	22	Agriculteur	0.34	0.8	4	0.30
2	3	24	Agriculteur	0.36	0.75	4	0.30
3	3	29	Ouvrier	0.43	0.6	5	0.38
4	4	26	Agriculteur	0.30	0.73	-	-
5	4	33	Ouvrier	0.38	0.5	-	-
6	4	45	Ouvrier	0.40	0.2	-	-

FIGURE 32 – Exemple fictif

Le score de propension est estimé avec les caractéristiques (métier et âge), il s'agit de la probabilité d'appartenir à la stratégie 3 pour chacun des 6 sujets. Le sujet 1 qui a 22 ans et est agriculteur et qui appartient réellement à la stratégie de traitement, a le score le plus élevé de 80%. Vient ensuite le sujet 2 de 24 ans et agriculteur également avec un score de 0.75. En revanche, le sujet 3 a le score le plus faible parmi son groupe-0.6- qui s'explique par le fait qu'il soit un peu plus âgé -29 ans- et ouvrier. Parmi les sujets du groupe 4, le sujet 4 a la probabilité d'appartenir au groupe 3 la plus élevée avec un score de 0.73. Le sujet 5 a un score de 0.5 et le sujet 6 de 0.2 (c'est l'assuré le plus âgé).
L'algorithme apparie ensuite chaque assuré de la stratégie 3 à un sujet de la stratégie 4 et lui associe son taux de conversion. Le sujet 1 est donc associé au sujet 4 (0.73 est le score le plus proche de 0.8), le sujet 2 est également associé au sujet 4 (0.73 est le plus proche de 0.75), le sujet 3 est apparié au sujet 5 (0.5 est le plus proche de 0.6). Le sujet 6 est évincé du matching puisqu'il est considéré par l'algorithme comme trop éloigné des sujets de la stratégie 3 en termes de profil. Au final, la moyenne des taux de conversion des sujets de la stratégie 3 s'ils appartenaient à la stratégie 4 est inférieure à la moyenne des taux de conversion des sujets de la stratégie 4.
Deux conditions sont indispensables pour que le jumelage par score de propension fonctionne :

1. Il faut que le modèle qui estime le score de propension soit performant pour exclure les individus de contrôle dont le profil est trop éloigné de ceux du groupe de traitement.
2. Il faut que des sujets du groupe de contrôle ressemblent aux profils du groupe de traitement, autrement le jumelage n'a pas de sens.

Sujet	stratégie	âge	csp	taux de transformation	Score de propension	jumeau	taux de transformation_jumeau
1	3	22	Agriculteur	0.34	0.8	4	0.30
2	3	24	Agriculteur	0.36	0.75	5	0.38
3	3	29	Ouvrier	0.43	0.3	6	0.40
4	4	26	Agriculteur	0.30	0.95	-	-
5	4	33	Ouvrier	0.38	0.5	-	-
6	4	45	Ouvrier	0.40	0.2	-	-

FIGURE 33 – Illustration de la condition 1.

La condition 1 fait écho au principe de non-confusion et suggère le fait que :

— le modèle doit capter les facteurs discriminant le groupe de traitement du groupe de contrôle afin de faire mieux que l'appariement aléatoire initial qui se base sur l'hypothèse que les individus soient tous interchangeables.

— le modèle doit pouvoir sélectionner les facteurs les plus importants et les bonnes combinaisons d'interactions pour ne pas répartir les sujets de manière erronée.

Si le score de propension est mal estimé pour les sujets 3 et 4, passant de 0.6 et 0.73 à 0.3 et 0.95 alors l'appariement est complètement modifié. Le sujet 1 est toujours jumelé au sujet 4 mais le sujet 2 est associé au sujet 5 et le sujet 3 au sujet 6 qui entre maintenant dans le matching. Au final, aucun assuré du groupe de contrôle n'a été éliminé : l'algorithme considère que chaque sujet de la stratégie 4 peut avoir un comportement similaire à celui de la stratégie 3. La moyenne des taux de transformation avant et après le jumelage est la même ($\frac{0.3+0.38+0.4}{3}$).

Sujet	stratégie	âge	csp	taux de transformation	Score de propension	jumeau	taux de transformation_jumeau
1	3	22	Agriculteur	0.34	0.8	4	0.30
2	3	24	Agriculteur	0.36	0.75	4	0.30
3	3	29	Agriculteur	0.43	0.7	4	0.30
4	4	56	Ouvrier	0.30	0.3	-	-
5	4	43	Ouvrier	0.38	0.25	-	-
6	4	45	Ouvrier	0.40	0.2	-	-

FIGURE 34 – Illustration de la condition 2

Avant de vouloir modifier la politique tarifaire d'une population, l'assureur doit vérifier la cohérence de cette action. Attribuer à un bas risque un tarif exorbitant, qui est habituellement donné à un haut risque est puéril (un haut risque accepte mieux de payer une assurance élevée). Si l'exemple est modifié et que la stratégie 3 est exclusivement composée d'agriculteurs ayant une vingtaine d'années et que la stratégie 4 est composée d'ouvriers ayant plus de 40 ans, alors le comportement d'aucun des profils de la stratégie 4 ne peut être comparé à un profil de la stratégie 3, dans le cas où ils bénéficieraient du même tarif. Le modèle estime des scores de propension supérieurs à 70% pour le groupe 3 et inférieurs à 30% pour le groupe 4 (il n'existe pas de support commun de caractéristiques entre les deux groupes). Tous les sujets sont appariés au sujet 4 mais l'estimation du taux de transformation de 30% est erronée car la propriété de non-chevauchement des données n'est pas respectée.

4.1.5 Estimation robuste du score de propension

Une estimation précise du score de propension évite les problèmes de spécification. En effet, une estimation pauvre conduit à de mauvais appariements. Des modèles comme les forêts aléatoires ou le XGBoost sont privilégiés dans ce cadre.

La qualité de l'appariement des observations se mesure en comparant les distributions de chaque variable au sein des groupes de traitement et de contrôle avant et après jumelage. Le but est de gommer les différences de caractéristiques du groupe de contrôle pour qu'ils deviennent interchangeables avec celles du groupe de traitement. L'exemple précédent est repris :

Sujet	stratégie	âge	csp	taux de transformation	Score de propension	jumeau	taux de transformation_jumeau
1	3	22	Agriculteur	0.34	0.8	4	0.30
2	3	24	Agriculteur	0.36	0.75	4	0.30
3	3	29	Ouvrier	0.43	0.6	5	0.38
4	4	26	Agriculteur	0.30	0.73	-	-
5	4	33	Ouvrier	0.38	0.5	-	-
6	4	45	Ouvrier	0.40	0.2	-	-

FIGURE 35 – Appariement pour l'exemple fictif

	Traitement	Contrôle avant matching	Contrôle après matching
Agriculteur	67%	33%	67%
Ouvrier	33%	67%	33%

	Traitement	Contrôle avant matching	Contrôle après matching
Age (moyenne)	25	34,67	28,33

FIGURE 36 – Comparaison des distributions avant et après le jumelage (exemple fictif). La distribution de la csp du groupe de contrôle après le jumelage est identique à celle du groupe de traitement. La correction du biais a bien eu lieu sur cette variable puisque les sujets sont devenus interchangeables entre les deux groupes. L'âge moyen s'est également rapproché de celui du groupe de traitement en passant de 28.33 à 34.67 ans (en pratique ce sont les fonctions de distributions qui sont comparées dans le cas de données continues).

4.1.6 Résultats du jumelage par score de propension

4.1.6.1 Courbes de sensibilité estimées

Sur chacun des graphes sont présentés la courbe des taux de transformation moyen de chaque groupe tarifaire en orange et la courbe de sensibilité de la stratégie de traitement considérée en bleu. Deux méthodes sont comparées : le GLM et le XGBoost.

— Plus la marge augmente (de la stratégie 1_2 à la stratégie 8) et plus la conversion diminue, et ce, peu importe les groupes de profils considérés.

— Plus on s'éloigne du groupe de traitement en termes de marge et plus la courbe de sensibilité s'éloigne de la courbe des taux de transformation moyen d'origine des groupes de stratégies. C'est dû au fait que l'algorithme élimine d'autant plus d'individus de contrôle qu'ils sont différents de ceux du groupe de traitement. C'est aussi pour cette raison que la stratégie de traitement n'est comparée qu'aux deux stratégies voisines : au-delà de cette limite, les résultats sont moins robustes et la démarche peut être incohérente.

— Le GLM est moins précis que le XGBoost dans l'estimation du score de propension. Les taux moyens estimés avant et après le jumelage sont égaux pour la plupart des stratégies voisines, ce qui signifie que le GLM n'a pas capté les facteurs de différenciation entre les groupes. Le jumelage n'a donc pas corrigé le biais émanent du fait que les profils ne soient pas interchangeables pour deux politiques tarifaires distinctes considérées.

— Le taux de conversion moyen des sujets de traitement est moins élevé que celui des sujets de contrôle lorsque ces derniers ont une marge plus faible. En revanche, le taux de conversion moyen des sujets de traitement est plus élevé que celui des sujets de contrôle lorsque ces derniers ont une marge plus forte. C'est d'autant plus vrai quand le groupe de traitement bénéficie d'une marge faible. Il est donc plus intéressant d'augmenter les marges pour les prospects à marge faible et de diminuer les marges pour les prospects à marge élevée si l'assureur cherche un levier pour accroître le volume de son portefeuille tout en conservant un niveau de marge correct.

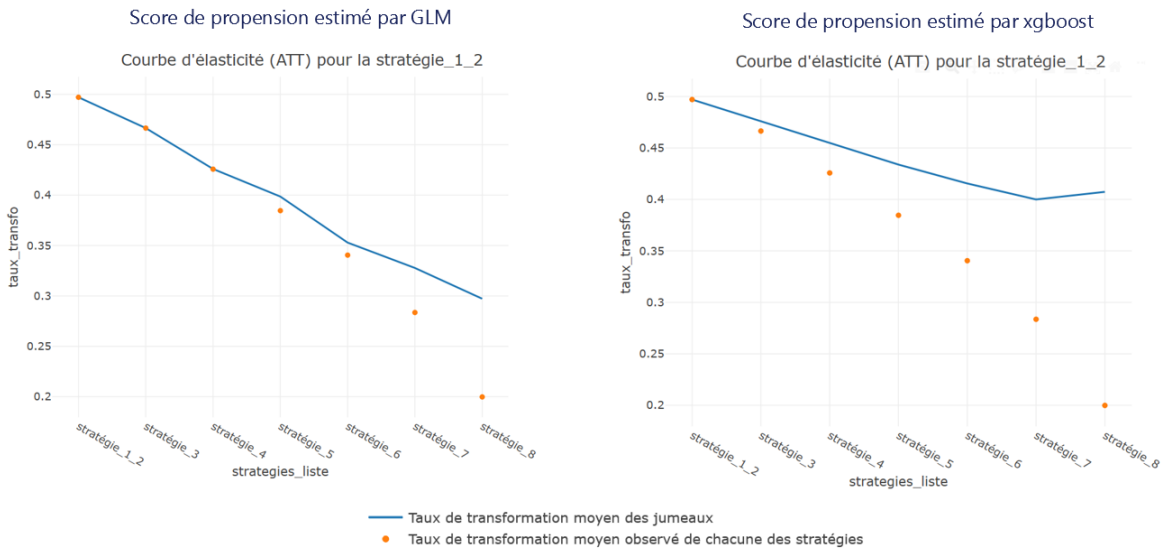


FIGURE 37 – Lorsque le score de conversion est estimé par GLM alors l'algorithme considère comme interchangeables les sujets des stratégies 1_2, 3 et 4 puisque les courbes sont confondues en ces points. Le matching n'a pas réussi à capter ce qui différencie les sujets de contrôle des sujets de traitement contrairement au XGBoost qui est plus performant. La conversion moyenne de la stratégie 1_2 est supérieure à celle des stratégies 3 et 4 à marge équivalente, l'élasticité au prix est évidemment négative mais plus faible en valeur absolue que si on comparait naïvement les sujets entre les classes.

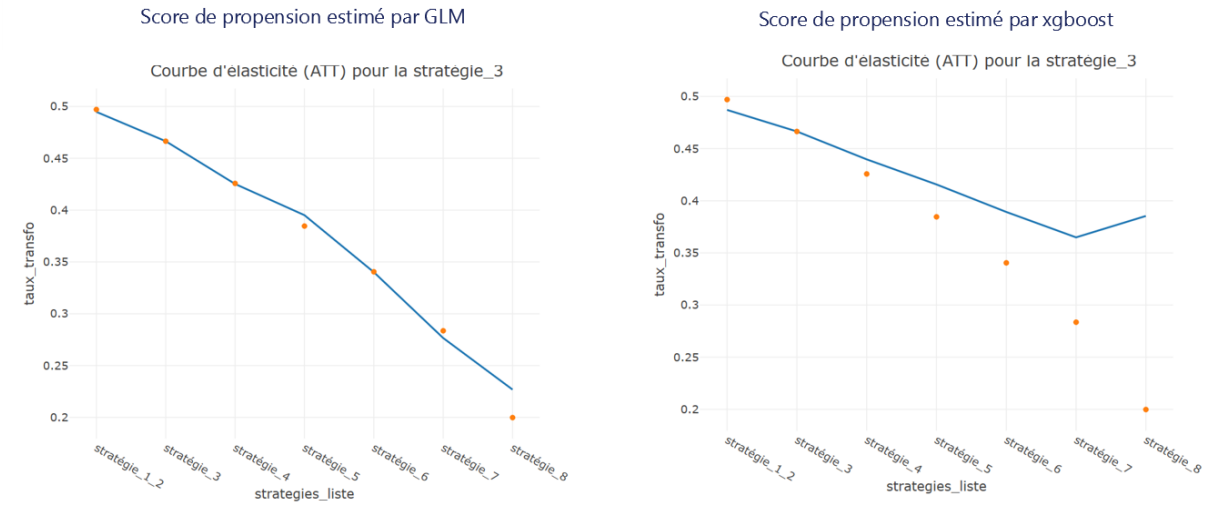


FIGURE 38 – Le GLM ne corrige aucunement le biais induit par la différence de profil entre les différentes politiques tarifaires, le modèle XGBoost est conservé. Si l'assureur diminuait la marge des sujets du groupe 3, la conversion serait moins bonne que celle des assurés ayant cette marge aujourd'hui. Le raisonnement est inversé s'il augmentait la marge.

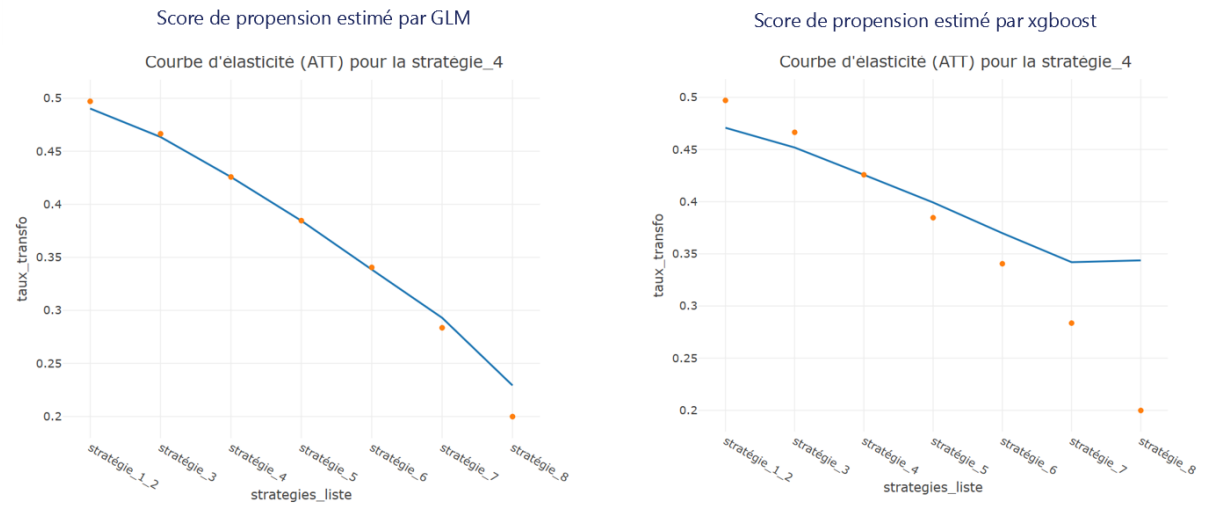


FIGURE 39 – Le GLM ne distingue pas les distinctions de profil et considère que l'élasticité au prix est la même pour tous les sujets du portefeuille. Du côté du XGBoost, le même constat que celui de la stratégie 3 est observé : il est plus judicieux d'augmenter la marge puisque la transformation n'atteint pas celle de ceux bénéficiant d'une marge élevée. En revanche rogner sur la marge ne permet pas d'accroître le portefeuille de la même façon que pour les personnes qui bénéficient d'un tarif plus compétitif.

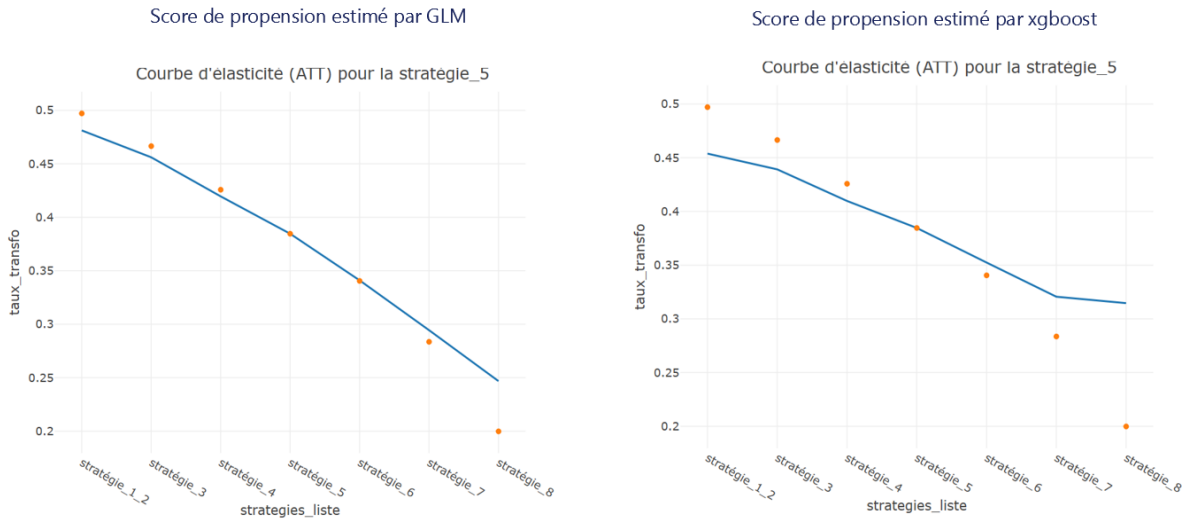


FIGURE 40 – Le GLM capte les distinctions de profils que dans le cas où l'écart est important : entre les assurés avec la politique tarifaire 5 et ceux avec la politique tarifaire 8.

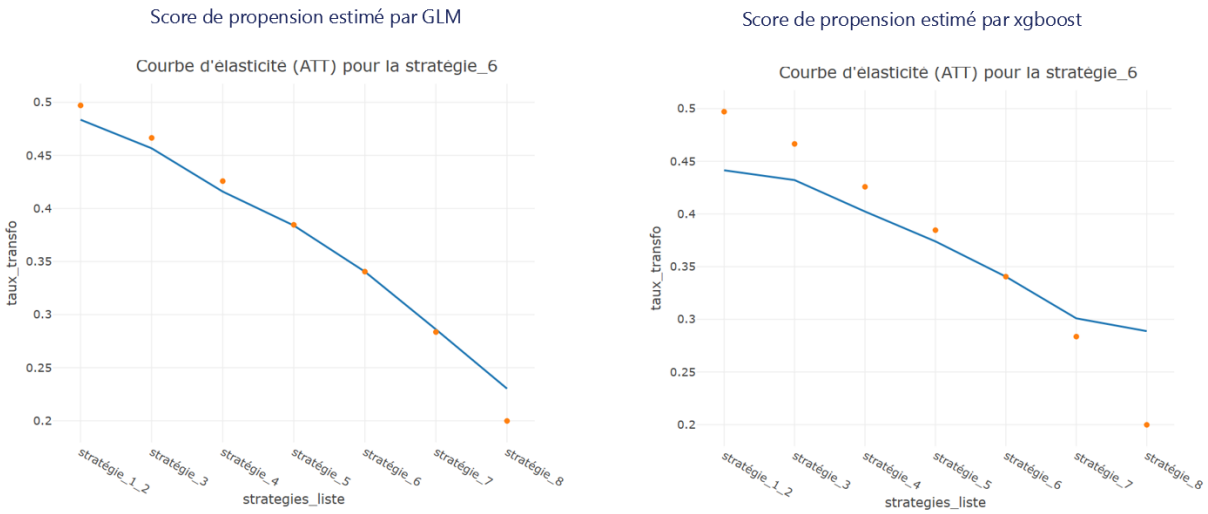


FIGURE 41 – C'est à partir de la stratégie 6 en tant que groupe de traitement que la courbe de sensibilité est décroissante pour la méthode avec XGBoost. Pour les marges plus faibles, elle remontait entre la stratégie 7 et la stratégie 8, signe que le nombre d'individus de contrôle était faible et que ceux sélectionnés avaient un taux de transformation plus élevé. L'interprétation n'était donc pas cohérente puisqu'il valait mieux appliquer un tarif plus élevé pour gagner en volume de portefeuille.

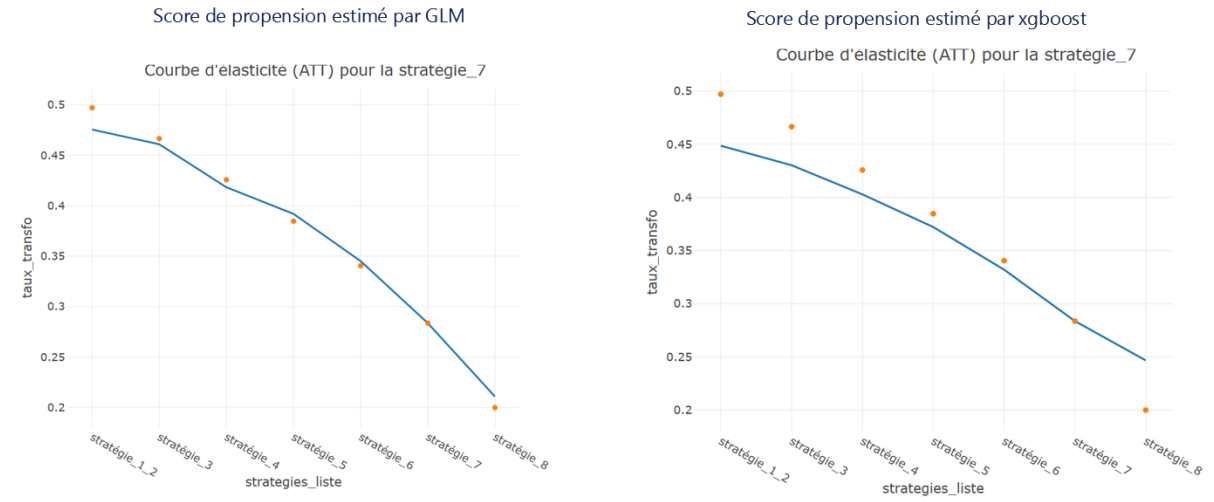


FIGURE 42 – Le GLM ne voit aucune distinction entre les groupes 6, 7 et 8. En moyenne, gain de conversion est intéressant si la marge des assurés dans la stratégie 7 était celle appliquée dans la stratégie 6.

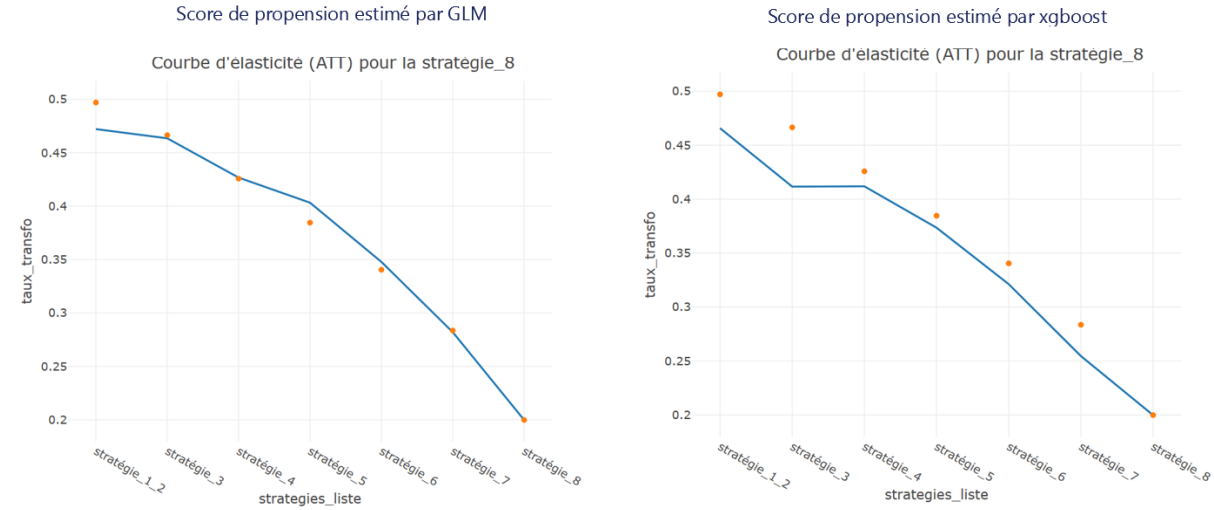


FIGURE 43 – Le gain de conversion dû à un abaissement de prix est en moyenne plus efficace lorsqu'il est effectué sur des prospects ayant à l'origine une marge très élevée. C'est le cas ici pour la stratégie 8 où on se rapproche davantage des points de référence orange.

4.1.6.2 Correction du biais : mesure de la qualité de l'appariement sur plusieurs exemples

Il s'agit de vérifier dans cette section que l'appariement a réellement permis de corriger le biais émanent des différences de profils entre les groupes tarifaires. Les fonctions de répartition entre la stratégie de traitement et de contrôle avant et après le jumelage sont analysées et ce, pour les estimations du score par GLM et XGBoost. Le même travail est effectué pour les variables catégorielles en comparant les différences de proportion. En bleu sera représentée la stratégie de traitement et en rouge la stratégie de contrôle. Les résultats suivants sont constatés :

— Le jumelage permet de rapprocher les prospects de contrôle à ceux de traitement en termes de caractéristiques. Les probabilités des facteurs discriminant les groupes deviennent si-

milaires sans dégrader les similarités dans les distributions similaires entre les groupes.

— La méthode n'est pas performante quand il s'agit de rapprocher des stratégies trop éloignées.

— Le XGBoost est plus performant dans la détection des sujets de contrôle éloignés des sujets de traitement.

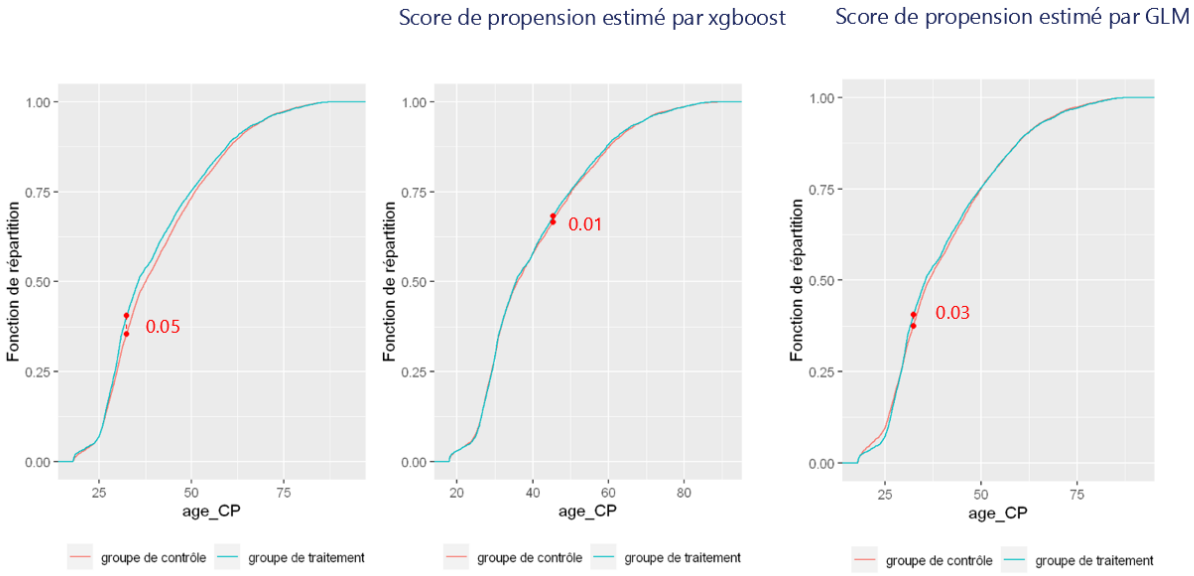


FIGURE 44 – La stratégie 1_2 est le traitement et la stratégie 3 est le contrôle. Le plus grand écart entre les fonctions de répartition de l'âge du conducteur principal des deux groupes est atténué par le jumelage, d'autant plus fortement que le modèle estimant le score d'appartenance à la stratégie 1_2 est précis.

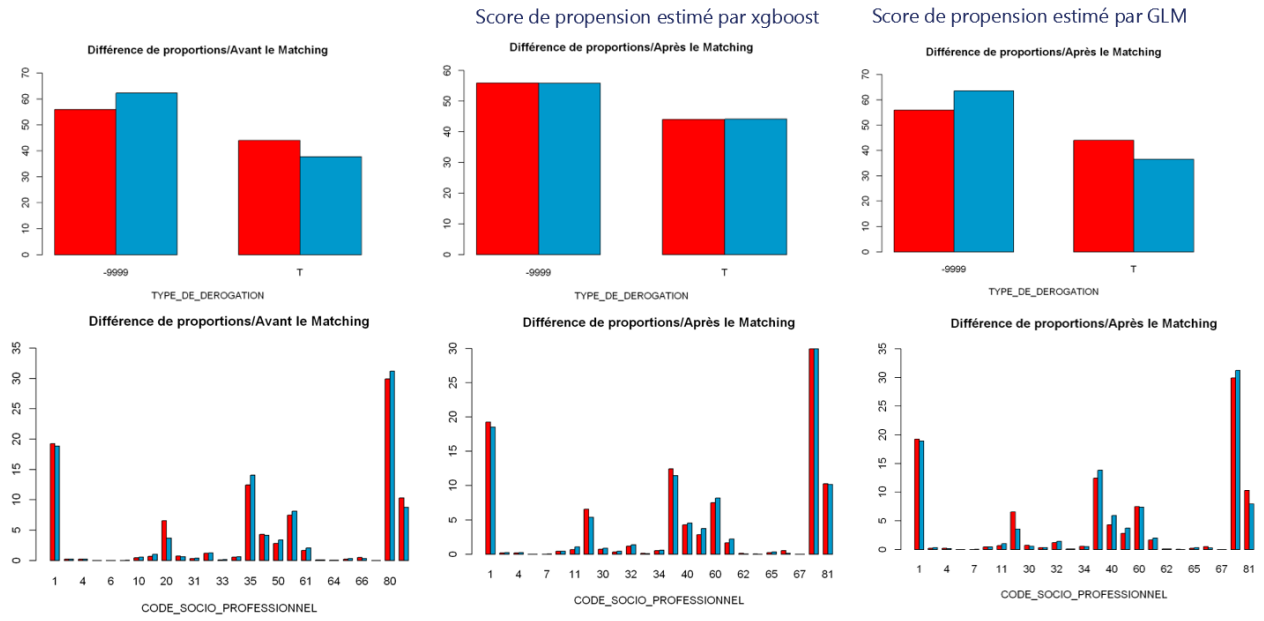


FIGURE 45 – La stratégie 1_2 est le traitement et la stratégie 3 est le contrôle. Les différences de proportion entre chaque catégorie est corrigée avec l'appariement XGoost, ce qui n'est pas le cas avec le GLM. Si on se focalise sur la catégorie socio-professionnelle majoritaire (80), le XGBoost a réussi à combler la faible différence entre les deux distributions des deux groupes et élimine donc avec précision les sujets de contrôle qui ne peuvent pas être jumelé à un sujet de traitement.

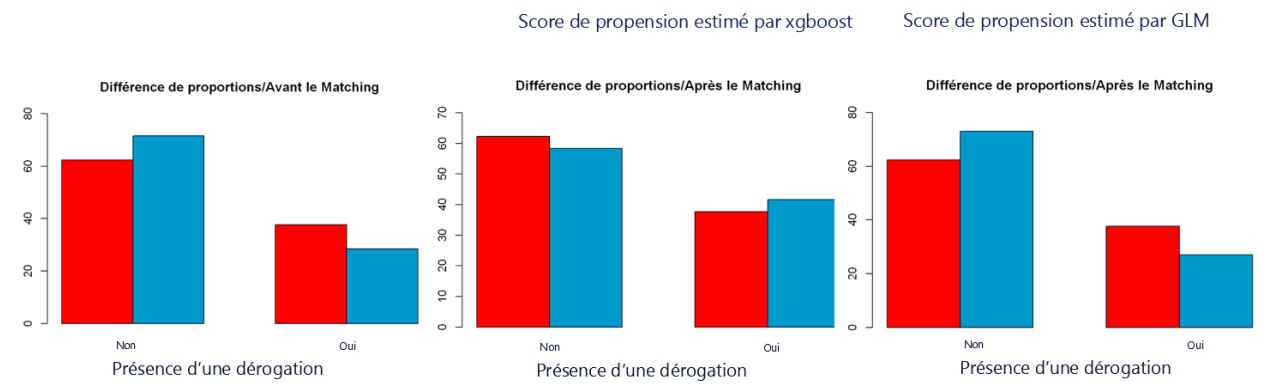


FIGURE 46 – La stratégie 3 est le traitement et la stratégie 4 est le contrôle. L'appariement avec introduction du XGBoost est parvenu à augmenter la proportion de sujets du groupe 4 n'ayant pas de dérogation.

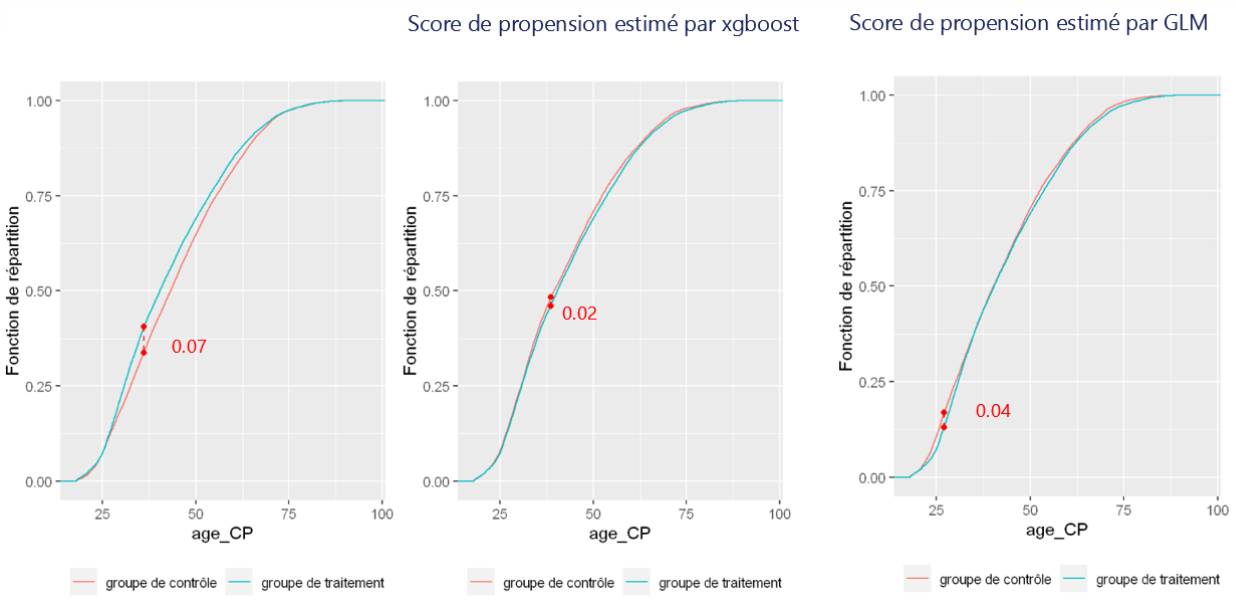


FIGURE 47 – La stratégie 4 est le traitement et la stratégie 5 est le contrôle. La similarité des deux groupes sur l'âge est plus forte après l'appariement.

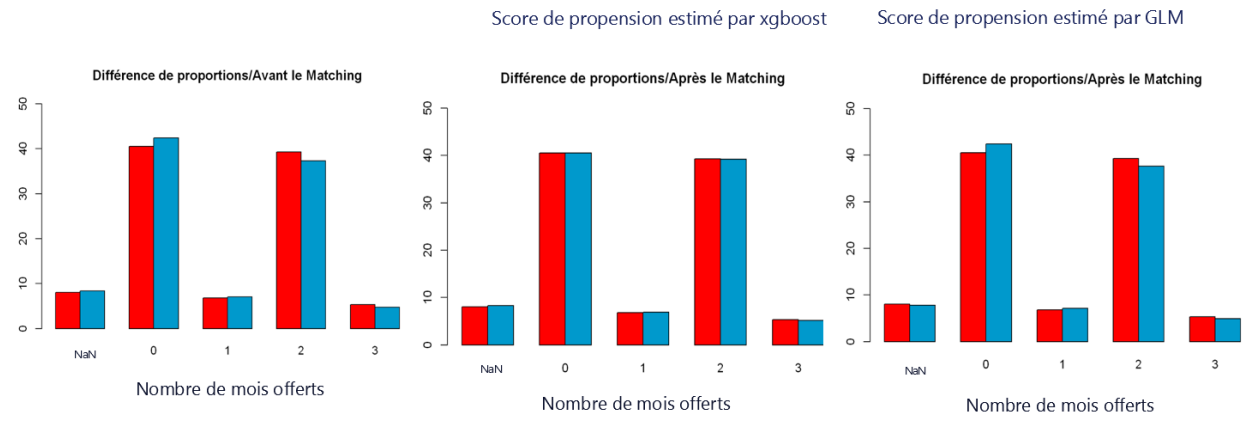


FIGURE 48 – La stratégie 4 est le traitement et la stratégie 5 est le contrôle. Le biais est atténué avec le XGBoost qui parvient à rendre les individus de traitement et de contrôle interchangeables sur cette variable (nombre de mois de réduction).

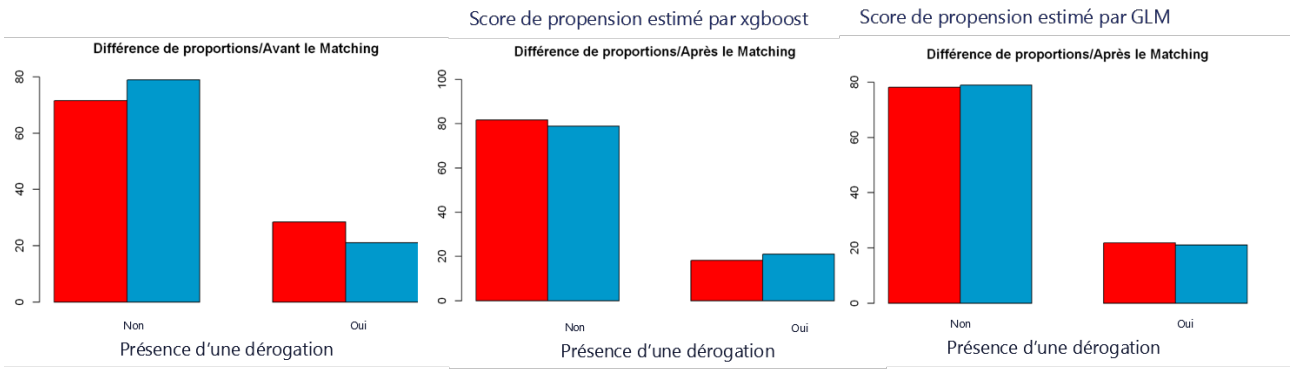


FIGURE 49 – La stratégie 5 est le traitement et la stratégie 4 est le contrôle. Le biais est corrigé avec les deux méthodes sur cette variable. La différence n'était pourtant pas significative entre les deux classes, ce qui prouve que le jumelage n'a pas uniquement permis de rapprocher les jumeaux sur les caractéristiques qui les distinguent mais a également stabilisé -voire amélioré- la similarité sur les points communs.

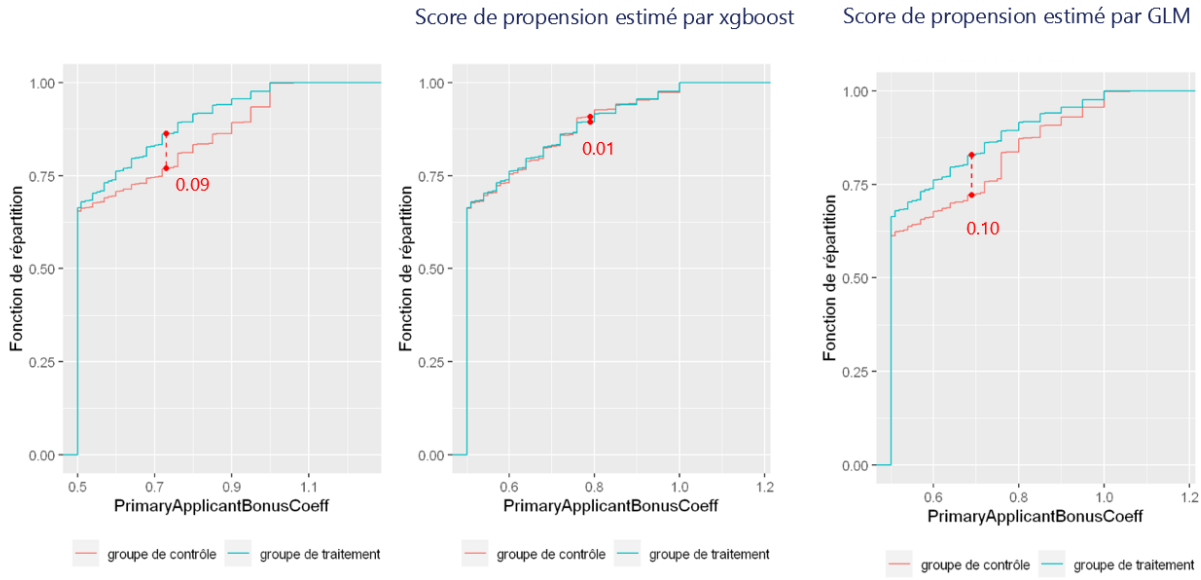


FIGURE 50 – La stratégie 5 est le traitement et la stratégie 7 est le contrôle. L'appariement avec la méthode du GLM dégrade la similarité entre les sujets sur la valeur du bonus/malus. En revanche, le XGBoost est très performant et permet d'obtenir deux courbes quasiment similaires.

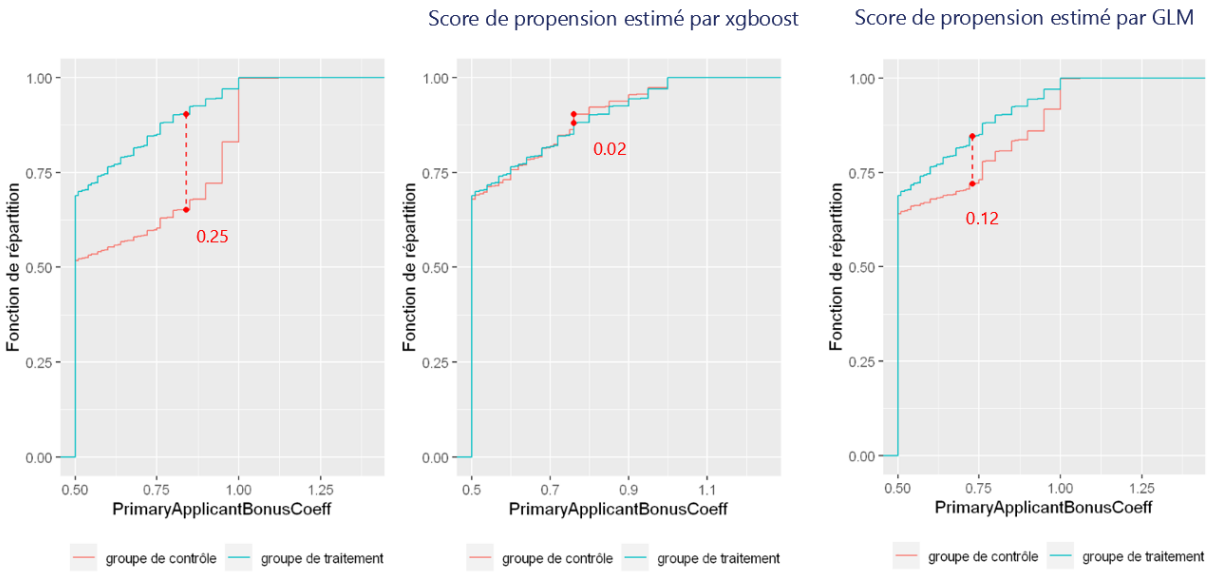


FIGURE 51 – La stratégie 6 est le traitement et la stratégie 8 est le contrôle. A l'origine, l'écart entre les deux groupes sur cette variable est plus important que celui observé pour les jumelages précédents. Le XGBoost obtient d'excellents résultats, compte tenu de l'écart de départ.

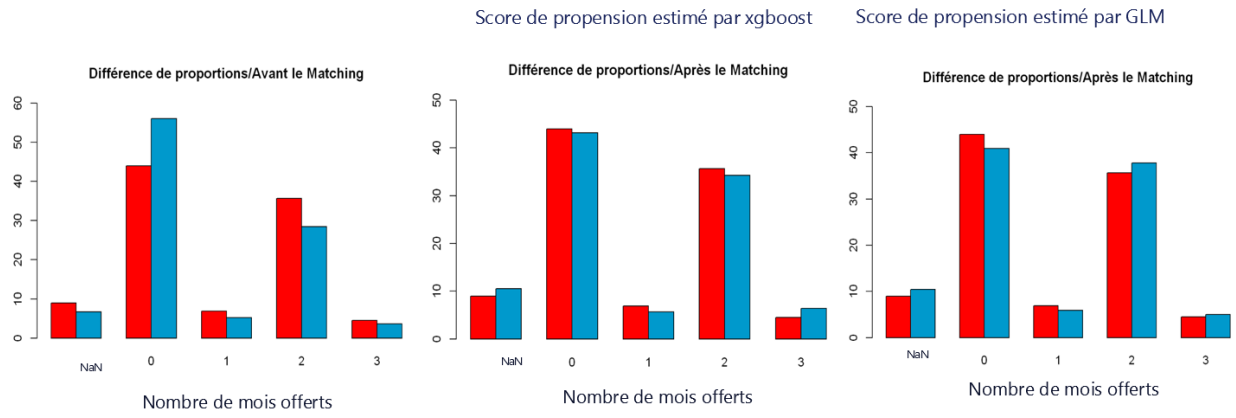


FIGURE 52 – La stratégie 6 est le traitement et la stratégie 8 est le contrôle. Les proportions des sujets du groupe 8 ayant 0 mois de réduction ont été augmentées au détriment de ceux ayant 2 mois de réduction, ce qui permet d'aligner la répartition avec celle du groupe 6.

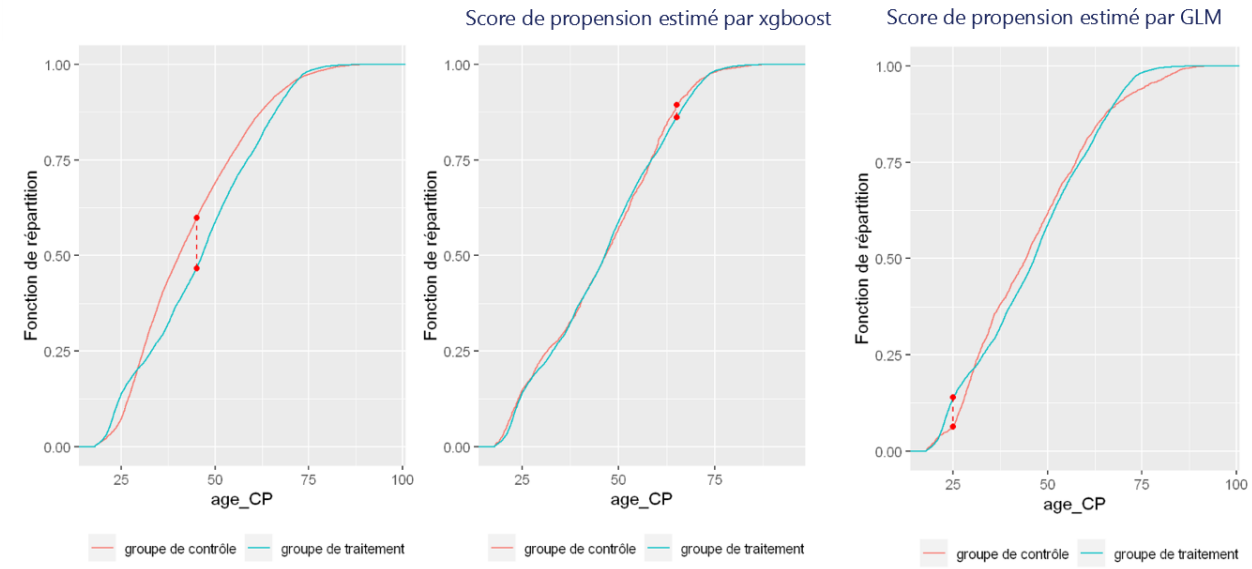


FIGURE 53 – La stratégie 7 est le traitement et la stratégie 4 est le contrôle. L'appariement est très précis avec le XGBoost, même si les deux stratégies ne sont pas "voisines". En revanche, le GLM ne corrige l'écart sur les âges jeunes et ne se concentre que sur le reste de la distribution.

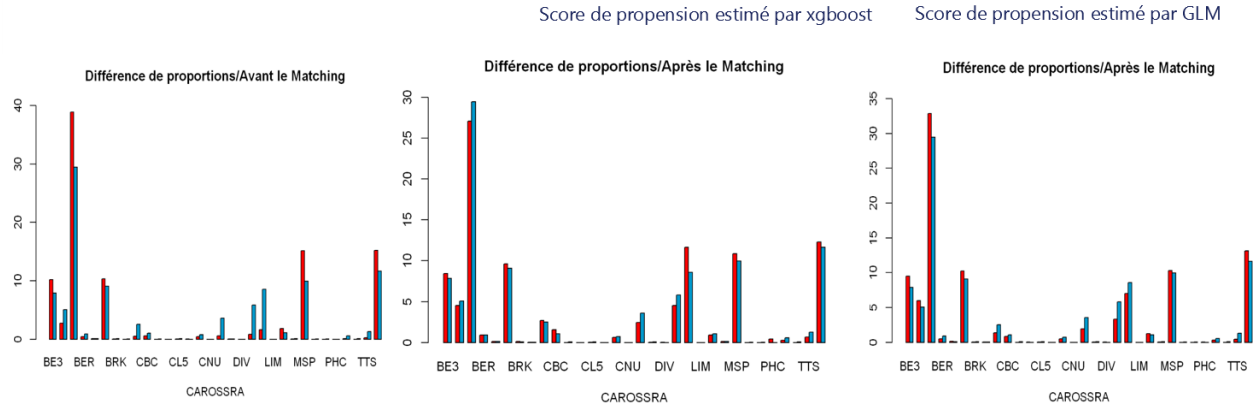


FIGURE 54 – La stratégie 7 est le traitement et la stratégie 4 est le contrôle. Le nombre d'individus de contrôle appariés ayant une berline ou un monospace a diminué par rapport à la population d'origine pour coller aux proportions du groupe de traitement.

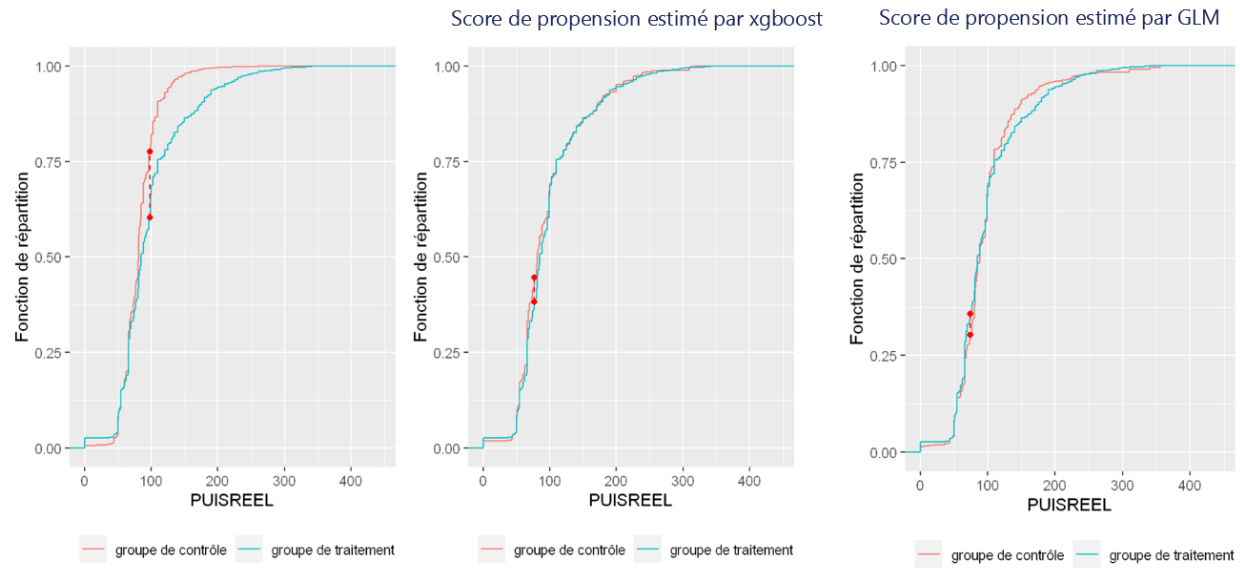


FIGURE 55 – La stratégie 8 est le traitement et la stratégie 6 est le contrôle. Le groupe 8 comporte davantage de véhicules puissants, ce phénomène est visible à l'écart entre les deux fonctions de répartition qui se situe au niveau des puissances moyennes (courbe plus pentue pour le groupe 6 qui converge vers 1 plus rapidement)

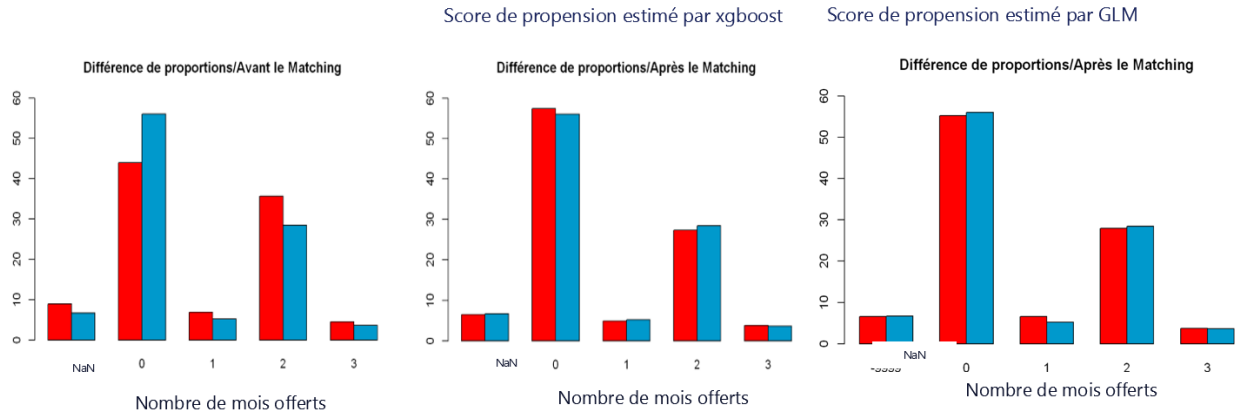


FIGURE 56 – La stratégie 8 est le traitement et la stratégie 6 est le contrôle.

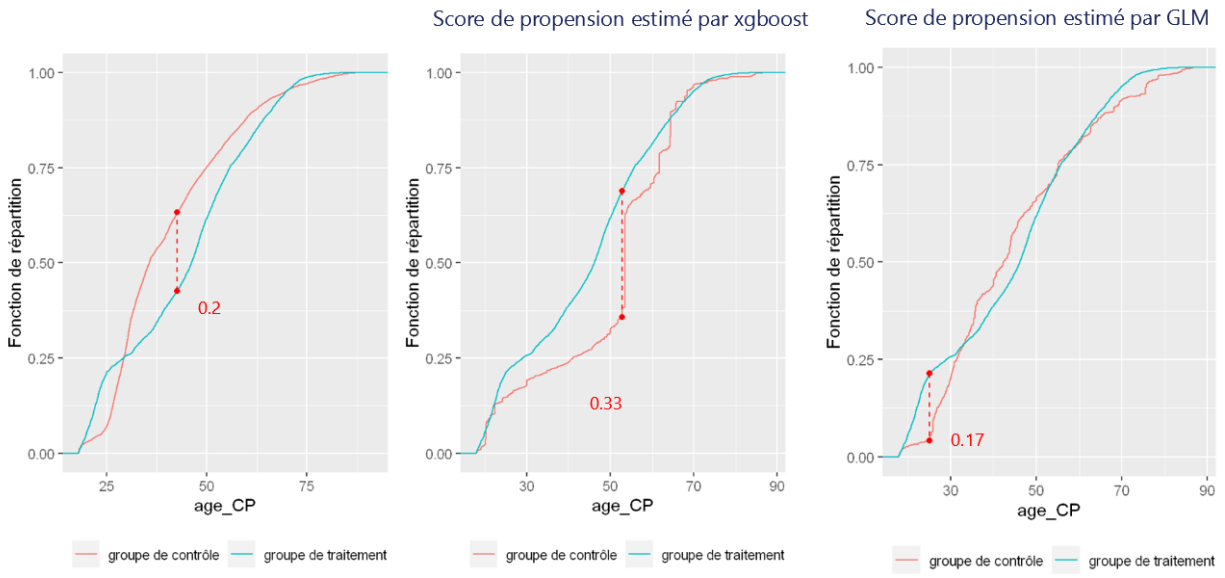


FIGURE 57 – La stratégie 8 est le traitement et la stratégie 1_2 est le contrôle. Il s'agit d'un contre-exemple permettant de démontrer l'instabilité de l'appariement de groupes trop éloignés en termes de politique tarifaire et donc de risque propre aux profils. Le groupe 8 est en moyenne plus âgé que le groupe 1_2. Le XGBoost corrige le premier écart au niveau des âges jeunes mais augmente aussi le nombre de sujets de 50 ans qui crée un saut conséquent. Le GLM est plus précis cette fois mais ne parvient pas à être fiable sur ce jumelage.

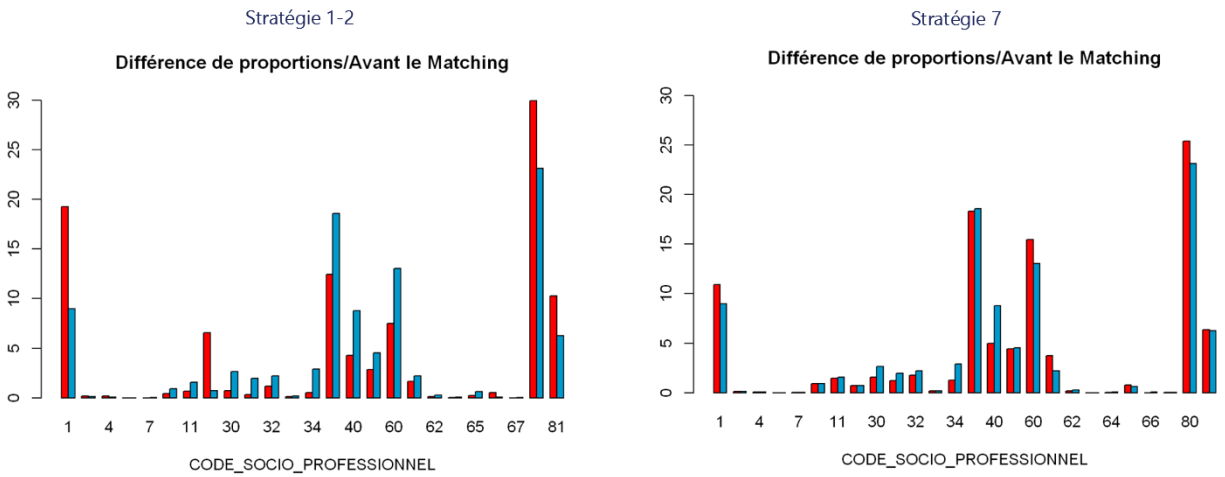


FIGURE 58 – Le groupe de traitement est la stratégie 8. Ce qui explique le mauvais appariement des stratégies éloignées provient de la nature des données : les différences de proportion sont plus faible pour des politiques tarifaires proches comme c'est illustré ici avec la catégorie socio-professionnelle.

4.1.6.3 Conclusion

Le travail d'analyse de la qualité de l'appariement a été réalisé sur l'ensemble des variables intervenant dans les modèles de score de propension pour les 42 jumelages (7 stratégies de traitement * 6 stratégies de contrôle). Parmi les deux techniques de jumelage -par score de propension avec une estimation XGBoost et GLM- l'apport du machine learning porte ses

fruits : lors de l'appariement des stratégies voisines (qui correspond à des variations tarifaires applicables sur le marché) le XGBoost est plus précis et régulier dans les résultats obtenus. Il parvient à rendre interchangeables les personnes de deux groupes à la fois sur les variables où l'écart était important tout en conservant la similarité sur les variables au support commun.

4.2 Méthodes alternatives au jumelage par score de propension

4.2.1 Mahalanobis

[16] Pour reprendre les notations de Cochran et Rubin (1973), la distance de Mahalanobis qui mesure la similarité entre deux individus est déterminée par :

$$md(X_i, X_j) = (X_i - X_j)' S^{-1} (X_i - X_j)^{1/2} \tag{38}$$

avec S la matrice de covariance de X composée des co-variables pour l'ensemble des sujets des deux groupes. La distance de Mahalanobis associe un poids à chaque facteur qui est égal à l'inverse de la covariance. Plus les variables sont corrélées ou possèdent une variance importante et moins elles ont de poids pour la mesure de distance. En effet, les composantes les plus dispersées admettent des écarts plus tolérés entre les deux sujets. En revanche, une variance faible exige une forte promiscuité entre les deux sujets concernant le facteur.

Pour estimer par jumelage avec remplacement l'effet d'appartenir à une stratégie par rapport à une autre pour les personnes appartenant à la dite stratégie, on apparie chacune d'entre elles avec la personne du groupe de contrôle la plus proche, selon la distance md().
Cette métrique est adaptée pour minimiser la distance entre les coordonnées de deux individus. Cependant, rapprocher deux individus selon une variable qui n'intervient pas nécessairement dans l'élaboration d'une politique tarifaire n'est pas judicieux.

4.2.2 Genetic matching

Combiner les méthodes du score de propension et de Mahalanobis peut s'avérer judicieux. Le premier sujet de traitement est sélectionné et lui sont associés les individus de contrôle dont l'écart de score de propension est en-dessous d'un certain seuil. Le choix de l'appariement final se fait ensuite avec la distance de Mahalanobis. La distance de Mahalanobis généralisée est définit par :

$$d(X_i, X_j) = (X_i - X_j)' S^{-1/2\prime} W S^{-1/2} (X_i - X_j)^{1/2} \tag{39}$$

où W est une matrice p*p définie positive composée de poids et $S^{-1/2}$ est issue de la décomposition Choleski de la matrice de variance/co-variance de X.
Le *genetic matching*[17] (réf. (7))allie les méthodes de jumelage sous le critère de la distance de Mahalanobis et jumelage par score de propension. Si W est égale à la matrice identité, alors la distance de Mahalanobis est retrouvée. Si W accorde des poids tels que $d(X_i, X_j)$ détermine l'écart de score de propension entre i et j alors la méthode par score de propension est retrouvée.

16. Mahalanobis, *On the generalised distance in statistics, Proceedings of the National Institute of Sciences (Calcutta)* , 1936

17. Diamond et Sekhon, *Genetic Matching for Estimating Causal Effects : A General Multivariate Matching Method for Achieving Balance in Observational Studies, Review of Economics and Statistics*, 2012

C'est notamment le cas si le score de propension est introduit comme co-variables au sein de X et que les composantes de W sont nulles pour les autres facteurs.

A chaque itération, les co-variables sont pondérées de manière différente. Cette technique garantie la convergence vers l'appariement optimal, l'objectif étant de pondérer les facteurs pour aboutir au meilleur appariement. Le critère à optimiser peut être la statistique de test de Kolmogorov-Smirnov, de la t de Student ou la comparaison de quantiles ($QQPlot$) sur chacune des distributions des co-variables entre celles du groupe de traitement et celles du groupe de contrôle après appariements. Selon le critère opté, on aboutit à un jumelage différent.

Test de Student sur échantillons appariés

H0 : Les deux séries ont la même moyenne/variance.
H1 : Les deux séries n'ont pas la même moyenne/variance.
Ce test est valide si les co-variables sont continues.
Analyse des quantiles

H0 : Les deux séries ont la même distribution. $\iff$ Les jumeaux sont interchangeables selon le facteur.
H1 : Les deux séries n'ont pas la même distribution. $\iff$ Les jumeaux ne sont pas interchangeables selon le facteur.
Ce test est valide si les co-variables sont continues.

Test de Kolmogorov-Smirnov

H0 : Les deux séries ont la même distribution. $\iff$ Les jumeaux sont interchangeables selon le facteur.
H1 : Les deux séries n'ont pas la même distribution. $\iff$ Les jumeaux ne sont pas interchangeables selon le facteur : comparer leur taux de conversion n'a donc pas de sens.
Soient F la fonction de répartition de la série du facteur du groupe de traitement et G la fonction de répartition de la série de ce même facteur des individus du groupe de contrôle ayant été sélectionnés, la statistique de Kolmogorov-Smirnov s'écrit :

$$|| F - G ||_\infty$$

L'objectif est d'obtenir une p-valeur la plus élevée possible pour avoir assez d'évidence statistique pour ne pas rejeter l'hypothèse nulle. Cela revient à souhaiter une valeur la plus faible possible de la statistique de test. A chaque itération, si l'algorithme va maximiser la p-valeur minimale entre toutes les co-variables.

Le genetic matching minimise alors la dispersion maximale peu importe la distribution de X, et donc, même si la propriété EPBR n'est pas vérifiée.

4.2.3 Difficultés opérationnelles

Même si le genetic matching permet d'obtenir de meilleurs résultats, le gain de précision obtenu ne supplante pas le coût algorithmique généré par l'optimisation de la plus petite p-valeur. Il est toujours possible de réduire le nombre d'observations impliquées dans l'optimisation à travers le paramètre *pop.size* de la fonction *GenMatch* (RStudio). Cependant, les théorèmes qui prouvent l'existence de bons résultats de cet algorithme ne sont prouvés que pour des échantillons de population choisis suffisamment grand [18] (réf. (15)). Il existe donc un arbitrage entre temps de calcul et précision. En revanche, même en contrôlant la taille de la population, le coût est tel que le genetic matching ne peut être une solution envisagée au regard des bonnes performances du jumelage par score de propension estimé par des techniques de machine learning.

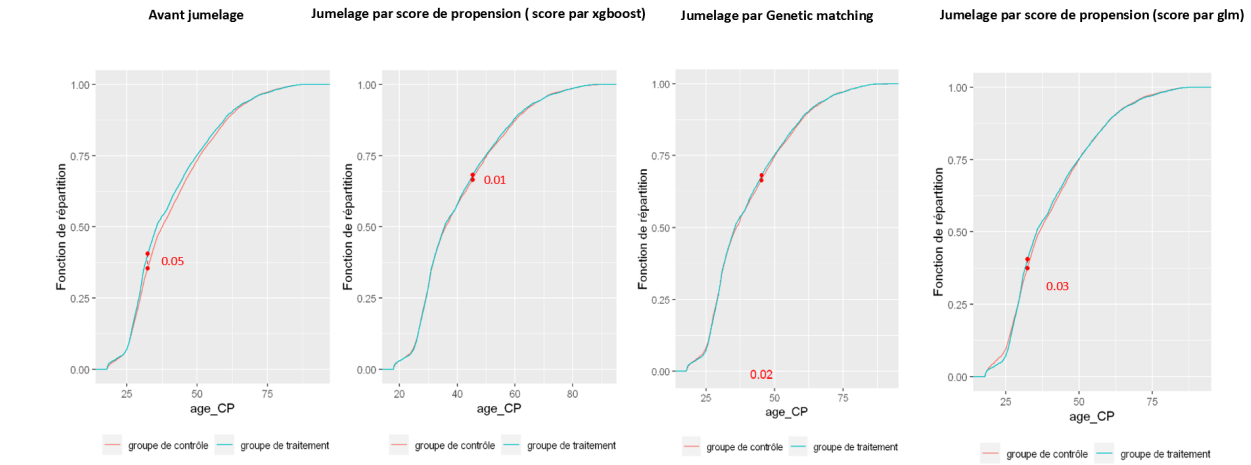


FIGURE 59 – Comparaison des différentes méthodes sur l'âge du conducteur avec le groupe de traitement étant la stratégie 1-2 et le groupe de contrôle étant la stratégie 3. Le *genetic matching* a été réalisé avec une taille de population de 2000 sujets et a nécessité de nombreuses heures de calcul uniquement sur ces deux groupes.

18. Nix et Vose,*Modeling Genetic Algorithms with Markov Chains, Annals of Mathematics and Artificial Intelligence*, vol. 5, p. 79–88, 1992

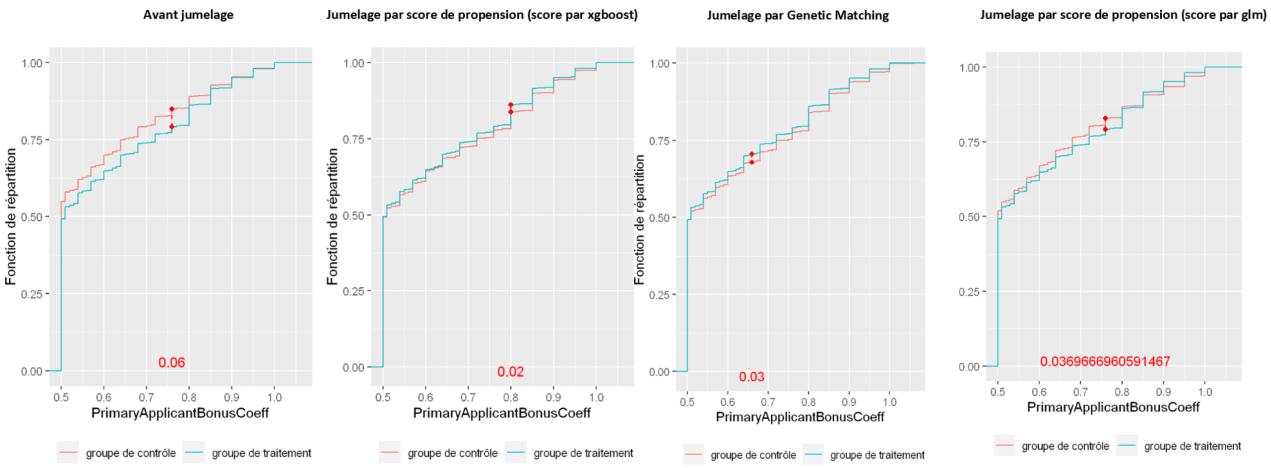


FIGURE 60 – Comparaison des différentes méthodes sur le bonus/malus avec le groupe de traitement étant la stratégie 1 et le groupe de contrôle étant la stratégie 3. Le *genetic matching* a été réalisé avec une taille de population de 2000 sujets et a nécessité de nombreuses heures de calcul uniquement sur ces deux groupes.

Le nombre de sujets impliqués dans l'optimisation ne permet pas de concurrencer le jumelage par score de propension estimé par XGBoost. Il est essentiel d'opter pour une méthode calibrée sur les données (volume de variables, nature des variables, volume d'observations), avec des jumelages de bonne qualité et qui soit réalisable dans les temps impartis. L'appariement via les scores de propension estimés par XGBoost remplit ces critères et permet d'accroître la connaissance du comportement client à la souscription en alimentant la base de devis d'origine de sujets connus avec de nouveaux tarifs proposés, sans avoir recours à des A/B tests. Cependant la confiance en cette simulation de price test ne sera possible que dans la limite d'une certaine évolution de la marge.

5 Modélisation de l'élasticité au prix et optimisation tarifaire

5.1 Contraintes sur le modèle d'élasticité au prix

5.1.1 Restriction sur les nouvelles marges à tester

Un premier modèle a été créé par XGBoost pour tester les performances qu'il est possible d'obtenir sur la base de test (30% de l'échantillon global). L'erreur s'est accrue de manière significative par rapport aux modèles de taux de transformation élaborés avant le jumelage. C'est dû au fait que des devis similaires en termes de caractéristiques de conducteur, de véhicule, d'options aient des taux de transformation différents uniquement du fait d'une variation du tarif TTC. Même si ce dernier peut avoir une plus grande importance dans le modèle, il ne peut à lui seul expliquer la conversion qui dépend des autres facteurs. De ce fait, l'erreur de prédiction s'accroît avec l'inclusion de devis dont l'écart de la nouvelle marge avec la marge d'origine augmente. En d'autres termes, pour un devis créé par jumelage, plus on s'éloigne de la marge d'origine et plus l'erreur de prédiction peut potentiellement augmenter.

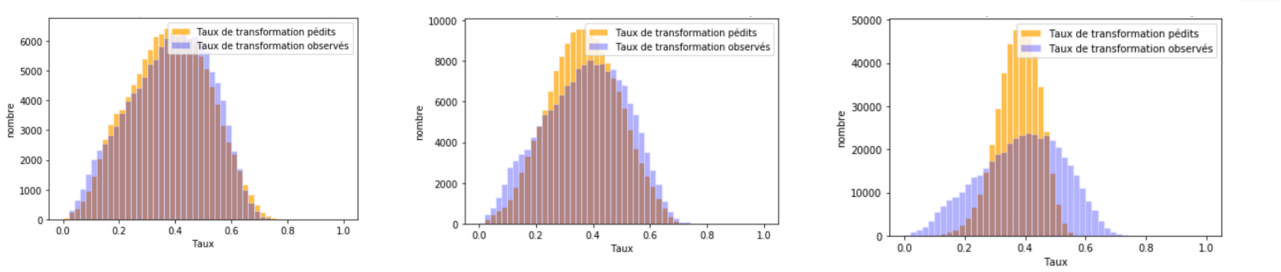


FIGURE 61 – Evaluation des erreurs de prédiction dans le cas où, de gauche à droite, la base devis n'a pas été augmentée de jumeaux, la base devis ne contient que des jumeaux n'excédant pas une évolution de plus de 20% de la marge, la base devis avec tous les jumeaux sans aucune restriction sur les nouveaux tarifs appliqués.

Après l'étude de différents seuils pour l'évolution possible de la marge entre un prospect et son jumeau, le seuil de 20% a été retenu.

— Au-delà de ce seuil, le modèle est peu fiable compte tenu de l'erreur générée.

— En-dessous de ce seuil, le gain en termes de performance n'est pas assez important au regard de l'apport de données supplémentaires.

— Une variation de 20% de la marge est suffisamment conséquente si l'assureur souhaite établir une politique tarifaire cohérente par rapport à celle d'origine.

Il s'agit ici d'une limite à la méthode du jumelage par score de propension. Le price test peut être simulé que dans un rayon autour de la politique tarifaire d'origine, la fiabilité de l'algorithme étant remise en question pour des appariements de sujets trop éloignés en termes de politique tarifaire.

In fine, la base de devis a vu son volume augmenter de 22% par rapport aux données transmises par le client après ajout des jumeaux ayant une marge n'excédant pas 20% d'évolution par rapport à la marge préalablement appliquée sur ce prospect.

5.1.2 Equation d'optimisation et critères sur le modèle à opter

Avec la modélisation de l'élasticité au prix, l'assureur peut simuler les principaux indicateurs financiers pour une nouvelle politique tarifaire, si elle n'excède pas une variation de 20%. Maintenant, il est également possible de chercher à maximiser la formule du profit (ou du chiffre d'affaires, ou du loss ratio) en cherchant l'allocation de prix optimale.

Le problème d'optimisation nécessite la modélisation de l'élasticité-prix qui devient donc la première étape de résolution. L'élasticité au prix doit être obtenue de la manière la plus précise pour plusieurs valeurs de prix potentiellement applicables par l'assureur pour l'assuré. C'est pour cette raison que la base d'apprentissage a été augmentée par des profils se trouvant originellement dans la base mais à qui d'autres tarifs ont été appliqués pour obtenir un nouveau score de transformation par *propensity score matching* adapté à la nouvelle politique tarifaire. La série de ces taux de transformation va être modélisée pour terminer la première étape de résolution de l'équation tarifaire qui pourra être réalisée par le client :

$$Profit = \sum_{i=1}^{n} (P_i - S_i) * \hat{f}(P_i, X_i) \tag{40}$$

$$\text{Sous contrainte d'un certain volume : } \sum_{i=1}^{n} \hat{f}(P_i, X_i) > \theta$$

- P_i : la variable endogène qui correspond au tarif Hors Taxes payé par l'assuré i
- S_i : correspond à tous les frais et la charge sinistre de l'assuré i
- $\hat{f}(P_i, X_i)$: la probabilité de conversion estimée de l'assuré i
- X_i : les caractéristiques de l'assuré i (âge, profession, bonus/malus...), les caractéristiques du contrat de i (présence de franchise, d'offres commerciales, de dérogation, marge appliquée...), les caractéristiques du marché pour le profil i (prix du concurrent, écart avec la médiane du marché...)

Pour trouver le tarif Hors Taxes optimal de chaque prospect i $\in \{1, ...n\}$ la fonction du taux de transformation doit pouvoir s'écrire en fonction de ce tarif. En d'autres termes, les propriétés de modularité et de simulabilité d'un modèle *glassbox* doivent être respectées :

- Simulabilité : à partir des résultats obtenus par le modèle, l'humain doit pouvoir retrouver la prédiction.
- Modularité : une portion significative du processus peut être interprétée indépendemment.

La modélisation ne peut donc pas être réalisée par un modèle comme le XGBoost dont l'intervention du tarif dans la prédiction ne peut pas être exprimée explicitement. Dans la section 2 ont été comparés le XGBoost et l'explainable boosting machine, ce dernier répondant aux critères précédents et s'approchant des performances du XGBoost. La comparaison n'aurait pas

pu être faite ici car étant donné que le tarif doit être exprimée de manière la plus concise possible, une contrainte sur le nombre de segmentation des variables doit être posée. Le XGBoost et l'EBM n'aurait donc être comparé équitablement. La modélisation porte sur le logit du taux de transformation afin de poser une loi normale sur la distribution d'erreurs à chaque itération du boosting de l'EBM, et donc :

$$\hat{y_i} = \hat{logit}(\pi_i) = \sum_{j=1}^{p}\sum_{h=1}^{H}\beta_{j,h}X_i^{j,h} = \sum_{j=1\ sauf\ j_0}^{p}\sum_{h=1}^{H}\hat{\beta}_{j,h}X_i^{j,h} + \sum_{h=1}^{H}\hat{\beta}_{j_0,h}P_i^h \tag{41}$$

avec pour le sujet i, j_0 l'indice de la variable du prix HT, H le nombre de fenêtres (découpages) maximal autorisé pour chacune des variables, p le nombre de variables.

$$\hat{f}(X_i, P_i) = \frac{\exp \hat{y_i}}{1 + \exp \hat{y_i}} = \frac{\exp\left(\hat{y_i} - \sum_{h=1}^{H}\hat{\beta}_{j_0,h}P_i^h + \sum_{h=1}^{H}\hat{\beta}_{j_0,h}P_i^h\right)}{1 + \exp\left(\hat{y_i} - \sum_{h=1}^{H}\hat{\beta}_{j_0,h}P_i^h + \sum_{h=1}^{H}\hat{\beta}_{j_0,h}P_i^h\right)} \tag{42}$$

avec $\hat{y}$ et $\hat{\beta}$ déterminés par le modèle. L'équation d'optimisation s'écrit alors :

$$Profit(P) = \sum_{i=1}^{N}(P_i - S_i) * \frac{\exp\left(\hat{y} - \sum_{h=1}^{H}\hat{\beta}_{j_0,h}P_i^h + \sum_{h=1}^{H}\hat{\beta}_{j_0,h}P_i^h\right)}{1 + \exp\left(\hat{y} - \sum_{h=1}^{H}\hat{\beta}_{j_0,h}P_i^h + \sum_{h=1}^{H}\hat{\beta}_{j_0,h}P_i^h\right)} \tag{43}$$

$$\text{s.c. } \sum_{i=1}^{N}\pi_i >= \theta$$

avec N le nombre de prospects.

Il s'agit donc de déterminer les coefficients reliés à la variable de prix HT. Sont impliquées les co-variables :

— marge

— tarif TTC

— écart avec le tarif leader

qui doivent donc subir une transformation afin d'isoler le tarif HT dans l'équation.

5.2 Modélisation du score de conversion global

5.2.1 Modélisation par EBM

Le logit du taux de transformation a été estimé par EBM en contrôlant la largeur des fenêtres qui scindent la variable par tranche. Le paramètre correspondant permet d'obtenir des courbes plus lisses plus facilement exploitables.
L'avantage des EBM découle de cette segmentation automatique qui octroie une meilleure précision dans la prédiction. Aussi, les scores marginaux de chacun des segments permet d'établir des stratégies tarifaires. Ainsi, le score diminue avec l'augmentation de l'écart du tarif TTC proposé au tarif leader du marché. Ce score franchit le seuil de 0 à partir du seuil de 175 euros d'écart. A l'inverse, la conversion s'améliore avec l'âge avec un seuil à 32 ans. La conversion s'affaiblit sur la fin d'année tandis que les périodes à forte conversion coïncident avec la mise en place de réductions tarifaires et d'offres promotionnelles. Le modèle construit est un point d'appui pour les décisions de court terme relatives à l'activité commerciale de la compagnie.

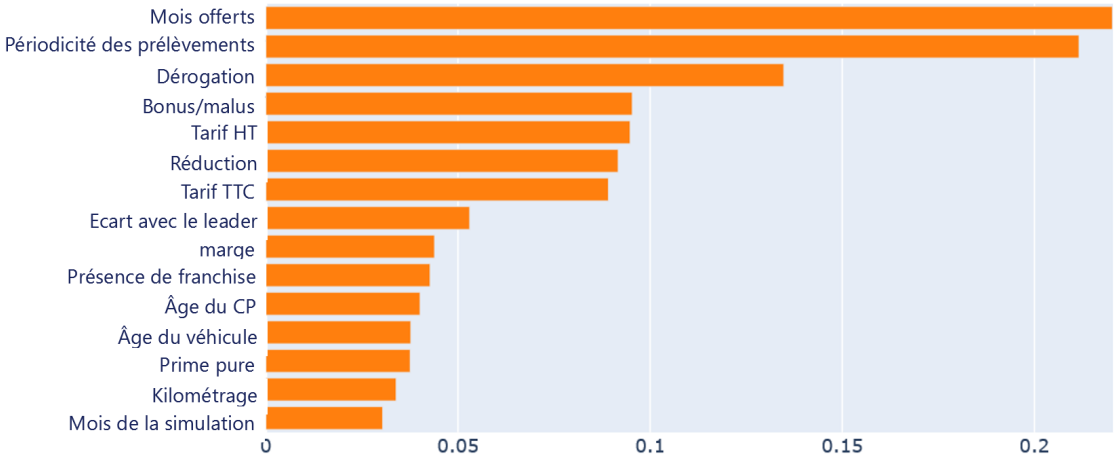


FIGURE 62 – Ordre d'importance des variables

FIGURE 63 – Score d'importance moyen obtenu pour chaque stratégie

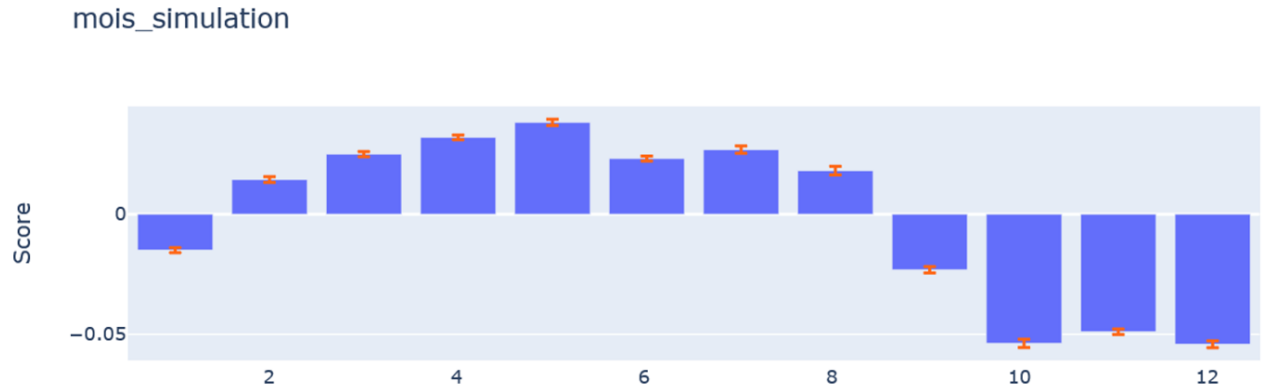


FIGURE 64 – Score d'importance moyen obtenu pour chaque mois de souscription

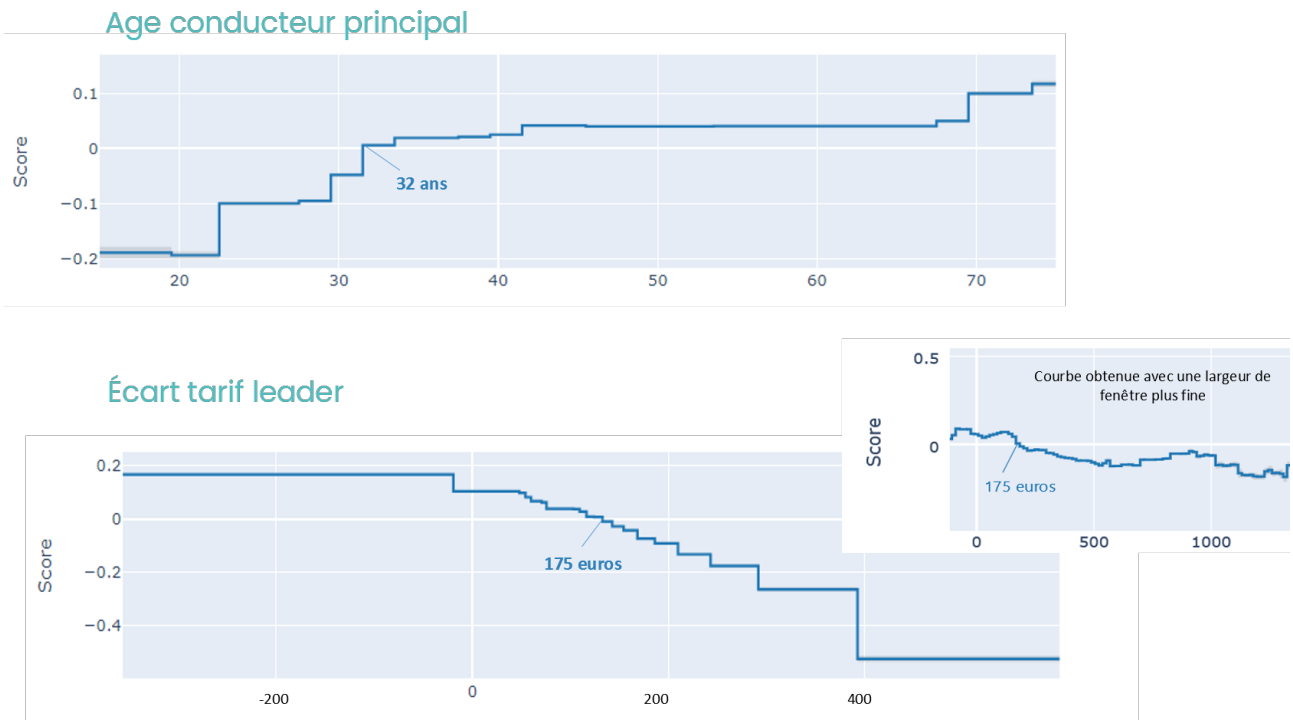


FIGURE 65 – Score d'importance en fonction de l'âge et de l'écart avec le tarif leader

FIGURE 66 – Score d'importance moyen obtenu pour chaque classe de prix et code kilométrique (du plus petit rouleur au plus grand)

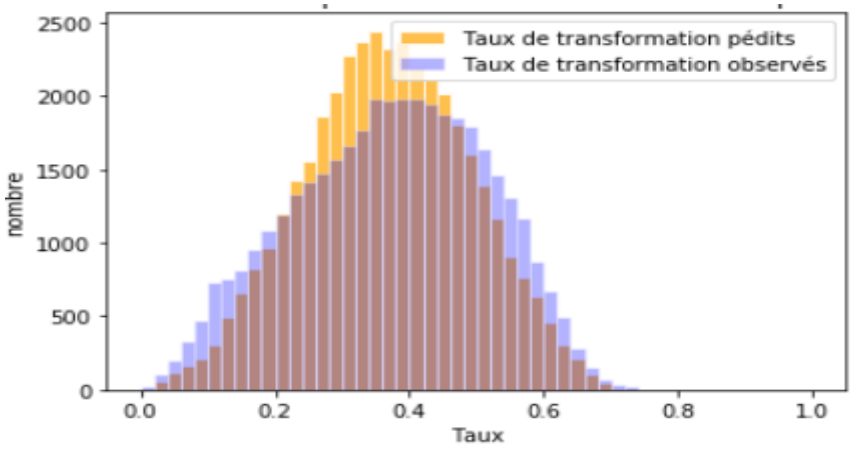


FIGURE 67 – Distribution des taux sur la base de test. Le MAE est de 0.05

Néanmoins l'EBM pose des difficultés techniques pour la réalisation de la résolution de l'équation d'optimisation tarifaire, qui est l'application majeure des travaux réalisés pour la mesure de l'élasticité au prix individuelle. Trouver l'allocation de prix optimaux, qui est la solution du lagrangien sous contraintes, impose une hypothèse de dérivabilité de l'équation par rapport au prix. La morphologie des courbes marginales impliquant le prix (marge, écart avec le leader etc) suggère des paliers selon le découpage de la variable. Une méthode pour passer outre ce problème serait de sélectionner plusieurs points et d'effectuer une moyenne des scores obtenus sur chacun d'entre eux.

5.2.2 Modélisation par GLM

Le GLM estime une pente moyenne pour chacune des variables de format continue. Un Lasso a été employé pour s'acquitter du problème de sur-apprentissage. Les résultats sont moins précis puisque le score n'est pas adapté à chaque profil - ce dont témoigne un MAE dégradé de 0.1-, mais l'optimisation est plus aisé à réaliser par la suite. En effet, les variables incluant le tarif HT ayant été laissées continues, leur dérivée est semblable sur l'ensemble des allocations de prix.

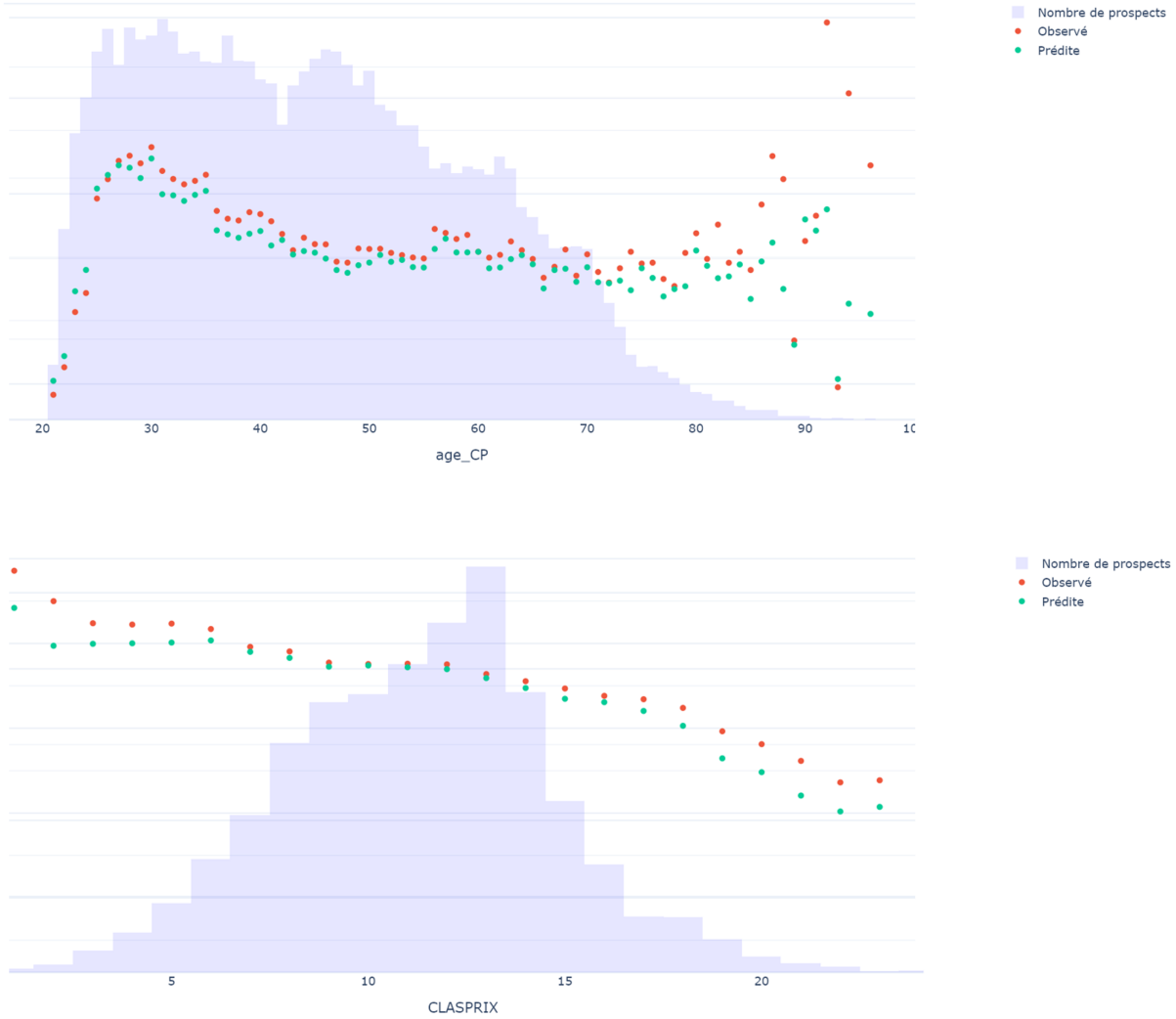


FIGURE 68 – Comparaison des taux observés et prédits sur la base de test pour l'âge et la classe de prix du véhicule

5.2.3 Courbes d'élasticité : cas de la catégorie socio-professionnelle

A partir des modèles GLM construits, une courbe des taux peut être tracée pour chaque profil à partir d'une liste de tarif à considérer. En particulier, il est intéressant de comparer la courbe des taux de transformation d'un segment de risque à l'autre. Il s'agit ici d'une première application des travaux pour la modélisation de l'élasticité au prix. L'ajout de données par jumelage par score de propension permet d'agrandir la fenêtre de tarifs qu'on peut proposer à un profil de client tout en ayant sa probabilité de conversion sur l'ensemble de cette fenêtre.

Cependant, pour des soucis de cohérence et de fiabilité dans les résultats de l'algorithme de jumelage, seul 20% de variation autour de la marge d'origine a été permis, pour ne pas affecter à un sujet un jumeau qui ne lui ressemble pas sous prétexte qu'aucun autre individu proche n'ait un tarif si éloigné. Il demeure donc certaines zones d'ombre dans la fenêtre de tarifs que l'on peut proposer à l'avenir, ce qui ne soulève pas de problématique puisqu'un changement brusque de tarif n'est pas envisageable lors de la mise en place d'une nouvelle politique tarifaire. Il convient simplement de signaler les délimitations de la fenêtre des tarifs envisageables dans

les exemples présentés.

A la demande de l'assureur, le cas de la catégorie socio-professionnelle est exposé ici pour laquelle ont été fixés les autres critères tarifaires :

— l'âge du conducteur se situe entre 30 et 45 ans ;

— le conducteur est marié ;

— le conducteur a un bonus de 0.5 ;

— le conducteur roule plus de 12000 km par an ;

— le véhicule est utilisé pour des trajets privés ou des trajets domicile/travail ;

— le véhicule est de classe SRA inférieure à M.

Les critères tarifaires principaux étant similaires, il s'agit maintenant d'évaluer toutes choses similaires par ailleurs, l'impact de chacune des CSP parmi celles des étudiants, retraités, cadres et des salariés, sur le taux de transformation à la souscription.

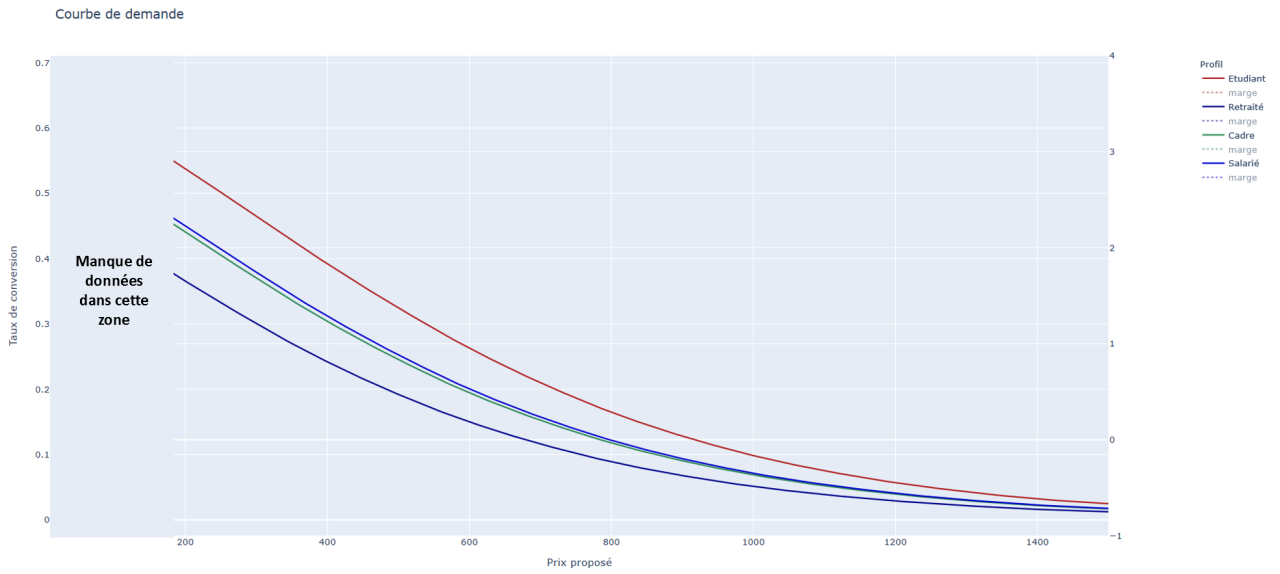
Comme énoncé plus haut, il convient de souligner que les courbes de demande (et donc le calcul d'élasticité) ne peuvent être analysées sur les extrêmes : en-deça de 170 euros environ et au-dessus de 1400 euros (borne restrictive incluant les 4 segments) par manque de données. En particulier, les conditions suivantes ne sont pas respectées :

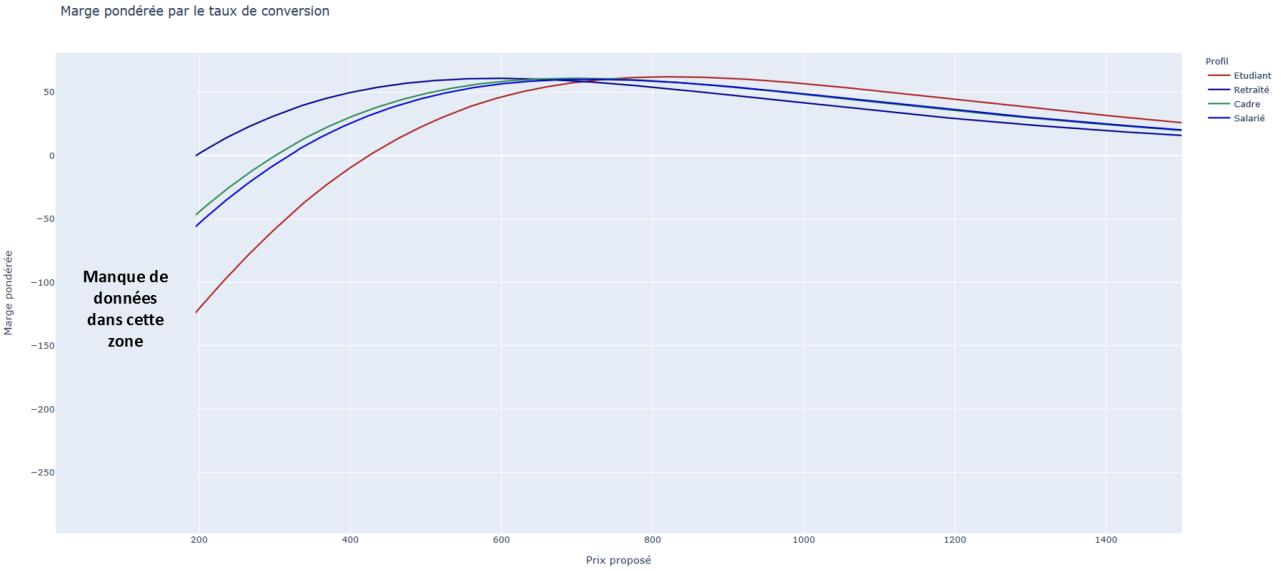
$$\lim_{P \to \infty} f(X, P) = 0$$

$$\lim_{P \to 0} f(X, P) = 1$$

La première convergence est plus observée par nos données : le coefficient lié au prix étant positif, tous les autres paramètres sont neutralisés lors du passage à la limite. Seule la vitesse de la convergence (à partir de quel tarif, le seuil ϵ est atteint) n'est pas garanti au-dessus de 1400 euros. En revanche, la limite finie en 0 n'est pas vérifiée puisqu'il n'existe aucun prospect donc le tarif a été mis à 0 ou proche de 0 et que les autres paramètres influent toujours sur la conversion.

Les bornes de la fenêtre ont été fixées en établissant pour chaque groupe de profession les prix les plus bas et les plus élevés intégrés dans le modèle d'apprentissage. A gauche, le maximum entre les groupes est choisi pour délimitation.





Les étudiants suivis des salariés, des cadres puis des retraités ont la courbe de sensibilité la plus haute, c'est-à dire que pour toute allocation de prix donnée, les quatre taux de conversion moyens sont décroissants dans cet ordre. Les courbes convergent ensemble lorsque le prix augmente. Les taux sont bornés entre 0.02 et 0.55 (maximum atteint par les étudiants).

Tout d'abord, le comportement des quatre courbes est similaire : plus le tarif augmente et plus la conversion s'amoindrit. En revanche, les courbes de sensibilité ont une dérivée seconde négative puisque la décroissance s'atténue au fur et à mesure que le prix augmente. Cela signifie qu'augmenter le prix lorsqu'il est bas a plus d'impact en moyenne qu'augmenter un prix déjà élevé (le pourcentage relatif est plus faible). Il n'est donc pas pertinent stratégiquement d'opter pour une diminution d'un tarif au-dessus d'un certain seuil (800-1000 euros selon le segment), à moins de pouvoir le diminuer suffisamment pour atteindre le volume de portefeuille cible. De la même manière, accroître le chiffre d'affaires sur les plus petits tarifs (en-dessous de 400 euros HT) entraîne une perte de volume importante. L'arbitrage entre les deux peut se déterminer à travers la marge pondérée :

$$marge\ pondérée_c = \frac{1}{N(c)} \sum_{i=1}^{N(c)} (\frac{tarif HT_i}{(1+frais)} - prime\ pure_i) * (Prédiction\ de\ la\ conversion_i);$$

avec N(c) le nombre de prospect sur la catégorie socio-professionnelle c.

L'activité devient rentable à partir d'un certain seuil de tarif. Les salariés étant moins risqués, ils deviennent rentables à partir d'un prix de 200 euros contre 420 euros pour les étudiants. Ensuite, augmenter le tarif devient judicieux car la marge moyenne s'accroît malgré une perte dans la transformation. C'est à partir de 600 euros pour les salariés - 800 euros pour les étudiants- que le gain en chiffre d'affaires ne supplante pas la perte liée à une plus faible transformation.

5.3 Equation d'optimisation

Les travaux issus de ce mémoire remplissent le cahier de charge pour la première étape de l'optimisation tarifaire à la souscription. L'élasticité-prix modélisée, il s'agit maintenant de dresser un aperçu des prochaines étapes à effectuer pour la résolution de l'équation.
De par les discussions précédentes, des modifications dans l'écriture du problème d'optimisation sont à apporter au regard des différentes conjectures énoncées :

— La modélisation de la probabilité de transformation par GLM satisfait la contrainte de dérivabilité sur l'ensemble du domaine de définition des prix. On alors le nombre de découpages de la variable continue de tarif HT H =1 et $\sum_{h=1}^{H} \hat{\beta_{j_0,h}} * P_i^h = \beta_{j_0} * P_i$. De plus, $\hat{y} - \sum_{h=1}^{H} \hat{\beta}_{j_0,h} P_i^h$ peut s'écrire comme une constante $\hat{c}$.

— Chaque prospect a été jumelé à un autre sujet sur des variations de marge n'excédant pas +- 20%. L'élasticité-prix modélisé par GLM a été construit à partir de ces variations. Une contrainte sur le prix de chaque prospect est donc primordial pour s'assurer que le taux de transformation utilisé dans l'optimisation soit fiable.

et donc :

$$Profit(P) = \sum_{i=1}^{N} (P_i - S_i) * \frac{\exp(\hat{y} - \sum_{h=1}^{H} \hat{\beta}_{j_0,h} P_i^h + \sum_{h=1}^{H} \hat{\beta}_{j_0,h} P_i^h)}{1 + \exp(\hat{y} - \sum_{h=1}^{H} \hat{\beta}_{j_0,h} P_i^h + \sum_{h=1}^{H} \hat{\beta}_{j_0,h} P_i^h)} \tag{44}$$

$$= \sum_{i=1}^{N} (P_i - S_i) * \frac{\exp(\hat{c} + \beta_{j_0} * P_i)}{1 + \exp(\hat{c} + \beta_{j_0} * P_i)} \tag{45}$$

$$s.c. \sum_{i=1}^{N} \pi_i >= \theta \tag{46}$$

$$P_i \in [P_{i,0} - (P_{i,0} - Pr_{i,0}^{HT}) * 20\%); P_{i,0} + (P_{i,0} - Pr_{i,0}^{HT}) * 20\%)] \tag{47}$$

avec N le nombre de prospects, $P_{i,0}$ le tarif HT originellement appliqué au prospect i dans la base devis d'entrée, $Pr_{i,0}^{HT}$ la prime pure HT du prospect i (prime pure + frais), θ un seuil minimal de volume de clients déterminé par l'assureur.

Cette deuxième contrainte serait considérée en remplaçant la descente de gradient classique en descente de gradient projetée [19] (réf. (2)). Le lagrangien s'écrit alors :

$\mathrm{L}(\mathrm{P}_1, ..., P_n, \lambda) == \sum_{i=1}^{N}(P_i - S_i) * \frac{\exp(\hat{c}+\beta_{j_0} * P_i)}{1+\exp(\hat{c}+\beta_{j_0} * P_i)} + \lambda(\sum_{i=1}^{N} \pi_i - \theta)$ avec $\lambda >= 0$.

A chaque itération t, si la deuxième contrainte n'était pas exigée, la descente de gradient s'écrit comme :

$$P_{t+1} = P_t - \nabla L(P_{1,t}, ..., P_{n,t}, \lambda) * \delta \tag{48}$$

et la descente de gradient projetée, s'adaptant à cette seconde contrainte qui borne les tarifs sur $\mathrm{E} = P_i \in [P_{i,0} - (P_{i,0} - Pr_{i,0}^{HT}) * 20\%); P_{i,0} + (P_{i,0} - Pr_{i,0}^{HT}) * 20\%)]$ n'est autre que :

$$P_{t+1} = projection_E(P_t - \nabla L(P_{1,t}, ..., P_{n,t}, \lambda) * \delta) \tag{49}$$

Le coefficient β_{j_0} étant strictement positif, $L(P_1, ..., P_n, \lambda)$ est convexe comme composée et produit de fonctions convexes. Il existe donc une solution au problème d'optimisation. En revanche, aucune fonction n'est prédéfinie pour la maximisation sous contrainte dans les langages de programmation classique, cette tâche peut donc s'avérer ardue aux vues de la complexité de la contrainte (46) qui englobe l'ensemble des prospects. Afin de faciliter la résolution du lagrangien, une grille de lambda est créée telle que, pour chacune des valeurs de la grille, une allocation P soit trouvée [20] (réf (17)). Concrètement :

Pour $\lambda \in$ Grille :
    Résoudre :$L(P_1, ..., P_n, \lambda)$ par descente de gradient projetée
        Vérifier si $\sum_{i=1}^{N} \pi_i >= \theta$:
            Fin si vérifié

La solution trouvée ne serait donc pas forcément la plus optimale, la contrainte globale n'étant pas saturée. A noter que plus λ augmente et plus la contrainte de volume impose son poids et donc la transformation a plus de valeur et le tarif diminue.

Une fois l'allocation déterminée, le niveau de marge est amélioré sur l'ensemble du portefeuille et il est possible de réaliser une optimisation ratebook pour déterminer à partir des critères tarifaires, le tarif optimal pour chacun des prospects effectuant un devis.

19. Antoine, Dreyfuss et Privat, *Introduction à l'optimisation : aspects théoriques,numériques et algorithmes*, p. 81, 2006-2007

20. Sourisseau, *Modélisation du taux de transformation et de l'élasticité prix en affaire nouvelle pour l'assurance automobile*, p 59-62, 2003

6 Conclusion

Le traitement du biais dans la mesure de l'élasticité-prix a été mené avec succès dans les différentes méthodes employées. Les champs d'application de ce mémoire se répartissent sur plusieurs pans.

En premier lieu, avec un modèle précis d'élasticité au prix, il est alors possible pour l'assureur de simuler les impacts financiers - chiffre d'affaires, loss ratio, rentabilité - des changements tarifaires envisagés de manière rapide et fiable. Il permet alors d'évaluer des mesures commerciales. L'élasticité au prix fournit un outil d'aide à la décision précieux dans l'atteinte des objectifs d'un assureur une fois la vision du risque retranscrite à travers la prime pure.
Ensuite, ce qui vient déterminer les mesures commerciales, le modèle permet de mieux comprendre les comportements des clients. L'analyse de la transformation par segment de risque pointe un ciblage précis sur lequel l'assureur peut entreprendre des actions commerciales. Les critères impactant la demande constituent les leviers aptes à faire face à la mise en concurrence accrue des acteurs sur le marché de l'assurance automobile. Entre autres, les remises commerciales et ajustements tarifaires expliquent la hausse de la conversion ainsi que les segments de risque sur lesquels le tarif proposé est bien situé par rapport au marché. De manière générale, l'élasticité permet de mieux comprendre les différences tarifaires avec les concurrents et indique donc si un changement de politique tarifaire doit être effectué.
Enfin, les courbes d'élasticité-prix fournissent les éléments, à travers la marge pondérée, aboutissant à l'établissement du tarif optimal à proposer pour atteindre le meilleur niveau de rentabilité.

La mesure de l'élasticité-prix a nécessité de pallier à un problème de manque de données pour qu'elle puisse être prédite avec une meilleure justesse. Grâce au respect de la propriété de chevauchement des données, des techniques d'appariement ont été employées, notamment le jumelage par score de propension estimé par XGBoost qui a montré de bons résultats et qui n'a n'a pas été détrôné par le *Genetic Matching* en raison de difficultés opérationnelles quant aux ressources de calcul. Bien que les résultats soient satisfaisants, la robustesse de l'algorithme n'excède pas 20% d'évolution autour de la marge d'origine et ceci constitue une limite de la méthode. Le jumelage ne peut remplacer un price test qui pourrait indiquer les mouvements même au-delà de ce seuil, et apporter des données exactes également à l'intérieur de ce seuil.

Pour aller plus loin, l'estimation de l'élasticité au prix peut également conduire à la réalisation d'une optimisation tarifaire, qui impose l'appel de modèles interprétables. Le tarif optimal choisi peut ensuite être modélisé à partir des caractéristiques de la base devis. Une optimisation ratebook mise en production octroierait un aperçu des segments efficients du tarif à atteindre.

Cependant l'optimisation tarifaire repose sur des pré-requis qui n'ont pas tous été travaillés ici :

1. Un bon modèle de prime pure.

2. Une bonne gestion des frais et des charges.
3. Un bon modèle de tarif concurrent.
4. Une bonne étude des offres concurrentes disponibles sur le marché.
5. Une bonne adéquation entre l'assuré et la couverture proposée.
6. Un bon modèle d'élasticité-prix à la souscription.
7. Un bon modèle de rétention client (taux de résiliation, élasticité au prix...). Le profil de la cible ne tient pas que de sa probabilité de survenance de sinistre et de son coût moyen. Son comportement sur le long terme est également à envisager pour mesurer l'importance de l'investissement à déployer de la part de l'assureur (prix d'entrée compétitif, ristournes commerciales, plan marketing renforcé...). Les modèles de rétention qui estiment le niveau de fidélité d'une classe sont une aide précieuse pour déterminer la valeur client.
8. Une image de marque soignée.

C'est la réunion de ces éléments qui garantit la bonne santé financière d'un organisme d'assurance et dirige les stratégies commerciales et marketing permettant de modifier la transformation d'une cible à son avantage.

Références

[1] Akerlof, *The Market for "Lemons" : Quality Uncertainty and the Market Mechanism, Quarterly Journal of Economics*, p. 488–500, 1970

[2] Antoine, Dreyfuss et Privat, *Introduction à l'optimisation : aspects théoriques,numériques et algorithmes*, p. 81, 2006-2007

[3] Berk, *Regression Analysis : A Constructive Critique*, Chap. 5, 2004.

[4] Charpentier, Denuit, Elie, *Pricing Game*, 100% ACTUAIRES / 100% DATA SCIENCE, 2015

[5] Dawid,*Conditional Independence in Statistical Theory, J. R. Statist. Soc. B 41*, p.1-31, 1979

[6] De Larrard,*Commercial price optimization strategies in car insurance*, 2016

[7] Diamond et Sekhon, *Genetic Matching for Estimating Causal Effects : A General Multivariate Matching Method for Achieving Balance in Observational Studies, Review of Economics and Statistics*, 2012

[8] Friedman, *Greedy function approximation : a gradient boosting machine, The Annals of Statistics*, vol. 29, p. 1189-1232, 1999

[9] Kim & Guelman, *Propensity Scoring : Theory and Applications*, Dallas, 2015

[10] Lorenz, *Methods of measuring the concentration of wealth, American Statistical Association*,vol. 9, p. 209-219, 1905

[11] Lou & al., *Intelligible Models for Classification and Regression, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 150-158, 2012

[12] Lundberg et Lee, *A unified approach to interpreting model predictions, NIPS'17 : Proceedings of the 31st International Conference on Neural Information Processing Systems (USA)*, 2017

[13] Mahalanobis, *On the generalised distance in statistics, Proceedings of the National Institute of Sciences (Calcutta)* , 1936

[14] Mbengue, *Faut-il brûler les tests de signification statistique ?, M@n@gement*, Vol. 13, p. 100-127, 2010

[15] Nix et Vose,*Modeling Genetic Algorithms with Markov Chains, Annals of Mathematics and Artificial Intelligence*, vol. 5, p. 79–88, 1992

[16] Rosenbaum & Rubin, *The central role of the propensity score in observational studies for causal effects, Biometrika* , p. 41-55, 1983

[17] Sourisseau, *Modélisation du taux de transformation et de l'élasticité prix en affaire nouvelle pour l'assurance automobile*, p 59-62, 2003

[18] Youness, *Contributions à une méthodologie de comparaison de partitions* ,2004

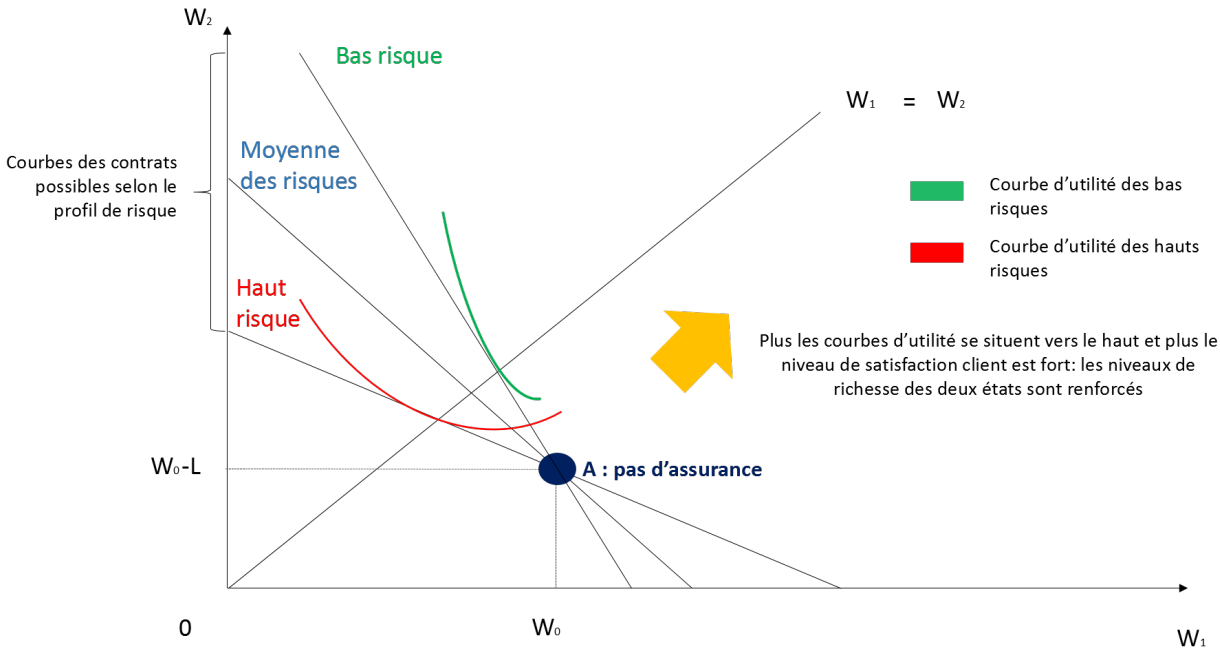
Annexes

Annexe A Prix d'équilibre en assurance : le principe de sélection adverse

A.1 Le Modèle d'Akerlof, Rothshild et Stiglitz (1976)

Pour illustrer ces propos, on se place temporairement dans un marché ou segment tarifaire où il existe deux profils de risque : les hauts risques et les bas risques. Nous sommes en situation d'asymétrie d'information : on ne sait pas précisément évaluer le risque de chaque assuré. On adopte les notations suivantes, et la modélisation qui s'ensuit nous vient de Akerlof, Rothshild et Stiglitz.

- w_0 est la richesse initiale des agents souscripteurs.
- w_f est la richesse finale des agents souscripteurs.
- w_1 la richesse dans le monde 1 sans accident et w_2 la richesse dans le monde 2 avec accident.
- L est la perte en cas d'accident
- En particulier, q_H est la probabilité d'avoir un accident pour les hauts risques et q_B est la probabilité d'avoir un accident pour les bas risques. On a $q_H > q_B$. q est la probabilité d'occurrence de sinistre, on a q = $\lambda * q_H + (1 - \lambda) * q_B$, avec λ la proportion de hauts risques.
- I est l'indemnité que verse l'assurance en cas de sinistre aux assurés.
- σ est le coefficient de chargement, ie, un facteur qui prend en compte les frais à charge de l'assureur et l'élaboration de sa marge.
- P est la prime d'assurance où P = $(1 + \sigma)qI$
- La prime actuarielle est qI. Elle est inférieure à la prime d'assurance.



Un équilibre du marché de l'assurance -au sens de Rothschild et Stiglitz- est un menu de contrats d'assurance $(P_1, I_1), ..., (P_n, I_n)$ tel que :

1. Chaque contrat du menu doit réaliser un profit strictement positif (sinon, les assureurs proposant cette police la retireraient).
2. Aucun contrat supplémentaire arrivant sur le marché ne peut être créé et générer des bénéfices strictement positifs.

On établira deux approches : le cas d'un équilibre mélangeant où l'on ne segmente pas, et le cas d'un équilibre séparateur. Dans cette dernière, c'est la quantité d'assurance qui discrimine les assurés entre eux au sein d'un même segment chez Rothschild et Stiglitz. Les hauts risques sont prêts à payer d'avantage pour plus de couverture : on pratique alors de l'auto-sélection.

Equilibre pooling

Sur les figures ci-dessous, le point A correspond toujours au point de non-assurance. Le point B est proposé comme un équilibre mélangeant. Il ne peut se situer en dessous de la droite M ou alors il existerait un contrat profitable qui attirerait les deux types de profil. On ne peut proposer une couverture complète aux deux types d'individus sinon les bas risques préfèrent ne pas s'assurer. La deuxième possibilité n'est pas efficiente non plus car il existe un contrat dans la zone orangée qui n'attirerait que les bas risques et serait donc profitable (car en-deça de la droite B).

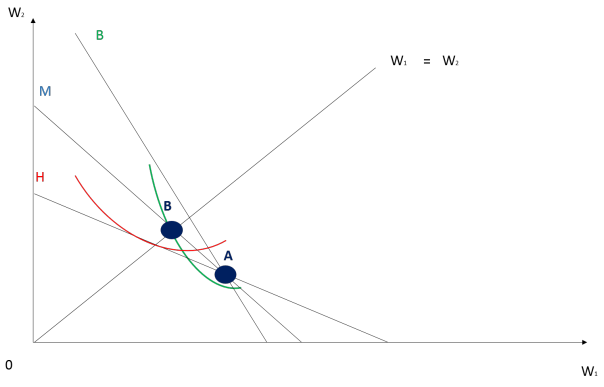


FIGURE 69 – Couverture complète

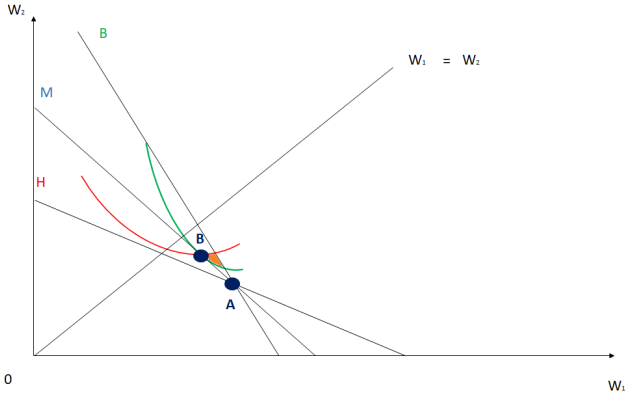


FIGURE 70 – Couverture partielle

En assurance, un contrat conçu pour un assuré moyen attire seulement les plus risqués et mène au déficit. C'est pour cette raison qu'une segmentation trop grossière aboutit à la perte d'une part de marché avec une mauvaise définition de prix.

Equilibre séparateur

On tente de proposer deux contrats différents : l'objectif est qu'aucun profil ne puisse dévier afin qu'ils s'auto-sélectionnent et délivrent de l'information concernant leur niveau de risque. On propose une couverture complète C aux hauts risques, située sur sa droite des loteries possibles H. Pour qu'aucune déviation n'ait lieu, on choisit le contrat des bas riques B comme intersection entre la droite B est la courbe d'utilité des hauts risques. Dans le cas où la proportion de hauts risques serait suffisamment élevée, il n'existe pas de contrat profitable pouvant destabiliser l'équilibre.

En revanche, Rothshild et Stiglitz montrent que le concept d'équilibre qu'ils proposent ne permet pas d'obtenir une situation stable si la proportion de hauts risques est faible. Dans ce cas, on peut en effet au moins trouver un contrat dans la zone orangée qui attire tous les individus.

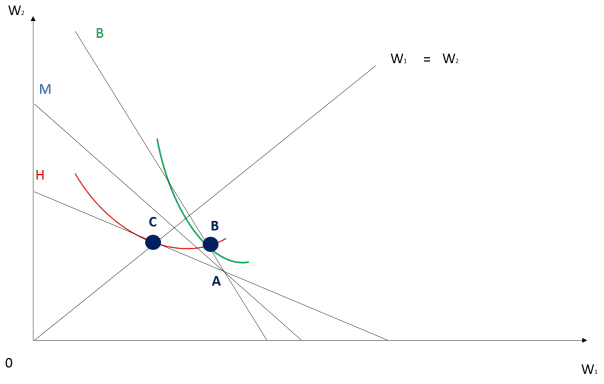


FIGURE 71 – λ élevé

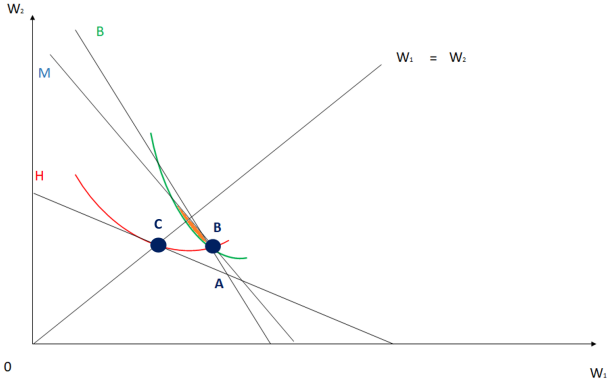


FIGURE 72 – λ faible

A.2 Equilibre de Wilson (1977)

Wilson modifie les hypothèses de comportement des acteurs. Les assureurs peuvent répondre à un assureur tiers qui aurait déstabilisé l'équilibre. La dimension stratégique ajoutée au comportement de l'assureur aboutit à éliminer la tentation de déstabiliser le contrat mélangeant en écrémant les hauts risques. Wilson propose une nouvelle dimension de l'équilibre, en ajoutant l'hypothèse de libre sortie des acteurs.

Un équilibre du marché de l'assurance -au sens de Wilson- est un menu de contrats d'assurance $(P_1, I_1), ..., (P_n, I_n)$ tel que :

1. Chaque contrat du menu doit être profitable (sinon, les assureurs proposant cette police la retireraient).
2. Aucun contrat supplémentaire arrivant sur le marché ne peut être créé et générer des bénéfices strictement positifs, même lorsque les premiers contrats (déficitaires) existant sur le marché sont retirés.

La définition de l'équilibre adoptée par Wilson intègre donc l'anticipation par un assureur de la réaction de ses concurrents qui retirent du marché des contrats devenus déficitaires, du fait de sa propre entrée sur le marché. Cela peut se produire dans le cas de contrats jugés peu rentables ou négligeables au regard de leur poids sur le marché. Cette hypothèse permet de rétablir l'équilibre pooling présent sur la figure 70, puisque si les premiers contrats se retirent, alors le nouvel entrant captera à la fois les hauts mais aussi les bas risques et ne serait donc pas profitable, puisque le contrat proposé se situerait en-deça de la droite de faisabilité pour les équilibres mélangeants. La dimension stratégique attribuée au comportement des assureurs aboutit à éliminer la tentation de déstabiliser le contrat mélangeant en écrémant les hauts risques.

A.3 Equilibre de Miyasaki-Spence-Wilson (1977)

L'équilibre de Miyasaki conserve le principe stratégique à la Wilson mais lève les contraintes de profit des assureurs. Jusqu'ici, chaque contrat proposé devait être profitable individuellement. Au lieu que chaque contrat soit profitable, on émet maintenant la possibilité d'une profitabilité globale : c'est le principe de mutualisation des risques. Le but est de compenser des pertes au niveau des hauts risques avec des gains sur les bas risques. Par définition, il n'existe donc plus d'équilibre mélangeant, il est nécessairement séparateur.

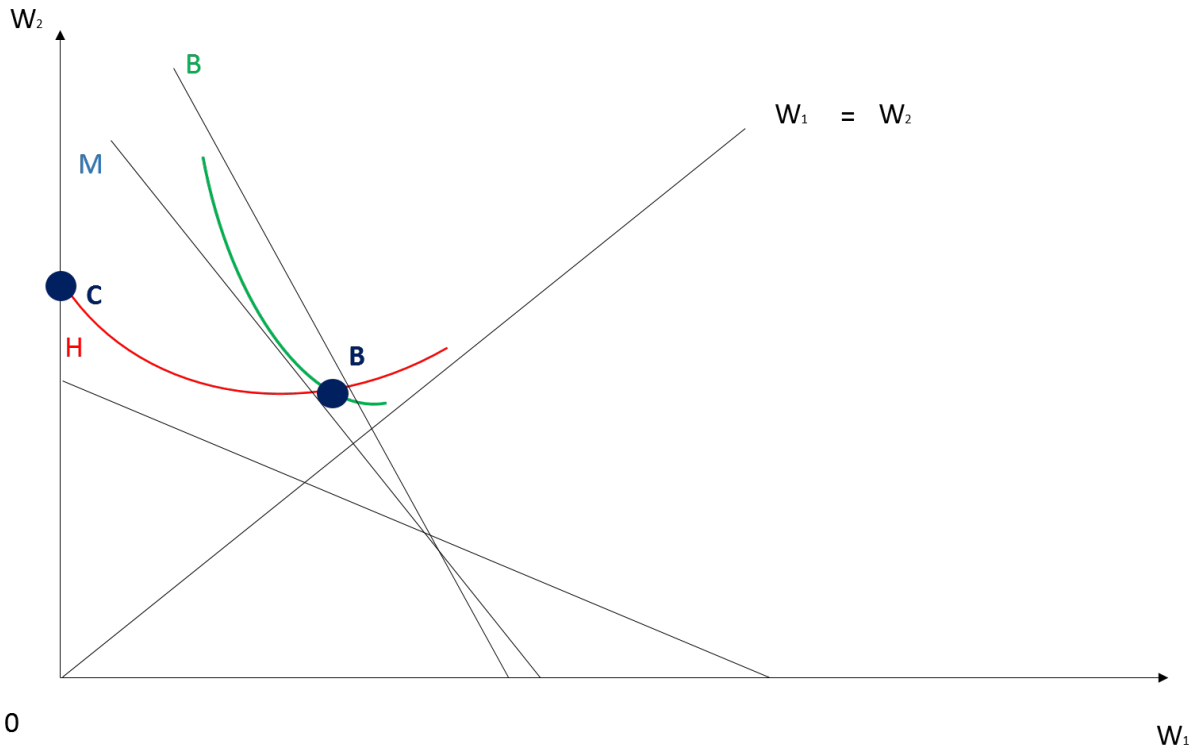


FIGURE 73 – λ faible

Il existe alors un équilibre dans le cas de la figure 72 où la proportion de hauts risques est relativement peu élevé : on propose le couple (B,C) de la figure 73. En effet, il n'existe pas d'autre contrat préféré par les bas risques qui n'attire pas directement ou indirectement les hauts risques. Aussi, si on offre un contrat préféré par les deux catégories d'individu, il réalise un profit négatif.Si on offre un contrat préféré par les bas risques mais pas par les hauts risques, B est délaissé par les bas risques et C est déficitaire et donc retiré du marché. Le nouveau contrat (selon la règle de Wilson) ne représente donc pas une menace crédible car la disparition de C le rendra déficitaire. Enfin, l'équilibre mélangeant n'est pas possible puisque les bas risques peuvent s'attendre à un minimum de niveau d'utilité.

L'équilibre existe alors quelle que soit la proportion de hauts risques dans la population. Son émergence nécessite cependant que les firmes soient suffisamment coordonnées pour accepter de vendre des contrats déficitaires et pour se refuser à mettre en œuvre une stratégie déstabilisante en cherchant à attirer des bas risques.

Annexe B Matrice de confusion et critères associés

A partir des matrices de confusion, on peut évaluer le score d'exactitude :

$$score\ d'exactitude = \frac{Nombre\ de\ prédictions\ correctes}{Nombre\ de\ prédictions} \tag{50}$$

Quelque fois le score d'exactitude se traduit en taux d'erreur :

$$taux\ d'erreur = 1 - \frac{Nombre\ de\ prédictions\ correctes}{Nombre\ de\ prédictions} \tag{51}$$

La qualité de prédiction ne doit pas être présente uniquement à l'échelle globale : le score d'exactitude par classe doit également attirer notre intention.

$$sensibilité = \frac{Nombre\ de\ devis\ transformés\ correctement\ prédits}{Nombre\ de\ devis\ transformés} \tag{52}$$

Il s'agit de la probabilité que la classification conclut à un devis transformé si tel est bien le cas. Cette quantité est appelée sensibilité.
Symétriquement, la même formule est employée pour les devis non transformés (Y = 0) et définit la spécificité.

$$spécificité = \frac{Nombre\ de\ devis\ non\ transformés\ correctement\ prédits}{Nombre\ de\ devis\ non\ transformés} \tag{53}$$

La sensibilité n'a de sens que lorsqu'on la compare à la spécificité. Lorsque la somme de la spécificité et de la sensibilité vaut 1 alors la prédiction n'a rien à voir avec le fait que le devis soit transformé.

En situation réelle, c'est le résultat prédictif du modèle qui est disponible et c'est à partir de celui-ci que l'assureur doit évaluer si l'assuré qui a remplit le devis fera parti des affaires nouvelles. La valeur prédictive positive est la probabilité que la maladie soit présente lorsque le test est positif :

$$valeur\ prédictive\ positive = \frac{Nombre\ de\ devis\ transformés\ correctement\ prédits}{Nombre\ de\ devis\ prédits\ comme\ transformés} \tag{54}$$

Symétriquement, la valeur prédictive négative est la probabilité que la souscription ne soit pas finalisée lorsque la prédiction est négative. Les valeurs prédictives dépendent de la prévalence de la souscription. Ainsi, pour une même sensibilité et spécificité, la valeur prédictive négative d'une prédiction donnée va s'améliorer d'autant que la souscription est rare (peu prévalente) et la valeur prédictive positive de la même prédiction va s'améliorer d'autant que la souscription est fréquente.

Le score F1 est une sorte de moyenne entre le score d'exactitude par classe et la valeur prédictive par classe.

$$score\ F1 = 2 * \frac{valeur\ prédictive\ *\ score\ d'exactitude}{valeur\ prédictive\ +\ score\ d'exactitude} \tag{55}$$

Annexe C Intégration de données concurrentes

Sur les sites comparateurs, à la fin du recueil d'informations fournies par les clients, ces derniers pourront obtenir un aperçu des tarifs proposés par ordre décroissant de prix, le leader en tête. Les assureurs directs, qui disposent de frais amoindris, avancent généralement des tarifs extrêmement compétitifs tout en bénéficiant d'une marge.

Le leader, c'est- à dire celui qui a le prix le plus bas sur un segment et pour une typologie de produit, doit avoir une meilleure segmentation du risque : les variables optées et le découpage de celles-ci sont plus efficients. Le leader, à produit similaire, représente un objectif à atteindre sur le marché. [21]

La collecte automatisée de données provenant d'un site comparateur d'assurances automobiles pourra permettre de reproduire ce tarif leader afin d'obtenir le meilleur compromis entre hypersegmentation et mutualisation du risque, par des techniques de reverse engineering.

Le principe est d'alimenter une base de données fictives à partir d'un large spectre de profils simulés pour lesquels on extrait les dix meilleurs tarifs (ce qui constitue notre panier) à une date t pour chaque type de couverture proposé.

C.1 Collecte automatisée de données tarifaires

C.1.1 Cadre légal

Le processus de scraping en lui-même n'est pas répréhensible. Il consiste à extraire des informations accessibles en ligne. Prohiber le scraping, cela reviendrait donc à prohiber tout Internet. Le scraping est donc une activité légale, c'est la récupération de données publiques et disponibles. Le scrapeur pratiquant le plus reste Google dont la plupart des services sont construits autour de cette pratique. Le droit n'interdit pas l'opération en elle-même. Par contre, la façon dont on collecte l'information et dont on la réutilise peut être condamnable. Il existe trois axes de condamnation possibles liés à l'aspiration d'information sur un site internet :

1. Tout d'abord, si un site Internet interdit explicitement le scraping dans ses conditions générales d'utilisation (CGU), alors il peut attaquer en justice au nom du droit des contrats. Dans les conditions d'utilisation de notre comparateur, aucune spécification quant à l'utilisation du site par un robot n'est mentionnée.
2. Le propriétaire d'un site Internet peut également saisir la justice pour violation de la propriété intellectuelle dès lors qu'il est capable de prouver qu'il a consacré un investisse-

21. En revanche, lorsqu'on se place en tant que leader, on a intérêt à ne pas trop s'éloigner de son concurrent direct, deuxième sur le marché. Augmenter son tarif aura une influence positive sur le Loss Ratio et la marge de rentabilité tandis que le chiffre d'affaires sera accru toutes choses égales par ailleurs, puisque le leader disposera du même nombre de parts du marché. Là encore, le delta de manoeuvre nécessite la connaissance des tarifs concurrents.

ment substantiel, financier, humain ou matériel pour agencer ses données sous forme de base structurée. Or, le comparateur se contente d'obtenir les tarifs de ses partenaires et perçoit une commission lors de la souscription d'un contrat. Il s'agit en effet d'un intermédiaire pour la distribution de produits d'assurance. De plus, l'aspiration de tarifs par devis envoyé est employée à des fins d'analyse et de benchmark tarifaire, il ne s'agit pas de concurrence déloyale envers le site comparateur.

3. La pratique du web scraping pourrait être considérée comme un "vol de données" (atteinte au STAD) en s'appuyant sur l'article 323-3 du Code pénal qui énonce :

   "Le fait d'introduire frauduleusement des données dans un système de traitement automatisé, d'extraire, de détenir, de reproduire, de transmettre, de supprimer ou de modifier frauduleusement les données qu'il contient est puni de cinq ans d'emprisonnement et de 150.000 € d'amende."

Sur le dernier point, il conviendrait alors de caractériser l'intention frauduleuse du web scraping. En effet, l'accès et le maintien frauduleux dans un système informatique consistent à pénétrer sans droit dans un système en en forçant l'accès. Les dispositions du Code pénal permettent de lutter contre les intrusions frauduleuses (connexion pirate, appel d'un programme ou d'un fichier sans autorisation etc), le maintien frauduleux, l'entrave d'un système ou l'altération de son fonctionnement (virus, mail bombing etc), ainsi que l'altération, la suppression ou l'introduction de données pirates.

Le site du comparateur est en libre accès et en aucun cas la démarche envisagée vise à modifier la structure de son contenu. En revanche, il faut veiller à ne pas attaquer le comparateur en envoyant un trop grand nombre de requêtes en très peu de temps, ce qui pourrait aboutir à un dysfonctionnement de leur service. Les comparateurs utilisent les captcha afin de lutter contre ces types de cyber attaques. Notre web-scraper n'envoie que très peu de requêtes par jour, le processus n'a donc aucun impact sur leur serveur.

C.1.2 Simulation de profils

Le site comparateur se compose d'un tronc commun d'une trentaine de questions. En fonction des réponses que l'on fournit, d'autres sous-questions peuvent se débloquer. Par exemple, si l'on compte acheter le véhicule, et si et seulement si on compte l'acheter, la question du type de paiement apparaît. De manière similaire, c'est uniquement lorsque l'on déclare un conducteur secondaire qu'une liste déroulante de choix pour l'identification de celui-ci apparaît.
On prend garde à élaborer des profils complets en recensant toutes les possibilités de questions pour le formulaire. La vérification de l'exhaustivité et la cohérence des profils créés est essentiel pour obtenir une base représentative du marché. Le recueil de tarifs doit couvrir suffisamment de segments pour pouvoir construire un modèle robuste et compléter la base d'individus fictifs en extrapolant parmi ceux les plus proches en termes de caractéristiques.

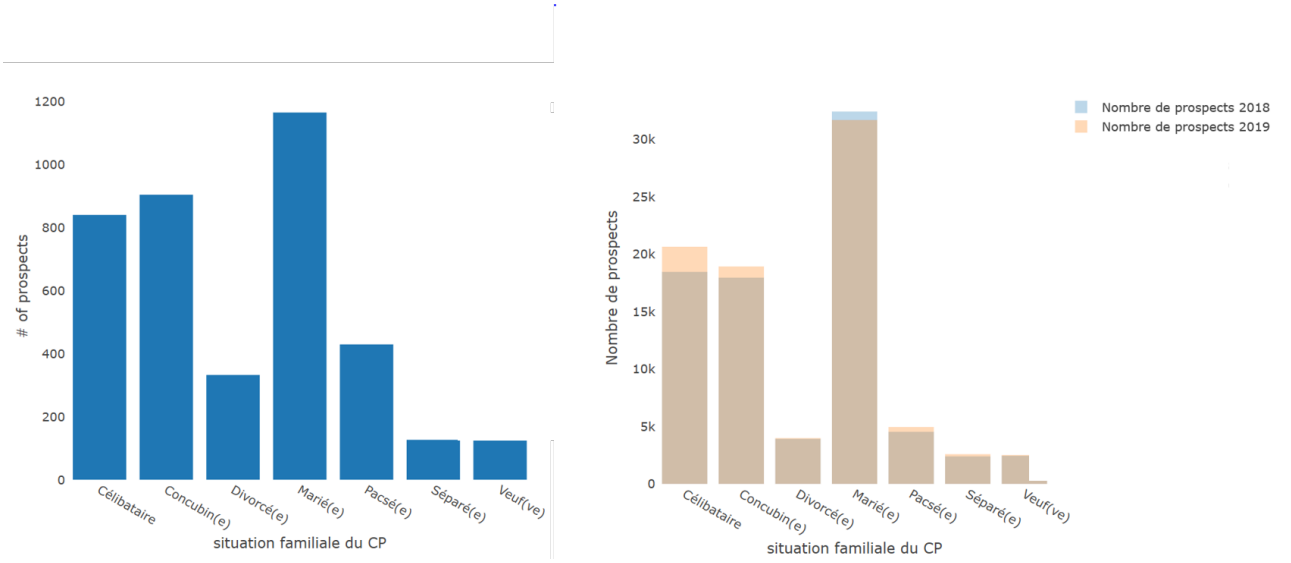


FIGURE 74 – Couverture des segments de la situation familiale au sein de la base des profils simulés (gauche) dans des proportions équivalentes à celles de la base devis de l'assureur (à droite). Les personnes mariées sont les plus représentées sur le marché, suivies des concubinages et des célibataires.

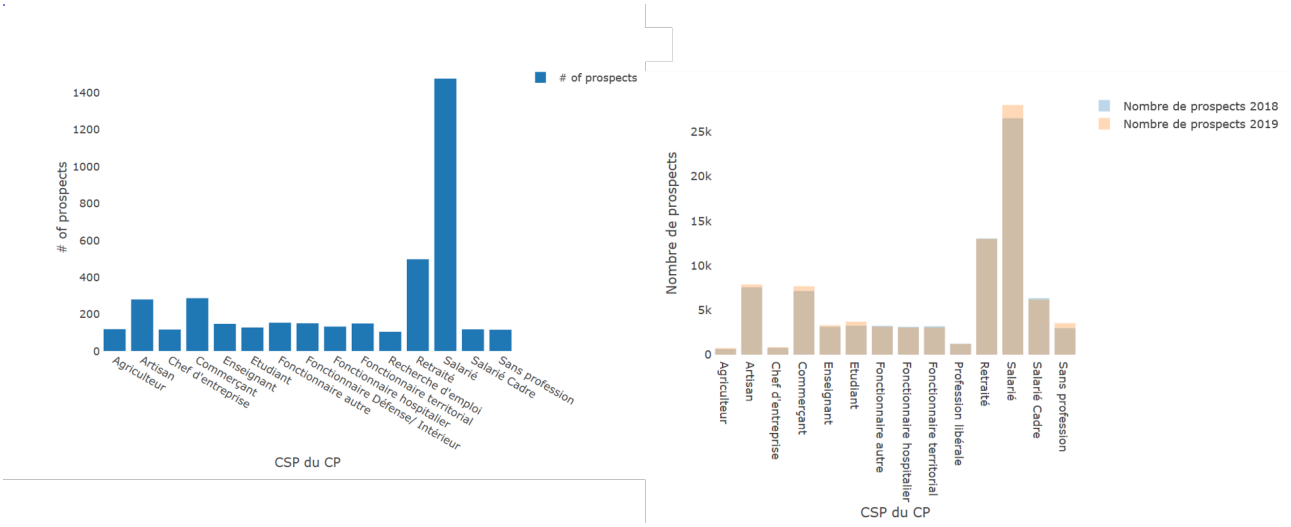


FIGURE 75 – Couverture des segments de la catégorie socio-professionnelle au sein de la base des profils simulés (gauche) dans des proportions équivalentes à celles de la base devis de l'assureur (à droite). Les salariés sont la classe sur-représentée. Les retraités sont très présents dans la demande de devis (également dans le portefeuille de l'assureur). Les probabilités d'occurrence ont donc été adaptées.

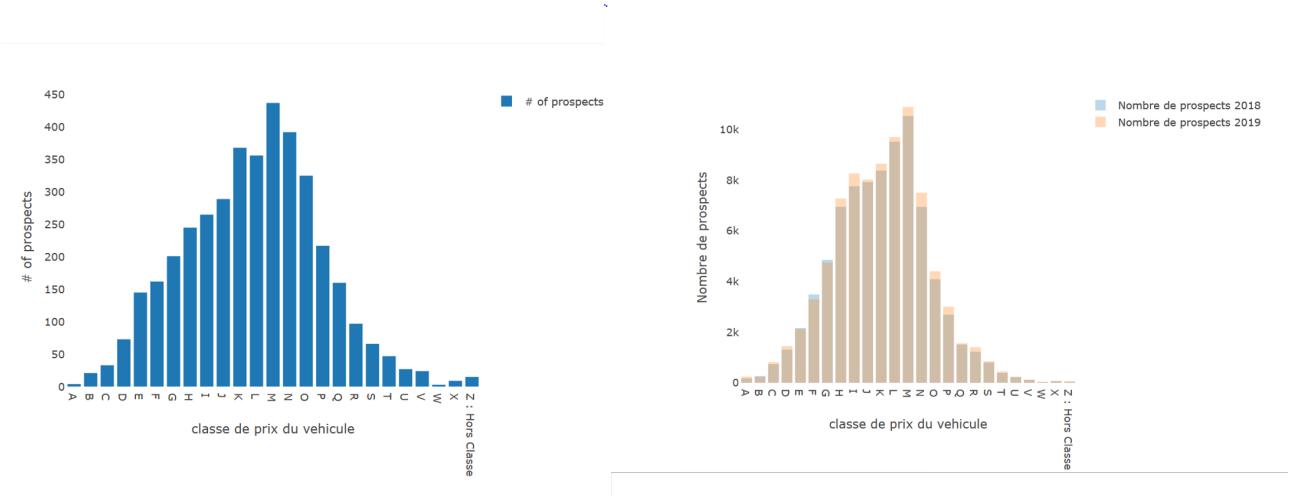


FIGURE 76 – Couverture des segments de prix du véhicule au sein de la base des profils simulés (gauche) dans des proportions équivalentes à celles de la base devis de l'assureur (à droite). Le groupe A représente les véhicules bon marché tandis qu'on avance avec l'alphabet vers les plus onéreux. Les automobiles intermédiaires sont majoritaires.

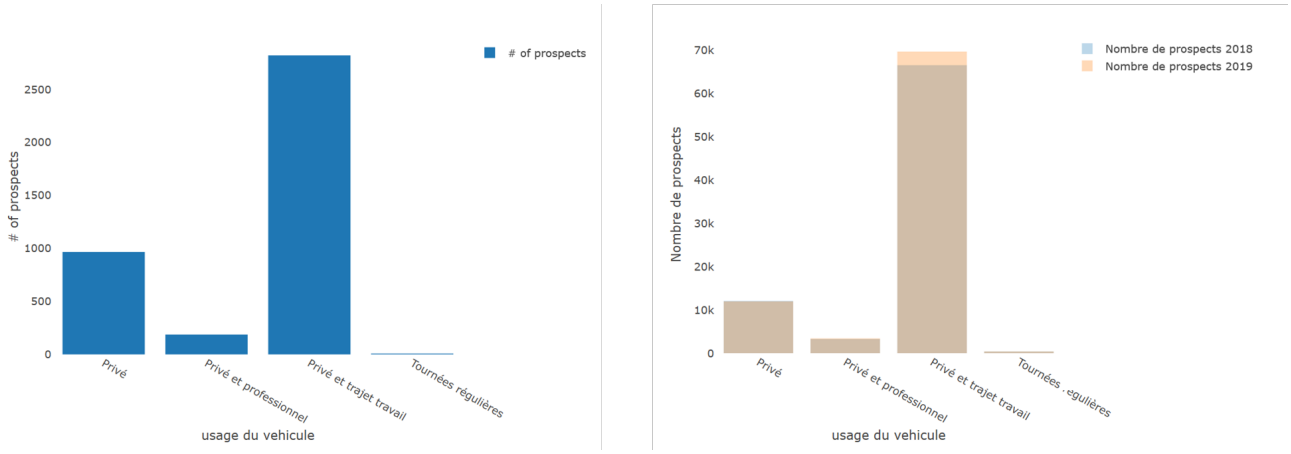


FIGURE 77 – Couverture des segments correspondant à l'usage du véhicule au sein de la base des profils simulés (gauche) dans des proportions équivalentes à celles de la base devis de l'assureur (à droite). Une majorité écrasante des clients utilisent leur véhicule pour des trajets domicile/travail. Une minorité l'emploie à des fins professionnelles.

La création des profils ne peut être aléatoire : une personne de 20 ans a rarement 3 enfants. En fonction des premières variables tirées, on affine les probabilités des suivantes. Le contrôle de la cohérence des profils simulés s'appuie sur des analyses statistiques et la connaissance du marché.

Afin de palier à quelques difficultés, notamment celle d'associer un type de véhicule aux caractéristiques d'un profil, on exerce un mapping entre la base de l'assureur et le comparateur pour l'adapter aux conformités du questionnaire (format des réponses, complétion des questions manquantes...). Cet exercice permet d'identifier les corrélations entre les modalités de réponse et le véhicule à assurer. Ainsi, on adapte la base marché afin qu'elle soit représentative du portefeuille client.

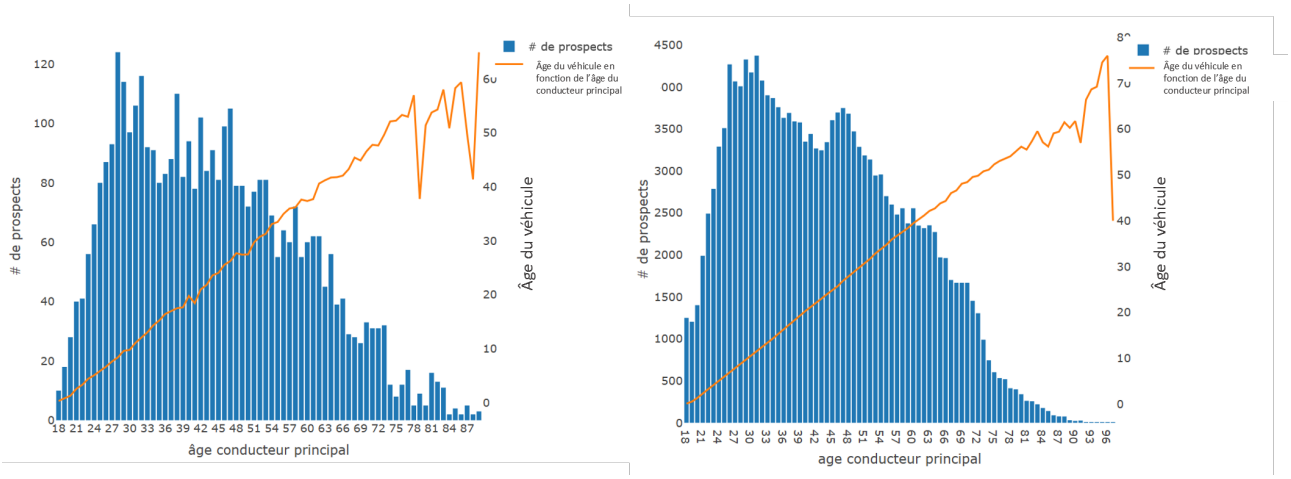


FIGURE 78 – On retrouve la corrélation linéaire positive entre l'âge du véhicule et du conducteur principal au sein de la base des profils simulés (gauche) et la base devis de l'assureur (droite)

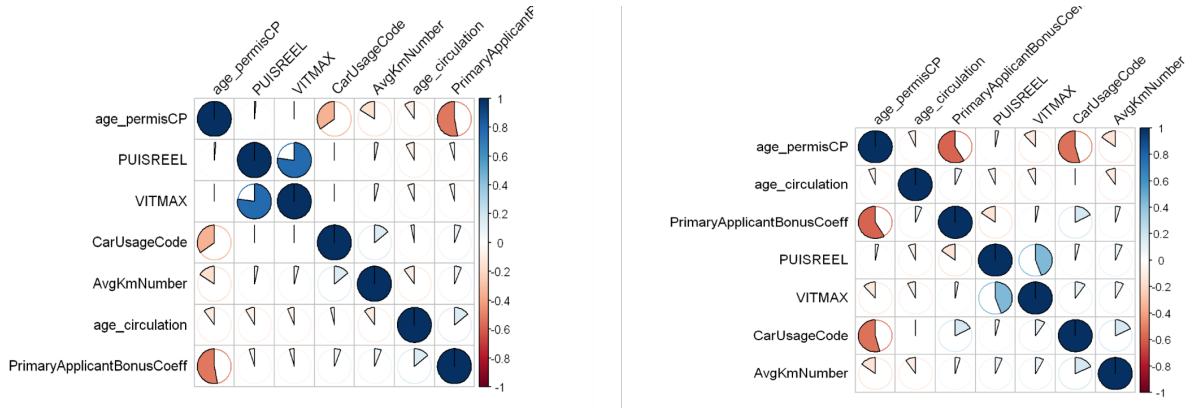


FIGURE 79 – Matrice des corrélations linéaires de la base fictive (gauche) et la base devis (droite) de plusieurs variables continues. Plus l'âge du permis de conduire augmente et plus le coefficient de bonus/malus diminue, ce qui signifie que l'expérience du conducteur influe positivement sur sa fréquence de sinistre. En revanche, plus l'âge augmente et plus l'utilisation du véhicule est détournée des raisons professionnelles et demeure privée. L'utilisation du véhicule est corrélée positivement avec le nombre de kilomètres parcourus, les petits rouleurs affectionnant davantage les trajets privés exclusivement. Enfin, la vitesse et la puissance sont deux grandeurs allant dans le même sens. Le lien entre les variables de la base simulée est cohérent aux vues de celui retrouvé dans la base devis.

D'autres bases disponibles en open source permettent de renforcer l'exactitude des individus crées :

— La base SRA permet l'identification des caractéristiques des véhicules sur le marché.

— Pour l'élaboration de zonier, on précise l'adresse du lieu de travail et du domicile parmi la liste des communes françaises disponible en Open Data sur des sites publics.

C.1.3 Solutionner les problèmes de Captcha

L'adresse IP est un numéro d'identification dont dispose chaque appareil connecté à un réseau utilisant Internet. Lorsqu'on envoie un nombre élevé de requêtes sur le même site, on peut trahir notre activité, anormale pour un particulier.
Le site comparateur limite son site à un nombre de requêtes par IP. Au-delà d'un certain seuil, un Captcha apparaît, qui est difficile à passer et qu'il est possible d'éviter. Un captcha est un test qui permet au site de vérifier que l'utilisateur n'est pas un robot. Ce type de test est difficile à passer lors d'un remplissage automatique de questionnaire.

Première étape : "humaniser" l'algorithme
D'autres facteurs sont à l'origine de l'apparition des Captcha comme les *fingerprints* (empreintes) et le comportement. Les *fingerprints* sont des informations que l'on donne à un site lorsqu'on le consulte (données de configuration de notre ordinateur -langage par exemple- , ce que l'on tape dans la barre de recherche, *cookies*) et plus ces informations sont redondantes (toujours les mêmes recherches dans le navigateur, en langue française..) plus notre identité est menacée car on se singularise.
Le comportement de l'utilisateur permet de détecter s'il s'agit d'une machine ou d'un humain. Des solutions existent pour "s'humaniser" comme ne pas répondre aux questions dans le bon ordre, se tromper dans la réponse à une question et revenir dessus ou accroître le temps de réponse.

Deuxième étape : l'utilisation de proxies
Un proxy agit comme un intermédiaire entre deux ordinateurs. Cet entremetteur sera utilisé de deux manières : d'une part pour pouvoir accéder à la page devis même en franchissant le seuil clé de requêtes déclenchant le Captcha. Il permet de camoufler l'adresse IP en en empruntant une autre. Il existe différents types de proxy, qui peuvent répondre à différentes problématiques. On cherche un regroupement d'adresses IP françaises n'étant pas bloqué par le site et rapide à utiliser. Deux natures de proxy ont été recensées :

1. *Proxy résidentiel* Le principe est de masquer l'adresse IP utilisée derrière une autre appartenant à un particulier. C'est là que réside l'avantage du proxy résidentiel : le comparateur n'a aucun moyen de nous bloquer car il pense que nous sommes des personnes privées qui remplissent le questionnaire.
2. *Centre de données* Le principe est de masquer l'adresse IP utilisée derrière celle de la compagnie ou le centre de données associé (lieu physique où sont regroupés différents équipements informatiques). L'inconvénient est que si une adresse IP est détectée, un sous-réseaux d'adresses IP n'est plus exploitable. Les sites détectent plus facilement ces adresses et les bloquent.

C.1.4 Algorithmie

Le questionnaire à remplir par le client se déroule en plusieurs pages, chacune d'elle conduisant à la page suivante ou -pour la dernière- à la page où figurent les tarifs proposés par le

marché dans l'ordre croissant.
L'objectif, à l'aide d'un langage de programmation, est de pouvoir remplir de manière automatisée à un questionnaire en ligne comportant plusieurs pages à partir d'un profil simulé, pour enfin extraire de la page de devis obtenu les tarifs leader du marché avec les options, franchises et frais de dossier associés pour chaque type de couverture. On effectue du web scraping en dynamique, c'est-à dire que l'adresse de la page à web scraper ne s'obtient pas de façon statique : il ne suffit pas d'accéder à un site pour en extraire l'information, il faut pouvoir naviguer tel un humain pour aboutir aux données souhaitées.

On choisit pour ce faire *selenium*, disponible sur python qui permet d'interagir avec différents navigateurs web de même que le ferait un utilisateur de l'application. Il entre ainsi dans la catégorie des outils de test dynamique (à l'inverse des tests statiques qui ne nécessitent pas l'exécution du logiciel). L'extraction des données s'exécute grâce à la bibliothèque *Beautiful Soup* qui permet d'extraire des données de fichiers HTML.

Un site Web est un ensemble de pages codées en HTML qui permet de décrire à la fois le contenu et la forme d'une page Web.
Les balises sont des codes qui permettent de structurer le contenu d'une page html. Elles vont par paire : une «balise ouvrante» et une «balise fermante». (par exemple < p> et < /p>). C'est grâce à ces balises que l'on peut situer un élément à l'intérieur d'une page web. Lorsque l'on cherche à remplir un questionnaire en ligne, il faut localiser la réponse pour pouvoir cliquer dessus ou la zone de remplissage de texte dans le cas d'une réponse libre (adresse, nom, prénom...). Une expression *XPath* est un chemin de localisation facilement obtenu grâce à la structure descendante d'un document XML [22]. L'*XPath* est repérable manuellement en analysant l'application du site comparateur : l'objet recherché (liste déroulante, radio...) dispose de son propre *XPath*. On se servira le plus souvent de l'*id* lorsqu'il est définit, qui est unique, pour lequel on affectera une valeur (*value*).

C.2 Modélisation du tarif leader

Le tarif minimum de la garantie "Tous risques" rencontré pour chaque individu de la base marché simulée est posé comme objectif. Le leader influence de manière incontestable les choix des consommateurs, dont l'oeil est potentiellement attiré par le premier prix de la liste. En effet, si l'écart entre la première et la deuxième position est important, l'assuré se dirigerait plutôt vers la prime la plus basse et l'effort pour le conduire vers une autre offre est de ce fait accru, même avec une bonne image de marque. Si la première place est en revanche chahutée, avec une faible différence entre les deux concurrents leaders, alors modéliser la moyenne des deux tarifs ou le tarif leader est similaire pour notre étude.

22. Ludovid Roland *Structurez vos données avec HTML*

C.2.1 Calibration d'un modèle XGBoost

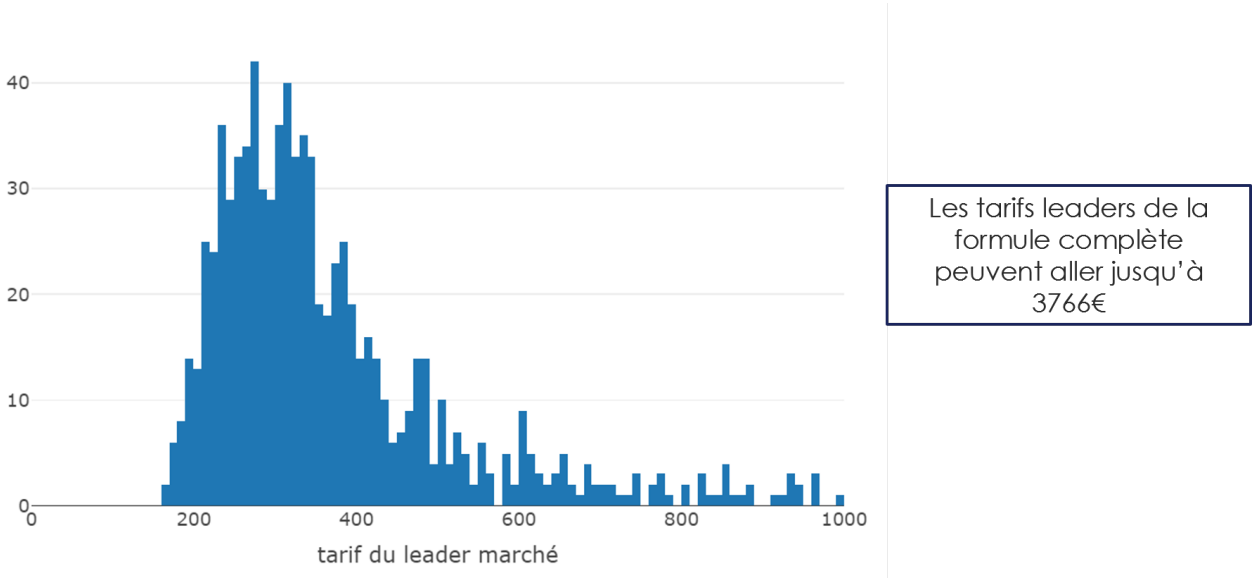
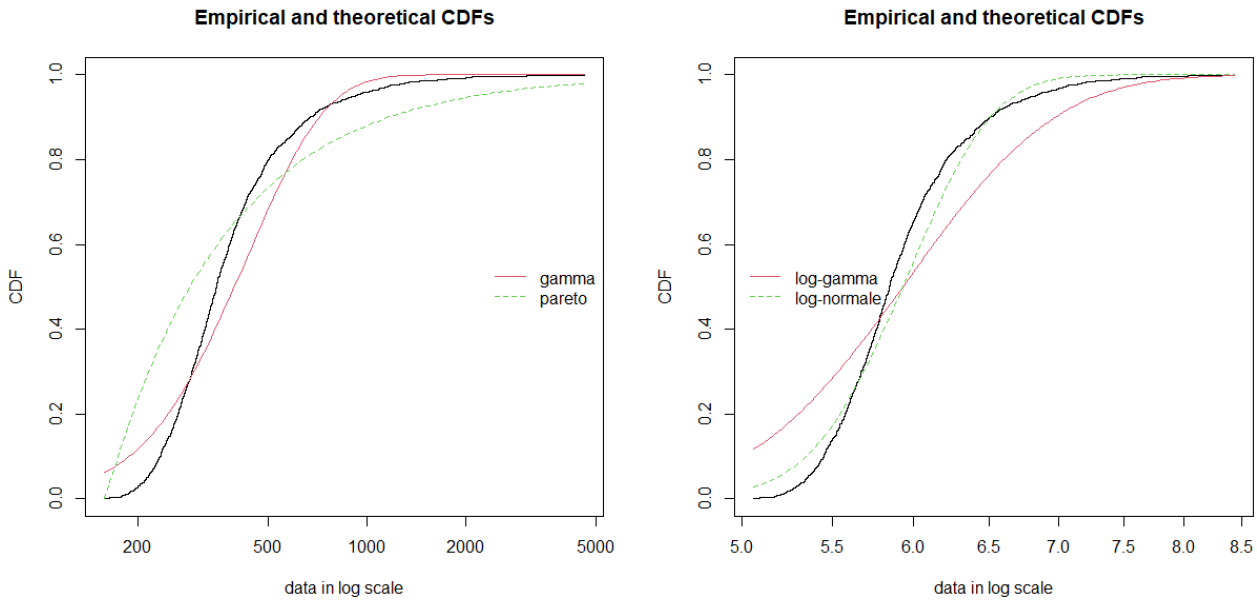


FIGURE 80 – Distribution du tarif le plus compétitif sur le marché

Le comportement du tarif leader de la formule "Tous risques" issu de notre base de données fictives peut être modélisé par une loi gamma ou une loi log-normale. Tout d'abord, les paramètres sont estimés de manière consistente par log-vraisemblance pour les distributions pareto, log-gamma, log-normale et gamma. Les tests de Kolmogorov-Smirnov et Wilcoxon, qui émettent l'hypothèse nulle d'égalité de loi et d'égalité des rangs respectivement, éliminent les lois de pareto et de log-gamma : il y a assez d'évidence statistique pour rejeter l'hypothèse nulle. Des études graphiques des fonctions de répartition et des densités confirment la pré-sélection des lois gamma et log-normale pour ajuster la série de tarifs, la loi log-normale en tête. L'étude des *skewness* et *kurtosis* nous aiguille en revanche vers la loi gamma.



Lois	Kolmogorov-Smirnov	Wilcoxon (non-apparié)
Log-normale	0.2896	0.3249
Gamma	0.2776	0.01903
Log-gamma	2.2e-16	2.2e-16
Pareto	2.2e-16	2.2e-16

TABLE 11 – p-valeurs enregistrées pour les tests de loi

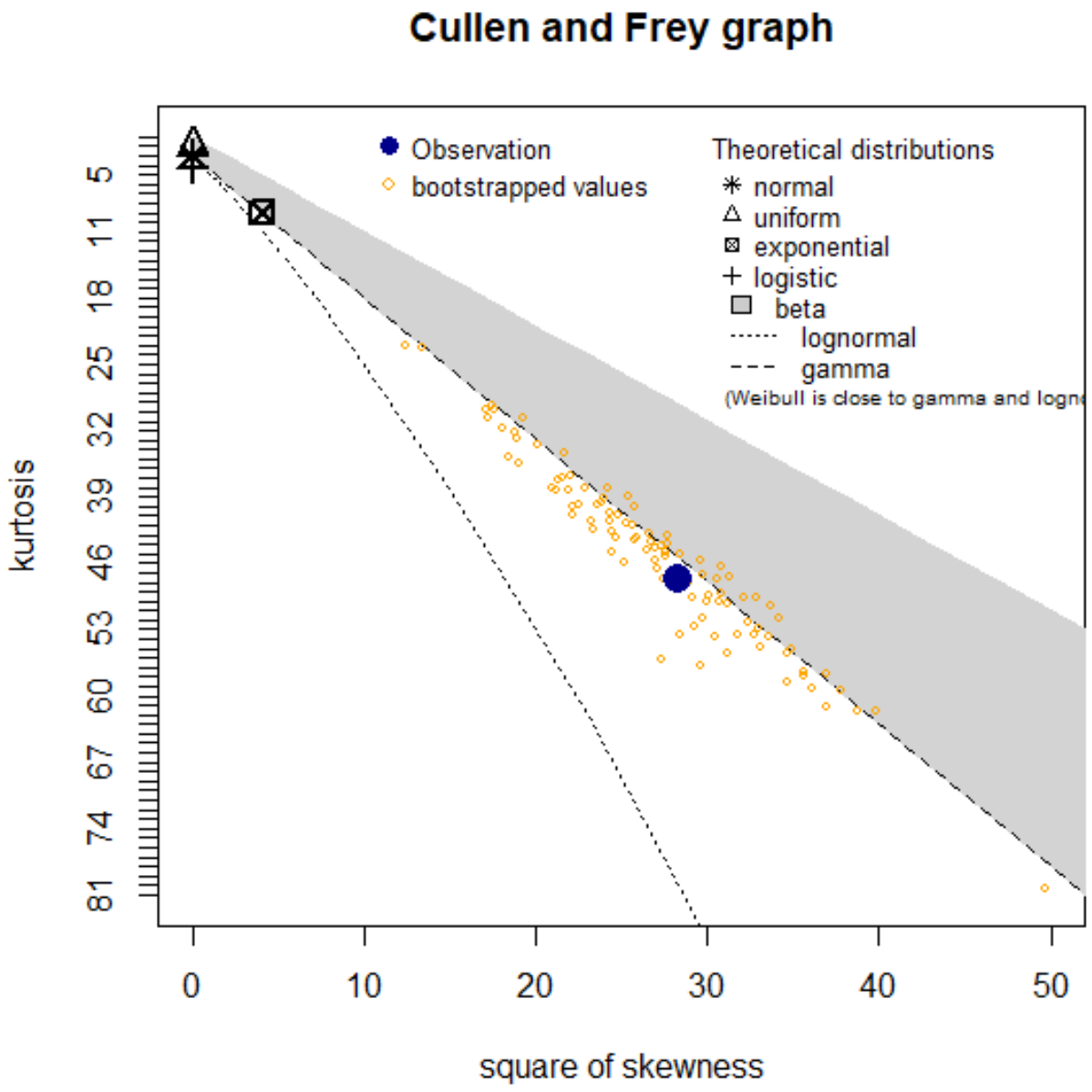


FIGURE 81 – Comparaison moments centrés d'ordre 3 et 4 de différentes distributions et des tarifs boostrapés

Le but recherché est de pouvoir attribuer à chaque sujet de la base devis, le tarif préférentiel qu'il rencontre sur un site comparateur. L'estimation de l'élasticité-prix requiert la prise en

compte du positionnement concurrent. L'approximation du tarif leader se doit d'être la plus précise possible aux vues de nos enjeux. Parmi l'éventail des modèles existant, le XGBoost est opté pour ses bonnes propriétés prédictives, ayant déjà fait ses preuves dans de nombreux concours. Le XGBoost est en réalité issu du *Gradient Boosting*[23] auquel est ajouté des techniques de parallélisation afin d'améliorer les coûts algorithmiques :

1. Un premier modèle est construit pour prédire la valeur cible. Ce modèle est en sur-apprentissage.
2. A partir du modèle construit précédemment, des poids plus importants sont attribués aux observations mal prédites.
3. Un nouveau modèle est construit. Il renforce la performance de prédiction au niveau des observations ayant davantage de poids.
4. L'étape 2 et 3 sont répétées autant de fois que l'utilisateur l'aura suggéré : le *Gradient Boosting* est caractérisé comme séquentiel.

Le résultat final est obtenu par agrégation de l'ensemble des modèles. Afin de profiter de la robustesse de cet algorithme, la bonne paramétrisation est primordiale. Les hyperparamètres se décomposent en plusieures catégories :

— Les paramètres généraux qui incluent le modèle employé, le contrôle de l'apparition de messages d'erreur, les contraintes de parallélisation. Le modèle choisi est l'arbre de régression.

— Les paramètres de boosting qui incluent notamment le poids minimal qu'un noeud de l'arbre doit contenir, la profondeur maximal de l'arbre, le nombre maximal de noeuds terminaux, le seuil du taux d'apprentissage (seuil à partir duquel on considère qu'il y a eu une amélioration de la fonction de perte), la proportion du sous-échantillon utilisé pour la construction du modèle, le nombre de facteurs maximal intervenant dans le découpage d'un noeud et enfin, la pose de contraintes de type L1 ou L2. La calibration de ces paramètres permet un meilleur arbitrage entre biais et variance et visent à éviter le sur-apprentissage. Un ensemble de valeurs est proposée pour chacun d'entre eux et la convergence vers un unique jeu de paramètres sera obtenue par validation croisée sur la base d'apprentissage (procédé de *tuning*).

— Les paramètres d'apprentissage qui comprennent la définition de la fonction de perte et la métrique d'évaluation. La distribution appliquée est une loi gamma (justifiée ci-dessous) et le critère d'évaluation pour l'apprentissage sera définit par *tuning* entre la norme L1 et la norme L2, mesurant ainsi la distance entre l'échantillon observé et prédit.

La métrique d'évaluation choisie pour la base de test, qui intervient dans la validation croisée pour la calibration des paramètres et la validation générale du modèle, est celle de gini (Corrado Gini). Elle permet d'évaluer le niveau d'inégalité entre l'échantillon observé et prédit en tenant compte de la distribution des tarifs. Mathématiquement, la métrique de gini se définit par la différence de l'aire des courbes de Lorenz[24] (réf. (10)) :

23. Friedman, *Greedy function approximation : a gradient boosting machine*, 1999

24. Lorenz, *Methods of measuring the concentration of wealth, American Statistical Association*,vol. 9, p. 209-219, 1905

Courbe de $Lorenz_y \equiv$ (x,y') $\equiv$ (part d'observations cumulées (ajoutées par ordre croissant en fonction de y), part de y détenue par ce cumul d'observations)

$Gini_{Tarifs\ leaders\ observés, Tarifs\ leaders\ prédits} \equiv$ aire de la Courbe de $Lorenz_{Tarifs\ leaders\ observés}$ - aire de la Courbe de $Lorenz_{Tarifs\ leaders\ prédits}$

Paramètres	Résultats du tuning
Pénalisation L1	oui
Proportion de facteurs utilisée pour chaque arbre	70%
Réduction minimale de la fonction de perte pour effectuer un découpage	0.01
Taux d'apprentissage	0.2
Profondeur maximale d'un arbre	3
Poids minimal dans un noeud	5
Métrique d'erreur (apprentissage)	MAE
Proportion de l'échantillon utilisé pour chaque arbre	50%
Nombre d'arbres utilisés pour le boosting	200

TABLE 12 – Hyperparamètres sélectionnés

C.2.2 Création d'un zonier

Le modèle est entraîné sur un sous-échantillon (base d'apprentissage) de la base marché simulée et *scrapée*. Les variables du site comparateur qui ne sont pas présentes dans la base devis de notre assureur ou qui sont hors de notre champ d'étude (facteurs liés aux conducteurs novices, aux conducteurs secondaires) sont retirées pour que le modèle final puisse être applicable à nos assurés. A l'opposé, des données externes liées à la situation géographique du domicile sont retranscrites à travers l'élaboration d'un zonier et incorporées au modèle. L'objectif d'un zonier est d'expliquer une partie des résidus du modèle par la localisation même de l'assuré. La méthodologie appliquée est la suivante :

1. Elaborer un fichier de données externes. Chaque ligne correspond à un code INSEE. Des exemples de variables sont la densité, le nombre de cadres, la température maximale, le nombre de jours de gel, les revenus moyen, la présence d'hôpitaux ou d'écoles etc.
2. Faire un XGBoost pour la prédiction du tarif leader du marché.
3. Extraire les résidus. Ces résidus portent l'information géographique. La différence entre l'observé et l'estimé représente la valeur continue du zonier.
4. Utiliser des données externes géographiques pour expliquer cette variable (résidu) par forêt aléatoire.
5. Lisser la sinistralité sur la totalité du territoire à l'aide des données externes significatives

par krigeage [25]. En effet des codes INSEE peuvent manquer dans notre base de données externes.

6. Relancer le modèle avec ajout de la variable zonier dans le XGBoost.

La constitution d'un zonier, ainsi que la méthode de krigeage utilisée au sein de celle-ci, sortent du cadre de ce mémoire et sont donc détaillées dans les publications en bas de page pour le lecteur intéressé. Néanmoins, les principaux résultats sont rapportés. Ainsi, l'importance des variables obtenue par la prédiction des résidus par forêt aléatoire, agrémentée par la matrice des corrélations des variables externes, octroie à la densité, au nombre de radars à moins de 10km et à l'ensoleillement la robustesse suffisante pour constituer le zonier. Il faut savoir que la latitude et la longitude sont les facteurs discriminants permettant le lissage géospatial effectué par krigeage.

25. Gratton, *Le krigeage : la méthode optimale d'interpolation spatiale, Les Articles de l'Institut d'Analyse Géographique.*, 2002

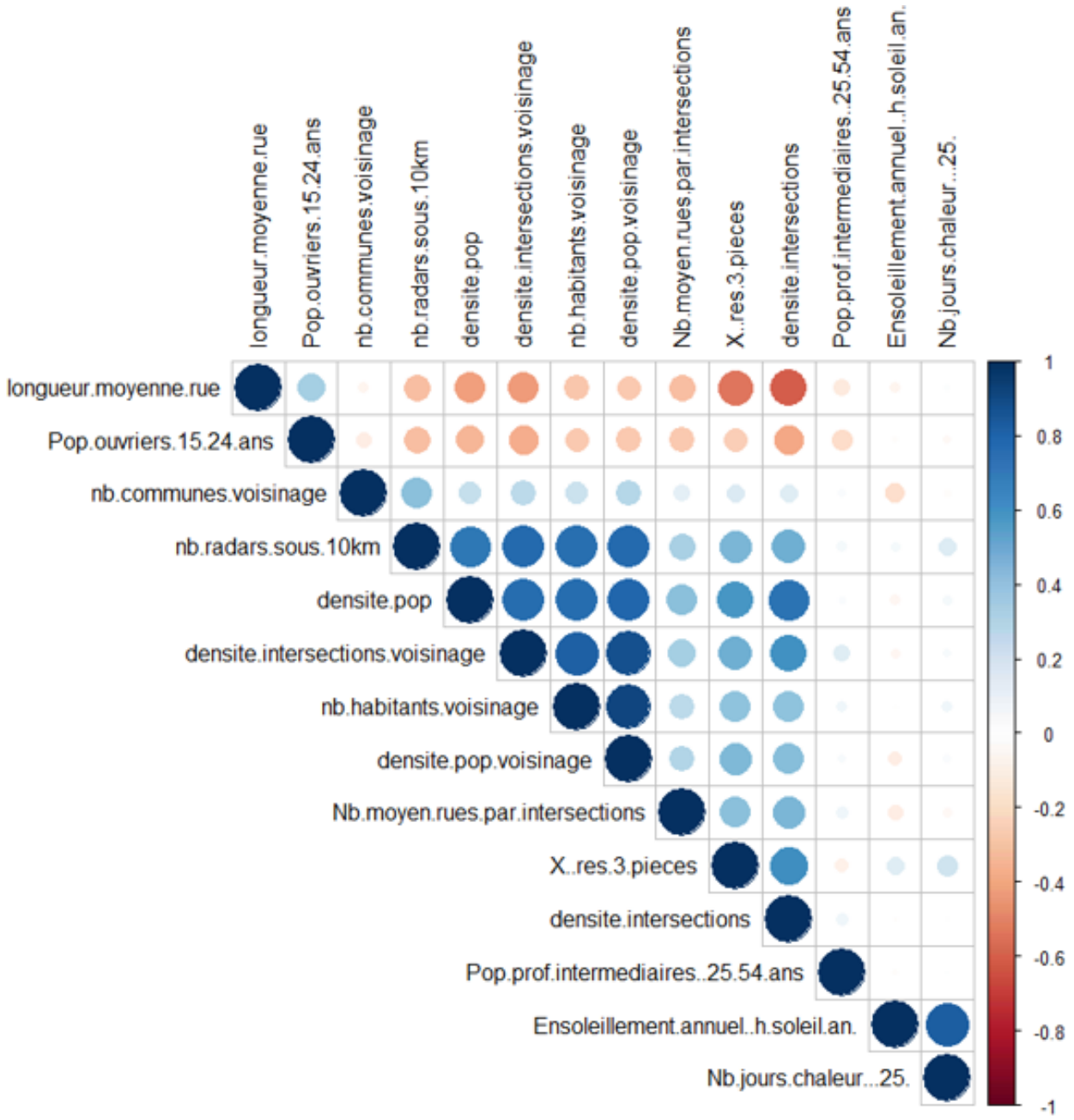


FIGURE 82 – Corrélation linéaire des données externes. Le nombre de radars, les densités, le nombre de communes alentours et la taille restreinte des habitations sont corrélés positivement. La population ouvrière est davantage installée dans les zones à faible densité. La densité et la longueur moyenne de la rue ont une corrélation négative forte. Les données météorologiques sont fortement corrélées.

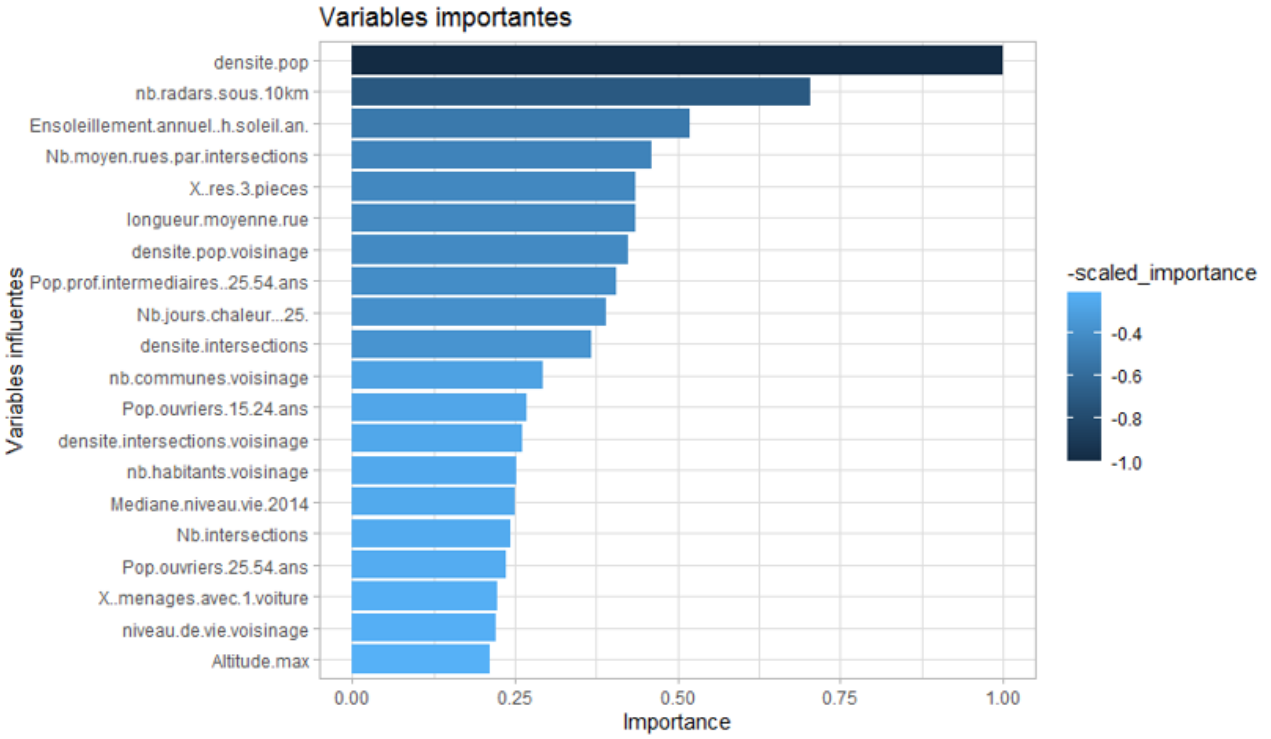


FIGURE 83 – Importances des variables (effectuées par permutation) dans la prédiction des résidus provenant d'un modèle XGBoost pour la prédiction du tarif leader du marché. Les trois premières variables sont retenues aux vues des corrélations existantes.

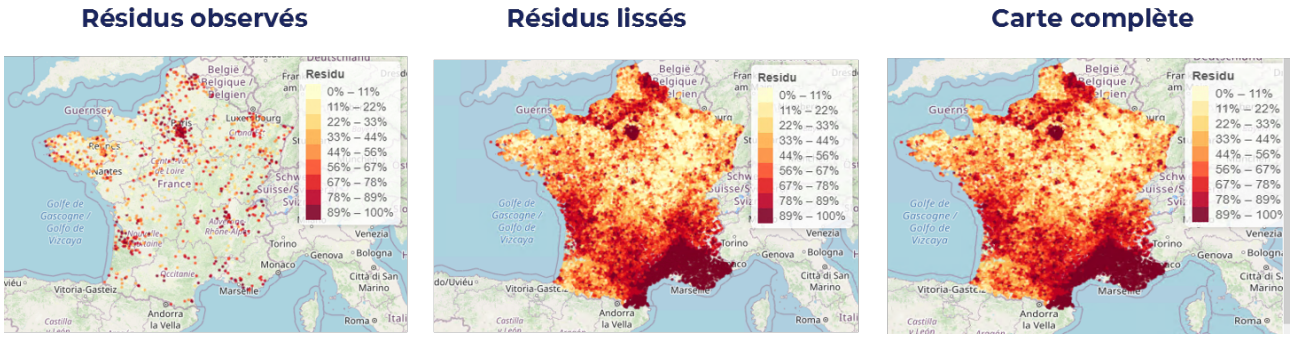


FIGURE 84 – Valeur des résidus (au sein des quantiles) en fonction de la zone géographique. Plus la couleur de la zone est foncée et plus on sous-estime le montant du tarif. Le zonier permet d'expliquer une partie de l'éloignement des résidus de la valeur 0. Le zonier est réputé comme primordial pour une tarification juste dans la région de Marseille.

C.2.3 Critères d'évaluation

Critères d'évaluation	Base d'apprentissage		Base de test	
	avant zonier	après zonier	avant zonier	après zonier
Gini	0.96	0.98	0.92	0.94
MAE	46.8	34.3	64.36	47.69
RMSE	85.11	68.84	140.57	82.48

Le zonier améliore le modèle que ce soit sur l'apprentissage ou la prédiction, ce qui conforte son incorporation. La moyenne des erreurs obtenue est égale à 2, le XGBoost intégrant des pénalisations qui empêchent la nullité de la perte moyenne. L'indice de Gini est bon, l'écart de tarif au sein des échantillons observés et prédits est de 6% seulement. Le MAE est faible compte tenu de la distribution du tarif leader.

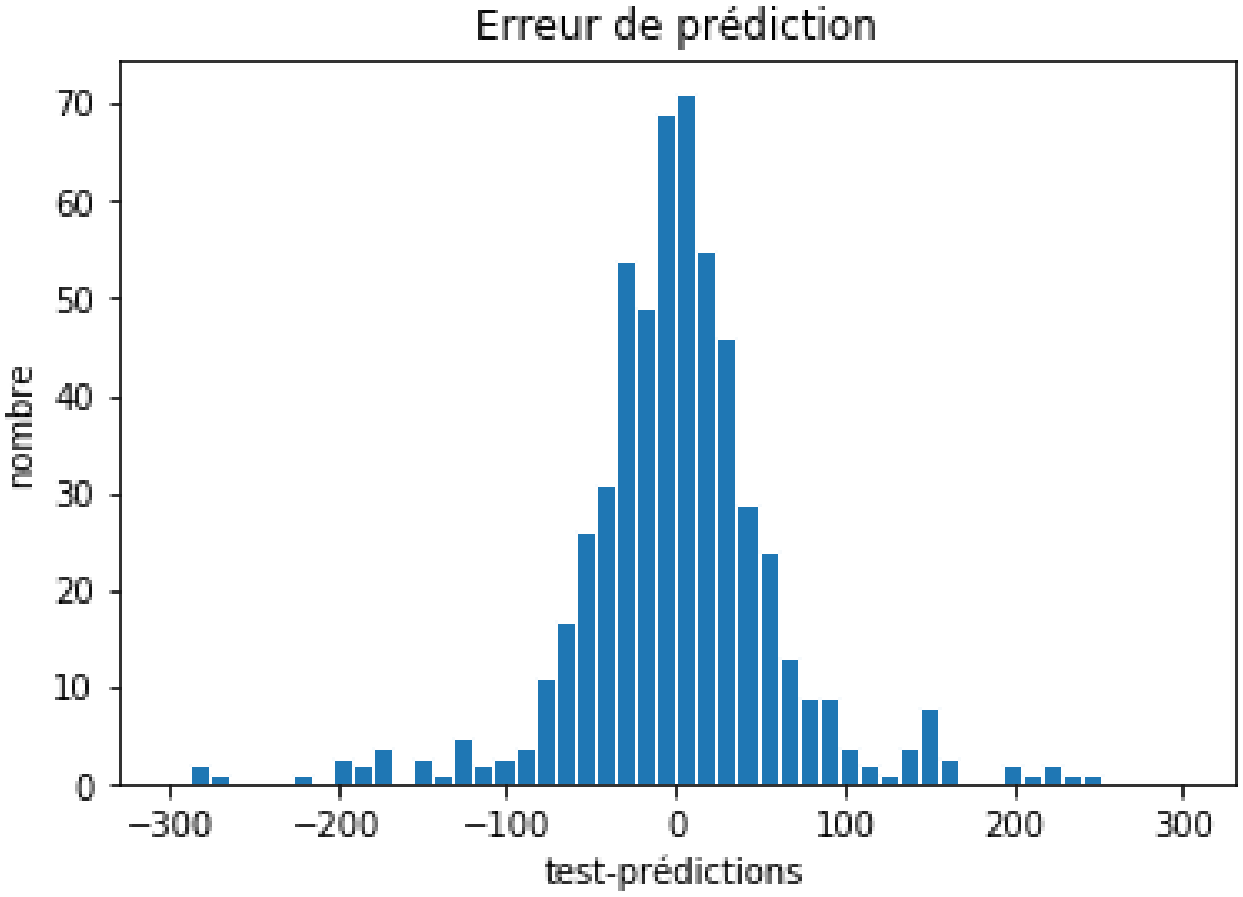


FIGURE 85 – Histogramme des erreurs sur la base de test (modèle incluant le zonier)

C.2.4 Interprétation du modèle

Pour comprendre la prédiction du modèle, des graphes d'importance sont réalisés. Selon la définition de l'importance, le classement des variables est modifié. Le principe *cover* caractérise l'importance d'une variable comme la couverture moyenne en termes de présence dans la

détermination d'un découpage. Le principe *gain* quant à lui, comptabilise l'importance comme la somme totale des gains en termes de réduction de la fonction de pertes à chaque découpage où le facteur est impliqué. L'ordre des facteurs selon leur importance est instable suivant la méthode empruntée.

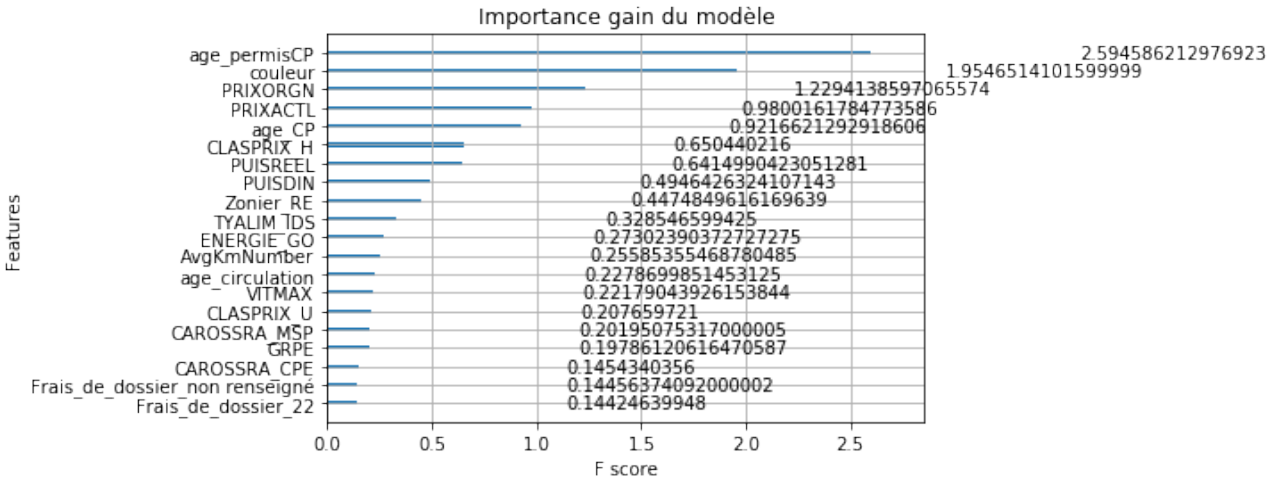


FIGURE 86 – Importance "gain" des 15 premières variables (modèle avec zonier)

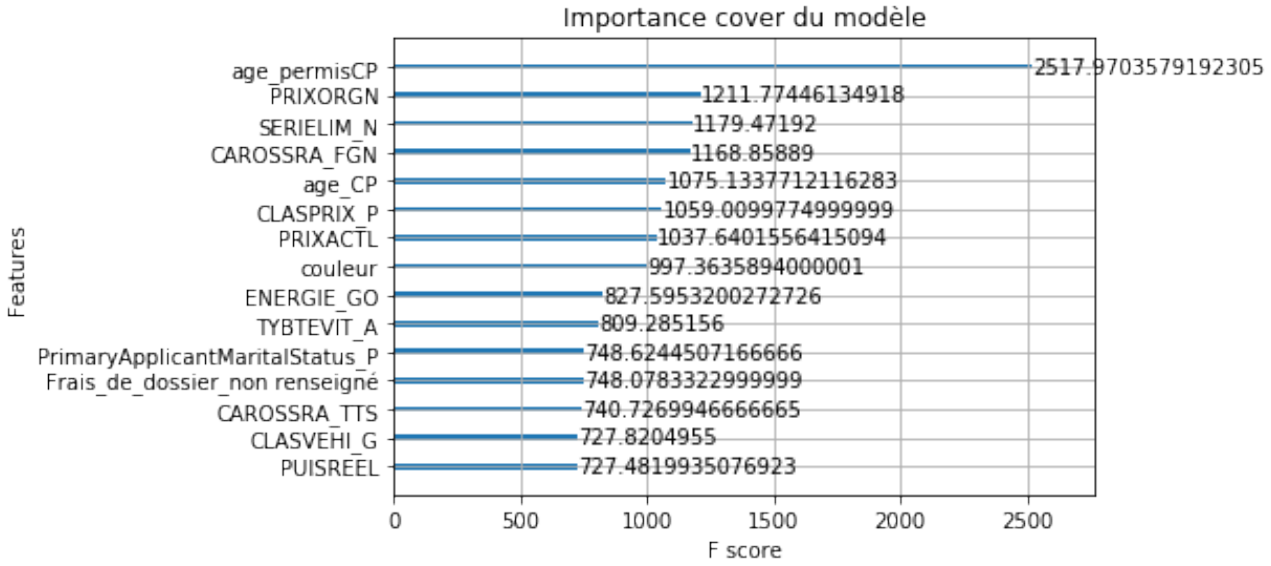


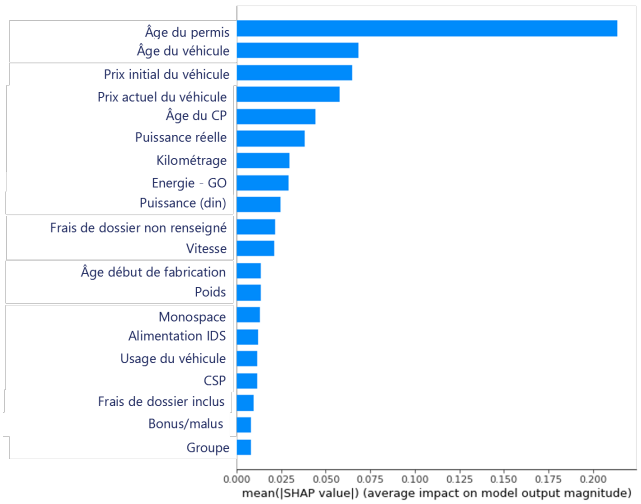
FIGURE 87 – Importance "cover" des 15 premières variables (modèle avec zonier)

Les valeurs de Shapley, appréciées pour leur consistence (contrairement aux méthodes *gain*, *cover* ou LIME), stabilisent l'ordre d'importance des variables. Elles font parties des approches additives (comme LIME) : la somme des contributions de chaque facteur intervenant pour une prédiction en particulier est égal à la valeur de la prédiction. La contribution prédictive se calcule alors localement et pourra être agrégée. Soit F l'ensemble des facteurs présents dans le modèle et soit k un facteur dont on souhaite calculer la contribution. On nomme par S un sous-ensemble quelconque de F ne contenant pas k. On note par $f_K(x_K)$ la prédiction des

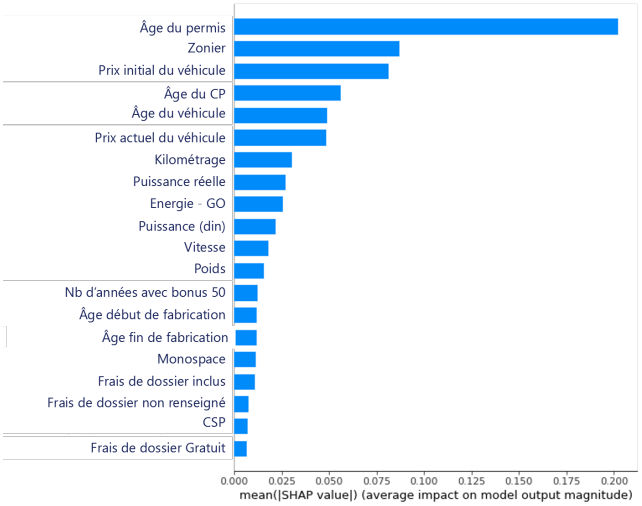
observations par le modèle entraîné avec l'ensemble de facteurs K.

$$\phi_k = \sum_{S \in F} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup k}(x_{S \cup k}) - f_S(x_S)] \tag{56}$$

Autrement dit, la contribution du facteur k est égale à la somme pondérée des contributions marginales de k dans chaque coalition S. La valeur de Shapley offre donc un poids marginal à chaque variable, toute chose égale par ailleurs. Ce concept est issu de la théorie des jeux. Des approximations des valeurs de Shapley existent pour réduire le coût algorithmique sous certaines hypothèses sur les covariables. Le lecteur intéressé pourra se référer à l'article en bas de page pour des approfondissements autour de la théorie sur le calcul de ces contributions [26].



Valeur de shapley pour les 20 premiers facteurs (modèle sans zonier)



Valeur de shapley pour les 20 premiers facteurs (modèle avec zonier)

L'effet marginal du *zonier*, sans bousculer réellement l'ordre d'importance général des autres facteurs, vient se hisser à la deuxième place du classement. Cela signifie que, toute chose égale par ailleurs, le zonier est la seconde variable discriminante au sein du modèle. Indétrônable peu importe la méthode de calcul d'importance ou l'introduction de données externes géographiques, l'âge du permis de conduire (*age_permisCP*) qui détermine le niveau d'expérience du conducteur. Cette variable intervient notamment dans les modèles de fréquence car un conducteur expérimenté provoque moins d'accidents (à l'exception des âges élevés). Ensuite le prix d'origine du véhicule (*PRIXORGN*) qui impacte directement le coût moyen lors d'une réparation ou d'un remplacement. L'âge du conducteur (*age_CP*) est corrélé avec l'âge du permis et sépare de la même manière les jeunes conducteurs, les conducteurs expérimentés et les grands âges. L'âge du véhicule (*age_circulation*) indique le niveau de technologie implémenté que ce soit pour une réduction de la probabilité d'occurrence de sinistre ou un coût moyen accru dû à la sophistication des pièces. S'ensuivent des facteurs explicatifs, qui traduisent eux aussi des comportements de l'assuré ou liés à l'usage de l'automobile, comme le nombre de kilomètres moyens annuels parcourus (*AvgKmNumber*), la vitesse maximale du véhicule (*VITMAX*), la puissance (*PUISREEL*), le poids du véhicule (*POIDSVID*), le bonus/malus (*PrimaryApplicantBonusCo-*

26. Lundberg et Lee, *A unified approach to interpreting model predictions*, 2017

eff), le montant des frais de dossier (*Frais_ de_ dossier*) ou la catégorie socio-professionnelle (*PrimaryApplicantOccupationCode*).

Dans la figure ci-dessous, sont représentées les contributions de Shapley par facteur explicatif. Chaque poids correspond à une observation, l'épaisseur indique alors le poids d'une valeur de Shapley, et la couleur traduit la valeur du facteur. Les contributions sont positives pour les jeunes conducteurs (*age_ permisCP*), ce qui signifie que le tarif leader est accru par des valeurs faibles de l'âge du permis. En revanche, des valeurs élevées de l'âge du permis orientent le tarif préférentiel à la baisse (contribution négative) dans nos prédictions. L'effet marginal du zonier (*Zonier_ RE*) est naturel dans son interprétation puisque dans le cas où le tarif leader prédit est en-deça du tarif leader observé (sous-tarification) alors la valeur de zonier est importante et participe à un accroissement du tarif leader prédit in fine (les points rouge sont dans la partie positive des contributions). La vitesse et la puissance du véhicule (*VITMAX* et *PUISREEL*), lorsqu'elles augmentent, font accroître le tarif leader et le font décroître à l'opposé, lorsqu'elles diminuent. Un véhicule bas de gamme (*PRIXORGN*) participe à la baisse du tarif minimal du marché tandis que celui-ci part à la hausse lorsque le véhicule monte en gamme. Le modèle est cohérent du point de vue de nos connaissances du marché et aucun élément ne constitue une contradiction avec nos croyances.

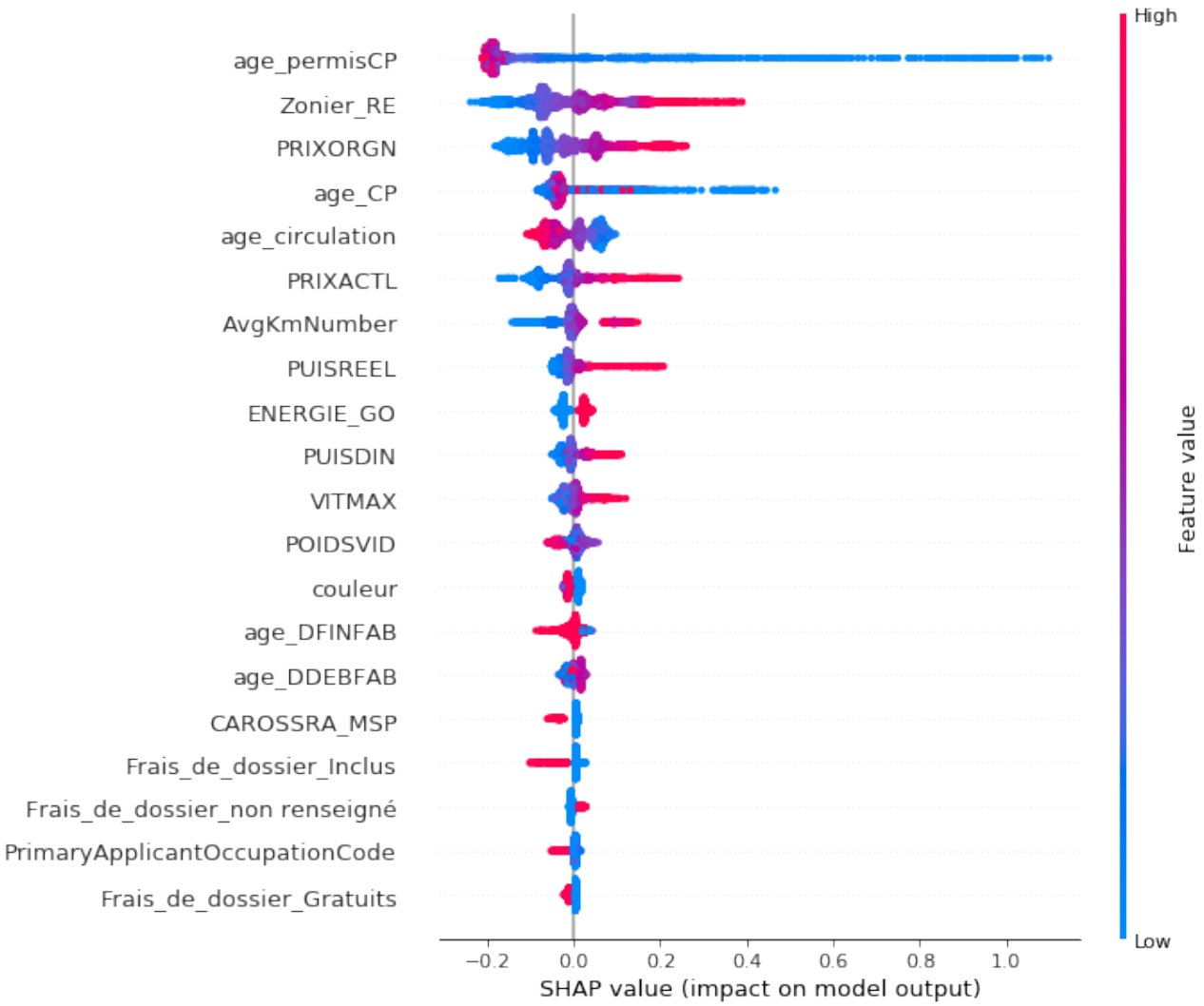


FIGURE 88 – Contributions de Shapley sur la valeur cible pour le modèle incluant le zonier.

Pour terminer l'allocation d'un prix concurrent aux observations de la base devis, un *mapping* est réalisé au sein de celle-ci. Le modèle de tarif leader ne comporte que les variables communes aux deux bases. La segmentation de ces variables doit être également compatible. Le regroupement des modalités ou, au contraire, leur dé-segmentation a dû être réfléchi et discuté avec l'assureur pour que les assurés de la base devis collent au mieux à leur nouvelle représentation.

7 Note de synthèse

7.1 Contexte et problématique

Les assureurs de taille modeste s'inscrivent généralement dans une optique d'accroissement du portefeuille d'assurés. La digitalisation croissante accorderait un poids plus impactant à la souscription en ligne qui rend la compétition entre acteurs de l'assurance plus ardue. Capter de nouvelles cibles conduit à une course effrénée aux techniques les plus innovantes en termes de critères collectés et de méthodes employées. Poser le tarif adéquat ne se résume pourtant pas à la mesure du risque caractérisée par la prime pure, mais doit tenir compte du positionnement de l'assureur sur le marché, auquel cas atteindre un bon niveau de marge sur l'ensemble du portefeuille n'est pas garanti. L'objectif de ce mémoire est de modéliser l'élasticité au prix individuelle avec le plus de précision possible afin de :

— mieux comprendre les comportements des clients

— mieux comprendre les différences tarifaires avec les concurrents présents sur le marché

— simuler les impacts financiers (Profit et chiffre d'affaires...) des changements tarifaires envisagés afin de ne pas dégrader la rentabilité avec des politiques tarifaires compétitives

La rentabilité globale correspond à la somme des marges individuelles des prospects qui sont effectivement acquises lors d'une conversion en affaire nouvelle. L'élasticité au prix informe l'assureur de la probabilité de transformation d'un prospect, à tarif et caractéristiques fixés :

$$Profit = \sum_{i=1}^{n} (P_i - S_i) * \hat{f}(P_i, X_i) \tag{57}$$

P_i, la variable endogène qui correspond au tarif payé par l'assuré i ; S_i, correspond à tous les frais et la charge sinistre de l'assuré i ; $\hat{f}(P_i, X_i)$, l'élasticité au prix de l'assuré i ; X_i, les caractéristiques de l'assuré i. Dans le cas où l'assureur souhaite optimiser son allocation de prix, le profit peut être optimisé par lagrangien en ajoutant une contrainte sur le volume minimal imposé comme objectif.

7.2 Jeu de données

La base regroupe l'ensemble des 165000 devis réalisés sur les années 2018 et 2019 comprenant les caractéristiques du conducteur (âge, profession, bonus/malus, adresse etc) et de son véhicule (prix, groupe SRA, vitesse, puissance, âge etc), du contrat (franchise, remises, options, tarif) et si, oui ou non, le devis s'est transformé en affaire nouvelle. Les seuls contrats considérés sont ceux des particuliers de la garantie « Tous risques» du produit automobile.

La modélisation de la réaction des souscripteurs face à la concurrence nécessite l'intégration des prix, à la fois de la concurrence et de l'assureur dans l'équation de demande. Une base marché, constituée de profils fictifs représentatifs du marché, a été créée dans le but de collecter des données tarifaires externes provenant de sites comparateurs. Le tarif leader (le tarif minimal rencontré pour chacun des profils fictifs) constitue un baromètre auquel l'assureur peut se comparer : il est incorporé au sein de la base. Parmi les données tarifaires, seront mentionnées :

— La prime pure : le coût annuel de l'assuré vu par l'assureur

— La prime hors taxe : la prime pure additionnée des frais

— Le tarif hors taxe : la prime hors taxe additionnée de la marge

— Le tarif TTC : le tarif hors taxe augmenté des taxes

— Le tarif TTC leader du marché (collecté en 2020), diminué de 3.4% pour les devis de 2019 et encore de 3.5% pour ceux de 2018.

La marge individuelle de chaque prospect définit la politique tarifaire qui lui est appliquée. C'est sur la marge que l'assureur peut jouer pour faire fluctuer son volume de portefeuille.

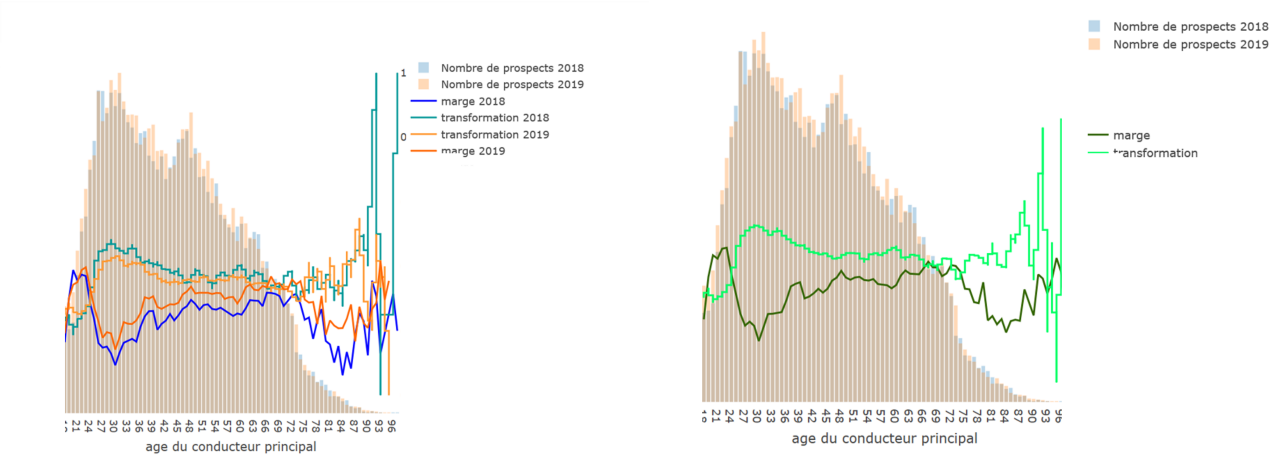


FIGURE 89 – Taux de transformation et politique tarifaire par segment d'âge : comment les fluctuations de la marge impactent la conversion des segments.

7.3 Modélisation du taux de transformation

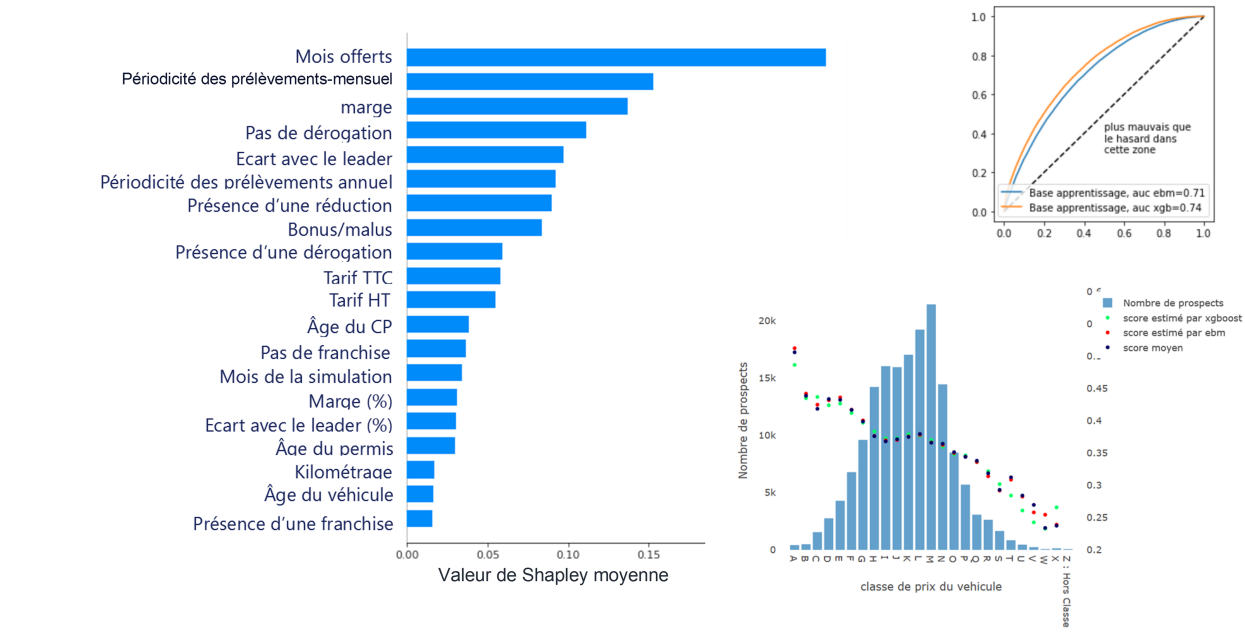
Est notée Y la variable qui décrit si l'assuré transforme son devis en souscription ou non.

Donc

$$Y = \begin{cases} 1 & si\ l'individu\ souscrit, \\ 0 & sinon. \end{cases} \tag{58}$$

Un modèle XGBoost a été choisi pour avoir le score de conversion par devis réalisé le plus précis possible. L'*Explainable Boosting Machine*, qui allie performance et interprétabilité, a également été testé pour juger de sa bonne qualité prédictive.

Critères	Performances sur la base d'apprentissage	
Modèle	XGBoost	EBM
Gini normalisé entre score estimé et variable-cible	47.8%	41.3%
Score moyen estimé de la base devis	37.4%	37.3%



Bien que l'EBM soit plus proche du taux de transformation moyen sur l'ensemble de la base, qui s'explique par les pénalisations inhérentes au XGBoost, les mesures d'erreur comme le Gini et la courbe ROC sont en faveur du XGBoost.
Le classement de l'importance des variables dans le modèle XGBoost de taux de transformation avec les valeurs de Shapley montre le fort impact des actions commerciales (mois offerts, dérogation, marge, tarifs...) et de la concurrence dans la conversion.

7.4 Problématique du biais de sélection

Le taux de transformation n'est pas suffisant pour estimer l'impact d'une nouvelle stratégie de tarif (augmentation ou diminution de la marge) sur le volume de portefeuille et sur les indicateurs financiers. En effet, la base devis servant à construire le modèle de taux de transformation ne prend en compte que les marges passées déjà appliquées pour chaque type de prospect. Se servir de ce modèle pour prédire un taux de conversion associé à une nouvelle marge et donc à un nouveau tarif conduit à un biais de sélection : les autres sujets ayant déjà bénéficié de ce nouveau tarif n'ont pas les mêmes caractéristiques que celui du prospect concerné, leur réaction face au prix ne peut donc être semblable.

Il s'agit de la différence entre taux de conversion et élasticité au prix, qui doit être estimée puisque nous sommes en absence d'A/B testing, c'est-à dire dans l'incapacité d'effectuer un price test pour connaître la réaction d'un client face à un nouveau prix. D'autres techniques doivent être employées pour pallier à ce manque de données expérimentales. La base comportera alors les devis fournis qui seront dupliqués pour agrémenter de nouvelles lignes de prospects où seule la marge (et donc le tarif HT et TTC) sera modifiée.

7.5 Principe du jumelage des sujets

Rosenbaun et Rubin (1983) proposent des algorithmes de jumelage par score de propension pour corriger le biais de sélection. La méthode, d'abord appliquée dans le cas d'évaluation de

traitements pharmaceutiques, emprunte son vocabulaire au corps médical. La première étape de cette technique consiste à former 2 groupes :

1. Le 1er groupe contient les prospects bénéficiant d'une marge m et à qui l'assureur voudrait connaître le taux de conversion s'il leur appliquait une marge m'. C'est le groupe de traitement.
2. Le 2nd groupe contient les prospects de la base devis ayant déjà une marge m'. Ils ont des profils différents de ceux du groupe de traitement. C'est le groupe de contrôle.

L'idée est d'apparier chaque sujet du groupe de traitement au sujet de contrôle qui lui "ressemble". Dans ce cadre, deux sujets se ressemblent s'ils ont la même probabilité conditionnelle d'assignation au traitement qui se définie comme :

$$\pi(X) = P(A|X) \tag{59}$$

avec A = "appartenir au groupe de traitement" et X les caractéristiques du sujet. L'estimation de π donne le score de propension.

Deux scores de propension similaires signifient même probabilité d'appartenir au groupe de traitement. Un prospect du groupe de traitement et son jumeau possèdent donc un support commun de caractéristiques qui crée un pont entre les deux stratégies : le sujet de contrôle ayant pu appartenir à la même stratégie que le sujet de traitement, le taux de transformation qui lui est associé pour la marge m' peut être attribué à ce dernier.
Deux conditions sont à respecter pour attester de l'atténuation du biais :

1. Il faut que certains sujets du groupe de contrôle puissent ressembler aux profils du groupe de traitement, autrement le jumelage n'a pas de sens : c'est la propriété de chevauchement des données.
2. Il faut que le modèle qui estime le score de propension soit performant pour garantir une bonne association : c'est la propriété de non-confusion.

Le protocole général est le suivant :

1. Estimer de manière la plus exacte possible la probabilité de transformation sur chaque ligne de la base de devis. Il s'agit de créer une série de taux quasiment observés qui seront utilisés lors du jumelage et estimés pour la résolution de l'équation d'optimisation (étape déjà réalisée).
2. Regrouper les sujets de la base devis en fonction de la politique tarifaire qui leur est appliquée.
3. Jumeler les sujets entre ces stratégies pour connaître l'élasticité-prix individuelle de chaque prospect si on lui appliquait un autre tarif
4. Modéliser le score de conversion avec ces nouvelles données.

7.6 Constitution des groupes de politique tarifaire

Dans un premier temps, il convient de regrouper les prospects qui ont la même politique tarifaire afin de mieux comprendre les stratégies appliquées par l'assureur (par l'étude des

profils de chaque politique) et de mieux maîtriser la bascule d'un profil sur une autre politique tarifaire. Afin d'établir des typologies de stratégies, la segmentation de la marge calculée se révèle indispensable pour restreindre le nombre de niveaux.

$$marge\ en\ \% = (Cotisations\ HT\ -\ (Prime\ pure\ +\ frais))/Cotisations\ HT \tag{60}$$

	Minimum (marge)	Maximum (marge)
Groupe 1-2	-2.42	-0.05
Groupe 3	-0.04	0.1
Groupe 4	0.11	0.22
Groupe 5	0.23	0.3
Groupe 6	0.31	0.42
Groupe 7	0.43	0.59
Groupe 8	0.6	0.9

Grâce à un algorithme de classification ascendante hiérarchique, 7 sous-échantillons de sujets homogènes en termes de marge ont été construits.

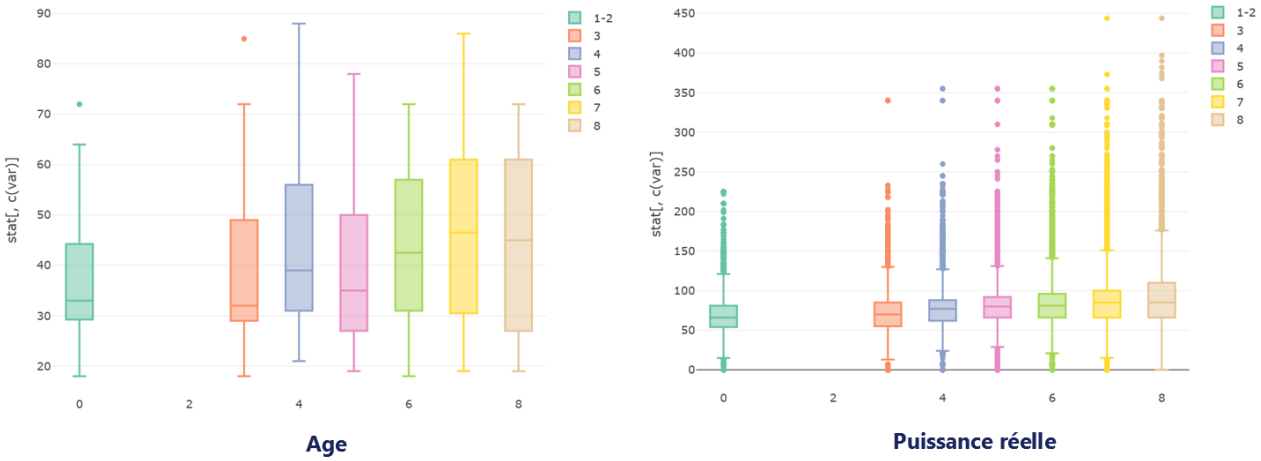


FIGURE 90 – Âges et puissance réelle du véhicule par groupe de stratégie. Lorsque les groupes de stratégie sont comparés, les moyennes et les dispersions ne sont pas les mêmes à travers l'ensemble des caractéristiques. En revanche, la propriété de chevauchement des données semble être respectée : sur chaque variable, tous les individus à l'exception de quelques extrêmes peuvent trouver un assuré d'un autre groupe tarifaire ayant la même valeur pour cette variable.

7.7 Jumelage par score de propension

L'algorithme fonctionne comme suit, sachant que l'on souhaite déterminer pour chaque sujet de la stratégie m $\in \{1-2,3,4,5,6,7,8\}$ son score de conversion s'il appartenait à la stratégie m' $\in \{1-2,3,4,5,6,7,8\}\backslash\{m\}$ (7*6 = 42 appariements à effectuer).

1. Estimer le score de propension π pour chaque individu des groupes m et m'.
2. Choisir le premier sujet du groupe m et lui associer le sujet du groupe m' qui a le score de propension le plus proche en valeur absolue.
3. Affecter la marge et le score de conversion du jumeau à ce premier sujet.

4. Recommencer la deuxième étape pour le second sujet du groupe etc.

Il est important de mentionner pour la compréhension de l'algorithme que deux sujets du groupe de traitement peuvent avoir le même jumeau. Ainsi, les sujets du groupe de contrôle qui ressemblent davantage à ceux du groupe de traitement seront davantage sollicités pour le jumelage tandis que les individus éloignés en termes de caractéristiques seront écartés : c'est ainsi que le biais est corrigé, en associant deux individus interchangeables selon le score de propension.Trois méthodes de jumelage par score de propension sont considérées :

— le jumelage par score de propension estimé par GLM

— le jumelage par score de propension estimé par XGBoost, qui confère une meilleure précision dans l'estimation de la probabilité d'appartenance à un groupe et améliore le jumelage : il s'agit de la méthode conservée

— le genetic matching, qui sélectionne certains individus de contrôle dont le score de propension (estimé par XGBoost) est suffisamment proche puis est choisi le jumeau en fonction de la corrélation de ses caractéristiques avec le sujet de traitement. Cette méthode est abandonnée car sa performance dépend du déploiement de ressources de calcul importantes.

Au sein de l'échantillon apparié, la mise en place d'une stratégie est indépendante des caractéristiques. Les jumeaux sont interchangeables car ils ont la même probabilité d'être affecté au traitement. Pour mesurer la qualité des appariements effectués grâce aux trois méthodes, les fonctions de répartition des caractéristiques sont étudiées pour le groupe de traitement, le groupe de contrôle, le groupe de contrôle après jumelage. Pour garantir une diminution du biais, les fonctions de répartition du groupe de contrôle après jumelage doivent être plus proches des fonctions de répartition du groupe de traitement que ne l'étaient les fonctions de répartition du groupe de contrôle initialement.

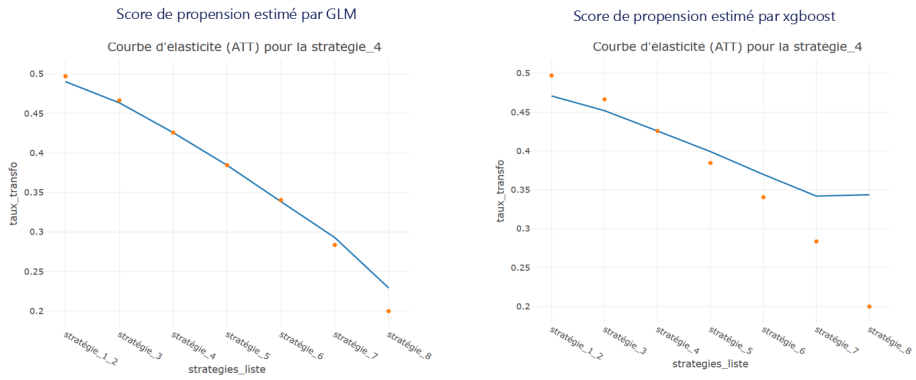


FIGURE 91 – En orange le taux de transformation moyen pour chaque stratégie et en bleu le taux de transformation moyen de la stratégie 4 si elle basculait sur une autre stratégie. Le GLM ne distingue pas les mécanismes d'assignation au traitement et considère que l'élasticité au prix est la même peu importe le profil considéré, pour toutes les stratégies. Le XGBoost lui, a épuré les sujets de contrôle puisque les moyennes de taux de transformation du groupe de contrôle ont été modifiées après le jumelage : il est judicieux d'augmenter la marge puisque la transformation est plus forte que celle de ceux bénéficiant déjà d'une marge élevée. En revanche rogner sur la marge ne permet pas d'accroître le portefeuille de la même façon que pour les personnes qui bénéficient d'un tarif plus compétitif.

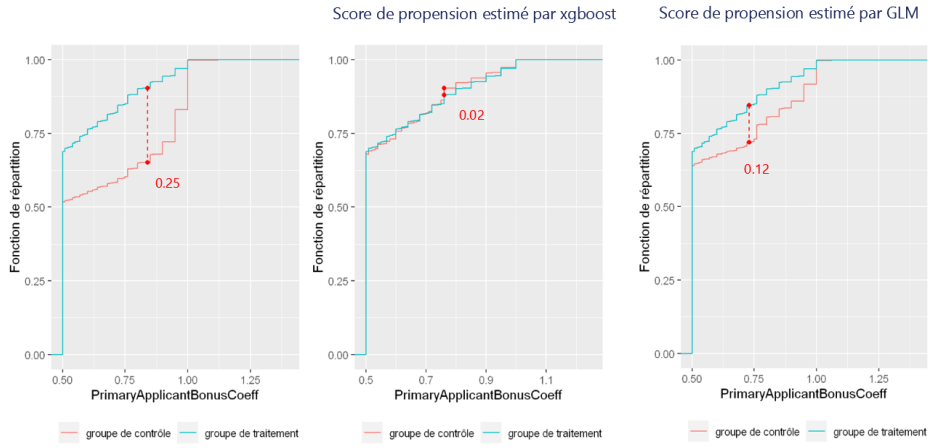


FIGURE 92 – Traitement : stratégie 6/ Contrôle : stratégie 8. A l'origine, l'écart entre les deux groupes sur la variable bonus/ malus est important. Le XGBoost obtient d'excellents résultats, compte tenu de l'écart de départ.

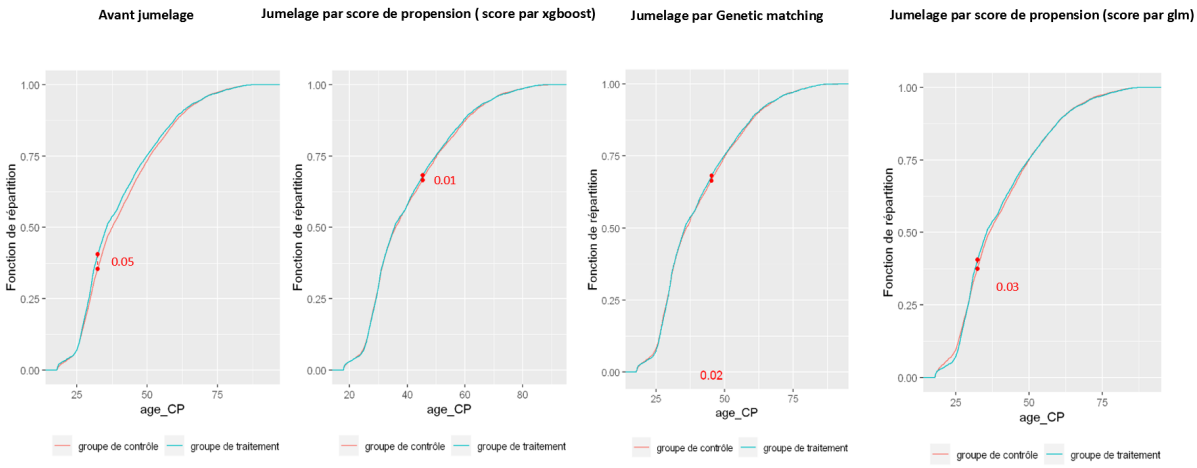


FIGURE 93 – Traitement : stratégie 1-2/ Contrôle : stratégie 3.Comparaison des différentes méthodes sur l'âge du conducteur. Le *genetic matching* a été réalisé avec une taille de population de seulement 2000 sujets et a nécessité de nombreuses heures de calcul uniquement sur ces deux groupes, ce qui explique la limite de ses résultats au regard des 2 autres méthodes.

7.8 Modélisation de l'élasticité au prix

La base initiale a été alimentée de 6 copies de chacun des devis, égales en toutes variables excepté la marge, les tarifs HT et TTC et le taux de transformation qui est celui d'un jumeau. Au cours de la modélisation de l'élasticité au prix, des erreurs plus importantes ont été constatées pour les devis s'éloignant davantage du tarif d'origine. Une évolution de 20% maximum autour de la marge d'origine est permise au sein de la base d'apprentissage parmi les devis rajoutés par jumelage. Au final, la base d'apprentissage a été augmentée de 22%

Dans le cas où l'assureur souhaite réaliser une optimisation tarifaire, le taux de transformation estimé doit être exprimé de manière explicite en fonction du prix et dérivable en celui-ci. L'*Explainable Boosting Machine* respecte la première condition de part les propriétés de modu-

larité et de simulabilité. En revanche, la seconde propriété est moins naturelle à l'algorithme. C'est pour cette raison que le GLM a été choisi (en laissant donc la variable de tarif HT continue) en dépit de ses moins bonnes performances prédictives.

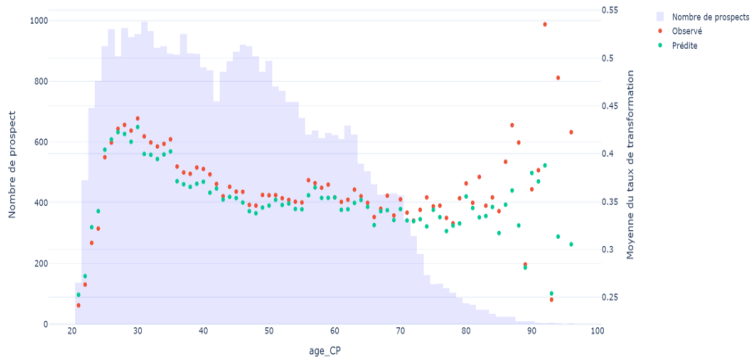
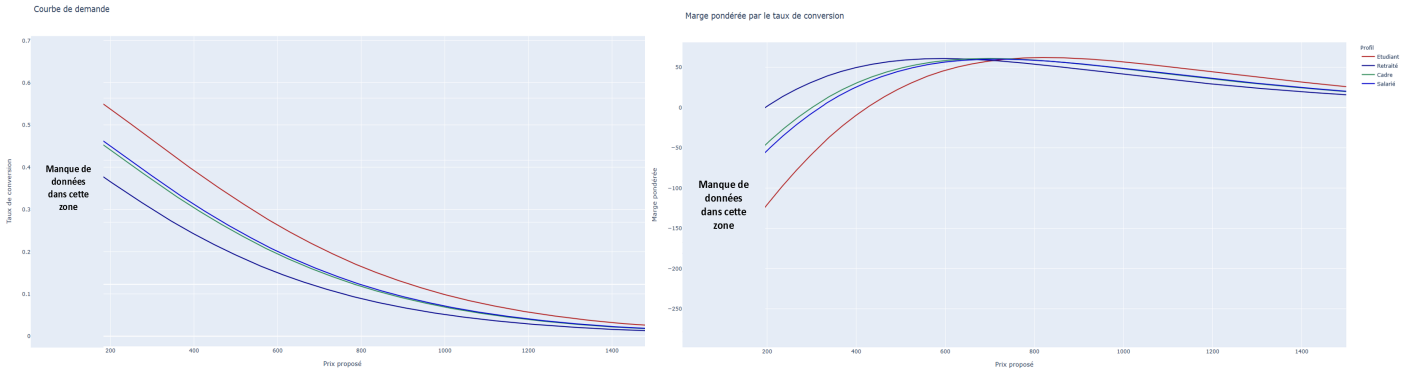


FIGURE 94 – Comparaison des taux observés et prédits (GLM) sur la base de test pour l'âge

Sont présentées les courbes de demande et les marges pondérées par le taux de transformation estimé par type de profession, les autres caractéristiques étant figées. Le point le plus haut indique le prix à proposer pour accroître la marge sur un segment.



Pour conclure, l'une des principales difficultés pour l'activité d'assurance est d'appréhender les décisions prises par les assurés. Ce travail a alors tenté d'apporter une pierre à l'édifice, puisqu'il fournit la connaissance des facteurs qui influencent la décision de souscrire à un contrat d'assurance. Pour aller plus loin, l'élasticité-prix a été estimée grâce au jumelage de données par score de propension. D'autres techniques existent comme le genetic matching mais n'ont pu aboutir faute de ressources. Grâce aux courbes de demande construites, le tarif peut être optimisé pour une cible en particulier, là où l'assureur n'est pas rentable ou au contraire, pas assez compétitif. Les facteurs financiers futurs comme le chiffre d'affaires, le loss ratio ou la rentabilité peuvent être simulés après le changement de politique tarifaire sur la cible à partir des nouveaux prix et de la demande future (prédiction du modèle).
L'exactitude de cette simulation repose sur la robustesse du modèle d'élasticité. Le price test a pu être reproduit mais seulement autour d'un seuil de 20% d'évolution de la marge à l'intérieur duquel les erreurs sont contenues. Des modèles de machine learning plus performants (EBM) et interprétables ont également été testés dans cette optique.
Enfin, l'estimation de l'élasticité-prix permet la résolution de l'équation d'optimisation, qui peut être complétée par un modèle de frais et l'estimation de la durée de rétention.

8 Executive summary

8.1 Context and problem

Smaller insurers are generally looking to increase their portfolio of insureds. Increasing digitalization is giving more impact to online underwriting, which makes the competition between insurance players more difficult. Capturing new targets leads to a frantic race for the most innovative techniques in terms of criteria collected and methods used. However, setting the right rate is not only a matter of measuring the risk in terms of the pure premium, but must also take into account the insurer's market positioning, in which case achieving a good margin level on the entire portfolio is not guaranteed. The objective of this paper is to model the individual price elasticity as accurately as possible in order to :

— better understand customer behavior

— better understand the price differences with competitors present on the market

— simulate the financial impacts (profit and turnover...) of the planned tariff changes in order not to degrade profitability with competitive tariff policies

Overall profitability is the sum of individual prospect margins that are effectively earned from a conversion to new business. The price elasticity informs the insurer of the probability of a prospect's conversion, at a fixed rate and characteristics :

$$Profit = \sum_{i=1}^{n} (P_i - S_i) * \hat{f}(P_i, X_i) \tag{61}$$

P_i, the endogenous variable that corresponds to the rate paid by insured i ; S_i, corresponds to all the expenses and the claims burden of insured i ; $\hat{f}(P_i, X_i)$, the price elasticity of insured i ; X_i, the characteristics of insured i. In the case where the insurer wishes to optimize its price allocation, the profit can be optimized by Lagrangian by adding a constraint on the minimum volume imposed as a target.

8.2 Data set

The database includes all 165,000 quotations made in 2018 and 2019, including the characteristics of the driver (age, profession, bonus/malus, address, etc.) and his vehicle (price, SRA group, speed, power, age, etc.), the contract (deductible, discounts, options, rate) and whether or not the quotation has turned into a new business. The only contracts considered are those for private individuals under the "All Risks" guarantee of the automobile product.

Modeling the reaction of policyholders to competition requires the integration of both the competitor's and the insurer's prices into the demand equation. A market database, made up of fictitious profiles representative of the market, was created in order to collect external pricing data from comparison sites. The leading rate (the minimum rate encountered for each of the fictitious profiles) constitutes a barometer against which the insurer can compare itself : it is incorporated into the database. Among the tariff data, the following will be mentioned :

— The pure premium : the annual cost of the insured seen by the insurer

— The premium before tax : the pure premium added to the expenses

— The pre-tax rate : the pre-tax premium plus the margin

— Rate including tax : the rate excluding tax plus taxes

— The market-leading rate including tax (collected in 2020), reduced by 3.4% for 2019 quotes and by a further 3.5% for 2018 quotes.

The individual margin of each prospect defines the pricing policy applied to him. It is on the margin that the insurer can play to fluctuate its portfolio volume.

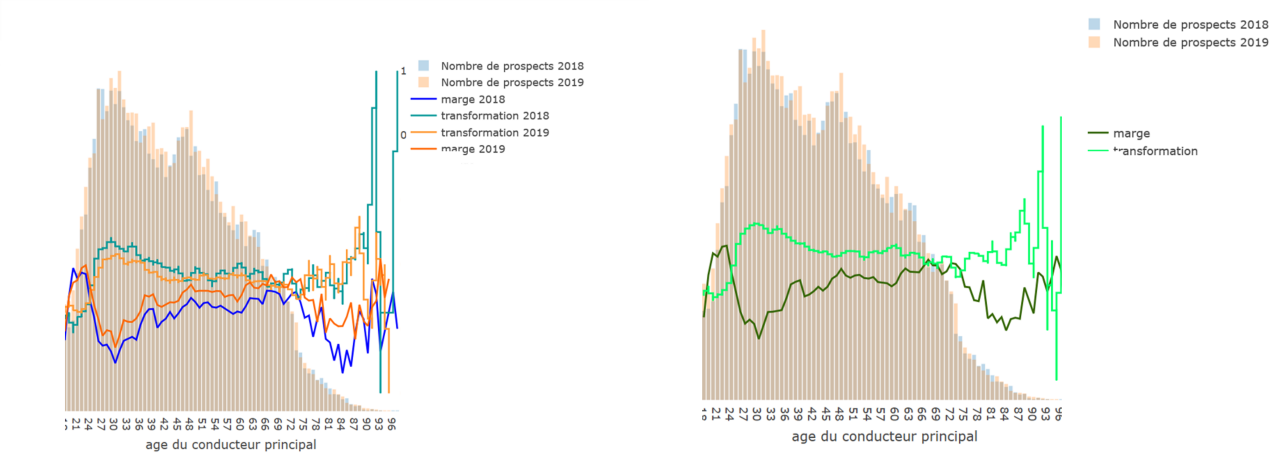


FIGURE 95 – Transformation rates and pricing policy by age segment : how margin fluctuations impact segment conversion.

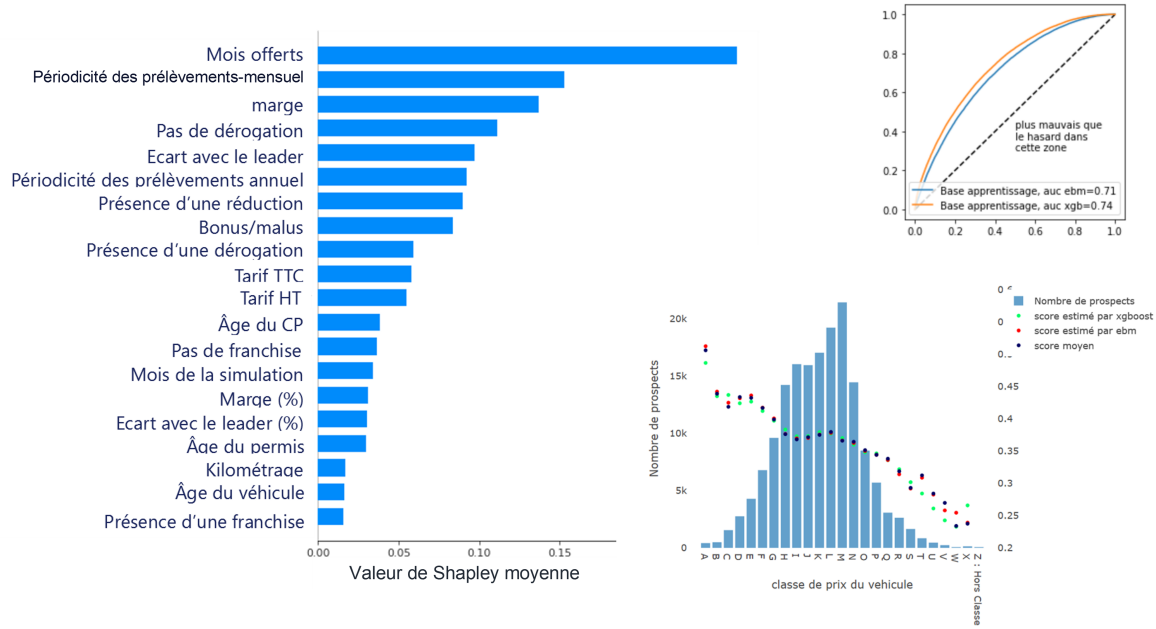
8.3 Modeling the transformation rate and the problem of selection bias

Let us note Y the variable which describes if the insured transforms his quote into a subscription or not.
Therefore

$$Y = \begin{cases} 1 & if\ individual\ subscribes, \\ & otherwise. \end{cases} \tag{62}$$

An XGBoost model was chosen to have the most accurate conversion score per quote. The *Explainable Boosting Machine*, which combines performance and interpretability, was also tested to judge its good predictive quality.
An XGBoost model was chosen to have the most accurate conversion score per quote. The Explainable Boosting Machine, which combines performance and interpretability, was also tested to judge its good predictive quality.

Criterias	Performance on the learning base	
Model	XGBoost	EE
Standardized Gini between estimated score and target variable	47.8%	41.
Estimated average score of the quotation base	37.4%	37.



Although the MVE is closer to the average transformation rate across the entire base, due to the penalties inherent in XGBoost, error measures such as the Gini and the ROC curve are in favor of XGBoost. Ranking the importance of variables in the XGBoost transformation rate model with Shapley values shows the strong impact of sales actions (months offered, waivers, margins, rates, etc.) and competition on conversion.

The transformation rate is not sufficient to estimate the impact of a new pricing strategy (increase or decrease in margin) on the portfolio volume and on the financial indicators. Indeed, the quotation base used to build the transformation rate model only takes into account the past margins already applied for each type of prospect. Using this model to predict a conversion rate associated with a new margin and therefore a new price leads to a selection bias : the other subjects who have already benefited from this new price do not have the same characteristics as the prospect in question, so their reaction to the price cannot be similar.

This is the difference between the conversion rate and the price elasticity, which must be estimated since we have no A/B testing, i.e. we are unable to carry out a price test to determine a customer's reaction to a new price. Other techniques must be used to compensate for this lack of experimental data. The database will then contain the quotes provided, which will be duplicated in order to create new lines of leads where only the margin (and therefore the price excluding tax and VAT) will be modified.

8.4 Principle of subject matching

Rosenbaun and Rubin (1983) propose propensity score matching algorithms to correct selection bias. The method, which was first applied to the evaluation of pharmaceutical treatments, borrows its vocabulary from the medical profession. The first step of this technique consists in forming 2 groups :

1. The first group contains the prospects with an m-margin and for whom the insurer would like to know the conversion rate if it applied an m'-margin. This is the treatment group.
2. The 2nd group contains the prospects from the quotation base who already have an m' margin. They have different profiles from those of the treatment group. This is the control group.

The idea is to match each subject in the treatment group to the control subject who "resembles" him. In this framework, two subjects are similar if they have the same conditional probability of assignment to the treatment which is defined as :

$$\pi(X) = P(A|X) \tag{63}$$

with A = "belonging to the treatment group" and X the characteristics of the subject. The estimation of π gives the propensity score.
Two similar propensity scores mean the same probability of belonging to the treatment group. A prospect in the treatment group and his twin therefore have a common support of characteristics which creates a bridge between the two strategies : the control subject having been able to belong to the same strategy as the treatment subject, the transformation rate associated with him for the margin m' can be attributed to the latter. Two conditions have to be met in order to attest to the mitigation of the bias :

1. Some subjects in the control group must be able to resemble the profiles in the treatment group, otherwise the matching is meaningless : this is the data overlap property.
2. The model that estimates the propensity score must perform well to guarantee a good association : this is the non-confusion property.

The protocol to be followed is as follows

as accurately as possible the probability of transformation on each line of the quotation base. This involves creating a series of quasi-observed rates which will be used during the matching and estimated for the resolution of the optimization equation (step already performed). the subjects of the quote base according to the rate policy applied to them. the subjects between these strategies to find out the individual price elasticity of each prospect if another price was applied to it the conversion score with this new data.

8.5 Constituting the pricing policy groups

First of all, it is advisable to group together the prospects who have the same pricing policy in order to better understand the strategies applied by the insurer (by studying the profiles of each policy) and to better control the switch of a profile to another pricing policy. In order

to establish typologies of strategies, the segmentation of the calculated margin proves to be essential in order to restrict the number of levels.

$$margin\ en\ \% = (Dues\ taxes\ excluded - (Pure\ premium + fees))/Dues\ taxes\ excluded\ taxes \tag{64}$$

Using a hierarchical bottom-up clustering algorithm, 7 subsamples of subjects homogeneous in terms of margin were constructed.

	Minimum (margin)	Maximum (margin)
Group 1-2	-2.42	-0.05
Group 3	-0.04	0.1
Group 4	0.11	0.22
Group 5	0.23	0.3
Group 6	0.31	0.42
Group 7	0.43	0.59
Group 8	0.6	0.9

When studying the characteristics of the different sub-groups of strategies, the first remark must be that there are similarities : on each variable, all individuals except a few extremes can find an insured from another tariff group with the same value for this variable. On the other hand, the means and proportions are not the same for all characteristics.

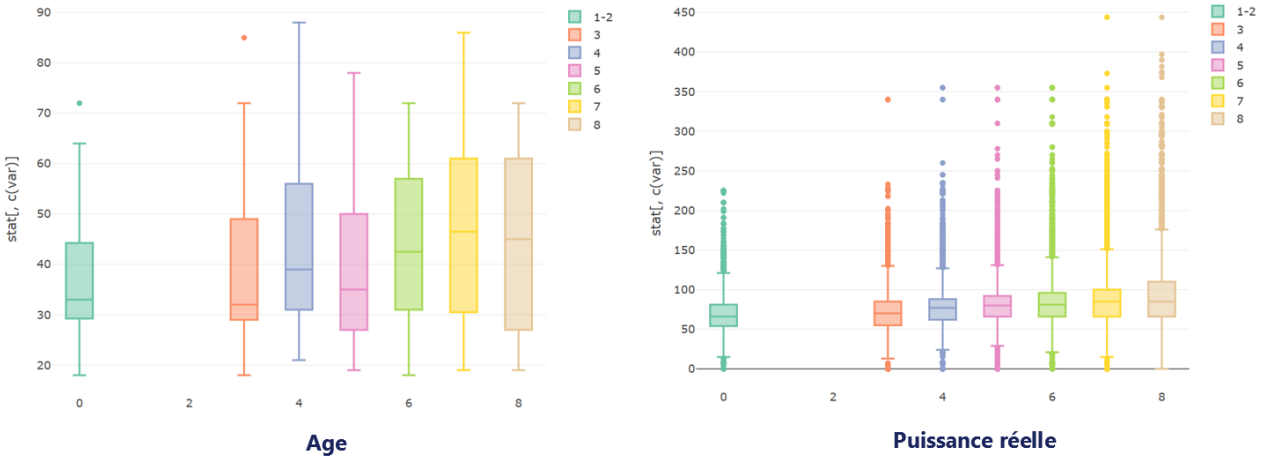


FIGURE 96 – Ages and actual vehicle power by strategy group. When strategy groups are compared, the means and dispersions are not the same across all characteristics. On the other hand, the overlap property of the data appears to be satisfied : on each variable, all but a few extreme individuals can find an insured from another rate group with the same value for that variable.

8.6 Matching by propensity score

The algorithm works as follows, knowing that we wish to determine for each subject of strategy m $\in \{1-2,3,4,5,6,7,8\}$ its conversion score if it belonged to strategy m' $\in \{1-$

$2, 3, 4, 5, 6, 7, 8\} - \{m\}$.

1. Estimate the propensity score π for each individual in groups m and m'.
2. the first subject of the group m and associate him the subject of the group m' which has the closest propensity score in absolute value.
3. Assign the margin and the conversion score of the twin to this first subject.
4. Repeat the second step for the second subject of the group etc.

It is important to mention for the understanding of the algorithm that two subjects in the treatment group can have the same twin. Thus, the subjects of the control group who are more similar to those of the treatment group will be more solicited for the matching while the individuals who are distant in terms of characteristics will be discarded : this is how the bias is corrected, by matching two interchangeable individuals according to the propensity score.
Three methods of matching by propensity score are considered :

— propensity score matching estimated by GLM

— propensity score matching estimated by XGBoost, which gives a better precision in the estimation of the probability of belonging to a group and improves the matching : it is the retained method

— genetic matching, which selects control individuals whose propensity score (estimated by XGBoost) is sufficiently close and the twin is chosen according to its characteristics correlation with the treatment subject. This method is abandoned because its performance depends on the deployment of significant computing resources.

Within the matched sample, strategy implementation is independent of characteristics. Twins are interchangeable because they have the same probability of being assigned to the treatment. To measure the quality of the matches made with the three methods, the distribution functions of the characteristics are studied for the treatment group, the control group, the post-match control group. To ensure that bias is reduced, the post-match control group distribution functions should be closer to the treatment group distribution functions than the control group distribution functions were initially.

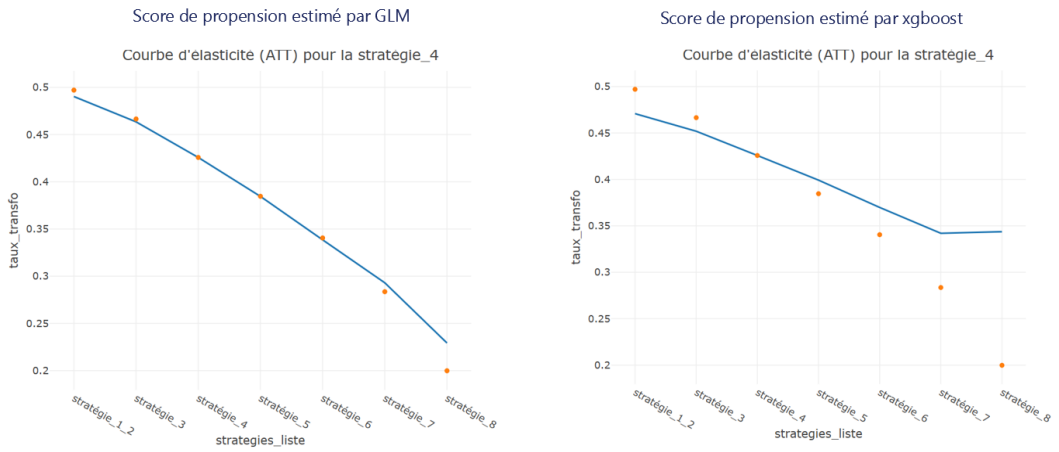


FIGURE 97 – In orange the average transformation rate for each strategy and in blue the average transformation rate of strategy 4 if it switched to another strategy. The GLM does not distinguish between the mechanisms of assignment to the treatment and considers that the price elasticity is the same regardless of the profile considered, for all strategies. XGBoost, on the other hand, has cleaned up the control subjects since the transformation rate averages of the control group have been modified after the matching : it makes sense to increase the margin since the transformation is higher than that of those already benefiting from a high margin. On the other hand, cutting the margin does not increase the portfolio in the same way as for those with a more competitive rate.

.

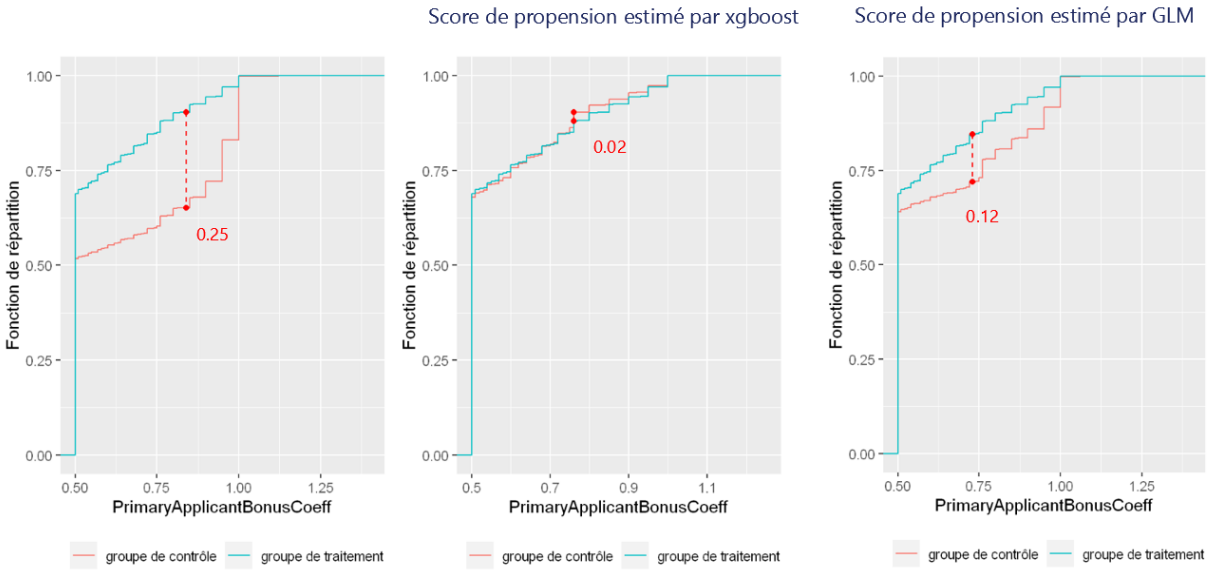


FIGURE 98 – Treatment : strategy 6/ Control : strategy 8. Initially, the gap between the two groups on the bonus/malus variable is large. The XGBoost obtains excellent results, considering the initial gap

.

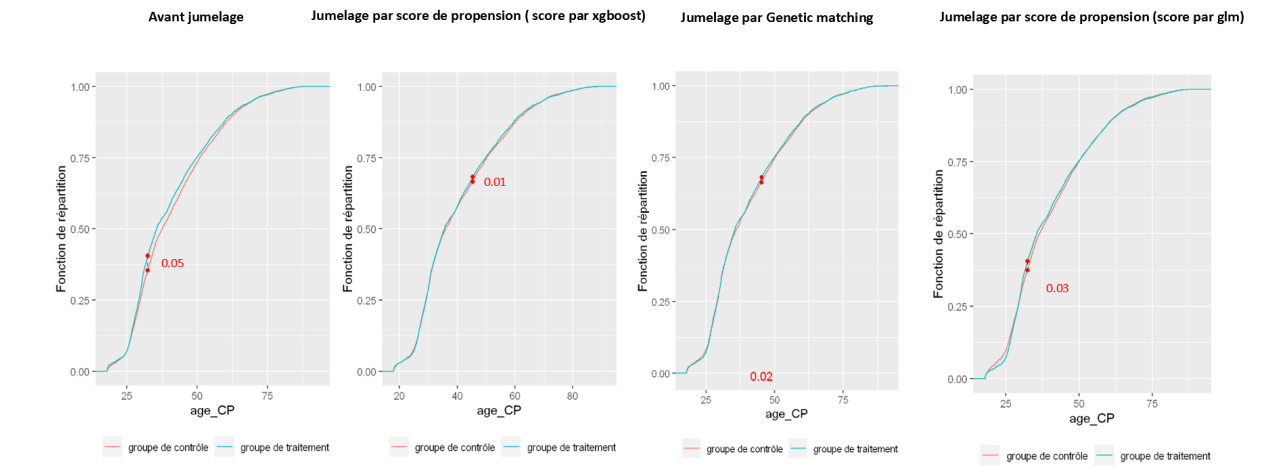


FIGURE 99 – Treatment : strategy 1-2/ Control : strategy 3.Comparison of the different methods on the driver's age. The *genetic matching* was performed with a population size of only 2000 subjects and required many hours of computation only on these two groups, which explains the limitation of its results compared to the 2 other methods.

.

8.7 Modeling the conversion score with the new data

The initial base was populated with 6 copies of each of the quotes, equal in all variables except the margin, the rates before and after tax and the transformation rate which is that of a twin. During the price elasticity modeling, larger errors were found for quotes that were further away from the original price. A maximum evolution of 20% around the original margin is allowed within the learning base among the quotes added by twinning. In the end, the learning base was increased by 22%.

In the case where the insurer wishes to carry out a price optimization, the estimated transformation rate must be expressed explicitly as a function of the price and derivable from it. The Explainable Boosting Machine respects the first condition because of its modularity and simulability properties. On the other hand, the second property is less natural to the algorithm. For this reason, the GLM was chosen (leaving the HT rate variable continuous) in spite of its less good predictive performance.

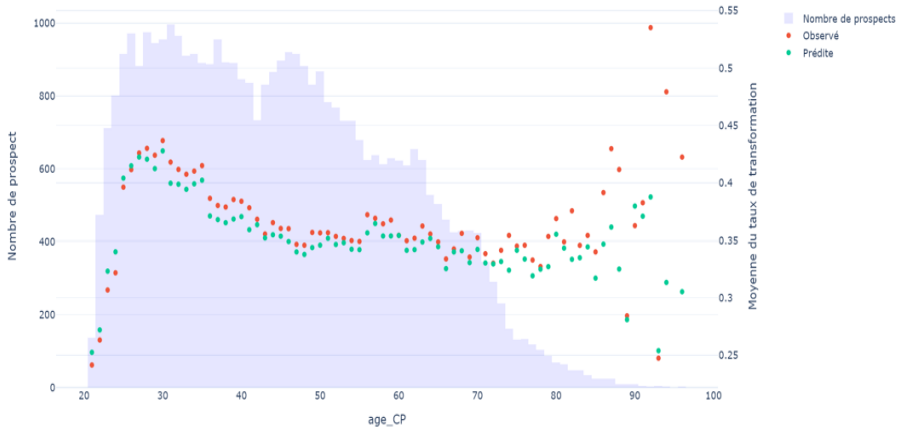
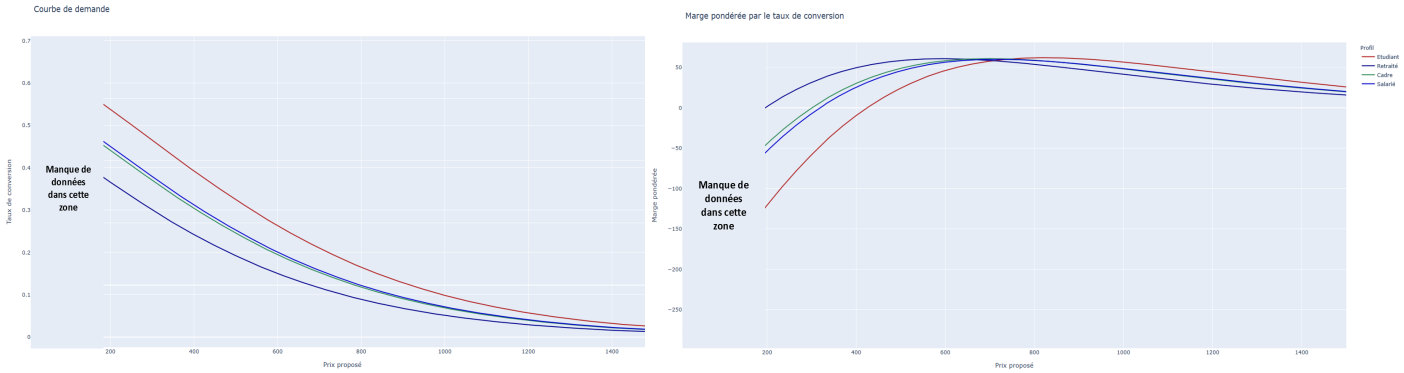


FIGURE 100 – Comparison of observed and predicted rates based on test for driver age

Demand curves and margins weighted by the estimated transformation rate by type of occupation are presented, with other characteristics frozen. The highest point indicates the price to be offered to increase the margin in a segment.



In conclusion, one of the main difficulties for the insurance business is to understand the decisions made by the insured. This work has therefore attempted to contribute to this problem by providing knowledge of the factors that influence the decision to subscribe to an insurance contract. The importance of justifying the results obtained was at the heart of the work with the use of interpretable algorithms or Shapley values. The variations of the demand are explained in particular by the competing prices and the commercial actions. Thanks to the demand curves constructed, the tariff can be optimized for a particular target, where the insurer is not profitable or on the contrary, not competitive enough. Future financial factors such as turnover, loss ratio or profitability can be simulated after the change of tariff policy on the target from the new prices and future demand (model prediction). The accuracy of this simulation relies on the robustness of the elasticity model. The price test could be reproduced around a certain threshold of margin evolution within which the errors are contained. More powerful and interpretable machine learning models (EBM) have also been tested in this respect.

Finally, the estimation of the price elasticity allows the solution of the optimization equation, which can be completed with a cost model and the estimation of the retention period.