Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 18/03/2021

Par : **Merieme AMINE**

Titre : **Orientation et réparation en garages partenaires en assurance automobile**

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
Caroline HILLAIRET

Entreprise : AXA France

Signature :

Membres présents du jury de l'Institut des Actuaires

Directeur du mémoire en entreprise :

Nom : Edouard VICAIRE
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

Résumé

Mots-clés : garages agréés, orientation, réparation, sévérité, Bagging, Boosting, CatBoost, GLM, GAM.

Dans un environnement de forte concurrence et de maturité des marchés, les actuaires travaillent en permanence pour ajuster les primes calculées aux vrais coûts et utilisent des méthodes bénéficiant de la grande quantité de données disponibles pour améliorer les prévisions. Notre travail vise à contribuer à cette amélioration dans le cas des primes en assurance automobile.

Dans ce mémoire, nous nous focaliserons sur une étape spécifique du cycle de vie des sinistres en assurance automobile : la réparation du véhicule, en particulier le choix du garage automobile, et ses conséquences sur les coûts de réparation. Avant 2014, les assureurs ont pu maîtriser les coûts de réparation du véhicule en orientant leurs clients vers des garages *agréés*, avec lesquels une convention d'agrément est négociée pour réduire ces coûts, en échange d'un volume élevé de clients. Mais depuis l'adoption de la loi Hamon, les assurés ont désormais le droit de choisir librement l'atelier de réparation automobile, ce qui peut entraîner jusqu'à 50% de charges supplémentaires. Ainsi, dans ce mémoire, nous allons tenter de comprendre et d'anticiper les déterminants du choix du carrossier de réparation pour inclure les conséquences de ce choix dans les primes futures.

Notre étude se fera en trois étapes. Tout d'abord, nous commençons par l'étude de l'orientation en garages partenaires, variable indiquant si l'agent contacté pour la déclaration du sinistre a orienté son client vers un garage agréé. Nous nous intéressons particulièrement à cette étape puisqu'elle influence la décision finale prise par l'assuré. Ensuite, en prédisant un score d'orientation pour chaque contrat, nous modélisons notre variable d'intérêt, à savoir le choix du réparateur, par des méthodes d'apprentissage automatique. Enfin, nous démontrons comment l'intégration du choix du garage dans les modèles de primes permet de constituer des tarifs plus exacts et de réduire les erreurs de prédiction des montants à verser par l'assureur.

Abstract

Keywords : insurer-recommended repair shop, orientation, repair, Claim Severity, Bagging, Boosting, CatBoost, GLM, GAM.

Reducing the uncertainty and predicting accurately the consequences of risks remain the main function of an actuary. In fact, in an environment of high competition and a mature market, actuaries are working constantly to adjust the calculated premiums to the actual costs and are using innovative methods that benefit from the high amount of available data to improve the predictions. Our work aims at contributing to this improvement in the case of car insurance premiums.

In this thesis, we will focus on a specific step of the life cycle of car insurance claims, often neglected in usual pricing models. The variable we are referring to is the reparation of the car, particularly the choice of the car shop and its consequences on the costs. Until 2014, insurers were able to control the reparation costs by referring their clients to "preferred shops", with which an agreement is signed to reduce these costs, in exchange of a high volume of clients. However, since the passing of the Hamon law, the insured have now the right to choose the car repair shop, which introduces a new source of randomness in the accident's costs. To ensure that the additional costs are reflected in the premiums, our research will attempt to understand the reasons behind the choice of the car shop and anticipate it.

Our research will be done in three steps. First, we focus on the variable indicating whether the agent contacted after the accident recommended to his client to repair in one of the insurer's approved car shops. We chose to study this variable since it will potentially influence the policy holder's final decision. Then, by predicting the orientation score for each contract, we model our variable of interest : the choice of the auto repair shop, using machine learning methods. Finally, we demonstrate how integrating the choice of the auto repair shop into premium models allows us to build more accurate rates and to reduce prediction errors of the amounts paid by the insurer.

Remerciements

Qu'il me soit permis, au terme de ce travail, d'exprimer ma reconnaissance et ma gratitude à mon directeur de mémoire et tuteur en entreprise, Edouard VICAIRE, pour son accompagnement et pour tout le temps qu'il m'a consacrée tout au long de l'élaboration de ce travail, et pour m'avoir fait confiance sur ce sujet de mémoire très intéressant.

Je tiens également à remercier toute l'équipe Technical & renewal pricing : Mikaël PATRIER-LEITUS, Gabrielle VILA et Apollinaire BARME pour leurs conseils et leur disponibilité durant toute la réalisation de ce mémoire. Je remercie aussi toute l'équipe actuariat et développement des produits auto d'AXA France IARD pour leur accueil et leur bonne humeur quotidienne.

Je tiens aussi à témoigner mes remerciements à ma tutrice à l'ENSAE, Caroline HILLAIRET, pour son suivi et ses conseils durant l'élaboration de cette étude. Je remercie finalement tous ceux qui ont contribué de près ou de loin à l'accomplissement de ce mémoire.

Table des matières

Introduction

L'évaluation des coûts futurs engendrés par un contrat reste l'une des fonctions principales de l'actuaire. En effet, élaborer un tarif qui couvre efficacement le risque souscrit est primordial pour le fonctionnement correct du secteur des assurances et pour ce faire, plusieurs variables sont à inclure dans la prédiction des coûts et le calcul de la prime. Pour notre cas, nous nous intéresserons particulièrement aux coûts de réparation du véhicule en assurance automobile.

Si l'assureur maîtrisait ces coûts en orientant ses assurés vers des garages partenaires avec lesquels les prix ont été négociés, depuis 17 mars 2014, la loi Hamon a introduit l'article L. 211-5-1 dans le Code des assurances, renforçant le droit de l'assuré. En effet, cette loi lui donne la liberté en cas de sinistre de réparer son véhicule dans le garage de son choix et impose à l'assureur de mentionner dans le contrat ce droit de s'orienter vers le réparateur qui convient à l'assuré, et au moment de la déclaration du sinistre, l'assureur doit également rappeler à son assuré sa liberté de choix du garage. Il est donc intéressant de comprendre les facteurs qui expliquent ce choix et leur impact sur les coûts assurés.

En réparant son véhicule en garage agréé par son assureur, l'assuré bénéficie de plusieurs avantages. En effet, aucune avance des frais de réparation n'est à payer et seule la franchise prévue dans le contrat reste à la charge de l'assuré. De plus, les démarches administratives sont simplifiées et les réparations sont garanties par l'assureur qui peut également fournir un véhicule de prêt à son assuré pour garantir sa mobilité durant la période de réparation.

Cependant, l'assuré peut préférer confier son véhicule à un garage non partenaire pour différentes raisons. Si la prise en charge du véhicule est rapide par un garage agréé, les durées d'intervention peuvent quant à elle varier selon sa disponibilité. De plus, les compétences du garagiste peuvent être un facteur expliquant le choix d'un garage non-partenaire qui peut avoir une meilleure réputation ou avoir l'avantage de connaître mieux la marque du véhicule et les réparations nécessaires. Ensuite, les coûts de réparation peuvent être réduits chez un garage non agréé et ainsi la franchise calculée sur ce montant est plus faible, dans le cas d'une franchise proportionnelle. Certains garages indépendants offrent également des services annexes intéressants (nettoyage gratuit après réparation, rembour-

sement de la franchise...) et une meilleure expérience client globalement.

Notre étude va porter ainsi sur les déterminants du choix entre un garage agréé ou indépendant. Nous allons tout d'abord commencer par étudier l'orientation de l'assuré vers les garages agréés au moment du sinistre. En effet, après survenance d'un accident, l'agent peut encourager l'assuré à privilégier le garage partenaire pour profiter des avantages cités précédemment. Nous allons donc prédire ce score d'orientation puisqu'il est déterminant dans le choix fait par l'assuré du réparateur puis l'inclure, ainsi que les différentes variables dont nous disposons, dans le calcul d'un score de réparation dans un garage partenaire ou non pour chaque assuré. Ce score permet d'avoir une meilleure vision des coûts futurs des sinistres et sera donc inclus dans la prédiction de ces charges.

Quant aux modèles employés, nous allons comparer deux méthodes ensemblistes : le bootstrap aggregating ou Bagging, et le Boosting. En effet, si ces deux méthodes sont basées sur des arbres de décision, elles diffèrent dans la façon de construction de ces arbres, soit en effectuant des tirages aléatoires des observations et variables à chaque itération (Bagging) soit en appliquant une variante de la descente de gradient sur l'ensemble des arbres (Boosting). Comme méthode de Bagging, nous avons choisi d'appliquer un modèle de *forêts aléatoires* pour les deux scores, et pour le Boosting, deux méthodes ont été testées : le *XGBoost* qui est une application du modèle de Boosting classique et un nouveau modèle : le *CatBoost*, qui corrige le biais causé par le Boosting et introduit une nouvelle méthode d'encodage des variables catégorielles. Nous allons présenter chacun de ces modèles d'apprentissage automatique, et les appliquer dans la modélisation de nos scores.

Enfin, ce score prédit nous permettra de modéliser la sévérité selon le choix du garage de réparation, en considérant la combinaison linéaire des sévérités par type de garages, pondérées par les probabilités prédites pour l'assuré de choisir un garage agréé ou non agréé. Ce modèle sera comparé à la sévérité qui ne prend pas en compte le type de garage de réparation et nous démontrons que notre approche réduit les erreurs de prédiction des sévérités et améliore ainsi le calcul de la prime.

Chapitre 1

Contextualisation et problématique

Dans cette première partie, nous allons introduire le contexte général de notre étude pour comprendre les besoins et enjeux dont il est question et analyser les différentes facettes de la problématique de ce mémoire.

1.1 L'assurance automobile en France

Depuis 1958, tout véhicule terrestre à moteur, roulant et non roulant, doit être couvert au minimum par une garantie responsabilité civile [1]. Cette garantie se substitue au conducteur pour indemniser les dommages corporels ou matériels causés à des tiers dans le cas d'un sinistre non intentionnel causé par un conducteur identifié. Le cas échéant, ou lorsque le conducteur est non assuré ou si son assureur est insolvable, c'est le Fonds de Garantie des Assurances Obligatoires de dommage, alimenté principalement par les cotisations des entreprises d'assurance, qui intervient pour indemniser les victimes du sinistre.

Étant donné cette nature obligatoire, l'assurance automobile concerne environ 42 millions de véhicules de première catégorie (des véhicules de quatre roues), faisant d'elle la branche non-vie la plus importante en termes des cotisations qu'elle génère. Cette branche représente un peu plus de 39% du chiffre d'affaire de l'assurance de biens et de responsabilité (voir figure 1.1).

1. article L211-1 du Code des assurances et article L324-1 du Code de la route.

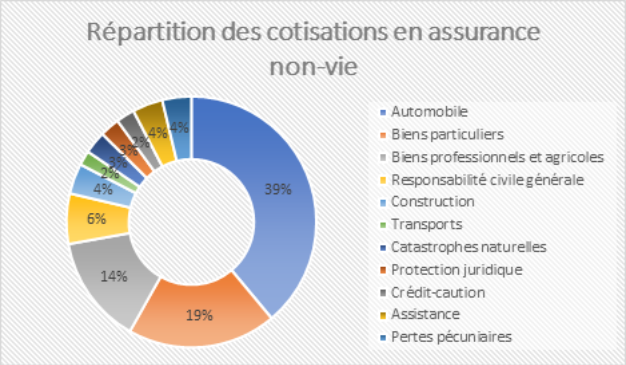


FIGURE 1.1: Les cotisations par branche d'assurance non-vie en France

Source: Rapport FFA 2019

En 2019, la branche non-vie a généré environ 22,8 Md€ de cotisations, en augmentation de 3,17% par rapport à 2018. En termes des prestations, elles fluctuent entre 2015 et 2019, avec une augmentation annuelle de 6,25% en 2019 à 18,7 Md€. Quant au ratio combiné net de réassurance, défini comme le rapport des prestations versées pour sinistres, des dotations et des frais généraux sur le chiffre d'affaires total, il est en augmentation de 2% par rapport à 2018 pour s'établir à 102,0%, remettant la branche en situation de déséquilibre, bien que la fréquence des sinistres soit globalement en baisse.

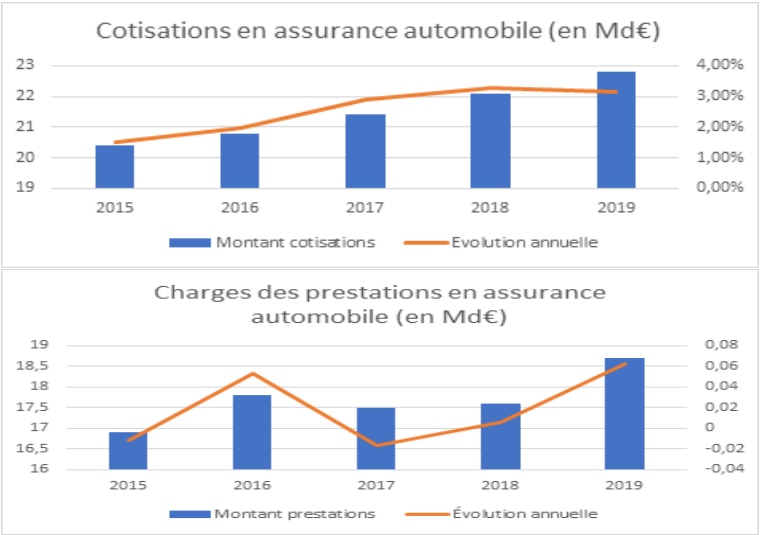


FIGURE 1.2: Évolution des cotisation et des charges en assurance automobile en France

Source: Rapport FFA 2019

Cette hausse des charges est en effet globalement liée à la hausse des coûts de réparation et pièces détachées, représentés par l'indice SRA (Sécurité et réparation automobiles).

Cet indice, mesuré trimestriellement, est la moyenne de trois composantes : les coûts des pièces de rechange, la main d'oeuvre et peinture. Et depuis 2015, ces trois composantes sont en hausse continue, ce qui contribue au déséquilibre de la branche automobile. Cette hausse, selon Frédéric Maisonneuve, ancien président du SRA, est principalement due à « l'intégration de nouveaux équipements technologiques améliorant la sécurité des véhicules, avec notamment la multiplication des capteurs ». C'est dans ce contexte de hausse des coûts de réparation que s'inscrit notre étude. En effet, pour maîtriser ces charges et les couvrir par les primes versées par les assurés, les assureurs recourent à des agréments avec les réparateurs partenaires pour fixer une grille tarifaire. Cependant, comme le choix du garage reste aux mains de l'assuré, la maîtrise des coûts n'est pas toujours possible. Il faut ainsi comprendre les déterminants du choix du garage par l'assuré pour en prendre compte dans la prime à payer. Ce choix de garage sera détaillé et modélisé dans la suite du mémoire.

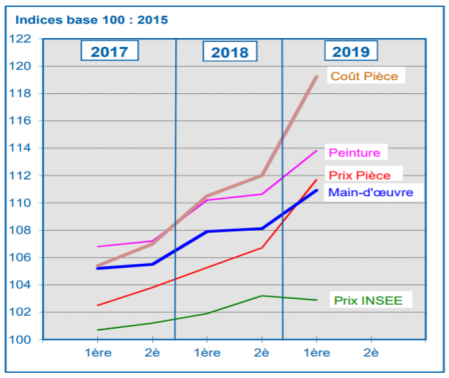


FIGURE 1.3: Évolution des coûts des composantes de la réparation

Source: Communication statistique SRA 2019

1.2 AXA France sur le marché de l'assurance automobile

En assurance automobile, la filiale AXA France, entreprise du groupe AXA, est classée deuxième en France en termes de chiffre d'affaires en 2019. Dans cette branche, plusieurs garanties classiques sont proposées : responsabilité civile automobile, défense pénale et recours suite à accident, garantie du conducteur dans le cas des sinistres corporels, assistance, bris de glace, dommages tous accidents... Elle offre également d'autres garanties moins usuelles, comme l'option Joker incluse dans tous les contrats auto pour les jeunes, et qui consiste à bénéficier de 5 trajets par an en taxi dans le cas de l'incapacité de conduire,

quelle que soit la cause de cette incapacité.

Dans un contexte de très forte concurrence et avec la montée en puissance des bancassureurs qui bénéficient d'un large réseau de distributeurs, AXA lance une nouvelle offre d'assurance auto en 2018, dite MonAuto, qui vient remplacer les contrats Axa Auto Référence. Cette nouvelle assurance se veut plus simple, en proposant six formules, et transparente puisque le client observe l'évolution des prix selon l'option choisie. Cette nouvelle offre a permis de réduire les prix d'affichage de 30%, selon le responsable du marché IARD des particuliers et des professionnels d'Axa France, Henry de Courtois. Enfin, pour fidéliser les clients, cette offre propose une réduction progressive de la franchise.

C'est dans ce cadre de marché très compétitif et de recherche d'amélioration des primes en assurance automobile que s'inscrit notre étude. En intégrant le type de réparation dans le tarif, le risque sous-jacent est mieux appréhendé et le tarif ainsi segmenté permettrait une meilleure maîtrise des coûts et d'éviter des problèmes d'antisélection.

1.3 Loi Hamon

La loi n° 2014-344 du 17 mars 2014 relative à la consommation, dite loi Hamon, vise à renforcer le droit du consommateur face aux professionnels dans plusieurs domaines. En effet, cette loi protège le consommateur en définissant les responsabilités du professionnel dans le cas des contrats conclus à distance par exemple, en allongeant les délais de rétraction de 7 à 14 jours. Dans le cadre de la vente en ligne, elle impose le respect des délais de livraison et interdit les cases pré-cochées. De plus, elle recadre le démarchage par téléphone en interdisant les appels en numéro masqué d'opérateurs et en permettant aux consommateurs qui le souhaitent de s'inscrire dans une liste d'opposition au démarchage. Et étant donné le volume des contrats souscrits en assurance, ce secteur a été particulièrement impacté par la loi Hamon qui visait à renforcer la concurrence dans ce domaine et de réduire l'augmentation des tarifs.

Loi Hamon en Assurance

La loi Hamon a introduit plusieurs nouveautés dans le secteur assurantiel. En effet, les modalités de résiliation des contrats auto, moto, multirisque habitation et affinitaires ont changé pour permettre à l'assuré de résilier son contrat, après un an d'engagement, à tout moment et sans frais ni pénalités. Il suffit d'informer l'assureur de la volonté de résí-

liation, ou de souscrire un contrat chez un nouvel assureur qui s'occupera des démarches nécessaires. Cela a eu pour effet l'augmentation du taux de perte de clientèle chez les assureurs de 16% durant l'année suivant l'entrée en application de la loi, taux qui était stable depuis 2013 [2].

La loi Hamon a également modifié l'assurance prêt immobilier et vient complémenter la loi Lagarde qui permet le libre choix de l'assureur emprunteur. En effet, elle donne à l'assuré la liberté de changer d'assureur sur l'année suivant la signature du contrat de prêt immobilier. En ce qui concerne les assurances affinitaires, qui incluent les garanties souscrites après la vente d'un produit ou d'un service, et en plus de la possibilité de résilier après un an, le client peut se rétracter dans un délai de 14 jours.

Et c'est l'article L. 211-5-1 introduit par la loi Hamon dans le code des assurances qui nous intéresse particulièrement. Cet article, qui concerne tous les contrats responsabilité civile automobile, donne à l'assuré la liberté de choisir le réparateur dans le cas d'un sinistre. En effet, avant la promulgation de la loi Hamon, les assureurs orientaient automatiquement les assurés sinistrés vers des garages agréés, avec lesquels une convention tarifaire fixant les prix de réparation a été signée. Et en échange de cette réduction des tarifs, les garages profitent du volume élevé de clients de l'assureur.

L'article L. 211-5-1 impose à l'assureur non seulement de laisser le libre choix du carrossier à l'assuré mais également de lui rappeler ce droit au moment de déclaration du sinistre. En effet, si le véhicule du sinistré nécessite des réparations couvertes par la garantie, l'assureur est dans l'obligation de mentionner la liberté du choix du garage selon des modalités précisées par un arrêté du 17 juin 2016 : soit par une mention dans le constat européen d'accident, par courrier électronique ou SMS, pour assurer la traçabilité de l'information.

Si cette option permet à l'assuré d'avoir plus de liberté, elle réduit la maîtrise du risque par l'assureur. En effet, un assuré qui répare en garage partenaire coûte moins cher à l'assureur qu'un autre qui choisit un garage indépendant, bien que le sinistre subi par les deux profils soit le même, différence qui peut atteindre jusqu'à 50% de charges supplémentaires. C'est ainsi qu'il serait intéressant d'anticiper ce choix et de l'intégrer au préalable dans la prime à payer, ce qui est l'objectif de notre étude.

2. Étude menée par l'institut d'études quantitatives on-line Arcane Research en 2015 : `https://www.arcane-research.com/etude/etude-assurance-auto-3/`

1.4 Problématique et enjeux

L'assurance automobile est un contrat entre un assureur et assuré qui protège ce dernier des dégâts financiers dans le cas de survenance de sinistre. En échange de cette couverture, l'assuré paie à l'assureur une prime déterminée par ce dernier, selon les caractéristiques de l'assuré, son véhicule et la police souscrite. Cette assurance concerne trois principales parties : la responsabilité civile qui couvre les dommages, corporels et matériels, causés à autrui ; les dommages matériels et vol subis par le conducteur et enfin les préjudices corporels dus au sinistre : les dépenses de santé actuelles et futures, les coûts d'assistance par tierce personne, les pertes de revenu ...

Pour estimer les coûts futurs potentiels d'un assuré et déterminer la prime nécessaire pour couvrir ces charges, un modèle collectif est appliqué. En effet, pour un assuré, la charge totale des sinistres sur une période donnée peut s'écrire :

$$\begin{cases} S = \sum_{i=1}^{N} X_i & \text{si } N > 0 \\ S = 0 & \text{si } N = 0 \end{cases}$$

tel que : N est le nombre de sinistres survenus, variable aléatoire inconnue au moment de la souscription de la garantie, et X_i est le coût du sinistre i. Ces coûts de sinistres $X_1, X_2, \ldots$ sont supposés indépendants entre eux et identiquement distribués, et indépendants du nombre de sinistres survenus N. La prime pure se définit comme l'espérance des coûts $\mathbb{E}(S)$, et se calcule ainsi selon la première formule de Wald : $\mathbb{E}(S) = \mathbb{E}(N) \times \mathbb{E}(X)$.

Deux principes fondamentaux sont au coeur de l'activité des assurances. Le premier concept est la **mutualisation** des risques qui consiste à constituer un grand portefeuille d'assurés de risques homogènes et indépendants. En appliquant ainsi la loi des grands nombres, ce principe implique la convergence de la moyenne empirique des coûts des sinistres des assurés vers une valeur constante prévisible égale à la prime pure.

Cette mutualisation des risques suppose que la population est à risques identiques, ce qui justifie le second principe de différenciation entre assurés et de création de groupes homogènes de risques, il s'agit de la **segmentation**. Cette dernière est importante dans les secteurs hautement concurrentiels, comme est le cas de l'assurance automobile. En effet, en supposant que les profils des assurés sont semblables et en présence d'assureurs qui différencient entre les risques, l'assureur risque de souffrir de problèmes d'antisélection :

en proposant un même tarif pour tous ses contrats, il fait fuir les bons profils d'assurés et attire les mauvais risques.

Notre étude a pour but d'améliorer la segmentation des assurés et d'éviter ces problèmes de sélection adverse en considérant un nouveau critère de segmentation : la réparation en garages partenaires. En effet, après un sinistre, l'assuré peut réparer son véhicule soit dans un garage partenaire avec lequel l'assureur a conclus un agrément fixant les tarifs à un niveau inférieur par rapport aux tarifs du marché, soit dans un garage de son choix. Et ce choix impacte fortement les coûts futurs dans le cas de survenance de sinistres. Nous pouvons ainsi parler de deux profils d'assurés : ceux qui choisissent des garages agréés et ceux qui préfèrent des garages libres. Notre problématique peut se formuler ainsi : Comment distinguer entre un assuré qui choisirait un garage partenaire d'un autre qui s'orienterait vers un garage indépendant ? Et comment segmenter le tarif selon ces deux profils ?

Puisque le choix du garage intervient après la survenance du sinistre, nous choisissons d'inclure ce choix dans le modèle de sévérité uniquement. Nous cherchons donc à modéliser une probabilité de réparation en garage agréé et d'inclure ce score prédit dans les modèles de sévérité.

Pour inclure ce choix, nous différencions entre deux variables aléatoires des coûts des sinistres : X_1 si la réparation est en garage partenaire et X_0 sinon. En posant Y le choix de l'assuré, variable binaire telle que $Y = 1$ indique que l'assuré a réparé en garage partenaire et $Y = 0$ sinon, la sévérité d'un sinistre s'écrit :

$$\mathbb{E}(X) = \mathbb{E}(X_1|Y = 1) \times \mathbb{E}(Y) + \mathbb{E}(X_0|Y = 0) \times (1 - \mathbb{E}(Y))$$

Où : $\mathbb{E}(Y)$ est la probabilité de réparation en garage agréé, score que nous allons modéliser. Et chaque sévérité X_0 et X_1 est à modéliser sur le sous-groupe des sinistres correspondant.

Avant de modéliser la variable du choix du garage de réparation, nous disposons également d'une variable d'orientation. Cette variable binaire indique pour un sinistre si l'agent a orienté l'assuré vers un des garages agréés et lui a rappelé les avantages de la réparation dans ces garages : un véhicule de prêt durant la période de réparation, payer uniquement la franchise de la garantie et des réparations garanties par l'assureur. Orienter un assuré vers un des réparateurs agréés est déterminant dans le choix final de l'assuré, c'est

pourquoi nous allons commencer par la modélisation de cette orientation pour ensuite l'intégrer dans le choix du garage de réparation comme variable explicative. Enfin, nous intégrons ce score dans les modèles de sévérité et démontrons l'utilité de distinguer entre les deux profils d'assurés dans l'amélioration de ces prédictions de sévérité.

1.5 Construction des bases de données

Dans cette partie, nous décrivons la méthode de construction des deux bases de données d'orientation et de réparation en garages partenaires. Notre but est d'avoir pour chaque police toutes les informations disponibles à la date du terme du contrat, par exemple celles relatives à l'assuré, son véhicule ou la police souscrite, et pouvant expliquer en cas de sinistre les deux variables binaires : si l'assuré a été orienté vers un garage partenaire AXA après avoir déclaré son sinistre auprès de son agent, et si la réparation a été effectivement effectuée dans un de ces garages. Nous allons également fixer un périmètre d'étude selon les données disponibles.

Nous disposons de deux bases de données « garage » : Orientation et Réparation. Elles renseignent, pour les sinistres survenus en 2018 et 2019, le numéro du contrat, la date du sinistre et les deux variables de l'étude : l'orientation vers un garage agréé et le choix du type de réparation. Ces deux variables sont collectées auprès de l'agent du client qui informera l'assureur s'il a orienté l'assuré vers l'un des garages partenaires et le type du garage choisi après cette orientation. Nous nous limitons aux polices disponibles dans le portefeuille de l'assureur durant les deux années 2018 et 2019, correspondantes aux dates de disponibilité des informations sur les garages.

Les informations sur ces polices sont renseignées dans des bases dites Business View (BV). Elles recensent les polices du portefeuille automobile des clients particuliers d'AXA et leurs informations actualisées à la fin de chaque mois. Les variables de ces bases de données ainsi que leurs caractéristiques seront traitées en détail dans la partie analyse descriptive du mémoire 3.1.2. Ces variables peuvent être regroupées en six principales catégories :

— Les variables décrivant la police souscrite : les garanties sélectionnées, la prime et la modalité du paiement choisie (annuelle, mensuelle...)...
— Les informations relatives à l'assuré : son âge, sa profession, le nombre de contrats vie et non vie souscrits...

— Le comportement de l'assuré : son mode de logement, son état matrimonial, le nombre d'enfants à sa charge...
— Les caractéristiques du véhicule assuré : la marque, le nombre de cylindres, le poids...
— Les informations sur l'agence de souscription du contrat : sa localisation, l'identifiant de l'agent et du portefeuille, son nom...
— Les informations relatives aux sinistres passés du contrat : leurs dates de survenance, types, coûts...

Il faut noter cependant que ces bases BV donnent une vision ponctuelle des contrats, en décrivant les données sur le portefeuille automobile à sa date d'actualisation. Il faut ainsi récupérer plusieurs bases BV, sur chaque date d'échéance, pour avoir l'information nécessaire pour la prédiction. Et puisque ces bases sont actualisées mensuellement, nous nous contentons de l'information à ces dates pour décrire le contrat, telles que sur chaque mois des deux années 2018 et 2019, nous collectons les données sur les contrats arrivés à leur échéance durant ce mois.

Pour sélectionner ces contrats, nous disposons d'une dernière base contenant deux informations : le numéro de contrat, le mois et l'année d'échéance. Nous sélectionnons ainsi pour chaque contrat et date renseignés dans cette base les données dans la BV correspondante, parmi les 24 BV collectées. Cette base obtenue après jointure nous donne une vision rétrospective sur les sinistres à chaque date d'échéance. Cependant, nous cherchons à prédire à cette date du terme du contrat l'orientation et la réparation en garages partenaires en cas de survenance d'un sinistre, un dernier traitement doit être ainsi effectué sur la base. Ce dernier consiste à récupérer les informations du contrat au moment d'échéance et les informations sur les sinistres un an après.

Enfin, puisque nous ne disposons des informations d'orientation et de réparation que sur les deux années 2018 et 2019, et pour avoir un périmètre d'observation égal pour chaque contrat, nous nous limitons aux contrats arrivés à leur terme durant l'année 2018. En effet, comme nous pouvons observer sur la figure 1.4, les contrats de date d'échéance en 2019 ne seront pas observés sur une année complète, notamment si le sinistre survient en 2020. Nous allons ainsi sélectionner les contrats d'échéance entre le 01 Janvier et le 31 Décembre de l'année 2018, leurs sinistres survenus sur une année et les informations orientation et réparation correspondantes :

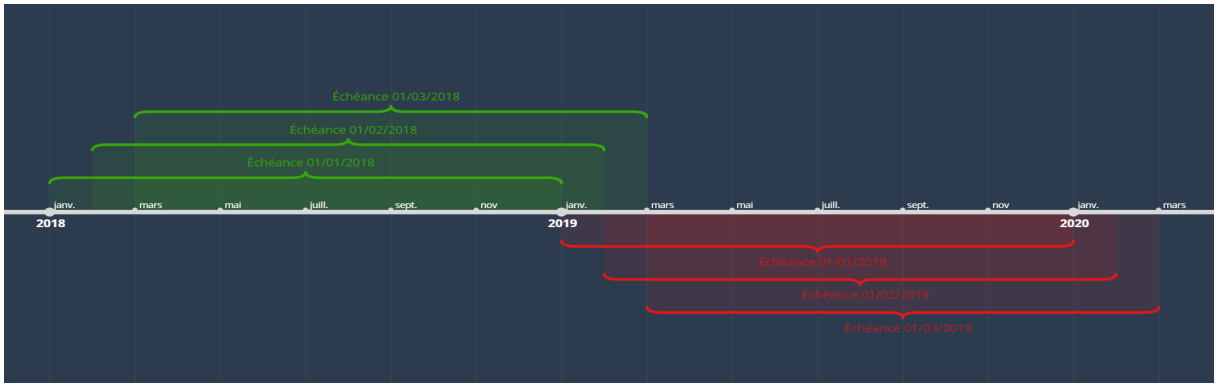


FIGURE 1.4: Durée d'observation du contrat. En vert, l'information sur les sinistres est disponible, alors qu'en rouge, elle est incomplète sur l'année.

La base d'observations obtenue ainsi contient 85 230 observations, et sera découpée aléatoirement en deux sous-bases : une base d'apprentissage avec 80% des observations sur laquelle les modèles sont estimés et une base test de 20% pour évaluer leurs performances.

Chapitre 2

Présentation des modèles

Dans ce chapitre, nous allons introduire les modèles de classification utilisés pour prédire nos deux scores d'orientation et de réparation. Nous avons choisi de travailler avec des méthodes ensemblistes, Bagging et Boosting, basées sur des arbres de décision, et qui diffèrent dans la méthode de construction de ces arbres. Pour l'exemple du Bagging, nous allons travailler avec des forêts aléatoires alors que pour le Boosting, nous allons employer le XGBoost, et le comparer à une nouvelle méthode de Boosting, CatBoost, qui introduit deux nouveautés par rapport au Boosting classique : *ordered boosting* et *ordered encoding* pour réduire le biais des prédictions. Nous allons ainsi présenter la théorie essentielle derrière chaque modèle afin de mieux comprendre l'apport de chacun.

2.1 Forêts aléatoires

Dans cette partie, nous présentons le premier modèle de prédiction employé dans notre étude. Il s'agit des forêts aléatoires, qui est une application de la méthode d'apprentissage ensembliste du bootstrap aggregating ou bagging, appliquée sur les arbres binaires de décision CART. Introduit en 2001 par Breiman [3], ce modèle devient très populaire et surpasse les performances des autres méthodes de classification sur un grand nombre de bases de données populaires [7]. Étant donné que la construction des forêts aléatoires se base principalement sur les deux notions d'arbres CART et de méthode de bagging, nous allons commencer par présenter chacune des deux méthodes, et leur application dans le modèle de forêts aléatoires.

2.1.1 CART

Classification And Regression Trees, comme son nom l'indique, est une méthode d'apprentissage non-paramétrique de construction d'arbres de classification et de régression, développée par Breiman et al. [12] en 1984. Ces arbres sont binaires, où l'ensemble d'observations, dit *noeud*, est divisé à chaque étape en deux noeuds plus homogènes selon la nature des données et le critère d'homogénéité choisi.

L'enjeu est de choisir parmi p variables $X_1, X_2, \ldots, X_p$, celles permettant d'expliquer une variable Y, quantitative ou qualitative, et d'associer à chaque combinaison des valeurs de ces variables une valeur ou catégorie unique de Y. Étant donné la nature des deux variables de notre étude, nous allons nous placer dans le cadre de la classification binaire et présenter la méthode de construction d'arbres de ce type.

En partant de l'échantillon complet, CART cherche parmi toutes les variables X et leurs valeurs celle qui permettra de créer deux noeuds plus homogènes que le noeud parent au sens d'un critère portant sur la variable Y. Et de façon itérative, ces deux noeuds seront également répartis en deux autres, jusqu'à obtenir des noeuds terminaux, dits *feuilles*, et auxquels sera associée la classe prédite de Y.

Homogénéité du noeud

Deux critères d'homogénéité principaux peuvent être employés. Le premier est l'indice de **Gini** qui mesure l'hétérogénéité du noeud et se calcule sur un noeud t comme suit :

$$i(t) = 1 - p_{t,0}^2 - p_{t,1}^2$$

Où : $p_{t,0}$ et $p_{t,1}$ sont les proportions des deux classes dans le noeud. En effet, si le noeud contient uniquement une seule classe, la fonction est nulle et le noeud est *pur*. Cette fonction est maximale pour la valeur $p = 0.5$, où les deux classes sont présentes à proportions égales dans le noeud et le noeud est dit *impur*.

Ainsi, nous cherchons parmi tous les découpages possibles d'un noeud en deux, celui qui minimisera cette fonction sur les deux noeuds enfants. Cela revient à maximiser la diminution d'impureté après division, calculée comme suit :

$$\text{Gini(noeud parent)} - p_1 \times \text{Gini(noeud enfant droit)} - p_2 \times \text{Gini(noeud enfant gauche)}$$

telle que p_i est la proportion d'observations du noeud parent qui sont regroupées dans le noeud i. Ainsi, en parcourant toutes les valeurs pour chaque variable X, les observations sont regroupées en deux noeuds, la perte d'impureté est calculée et la meilleure division retenue est celle qui maximisera cette perte.

La deuxième mesure d'impureté d'un noeud est l'**entropie**, qui introduit le logarithme des probabilités des classes dans son équation :

$$\text{Entropie} = -p_1 log_2(p_1) - p_2 log_2(p_2)$$

Cette fonction conserve les mêmes propriétés de l'indice Gini et est minimale lorsque le noeud contient une seule classe (p_1 ou p_0 nul). Le gain de pureté après division sera également calculé comme la différence entre l'entropie du noeud parent et les deux entropies des noeuds enfants, pondérées par les proportions des observations de chaque noeud. Elle peut être cependant plus lourde à calculer par rapport à l'indice de Gini puisqu'elle nécessite le calcul du logarithme à chaque étape.

2.1.2 Construction de l'arbre

Maintenant que nous avons introduit une mesure d'estimation de l'impureté du noeud, nous construisons des groupes d'observations homogènes à chaque étape. Nous commençons par un seul noeud, regroupant toutes les observations, dit *racine*. Ensuite, CART cherche parmi toutes les valeurs possibles des X celle qui maximisera le gain de perte d'impureté et crée ainsi deux noeuds. Il faut noter que les observations vérifiant la condition de découpage du noeud seront regroupées dans le noeud enfant gauche comme suit :

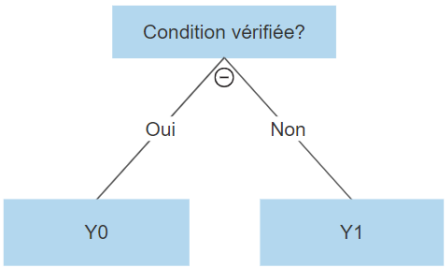


FIGURE 2.1: Exemple d'arbre de classification simple

Ensuite, sur les deux noeud créés, la même procédure est appliquée, en les divisant en deux noeuds à leur tour. Et c'est le noeud dont le gain de perte d'impureté est maximal qui sera effectivement réparti en deux. Ainsi, itérativement, plusieurs noeuds seront créés par des division successives de l'échantillon.

Pour éviter d'avoir des arbres très profonds et les problèmes de surapprentissage qui peuvent en découler, un ou des critères d'arrêt des divisions doivent être fixés. Le critère

d'élagage a priori peut s'agir par exemple du nombre minimal d'observations dans les feuilles, la profondeur maximale des noeuds ou encore un seuil minimal d'amélioration de l'homogénéité du noeud après découpage, pour éviter des divisions peu significatives.

CART présente plusieurs avantages, parmi lesquels nous pouvons citer :

- L'interprétabilité des résultats : étant donné la forme arborescente des résultats, il est facile de voir les variables qui interviennent à chaque étape et leurs effets sur la variable à prédire.
- Son caractère non paramétrique : CART ne nécessite aucun choix des lois des variables ou de leurs liens de dépendance. Il permet également de détecter les interactions entre les variables, sans le préciser dans le modèle.
- CART permet de détecter l'effet non linéaire d'une variable sans avoir à la transformer.

Cependant, ce modèle est souvent instable et la structure de l'arbre dépend fortement des observations utilisées. Un autre problème qui peut survenir est le surapprentissage, si les critères d'arrêts sont mal choisis. C'est ainsi que les méthodes ensemblistes, notamment le bootstrap aggregating, ont été développées pour améliorer la stabilité et la précision du modèle CART.

2.1.3 Bootstrap Aggregating

Supposons un échantillon d'apprentissage composé de n observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, où $X_i \in \mathbb{R}^p$ est l'observation i pour les p variables explicatives et Y_i la variable à prédire, qui est qualitative dans notre cas et prend deux valeurs : y_0 et y_1. Supposons également que cet échantillon est indépendant et identiquement distribué, de loi F.

Les méthodes d'agrégation consistent à regrouper plusieurs modèles construits sur des échantillons indépendants pour effectuer la prédiction finale. Cette méthode, bien qu'elle stabilise les résultats du modèle, nécessite une très grande quantité de données, ce qui n'est pas toujours possible en pratique. C'est ainsi que la méthode du Bagging construit B échantillons par échantillonnage aléatoire avec remise sur les observations d'origine. Ensuite, chacun des échantillons z_i où $i \in \{1,\ldots,\mathrm{B}\}$, tel que $card(z_i) \leq n$, est utilisé pour construire un modèle de prédiction de Y, qu'on notera $\hat{f}_k \in \{y_0, y_1\}$.

Pour obtenir la prédiction pour une nouvelle observation X_j, un vote à la majorité est effectué, en considérant la catégorie prédite par la majorité des modèles :

$$\hat{f}(X_j) = argmax_l N_l \text{ où } N_l = \{\text{nombre de modèles ayant prédit la classe } l\}$$

Cette méthode regroupe ainsi plusieurs modèles entraînés sur des observations différentes. Cependant, comme l'indique Breiman [6], une condition pour que cette méthode donne une meilleure performance par rapport à un seul modèle $\hat{f}$, est l'instabilité de ce modèle. En effet, si le changement de l'échantillon n'engendre aucune modification du modèle, alors les prédictions sur les arbres f seront les mêmes et le Bagging donne un résultat semblable à un seul modèle. Encore plus, Breiman démontre qu'appliquer cette méthode sur un modèle stable comme la régression linéaire peut donner des résultats moins bons que le modèle de base. Pour les arbres de décision, qui sont généralement instables, le Bagging peut améliorer les résultats par rapport à un modèle unique.

2.1.4 Forêts aléatoires

Les forêts aléatoires sont une version améliorée du Bagging, appliquée sur les arbres de décision. En effet, ce modèle fait intervenir une composante aléatoire dans la construction de chaque arbre, en sélectionnant non seulement des observations différentes mais aussi des variables différentes.

En effet, en plus du Bagging qui effectuera un tirage aléatoire avec remise des observations pour chaque modèle, les forêts aléatoires effectuent également pour chaque modèle un tirage aléatoire parmi les variables $X_1, \ldots, X_p$ et l'entraîne uniquement sur les variables sélectionnées. Ainsi, la recherche de la division optimale des noeuds d'un arbre se fera uniquement sur un échantillon des variables et des observations. Cela permet de créer des arbres non corrélés, car chacun entraîné sur une partie de la base d'apprentissage, et d'éviter ainsi le sur-ajustement du modèle final.

Lors de la création des forêts aléatoires, plusieurs paramètres sont à choisir pour améliorer la performance du modèle. Parmi ces paramètres, nous pouvons citer :

— Critère d'homogénéité : la fonction d'impureté, soit l'indice Gini soit l'entropie.
— Le nombre d'arbres à entraîner. Un grand nombre d'arbres améliore généralement la performance du modèle, mais rend l'apprentissage plus lent. Il faut ainsi choisir le nombre minimal d'arbres qui permet d'ajuster au mieux les données.

— La profondeur des arbres, définie comme la longueur du plus long chemin entre la racine de l'arbre et une feuille. Plus les arbres sont très complexes et profonds, plus le risque de surapprentissage est élevé.
— Le nombre de variables à tirer pour la division des noeuds : pour les problèmes de classification, le nombre de variables recommandé est $|\sqrt{p}|$ (voir [15]), mais en pratique, il faut le considérer comme un paramètre à régler pour chaque problème.
— Le nombre minimal d'observations par feuille : ici également, il faut éviter à la fois de créer des arbres très complexes en choisissant un petit nombre d'observations mais aussi pas très élevé afin que le modèle puisse détecter les dépendances entre les variables à partir des données.

Maintenant que nous avons analysé la méthode du Bagging à travers l'exemple des forêts aléatoires, nous allons introduire la deuxième méthode usuelle de construction d'arbres, à savoir le Boosting.

2.2 XGBoost

2.2.1 Introduction

Le Boosting est une méthode ensembliste de construction d'arbres de décision qui applique une variante de la descente du gradient à un espace fonctionnel F. En effet, à chaque étape, le prédicteur est actualisé en minimisant une fonction de perte sur un espace d'arbres de décision. Cette méthode diffère des forêts aléatoires par la méthode de construction des arbres. Si les forêts aléatoires créent des arbres parallèles tels que chacun est optimal sur une partie de la base de données, les méthodes de boosting construisent des prédicteurs sur les observations mal classées, en modélisant les erreurs des arbres précédents. Ainsi, cette méthode crée des prédicteurs faibles sur la base de données, tels que chaque arbre corrige les erreurs des arbres précédents.

eXtreme Gradient Boosting, ou XGBoost, améliore l'algorithme du Boosting classique en ajoutant une composante de régularisation à la fonction objective, ce qui permet de créer des arbres moins complexes et de réduire le surapprentissage. Il introduit également deux méthodes de choix des divisions des noeuds qui permettent de réduire le temps d'apprentissage sans pour autant réduire la précision des prédictions. Enfin, ce modèle propose une méthode de traitement des valeurs manquantes qui augmente également la

vitesse du modèle. Ces caractéristiques font du XGBoost l'un des modèles de prédiction les plus performants et utilisés sur la plateforme Kaggle.

2.2.2 Fonction objective et Boosting

Soit $D = \{x_i,\ y_i\}_{i=1,\dots,n}$ où $x_i \in \mathbf{R}^m$ sont les m variables explicatives de la i^{me} observation et $y \in \mathbf{R}$ la variable à prédire. Les modèles ensemblistes consistent à modéliser K arbres $f_1, f_2, \dots, f_K \in F$ l'espace des arbres de décision, tels que : $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$. Ces fonctions sont les solutions de la minimisation d'une fonction objective :

$$Obj(f) = L(f) + \Omega(f)$$

Où : L est la fonction de perte sur les données d'apprentissage, qui mesure le biais du modèle et Ω est une mesure de sa complexité. Cette dernière privilégie les modèles simples en pénalisant les arbres profonds et complexes. Ainsi, on retrouve la décomposition biais-variance classique, où la minimisation d'une des deux composantes entraîne la maximisation de l'autre.

Sur les données d'apprentissage D, la fonction objective s'écrit :

$$Obj(f) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Où : l est une fonction de perte, différentiable et convexe.

À chaque itération t, et comme en optimisation des paramètres en descente de gradient où on converge progressivement vers la solution, une nouvelle fonction f_t est modélisée de façon additive, telle que :

$$\hat{y_i}^{(t)} = \hat{y_i}^{(t-1)} + f_t(x_i)$$

Ainsi, on cherche un arbre f_t qui minimise la fonction suivante :

$$Obj(f) = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t) + cte$$

En appliquant la formule de Taylor au second degré sur l, on obtient :

$$l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) = l(y_i, \hat{y_i}^{(t-1)}) + f_t(x_i) \times \frac{\partial l(t,s)}{\partial s}(y_i, \hat{y_i}^{(t-1)}) + \frac{1}{2} \times f_t^2(x_i) \times \frac{\partial^2 l(t,s)}{\partial s^2}(y_i, \hat{y_i}^{(t-1)})$$

En posant : $g_i = \frac{\partial l(t,s)}{\partial s}(y_i, \hat{y_i}^{(t-1)})$ et $h_i = \frac{\partial^2 l(t,s)}{\partial s^2}(y_i, \hat{y_i}^{(t-1)})$, qui sont calculables à l'itération t, la fonction à minimiser à l'itération t devient :

$$Obj(f_t) = \sum_{i=1}^{n} f_t(x_i)g_i + \frac{1}{2}f_t^2(x_i)h_i + \Omega(f_t) + cte \tag{2.1}$$

Pour une complexité pénalisant le nombre de feuilles de l'arbre T et les scores prédits w_j dans chaque noeud j, la fonction se réécrit :

$$Obj(f_t) = \sum_{i=1}^{n} f_t(x_i)g_i + \frac{1}{2}f_t^2(x_i)h_i + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + cte$$

En posant I_j l'ensemble des observations du noeud j, f_t sur ce noeud est le score prédit w_j, et l'équation devient :

$$Obj(f_t) = \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i)w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda)w_j^2 \right] + \gamma T + cte$$

Pour résoudre ce problème, on suppose une structure d'arbre fixe et l'objectif revient à trouver les prédictions w_j^* sur chaque feuille qui minimisent l'équation. Il s'agir de la somme de T fonctions quadratiques indépendantes. Les prédictions optimales s'écrivent :

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

Ainsi, pour une structure d'arbre donnée, nous pouvons calculer les scores de prédiction sur chaque noeud et en déduire la fonction objective, ensuite la structure d'arbre minimisant cette fonction sera celle choisie par l'algorithme.

2.2.3 Choix de la structure d'arbre optimale

Comme il est impossible de vérifier toutes les structures d'arbre possibles, l'algorithme du Boosting crée des noeuds qui minimisent localement la fonction objective (greedy algorithm). En effet, au niveau de chaque noeud, l'algorithme trie les valeurs des variables, les parcourt et retient la division en deux noeuds qui minimise le plus la fonction de perte sur les deux branches créées.

Pour réduire le temps de calcul, le XGBoost modifie cette étape en vérifiant les divisions uniquement sur les quantiles des variables. Ces quantiles peuvent être calculés sur

toutes les observations au début de l'algorithme (global) ou être recalculés localement au niveau de chaque noeud ce qui est plus adapté pour les arbres profonds. Cependant, la méthode globale peut être tout aussi précise si le nombre de quantiles considérés est suffisant.

Enfin, pour le choix de l'arbre à retenir à l'itération t, deux méthodes sont proposées. La première consiste à s'arrêter de créer plus de noeuds lorsqu'aucune division ne permet de minimiser la fonction de perte. Si cette méthode est rapide, elle ignore les noeuds qui pouvaient créer des divisions futures plus précises. C'est ainsi que la seconde méthode crée des noeuds jusqu'à atteindre le nombre minimal de feuilles par noeud, puis regroupe les feuilles qui ne contribuent pas à l'amélioration du modèle.

Le XGBoost propose également une méthode pour traiter les valeurs manquantes. En effet, sur chaque noeud, ces valeurs sont regroupées par défaut soit avec les observations du noeud droit ou gauche. Pour choisir entre ces deux noeuds, l'algorithme, au moment de l'énumération des divisions possibles, calcule la fonction de perte pour les deux possibilités (gauche ou droite) et choisit la division optimale et le sens par défaut qui minimise la fonction de perte.

Bien que cette méthode donne de très bons résultats sur un grand nombre de base de données, elle nécessite le traitement préalable des variables catégorielles avant de pouvoir les inclure comme prédicteurs. Et étant donné la prédominance de ces variables dans les bases de données de notre étude, nous avons choisi d'étudier le modèle *CatBoost* pour un traitement automatique de ces données. Nous allons par la suite étudier les apports de ce modèle par rapport au Boosting classique et comment il corrige les biais causés par ce dernier.

2.3 CatBoost

Le Categorical Boosting ou CatBoost est la dernière méthode de boosting des arbres de décision, développée en 2017 par une équipe du groupe Yandex, le principal moteur de recherche en Russie. Cette méthode introduit deux principales nouveautés par rapport aux méthodes de Boosting classiques : ordered boosting ou boosting ordonné et une nouvelle méthode d'encodage des variables catégorielles : ordered target statistic (TS). En effet, les auteurs du CatBoost démontrent que l'algorithme du boosting utilisé dans tous les modèles classiques (GBDT, XGBoost, LightGBM) cause un problème de généralisation, puisque le gradient calculé à chaque itération utilise les mêmes valeurs de la variable

cible sur lesquels il a été optimisé. De plus, les méthodes d'encodage se basant sur la valeur cible souffrent du même problème de target leakage. Dans cette partie, nous allons donc explorer ces deux nouveaux principes et comment ils permettent d'améliorer les prédictions.

2.3.1 Categorical encoding

Plusieurs problèmes de classification font intervenir des variables catégorielles comme prédicteurs. Cependant, la majorité des modèles de prédiction ne peuvent pas utiliser l'information contenue dans ces variables sans traitements préalables. Pour effectuer ces transformations, plusieurs méthodes sont disponibles, dont nous citons :

Encodage non supervisé

Ces méthodes se basent uniquement sur les valeurs de la variable à encoder pour la transformation :

Label Encoder : Aux p catégories de la variable, cette méthode associe des valeurs de 0 à $p-1$. Ainsi, elle introduit un ordre pour la nouvelle variable créée. Elle peut être donc non adaptée, si par exemple la variable catégorielle n'est pas ordinale.

One Hot Encoder : Cette méthode transforme la variable X à p modalités en p variables binaires, en créant une variable par catégorie. Pour les modèles sensibles à la multicolinéarité des variables (modèles linéaires simples ou généralisés), l'une des colonnes est à garder comme modalité de référence et seulement $p-1$ colonnes sont créées. Cette méthode est simple à appliquer et peut être très utile lorsque de nouvelles modalités peuvent s'ajouter, puisqu'elle ne nécessite aucune transformation des autres colonnes déjà créées. Cependant, lorsque la variable à encoder a un grand nombre de modalités, par exemple les codes postaux ou les marques des véhicules, elle réduit la capacité du modèle à détecter les dépendances entre la variable à prédire et les colonnes, surtout si le nombre d'observations est faible. Elle augmente également le temps de calcul du modèle puisque le nombre de paramètres à estimer devient très élevé. Pour faire face à ces problèmes, une solution serait de rassembler les catégories à faibles fréquences pour réduire le nombre de colonnes, mais elle reste non optimale puisqu'elle cause une perte de l'information sur ces catégories.

Frequency encoding : Cette méthode consiste à remplacer chaque catégorie par sa fréquence. Bien qu'elle soit simple, cette méthode peut fausser l'analyse, si par

exemple deux catégories à effets différents sur la variable cible sont présentes à proportions égales dans la base.

Encodage supervisé

Ce type d'encodage utilise les valeurs de la variable à modéliser Y pour transformer les catégories. En effet, pour encoder la catégorie x_k^i de la variable i pour l'observation k, une statistique basée sur la valeur de Y est calculée. Souvent, il s'agit d'une approximation de l'espérance de Y sur la même catégorie : $\mathbb{E}(Y|X = x_k^i) = P(Y = 1|X = x_k^i)$ si la variable Y est binaire.

Greedy Target Statistic : Pour approcher l'espérance de la variable cible, cette méthode propose de considérer sa moyenne simple sur les observations appartenant à la même catégorie :

$$\hat{x}_k^i = \frac{\sum_{i=1}^{n} 1_{\{x_j^i = x_k^i\}} y_j}{\sum_{i=1}^{n} 1_{\{x_j^i = x_k^i\}}}$$

Pour les catégories à faible fréquence, cette statique peut être très volatile. Pour lisser ainsi ces valeurs, on considère une combinaison de deux probabilités de Y : une probabilité a priori p, calculée par la moyenne de Y sur toute la base, et une probabilité a posteriori sachant $X = x_k^i$:

$$\hat{x}_k^i = \frac{\sum_{i=1}^{n} 1_{\{x_j^i = x_k^i\}} y_j + ap}{\sum_{i=1}^{n} 1_{\{x_j^i = x_k^i\}} + a} \tag{2.2}$$

où n est le nombre d'observations sur la base d'apprentissage et a est un paramètre à fixer, tel que plus a est élevé, plus cet estimateur est régularisé. Pour la base test, le même encodage des catégories est appliqué si la catégorie est présente dans la base d'apprentissage, sinon c'est la moyenne p qui sera considérée comme valeur pour la catégorie.

Si cette méthode est souvent utilisée dans plusieurs modèles de prédiction, elle présente des problèmes de généralisation. En effet, puisque la variable fait intervenir les valeurs de la variable Y dans sa construction, elle peut causer le surapprentissage du modèle. Les auteurs du CatBoost démontrent ce problème sur un exemple simple que nous allons présenter ici. Supposons que la variable X a des valeurs uniques dans la base d'apprentissage. Alors, la statistique calculée par l'équation 2.2 donne $\hat{x}_k^i = \frac{y_k + ap}{1+a}$. Et si de plus, la probabilité de Y est de $0,5$ pour toutes les catégories, alors en considérant un seuil de $t = \frac{0.5+ap}{1+a}$, toutes les observations de la base

d'apprentissage sont parfaitement prédites. Sur la base test, $x_k^i = p$, puisque les catégories ne sont pas présentes dans la base d'apprentissage. Et en considérant le même seuil t, toutes les observations sont regroupées dans le même noeud et la précision du modèle est de $0,5$ (équivalent à un lancer de pièce).

Plusieurs méthodes ont été proposées pour résoudre ce problème. Elles consistent principalement à exclure l'observation k dans l'équation 2.2 de l'estimateur de x_k^i et de considérer un sous-échantillon $D_k \subset D \setminus \{x_k\}$:

$$\hat{x}_k^i = \frac{\sum_{x_j \in D_k} 1_{\{x_j^i = x_k^i\}} y_j + ap}{\sum_{x_j \in D_k} 1_{\{x_j^i = x_k^i\}} + a} \tag{2.3}$$

Hold-out TS : Pour exclure l'observation sur laquelle l'estimation est effectuée, cette méthode répartit la base d'apprentissage en deux sous-bases : la première servira à calculer la statistique $\hat{x}_k^i$ et le modèle de prédiction se fera sur la seconde. Cette méthode, bien qu'elle résolve le problème de target leakage, nécessite beaucoup d'observations, ce qui n'est pas toujours possible en pratique.

Leave-one-out TS : Pour utiliser toutes les observations de la base, cette méthode propose d'enlever uniquement l'observation sur laquelle on effectue le calcul de la statistique et de considérer l'ensemble : $D_k = D \setminus \{x_k\}$ dans l'équation 2.3. Les auteurs du Catboost démontrent que cette méthode ne permet toujours pas de résoudre le problème de target leakage, en prenant l'exemple d'une variable catégorielle constante, pour laquelle l'estimateur donne :

$$\hat{x}_k^i = \frac{n^+ - y_k + ap}{n - 1 + a}$$

où n^+ est le nombre d'observations de la catégorie x_k telle que $y = 1$. En posant le seuil t à $t = \frac{n^+ - 0.5 + ap}{n - 1 + a}$, alors $x < t \Rightarrow y > 0.5$ et les observations sont parfaitement classées, alors que la variable X n'a aucun effet sur Y, car elle est constante.

Ordered TS : Pour résoudre tous les problèmes évoqués précédemment, le CatBoost emploie une nouvelle stratégie d'encodage, dite **Ordered Target statistic**. Inspirée des modèles séquentiels où l'ordre des observations est important (séries temporelles, l'analyse des textes...), cette méthode crée un temps artificiel en effectuant une permutation σ de la base d'apprentissage. Ensuite, on calcule la statistique de la formule 2.3 en utilisant uniquement les observations précédant celle à encoder $D_k = \{x_j : \sigma(j) < \sigma(k)\}$. Enfin, pour réduire la volatilité de cette estimation, plusieurs permutations sont effectuées.

2.3.2 Ordered boosting

Dans des modèles de prédiction, on cherche à estimer une fonction F qui minimise l'espérance d'une fonction de perte différentiable et convexe $L(y, F(x))$:

$$l(F) = \mathbb{E}L(y, F(x))$$

Le Gradient Boosting est une application de l'algorithme de la descente du gradient aux espaces fonctionnels, en optimisant non pas des paramètres mais des fonctions F^t. En effet, elles sont estimées itérativement en construisant à chaque étape un modèle t sur les résidus du modèle précédent tels que la prédiction est obtenue par $F^t = F^{t-1} + \alpha h^t$. Ces modèles h^t minimisent ainsi la fonction :

$$h^t = \text{argmin}_{h \in H} \mathbb{E}L(y, F^{t-1}(x) + h(x))$$

Pour résoudre ce problème, les modèles de Boosting classiques estiment l'espérance de la perte sur la base d'apprentissage par la moyenne simple sur ces observations :

$$\frac{1}{n} \sum_{i=1}^{n} L(y_i, F^{t-1}(x_i) + h(x_i)) \tag{2.4}$$

En utilisant la formule de Taylor au second ordre à F^{t-1}, l'équation peut s'écrire (voir la partie XGBoost 2.1, en prenant L la somme des carrés des erreurs) :

$$h_t = \text{argmin}_{h \in H} \mathbb{E}(-g^t(x, y) - h(x))^2$$

Où g est le gradient $g^t(x, y) = \frac{\partial L(t,s)}{\partial s}(y_i, F^{t-1}(x_i))$. Et c'est cette dernière quantité qui cause un biais dans la prédiction des modèles de boosting. En effet, ces gradients g^t sont calculés en utilisant les mêmes observations sur lesquelles le modèle F^{t-1} a été estimé. La distribution de ces gradients dans la base d'apprentissage est donc décalée par rapport à leur vraie distribution ce qui biaise les estimations finales de F^t. Pour un cas particulier de régression, les auteurs du Catboost démontrent que lorsque la même base d'apprentissage est utilisée à chaque étape d'optimisation, le modèle est biaisé et le biais est inversement proportionnel à la taille de la base d'apprentissage. Ainsi, les modèles de Boosting peuvent ne pas être adaptés aux petites bases.

Pour résoudre ce problème de généralisation, le même principe de « ordering » utilisé dans l'encodage des variables catégorielles a été appliqué. En effet, après permutation des

observations σ, n modèles seront construits pour chaque itération, tels que chaque modèle j ne prend en entrée que les j premières observations de la base permutée, ce qui permet d'éviter d'estimer le gradient avec la même observation qui l'a construite. Pour réduire la complexité du modèle, c'est une modification de cet algorithme qui est appliqué, en créant $log(n)$ modèles au lieu de n, sur $\{1, ..., log_2(n)\}$.

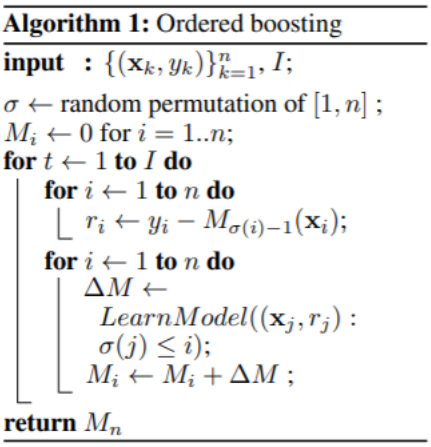


FIGURE 2.2: Algorithme d'*ordered boosting*

Source: CatBoost : unbiased boosting with categorical features, 2018

Cette modification du modèle d'optimisation du Boosting a démontré son efficacité dans les modèles de prédiction. En effet, le Catboost donne de meilleurs résultats par rapport au XGBoost et LightGBM sur un grand nombre de bases de données classiques et permet également de réduire le temps de calcul en construisant des arbres symétriques, tels qu'à chaque niveau de l'arbre, le même critère de division est considéré pour tous les noeuds. Le Catboost permet aussi d'étudier automatiquement les dépendances entre variables catégorielles. Pour cela, sur chaque noeud, les variables catégorielles sont combinées avec les variables présentes dans les noeuds précédents de l'arbre. Enfin, cette méthode permet de traiter les valeurs manquantes. En effet, pour les variables catégorielles, elles sont considérées comme une des catégories alors que pour les variables continues, les valeurs manquantes sont regroupées avec le minimum ou le maximum des observations.

Maintenant que nous avons présenté les modèles que nous allons employer pour classifier les assurés, nous introduisons les critères pour mesurer les performances de chacun de ces modèles et choisir celui qui s'ajuste à nos données.

2.4 Comparaison des modèles de classification

Pour comparer les performances des modèles sur nos deux problèmes de classification, plusieurs méthodes peuvent être employées. Nous allons présenter dans cette partie les métriques d'évaluation les plus utilisées et choisir celle adaptée à notre cas.

2.4.1 Matrice de confusion, précision et rappel

La matrice de confusion, comme son nom l'indique, mesure la «confusion» du modèle en calculant pour les deux classes «positive» et «négative», les observations mal prédites. Pour cela, quatre métriques de classification sont calculées :

TP (resp. TN) le nombre d'observations de la classe positive (resp. négative), correctement classées.

FP (resp. FN) le nombre d'observations de classe négative (resp. positive) prédites comme positives (resp. négative).

Et la matrice de confusion prend la forme suivante :

		Classe réelle	
		Positive	Negative
Classe prédite	Positive	TP	FP
	Negative	FN	TN

TABLE 2.1: Matrice de confusion

Cette matrice nous permet ainsi de visualiser les erreurs du modèle, dans les cases FP et FN. À partir de la matrice de confusion, nous pouvons mesurer la proportion des classes correctement prédites, pour les deux catégories :

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

Si cette métrique est facile à interpréter, elle est mal adaptée aux problèmes où les classes sont non équilibrées. Nous allons illustrer cela par un exemple. Pour 110 observations, telles que 10 sont positives et 100 sont négatives. Si le modèle prédit toutes les classes comme négatives alors : $Accuracy = \frac{100}{110} \approx 91\%$: Donc bien que le modèle ne distingue pas entre les deux classes, la métrique est élevée.

En se basant sur cette matrice, d'autres métriques peuvent être calculées :

Recall $= \frac{TP}{TP+FN}$: parmi les observations positives, le taux des observations correctement classées.

Precision $= \frac{TP}{FP+TP}$: parmi les observations classées comme positives, lesquelles sont effectivement des observations positives.

Choisir parmi la précision et le rappel dépend de l'objectif de l'étude. En effet, si l'erreur FN qui consiste à prédire des classes positives comme négatives a plus d'impact, comme est le cas par exemple pour les dépistages des maladies, alors le coût des FN est plus élevé que les FP, et le rappel est la métrique de référence.

Un troisième score permet de considérer un compromis entre les deux dernières métriques, en prenant leur moyenne harmonique. Il s'agit du **F1-score**, qui est maximal lorsque les deux erreurs FP et FN sont faibles, et permet ainsi de comparer les modèles en terme de précision et rappel simultanément en donnant aux deux scores la même importance.

2.4.2 Courbe ROC et AUC

Les scores précédents dépendent d'un seuil de probabilité à fixer, tel que les observations sont classées comme positives si la probabilité prédite dépasse ce seuil. Pour considérer la performance du modèle pour différents seuils $\in$ [0,1], la courbe **ROC** représente les taux de vrais positifs (TPR ou rappel) en fonction des taux de faux positifs (FPR) pour ces seuils. Elle permet ainsi de savoir si un modèle surperforme un autre quelque soit le seuil considéré, et permet également une visualisation de ces résultats.

Pour résumer l'information contenue dans la courbe ROC, l'aire sous cette courbe, dite **AUC**, est calculée. Cette métrique permet une comparaison entre les modèles, pour tous les seuils de classification possibles et mesure la capacité du modèle à distinguer les deux classes. En effet, dans "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve", A. Hanley et J. McNeilElle démontrent que l'AUC mesure, dans le cas d'un tirage alétoire d'une observation positive et d'une négative, la probabilité que le modèle classe correctement ces deux observations. De plus, puisqu'elle prend des valeurs entre 0 et 1, une approche standard est possible pour comparer la performance des modèles, selon les valeurs de l'AUC : si l'AUC dépasse 0,5, et plus il est proche de 1, plus il permet de classifier les catégories mieux qu'un classifieur aléatoire. C'est ainsi que nous allons considérer cette dernière métrique comme mesure pour comparer nos modèles de classification pour les deux variables d'orientation et de réparation.

Chapitre 3

Modélisation de l'orientation et de la réparation en garages agréés

3.1 Orientation en garages partenaires

Dans cette partie, nous nous intéresserons à la première variable de notre sujet. Il s'agit de l'orientation en garages partenaires. En effet, au moment de la déclaration du sinistre, bien que l'agent doive en premier lieu rappeler au client son droit de choisir le garage pour effectuer la réparation, il peut également lui conseiller de réparer dans un des garages partenaires de l'assureur pour bénéficier de différents avantages. Nous allons donc décrire le phénomène d'orientation en garages partenaires, comprendre les raisons pour lesquelles un agent pourrait ne pas orienter un assuré vers l'un de ces garages pour enfin construire des modèles de prédiction de ce score pour chaque contrat.

3.1.1 Le choix du garage, un mécanisme en deux étapes

Après un sinistre, l'assuré entre en contact avec son agent pour lui déclarer l'accident. À cette étape, l'agent décide d'orienter son assuré vers un garage partenaire ou vers un garage non-partenaire. Nous nous intéressons particulièrement à cette variable puisqu'elle devrait influencer le choix final de l'assuré. En effet, un assuré qui a été orienté vers un garage partenaire devrait plus probablement effectuer la réparation dans ce dernier. Nous allons donc modéliser cette variable et l'intégrer dans le modèle de réparation pour tenir compte du choix de l'agent dans la décision finale de l'assuré.

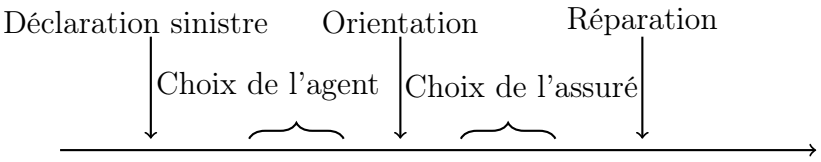


FIGURE 3.1: Choix du garage

3.1.2 Description de la variable d'orientation

Notre base de données des sinistres, dont nous avons décrit la construction précédemment, recense les contrats arrivés à leur échéance durant l'année 2018 et ayant eu un ou plusieurs sinistres durant un an d'observation. À cette base de données, nous ajoutons la variable d'orientation, lorsqu'elle est disponible. Cette variable décrit, pour chaque sinistre, si l'agent a orienté l'assuré vers un garage partenaire ou non.

La base de données obtenue ainsi contient **85 230 contrats sinistrés**, l'information relative à chaque contrat et le nombre de sinistres survenus. Nous calculons ensuite, pour ces contrats, la proportion des sinistres orientés. Les moyennes d'orientation obtenues sont résumées dans le tableau 3.1 suivant :

Moyenne d'orientation	Nombre d'observations
1.0	70141
0.0	14040
0.5	882
0.67	87
0.33	61
0.75	13
0.8	4
0.2	1
0.25	1

TABLE 3.1: Moyenne des orientations des contrats sinistrés

La majorité des sinistres de notre base de données sont des sinistres orientés vers les garages partenaires. Nous remarquons également que les contrats dont les sinistres ont été partiellement orientés représentent un faible pourcentage de la base de données, environ 1,23% des contrats. De plus, ce sont les contrats qui n'ont pas été orienté qui nous intéressent le plus, puisqu'ils engendrent des coûts élevés. C'est ainsi que nous rassemblons les contrats partiellement orientés avec les contrats orientés pour créer une variable binaire Y telle que :

$$\begin{cases} Y = 0 : \text{Le contrat dont les sinistres ne sont jamais orientés par l'agent.} \\ Y = 1 : \text{Le contrat dont les sinistres ont été orientés au moins une fois par l'agent.} \end{cases}$$

La moyenne de cette variable obtenue ainsi est de 83,57%. Le jeu de données est donc déséquilibré et la classe des contrats orientés est la plus répandue. Avant de commencer la

modélisation de cette variable, nous allons effectuer une analyse univariée des différentes variables dont nous disposons.

Les variables d'agent

Nous disposons dans notre base de données de 28 variables décrivant les caractéristiques de l'agent responsable du contrat : son nom et identifiant, l'adresse de l'agence, le numéro de portefeuille ... Nous nous intéressons particulièrement à ces variables puisque c'est l'agent qui choisit d'orienter ou pas un assuré. Avant de modéliser l'orientation en considérant l'ensemble des variables de la base de données (sur l'assuré, le véhicule, l'agent et la police), nous testons l'hypothèse selon laquelle l'orientation dépend entièrement de l'agent, et non pas du profil de l'assuré. En effet, si cette hypothèse est vérifiée, chaque agent choisit soit d'orienter tous ses clients, soit de ne pas les orienter et il suffirait ainsi de prendre les décisions antérieures de chaque agent comme prédiction finale du score d'orientation. Pour tester cela, nous calculons les moyennes et écarts-type d'orientation pour chaque numéro d'agent. Nous représentons dans le tableau 3.2 ces taux pour les agents à nombre significatif de contrats :

Numéro d'agent	Moyenne d'orientation	Écart-type	Nombre d'observations
1	0.796	0.403	1412
2	0.777	0.416	767
3	0.794	0.404	481
4	0.208	0.407	139

TABLE 3.2: Orientation par numéro d'agent

Nous remarquons qu'on s'écarte d'une orientation parfaite (0 ou 1 comme moyenne) par agent et que les écarts-type des orientations sont significatifs. Donc, l'orientation ne dépend pas uniquement de l'agent mais également d'autres variables. Nous rejetons ainsi cette première hypothèse.

En utilisant les adresses fournies dans la base de données, nous représentons sur la carte l'ensemble des agents AXA. Ces agences sont réparties sur l'ensemble du territoire français, avec une forte concentration observée dans la région d' Île-de-France.

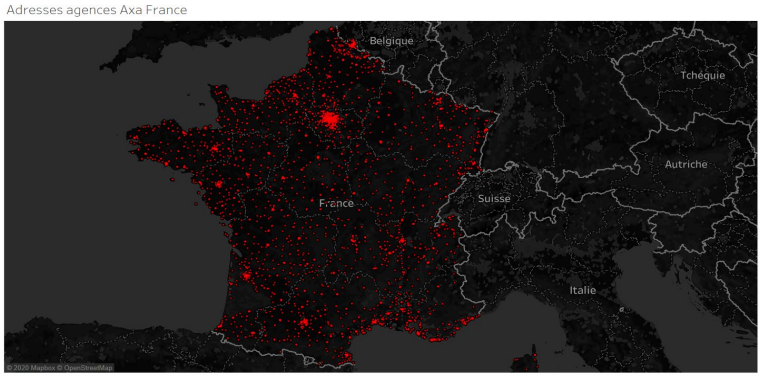


FIGURE 3.2: Agences contactées après sinistres

Parmi les variables dont nous disposons, nous avons une variable *Segment* qui classe les agences selon le nombre de garages partenaires à proximité. Nous comparons les moyennes des orientations selon cette variable dans le tableau 3.3 et remarquons qu'en partant du segment A, qui représente la classe des agents avec un grand nombre de garages partenaires à proximité, au segment E, le taux d'orientation diminue. De plus, le volume des observations de chaque classe est significatif. Nous pouvons déduire donc que le nombre de garages agréés à proximité des agents pourrait expliquer leur propension à orienter leurs clients vers ces garages.

Segment	Moyenne d'orientation	Nombre d'observation
E	0.70	11834
D	0.81	13361
C	0.87	40621
B	0.92	6132
A	0.95	842

TABLE 3.3: Orientation par segment d'agent

Les variables du véhicule

Dans notre base de données, nous disposons de 57 variables décrivant le véhicule de l'assuré : la marque du véhicule, le prix, la cylindrée, la date de mise en circulation... Nous incluons ces variables dans nos modèles d'orientation puisque c'est le véhicule qui fera l'objet de la réparation. De plus, selon le type du véhicule, les garages sont plus ou moins adaptés pour effectuer la réparation. En effet, dans l'absence de spécialiste agréé, l'agent pourrait orienter son client vers un garage concessionnaire qui connaît la marque et qui dispose des pièces de rechange adaptées au véhicule.

Pour réduire le nombre de ces variables, nous commençons par étudier les corrélations entre elles. Nous représentons un heatmap de la valeur absolue des corrélations de Pearson (voir graphique A.1), et remarquons que quelques variables sont très corrélées entre elles. Ainsi, pour tout couple de variables où le coefficient de corrélation dépasse 0,9, nous gardons une seule variable.

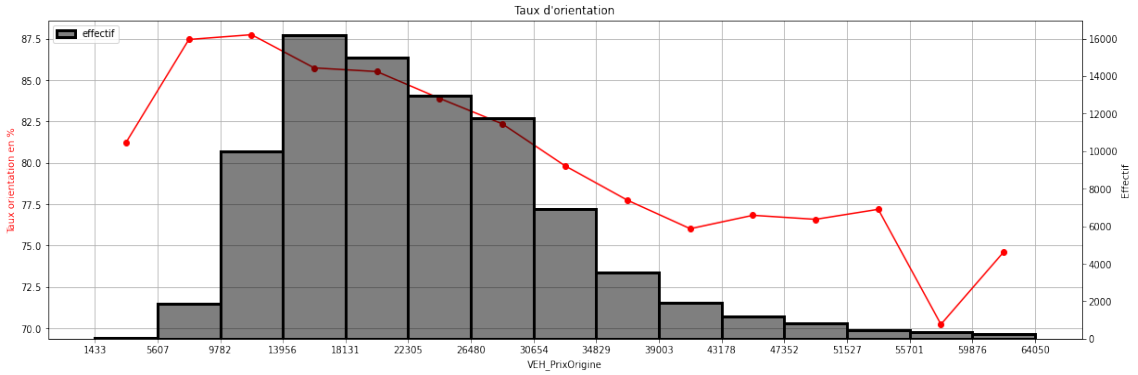


FIGURE 3.3: Moyenne des orientations par prix du véhicule

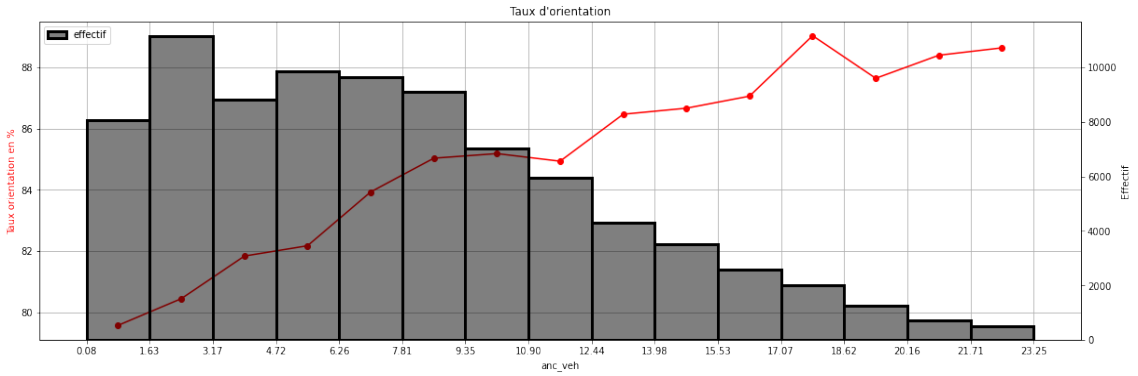
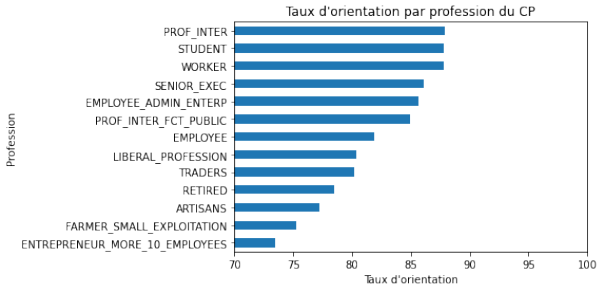


FIGURE 3.4: Moyenne des orientations par ancienneté du véhicule

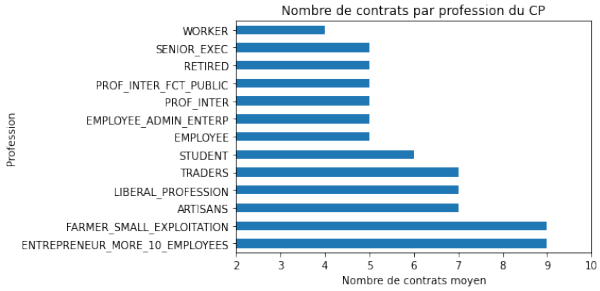
Nous représentons les taux d'orientation selon quelques caractéristiques du véhicule (figures 3.3, 3.4, A.2 et A.3) et remarquons que plus le véhicule est cher, moins il est orienté vers un garage agréé. De plus, les nouvelles voitures, souvent équipées de pièces électroniques, sont également moins orientées que les anciennes. De même, les taux d'orientation décroissent avec la puissance et le poids du véhicule, deux variables étroitement liées à la marque (voir A.4). Nous allons ainsi inclure les variables relatives au véhicule dans notre modélisation de l'orientation pour tenir compte de ces effets.

Les variables de l'assuré

237 variables de notre base de données décrivent les caractéristiques de l'assuré : l'âge, la profession, le nombre de contrats détenus par type de garantie... Après réduction du nombre de variables, en enlevant une variable si elle est fortement corrélée à une autre (voir matrice des corrélations A.5), et en ne gardant que la variable globale si elle résume l'information contenue dans une autre (par exemple, nombre de contrats et nombre de contrats souscrits sur une année), nous commençons par comparer les moyennes des orientations selon les variables gardées. Dans les graphiques A.6, A.7 et A.8, nous représentons la moyenne des orientations selon la distribution des trois variables : âge, ancienneté du client et le nombre de contrats, tous types de garanties confondus. Nous remarquons que le taux d'orientation est décroissant selon ces trois variables. Nous calculons également la moyenne des orientations par profession, et lions cette dernière variable au nombre de contrats :



(a) Moyenne d'orientation par profession



(b) Nombre de contrats par profession

Selon la profession de l'assuré, les agents choisissent plus ou moins de l'orienter vers des garages agréés. En effet, les professions dont les assurés détiennent le plus de contrats (entrepreneurs, agriculteurs...) sont les moins orientés, ce qui explique les taux des orientations selon les professions et la forme du graphique A.8. De cette première analyse, nous tirons une première conclusion : plus l'assuré entre en contact avec l'agent, moins il est orienté pour effectuer sa réparation dans un garage agréé. Pour comprendre cet effet, il est nécessaire d'analyser le fonctionnement des agents. Ces mandataires de la société d'assurance perçoivent des commissions selon le volume des contrats gérés et de leur valeur et ainsi fidéliser leurs clients est primordial. Et pour assurer que les clients qui détiennent beaucoup de contrats soient satisfaits des réparations, l'agent les oriente vers des garages connus pour leur qualité de service bien qu'ils ne soient pas partenaires.

Présélection des variables

Dû au nombre élevé des variables présentes dans notre base de données et pour réduire la complexité des modèles de prédiction, une première présélection des variables est primordiale. Pour ce faire, nous allons créer des forêts aléatoires sur chaque catégorie de variables : sur les variables d'agent, de police, de véhicules... et garder les variables significatives sur chaque catégorie.

Le modèle de forêts aléatoires a été privilégié puisqu'il introduit de l'aléatoire dans le choix des variables dans chaque noeud et est robuste face aux variables colinéaires. Cette première étape va nous permettre ainsi de réduire le nombre de variables et de sélectionner les plus pertinentes pour notre modélisation de l'orientation.

Pour classer les variables et choisir les plus significatives, nous nous sommes basés sur la *mean SHAP value* qui mesure la contribution marginale moyenne de chaque variable dans les prédictions finales. En effet, la méthode SHAP (SHapley Additive exPlanation) est un algorithme permettant d'interpréter les résultats des modèles complexes. Cette méthode est basée sur les valeurs de Shapley, un concept issu de la théorie des jeux. En effet, ces valeurs permettent de quantifier l'effet marginal d'une variable (le joueur) sur la prédiction finale pour chaque observation (le jeu) pour enfin prendre la moyenne de cette contribution sur l'ensemble des observations. Cette méthode a été préférée aux importances des variables données par défaut par les forêts aléatoires, les métriques *mean decrease in impurity*, puisque ces dernières donnent plus de poids aux variables continues ou à cardinalité élevée (voir [14]).

Les résultats de cette présélection sont représentées dans les graphiques A.9 à A.12 et les performances des modèles obtenus sont résumées dans le tableau A.1. Cette méthode nous permet de réduire le nombre de variables de 580 à 65 variables, sur lesquelles nous allons tester et comparer les trois modèles : forêts aléatoires, XGBoost et CatBoost, combinés avec différentes méthodes d'encodage des variables catégorielles.

3.1.3 Modélisation et résultats

Maintenant que nous avons fixé les variables explicatives et les modèles à tester, nous modélisons notre variable d'orientation et choisissons le modèle qui s'ajuste le mieux à nos données. Comme expliqué dans la section 2.4.2, nous allons nous baser sur la métrique

AUC pour cette comparaison.

Nous commençons par tester les deux modèles : forêts aléatoires et XGBoost, avec deux méthodes d'encodage : one hot encoding et label encoder. Pour ces deux modèles, il est important de fixer les hyperparamètres relatifs aux arbres pour améliorer leurs performances sur la base test tout en contrôlant tout en contrôlant le sur-ajustement sur la base d'apprentissage. Pour cela, nous avons effectué un *Random search*, en testant différentes valeurs pour chaque paramètre pour en déduire le meilleur paramétrage. Les résultats de cette recherche de paramètres pour les forêts aléatoires sont présentés dans les graphiques A.13a à A.14c.

Les AUC des différents modèles testés sont résumés dans le tableau suivant :

	Modèle	**AUC**
Forêts aléatoires	*Onehotencoder*	0,680
	LabelEncoder	0,685
XGBoost	*Onehotencoder*	0,687
	LabelEncoder	0,695

TABLE 3.4: Comparaison des modèles

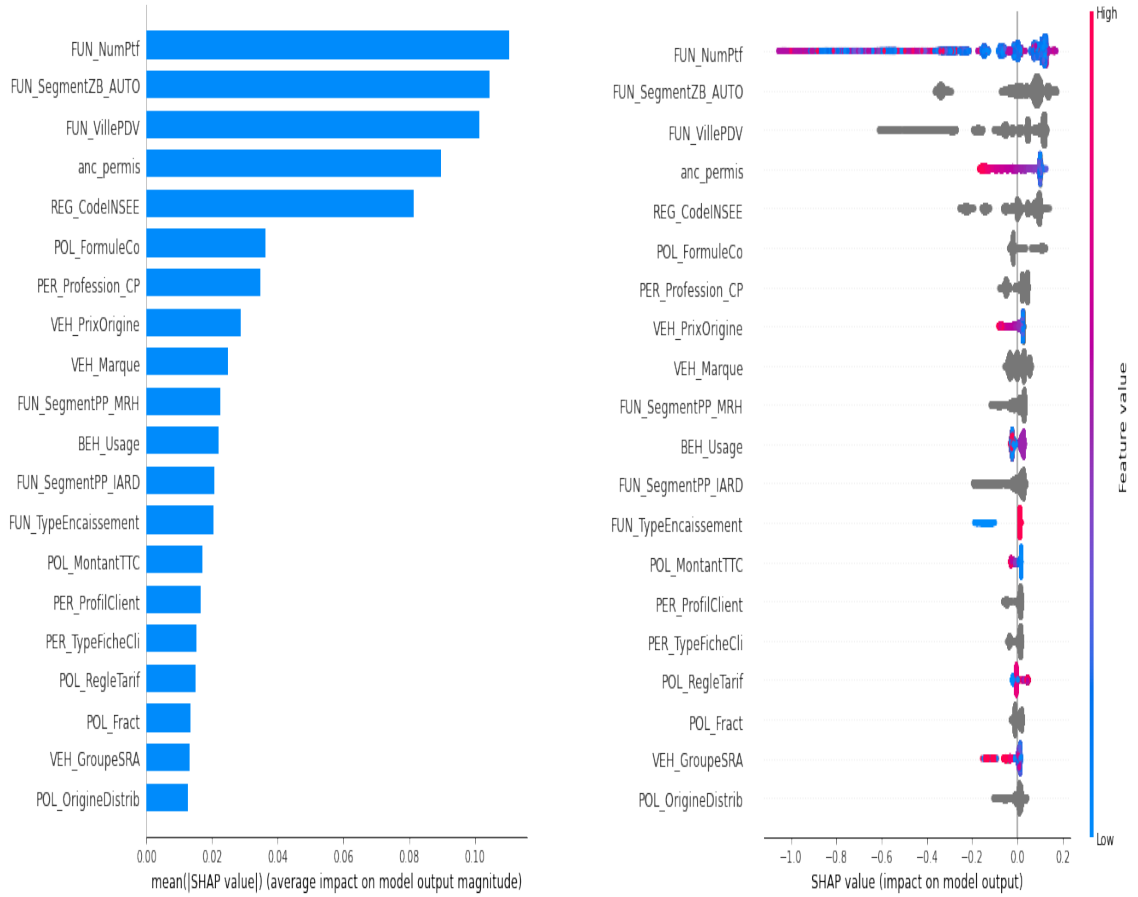
Nous remarquons que la deuxième méthode d'encodage LabelEncoder permet d'améliorer le modèle. En effet, les variables de notre base de données sont majoritairement qualitatives à grand nombre de modalités. Ainsi, le *onehotencoder*, contrairement au *labelencoder*, créé beaucoup de colonnes ce qui réduit la performance du modèle.

Étant donné cette prédominance des variables catégorielles et pour éviter de supposer un ordre entre les modalités, nous testons le modèle CatBoost. En plus d'un encodage automatique des variables, le CatBoost permet de suivre l'évolution de l'AUC par nombre d'arbres (figure A.15) et d'appliquer un critère d'arrêt pour éviter le surapprentissage. L'AUC après optimisation des hyperparamètres du modèle est de $\mathbf{0,705}$, et donc dépasse ceux des deux modèles précédents.

Enfin, nous allons incorporer les variables identifiantes de l'agent dans ce dernier modèle (son numéro et adresse), puisqu'il est le maillon clé de la variable d'orientation. Cela permet d'améliorer la performance du dernier modèle de CatBoost, et l'AUC passe de

0,705 à $\mathbf{0,717}$. L'importance des variables de ce modèle 3.6 nous confirme l'importance de l'agent dans l'étape d'orientation des sinistres. En effet, les trois variables les plus significatives sont relatives à ce dernier (variables précédées par "FUN"). Selon le numéro de l'agent, son emplacement et la densité des garages partenaires à sa proximité, ce dernier choisit plus ou moins d'orienter ses clients vers des garages partenaires.

Pour voir le sens de la contribution de chaque variable dans l'orientation, nous représentons les SHAP values. La couleur rouge indique les valeurs élevées de la variable explicative alors que l'axe des abscisses représente la différence de la prédiction de l'orientation par rapport à la moyenne. Cette analyse des effets des valeurs des variables se fera sur les variables numériques uniquement. Pour les variables catégorielles, c'est l'ampleur de l'effet qui nous intéressera. Nous remarquons que les assurés à anciennetés de permis élevées et les véhicules les moins chers ont un faible taux d'orientation, ce qui confirme ce que nous avons observé dans les graphiques B.1 et B.4.



(a) (b)

FIGURE 3.6: SHAP values

Maintenant que nous maîtrisons les paramètres d'orientation pour les assurés, nous allons étudier l'effet de cette orientation sur le choix final de l'assuré en modélisant ce dernier en fonction du score d'orientation et des autres variables de la base de données.

3.2 Réparation en garages partenaires

Cette deuxième partie est consacrée à la modélisation de la variable de réparation en garages partenaires. Nous allons en premier lieu analyser les facteurs pouvant expliquer le choix du garage. Ensuite, nous allons comparer les différentes variables de notre base selon le choix de réparation de l'assuré. Enfin, nous modéliserons cette variable en incluant comme prédicteur le score d'orientation prédit pour chaque contrat.

3.2.1 Première analyse du choix de l'assuré

Après survenance du sinistre et orientation de l'agent, l'assuré dont le véhicule a été endommagé décide du garage où sera effectuée la réparation. Ce choix peut dépendre de plusieurs facteurs. En garages partenaires, et selon chaque compagnie d'assurance, il bénéficie de nombreux avantages. En effet, un des points forts de ce type de garage est la non-avance des frais. Le règlement des coûts de réparation est directement effectué par l'assureur, et l'assuré doit verser uniquement la franchise du contrat souscrit. De plus, pour l'obtention d'un agrément, l'assureur exige au garage de répondre à un nombre de critères de qualité pour avoir la garantie d'un travail bien fait. Ensuite, en déposant son véhicule dans un des garages partenaires, l'assuré bénéficie d'un véhicule de prêt pour assurer sa mobilité durant la période de réparation.

Face à ces nombreux avantages, et pour renforcer le libre choix de l'assuré, un mécanisme de cession de créance a été instauré pour les garages non partenaires. Ce dernier consiste en un contrat qui permet le versement direct des coûts de réparation au garage par l'assureur. Cependant, la signification par huissier pour conclure ce contrat était exigée, ce qui réduisait l'efficacité du mécanisme. Mais depuis 2016, cette procédure a été simplifiée et il suffit de notifier l'assureur par lettre recommandée pour conclure le contrat avec un garage hors du réseau partenaire. Parmi les autres facteurs qui peuvent pousser l'assuré à réparer dans un garage libre est la rapidité et le soin du service, puisque le rythme de travail est beaucoup moins soutenu pour les garages non agréés qui ne reçoivent pas les véhicules des assureurs. L'assuré peut également souhaiter d'effectuer la réparation chez son garage habituel. Enfin, ces garages indépendants peuvent offrir également d'autres avantages pour garder une longueur d'avance sur la concurrence. C'est l'exemple d'un des réseaux de réparation, *ZeCarrossery*, qui en plus d'un système de fidélité, propose de rembourser une partie de la franchise aux assurés, pour les garanties tous risques ou bris

de glace, si le sinistre en question est responsable .

De cette partie, nous comprenons que le choix du garage de réparation est un phénomène complexe lié à plusieurs paramètres. Et étant donné ses conséquences sur les coûts de réparation, il est important de détecter les déterminants de ce choix pour l'anticiper et segmenter selon ce dernier. C'est ainsi que nous allons modéliser cette variable par des méthodes d'apprentissage automatique pour capter l'interaction des différentes variables de notre base de données et prédire ce choix de réparation.

3.2.2 Statistiques descriptives

La variable binaire de réparation, indique pour chaque sinistre, si la réparation a été effectuée dans un des garages partenaires d'AXA. Puisque l'information sur le sinistre est inconnue au moment du renouvellement du contrat, nous allons passer en vision contrat, en prenant la moyenne de cette variable pour chaque assuré. Ainsi, nous créons la variable à modéliser Y telle que :

$$\begin{cases} Y = 0 : \text{Tous les sinistres du contrat ont été réparés dans un garage non partenaire.} \\ Y = 1 : \text{Le contrat dont les sinistres ont été réparés au moins une fois dans un garage partenaire.} \end{cases}$$

Dans notre base de données, la proportion des sinistres réparés en garages agréés dépasse celle des non agréés, avec une moyenne de 59,63%.

Puisque les variables présentes dans la deuxième base de données sont similaires aux variables de la première, les mêmes étapes de présélection basées sur les variables corrélées seront appliquées et nous nous contentons ici d'analyser les liens entre ces variables et la variable de réparation créée Y.

Nous commençons par étudier les variables de l'assuré, puisque c'est ce dernier qui choisit le garage où effectuer les réparations. Nous représentons les taux de réparation en garage partenaire selon ces variables dans les figures B.1 à B.3. Ces taux sont décroissants selon l'âge de l'assuré et le nombre de contrats souscrits : nous constatons qu'à partir de 50 ans, les assurés réparent moins en garages partenaires par rapport à la moyenne. L'effet de l'âge constaté est dû à plusieurs facteurs, par exemple, pour les plus âgés, un revenu plus élevé leur permet de choisir le garage non agréé puisque le paiement des coûts de réparation en avance est moins contraignant que pour les plus jeunes. De plus, cette

catégorie pourrait comporter la clientèle habituelle des garagistes indépendants, fidèle à ces derniers. Nous remarquons aussi que les taux de réparation en garages partenaires sont décroissants en fonction du nombre de contrats souscrits et selon la valeur du client, variable qui attribue une valeur à chaque assuré selon le nombre et la valeur des contrats détenus par ce dernier.

Ensuite, nous comparons les taux de réparation selon les caractéristiques du véhicule assuré dans les figures B.4 à B.6. Nous remarquons que les véhicules les moins chers et anciens sont les plus réparés chez des garages partenaires. Cela peut être dû au fait que ces véhicules ne nécessitent aucune réparation complexe ou une intervention de spécialistes. Nous constatons aussi que le taux de réparation est décroissant avec le poids du véhicule, variable liée à la marque (voir A.4).

Enfin, nous représentons les taux de réparation selon le score d'orientation précédemment modélisé :

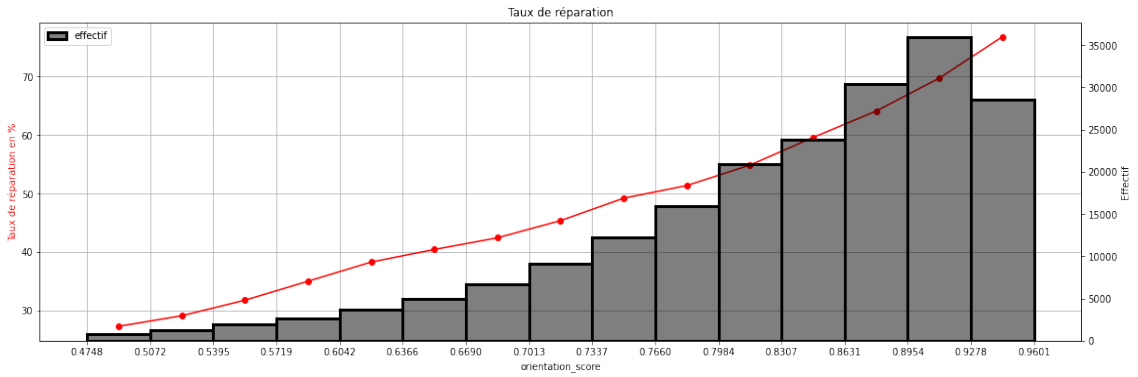


FIGURE 3.7: Moyenne des réparations en garage partenaire par probabilité d'orientation du sinistre

Nous remarquons la croissance des taux de réparation selon la probabilité de l'orientation par l'agent vers un garage partenaire. Ainsi, plus le client est susceptible d'être orienté vers un carrossier agréé après survenance d'un sinistre, plus il a tendance à réparer effectivement dans ces garages, ce qui est conforme à notre hypothèse de départ formulée dans la partie 3.1.1. Nous allons ainsi inclure ce score comme variable explicative dans la prédiction de la variable de réparation et étudier sa significativité dans ce choix de garage.

Avant de modéliser le score de réparation et pour inclure des variables pertinentes dans le choix du garage, nous allons procéder à la présélection des variables. Pour cela, nous allons procéder de façon similaire à celle appliquée pour l'*orientation*. Nous allons

modéliser par des forêts aléatoires la réparation en garages partenaires Y avec les variables de chaque catégorie et garder les plus significatives sur chaque modèle. Et comme critère d'importance des variables, la *mean SHAP value* sera employée. Les résultats de ces modélisations sont représentés dans les graphiques B.7 à B.10. Cette étape de présélection nous permet de réduire le nombre de variables de 323 à 43 variables, sur lesquelles nous allons appliquer nos trois modèles de classification choisis : forêts aléatoires, XGBoost et CatBoost.

3.2.3 Modélisation et résultats

Après avoir fixé les variables à inclure dans les modèles de réparation, nous testons nos modèles de prédiction. Pour le modèle de forêts aléatoires, nous testons deux méthodes d'encodage : label encoder et target encoder, décrites dans la partie 2.3.1. L'optimisation des hyperparamètres pour ces deux modèles est représentée dans les figures B.11a à B.11f. En comparant les performances des modèles avec les deux méthodes d'encodages des variables catégorielles, nous pouvons remarquer le surapprentissage avec la méthode target encoding. En effet, en utilisant la variable Y pour encoder les catégories, on introduit une fuite des données de la variable cible et l'algorithme devient très dépendant des variables encodées. Ainsi, en appliquant le modèle sur de nouvelles observations, sa performance est beaucoup moins élevée qu'en utilisant un encodage simple.

Les performances des deux modèles forêts aléatoires et XGBoost, mesurées par l'AUC, sont résumées dans le tableau suivant :

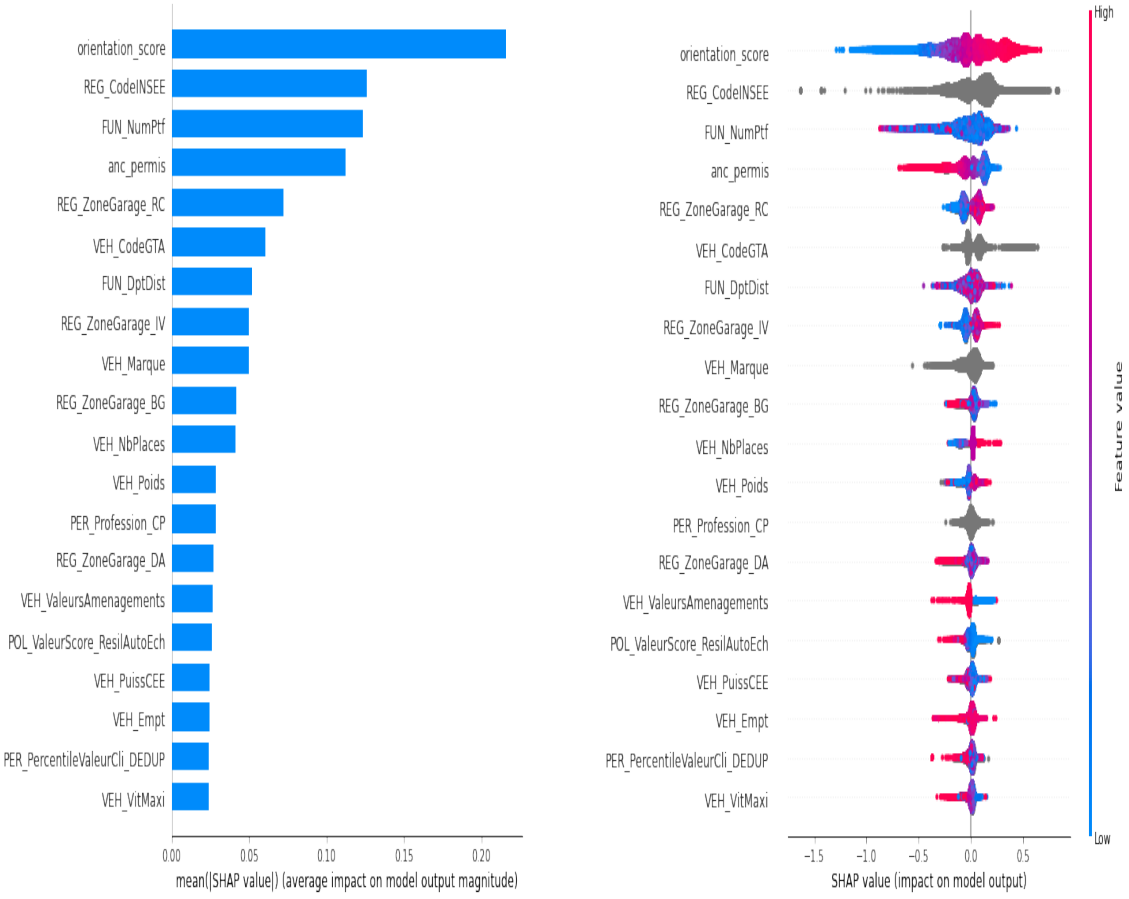
	Modèle	**AUC**
Forêts aléatoires	$LabelEncoder$	0,658
	$TargetEncoder$	0,632
XGBoost	$LabelEncoder$	0,655
	$TargetEncoder$	0,599

TABLE 3.5: Comparaison des modèles

Nous remarquons que le modèle des forêts aléatoires est un peu plus robuste face au surapprentissage causé par le target encoder que le modèle XGBoost. Ceci est dû à l'échantillonnage aléatoire des variables et observations sur chaque noeud. Nous remarquons également qu'en utilisant le label encoder, les deux modèles forêts aléatoires et

XGBoost donnent des performances similaires.

Nous testons enfin le modèle CatBoost sur le modèle de réparation. Après optimisation des hyperparamètres, nous obtenons un AUC de 0,68 (voir B.12). Pour comparer les contributions des variables explicatives dans ce modèle, nous représentons les shap values et leurs moyennes dans la figure 3.8 suivante :



(a) (b)

FIGURE 3.8: SHAP values

Nous remarquons que le score d'orientation que nous avons modélisé précédemment est la variable la plus significative dans le modèle de réparation. De plus, comme nous avons constaté dans la figure 3.7, la probabilité du choix de garage partenaire est croissante avec la probabilité d'orientation prédite. De même, le taux de réparation est plus élevé chez les assurés à faible ancienneté de permis (voir la figure B.1). Ensuite, c'est la région d'habitation de l'assuré qui justifie en second lieu le choix du garage. Cette variable région peut être liée à plusieurs facteurs : la non-disponibilité de garages agréés dans cette région, un niveau de vie faible poussant les assurés à réparer dans les garages agréés pour bénéficier de la non-avance des frais...

Pour réduire le temps de calcul et avoir un modèle moins complexe, nous construisons le modèle Catboost sur les variables les plus significatives uniquement présentes dans la figure 3.8. Ce dernier modèle donne un AUC de 0,677, ce qui caractérise un niveau de discrimination moyen entre les deux classes de réparation. Cela peut être dû au manque de variables étroitement liées au choix du garage. En effet, si l'orientation par l'agent explique une partie de cette décision, d'autres variables comme le nombre de garages à proximité de l'assuré, le temps de réparation de ces derniers par rapport aux garages libres, leur qualité de service... sont absentes dans notre base de données. De plus, les informations dont nous disposons et relatives aux garages concernent uniquement les deux années 2018 et 2019, et les observations des assurés s'étendent sur une année uniquement ce qui est insuffisant pour étudier l'évolution de ce phénomène.

3.3 Conclusion

Dans cette partie, nous avons modélisé les deux variables d'orientation et de réparation en garages partenaires. Notre objectif était de prédire pour chaque contrat ces deux scores au moment du renouvellement du contrat. C'est ainsi que nous avons privilégié des modèles nécessitant peu de traitement poussé des observations tout en profitant du maximum d'informations contenues dans cette base de données. Et étant donné la prédominance des variables catégorielles dans cette dernière, nous avons choisi de tester le modèle CatBoost.

Compte tenu des variables de la base de données, ce dernier modèle donne une meilleure performance par rapport aux deux modèles des forêts aléatoires et XGBoost testés. Cependant, la précision des scores prédits peut toujours être améliorée. En effet, étudier ces scores sur un intervalle temporel plus large et intégrer des variables externes ayant un effet direct sur le choix de réparation (revenu de l'assuré, les types de garage des réparations précédentes, les types de garage à proximité du domicile de l'assuré...) permettrait aux modèles de distinguer plus les deux classes d'assurés.

Maintenant que nous pouvons prédire pour chaque assuré s'il choisirait un garage agréé dans le cas d'un sinistre, nous passons à la dernière étape de notre étude qui consiste à employer ce score pour différencier les tarifs proposés et les ajuster aux deux vrais risques.

Chapitre 4

Modélisation de la sévérité par type de réparation

4.1 Objectif

Dans cette partie, nous justifions l'utilité de notre modèle de prédiction du choix du garage de réparation dans l'amélioration des tarifs proposés. Étant donné que ce choix intervient après la survenance du sinistre et impacte principalement les coûts de réparation, nous allons nous intéresser à la modélisation de la sévérité des sinistres, en particulier des sinistres de type bris de glace, garantie prédominante dans notre base de données [1]. Deux sévérités seront ainsi comparées, une sévérité où la propension de réparer en garages partenaires de l'assuré n'est pas considérée comme critère de segmentation et une seconde de la forme :

$$\begin{aligned}\mathbb{E}(\text{Sévérité}) = \mathbb{E}(\text{Sévérité agréé}) \times \text{Probabilité de réparer en garage agréé} \\ + \mathbb{E}(\text{Sévérité non agréé}) \times (1 - \text{Probabilité de réparer en garage agréé})\end{aligned} \tag{4.1}$$

Pour modéliser ces sévérités et prendre en compte les effets non linéaires que peuvent avoir certaines variables tarifaires sur les coûts de réparation, nous allons introduire les modèles additifs généralisés (**GAM**). En effet, pour la tarification des contrats, les assureurs se basent souvent sur des modèles linéaires généralisés (**GLM**) qui supposent une relation linéaire entre l'espérance de la variable à prédire Y et les variables explicatives $X_1, X_2 \dots X_P$, ce qui n'est pas toujours vérifié en pratique.

Pour capter les effets non linéaires entre les variables continues et la variable réponse, une solution souvent adoptée par les assureurs est la création de groupes homogènes de risque tout en gardant des poids significatifs pour chaque classe pour assurer une mutualisation entre les assurés. Cependant, le choix de ces classes est souvent basé sur l'expertise de l'actuaire. Nous allons présenter ainsi une méthode orientée données de

1. Les garanties de notre base de données sont résumées dans le tableau C.1.

discrétisation des variables explicatives continues, proposée par K.Antonio et al. [17] et basée sur les modèles additifs généralisés et les arbres de régression.

4.2 Des modèles GLM aux GAM

Le modèle linéaire généralisé ou GLM est une généralisation du modèle linéaire gaussien. En effet, ce dernier suppose que la variable réponse est gaussienne et donc ne permet pas de modéliser des variables discrètes telle que la fréquence des sinistres, ou des variables continues positives asymétriques, ce qui est souvent le cas pour les sévérités. De plus, le modèle gaussien suppose un effet additif des variables explicatives, ce qui n'est pas toujours souhaitable. Le modèle linéaire généralisé résout ainsi ces problèmes en proposant plus de lois pour la modélisation de la variable réponse et une fonction de lien entre l'espérance de cette variable et les variables explicatives. Un modèle GLM s'écrit alors :

$$g(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

Où : Y est la variable réponse suivant une loi de distribution appartenant à la famille exponentielle [2], X_i sont les variables explicatives ou la composante déterministe et g est la fonction de lien entre ces deux composantes aléatoire et déterministe.

Un modèle GLM sera ainsi défini par le choix de ces trois composantes. Pour notre cas, étant donné que nous modélisons des sévérités continues, positives et asymétriques à gauche, nous nous intéresserons aux deux lois Gamma et log-normale et nous allons considérer le logarithme de la sévérité pour obtenir des effets multiplicatifs des variables tarifaires. Il faut noter que la loi log-normale n'appartient pas à la famille exponentielle et que pour modéliser la sévérité selon cette loi, il faut modéliser le logarithme de la sévérité selon la loi gaussienne appartenant à la famille exponentielle.

Le modèle additif généralisé est une extension du modèle GLM pour considérer des effets plus complexes des variables continues sur la variable Y. Il s'écrit ainsi :

$$g(\mathbb{E}[Y]) = \beta_0 + \sum \beta_j X_j^d + \sum f_l(X_l^c)$$

Où : X_j^d sont les variables binaires indiquant l'appartenance à une modalité d'une variable qualitative, X_l^c sont les variables explicatives continues et f_l sont des fonctions

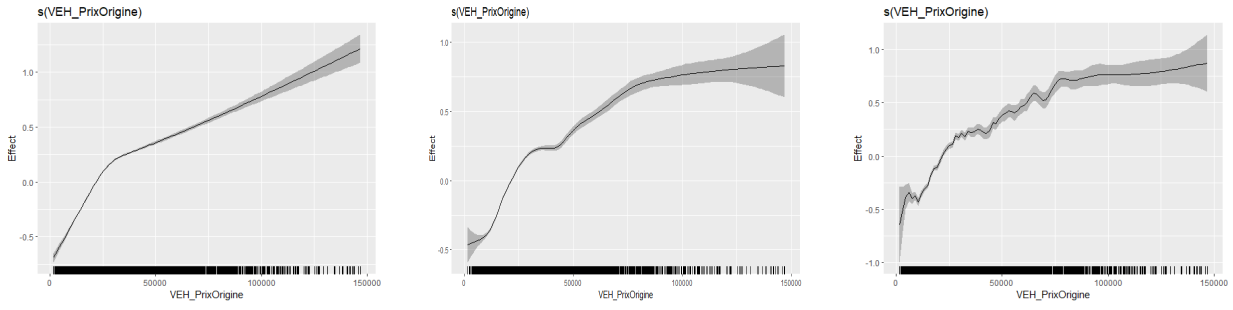
2. Cette famille de lois est définie en annexe C.3.

représentant le lien entre la variable X_l^c et la variable réponse, dites *fonctions lisses* ou *smooths*. Ces fonctions f_l peuvent avoir une forme paramétrique ou des splines en considérant des polynômes sur des intervalles de la variable X_l^c et en imposant des contraintes de continuité sur les limites de ces intervalles appelées *noeuds*.

Les fonctions lisses sont définies à l'aide d'une base de splines $(b_k(.))_{k=1,...,K}$ où K est la dimension de cette base ou le nombre de splines à considérer pour approcher les observations, et f s'écrit ainsi : $f(x) = \sum_{i=1}^{K} \beta_i b_i(x)$. La modélisation des f revient ainsi à l'estimation des paramètres à associer à chaque spline. Un exemple simple des fonctions lisses f est le cas polynomial d'ordre K : $f(x) = \sum_{i=1}^{K} \beta_k x^k$, ce qui revient à modéliser un GLM très flexible et qui peut entraîner le sur-apprentissage du modèle. Pour éviter ce sur-ajustement aux données, le modèle est pénalisé, et on maximise ainsi :

$$\text{log-vraisemblance} - \lambda * W$$

Pour un modèle à une seule variable explicative, W peut s'écrire : $\int [f'']^2$, et ainsi cette formule peut être interprétée comme une pénalisation de la courbure des fonctions lisses. λ est dit paramètre de lissage et contrôle le compromis entre l'ajustement du modèle aux données (log-vraisemblance) et son irrégularité (W) et est estimé par validation croisée pour maximiser la capacité du modèle à prédire sur des données sur lesquelles il n'est pas entraîné. Si λ est très faible, le modèle est non pénalisé et la fonction f devient très volatile, et si λ tend vers infini, il est sur-pénalisé et la fonction $\hat{f}$ devient linéaire. Nous visualisons cet effet dans le graphique 4.1, où nous modélisons la sévérité observée par la vitesse maximale du véhicule pour différentes valeurs de λ.



(a) $\lambda = 0$ (b) λ estimé (c) $\lambda = 100$

FIGURE 4.1: Comparaison des effets estimés du prix du véhicule sur la sévérité en fonction du paramètre de lissage.

Pour le choix de la dimension de la base des splines, et étant donné que le modèle est pénalisé, il suffit de choisir K assez large pour approcher suffisamment la vraie fonction f. Cette dimension est fixée aléatoirement dans la procédure *gam* de la librairie *mgcv* de R, et nous vérifions ainsi ce paramètre avec la procédure *gam.check* pour savoir si la dimension de base est adéquate : une p-value faible indique que la dimension choisie est faible et doit être ainsi augmentée. Cependant, en augmentant ce paramètre, les modèles GAM sont beaucoup plus lents que les modèles GLM, ce qui représente le principal désavantage de ces modèles.

4.3 Discrétisation des variables continues

Pour le calcul des primes dans une approche a priori, les fréquences et sévérités sont modélisées par des modèles linéaires généralisés et une discrétisation des variables continues est effectuée. Cela permet en effet de capter leurs effets non linéaires sans transformations de ces variables et de créer ainsi des modèles simples et faciles à implémenter, avec un nombre fini et interprétable de classes de risques.

Différentes approches peuvent être utilisées pour la création de ces segments. Une première méthode consiste à considérer des intervalles d'amplitudes égales. Pour k classes à créer, cette amplitude sera calculée ainsi : $\frac{max(valeurs)-min(valeurs)}{k}$. Bien que simple, cette méthode appliquée sur des données non uniformes peut créer des classes à nombre très élevé ou faible d'observations. De plus, étant donné que cette méthode est non supervisée et ne prend pas en compte la variation de la variable à prédire (fréquence ou sévérité) lors de la création des intervalles de la variable à discrétiser, des classes différentes peuvent concerner des niveaux de risques égaux.

Une deuxième méthode pour discrétiser les variables tarifaires continues consiste à créer des intervalles d'observations à effectifs égaux. Si cette méthode permet de créer des classes à nombre significatif d'observations, elle ne garantit pas l'homogénéité des risques au sein de chaque classe.

Pour créer des groupes homogènes de sévérité se basant sur les deux méthodes précédentes, des modèles linéaires généralisés sont appliqués en considérant une première discrétisation des variables continues. Ensuite, les catégories non significatives dans le modèle seront regroupées avec la classe de référence. Cette approche peut être assez laborieuse dans la présence d'un grand nombre de variables tarifaires continues. Une autre approche

possible consiste à modéliser un modèle linéaire généralisé avec des variables continues non discrétisées, par exemple l'âge de l'assuré, et à découper ensuite ces variables selon leurs effets sur la sévérité, dans ce cas $\hat{\beta}\times$âge, à l'aide d'arbres de régression [3]. Cependant, cette méthode ne prend pas en compte les effets non linéaires que peuvent avoir certaines variables sur la sévérité.

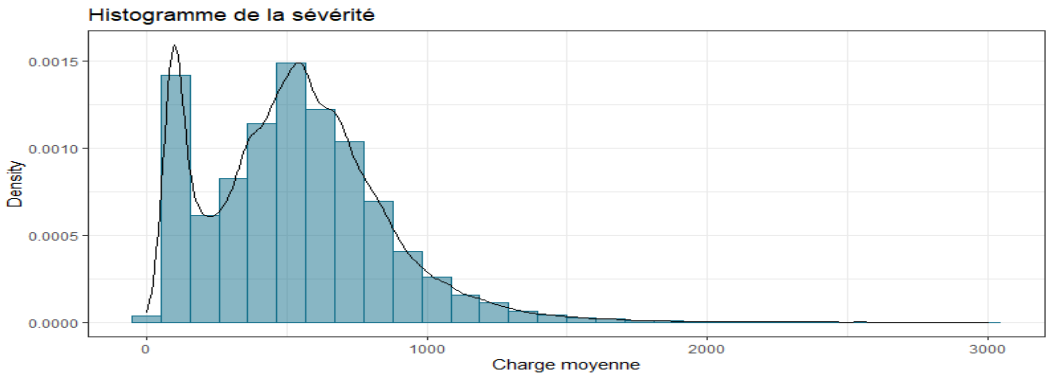
C'est ainsi que dans "*A data driven strategy for the construction of insurance tariff classes*", Antonio et al. [17] proposent d'employer des modèles additifs généralisés, plus flexibles, pour modéliser la sévérité avec l'ensemble des variables tarifaires continues et discrètes, et discrétiser ensuite chaque variable selon son effet prédit, par exemple : l'âge de l'assuré selon $\hat{f}$(âge) , par des arbres de régression. En contrôlant le nombre minimal d'observations par noeuds, cette méthode permet de créer des classes de variables significatives et homogènes en terme de sévérité. Enfin, un GLM est modélisé sur l'ensemble des variables catégorielles et continues discrétisées obtenues. Cette méthode sera testée par la suite sur nos données et comparée à une modélisation par GLM sans discrétisation préalable des variables continues.

4.4 Statistiques descriptives

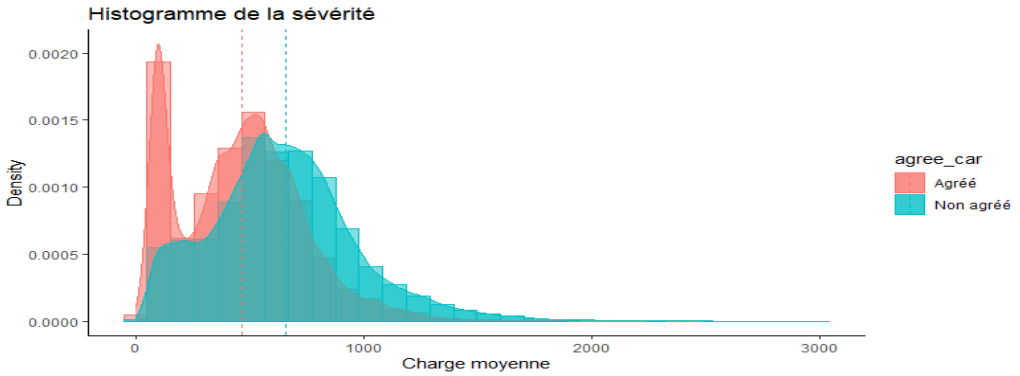
À partir de la base de données d'apprentissage utilisée pour modéliser les probabilités de réparation précédentes, nous sélectionnons les assurés ayant eu un sinistre de type bris de glace durant la période d'observation. Étant donné que cette dernière correspond à une année pour tous les assurés, l'exposition est égale à un et donc ne sera pas spécifiée dans les modèles de la sévérité, et cette dernière correspond dans ce cas à la charge totale pour chaque assuré.

La base de données d'apprentissage obtenue ainsi est composée de **63 000 observations** d'assurés ayant eu un sinistre bris de glace et contient les informations sur le coût de ce sinistre, le type du garage choisi, la probabilité de réparation en garages partenaires prédite ainsi que les différentes informations décrites dans la partie 1.5. 63% de ces observations sont des sinistres réparés en garages partenaires et 37% en garages non partenaires. Nous représentons dans le graphique 4.2 les histogrammes de ces sévérités bris de glace.

3. La construction de ces arbres est définie en annexe C.4.



(a) Tous types de réparation confondus



(b) Par type de réparation

FIGURE 4.2: Histogramme de la sévérité bris de glace

La sévérité moyenne des sinistres est d'environ 500 euros, et elle est plus élevée pour les sinistres réparés en garages non partenaires. En effet, la moyenne de ces charges en garages non partenaires dépasse d'environ 40% celles en garages partenaires. Nous remarquons également un pic aux alentours des 100 euros correspondant à un montant forfaitaire de réparation de pare-brise, et est observé pour les sinistres réparés en garages partenaires. Ces coûts dépendent du type de prestation (remplacement de pare-brise, réparation de pare-brise et intervention hors pare-brise [18]), donnée indisponible dans notre base de données. Nous enlevons ces montants fixes des observations pour une meilleure adéquation des modèles. Les sévérités obtenues ainsi sont représentées dans le graphique C.1.

Pour comparer l'adéquation de ces sévérités aux deux loi Gamma et log-normale, nous les modélisons selon ces lois et comparons le critère d'information Akaike (AIC) dans le tableau 4.1 ainsi que les distributions observées et prédites, représentées dans le graphique C.2. Le critère AIC évalue l'ajustement du modèle aux données en ajoutant un terme de pénalité pour la complexité du modèle. Il s'écrit ainsi : $AIC = -2log(L) + 2k$ où L est

la vraisemblance du modèle et k est le nombre de paramètres, ici égal à 2. Le modèle retenu sera celui qui minimisera ce critère. Pour notre cas, nous remarquons que pour les trois sévérités, la loi Gamma s'ajuste mieux aux données que la loi log-normale. Cette distribution sera ainsi retenue pour la suite de notre étude.

Type de sévérité	AIC Gamma	AIC Log-normale
Tous types de réparation confondus	773 494	780 796
Réparation en garage agréé	455 731	460 259
Réparation en garage non agréé	315 565	318 891

TABLE 4.1: Comparaison des AIC

Pour une première présélection des variables à inclure dans nos modèles, nous commençons par comparer la sévérité selon les différentes variables de notre base de données, et gardons les variables ayant un effet significatif sur la sévérité. Nous nous contentons de cette méthode simple de présélection des variables puisque notre objectif principal est la comparaison de modèles avant et après différenciation par le type du choix de réparation.

Nous représentons ainsi dans les graphiques C.3 à C.6 des nuages de points de la sévérité bris de glace selon certaines variables, la courbe en traits plein représente une régression locale lissée pour visualiser la tendance du lien entre la sévérité et chaque variable. Étant donné que la garantie bris de glace concerne principalement les surfaces vitrées du véhicule, la sévérité est croissante selon le prix de pare-brise et des optiques et moins élevée pour les anciens véhicules.

Nous représentons également dans le graphique C.7 la sévérité pour chaque code postal des assurés, après géolocalisation de ces derniers et découpage des sévérités selon les quantiles : 0.2 et 0.8 pour obtenir trois classes : sévérité faible, moyenne et élevée. Étant donné le faible effectif de sinistres par code postal dans notre base de données, il est difficile de tirer des conclusions sur la sévérité par code postal. Pour réduire la complexité et le temps d'apprentissage des modèles, nous choisissons de considérer la région du sinistre comme variable à inclure dans nos modèles de sévérité et comparons dans le tableau 4.3 cette dernière selon les régions présentes dans notre base de données. Nous remarquons que la sévérité est plus importante dans les régions du Grand-Est, Hauts-de-France et Normandie, mais que l'écart entre ces sévérités reste relativement faible.

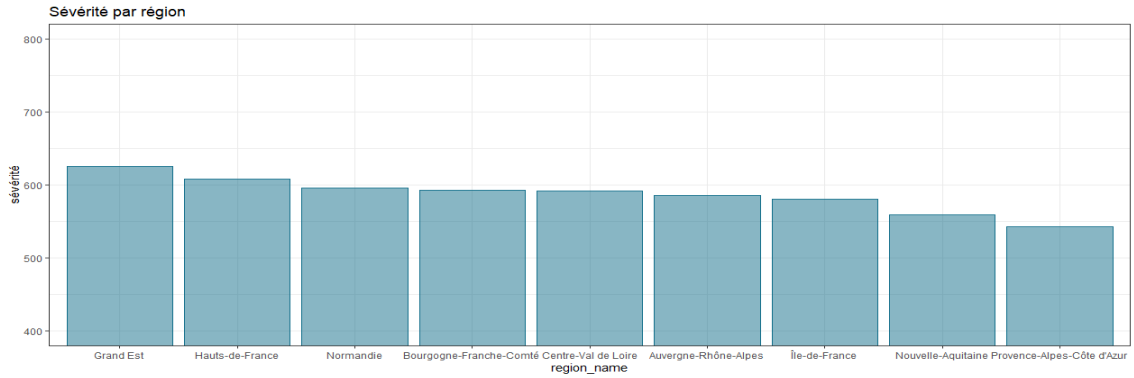


FIGURE 4.3: Sévérité par région

Nous représentons la matrice des corrélations entre nos différentes variables continues dans le graphique C.8. Puisque nous allons modéliser la sévérité par des modèles linéaires, nous vérifions que ces corrélations ne sont pas très élevées (coefficient de corrélation de Pearson inférieur à 0.9 en valeur absolue) pour éviter des problèmes de multicolinéarité impactant la significativité des paramètres de chaque variable.

4.5 Modélisation et résultats

Comme nous avons expliqué précédemment, notre objectif est de comparer entre deux modèles : un modèle sans prise en compte du type de réparation choisie et un modèle où ce choix est intégré. Pour cela, nous allons modéliser trois sévérités par des modèles GAM :

$$\begin{cases} S : \text{sévérité tous types de réparation confondus} \\ S_0 : \text{sévérité des sinistres réparés en garages non partenaires} \\ S_1 : \text{sévérité des sinistres réparés en garages partenaires} \end{cases}$$

L'approche retenue pour la sélection des variables significatives est la méthode *Backwards* qui consiste à intégrer toutes les variables pré-sélectionnées dans le modèle, puis nous éliminons les variables non significatives et dont la suppression réduit l'AIC du modèle. À l'issue de cette méthode, nous obtenons 14 variables significatives, dont les coefficients sont représentés dans les tableaux C.2, C.3 et C.4 et les fonctions lisses estimées sont présentées dans les graphiques C.10 et 4.4 suivants :

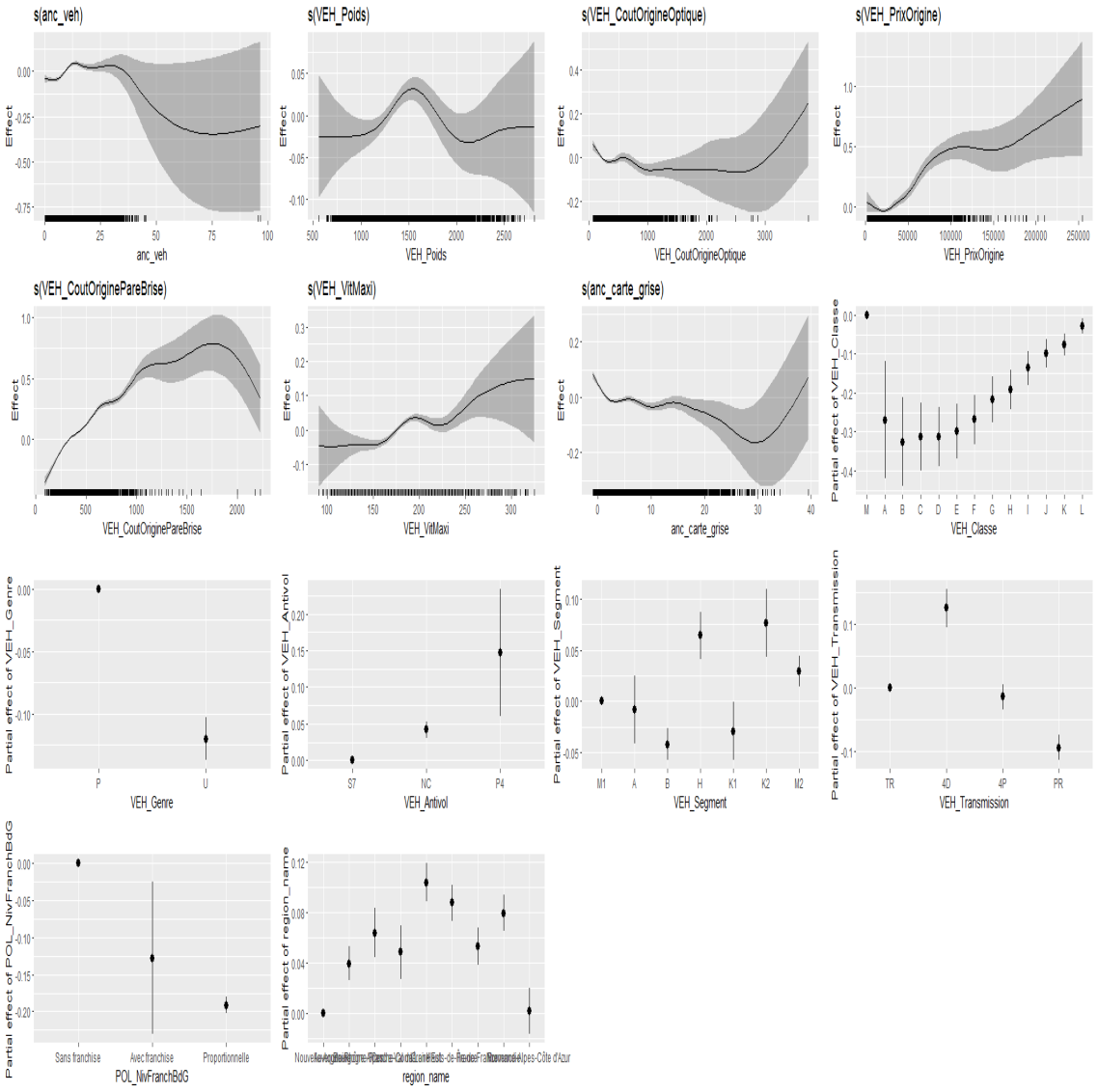


FIGURE 4.4: Les composantes estimées des modèles GAM de sévérité

Nous remarquons que ces modèles dépendent principalement des caractéristiques du véhicule. En effet, seules deux des variables retenues par les modèles de sévérité dépendent des caractéristiques de l'assuré : le type de franchise souscrite et la région. Les assurés ayant souscrits des garanties bris de glace avec franchise ont tendance à avoir des sinistres moins coûteux que les assurés préférant des contrats sans franchise. Cela peut être expliqué par un niveau de risque plus élevé chez ces derniers, puisqu'ils choisissent des contrats tels que la somme restant à leur charge soit nulle. De plus, la sévérité est moins élevée pour les contrats à franchise proportionnelle, où le montant à verser par l'assuré dépend de la sévérité du sinistre, que pour ceux à franchise fixe.

Pour les caractéristiques du véhicule de l'assuré, la sévérité bris de glace dépend fortement du prix de pare-brise, optiques du véhicule et son prix total. En effet, étant donné que les coûts moyens modélisés concernent le dédommagement après sinistre impactant les éléments de glace ou de verre du véhicule (pare-brise, la lunette arrière, le toit, les vitres latérales, et les feux avants), ils sont croissants avec le prix de ces composantes.

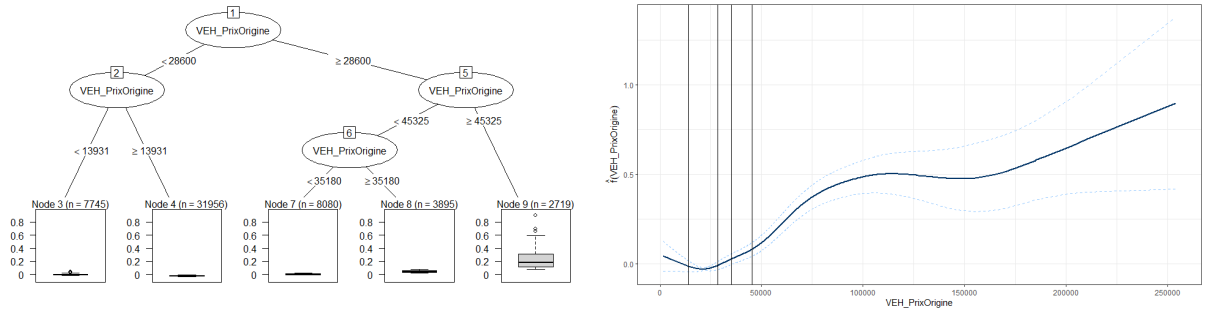
Avant de démontrer l'impact de l'inclusion du choix du type de garage de réparation dans la modélisation de la sévérité et étant donné que des tarifs avec des variables continues sont moins souhaitables en assurance, nous comparons les deux modèles GLM et GAM sur la sévérité S (tous types de réparations confondus) et appliquons la méthode décrite dans la partie 4.3 pour obtenir des tarifs par classes de risques, et donc plus faciles à implémenter en pratique et à communiquer aux agents.

Nous discrétisons ainsi chaque variable continue par des arbres de régression selon son effet sur la sévérité estimé $\hat{f}$ du modèle GAM précédent. Par exemple, pour le prix du véhicule, nous construisons un arbre de régression sur les données suivantes :

Prix du véhicule	$\hat{f}$(prix)	Poids ou nombre d'observations
1663	0.0415	1
2287	0.0386	1
2348	0.0383	3
2439	0.0379	1
2592	0.0372	1
2897	0.0358	1
...	...	...

TABLE 4.2: Effet du prix du véhicule sur la sévérité

Pour contrôler la taille des arbres, nous imposons un nombre minimal de 5% des observations dans chaque feuille et nous limitons la profondeur des arbres à 5. Les résultats de cette discrétisation pour la variable du prix du véhicule sont représentés dans le graphique 4.5 suivant :



(a) Noeuds de l'arbre de régression

(b) Effet du prix du véhicule discrétisé

FIGURE 4.5: Discrétisation du prix du véhicule

Le prix du véhicule a été discrétisé selon son effet marginal sur la sévérité bris de glace en 5 classes. Nous remarquons que la dernière classe obtenue concerne un intervalle de prix assez large et à effets non stables sur la sévérité. Cela est dû au faible nombre d'observations de véhicules de prix supérieur à 45 000 euros. Nous obtenons ainsi des classes de variables continues dont les observations sont plus homogènes en terme de sévérité. L'ensemble des variables continues discrétisées ainsi est représenté dans le graphique C.11.

Les variables discrétisées ainsi que les variables qualitatives du modèle GAM précédent seront employées pour obtenir un modèle linéaire généralisé simple de la sévérité. Les coefficients de ce dernier modèle sont représentés dans le tableau C.5 et nous remarquons que toutes les classes créées par cette approche sont significativement différentes de la classe de référence et entre elles. Nous comparons dans le tableau 4.3 les performances de trois modèles : le modèle GAM précédent, un modèle GLM sans discrétisation des variables continues et le dernier modèle GLM à variables discrétisées, basées sur le critère AIC :

	GLM avant discrétisation	GLM après discrétisation	GAM
AIC	760 125	759 759	759 651

TABLE 4.3: Comparaison des AIC

Nous remarquons que le modèle GLM à variables continues est moins performant que les deux autres modèles puisqu'il ne prend pas en compte les effets non linéaires des variables sur la sévérité. Nous observons également que cette approche de discrétisation permet aux modèles GLM d'approcher les modèles GAM, tout en réduisant le temps de calcul et simplifiant les tarifs proposés. Pour la suite de l'étude et puisque notre objectif

est la comparaison entre les deux approches de modélisation de la sévérité, nous allons nous baser sur les modèles GAM précédents pour les trois sévérités S, S_0 et S_1.

4.6 Résultats et conclusion de l'analyse

Dans cette partie, nous comparons les performances des modèles de prédiction de la sévérité précédents :

$$\begin{cases} S : \text{Sévérité tous types de réparation confondus} \\ S^* : \text{Sévérité qui inclut le choix du garage de réparation (formule 4.1)} \end{cases}$$

Étant donné que nous avons modélisé les sévérités selon le choix du garage de réparation sur des bases de données de tailles différentes, les mesures basées sur la vraisemblance (AIC et BIC), qui dépend à son tour de la taille de l'échantillon, ne permettent pas de comparer ces sévérités. Ainsi, nous nous baserons sur des mesures des écarts entre les observations et les prédictions de sévérités sur la base **test** composée de **13 500** observations de sinistres bris de glace. Notons que pour évaluer correctement les performances des modèles sur de nouvelles observations, cette base de données n'a pas été employée dans la modélisation des probabilités du choix du garage de réparation.

Pour comparer les deux approches de modélisation de la sévérité, nous calculons l'erreur quadratique moyenne (RMSE) entre les valeurs réelles y et prédites $\hat{y}$ pour les n observations de la base de données. Pour que les écarts positifs et négatifs ne se compensent pas, le carré des erreurs est considéré et la racine carrée permet à la RMSE d'être de même unité que la variable observée :

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

Cette dernière métrique est sensible aux valeurs extrêmes puisqu'elle donne plus de poids aux erreurs élevées. Nous allons ainsi également comparer l'erreur absolue moyenne (MAE) définie comme la moyenne des valeurs absolues des erreurs :

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

Nous comparons dans le tableau 4.4 les deux approches de calcul de sévérité en employant les deux erreurs précédentes :

	Prédiction sans choix de garage	Prédiction avec choix de garage	Différence
RMSE	272	254	**-7,08%**
MAE	189	174	**-8,62%**

TABLE 4.4: Comparaison des erreurs avant et après segmentation par le choix du garage de réparation

En prédisant le type de garage de réparation pour chaque assuré et en différenciant la sévérité selon ce choix, l'erreur des prédictions moyenne a été réduite d'environ **8%**. Nous comparons également dans le tableau 4.5 les moyennes des sévérités observées et prédites selon les deux approches pour comprendre cette réduction des erreurs :

Choix de l'assuré	Sévérité observée	S* prédite moyenne	S prédite moyenne
Garages partenaires	520,50	524,84	568,95
Garages non partenaires	656,53	661,58	596,08

TABLE 4.5: Comparaison des sévérités selon le type de garage choisi par l'assuré

Nous remarquons que la prédiction de la sévérité est plus proche aux coûts observés lorsque le choix du garage de réparation est anticipé et intégré dans le modèle. En effet, la différence entre la sévérité moyenne prédite et celle observée passe de 48 euros à 4 euros pour les assurés qui choisissent des garages partenaires et de 60 euros à 5 euros pour les garages non partenaires. En termes de charge totale des sinistres de l'ensemble de la base de données, la différence entre la somme de la sévérité observée et prédite est passée de 71 340 à 62 260 euros, réduisant également l'écart d'environ 15% par rapport au modèle sans segmentation par type de garage de réparation.

Les résultats précédents démontrent d'une part que nous avons bien pu classifier les assurés selon leur choix du type de garage de réparation avec les modèles du chapitre 3. En effet, les sévérités S* pondérées par les probabilités prédites de réparer en garages partenaires (formule 4.1) tendent vers la sévérité réelle observée. D'autre part, le nouveau modèle de sévérité permet de créer des groupes plus homogènes d'assurés. En effet, en considérant une fréquence égale à 1 pour les deux types d'assurés (puisque le choix de garage n'intervient qu'après survenance du sinistre), nous proposons des primes moins

élevées (-7,74% en moyenne) pour les assurés ayant tendance à réparer leurs véhicules dans des garages partenaires et reflétons les charges supplémentaires des réparations dans les primes du second type d'assurés.

Supposons que nous sommes en présence de deux assureurs : l'assureur A ne prend pas en compte le choix potentiel du garage de réparation du client et ne segmente pas ses tarifs selon ce critère. Il propose ainsi le tarif S du tableau 4.5. Le deuxième assureur B choisit quant à lui d'appliquer une segmentation plus fine et propose la prime S^* à ses clients. Les bons risques, dans ce cas les assurés qui réparent leurs véhicules en garages partenaires et qui ont une sévérité moyenne observée égale à 520 euros, préfèrent l'assureur B puisqu'il propose une prime moins élevée (525 euros $<$ 569 euros). De plus, cet assureur reste en moyenne en situation d'équilibre alors que l'assureur A perd de l'argent en attirant les mauvais risques (assurés qui réparent en garages non partenaires) puisque la prime qu'il propose (596 euros) est inférieure au vrai risque de ce type d'assurés (657 euros). Cette situation est un exemple du phénomène de sélection adverse en assurance.

Étant donné que les bons risques (assurés réparant en garages partenaires) sont les types d'observations prédominants dans notre base de données, cette segmentation des primes peut être avantageuse puisqu'elle engendre une baisse des primes pour ces assurés et permet ainsi de les fidéliser. Cependant, cela a pour effet d'augmenter les primes de la deuxième tranche d'assurés ce qui peut réduire la part de marché de l'assureur B. Ainsi, une étude de la sensibilité des deux types d'assurés à des variations de prix ainsi que de leur rentabilité sur l'ensemble des garanties souscrites serait intéressante pour mesurer l'impact de la segmentation que nous proposons.

Une deuxième approche pour réduire les coûts de réparation déboursés par l'assureur consiste à inciter les agents à orienter plus les assurés vers des garages partenaires. En effet, comme nous avons démontré, les assurés orientés vers des garages partenaires par les agents ont tendance à réparer plus dans ce type de garage que les assurés non orientés. Cette méthode consisterait ainsi à calculer les moyennes d'orientation du portefeuille d'assurés de chaque agent et d'augmenter les primes de ceux ayant des faibles taux d'orientation vers des garages agréés, les poussant ainsi à orienter plus vers ces garages pour éviter une fuite des clients vers d'autres agents.

Conclusion

Dans ce mémoire, nous nous sommes focalisés sur la modélisation du choix de l'agent d'orienter un accident vers des garages agréés et le choix du client de réparer son véhicule dans lesdits garages, en raison de leur lourd impact sur les coûts de réparation assurés. Pour ce faire, nous avons privilégié les méthodes d'apprentissage automatique, pour bénéficier de la grande quantité de données disponibles, capter les interactions entre les variables et détecter les plus importantes dans la prédiction des deux scores. Nous avons testé des méthodes de classification usuelles basées sur le Bagging et le Boosting, ainsi qu'un nouveau modèle CatBoost qui donne de meilleurs résultats lorsqu'un grand nombre de prédicteurs catégoriels est présent dans la base de donnée.

Après modélisation du choix du garage de réparation, nous avons pu prédire pour chaque assuré sa propension de réparer en garages partenaires dans le cas de survenance d'un sinistre. À l'aide de cette probabilité, nous proposons un nouveau modèle de sévérité s'ajustant mieux aux observations et permettant ainsi une meilleure prédiction des sommes à payer par l'assureur et des risques du portefeuille.

Notre étude permet ainsi de segmenter plus finement les tarifs proposés et de construire des classes plus homogènes de risque en anticipant le choix du garage de réparation dans le cas d'un sinistre. De façon similaire et en analysant le comportement des assurés, d'autres pistes d'amélioration des primes sont toujours à explorer. Cependant, le nouveau critère utilisé doit respecter un ensemble de règles du code de déontologie de la profession, pour ne pas inclure des variables discriminatoires dans les modèles et permettre un certain degré de solidarité entre les assurés.

Bibliographie

[1] Tianqi Chen. [2014]. Introduction to Boosted Trees. Présentation consultée sur : https://web.njit.edu/~usman/courses/cs675_spring20/BoostedTree.pdf

[2] Tianqi Chen et Carlos Guestrin [2016], XGBoost : A Scalable Tree Boosting System. Publié dans : Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785-794.

[3] Leo Breiman [2001], Random Forests. Publié dans Machine Learning, Volume 45, Number 1 - SpringerLink.

[4] Marie Chavent [2016] à partir des cours d'Adrien Todeschini et Robin Genuer, Master MIMSE - Université de Bordeaux http://www.math.u-bordeaux.fr/~mchave100p/wordpress/wp-content/uploads/2013/10/Apprentissage_C3_pres.pdf

[5] Agrégation de modèles, Institut de Mathématiques de Toulouse, https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-agreg.pdf

[6] Leo Breiman [1996], Bagging predictors. Machine Learning 24, 123–140. https://doi.org/10.1007/BF00058655

[7] Manuel Fernandez-Delgado, Eva Cernadas and Senén Barro [2014], Do we Need Hundreds of Classifers to Solve Real World Classification Problems ? The Journal of Machine Learning Research, January 2014.

[8] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin [2017], CatBoost : unbiased boosting with categorical features, NeurIPS 2018. https://catboost.ai/

[9] Scott M. Lundberg and Su-In Lee [2017], A Unified Approach to Interpreting Model Predictions. Publié dans : NIPS'17 : Proceedings of the 31st International Conference on Neural Information Processing Systems, Pages 4768–4777. https://shap.readthedocs.io/

[10] Hastie, T. J. and Tibshirani, R. J. [1990]. Generalized additive models, volume 43. CRC Press.

[11] Simon Wood. [2006]. Generalized additive models : an introduction with R. Chapman and Hall/CRC Press.

[12] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. [1984]. Classification and regression trees. CRC press.

[13] Maxime Clijsters [2014]. Dealing with continuous variables and geographical information in non-life insurance ratemaking. Faculty Of Economics AND Business Faculty Of Science.

[14] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis and Torsten Hothorn [2007]. Bias in random forest variable importance measures : Illustrations, sources and a solution.

[15] Hastie Trevor, Tibshirani Robert and Friedman, Jerome [2008]. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. (second edition) ; Springer.

[16] Katrien Antonio, Roel Henckaerts and Roel Verbelen [2020], Insurance pricing analytics with R , Workshop on Pricing Analytic.

[17] Katrien Antonio, Roel Henckaerts and Roel Verbelen [2018], A data driven strategy for the construction of insurance tariff classes. Scandinavian Actuarial Journal.

[18] Eric VA. Modélisation du coût de la garantie Bris de Glace automobile. Mémoire d'actuariat, 2017.

[19] `https://towardsdatascience.com/`

[20] `https://medium.com/`

[21] `https://machinelearningmastery.com/`

Annexe A

Orientation

A.1 Véhicules

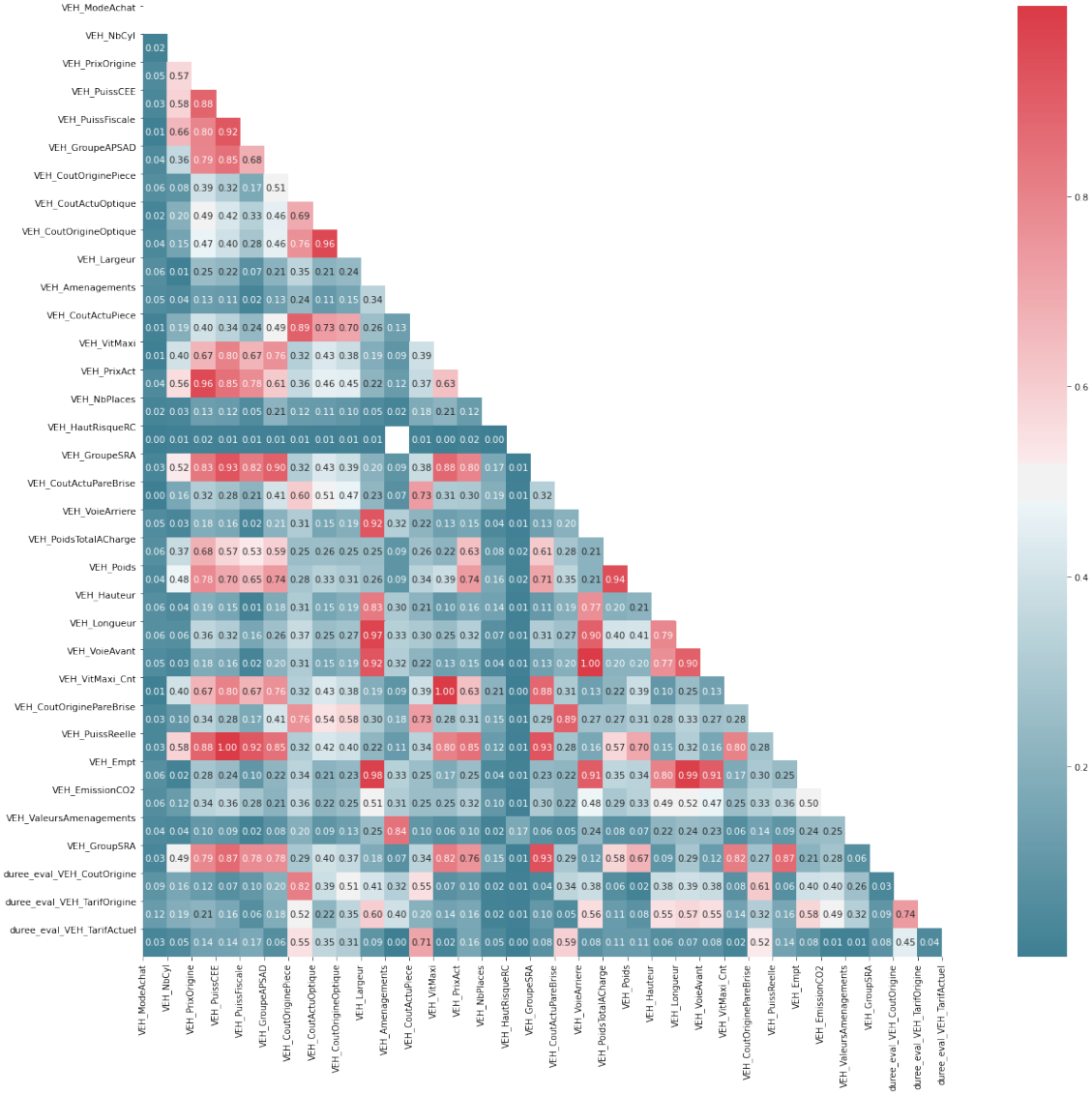


FIGURE A.1: Corrélations de Pearson des variables du véhicule

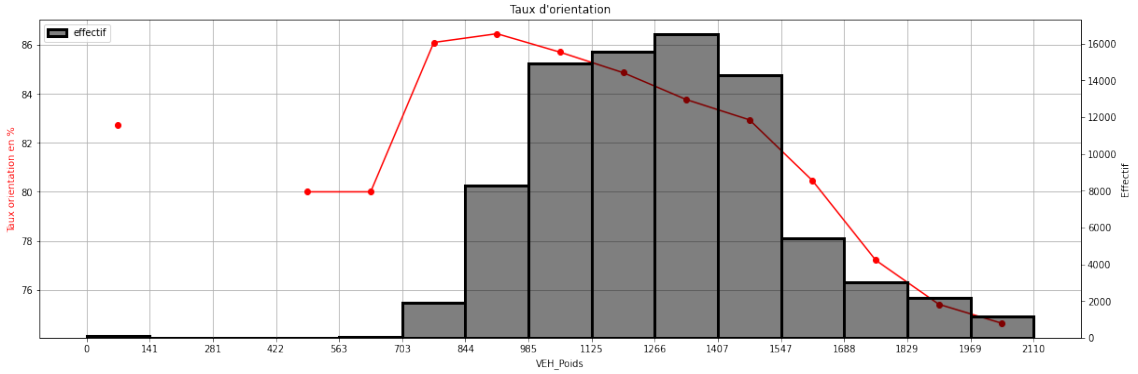


FIGURE A.2: Moyenne des orientations par poids du véhicule

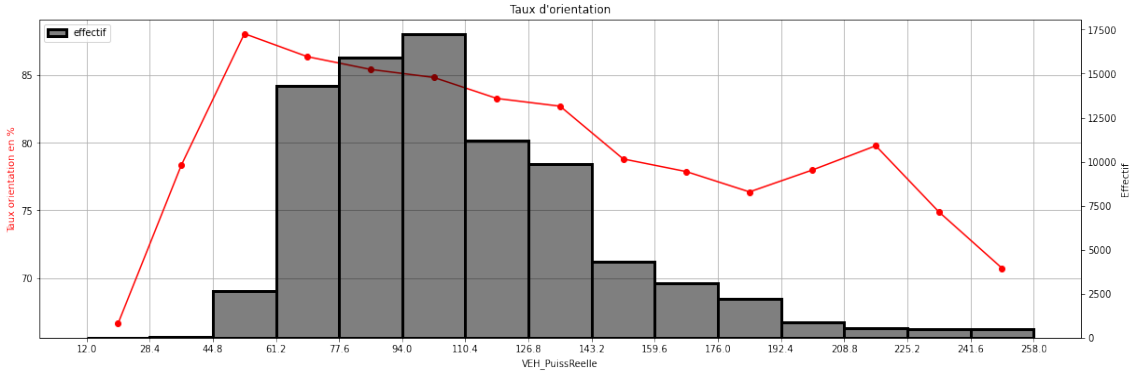
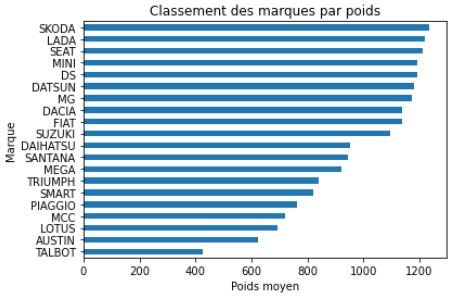
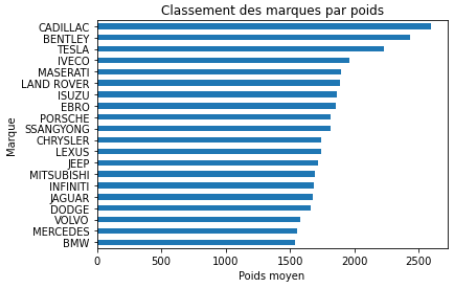


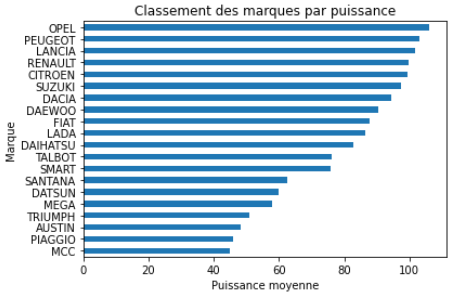
FIGURE A.3: Moyenne des orientations par puissance du véhicule



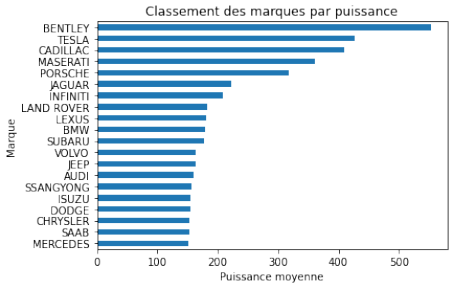
(a) Poids moyen par marque, les 20 moins lourds



(b) Poids moyen par marque, les 20 plus lourds



(c) Puissance moyenne par marque, les 20 moins puissants



(d) Puissance moyenne par marque, les 20 plus puissants

FIGURE A.4: Poids et puissances par marque de véhicule

A.2 Assuré

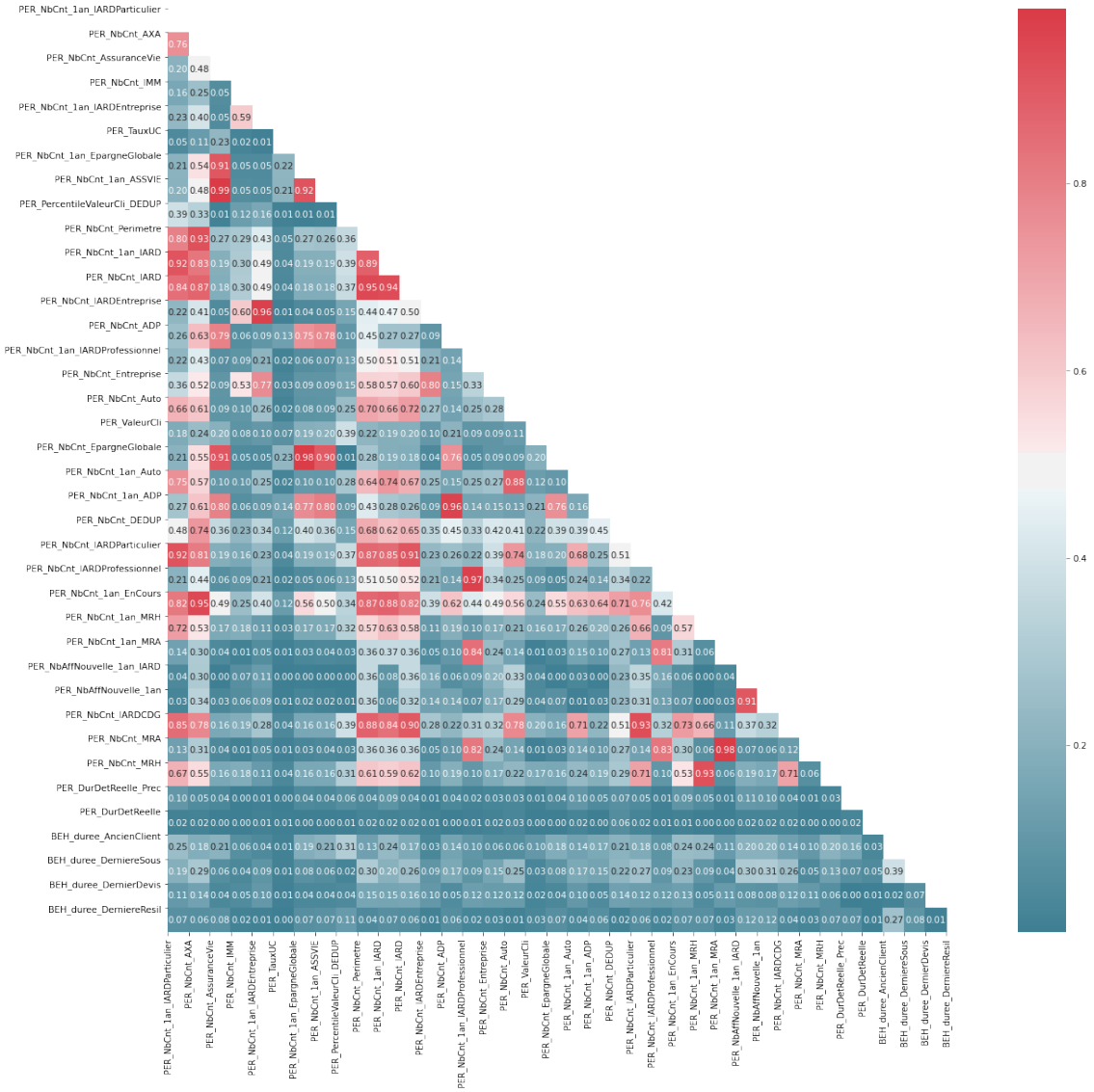


FIGURE A.5: Corrélations de Pearson des variables de l'assuré

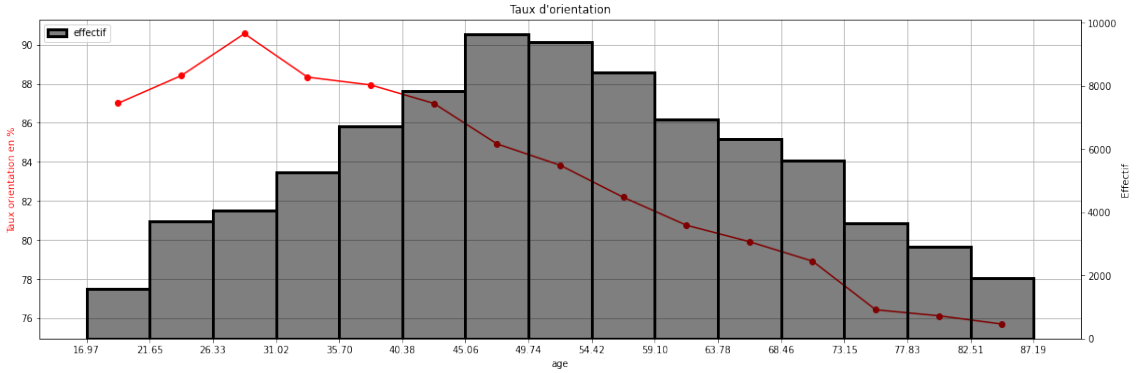


FIGURE A.6: Moyenne des orientations par âge

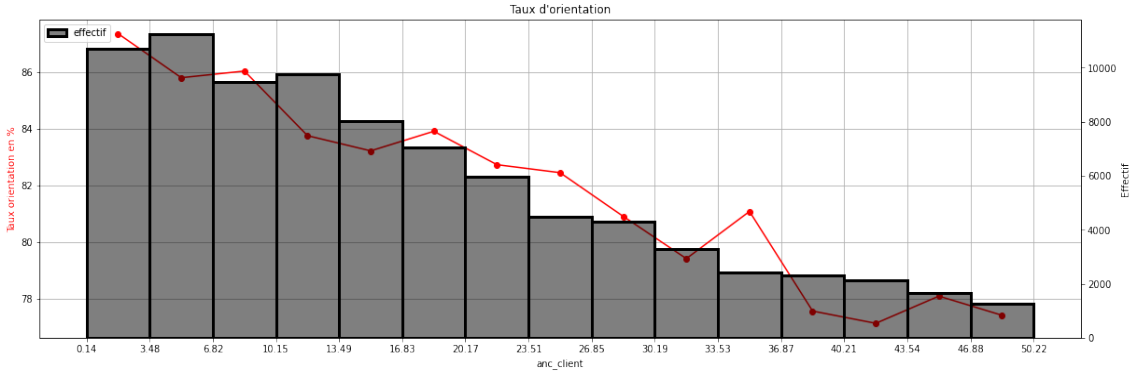


FIGURE A.7: Moyenne des orientations par ancienneté client

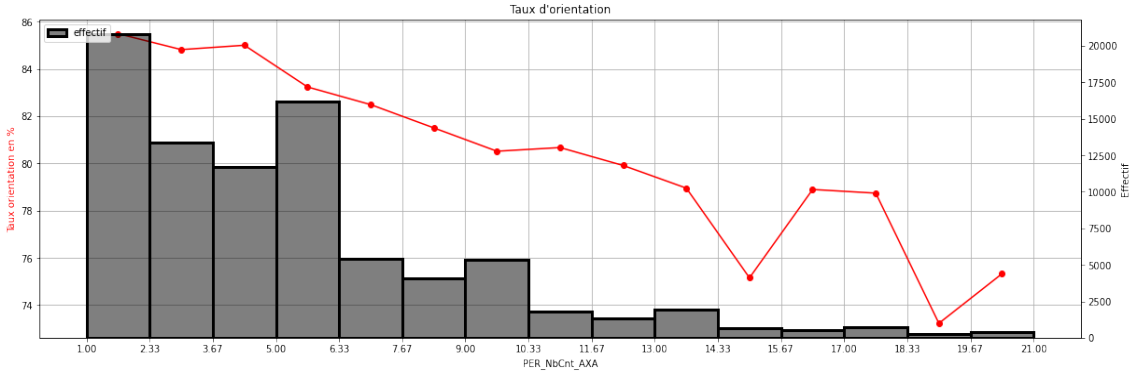


FIGURE A.8: Moyenne des orientations par nombre de contrats

A.3 Présélection des variables

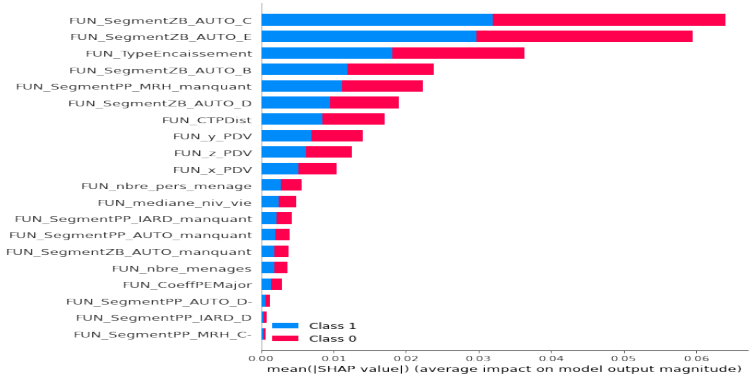


FIGURE A.9: Importance des variables : Agent

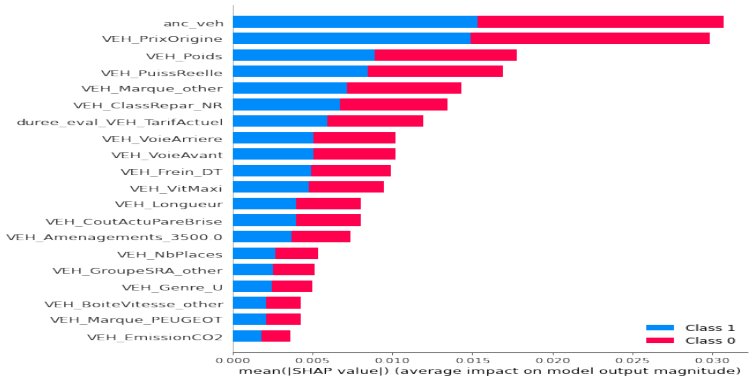


FIGURE A.10: Importance des variables : Véhicules

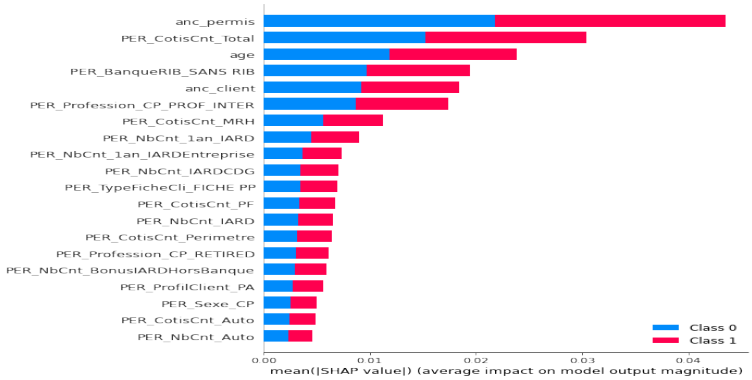


FIGURE A.11: Importance des variables : Assuré

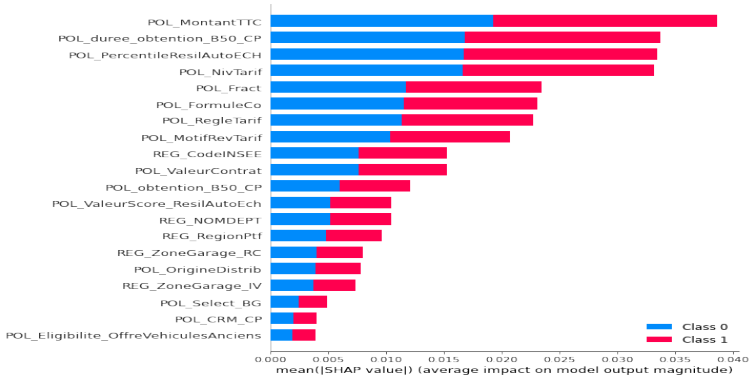
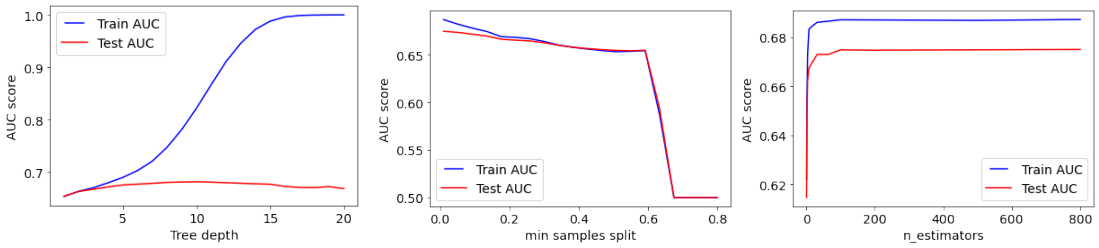


FIGURE A.12: Importance des variables : Police

A.4 Comparaison des modèles

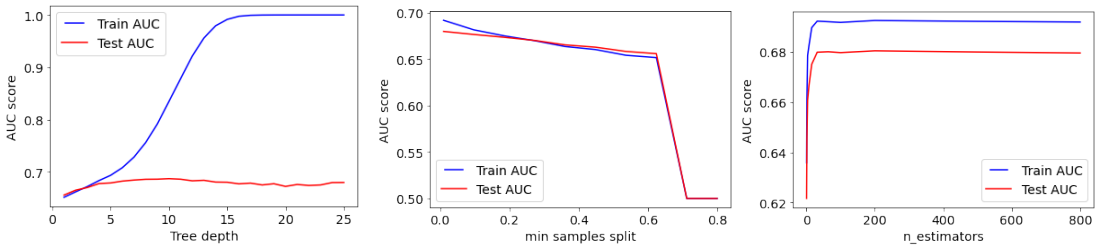
Catégorie de variables	Agent	Assuré	Véhicule	Police
AUC	0,642	0,641	0,60	0,633

TABLE A.1: AUC des modèles de forêts aléatoires



(a) Profondeur des arbres de la forêt (b) Nombre d'observations minimal (c) Nombre d'arbres construits

FIGURE A.13: One hot encoding : Hyperparamètres du modèle de forêts aléatoires



(a) Profondeur des arbres de la forêt (b) Nombre d'observations minimal (c) Nombre d'arbres construits

FIGURE A.14: Label encoder : Hyperparamètres du modèle de forêts aléatoires

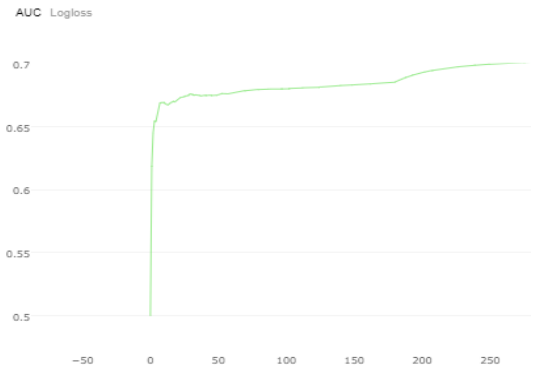


FIGURE A.15: AUC par itération pour le modèle CatBoost

Annexe B

Réparation

B.1 Assuré

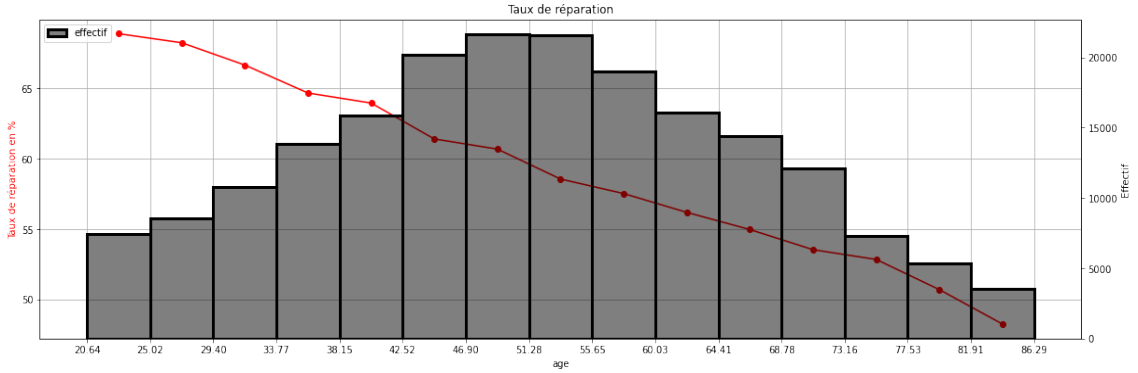


FIGURE B.1: Moyenne des réparations en garage partenaire par âge de l'assuré

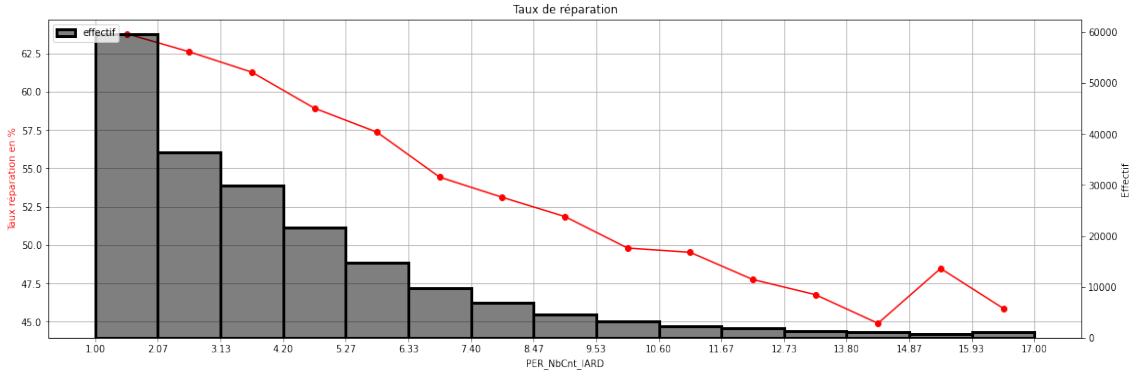


FIGURE B.2: Moyenne des réparations en garage partenaire par nombre de contrats IARD souscrits

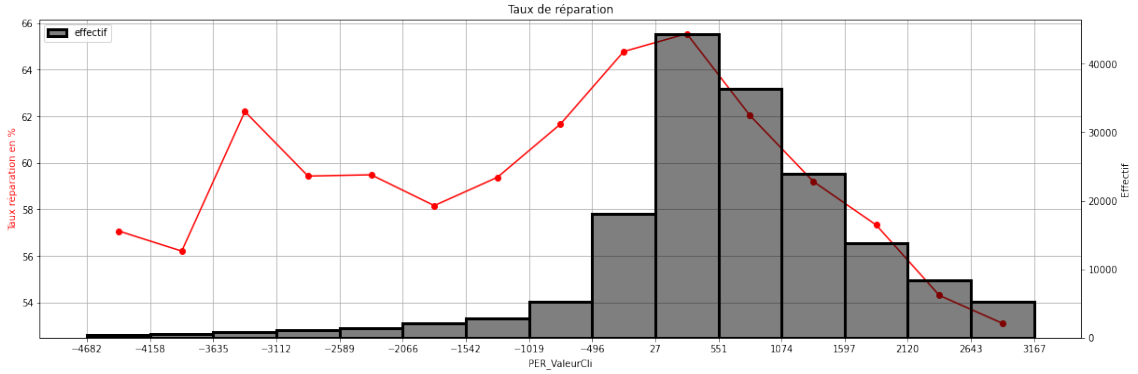


FIGURE B.3: Moyenne des réparations en garage partenaire par valeur client

B.2 Véhicule

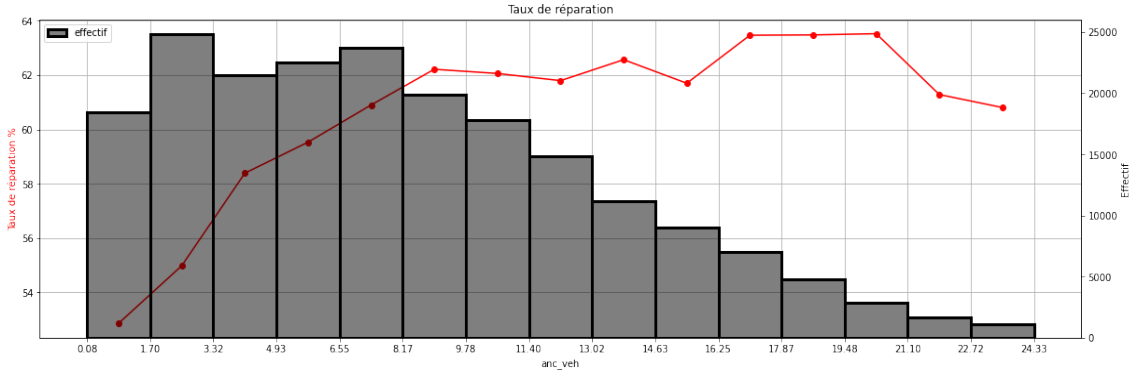


FIGURE B.4: Moyenne des réparations en garage partenaire par ancienneté du véhicule

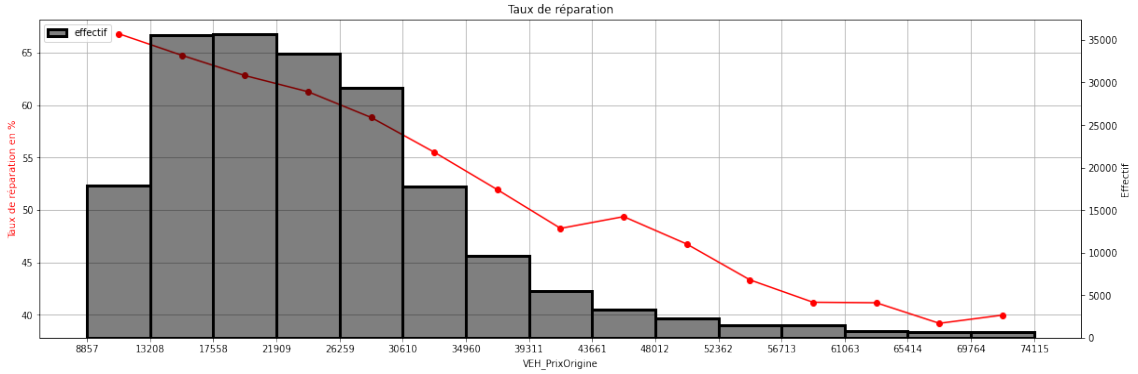


FIGURE B.5: Moyenne des réparations en garage partenaire par prix du véhicule

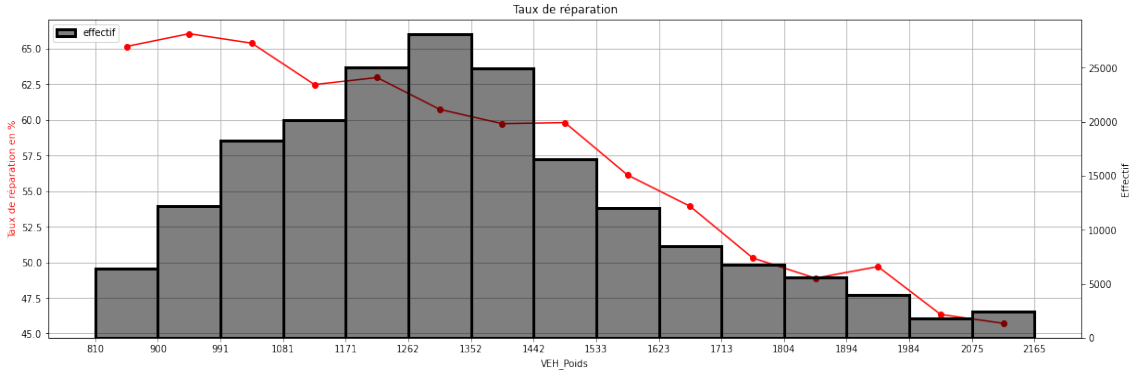


FIGURE B.6: Moyenne des réparations en garage partenaire par poids du véhicule

B.3 Présélection des variables

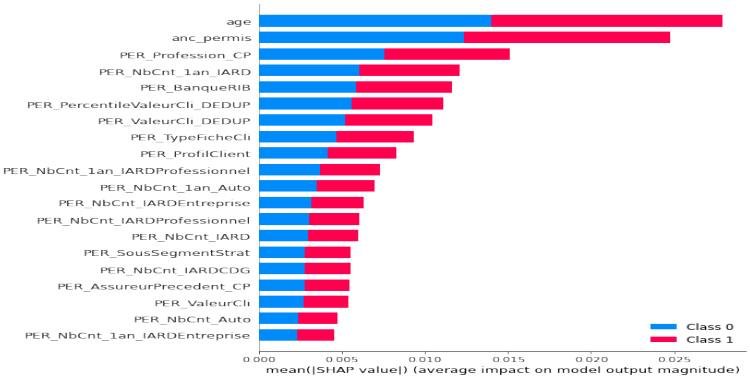


FIGURE B.7: Importance des variables : Assuré

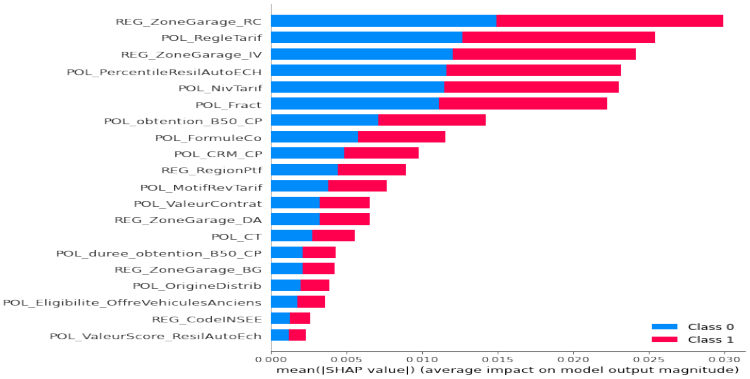


FIGURE B.8: Importance des variables : Police

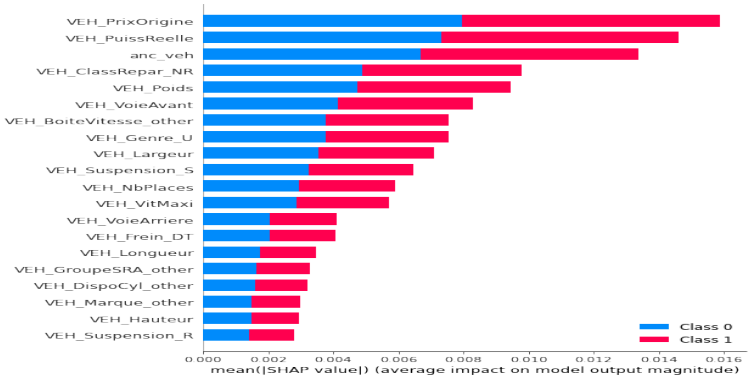


FIGURE B.9: Importance des variables : Véhicule

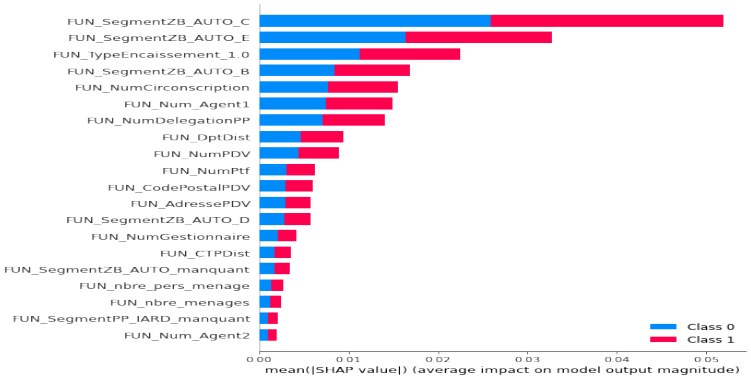
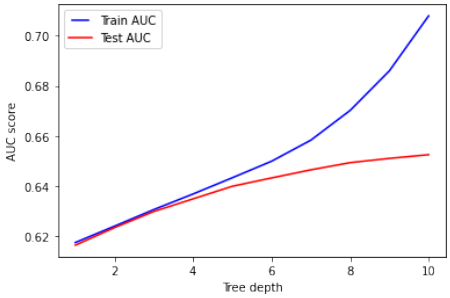
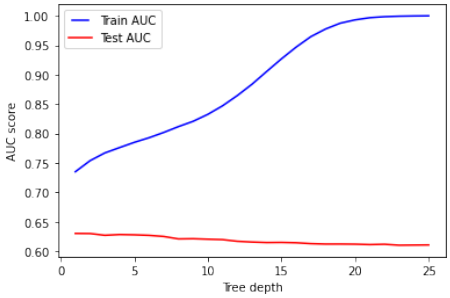


FIGURE B.10: Importance des variables : Agent

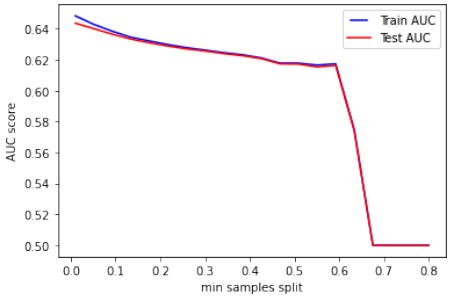
B.4 Modélisation



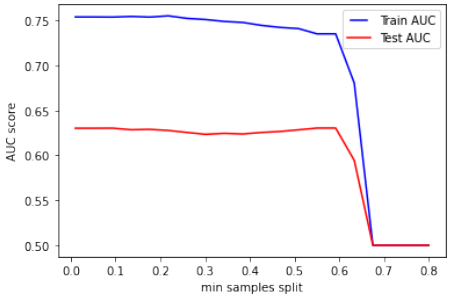
(a) Profondeur des arbres, label encoder



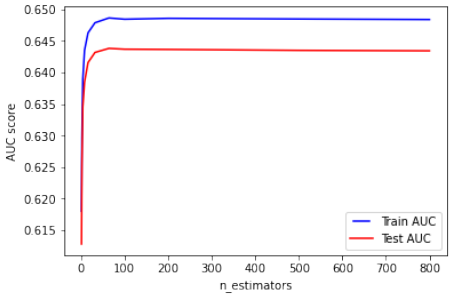
(b) Profondeur des arbres, target encoder



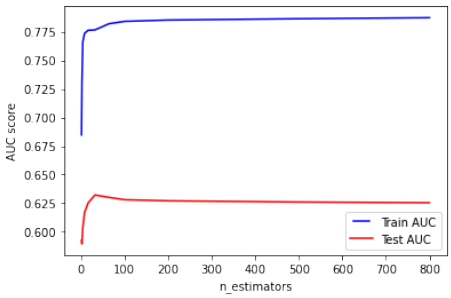
(c) Nombre d'observations, label encoder



(d) Nombre d'observations, target encoder



(e) Nombre d'arbres, label encoder



(f) Nombre d'arbres, target encoder

FIGURE B.11: Comparaison du label encoder et target encoder

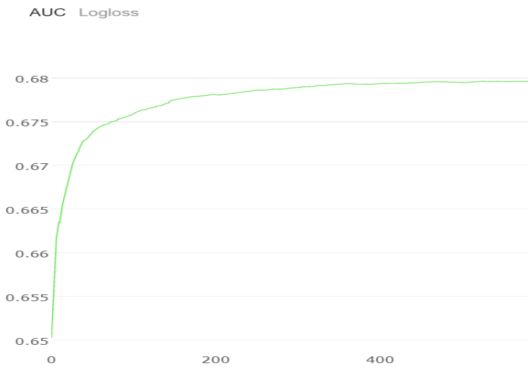


FIGURE B.12: AUC par itération pour le modèle CatBoost

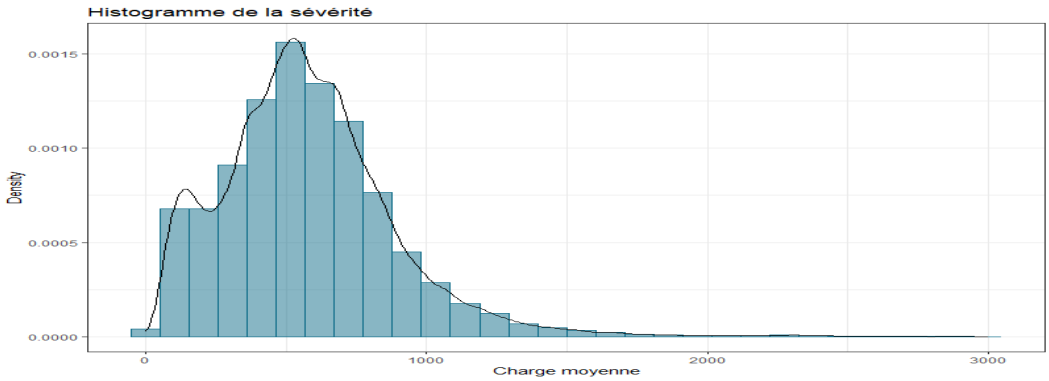
Annexe C

Modélisation de la sévérité

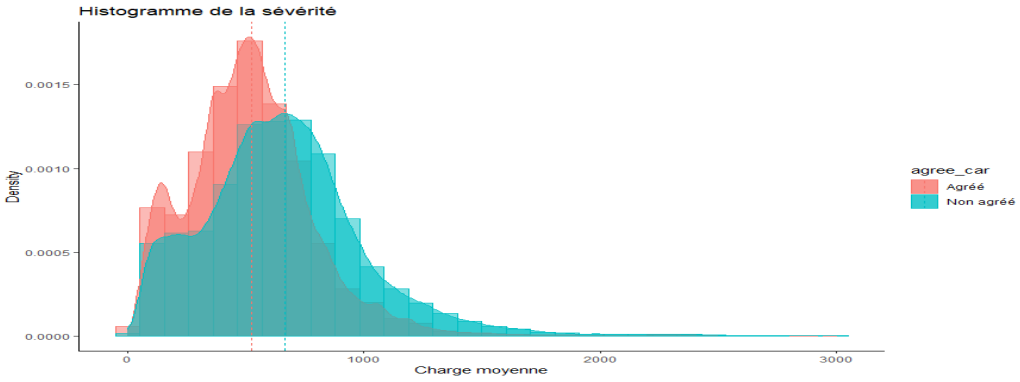
C.1 Distribution de la sévérité bris de glace

Garantie	Nombre d'observations
Bris de glace	98 814
Dommages	28 964
Responsabilité civile	25 264
Grêle	13 230
Vol	867
Incendie	99

TABLE C.1: Nombre d'observations par garantie

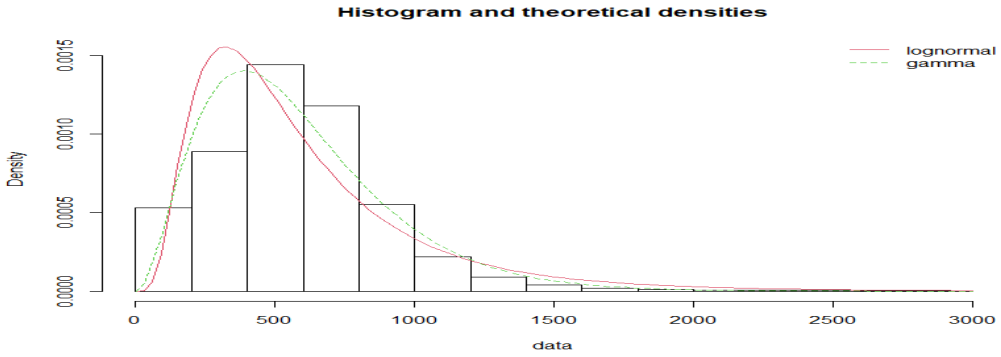


(a) Tous types de réparation confondus

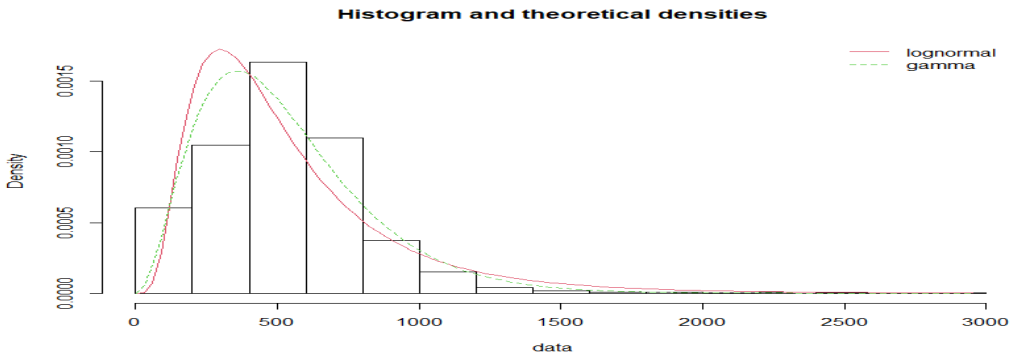


(b) Par type de réparation

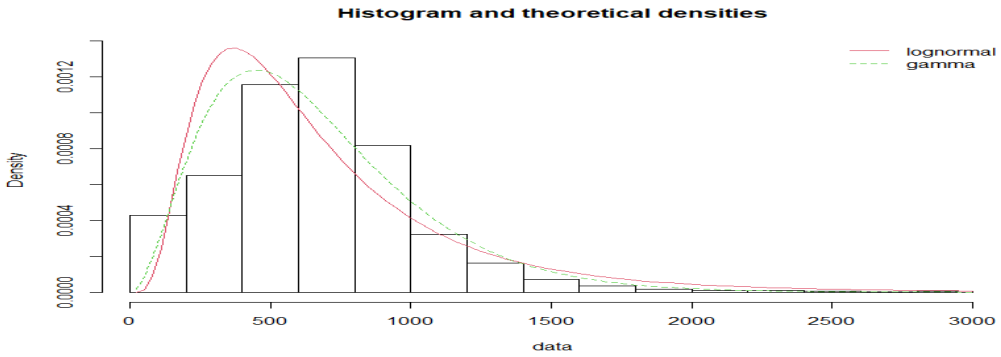
FIGURE C.1: Histogramme de la sévérité BG



(a) Tous types de réparation confondus



(b) Réparation en garage agréé



(c) Réparation en garage non agréé

FIGURE C.2: Adéquation aux deux lois Gamma et log-normale

C.2 Sévérité bris de glace selon différentes variables

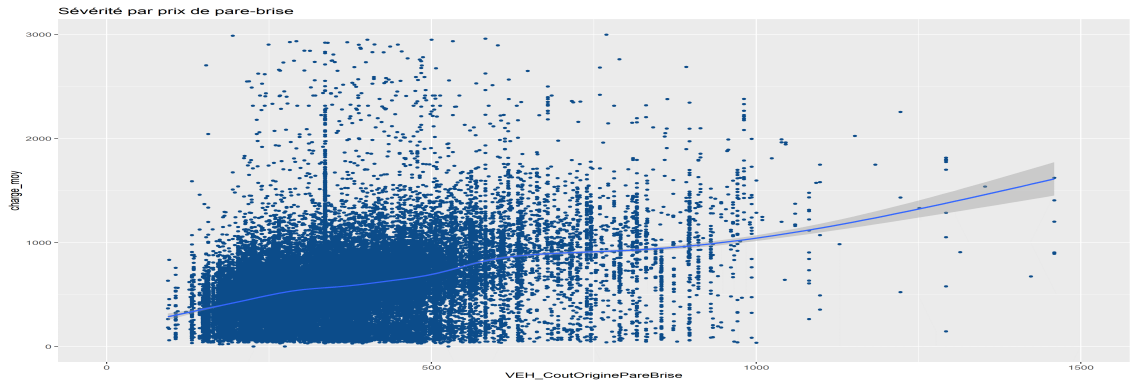


FIGURE C.3: Sévérité par prix de pare-brise

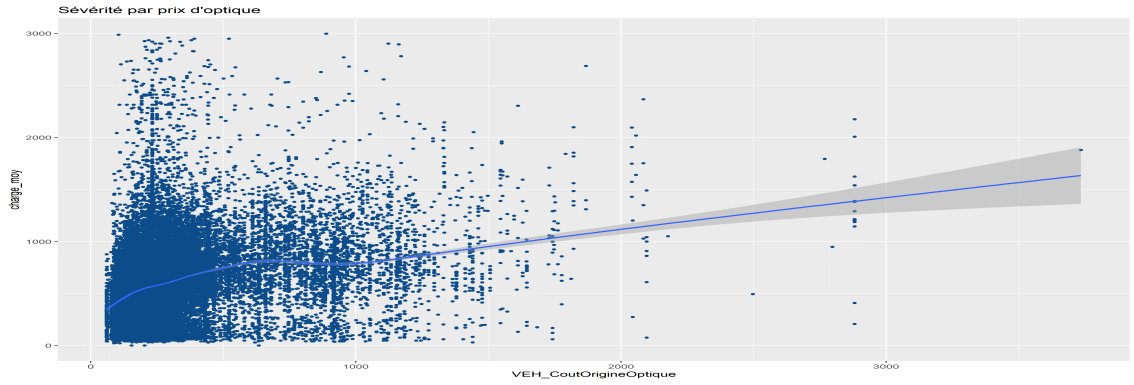


FIGURE C.4: Sévérité par prix des optiques du véhicule

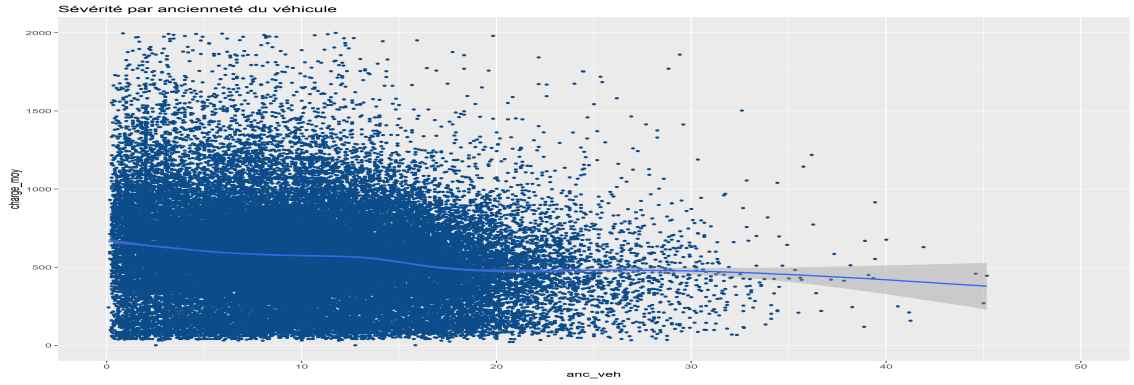


FIGURE C.5: Sévérité par ancienneté du véhicule

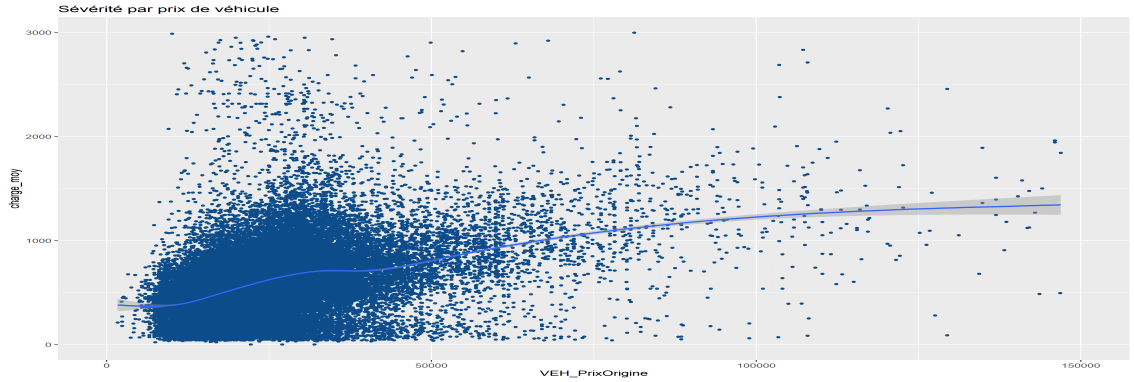


FIGURE C.6: Sévérité par prix du véhicule

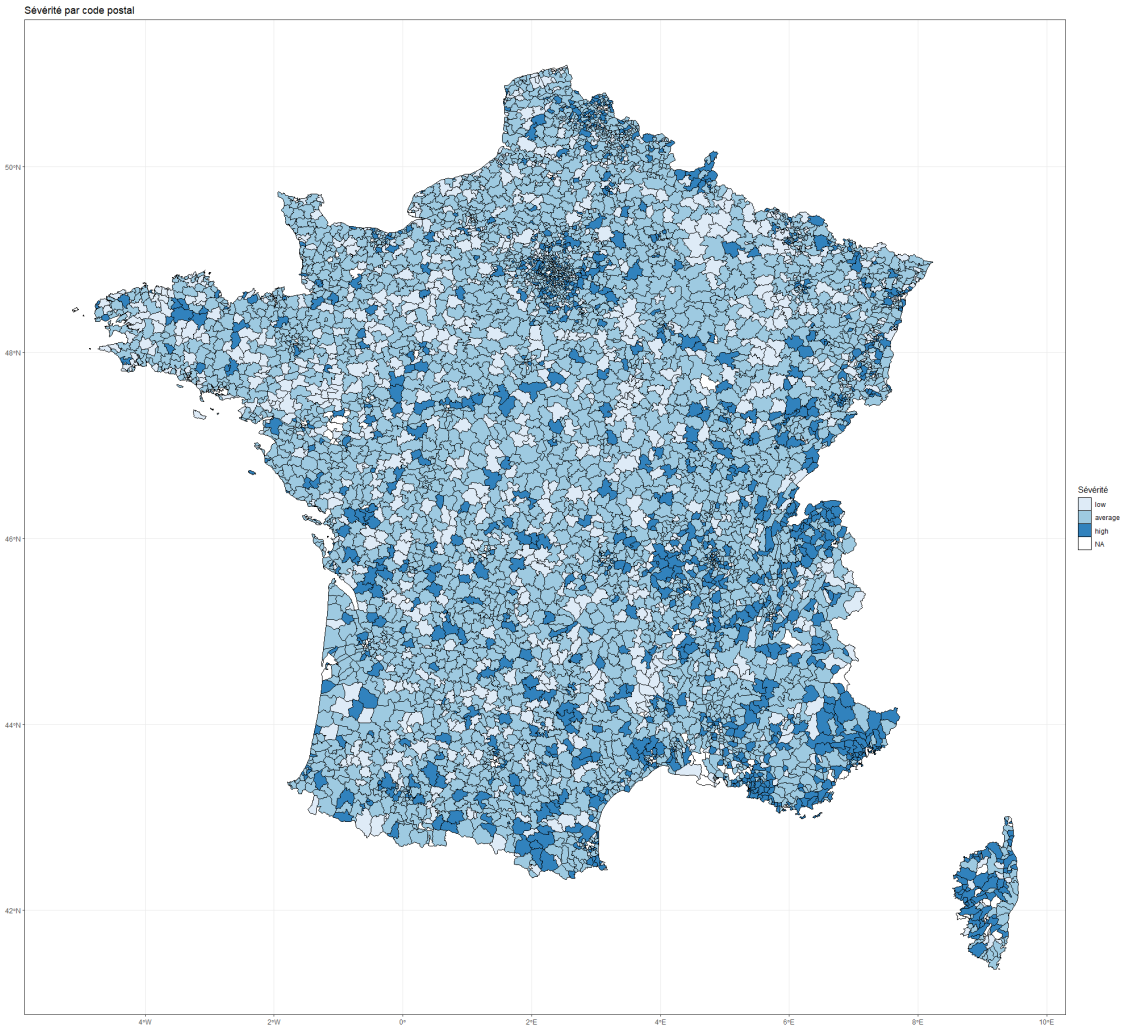


FIGURE C.7: Sévérité par code postal

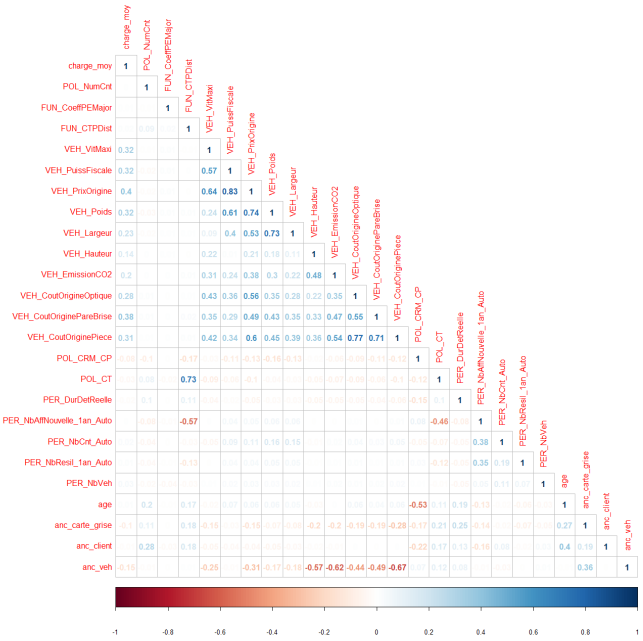


FIGURE C.8: Corrélations de Pearson des variables

C.3 La famille exponentielle

La loi de probabilité P appartient à une famille de lois de type exponentielle $\{P_\theta\}_{\theta \in \mathbb{R}^p}$ s'il existe une mesure dominante μ (Lebesgue ou mesure de comptage le plus souvent) telle que les lois P_θ admettent pour densité par rapport à μ :

$$f_\theta(y) = c(\theta)h(y)exp\left(\sum_{j=1}^{p} \alpha_j(\theta)T_j(y)\right), y \in Y$$

où $T_1, ..., T_p, \alpha_1, ..., \alpha_p$ sont des fonctions mesurables et Y l'ensemble de définition de f_θ.

Dans le cadre des modèles linéaires généralisés, nous nous intéressons à une forme particulière de la famille de loi exponentielle. La loi de probabilité possède une densité par rapport à μ qui s'écrit :

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

$a(.), b(.)$ et $c(.)$ sont des fonctions connues et dérivables, $b(.)$ est trois fois dérivable et sa dérivée première est inversible, et le couple (θ, ϕ) appartient à $\Omega \subset \mathbb{R}^2$. θ est parfois

appelé le paramètre naturel et ϕ est appelé le paramètre de dispersion. Ce dernier est estimé séparément puis est supposé connu et fixé. Nous pouvons aussi démontrer que : $\mathbb{E}(Y) = b^{'}(\theta)$ et $Var(Y) = b^{''}(\theta)a(\phi)$. On constate que l'espérance et la variance sont toutes les deux fonctions de θ et donc liées. Nous noterons ainsi : $\mathbb{E}(Y) = \mu$ et $Var(Y) = a(\phi)V(\mu)$.

Les principales lois de la famille exponentielle utilisées dans la modélisation des fréquences et sévérités sont représentées dans le tableau suivant :

Loi	θ	ϕ	a(x)	b(x)
Loi Normale $\mathcal{N}(\mu, \sigma^2)$	μ	σ^2	x	$\frac{x^2}{2}$
Loi Gamma $\mathcal{G}(\alpha, \beta)$	$-\frac{\beta}{\alpha}$	$\frac{1}{\alpha}$	x	$-log(-x)$
Loi Poisson $\mathcal{P}(\mu)$	$log(\mu)$	1	1	e^x
Loi Binomiale Négative $\mathcal{BN}(r, p)$	$log(p)$	1	1	$-rlog(1-p)$

FIGURE C.9: Paramètres de la famille exponentielle

C.4 Arbres de régression

Un arbre de régression est une méthode itérative qui consiste à partitionner un ensemble X d'observations en groupes homogènes, dit noeuds, à travers une séquence de divisions binaires. Ainsi, sur chaque nœud terminal, appelée feuille, les observations sont homogènes et la prédiction de la variable réponse est constante. Cette prédiction sera pour notre cas la moyenne de la variable d'intérêt sur chaque noeud. En effet, pour chaque noeud t comportant $N(t)$ observations (x_n), la valeur $\hat{y}$ de y_n qui minimise la somme carrée des erreurs "intra-classe" :

$$R(t) = \frac{1}{N(t)} \sum_{x_n \in t} (y_n - \hat{y})^2$$

est la moyenne :

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{x_n \in t} y_n$$

Deux critères restent à être déterminés pour la construction d'un arbre binaire : le critère de partitionnement d'un nœud en deux et un critère d'arrêt.

Pour le choix des deux sous-groupes d'un noeud, on sélectionne parmi toutes les divisions possibles S celle qui réduit le plus la somme des carrées des erreurs sur tous les noeuds de l'arbre T : $R(T) = \sum_t R(t)$. Ceci est équivalent à maximiser la diminution

d'erreur lors de la subdivision du noeud t :

$$s^* \text{ tel que } R(t) - R(t_{left}) - R(t_{right}) \text{ est maximale.}$$

En ce qui concerne la règle d'arrêt et la sélection de l'arbre optimal, on impose une taille minimale d'observations par feuille (souvent 1 % de la table d'apprentissage) ainsi que la profondeur de l'abre maximale.

C.5 Résultats modèles GAM

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	6.4169	0.0130	495.0652	< 0.0001
VEH_ClasseA	-0.2704	0.0766	-3.5288	0.0004
VEH_ClasseB	-0.3260	0.0584	-5.5781	< 0.0001
VEH_ClasseC	-0.3140	0.0445	-7.0542	< 0.0001
VEH_ClasseD	-0.3127	0.0389	-8.0398	< 0.0001
VEH_ClasseE	-0.2983	0.0361	-8.2586	< 0.0001
VEH_ClasseF	-0.2694	0.0326	-8.2605	< 0.0001
VEH_ClasseG	-0.2172	0.0296	-7.3413	< 0.0001
VEH_ClasseH	-0.1921	0.0262	-7.3166	< 0.0001
VEH_ClasseI	-0.1369	0.0228	-6.0111	< 0.0001
VEH_ClasseJ	-0.0982	0.0185	-5.3070	< 0.0001
VEH_ClasseK	-0.0762	0.0143	-5.3417	< 0.0001
VEH_ClasseL	-0.0282	0.0103	-2.7381	0.0062
VEH_GenreU	-0.1202	0.0088	-13.6158	< 0.0001
VEH_AntivolNC	0.0419	0.0059	7.0908	< 0.0001
VEH_AntivolP4	0.1470	0.0446	3.2955	0.0010
VEH_SegmentB	-0.0418	0.0080	-5.2120	< 0.0001
VEH_SegmentH	0.0641	0.0116	5.5328	< 0.0001
VEH_SegmentK1	-0.0290	0.0145	-2.0071	0.0447
VEH_SegmentK2	0.0762	0.0169	4.4986	< 0.0001
VEH_SegmentM2	0.0293	0.0078	3.7636	0.0002
VEH_Transmission4D	0.1259	0.0156	8.0932	< 0.0001
VEH_Transmission4P	-0.0237	0.0103	-1.3289	0.0039
VEH_TransmissionPR	-0.0940	0.0101	-9.2737	< 0.0001
POL_NivFranchBdGAvec franchise	-0.1284	0.0522	-2.4609	0.0139
POL_NivFranchBdGProportionnelle	-0.1912	0.0055	-34.5183	< 0.0001
region_nameAuvergne-Rhône-Alpes	0.0392	0.0069	5.6815	< 0.0001
region_nameBourgogne-Franche-Comté	0.0637	0.0100	6.3912	< 0.0001
region_nameCentre-Val de Loire	0.0483	0.0109	4.4408	< 0.0001
region_nameGrand Est	0.1038	0.0079	13.1342	< 0.0001
region_nameHauts-de-France	0.0873	0.0075	11.7080	< 0.0001
region_nameÎle-de-France	0.0531	0.0075	7.0805	< 0.0001
region_nameNormandie	0.0794	0.0073	10.8787	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(anc_veh)	6.1070	9.0000	14.3277	< 0.0001
s(VEH_Poids)	4.8810	9.0000	4.4296	< 0.0001
s(VEH_CoutOrigineOptique)	5.8879	9.0000	4.8303	< 0.0001
s(VEH_PrixOrigine)	5.8135	9.0000	20.0993	< 0.0001
s(VEH_CoutOriginePareBrise)	7.1295	9.0000	171.9742	< 0.0001
s(VEH_VitMaxi)	5.4566	9.0000	8.4165	< 0.0001
s(anc_carte_grise)	6.3013	9.0000	17.7799	< 0.0001

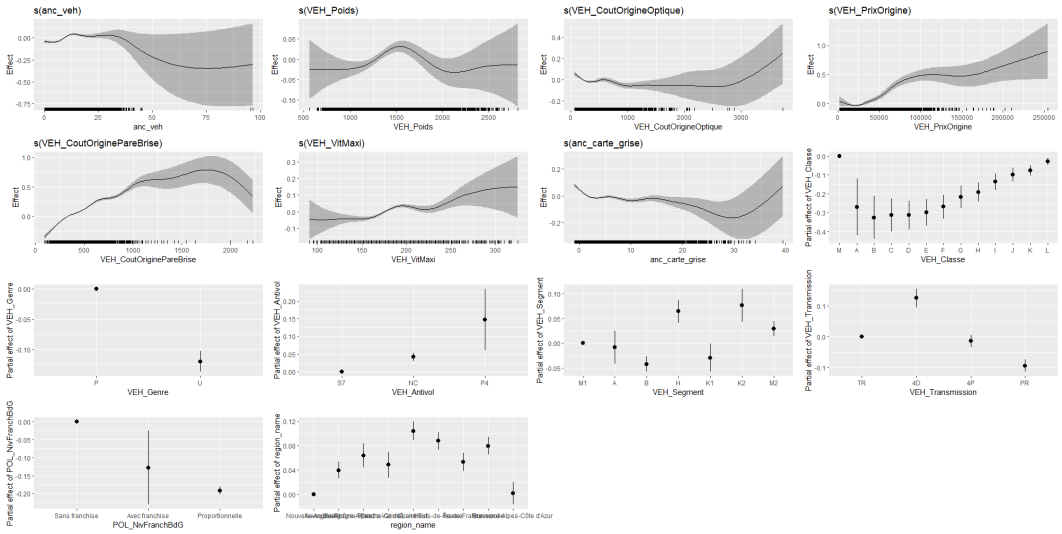
TABLE C.2: Coefficients du modèle GAM de sévérité, tous types de réparation confondus.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	6.3519	0.0150	423.1201	< 0.0001
VEH_ClasseA	-0.1752	0.0888	-1.9728	0.0485
VEH_ClasseB	-0.2988	0.0673	-4.4431	< 0.0001
VEH_ClasseC	-0.3103	0.0509	-6.1007	< 0.0001
VEH_ClasseD	-0.2700	0.0433	-6.2418	< 0.0001
VEH_ClasseE	-0.2571	0.0398	-6.4604	< 0.0001
VEH_ClasseF	-0.2171	0.0355	-6.1091	< 0.0001
VEH_ClasseG	-0.1934	0.0319	-6.0661	< 0.0001
VEH_ClasseH	-0.1703	0.0280	-6.0747	< 0.0001
VEH_ClasseI	-0.1303	0.0243	-5.3670	< 0.0001
VEH_ClasseJ	-0.0933	0.0200	-4.6743	< 0.0001
VEH_ClasseK	-0.0694	0.0160	-4.3266	< 0.0001
VEH_ClasseL	-0.0310	0.0123	-2.5164	0.0119
VEH_GenreU	-0.1647	0.0114	-14.4049	< 0.0001
VEH_AntivolNC	0.0377	0.0076	4.9665	< 0.0001
VEH_AntivolP4	0.1301	0.0569	2.2854	0.0223
VEH_SegmentA	-0.0157	0.0209	-0.7509	0.0427
VEH_SegmentB	-0.0470	0.0100	-4.6922	< 0.0001
VEH_SegmentH	0.0856	0.0154	5.5599	< 0.0001
VEH_SegmentK1	-0.0276	0.0191	-1.4427	0.1491
VEH_SegmentK2	0.0546	0.0229	2.3868	0.0170
VEH_SegmentM2	0.0417	0.0101	4.1486	< 0.0001
VEH_Transmission4D	0.1620	0.0213	7.6194	< 0.0001
VEH_Transmission4P	-0.0168	0.0138	-0.4898	0.0243
VEH_TransmissionPR	-0.1285	0.0137	-9.3942	< 0.0001
POL_NivFranchBdGAvec franchise	-0.0667	0.0637	-1.0463	0.0254
POL_NivFranchBdGProportionnelle	-0.2906	0.0072	-40.5000	< 0.0001
region_nameAuvergne-Rhône-Alpes	0.0199	0.0088	2.2637	0.0236
region_nameBourgogne-Franche-Comté	0.0691	0.0129	5.3666	< 0.0001
region_nameCentre-Val de Loire	0.0518	0.0140	3.7011	0.0002
region_nameGrand Est	0.0899	0.0105	8.5814	< 0.0001
region_nameHauts-de-France	0.0656	0.0095	6.9152	< 0.0001
region_nameÎle-de-France	0.0150	0.0093	1.6079	0.1079
region_nameNormandie	0.0566	0.0105	5.3786	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(anc_veh)	4.8356	9.0000	11.2789	< 0.0001
s(VEH_Poids)	4.4410	9.0000	2.4668	0.0001
s(VEH_CoutOrigineOptique)	3.9923	9.0000	1.9465	0.0006
s(VEH_PrixOrigine)	3.6806	9.0000	3.6070	< 0.0001
s(VEH_CoutOriginePareBrise)	6.7861	9.0000	128.4079	< 0.0001
s(VEH_VitMaxi)	4.4942	9.0000	3.6757	< 0.0001
s(anc_carte_grise)	5.3983	9.0000	3.9054	< 0.0001

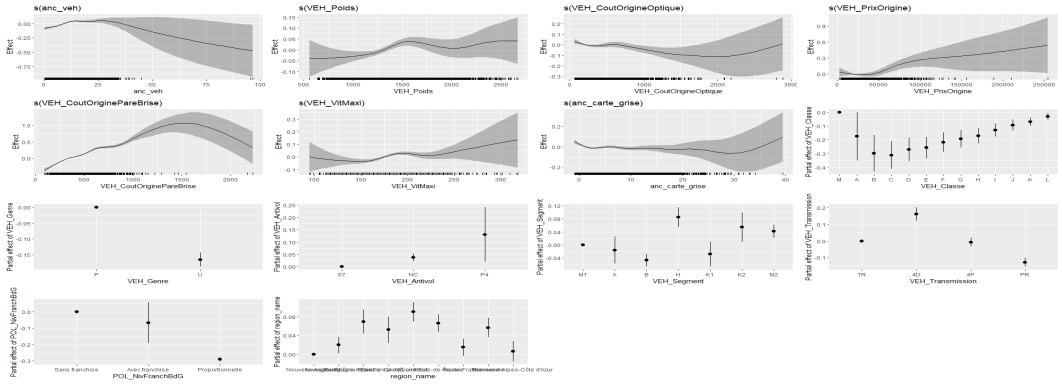
TABLE C.3: Coefficients du modèle GAM de sévérité des sinistres réparés en garages agréés.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	6.5128	0.0180	361.9641	< 0.0001
VEH_ClasseA	-0.4566	0.1318	-3.4645	0.0005
VEH_ClasseB	-0.3325	0.0978	-3.4004	0.0007
VEH_ClasseC	-0.3094	0.0703	-4.3978	< 0.0001
VEH_ClasseD	-0.3759	0.0618	-6.0851	< 0.0001
VEH_ClasseE	-0.3480	0.0570	-6.1054	< 0.0001
VEH_ClasseF	-0.3415	0.0507	-6.7377	< 0.0001
VEH_ClasseG	-0.2487	0.0454	-5.4748	< 0.0001
VEH_ClasseH	-0.2182	0.0397	-5.4893	< 0.0001
VEH_ClasseI	-0.1525	0.0342	-4.4616	< 0.0001
VEH_ClasseJ	-0.1121	0.0278	-4.0405	0.0001
VEH_ClasseK	-0.0836	0.0214	-3.9097	0.0001
VEH_ClasseL	-0.0224	0.0154	-1.4505	0.0469
VEH_GenreU	-0.0937	0.0130	-7.1955	< 0.0001
VEH_AntivolNC	0.0540	0.0088	6.1514	< 0.0001
VEH_AntivolP4	0.1910	0.0672	2.8404	0.0045
VEH_SegmentA	-0.0128	0.0259	-0.4937	0.0216
VEH_SegmentB	-0.0429	0.0122	-3.5243	0.0004
VEH_SegmentH	0.0454	0.0164	2.7780	0.0055
VEH_SegmentK1	-0.0391	0.0204	-1.9200	0.0549
VEH_SegmentK2	0.0812	0.0233	3.4805	0.0005
VEH_SegmentM2	0.0174	0.0114	1.5290	0.1263
VEH_Transmission4D	0.0714	0.0213	3.3518	0.0008
VEH_Transmission4P	-0.0237	0.0147	-1.6141	0.0165
VEH_TransmissionPR	-0.0629	0.0141	-4.4722	< 0.0001
POL_NivFranchBdGAvec franchise	-0.2240	0.0857	-2.6123	0.0090
POL_NivFranchBdGProportionnelle	-0.0940	0.0083	-11.3898	< 0.0001
region_nameAuvergne-Rhône-Alpes	0.0791	0.0105	7.5155	< 0.0001
region_nameBourgogne-Franche-Comté	0.0633	0.0149	4.2565	< 0.0001
region_nameCentre-Val de Loire	0.0414	0.0163	2.5383	0.0111
region_nameGrand Est	0.1055	0.0114	9.2644	< 0.0001
region_nameHauts-de-France	0.1191	0.0114	10.4501	< 0.0001
region_nameÎle-de-France	0.1437	0.0120	11.9807	< 0.0001
region_nameNormandie	0.0533	0.0097	5.4888	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(anc_veh)	4.6936	9.0000	9.3783	< 0.0001
s(VEH_Poids)	3.7605	9.0000	2.0037	0.0003
s(VEH_CoutOrigineOptique)	1.3874	9.0000	0.6856	0.0102
s(VEH_PrixOrigine)	6.1007	9.0000	15.5274	< 0.0001
s(VEH_CoutOriginePareBrise)	5.3308	9.0000	63.0335	< 0.0001
s(VEH_VitMaxi)	4.7842	9.0000	5.1576	< 0.0001
s(anc_carte_grise)	3.5422	9.0000	19.0165	< 0.0001

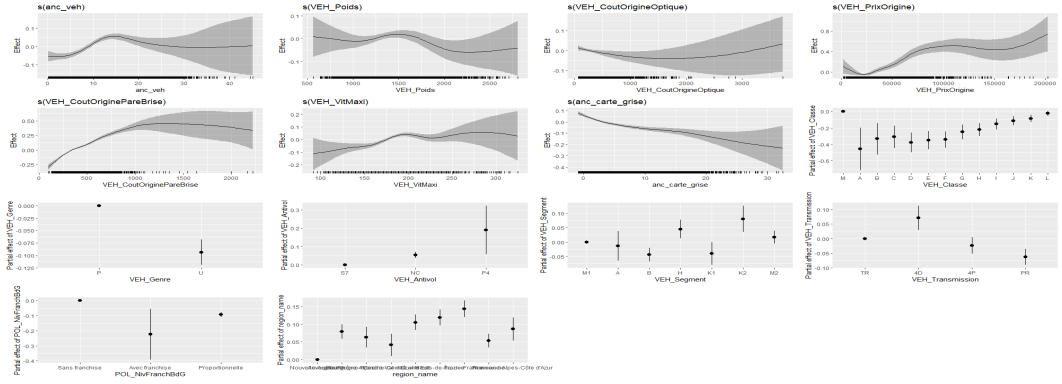
TABLE C.4: Coefficients du modèle GAM de sévérité des sinistres réparés en garages non agréés.



(a) Tous types de réparation confondus

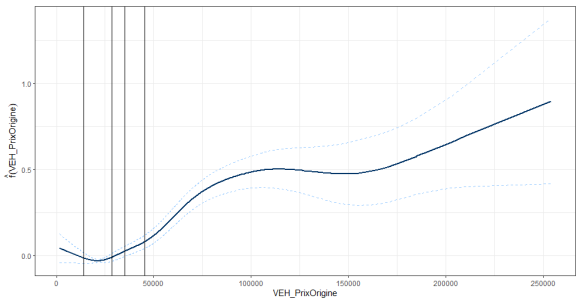


(b) Réparation en garage agréé

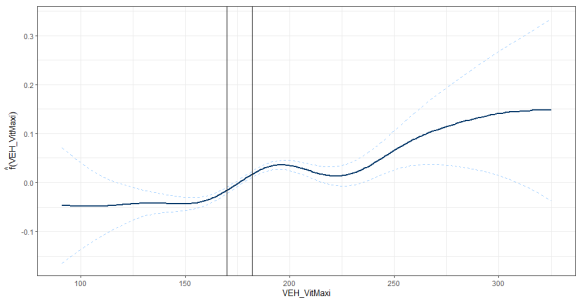


(c) Réparation en garage non agréé

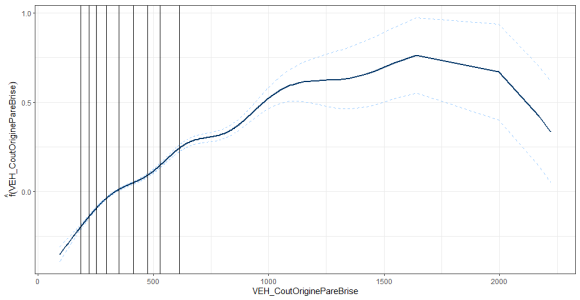
FIGURE C.10: Les composantes estimées des modèles GAM de sévérité



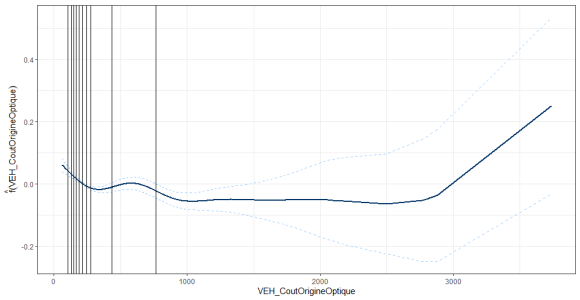
(a) Prix du véhicule



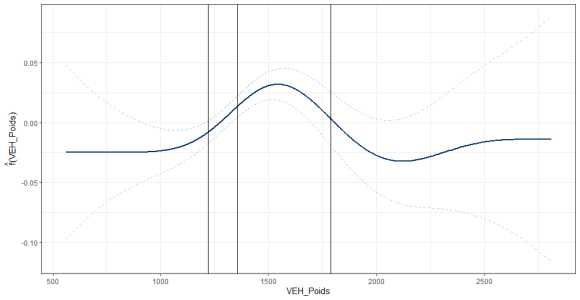
(b) Vitesse maximale du véhicule



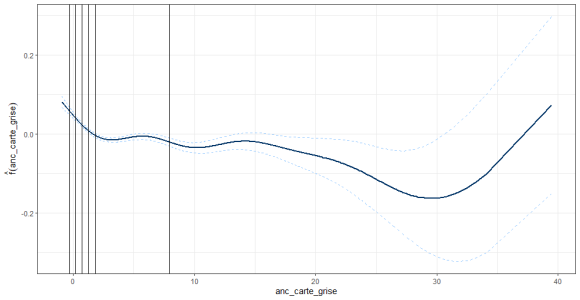
(c) Prix du pare-brise



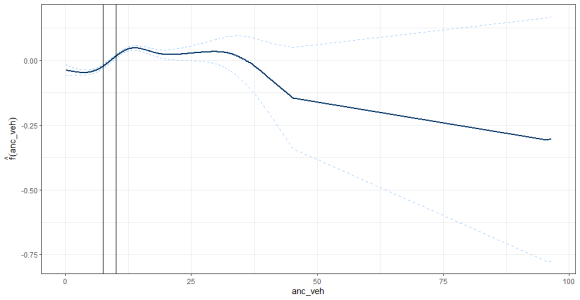
(d) Prix des optiques



(e) Poids du véhicule



(f) Ancienneté carte grise



(g) Ancienneté du véhicule

FIGURE C.11: Les variables continues discrétisées selon leurs effets sur la sévérité

	Estimate	Std. Error	t value	Pr(> | t|)
(Intercept)	6.14	0.02	288.45	0.00
anc_veh[7.46,10)	0.04	0.01	6.63	0.00
anc_veh[10,96.5]	0.06	0.01	8.30	0.00
VEH_Poids[1.22e+03,1.36e+03)	0.03	0.01	3.54	0.00
VEH_Poids[1.36e+03,1.79e+03)	0.06	0.01	5.65	0.00
VEH_Poids[1.79e+03,2.81e+03]	0.03	0.01	2.33	0.02
VEH_CoutOrigineOptique[106,149)	-0.02	0.01	-1.49	0.04
VEH_CoutOrigineOptique[149,168)	-0.03	0.01	-1.91	0.03
VEH_CoutOrigineOptique[168,190)	-0.03	0.01	-2.36	0.02
VEH_CoutOrigineOptique[190,217)	-0.06	0.01	-3.97	0.00
VEH_CoutOrigineOptique[217,245)	-0.04	0.01	-2.53	0.01
VEH_CoutOrigineOptique[245,277)	-0.07	0.01	-4.50	0.00
VEH_CoutOrigineOptique[277,437)	-0.08	0.01	-5.38	0.00
VEH_CoutOrigineOptique[437,768)	-0.04	0.02	-2.60	0.01
VEH_CoutOrigineOptique[768,3.73e+03]	-0.09	0.02	-5.16	0.00
VEH_PrixOrigine[1.66e+03,1.39e+04)	0.01	0.01	0.72	0.04
VEH_PrixOrigine[2.86e+04,3.52e+04)	0.04	0.01	4.54	0.00
VEH_PrixOrigine[3.52e+04,4.53e+04)	0.04	0.01	3.77	0.00
VEH_PrixOrigine[4.53e+04,2.54e+05]	0.23	0.02	14.93	0.00
VEH_CoutOriginePareBrise[185,221)	0.11	0.01	8.86	0.00
VEH_CoutOriginePareBrise[221,254)	0.16	0.01	11.99	0.00
VEH_CoutOriginePareBrise[254,298)	0.17	0.01	12.43	0.00
VEH_CoutOriginePareBrise[298,351)	0.25	0.01	18.81	0.00
VEH_CoutOriginePareBrise[351,414)	0.27	0.01	19.14	0.00
VEH_CoutOriginePareBrise[414,475)	0.31	0.01	20.65	0.00
VEH_CoutOriginePareBrise[475,530)	0.35	0.02	22.60	0.00
VEH_CoutOriginePareBrise[530,614)	0.45	0.02	26.86	0.00
VEH_CoutOriginePareBrise[614,2.22e+03]	0.59	0.02	34.55	0.00
VEH_VitMaxi[170,182)	0.06	0.01	8.13	0.00
VEH_VitMaxi[182,325]	0.08	0.01	9.68	0.00
anc_carte_grise[0.22,0.72)	-0.04	0.01	-3.36	0.00
anc_carte_grise[0.72,1.24)	-0.06	0.01	-5.22	0.00
anc_carte_grise[1.24,1.86)	-0.07	0.01	-6.15	0.00
anc_carte_grise[1.86,7.92)	-0.07	0.01	-7.92	0.00
anc_carte_grise[7.92,39.4]	-0.09	0.01	-9.01	0.00

TABLE C.5: Modèle GLM après discrétisation des variables continues. Pour des raisons de lisibilité, seules les coefficients de ces dernières sont présentés ici.

Note de synthèse

Mots-clés : Garages agréés, orientation, réparation, sévérité, Bagging, Boosting, CatBoost, GLM, GAM.

Introduction et problématique

Dans le cadre d'un marché très compétitif et mature, et avec la montée en puissance des bancassureurs bénéficiant d'un large réseau de distributeurs, les assureurs travaillent en permanence pour améliorer les tarifs proposés, fidéliser leurs clients et éviter des problèmes d'antisélection. Notre travail a pour but de contribuer à cette optimisation des primes pour le cas de l'assurance automobile, en considérant une segmentation plus fine selon le choix potentiel du garage de réparation.

Après survenance d'un sinistre, l'assuré a le choix entre deux types de garages : des garages agréés avec lesquels l'assureur a conclu une convention d'agrément visant à réduire les coûts des réparations en contrepartie d'un grand volume de clients, et des garages non agréés indépendants. Et depuis 2014, à la suite de la loi Hamon, l'assureur ne peut plus imposer le garage de réparation à son client et doit également rappeler à ce dernier, au moment de déclaration du sinistre, son droit de choisir librement le carrossier. Cette liberté de choix a comme effet direct l'augmentation des coûts de réparation remboursés par l'assureur, pouvant aller jusqu'à 50% de charges supplémentaires. Ainsi, pour inclure cette hausse des coûts dans les primes des assurés ayant une faible propension de réparer dans des garages agréés et récompenser les profils qui choisissent les réparateurs partenaires, il faut comprendre les déterminants de ce choix pour l'anticiper et différencier entre les deux profils d'assurés.

Ainsi, notre travail consiste à modéliser la probabilité de réparation en garages agréés pour l'inclure comme paramètre dans les modèles de prime pour une meilleure maîtrise des coûts. Notre étude se fera donc en trois étapes. Tout d'abord, nous commençons par l'étude de l'orientation en garages partenaires, variable indiquant si l'agent contacté pour la déclaration du sinistre a orienté son client vers un garage agréé. Cette orientation influencera le choix de l'assuré, hypothèse que nous confirmons dans notre étude. Ensuite, en prédisant un score d'orientation pour chaque contrat, nous modélisons notre variable

d'intérêt, à savoir le choix du réparateur, par des méthodes d'apprentissage automatique. Enfin, nous démontrons comment l'intégration du choix du garage dans les modèles de primes permet de constituer des tarifs plus exacts et de réduire les erreurs de prédiction des montants des sinistres.

Le choix du garage, un mécanisme en deux étapes

Après déclaration du sinistre, l'agent décide d'orienter son assuré vers un garage partenaire ou vers un garage non-partenaire selon ses caractéristiques, celles de l'assuré ou du véhicule. Nous nous intéressons particulièrement à l'orientation par l'agent puisqu'elle devrait influencer le choix final de l'assuré. L'hypothèse que nous démontrons est qu'*un assuré qui a été orienté vers un garage partenaire devrait plus probablement effectuer la réparation dans ce dernier.*

Le choix du garage de réparation est également lié aux avantages offerts par chaque type de garage. En garages partenaires, et selon chaque compagnie d'assurance, l'assuré bénéficie de nombreux avantages. En effet, un des points forts de ce type de garage est la non-avance des frais : le règlement des coûts de réparation est directement effectué par l'assureur, et l'assuré doit verser uniquement la franchise du contrat souscrit. De plus, pour l'obtention d'un agrément, l'assureur exige au garage de répondre à un nombre de critères de qualité pour avoir la garantie d'un travail bien fait. Ensuite, en déposant son véhicule dans un des garages partenaires, l'assuré bénéficie d'un véhicule de prêt pour assurer sa mobilité durant la période de réparation.

Face à ces nombreux avantages, et pour renforcer le libre choix de l'assuré, un mécanisme de cession de créance est allégé pour les garages non-partenaires pour permettre à ces derniers de recevoir un versement direct des coûts de réparation par l'assureur. Parmi les autres facteurs qui peuvent pousser l'assuré à réparer dans un garage libre est la rapidité et le soin du service, puisque le rythme de travail est beaucoup moins soutenu pour les garages non agréés qui ne reçoivent pas les véhicules des assureurs. L'assuré peut également souhaiter d'effectuer la réparation chez son garage habituel. Enfin, ces garages indépendants peuvent offrir également d'autres avantages pour garder une longueur d'avance sur la concurrence (remboursement de la franchise, programmes de fidélité...).

Cette première analyse nous permet de comprendre la complexité du choix du garage de réparation et sa dépendance de plusieurs variables liées aux caractéristiques de l'agent,

de l'assuré ou du véhicule assuré... C'est ainsi que nous allons modéliser ce phénomène par des méthodes d'apprentissage automatiques pour capter ces dépendances et prédire avec précision le choix du carrossier pour chaque contrat.

Modélisation et résultats

Notre étude concerne un portefeuille de contrats sinistrés renouvelés en 2018, leurs sinistres survenus sur une année d'observation, selon la date d'échéance, et les données d'orientation et de réparation sur cette période. La première variable d'orientation est une variable binaire décrivant si l'agent a orienté l'assuré vers un garage agréé, et la seconde si l'assuré effectue effectivement la réparation dans l'un de ces garages.

Pour modéliser ces variables, trois modèles ensemblistes de classification ont été testés. Les *forêts aléatoires*, exemple de la méthode du bootstrap aggregating ou *Bagging* et basés sur les arbres CART, effectuent un tirage aléatoire avec remise des observations pour chaque modèle mais également parmi les variables explicatives $X_1, \ldots, X_p$ pour chaque arbre. Cela permet de créer des arbres non corrélés, car chacun entraîné sur une partie de la base d'apprentissage, et de réduire ainsi le sur-ajustement du modèle final.

La deuxième méthode de construction des arbres est le *Boosting*, qui applique une variante de la descente du gradient à un espace fonctionnel d'arbres de décision et actualise le prédicteur en minimisant une fonction de perte. Comme exemple de cette méthode, nous choisissons d'appliquer le XGBoost qui améliore l'algorithme du Boosting en ajoutant une composante de régularisation à la fonction objective, ce qui permet de créer des arbres moins complexes et de réduire le surapprentissage.

Ces deux derniers modèles nécessitent un encodage préalable des variables catégorielles avant de les inclure comme prédicteurs, ce qui conduit souvent à une perte d'information, comme est le cas pour *OneHotencoder* qui nécessite le regroupement des modalités si elles sont nombreuses, ou introduit un ordre incorrect dans des variables non ordinales si le *labelencoder* est utilisé. C'est ainsi que nous avons testé une nouvelle méthode améliorée du Boosting intitulée *CatBoost*, et qui introduit deux modifications par rapport aux autres modèles : *ordered boosting* et *ordered target statistic*. En effet, les auteurs du CatBoost démontrent que l'algorithme du boosting utilisé dans tous les modèles classiques cause un problème de généralisation, puisque le gradient calculé à chaque itération utilise les

mêmes valeurs de la variable cible sur lesquels il a été optimisé.

Pour résoudre ce problème, une permutation des observations σ est effectuée et chaque arbre j ne prend en entrée que les j premières observations de la base permutée, ce qui permet d'éviter d'estimer le gradient avec la même observation qui l'a construite. Le même principe de permutation est considéré pour encoder les variables catégorielles, qui modifie le *target encoder* en créant un temps artificiel et en utilisant uniquement les observations précédant celle à encoder pour calculer la moyenne de la variable Y à modéliser sur cette catégorie, et prendre cette moyenne comme encodage.

Ces trois modèles appliqués sur nos observations pour la variable d'orientation puis pour la variable de réparation donnent les résultats suivants :

	Modèle	**AUC**
Orientation	Forêts aléatoires	0,685
	XGBoost	0,695
	CatBoost	0,717
Réparation	Forêts aléatoires	0,658
	XGBoost	0,655
	CatBoost	0,680

TABLE 4.6: Comparaison des modèles

Le modèle CatBoost donne une meilleure performance par rapport aux deux autres modèles, et ce pour les deux variables d'orientation et de réparation. En termes d'importance des variables, ce sont les caractéristiques de l'agent qui ressortent en premier pour l'orientation. En effet, maillon clé de la variable d'orientation, l'agent choisit plus ou moins d'orienter ses clients selon son identifiant, son emplacement et la densité des garages partenaires à sa proximité. L'orientation dépend également de la profession et du nombre de contrats souscrits.

En ce qui concerne la variable de réparation, la variable la plus importante est le score d'orientation prédit. De plus, la probabilité du choix de garage partenaire est croissante avec la probabilité d'orientation prédite, ce qui confirme notre hypothèse de départ. Ensuite, c'est la région d'habitation de l'assuré qui justifie en second lieu le choix du garage. Cette variable région peut être liée à plusieurs facteurs : la non-disponibilité de garages agréés dans cette région, un niveau de vie faible poussant les assurés à réparer dans les

garages agréés pour bénéficier de la non-avance des frais...

Enfin, nous justifions l'utilité de notre modèle de prédiction du choix du garage de réparation dans l'amélioration des tarifs proposés. Étant donné que ce choix impacte principalement les coûts de réparation, nous allons nous intéresser à la modélisation de la sévérité des sinistres. Pour ce faire, deux sévérités de type bris de glace seront comparées, une sévérité où la probabilité de réparer en garages partenaires est négligée et une sévérité mixte de la forme :

$$\begin{aligned}\mathbb{E}(\text{Sévérité}) = \mathbb{E}(\text{Sévérité agréé}) \times \text{Probabilité de réparer en garage agréé} \\ + \mathbb{E}(\text{Sévérité non agréé}) \times (1 - \text{Probabilité de réparer en garage agréé})\end{aligned}$$

Pour modéliser ces sévérités et prendre en compte les effets non linéaires que peuvent avoir certaines variables tarifaires sur les coûts de réparation, nous employons des modèles additifs généralisés et présentons une approche de discrétisation des variables continues basée sur les effets estimés de ces variables sur la sévérité. Pour comparer les performances de ces modèles, nous nous basons sur des mesures des écarts entre les observations et les prédictions.

En prédisant le type de garage de réparation pour chaque assuré et en différenciant la sévérité selon ce choix, l'erreur des prédictions moyenne est réduite d'environ **8%** et la prédiction de la sévérité est plus ajustée aux coûts observés chez les deux types d'assurés. Cela démontre que les modèles de classification précédents permettent de bien classifier les assurés selon leur choix du type de garage de réparation et que nous obtenons des sévérités reflétant les vrais niveaux de risque des assurés : elles sont moins élevées pour les assurés ayant tendance à réparer leurs véhicules dans des garages partenaires alors que les charges supplémentaires des garages non agréés sont prises en charge par le second type d'assurés.

Donc, les primes versées par chaque type d'assurés correspondent à ce que ces derniers engendrent comme coûts pour l'assureur. C'est pour cette raison que l'implémentation de cette modification de primes que nous proposons doit être précédée par une étude du type d'assurés prédominant dans le portefeuille de l'assureur (ceux qui réparent en garages agréés ou non agréés), de leur sensibilité à des variations de primes ainsi que de la rentabilité totale de chacun des deux segments, pour mesurer l'impact de cette nouvelle prime.

Conclusion

Notre objectif était de prédire pour chaque contrat les deux scores d'orientation et de réparation au moment du renouvellement et d'intégrer la probabilité de réparation prédite comme paramètre de la sévérité. C'est ainsi que nous avons privilégié des modèles nécessitant peu de traitement poussé des observations tout en profitant au maximum des informations contenues dans cette base de données. Et étant donné la prédominance des variables catégorielles dans cette dernière, nous avons choisi de tester le modèle CatBoost.

Nous avons constaté qu'une combinaison des caractéristiques de l'agent, de l'assuré et du véhicule était à l'origine de chaque phénomène et en les utilisant nous pouvons distinguer les deux profils d'assurés et prédire pour chaque contrat la probabilité qu'un client choisisse un garage agréé plutôt qu'un garage indépendant. Toutefois, la précision de ces prédictions peut toujours être améliorée. En effet, intégrer plus de variables ayant un effet direct sur le choix de réparation (le revenu de l'assuré, l'historique des types de garages des réparations passées, types de garage à proximité du domicile de l'assuré...) et considérer un intervalle temporel d'observation plus large permettraient aux modèles de distinguer plus les deux classes d'assurés.

À l'aide de cette probabilité, nous proposons un nouveau modèle de sévérité s'ajustant mieux aux observations et permettant ainsi une meilleure maîtrise des risques du portefeuille. Notre étude permet ainsi de segmenter plus finement les tarifs proposés et de construire des classes plus homogènes de risque en anticipant le choix du garage de réparation dans le cas d'un sinistre.

Executive summary

Keywords : : Insurer approved repair shop, orientation, repair, Bagging, Boosting, CatBoost, GLM, GAM.

In a highly competitive and mature market, and with the rise of bancassurers benefiting from a large network of distributors, insurers are constantly working to improve their premiums, build customer loyalty and avoid adverse selection problems. Our work aims to contribute to this optimization of premiums in the case of car insurance, by considering a new criterion to differentiate the premiums : the choice of the car repair shop.

After an accident, the insured can chose between two types of car repair shops : licensed garages with which an agreement is signed to reduce the reparation costs, in exchange of a high volume of clients, and independent garages. Since 2014, following the Hamon law, the insurer can no longer impose the repair shop on its clients, even more, it must remind the clients of their right to freely choose the repair shop when declaring an accident. This freedom of choice has as a direct effect the increase of the repair costs reimbursed by the insurer, which can go up to 50% of additional charges. Thus, to include this increase in costs in the premiums of policyholders with a low propensity to repair in approved garages and to reward the profiles that choose partner repairers, it is necessary to understand the reasons behind this choice in order to anticipate it and differentiate between the two types of policyholders.

Our work consists in modeling the probability of repairing in approved car shops to include it as a parameter in premium models for better cost understanding. Our study will therefore be carried out in three steps. First of all, we begin by studying the referral in partner garages, a variable indicating whether the agent contacted for the claim has referred his client to an approved garage. This orientation will influence the choice of the insured, an assumption that we confirm in our study. Then, by predicting an orientation score for each contract, we model our variable of interest, i.e. the choice of repairer, using machine learning methods. Finally, we show how the integration of the choice of garage in premiums' models allows for more accurate pricing and reduces errors in predicting claim amounts.

The choice of the garage, a two-step mechanism

After reporting the claim, the agent decides to refer the insured to a preferred car repair shop or to a a different type of car shop, depending on the characteristics of the insured and the vehicle. These insurance agents receive commissions depending on the volume of contracts they manage and their value, and therefore, building customer loyalty is essential to them. We can thus assume that to ensure that the most profitable customers are satisfied with the repairs, the agent directs them to automobile repair shops known for their quality of service even though they are not approved by the insurer. We are particularly interested in agent orientation since it should influence the final choice of the insured. Indeed, the hypotheses that we are going to prove is that *an insured who has been referred to an a car shop is more likely to perform the repair in the latter*.

The choice of repair shops is also linked to the advantages offered by each type. In approved garages, and for each insurance company, the client can benefit from various advantages. Indeed, one of the strong points of this type of garage is that the payment of the bill is directly made by the insurer, and the insured must pay only the deductible of the contract. In addition, to obtain an approval, the insurer requires the garage to meet a number of quality criteria to guarantee client satisfaction.

To reinforce the insured's freedom of choice, a similar mechanism is implemented for non approved car shops, allowing them to receive direct payment of repair costs from the insurer. Among other factors that may lead the insured to repair in an independent garage is the speed and care of service, since the pace of work is much slower for unlicensed garages that do not receive vehicles from insurers. The insured may also wish to have the repair done at his or her regular garage. Finally, these independent garages may also offer other advantages to stay ahead of the competition (reimbursement of the deductible, loyalty programs ...).

This first analysis allows us to understand the complexity of the choice of repair garage and its dependence on several variables related to the characteristics of the agent, the insured, the insured vehicle, etc. We will therefore model this phenomenon using machine learning methods to capture these dependencies and accurately predict the choice of the auto body repairer for each contract.

Models and Results

In our study, we will work on a portfolio of contracts renewed in 2018, their claims incurred over a one year period and the orientation and repair information over this period. The first binary variable of orientation describes whether the agent referred the insured to a licensed garage, and the second describes whether the client actually performed the repair in one of these garages.

To model these variables, three classification ensemble models were tested. The *Random Forests*, example of the bootstrap aggregating method or *Bagging* and based on CART trees, perform a random draw of observations for each model but also among the features $X_1, \ldots, X_p$ for each tree. This creates uncorrelated trees, as each one is trained on a part of the learning data frame, and thus reduces the over-fitting of the final model.

The second method to build trees is *Boosting*, which applies a variant of the gradient descent to a functional space of classification trees and updates the predictor by minimizing a loss function. As an example of this method, we choose to apply the XGBoost which improves the Boosting algorithm by adding a regularization component to the objective function, thus creating less complex trees and reducing over-fitting.

These last two models require a prior encoding of categorical features before including them as predictors, which often leads to a loss of information, as is the case for *OneHotencoder* which requires the grouping of categories when the feature has high cardinality, or introduces an incorrect order in non-ordinal features if the *labelencoder* is used. This is why we tested a new improved method of boosting called *CatBoost*, which introduces two new concepts : *ordered boosting* and *ordered target statistic*. In fact, the authors of the CatBoost show that the usual boosting algorithm causes a generalization problem, since the gradient calculated at each iteration uses the same values of the target variable on which it has been optimized.

To solve this problem, a permutation of the observations σ is performed and each tree j takes as input only the first j observations of the permuted data frame, thus avoiding to estimate the gradient with the same observation that constructed it. The same principle of permutation is considered for encoding categorical variables, which modifies the *target encoder* by creating an artificial time and using only the observations preceding the one to be encoded to compute the mean of the target feature Y on this category, and take

the calculated mean as the encoded value.

These three models applied to our observations for the orientation feature and then for the choice of the auto car shop give the following results :

	Model	**AUC**
Orientation	Random Forests	0,685
	XGBoost	0,695
	CatBoost	0,717
Repair	Random Forests	0,658
	XGBoost	0,655
	CatBoost	0,680

TABLE 4.7: Model comparison

The CatBoost outperforms the other two models on both the orientation and repair variables. In terms of the importance of the variables, it is the agent characteristics that are the most significant for orientation. Indeed, being a key actor in the orientation process, the agent chooses more or less to refer his customers to preferred car repair shops depending on his identifier, his address and the density of approved garages in his vicinity.

For the repair variable, the most important variable is the predicted orientation score. Moreover, the probability of choosing a partner garage increases with the predicted orientation probability, which confirms our initial hypothesis. Second, it is the address of the insured that justifies the choice of the garage in the second place. This feature can be linked to several factors : the unavailability of approved garages in this region, a low standard of living leading policyholders to repair in approved garages in order to not pay in advance the fees of the reparation ...

Finally, we justify the usefulness of our model for predicting the choice of repair garage in the improvement of the premiums. Given that this choice mainly impacts repair costs, we will focus on the modeling of the severity of claims. To do so, two severities of glass damage will be compared, one where the probability of repairing in partner garages is neglected and one that includes this repair choice.

To model these claim severities and to take into account the non-linear effects that certain tariff variables may have on repair costs, we use generalized additive models and

present a discretization approach for continuous variables based on the estimated effects of these variables on severity. To compare the performance of these models, we rely on measures of the differences between observations and predictions.

By predicting the type of the car repair shop for each insured and differentiating the severity based on that choice, the average prediction error is reduced by about **8%** and the severity prediction is closer to the observed claim costs for both types of insured. This shows that the previous classification models correctly classify the insureds according to their choice of repair garage type and that we obtain severities that reflect the true risk levels of the insureds : they are lower for insureds who tend to repair their vehicles in partner garages while the additional costs of non-approved garages are paid by the second type of insureds. Thus, the premiums paid by each type of insured would correspond to what they cost for the insurers. Therefore, implementing this new premium that we propose should be preceded by a study of the predominant type of insureds in the insurer's portfolio (those who repair in approved or non-approved garages), of their sensitivity to premium variations, and of the total profitability of each of the two segments in order to measure the impact of this new premium.

Conclusion

Our objective was to predict both orientation and repair scores for each contract at renewal and to integrate the predicted repair probability as a parameter of severity. Thus, we prefer models requiring less preprocessing of the features while also benefiting from the information available in the data frame. And given the predominance of categorical features in the database, we chose to test the CatBoost model.

This last model gives a better performance compared to the two other models tested. Nonetheless, the accuracy of the predicted scores can always be improved. Indeed, integrating more features that have a direct effect on the choice of the car repair shop (the insured's income, past repair garage types, types of garages near the insured's home ...) would allow the models to distinguish more between the two classes of insured.

Using this probability, we calculate a more accurate severity model that allows a better control of portfolio risks. Our study thus enables us to build more homogeneous risk classes by anticipating the choice of repair garage in the event of a claim.