Mémoire présenté le :

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

Par : SANGONG DONGFACK BLONDEL JAPHET

Titre : Construction d'un modèle de conversion en assurance automobile

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

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

Membre présents du jury de l'Institut des Actuaires — signature

F. PLANCHET
R. CAILLET

Membres présents du jury de l'ISFA

D. DOROBANTU
P. RIBEREAU

Entreprise :

Nom : ALLIANZ FRANCE

Signature :

Directeur de mémoire en entreprise :
Nom : David JAOUEN

Signature :

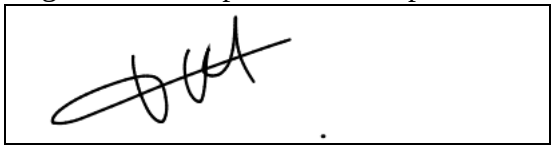
Invité :

Nom :

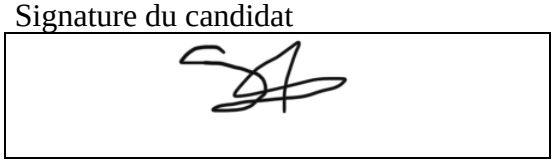
Signature :

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

Signature du responsable entreprise



Signature du candidat



Résumé

Le Groupe Allianz a à coeur de mettre l'excellence technique au centre de son activité. Cette vision est à l'origine du projet de fiabilisation et d'uniformisation du processus de tarification au sein des différentes entités. Sur le marché français de l'assurance automobile, ce projet se reflète par un accompagnement lors de la mise à jour des tarifs. Les directives données par le Groupe définissent un canevas pour le traitement des données, le modèle de tarification et le calcul des indicateurs de rentabilité. Un élément clé dans ce processus est l'estimation du taux de conversion. Le taux de conversion correspond à la proportion de devis convertis par rapport au nombre de devis réalisés. Estimer le taux de conversion donne une idée de la réponse des prospects au tarif proposé. Le thème de ce mémoire est la construction d'un modèle qui permet d'estimer le taux de conversion : ce modèle est appelé modèle de conversion. Dans la première partie, nous présentons le contexte dans lequel s'inscrivent nos travaux. Ensuite, nous décrivons les données utilisées dans le cadre de la modélisation. Dans cette section, sont mis en avant les indicateurs de compétitivité et les données externes utilisées pour la modélisation. Ces élements nous permettent de prendre en compte le positionnement tarifaire d'Allianz sur le marché, ainsi que la représentativité du réseau de distribution. Enfin, vient la phase de construction des modèles. Au terme de l'étude, le gradient boosting se démarque par une meilleure qualité de prédiction et un AUC plus élevé. Une fois implémenté ce modèle sera un appui pour orienter la politique tarifaire à mettre en place.

Mots-clés : devis, assurance automobile, taux de conversion, modèle de conversion, indicateurs de compétitivité, données externes, regression logistique, random forest, gradient boosting, AUC.

Abstract

The Allianz Group is committed to putting technical excellence at the heart of its business. This vision is at the origin the project to make the pricing process more reliable and standardised within the entities. On the french motor insurance market, this project is reflected in the support provided when updating rates. The guidelines issued by the Group define a framework for data processing, the pricing model and the calculation of performance indicators. A key element in this process is the estimation of the conversion rate. The conversion rate is to the proportion of quotations converted in relation to the number of quotations carried out. Its estimation gives an idea of the response of prospects to the proposed rate. Our topic is the construction of a model which enables to estimate the conversion rate : this model is called the conversion model. In the first part, we present the context in which our work takes place. Next, we describe the data used in the modelling. In this section, we highlight the competitiveness indicators and external data used for the modelling. These elements allow us to take into account Allianz's pricing position on the market, as well as the representativeness of the distribution network. Finally, comes the models construction. The gradient boosting is characterised by a better quality of prediction and AUC. Once implemented, this model will be a support to guide the tariff policy to be implemented.

Keywords : quote, motor car insurance, conversion rate, conversion model, competitiveness indicators, external data, logistic regression, random forest, gradient boosting, AUC.

Remerciements

La réalisation de ce mémoire a fait l'objet de la convergence d'efforts de plusieurs personnes, envers qui je tiens à manifester ma reconnaissance.

Je remercie tout d'abord mon tuteur M David JAOUEN pour m'avoir offert l'opportunité de monter en compétences, par sa disponibilité et ses conseils prolifiques. Je remercie également Mme Magali VACHEROT, à la tête de la Direction Standard Product Pricing, de m'avoir accueilli dans son équipe et de m'avoir proposé ce sujet de mémoire.

Je remercie également l'ensemble des collaborateurs de la Direction Standard Product Pricing, pour leur aide indispensable à la compréhension des sujets annexes à celui traité dans ce rapport. Particulièrement, Mme Claire LAMON, pour ses conseils et son implication durant la préparation de ce mémoire.

Enfin, je remercie l'ensemble du corps enseignant et du personnel administratif de l'I.S.F.A, pour les connaissances transmises et le suivi durant ma formation. Tout particulièrement mon tuteur, M Pierre RIBEREAU, pour le suivi et les conseils durant ma formation.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'un des projets au sein du Groupe Allianz en assurance non-vie consiste à fiabiliser et uniformiser le processus de tarification des contrats dans les différentes entités. A cet effet, le groupe a mis en place des directives afin de piloter le suivi de ce projet. L'un des aspects porte sur la mise à jour des tarifs des affaires nouvelles. En France, cette tâche a été confiée à la Direction Standard Product Pricing, qui s'occupe des marchés comme la multirisque habitation, le marché des professionnels, les flottes automobiles et l'automobile pour les particuliers. Les sujets traités dans ce mémoire portent sur l'assurance automobile des particuliers, particulièrement sur les véhicules quatre roues à moteur.

Dans le cade de la mise à jour tarifaire, le cahier des charges prévoit la mise en place d'un outil applelé **Scenario Testing New Business**. Il s'agit d'un outil d'aide à la décision lors de la détermination du nouveau tarif. Il revient donc à l'équipe de définir des indicateurs pertinents pour construire le Scenario Testing. C'est dans ce contexte que s'inscrit la problématique de ce mémoire : **« Comment intégrer la prédiction de la probabilité de conversion dans le processus de mise à jour tarifaire ?»**

Pour y répondre, notre mission a consisté à construire un modèle dit de conversion. Celui-ci sera utilisé pour projeter l'impact de la variation de tarif sur la prédiction de la probabilité de conversion d'un devis. Convertir un devis revient à le concrétiser en contrat ; l'on devient ensuite effectivement client de l'assureur. L'assureur peut ensuite déterminer le taux de conversion, c'est-à-dire la proportion de devis concrétisés en contrats sur l'ensemble des émissions de devis : $\frac{\text{nombre de devis convertis}}{\text{nombre de devis émis}}$.

D'autre part, notons que le marché de l'assurance automobile en France est un marché très concurrentiel. Cette concurrence réside sur le fait que l'assurance automobile est un produit d'approche, sur lequel les acteurs du marché s'appuient pour proposer d'autres produits à leurs clients. Cette concurrence est d'autant plus accrue avec l'arrivée sur le marché de nombreux acteurs, tandis qu'on observe une stagnation du parc automobile français sur la dernière décennie. Tout ceci a pour conséquence de rendre le marché dynamique. Dans de telles conditions, nous sommes tenus de nous adapter et de trouver des solutions afin de maintenir notre activité rentable. Ce qui passe par l'amélioration de nos modèles. C'est la raison pour laquelle dans nos travaux, nous avons cherché à exploiter des sources d'information externes, qui auraient potentiellement une influence sur la prédiction

de l'acte de conversion. Deux axes ont été pris en compte : l'environnement de l'assuré et l'effet de la concurrence.

Pour évaluer l'impact qu'aurait la concurrence, nous avons utilisé des indicateurs de compétitivité. Ceux-ci nous donnent une idée de notre positionnement tarifaire par rapport au marché. Nous avons également introduit la répartition géographique du réseau de distribution, sur l'ensemble de la France métropolitaine, comme source d'information. En plus de ces informations, nous mettons en avant la notion de potentiel d'une zone géographique. Il s'agit d'une estimation du montant des primes d'assurance consommées dans une zone géographique.

Selon les directives du projet, nous devons recourir à différentes méthodes de modélisation et déterminer l'approche présentant les meilleurs résultats. Trois méthodes ont été testées dans le cadre de ce mémoire. La première est la régression logistique. C'est la méthode la plus utilisée lorsque la variable réponse Y est binaire. Elle a l'avantage d'être simple à mettre en place et permet notamment d'interpréter les relativités entre les coefficients du modèle. Ensuite, nous nous sommes appuyés sur des algorithmes de machine learning, en l'occurrence le random forest et le gradient boosting.

Pour présenter l'ensemble de nos travaux, nous avons subdivisé ce rapport en quatre parties. La première nous permettra de mettre en avant le contexte dans lequel s'inscrivent nos travaux. Ensuite, nous expliciterons la construction des données utilisées pour la modélisation. Puis, nous aborderons les éléments théoriques sur lesquels reposent les différentes approches. Enfin, nous présenterons comment les différentes méthodes ont été appliquées et les résultats obtenus.

Chapitre 1

Contexte

Dans ce chapitre, il sera question pour nous de présenter le marché de l'assurance automobile en France et le contexte dans lequel s'inscrivent nos travaux. Nous montrerons que le parc automobile français alimente un marché de l'assurance très concurrentiel. Ensuite, nous présenterons l'offre d'assurance automobile Allianz. Ceci nous permettra de poser un cadre et donner par la suite, l'enjeu de la construction d'un modèle de conversion.

1.1 Le marché de l'assurance automobile

En France, sur la dernière décennie, on a observé une croissance faible, voire une stagnation du parc automobile. En effet, on constate que le parc automobile français a évolué d'à peine 6% entre janvier 2010 et janvier 2019. Il comptait en janvier 2019 un peu plus de 39,9 millions de véhicules, dont plus de 80% étaient des véhicules de particuliers. Ce qui constitue tout de même une masse assurable considérable pour le marché français de l'assurance automobile. Surtout que, depuis la loi du 27 février 1958 sur l'assurance automobile, reprise notamment aux articles L.211-1 du Code des assurances et L. 324-1 du Code de la route, il est obligatoire de souscrire une assurance pour « tout véhicule automoteur destiné à circuler sur le sol et qui peut être actionné par une force mécanique sans être lié à une voie ferrée, ainsi que toute remorque, même non attelée ».

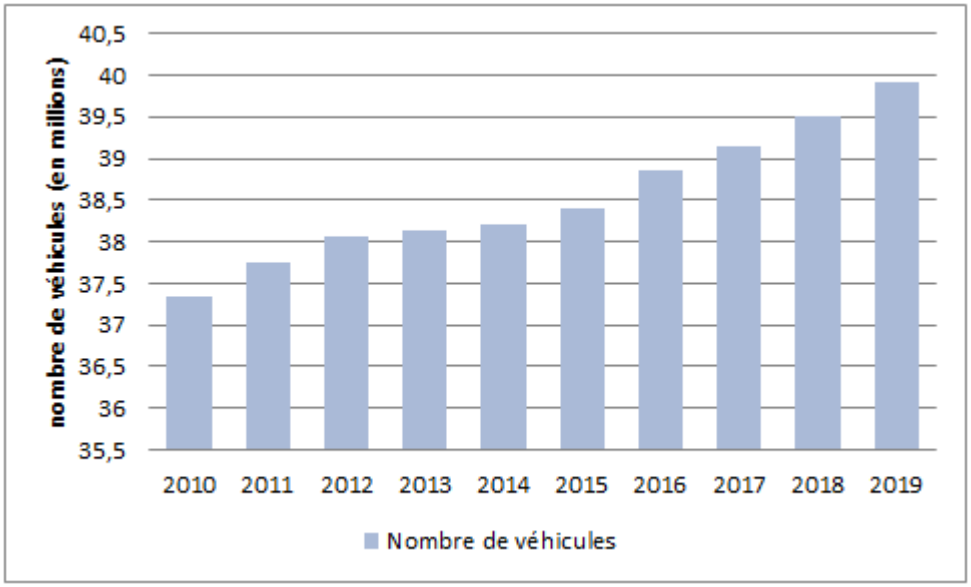


FIGURE 1.1 – Evolution du parc automobile en France entre 2010 et 2019

Le nombre de risques assurables évolue peu, tandis que les intervenants sur le marché se multiplient ; ce qui le rend extrêmement concurrentiel. De plus, la réglementation est en constante évolution. Elle oblige les acteurs à se réinventer. En l'occurrence en 2012, intervient la Gender Directive qui instaure le principe d'égalité entre hommes et femmes dans l'accès aux services. Les assureurs n'ont plus le droit de différencier le tarif des produits en fonction du sexe du prospect. Ce qui constitue pour eux un nouveau défi. De plus, depuis janvier 2015, la loi Hamon, autorise les assurés ayant au moins un an de contrat, à le résilier à la date de leur choix, sans préjudice financier. Pour certains, cette loi est source d'inquiétude en matière de résiliations, quand d'autres y voient l'opportunité de faire croitre leurs parts de marché. En ce sens, le cadre réglementaire nourrit le dynamisme du marché.

De nombreuses stratégies sont alors mises en place non seulement pour fidéliser les clients déjà en portefeuille mais aussi pour attirer de nouveaux clients. Il peut s'agir de promotions pour attirer de nouveaux clients, ou alors des réductions de tarifs pour les clients acquis. Les nombreux comparateurs de prix existants donnent aux clients accès aux tarifs proposés sur le marché, et ils s'en servent pour faire jouer la concurrence. Au sein du marché, les acteurs jouent énormément sur les tarifs et les avantages proposés afin de gonfler leur portefeuille.

Cette situation de marché ouvert et fortement concurrentiel crée pour les acteurs un nouveau risque. Ils doivent pouvoir proposer des tarifs en accord avec leurs objectifs de rentabilité, tout en étant compétitifs. C'est dans ce contexte que naît la problémtique traitée dans ce mémoire. Nos travaux s'inscrivent dans le cadre de la mise à jour tarifaire du produit d'assurance automobile d'Allianz France. L'objectif est de construire un modèle de conversion, qui servira d'appui à la construction du modèle tarifaire.

1.2 L'offre d'assurance automobile d'Allianz France

L'offre d'assurance automobile Allianz est composée de 5 formules (CECO, C1, C1+, C2 et C3), qui correspondent à différents niveaux de couverture. Les formules CECO et C1 sont assimilables à une couverture « au tiers », et la formule C3 à une couverture « tous risques ». les formules C1+ et C2 sont des niveaux de couverture intermédiaires, offrant plus de sécurité que les formules « au tiers » mais moins que la formule « tous risques ».

Sur l'ensemble des formules, l'assuré est couvert pour la responsabilité civile, les défenses civile et pénale suite à un accident et dispose également d'une garantie conducteur. Cette garantie est très importante car elle protège le conducteur en cas d'accident, même s'il en est responsable. Il faut savoir que, parmi les victimes de la route, le cadre réglementaire n'a pas prévu de protection pour le conducteur responsable de l'accident. Il est de sa responsabilité de souscrire une assurance pour couvrir les conséquences de l'accident. Cette garantie peut couvrir les frais médicaux, chirurgicaux ou d'hospitalisation, mais également le préjudice financier lié à un arrêt de travail ou une incapacité permanente.

Les formules CECO et C1, offrent un socle minimal de couverture. Contrairement aux autres formules, elles ne prennent pas en charge le bris de glace, les catastrophes naturelles ou technologiques, le vol et l'incendie. Pour aller plus loin en termes de couverture, toutes les formules proposent des packs et d'autres garanties en option. Il existe trois types de packs à savoir :

- Les Packs Mobilité Allianz : ils permettent de bénéficier d'un dépannage à 0 Km et d'un véhicule de remplacement. Ils sont disponibles sur toutes les formules, sauf la CECO.
- Le Pack Réparation Allianz : l'intérêt de ce pack est que l'assureur prend en charge la réparation en cas de panne. Lorsqu'un assuré souscrit le Pack Réparation, il est conseillé de lui proposer la souscription conjointe des Packs Mobilité, pour qu'il puisse bénéficier d'un dépannage 0km. Le pack réparation n'est disponible que pour la formule C3.

Concernant les options, on en compte cinq :

- L'option contenu : elle est accessible en formule C3 et en C2. Elle permet de couvrir les effets personnels, les équipements de loisirs (ex : bicyclette, planche à voile), le matériel professionnel (ex : outillage d'artisans)... Cette garantie entre en jeu dans des cas spécifiques, par exemple si les objets ont été volés ou endommagés en même temps que le véhicule.
- L'option équipements : elle permet de garantir tout ce qui a été ajouté au véhicule aux fins d'enjolivement, d'aménagement fonctionnel ou d'amélioration de confort.

Elle permet également de garantir les aménagements professionnels des véhicules utilitaires de moins de 3,5 tonnes.

- La protection juridique automobile : elle est utile aussi bien en prévention qu'en présence de litige. Elle bénéficie au souscripteur du contrat, au propriétaire du véhicule ou à toute personne autorisée à le conduire.
- Le dépannage 0km : cette option supprime la franchise de 25 km en cas de panne. Elle est ouverte à toutes les formules sauf la CECO.
- L'option Allianz conduite connectée : elle permet d'accéder aux indicateurs de comportement de conduite et d'usage du véhicule. Elle donne droit à une tarification à l'usage et personnalisée.

Ci-dessous, vous trouverez la structure globale de l'assurance automobile Allianz :

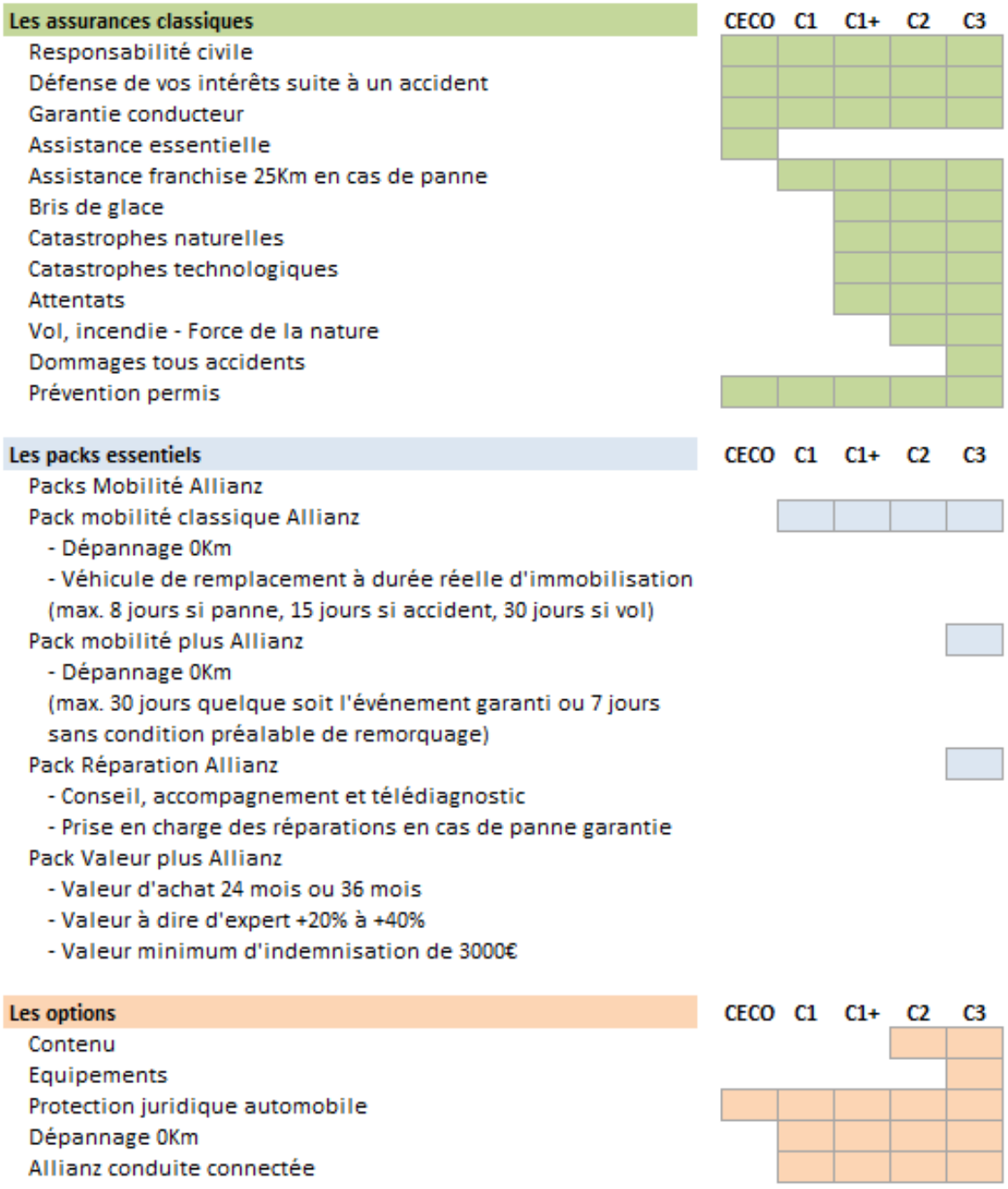


FIGURE 1.2 – Structure de l'offre d'Allianz France

1.3 Enjeux de l'étude pour Allianz

1.3.1 Mettre à jour la grille tarifaire

Dans le cadre des missions de l'équipe tarification, d'importants travaux sont menés afin de proposer aux prospects la cotisation la plus juste. Il s'agit de celle qui reflète exactement le coût que celui-ci représenterait pour l'entreprise : elle est appelée la prime pure. Elle est définie par :

$$\text{Prime pure} = E(X)$$

Où X est une variable aléatoire représentant la somme des coûts des sinistres de l'assuré sur la période de couverture.
Cette prime est assez souvent déterminée en utilisant deux modèles linéaires généralisés. Le premier captant la fréquence des sinistres et le second, le coût-moyen. Par conséquent, on peut écrire X de la manière suivante :

$$X = \sum_{i=1}^{N} W_i$$

Où :

- N est la variable aléatoire à valeur entière représentant le nombre total de sinistres ;
- W_i est le coût du i-ème sinistre. Les W_i sont des variables aléatoires positives, indépendantes et identiquement distribuées, toutes indépendantes de N.

On en déduit que

$$\text{Prime pure} = E(X) = E(N) * E(W)$$

Donc en multipliant les résultats du modèle de fréquence et ceux du modèle de sévérité, on obtient notre prime pure. En pratique, cette modélisation est faite pour chaque garantie puis agrégée afin de déterminer la prime pure globale. Appelons grille tarifaire, le tableau sur lequel on retrouve l'ensemble des coefficients d'un modèle de prime pure. Ci-dessous, un extrait de la grille tarifaire pour la garantie dommage.

GRILLE TARIFAIRE DOMMAGE	
Base	207,45
Statut marital	
Célibataire	0,96
Divorcé	1,20
Marié / Pacsé / Union libre	1,00
Veuf	9,98
Non Renseigné	1,00
Forfait kilométrique	
Moins de 4000 Km	0,69
Moins de 7000 Km	0,80
Moins de 9000 Km	0,98
Illimité	1,00
Usage	
Au repos	0,05
Déplacements privés	1,00
Déplacements privés et professionnels	1,00
Tous déplacements	1,46
Auto-école	0,99
Taxi	1,30
VTC conducteus exclusif	1,97
VTC multi-conducteur	2,44
Ambulance	1,49
Transport public de marchandises	1,88
Transport funéraire	1,35

FIGURE 1.3 – Exemple de grilles tarifaires pour la garantie dommage

Remarquons que le risque que représente un individu varie dans le temps. En effet, lorsqu'un conducteur gagne de l'expérience, il a tendance à causer moins d'accidents. Par contre, la puissance du véhicule a plutôt tendance à faire augmenter la prime. Tout ceci implique que l'on doive faire évoluer les tarifs au cours du temps. Ce qui entrainera des changements tout aussi bien à la hausse comme à la baisse des tarifs proposés. D'autre part, les tarifs doivent tenir compte des objectifs de rentabilité de l'entreprise et des interactions avec le marché. En effet, si l'assureur propose des tarifs trop élevés par rapport aux concurrents, il expose son produit au rejet de potentiels clients. A l'inverse, s'ils sont trop bas, il met en péril sa rentabilité technique : ce qui signifie qu'il n'aura pas réuni suffisamment de ressources pour faire face à ses engagements. Le problème qui se pose réside donc sur l'arbitrage nécessaire lors de la mise à jour des tarifs. Au sein de l'équipe tarification auto particulier, l'un des outils utilisés à cet effet, est le **Scenario Testing New Business**. Son fonctionnement sera présenté dans la section suivante.

1.3.2 Le Scenario Testing

Le Scenario Testing, conçu sous le logiciel Radar [1], est un projet dans lequel sont implémentés différents scenarii et servant à évaluer les indicateurs de rentabilité pour chacun d'eux. Les indicateurs calculés orientent la prise de décision lors de la mise à jour des tarifs. Dans ce mémoire, ne seront pas exposées les phases d'analyse de ces différents indicateurs.

Le Scenario Testing prend en entrée un ensemble de devis recueillis sur une période donnée, et détermine la rentabilité de ces affaires compte tenu des objectifs fixés par la compagnie. Il assure la projection des indicateurs suite aux variations tarifaires.

Ce projet entre dans le cadre de la certification du produit auto 4RP d'Allianz France. En effet, un audit interne est réalisé par une entité du Groupe Allianz appelée Global P&C, afin de certifier les produits d'Allianz France. Cette entité met en place des lignes directrices à chaque étape du processus de tarification afin d'aiguiller les équipes dans leurs objectifs de certification des tarifs. De ce fait, le processus de mise à jour tarifaire intègre des recommandations formulées par Global P&C. Le sujet de ce mémoire est la construction d'un modèle de conversion, pièce essentielle du Scenario Testing dont le schéma est présenté ci-dessous :

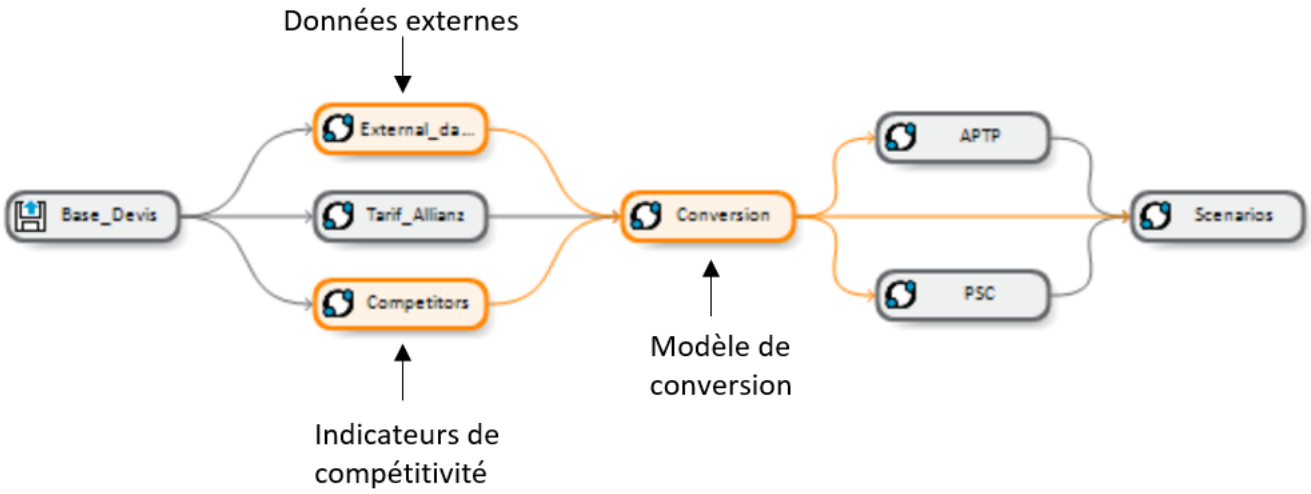


FIGURE 1.4 – Scenario Testing

Les différents éléments constituant le Scenario Testing seront tour à tour présentés dans la suite de ce mémoire.

1. Développé par le cabinet Willis Tower Watson.

1.4 Les indicateurs de rentabilité

Dans cette section, nous voulons présenter comment se fait le calcul des indicateurs du Scenario Testing. Pour ce faire, nous décrirons d'abord les paramètres utilisés pour les calculer. Puis, nous montrerons comment sont calculés ces indicateurs.

1.4.1 Eléments permettant de construire les indicateurs de rentabilité

Prime pure et Expected Ultimate Loss

Le premier élément qui entre en jeu dans le calcul des indicateurs de rentabilité est la prime pure. Pour un profil donné, elle est estimée à partir de l'expérience de sinistres de notre portefeuille. Cependant, cette quantité n'est pas suffisante pour donner une mesure réelle des coûts de sinistres. En effet, les coûts des sinistres peuvent évoluer dans le temps. De plus, l'estimation de la prime pure ne prend pas en compte les sinistres survenus mais non déclarés. Il convient alors de déterminer :

- les coûts des sinistres survenus mais déclarés tardivement : les IBNR ;
- les coûts des sinistres survenus mais qui ne sont pas assez provisionnés : les IBNER.

Ces informations nous sont utiles pour déterminer des coefficients de passage à l'ultime. Nous calculons ensuite l'EUL (Expected Ultimate Loss), qui est le Best Estimate (en français meilleure estimation) des coûts de sinistres futurs d'un contrat sur la période de couverture.

Dans le Scenario Testing, pour chaque garantie, on obtient l'EUL appliquant à la prime pure, un coefficient de passage à l'ultime :

$$EUL = \text{prime pure} * \text{coefficient de passage à l' ultime}$$

EUL et Technical Price

Le Technical Price, en abrégé TP, intègre des frais au best estimate des coûts des sinistres futurs . Il prend en compte l'EUL et y ajoute non seulement des frais d'acquisition, mais aussi des frais généraux et des frais de gestion de l'assureur.

$$TP = \text{EUL} + \text{autres frais non liés aux sinistres}$$

Technical Price et Commercial Price (CP)

Après avoir calculé le TP, l'assureur prend en compte ses objectifs de rentabilité, afin de déterminer le tarif à proposer au client. Ce dernier est appelé tarif commercial (Commercial Price). Pour l'obtenir, on effectue des ajustements sur les coefficients de la grille tarifaire.

Commercial Price et Actual Price

Dans les faits, le client ne paie pas toujours le tarif commercial. Il peut bénéficier d'offres promotionnelles et/ou d'une réduction de tarif liée à son profil. En effet, il peut arriver que les directives commerciales soient d'attirer des profils ciblés ou dans un cadre plus général, d'offrir une réduction à tout nouveau client (deux mois d'assurance gratuits par exemple). Ces avantages tarifaires sont appliqués au tarif commercial pour donner ce qu'on appelle l'Actual Price (en abrégé AP), qui est la prime réellement payée par le client.

1.4.2 Mesure de la rentabilité

Le Technical Price est utilisé pour orienter l'évolution de l'Actual Price. Les indicateurs influençant la prise de décision sont les suivants :

- **Le rapport AP/TP**

Si le ratio AP/TP est inférieur à 1, on conclut qu'il y a sous–tarification du produit. Le tarif proposé correspond à une sous-estimation de risque.

Si ce rapport est supérieur à 1, nous sommes dans une situation de sur-tarification du produit. Cette situation nuit à la compétitivité de l'entreprise ; ce qui aura un impact à la baisse sur les parts de marché de l'assureur.

S'il est de 1, alors la prime payée concourt à l'atteinte des objectifs de rentabilité. Dans ce cas, il faut éviter de dévier fortement du technical price, pour éviter de se retrouver dans une situation de sur-tarification ou de sous-tarification.

Toutefois, notons que les décisions à prendre en se basant sur le rapport AP/TP présentent quelques subtilités. L'assurance automobile est considérée comme un point d'approche. C'est-à-dire qu'à partir du moment où le prospect a concrétisé son devis d'assurance automobile, on peut lui proposer d'autres produits comme l'assurance habitation par exemple. On se dit qu'à partir du moment où il est client, il devient plus facile de lui conseiller de souscrire d'autres contrats. Pour attirer les prospects, nous pouvons leur proposer des tarifs en deçà du risque estimé qu'ils représentent. En conséquence, pour les contrats d'assurance automobile vieux de moins d'un an, on observe un AP/TP inférieur à

1. Avec le temps, la surveillance du portefeuille oriente les revalorisations tarifaires. L'objectif étant de rendre le portefeuille rentable. Cette surveillance nous permet alors d'identifier les contrats atypiques, auxquels on applique des hausses tarifaires assez fortes.

- **Le Projected S/C (PSC)**

Le PSC est défini comme le rapport de l'EUL et l'actual price.

$$PSC = \frac{EUL}{AP}$$

C'est un indicateur utilisé pour le suivi de la rentabilité technique du portefeuille. Il s'agit d'une estimation de l'évolution du ratio de sinistralité (le S/C). Le PSC d'un contrat donne une vision de de l'espérance du profit ou de la perte engendré(e) par celui-ci. Il évolue en sens inverse de la rentabilité. S'il est grand alors, la rentabilité est faible.

Par rapport au S/C observé, qui a du sens à la maile d'un portefeuille, le PSC est robuste lorsqu'on l'observe police par police. Il a également un caractère prospectif, contrairement au S/C qui se base sur une sinistralité observée.

1.5 Le modèle de conversion

Le modèle de conversion est une composante du Scenario Testing. Il a pour but de déterminer la probabilité qu'un prospect convertisse son devis en contrat. Dans le cadre du Scenario Testing, les résultats du modèle sont utilisés pour calculer les indicateurs de rentabilité.

Les informations fournies par le prospect lors de la souscription ainsi que les données externes font partie des variables explicatives de ce modèle. Afin d'illustrer l'influence des autres acteurs du marché, nous intègrerons aussi des indicateurs de compétitivité comme variables explicatives. L'objectif est dans un premier temps de déterminer quels sont les facteurs qui influencent le plus l'acte de souscription et ensuite, d'évaluer l'apport des indicateurs de compétitivité et des données externes dans un tel modèle. Nous utiliserons trois méthodes pour construire notre modèle de conversion : la régression logistique, le random forest et le gradient boosting. Les résultats du modèle le plus performant seront intégrés dans le Scenario Testing. Vous trouverez ci-dessous les étapes de la construction d'un modèle de conversion.

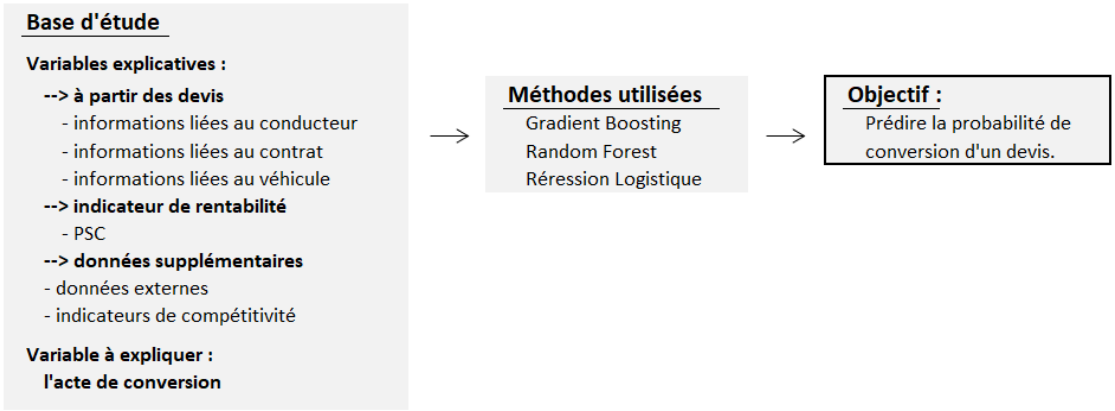


FIGURE 1.5 – Etapes du modèle de conversion

Dans le domaine de l'actuariat en France, la modélisation de l'acte de conversion n'est pas un sujet nouveau. Il a déjà été abordé par plusieurs actuaires et sur différents marchés. Par exemple, nous pouvons citer les travaux réalisés en 2019 par Théophile ROBERT qui portent sur la construction d'un modèle de conversion en multirisque habitation ou encore ceux réalisés en 2016 par Alice CHARGERAUD, portant sur la comparaison de diverses méthodes d'apprentissage pour la modélisation du taux de transformation en assurance automobile. Tout comme ces deux auteurs, nous abordons le sujet en utilisant diverses méthodes de modélisation. L'objectif final étant de déterminer la meilleure. Toutefois, nos travaux se démarquent de ceux de nos prédécesseurs par l'introduction de nouvelles sources de données. Celles-ci nous permettent de prendre en compte l'effet de la répartition de notre

réseau de distribution comme variable explicative. Nous donnerons plus de détails sur leur construction dans le chapitre 2 de ce mémoire.

Chapitre 2

Description des données

Ce chapitre est consacré à la présentation de la base d'étude. Premièrement, nous évoquerons le périmètre considéré ainsi que le détail des données. Puis, nous mettrons un accent sur la construction des indicateurs de compétitivité et des données externes. Et enfin, nous présenterons des résultats d'analyses exploratoires sur la base d'étude.

2.1 Construction de la base d'étude

La base d'étude est principalement constituée de données issues des devis. Il s'agit de devis émis auprès d'agents généraux Allianz ou de courtiers. Le périmètre de la base devis comprend des véhicules à moteur de quatre roues, appartenant à des particuliers mais aussi des véhicules d'auto-école, des ambulances, des taxis et des véhicules servant pour le transport funéraire. La période d'observation considérée va du 01/04/2019 au 31/03/2020. Cette restriction sur un an est due au fait qu'au mois d'avril, la principale mise à jour des tarifs est effectuée et vaut pour un an. Afin de ne pas biaiser les données, nous avons décidé de ne pas prendre en compte le mois de mars 2020 car les agences étaient fermées en raison de la crise de COVID-19. Chaque ligne de cette table correspond aux informations fournies par le prospect lors de la réalisation du devis ainsi que des informations contrats. On peut citer :

- les informations liées au conducteur : âge, ancienneté de permis, activité professionnelle...
- les informations liées au véhicule : date de mise en circulation, marque, modèle...
- les informations liées au contrat : garanties souscrites, fractionnement, prime TTC...

Cette base devis est ensuite enrichie de données externes et d'indicateurs de compétitivité. Ces sources d'informations seront détaillées dans les sections suivantes. C'est à l'aide de l'ensemble de ces variables que nous construisons notre base d'étude.

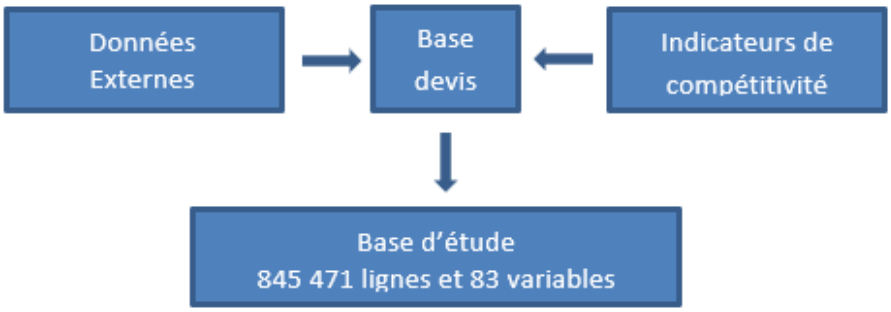


FIGURE 2.1 – Schéma de construction de la base d'étude

En résumé, la base d'étude compte 845 471 observations et 83 variables, y compris la variable à expliquer. Vous retrouverez en annexe la liste exhaustive de l'ensemble des variables de la base des devis. Les données externes et les indicateurs de compétitivité seront présentés dans la suite du mémoire.

La base d'étude est utilisée pour réaliser l'ensemble des modèles à savoir la régression logistique, les forêts aléatoires (random forest) et le gradient boosting.

Dans le cas de la régression logistique, les travaux ont été réalisés sous le logiciel EMBLEM. Il s'agit d'un logiciel de modélisation du cabinet Willis Towers Watson. Il a été choisi car il permet d'implémenter assez facilement des modèles linéaires généralisés. Implémenter un modèle sous EMBLEM nécessite de créer au préalable 2 fichiers ayant des formats particuliers :

- un fichier Bid (Binary Data File) : c'est l'équivalent de la base de données utilisée pour construire le modèle. Il renseigne sur les valeurs prises par les variables explicatives, les poids des observations et les valeurs de la variable cible ;
- un fichier Fac (Factor) : il est assimilable à un dictionnaire de données. Il donne la liste des variables, le nombre de modalités ainsi que les modalités de référence [2].

L'une des contraintes sous EMBLEM est que chaque variable doit être sous la forme qualitative et limitée à un maximum de 255 modalités. Nous avons donc fait les retraitements qui s'imposaient sur les données ainsi que la création des fichiers Fac et Bid en utilisant SAS Entreprise Guide. Pour les modèles machine learning, nous avons travaillé sous Python (version 3.6). Construits sous Python, les modèles machine learning ne sont pas soumis aux mêmes contraintes de formatage de données que le modèle de régression logistique. Dans la suite du mémoire, nous expliciterons la théorie derrière ces modèles.

2. Par défaut, sur EMBLEM, la modalité avec le plus grand effectif est la modalité de référence.

2.2 Construction des indicateurs de compétitivité

Pour calculer les indicateurs de compétitivité, nous nous sommes basés sur un marché comprenant quatre acteurs : Allianz et 3 concurrents. Nous les appellerons concurrent 1, concurrent 2 et concurrent 3. Afin d'estimer leurs tarifs, nous avons préalablement défini une liste de profils. Sur la base de cette liste, nous confions la tâche à un cabinet prestataire de déterminer la prime que proposerait chaque concurrent. L'idée est de pouvoir confronter les tarifs des différents acteurs du marché. Cependant, les formats des formulaires devis varient d'un assureur à l'autre. De ce fait, les questions posées ainsi que les réponses possibles ne sont pas les mêmes. Or, nous devons comparer les tarifs sur des critères qui sont propres à Allianz. Nous avons donc dû formuler des hypothèses qui nous permettent d'avoir un cadre universel pour pouvoir comparer les tarifs. Par exemple :

- Les concurrents ne proposent pas tous autant de formules qu'Allianz. Lorsqu'on n'a pas de tarif pour un concurrent car il ne propose pas la même formule qu'Allianz, on remplace la valeur manquante par une moyenne des primes des acteurs qui proposent cette formule ;
- Pour une variable tarifaire donnée, les possibilités proposées par les concurrents ne sont pas identiques à celles d'Allianz. Considérons à titre illustratif la variable marque du véhicule. Supposons qu'une marque proposée par Allianz ne figure pas dans la liste de celles proposées par un concurrent. Dans ce cas, nous identifions une marque qui produit des véhicules de la même gamme ou puissance, et on remplace la valeur manquante par une approximation faite sur cette marque.

Ainsi, pour chaque ligne de notre base devis, un tarif concurrent est calculé. Pour les concurrents 2 et 3, nous représentons ci-dessous le ratio tarif Allianz sur tarif Concurrent.

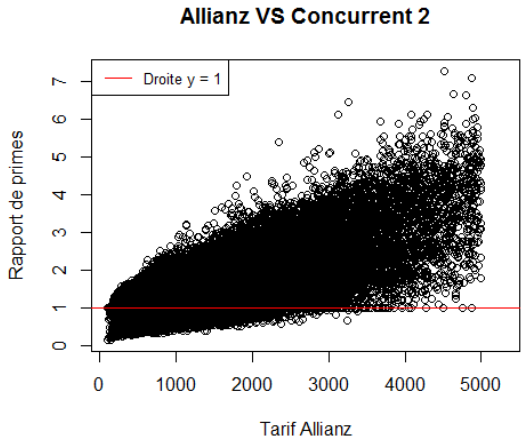


FIGURE 2.2 – Comparaison des primes Allianz et celles du concurrent 2

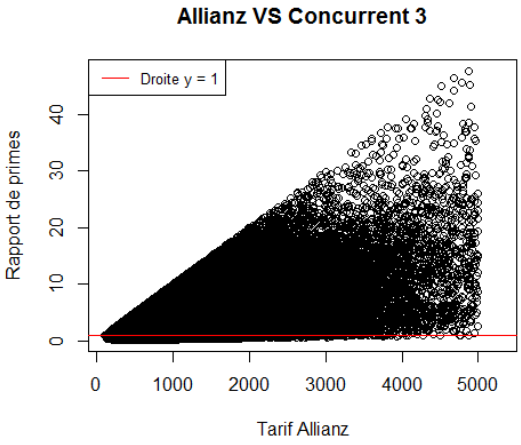


FIGURE 2.3 – Comparaison des primes Allianz et celles du concurrent 3

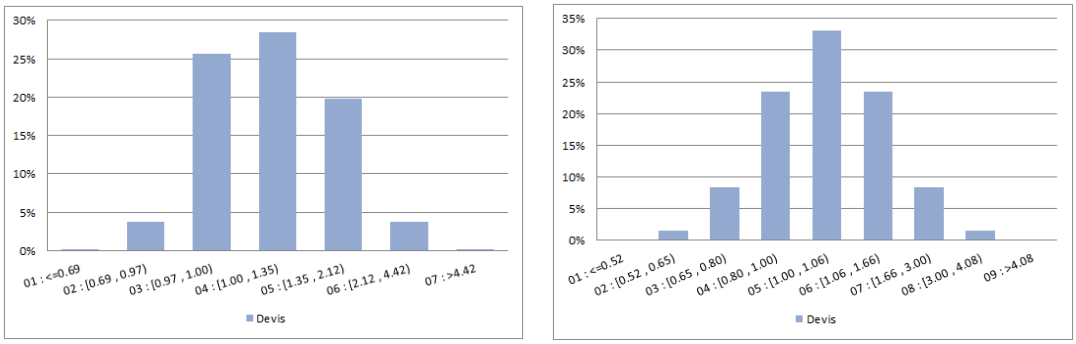
Qu'il s'agisse de la figure 2.2 à gauche ou encore de la figure 2.3 à droite, on constate que les points représentés sont majoritairement au-dessus de l'axe y=1. Sur cette droite, on retrouve l'ensemble des devis pour lesquels Allianz et le concurrent proposent le même tarif. Le fait de retrouver une forte proportion de points au-dessus de l'axe est assez caractéristique de notre marché, où le tarif Allianz est en moyenne plus élevé que celui des concurrents pris individuellement. Cette différence est encore plus nette lorsqu'on compare le tarif Allianz à celui du concurrent 3. En effet, la figure 2.3 nous montre qu'on a une forte densité de points au-dessus de l'axe y=10. Ce qui signifie, que pour un même profil, Allianz peut proposer un tarif 10 voire 40 fois plus élevé que le concurrent 3. Au vue des graphiques, il semblerait intéressant de discuter des hypothèses formulées sur les tarifs des concurrents mais, ce point n'est pas abordé dans ce mémoire. Bien que l'écart entre le tarif Allianz et celui du concurrent 3 nous semble peu cohérent, il traduit tout de même une réalité : pour la plupart des profils, le tarif Allianz est plus élevé. Nous utilisons les tarifs de l'ensemble des concurrents pour calculer nos indicateurs de compétitivité.

Les indicateurs de compétitivité renseignent sur le positionnement tarifaire d'Allianz par rapport au marché. Ayant notre marché et les tarifs des différents concurrents, on détermine :

- La prime moyenne du marché hors et y compris Allianz, respectivement P^{MWOA} et P^{MWA}
- Les extrema : la prime la plus élevée P^{Max} et la prime la moins élevée P^{Min} ;
- La deuxième la plus élevée du marché P^{Max2} et la deuxième prime la moins élevée P^{Min2}.

Après avoir déterminé ces 6 valeurs, nous évaluons comment elles se situent par rapport à la prime Allianz de deux manières :

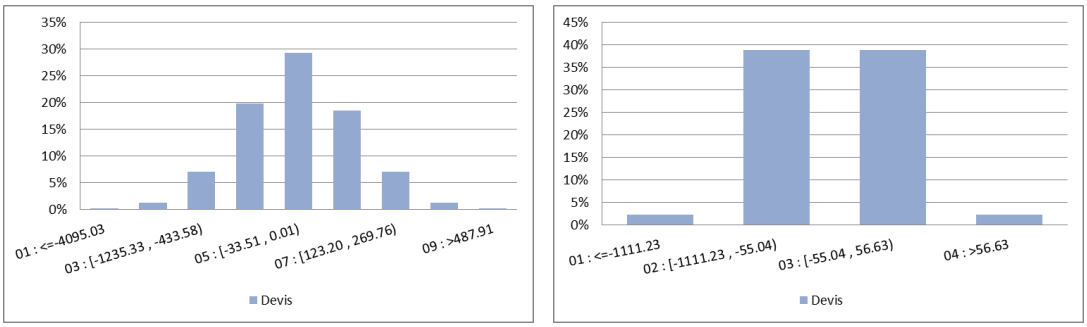
- En calculant le rapport : $\frac{p^i}{\text{Prime Allianz}}$
  Avec $i \in \{MWOA, MWA, Max, Min, Max2, Min2\}$
  Ensuite, nous étudions les distributions des résultats afin de les regrouper sous forme de classes. Suivant la distribution observée, nous faisons des regroupements selon une loi uniforme et/ou selon une loi normale comme le montrent les graphes suivants :



(a) par rapport à la moins élevée du marché (b) par rapport à la plus élevée du marché

FIGURE 2.4 – Distribution du rapport de la prime Allianz et des extrema du marché

- En faisant la différence : $P^i -$ prime allianz
  Avec $i \in \{MWOA, MWA, Max, Min, Max2, Min2\}$
  Comme précédemment, nous étudions les distributions des résultats puis nous créons des classes en fonction de celles-ci.



(a) par rapport à la moins élevée du marché (b) par rapport à la plus élevée du marché

FIGURE 2.5 – Distribution de la différence entre la prime Allianz et les extrema du marché

Nous sommes également tenus de faire ces regroupements car pour les modèles de régression logistique, nous utilisons le logiciel EMBLEM, qui n'admet pas de variable explicative continue.

En plus des indicateurs de compétitivité, nous créerons de nouvelles variables à l'aide des données externes. L'un des intérêts de l'étude est d'estimer l'apport de l'ensemble de ces variables dans la modélisation du taux de conversion. Dans la section suivante, nous présenterons les données externes dont nous disposons, et comment elles sont utilisées dans le cadre de ce mémoire.

2.3 Les données externes

Rappelons que notre problématique est de construire un modèle qui renseigne sur la probabilité qu'un prospect convertisse un devis. D'autre part, le réseau des agents fait partie du périmètre considéré. Nous avons alors décidé de nous intéresser à notre réseau de distribution mais également à celui de nos concurrents 1, 2 et 3. A cet effet, nous nous sommes rapprochés de la Direction Stratégie Commerciale, qui a nous a fourni :

- la répartition géographique de nos points de vente et de ceux des concurrents 1, 2 et 3 en France métropolitaine ;
- le **potentiel** des communes en France métropolitaine.

Avant de détailler ces deux points, nous présentons d'abord la notion de maille géographique, essentielle pour comprendre la granularité des informations traitées.

2.3.1 L'IRIS

L'acronyme **IRIS** désigne des **I**lots **R**egroupés pour l'**I**nformation **S**tatistique. Les IRIS sont issus du découpage des communes. En effet, la plupart des communes du plus de 5000 habitants sont découpées en "quartiers", dont la population est de l'ordre de 2000 habitants. Ce découpage sert d'échelle pour la diffusion des statistiques infracommunales fournies par l'Institut National de la Statistique et des Etudes Economiques (INSEE). Aussi, les communes non découpées en IRIS sont assimilées à IRIS. Ce qui permet de couvrir l'ensemble du pays. En 2019, la France comptait 48 590 IRIS répartis comme suit :

- 15 554 IRIS issus du découpage des communes ;
- 33 036 IRIS issus des communes non découpées.

La finesse du découpage du térritoire en IRIS, confère une grande précision aux données externes que nous utilisons. Elles peuvent ensuite être diffusées à différentes échelles. Vous trouverez en annexe les différentes mailles possibles.

2.3.2 La répartition des points de vente

Depuis le début des années 2000, l'**O**bservatoire **G**éographique des **R**éseaux d'**A**ssurances (en abrégé **OGRA**) fait l'inventaire des points de vente d'assurance en France métropolitaine et dans les Départements d'Outre-Mer. Il fournit une base de données utilisée par les assureurs dans l'analyse du maillage des réseaux de distribution de tous les acteurs du

marché. Elle permet entre autres de prendre en compte les implantations des points de vente des concurrents dans l'ajustement géographique de notre réseau de distribution. Il faut noter que ces données ne sont pas disponibles en accès libre. C'est par l'intermédiaire d'ESRI France, cabinet expert en traitement de données géographiques, que la Direction Stratégie Commerciale y a accès.

Pour chaque concurrent, nous déterminons le nombre d'agences dans les différentes communes de la métropole. Nous précisons que les communes sont issues du découpage administratif de 2017 et le recensement des points de vente date de 2018. A ces 3 variables nous ajoutons un comptage réalisé en 2020, des points de vente Allianz. Les quatre variables ainsi construites seront intégrées dans les modèles **avec indicateurs de compétitivité**.

2.3.3 La notion de potentiel

L'une des missions de la Direction Stratégie Commerciale, est d'optimiser la répartition géographique des points de vente Allianz sur le territoire français. Pour ce faire, l'un des éléments pris en compte est le **potentiel** de la zone géographique dans laquelle est implanté le point de vente. Le **potentiel** d'une zone géographique est le reflet de la somme des primes d'assurance payées par la masse assurable qui y est recensée. La méthode de calcul du potentiel est fonction du marché auquel on s'intéresse. Par exemple, le potentiel MRH[3] d'une localité est une estimation de la sommme des primes d'assurance payées par les ménages dans cette localité. L'étude des localités offrant le meilleur potentiel permet entre autres de :

- préconiser la délocalisation de points de vente peu performants ;
- préconiser l'ouverture de nouveaux points de vente.

Pour faire cette étude, on calcule le potentiel de différentes zones géographiques à une maille donnée. En fonction des valeurs obtenues, on définit des seuils B (bas) et H (haut) tels que $B < H$. Pour toutes les zones où le potentiel est inférieur à B, on considère qu'il est faible. Ces zones ne présentent pas un réel intérêt pour nous. Les localités ayant un potentiel compris entre B et H ont un potentiel dit modéré. On peut envisager d'y implanter un ou plusieurs points de vente, s'il n'en existe pas encore. Les zones géographiques les plus attracives sont celles qui ont potentiel supérieur à H. On utilise ensuite des fonds de carte comme celui présenté sur la figure suivante, pour visualiser ces différentes zones.

3. MRH : Multirisque habitation.

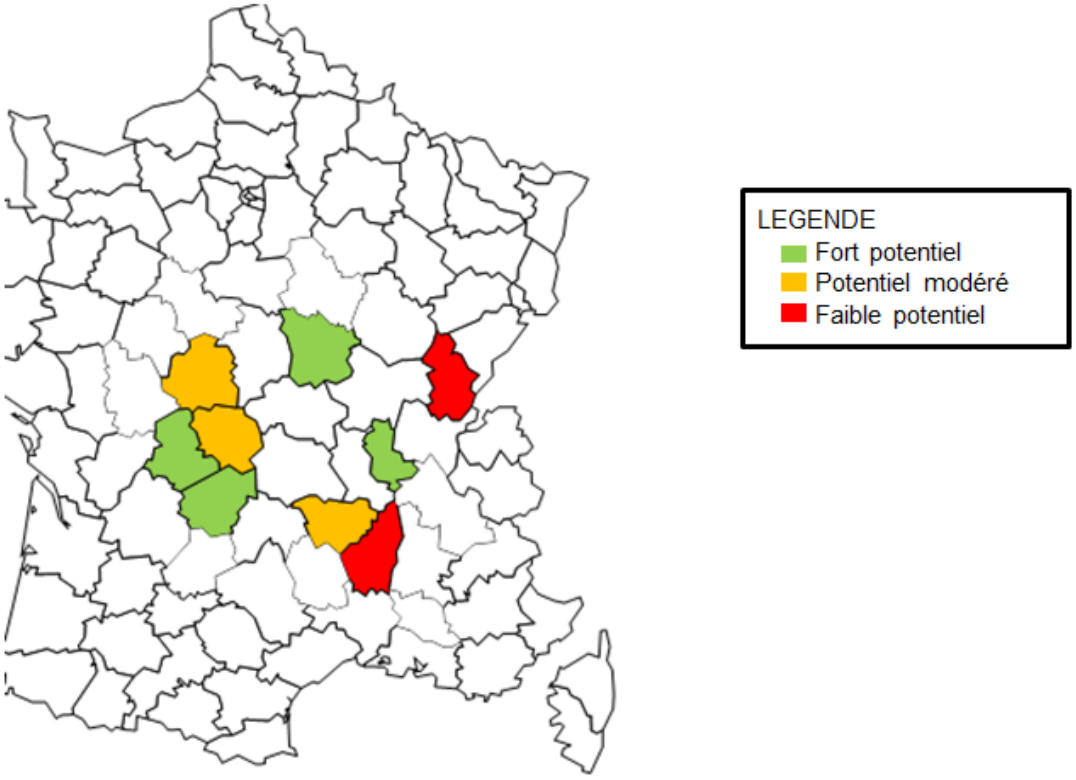


FIGURE 2.6 – Illustration du potentiel d'une localité

Pour le calcul des potentiels, au sein de la Direction Stratégie Commerciale, la maille géographique utilisée est l'**IRIS**. Cependant, dans notre base devis, l'information retenue pour le lieu de garage habituel du véhicule assuré est à l'échelle de la commune. Nous avons alors décidé d'agréger les potentiels à l'échelle des communes. Pour le besoin de notre étude, nous avons calculé les potentiels sur différents marchés à savoir :

- l'assurance automobile, l'assurance multirisque habitation et l'assurance santé : pour ces marchés, les potentiels sont issus des enquêtes INSEE sur les consommations des ménages en produits d'assurance.
- l'assurance des professionnels/TPE[4] : ces potentiels sont calculés à partir des données de prospects valorisées par les primes moyennes Allianz.

Nous retiendrons les potentiels en assurance automobile et sur le retail. Le potentiel retail étant une agrégation des potentiels des marchés cités précédemment (y compris l'automobile). Tout comme le comptage des points de vente, ces deux variables ne seront incluses que dans les modèles **avec indicateurs de compétitivité**.

4. Les TPE considérées sont des entreprises de moins de 10 salariés.

2.4 Statistiques descriptives

Dans cette partie, nous présenterons des analyses descriptives réalisées sur quelques variables de la base d'étude, afin de donner un aperçu de notre jeu de données.

- **Taux de conversion :**

Sur les 845 471 devis présents dans notre base, 330 415 ont été convertis ; ce qui nous donne un taux de conversion de 39,08%. Sur la période d'observation, c'est-à-dire d'avril 2019 à février 2020, nous présentons ci-dessous l'évolution du volume de devis et du taux de transformation.

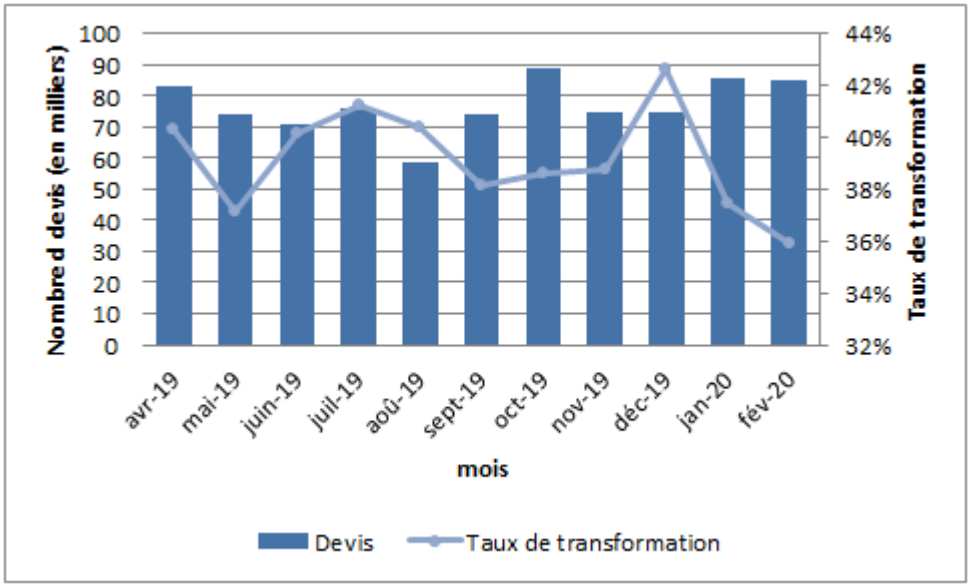


FIGURE 2.7 – Evolution du taux de conversion par mois

- **Age du conducteur :**

Les devis ont été réalisés sur une population dont l'âge varie entre 18 et 80 ans. Le nombre de devis réalisés baisse légèrement avec l'âge. On enregistre plus de devis auprès de conducteurs moins âgés, bien que les proportions soient à peu près stables entre 19 et 55 ans. Par contre, en observant l'évolution du taux de conversion, on constate que les valeurs les plus faibles sont enregistrées sur des conducteurs d'au plus 22 ans. Pour les conducteurs dont l'âge varie entre 18 et 22 ans, on observe un taux de conversion moyen de 31,65%. Ensuite ce taux évolue progressivement jusqu'à se stabiliser autour de 40% entre 27 et 60 ans. Puis il baisse continuellement entre 60 et 71 ans avant de remonter et d'atteindre son pic 44,69% sur la tranche 80 ans et plus.

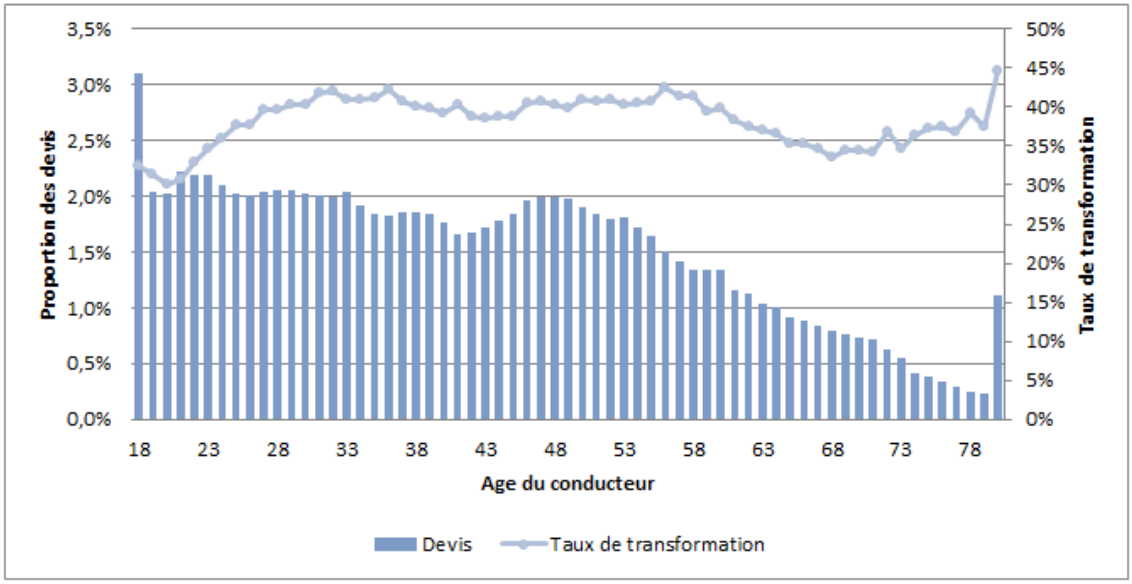


FIGURE 2.8 – Evolution du taux de conversion en fonction de l'âge du conducteur principal

- **Ancienneté de permis :**

En ce qui concerne l'ancienneté de permis, les modalités varient entre 0 et 60 ans. La répartition du nombre de devis en fonction de l'ancienneté de permis est proche de celle décrite pour l'âge du conducteur. A partir de 3 ans d'ancienneté, le nombre de devis baisse progressivement jusqu'à 59 ans. En ce qui concerne le taux de transformation, on remarque qu'il croit progressivement jusqu'à se stabiliser autour de 42% entre 6 et 40 ans d'ancienneté de permis. Ensuite, il baisse et se situe entre 32% et 37% pour des conducteurs ayant entre 41 et 59 ans d'ancienneté de permis. Il remonte pour les 60 ans et on enregistre un taux de transformation de plus de 40%.

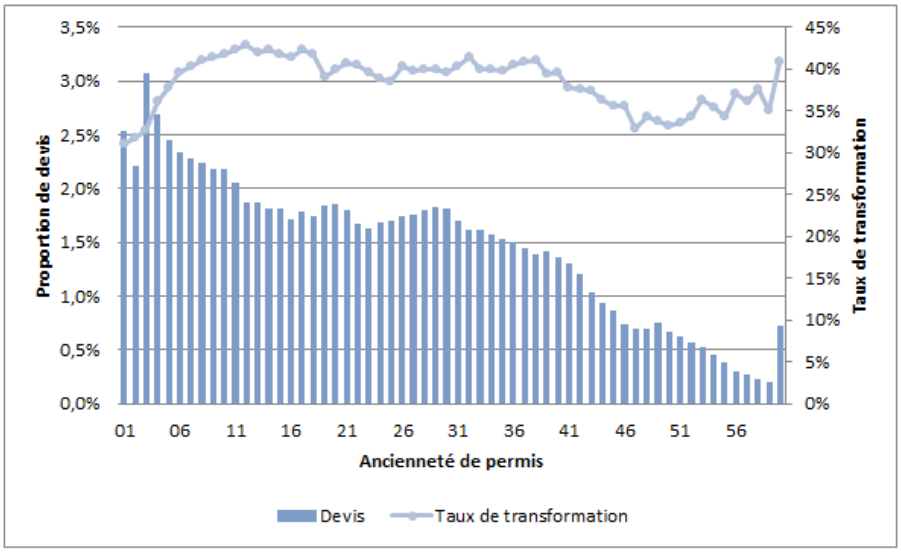


FIGURE 2.9 – Evolution du taux de conversion en fonction de l'ancienneté de permis du conducteur principal

- **Coefficient bonus-malus :**

Dans notre base d'étude, les modalités de la variable coefficient bonus-malus vont de 0.5 à >1, comme le montre le graphe ci-dessous. Près de 49% des devis sont enregistrés sur des profils dont le coefficient bonus-malus est de 0.5. Les autres modalités représentent individuellement moins de 10% des données, sauf la modalité 1 pour laquelle on a environ 13% des devis. Les profils ayant un coefficient bonus-malus supérieur à 1 sont peu représentés dans la base : environ moins de 2%. En ce qui concerne le taux de transformation, on remarque qu'on a des valeurs autour de 45% voire plus pour des profils dont le coefficient bonus-malus varie entre 0.51 et 0.8. Puis, il baisse progressivement pour les coefficients bonus-malus au-delà de 0.8.

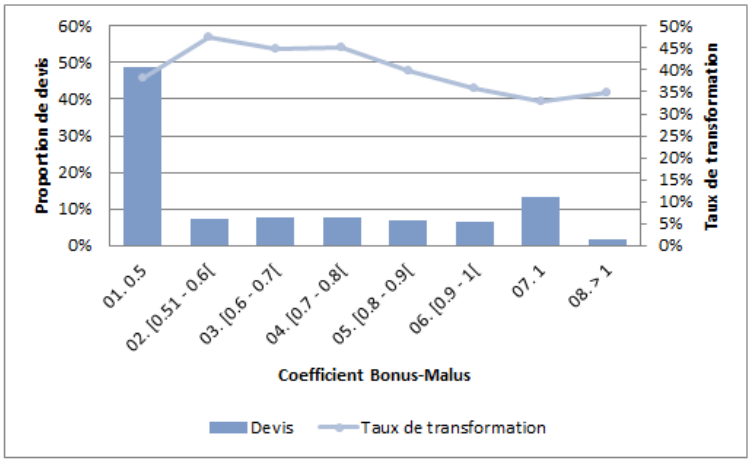


FIGURE 2.10 – Evolution du taux de conversion en fonction du coefficient bonus-malus du conducteur principal

- **Age du véhicule :**

Dans la base d'étude, on compte près de 12% de devis faits sur des véhicules neufs (Age Véhicule = 0). Ensuite, pour les modalités de 1 et 14 ans, la proportion de devis varie entre 4% et 5%. Ces proportions baissent peu à peu lorsque l'âge du véhicule augmente. Quant au taux de transformation, on remarque que pour les véhicules neufs, il est de 38,5%, proche du taux moyen sur toute la base qui est de 39%. C'est le taux le plus élevé pour les véhicules vieux de moins de 10 ans. Entre 1 et 10 ans, le taux de transformation évolue peu à peu de 33,9% à 37,9%. Au-delà de 10 ans, on observe toujours une croissance du taux de transformation. Il se stabilise autour 47%, entre 22 et 25 ans et baisse à 26 ans.

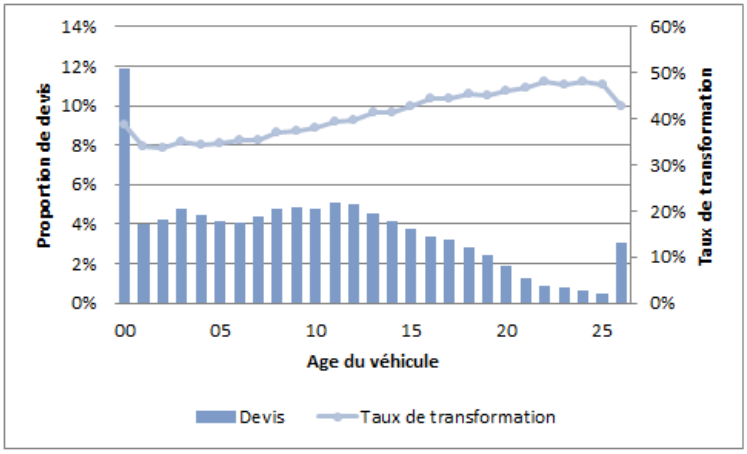


FIGURE 2.11 – Evolution du taux de conversion en fonction de l'âge du véhicule

Chapitre 3

Cadre théorique

3.1 Modèle linéaire généralisé

Dans cette partie, nous rappellerons les notions théoriques liées aux modèles linéaires généralisés, puis celle liées au cas particulier de la régression logistique. Dans le cadre de notre étude, la variable cible est dichotomique ; utiliser un modèle de régression logistique est donc envisageable.

Théorie du modèle linéaire généralisé

La théorie des modèles linéaires généralisés est introduite par Nelder et Wedderburn en 1972. Plus tard, d'autres travaux comme ceux de Nelder et Mc Cullagh (1983), Agresti (1983) ou encore Antoniadis et al. (1992) permettent d'y apporter plus de précision. L'idée sous-jacente est de pouvoir exprimer une variable cible en fonction d'une combinaison linéaire de variables explicatives.

Un modèle catalogué dans la classe des modèles linéaires généralisés présente trois composantes : une composante aléatoire, une composante systématique et une fonction dite de lien. Dans la suite, nous présenterons ces composantes dans un cadre général. Puis, nous donnerons les particularités du modèle de régression.

La distribution

La composante aléatoire donne une information sur la distribution de probabilités de la variable cible. On considère n variables aléatoires $Y_1, \ldots, Y_n$ indépendantes et dont les distributions ont une structure exponentielle. C'est-à-dire que les lois de ces variables sont dominées par une mesure commune dont les densités par rapport à cette mesure s'écrivent sous la forme :

$$f(y_i, \theta_i, \phi) = \exp\{\frac{y_i\theta_i - v(\theta_i)}{u(\phi)} + \omega(y_i, \phi)\} \tag{3.1}$$

Où

- θ_i est le paramètre naturel de la loi exponentielle ;
- u est une fonction non nulle et dérivable sur $\mathbb{R}$;
- v est une fonction trois fois dérivable sur $\mathbb{R}$ et sa dérivée première est inversible ;
- ω est une fonction définie sur $\mathbb{R}^2$.

Dans certains cas, la fonction u s'écrit :

$$u(\phi) = \frac{\phi}{\omega_i}$$

Où ω_i représente le poids des observations, souvent fixé à 1 pour simplifier. Dans ce cas, ϕ est appelé paramètre de dispersion ou de nuisance. En utilisant l'expression précédente de u, on peut écrire (3.1) sous sa forme canonique en posant :

$$\begin{aligned} Q(\theta) &= \frac{\theta}{\phi} \\ a(\theta) &= \exp\{-\frac{v(\theta)}{\phi}\} \\ b(y) &= \exp\{\omega(y,\phi)\} \end{aligned}$$

Et donc :

$$f(y_i, \theta_i, \phi) = a(\theta_i) * b(y_i) \exp\{y_i * Q(\theta_i)\} \tag{3.2}$$

Il faut noter que, la mesure commune précédemment citée change d'une structure exponentielle à l'autre. En effet, on considère la mesure de Lebesgues dans le cas de variables continues, tandis que pour des variables discrètes, on considère une combinaison de masses de Dirac.

Le prédicteur linéaire

Soit X la matrice du plan d'expérience, définie par :

$$X = \begin{pmatrix} 1 \\ x_1 \\ ... \\ x_p \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{n,p} \end{pmatrix}$$

Soit β un vecteur de p paramètres. Le prédicteur linéaire, encore appelé composante déterministe du modèle, est un vecteur de taille n défini par :

$$\eta = X * \beta$$

La fonction de lien

La troisième composante, appelée fonction de lien, permet d'établir une relation fonctionnelle entre la composante aléatoire et le prédicteur. La fonction de lien g, est une fonction monotone et différentiable définie par :

$$\eta_i = g(\mu_i), \ i = 1, ..., n$$

Où $\mu_i = E(Y_i)$ $i = 1, ..., n$

En partant de l'expression précédente, on peut écrire

$$g(\mu_i) = x_i' * \beta, \ i = 1, ..., n$$

Ce qui signifie qu'on peut écrire une fonction de la moyenne comme appartenant au sous-espace engendré par les variables explicatives. Si g est telle que $g(\mu_i) = \theta_i$ alors g est appelée fonction de lien canonique.

Significativité des variables

Afin d'évaluer la significativité des variables explicatives, on teste pour $j \in \{1, \ldots, p\}$ $H_0 : \beta_j = 0$ contre $\beta_j \neq 0$. Dans ce cas, on peut se référer au test de Wald. Ce dernier se base sur la normalité asymptotique des estimateurs de maximum de vraisemblance. La statistique du test est définie par

$$W = \frac{\hat{\beta}^2}{Var(\hat{\beta})}$$

Cette statistique suit asymptotiquement une loi de Chi-deux à un degré de liberté. En pratique, le seuil de significativité du test est de 5% : on rejette H_0 si $W > \chi^2_{95\%}$, $\chi^2_{95\%}$ étant le quantile à 95% de la loi de Chi-deux à un degré de liberté.

Outre le test de Wald, on peut également citer la déviance comme indicateur de la significativité des variables. Comme expliqué précédemment, elle permet de déterminer si l'ajout de variables explicatives permet d'améliorer la qualité du modèle. Sous l'hypothèse nulle, les coefficients des variables supplémentaires sont tous nuls.

Sélection des variables explicatives

Dans la recherche du meilleur modèle, on cherche à optimiser l'un des critères évaluant la qualité du modèle. Ceci est semblable à déterminer le vecteur de variables qui optimisent une fonction cible : la qualité du modèle. Pour faire cela, on peut procéder par une recherche exhaustive ou partir d'un point de départ et faire une recherche pas à pas. C'est cette deuxième alternative que nous avons choisie de développer. Nous présentons ci-dessous les algorithmes des méthodes ascendante et descendante de sélection des variables.

Considérons un jeu de données avec p variables explicatives et, l'AIC comme critère de qualité.

1. La méthode forward ou méthode ascendante :

- au départ, estimer le modèle 'nul' M_0 construit à partir d'une constante ;
- estimer l'ensemble des modèles construits à partir de la constante et de l'une des variables. Puis retenir le modèle M_1 qui minimise l'AIC ;
- partir du modèle retenu et estimer l'ensemble des modèles construits en lui ajoutant une nouvelle variable. On conservera celui qui minimise l'AIC ;
- reprendre l'étape précédente jusqu'à ce que l'ajout d'une nouvelle variable ne permette plus de réduire l'AIC : cela revient à ajouter une à une les variables les plus significatives.

2. La méthode backward ou méthode descendante :

- au départ, estimer le modèle 'complet' M_p construit à partir des p variables ;
- estimer l'ensemble des modèles construits à partir de $p-1$ variables. Puis retenir celui qui minimise l'AIC ;
- partir du modèle retenu et estimer l'ensemble des modèles construits en lui enlevant une variable. On conservera celui qui minimise l'AIC ;
- reprendre l'étape précédente jusqu'à ce que la suppression d'une variable ne permette plus de réduire l'AIC : cela revient à retirer une à une les variables les moins significatives.

Validation du modèle

Pour apprécier la qualité du modèle, nous pouvons faire une analyse des résidus. Pour chaque observation $i = 1, \ldots, n$ on définit :

- Les résidus de Pearson :

$$e_i^p = y_i - \hat{\pi}(x_i)$$

- Les résidus de Pearson standardisés :

$$e_i^{ps} = \frac{e_i^p}{\sqrt{1-h_i}} \text{ Où } h_i \text{ : i-ème terme de la matrice de projection.}$$

- Les résidus de Pearson standardisés et studentisés :

$$e_i^{pss} = \frac{e_i^{ps}}{\sqrt{\phi}} \text{ Où } \phi \text{ : le paramètre de dispersion.}$$

- Les résidus de déviance :

$$e_i^D = \sqrt{-2\ln(\hat{p}_i(x_i))},\ \text{si } y_i = 1$$
$$e_i^D = \sqrt{-2\ln(1-\hat{p}_i(x_i))},\ \text{si } y_i = 0$$

Remarquons que la déviance peut s'écrire comme une somme de résidus de déviance

$$D = \sum_{i=1}^{n} (e_i^D)^2$$

3.2 Régression logistique

La régression logistique est un cas particulier de modèle linéaire généralisé pour lequel la variable à expliquer Y est binaire. Y se définit par :

$$Y = \begin{cases} 1 & \text{, si le devis est converti} \\ 0 & \text{, s'il ne l'est pas} \end{cases}$$

En s'appuyant sur les données relatives au risque, l'idée est d'estimer les coefficients du modèle de régression, afin de prédire la probabilité de conversion d'un devis. On suppose que Y suit une loi de Bernouilli de paramètre $P(Y = 1)$. Les notions que nous développerons dans la suite sont propres à la régression logistique. Toutefois, elles peuvent être réadaptées dans le cadre de tout autre modèle linéaire généralisé. On a ainsi :

$$E(Y) = 1 * P(Y = 1) + 0 * P(Y = 0) = P(Y = 1)$$

c'est-à-dire,

$$E(YX = x) = P(Y = 1|X = x)$$

Par conséquent le modèle s'écrit :

$$g(P(Y = 1|X)) = X * \theta$$

La fonction de lien

Dans le cadre d'une régression logistique, on distingue trois principales fonctions de lien :

- Le lien logit : $g(y) = \ln(\frac{y}{1-y})$
- Le lien probit : $g(y) = \phi^{-1}(y)$, où ϕ est la fonction de répartition de la loi normale centrée réduite
- Le lien log-log : $g(y) = \log(-\log(1 - y))$

En théorie, aucune règle ne recommande l'utilisation d'une fonction plutôt qu'une autre. Cependant, la fonction logit est reconnue comme la plus utilisée en pratique, car elle présente de bonnes propriétés statistiques. En effet, la fonction logit est la fonction de lien canonique. Pour la suite, c'est bien la fonction logit que nous retiendrons. Par conséquent :

$$g(E(Y|X)) = \ln(\frac{P(Y = 1|X)}{P(Y = 0|X)}) = X * \beta \tag{3.3}$$

Estimation des paramètres

Pour estimer les paramètres β_i, nous utilisons la méthode de maximum de vraisemblance. Dans notre cas, la vraisemblance est donnée par :

$$L = \prod_{i=1}^{n} f(y_i|X = x_i), y \in \{0, 1\}$$

Où f est la densité d'une loi de Bernouilli. Posons $l_i = \ln(f(y_i|X = x_i))$.
Par conséquent, la log-vraisemblance s'écrit :

$$\mathcal{L} = \sum_{i=1}^{n} l_i$$

Ainsi, pour la maximiser il faut résoudre l'équation $\frac{\partial \mathcal{L}}{\partial \beta_j} = 0, \forall j$.
Calculons

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

Comme

$$\begin{aligned}
&\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - v'(\theta_i)}{u(\phi)} = \frac{y_i - \mu_i}{u(\phi)} \\
&\frac{\partial \mu_i}{\partial \theta_i} = v''(\theta_i) = Var(Y_i)/u(\phi) \\
&\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \text{ car } \eta_i = x_i'\beta \\
&\frac{\partial \mu_i}{\partial \eta_i} \text{ dépend de la fonction de lien } \eta_i = g(\mu_i)
\end{aligned}$$

On obtient

$$\frac{l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}, \ j = 1, ..., p$$

Ces équations sont non-linéaires en β et leur résolution requiert des méthodes itératives faisant intervenir le Hessien (pour Newton-Raphson) ou encore la matrice de d'information (pour les Scores de Fisher).

Remarquons que $\mu_i = y_i$ représente une solution particulière : il s'agit du modèle dit saturé. En pratique, ce modèle est inutilisable et aucun modèle estimé ne permet d'aboutir à de meilleurs résultats.

Comparaison des modèles

Pour comparer différents modèles, de nombreux critères peuvent être utilisés. Ils permettent d'évaluer la qualité d'ajustement du modèle sur la base des différences entre estimations et valeurs observées. Dans cette section, nous présenterons trois critères de comparaison très souvent utilisés : le critère Akaike Information Criterion (en abrégé AIC), le critère Bayesian Information Criterion (BIC) et la déviance.

- La déviance

Soient $\mathcal{L}_{max}$ et $\mathcal{L}_k$ les log-vraisemblances respectives du modèle saturé et du modèle à k degrés de liberté. La déviance est une mesure d'ajustement global du modèle au jeu de données. Sous sa forme normalisée, elle s'exprime de la manière suivante :

$$D(\beta_k) = -2 * (\mathcal{L}_k - \mathcal{L}_{max})$$

Dans le cas de la régression logistique, la variable cible est à valeurs dans 0, 1 et, la vraisemblance du modèle saturé est de 1. Par conséquent :

$$D(\beta_k) = -2 * (\mathcal{L}_k)$$

Donc maximiser la vraisemblance équivaut à minimiser la déviance. Cette dernière permet également de réaliser des tests dits de modèles emboités : test de rapport de vraisemblance. Ce test permet de comparer un modèle avec un modèle réduit. Le rapport de vraisemblance ou la différence de déviance capte l'apport de variables explicatives supplémentaires dans l'ajustement du modèle. Sous l'hypothèse nulle H_0, les deux modèles sont équivalents. La statistique du test est basée sur la différence des déviances entre les modèles emboités. Considérons des modèles emboités avec respectivement k_1 et k_2 degrés de liberté ($k_1 < k_2$). On a :

$$D(\beta_{k_2}) - D(\beta_{k_1} = 2 * (\mathcal{L}_{k_1} - \mathcal{L}_{max}) - 2 * (\mathcal{L}_{k_2} - \mathcal{L}_{max}) = 2 * (\mathcal{L}_{k_1} - \mathcal{L}_{k_2})$$

Cette statistique suit une loi de χ^2 à $k_2 - k_1$ degrés de liberté pour les lois à un paramètre (ex - poisson) et une loi de Fisher pour les lois à 2 paramètres (ex - loi normale).

- Les critères AIC et BIC

Ces critères permettent tous les deux de comparer des modèles entre eux. Ils sont fonction de la log-vraisemblance du modèle étudié et maximiser la vraisemblance conduit à les minimiser. Ils décroissent notamment lorsqu'on intègre au modèle des variables pertinentes. Le BIC donné par :

$$BIC = -2 * \mathcal{L} + \ln(n) * p$$

L'AIC est donné par :

$$AIC = -2 * \mathcal{L} + 2 * p$$

Où L est la log-vraisemblance maximisée du modèle et p le nombre de coefficients du modèle.

Bien que les deux critères soient assez proches, la complexité du modèle est davantage pénalisée dans le cas du BIC lorsque la taille n de l'échantillon augmente. Dans la comparaison de modèles en utilisant l'AIC ou le BIC, on retiendra celui qui minimise le critère considéré.

Interprétation des coefficients

En partant de la relation (3.3), on montre que

$$E(Y|X) = P(Y = 1|X) = \frac{\exp\{X * \beta\}}{1 + \exp\{X * \beta\}}$$

On constate que la relation précédente n'est ni linéaire, ni multiplicative. Ce qui pose un problème pour l'interprétation des coefficients. Par contre, en considérant la cote [4] (odds en anglais), on obtient

$$odds(x) = \frac{P(Y = 1|X)}{P(Y = 0|X)} = \exp\{X * \beta\} = \prod_{j=1}^{p} \exp\{\beta_j * x^{(j)}\}$$

Les variables explicatives ont alors un effet multiplicatif sur la cote. On a $odds(x)$ fois plus de chances qu'un prospect de caractéristiques x convertisse un devis, plutôt qu'il ne le convertisse pas. On définit également le rapport de cotes (odds-ratio) :

$$OR = \frac{odds(x)}{odds(x')}$$

Il correspond au rapport de chances entre deux individus ayant des caractéristiques x et x' différentes. $OR = \alpha$ signifie que le rapport de chances entre la conversion de devis ($Y = 1$) et la non conversion ($Y = 0$) est multiplié parα lorsqu'on change les caractéristiques liées au prospect. Un OR de 1 signifie qu'il n'y pas d'effet dû au changement des caractéristiques.

Dans le cadre de ce mémoire, le premier modèle employé pour modéliser la conversion de devis est la régression logistique. Pour s'assurer de la convergence d'un tel modèle, il convient de vérifier les potentiels effets de corrélations entre les variables explicatives.

4. La cote est un rapport de probabilité entre un événement et son complémentaire.

Lorsque des variables fortement corrélées sont introduites dans un modèle linéaire généralisé, ou en particulier dans une régression logistique, cela crée un biais sur les prédicteurs. Les modèles machine learning que nous développerons dans la suite, ne sont pas sujets à ce problème. On peut y intégrer l'ensemble des variables dont nous disposons, et l'algorithme se chargera de déterminer intuitivement les plus pertinentes. C'est-à-dire celles qui expliquent le mieux l'acte de conversion.

3.3 Algorithmes de machine learning

3.3.1 Algorithme CART

L'acronyme CART (Classification And Regression Trees) renvoie à la formalisation par Breiman et col. (1984) des méthodes de partitionnement récursif connues depuis les années 60. Il peut être utilisé aussi bien dans le cadre d'une classification que dans le cadre d'une régression. L'idée est de procéder à une suite récursive de splits au sein d'une population, afin de créer des sous-groupes de plus en plus homogènes. A la fin du partitionnement, on attribue soit une classe à chaque sous-groupe s'il s'agit d'une classification, soit une valeur, si la variable cible est quantitative. Puis vient l'étape d'optimisation, encore appelée élagage, de l'arbre de construit. Cette étape consiste à réduire la complexité de l'arbre, afin d'éviter le sur-apprentissage.

Principe de construction d'un arbre

Considérons un jeu de données constitué de p variables explicatives $(X_j)_{j=1,\dots,p}$ et d'une variable cible Y. Construire un arbre consiste à déterminer une suite de nœuds telle que :

- un nœud est construit en choisissant une variable et en définissant un critère de segmentation sur celle-ci afin de scinder la population en deux sous-ensembles. Ce qui implique qu'à chaque nœud, on crée deux sous-échantillons disjoints.

- le critère de division correspond à une valeur seuil si Y est quantitative, ou à un fractionnement en deux groupes de modalités si Y est qualitative.

Ces étapes sont appliquées à un nœud initial appelé racine de l'arbre, qui correspond à l'ensemble du jeu de données. Puis, elles sont itérées sur les sous-ensembles créés. Toutefois, il faut noter que pour pouvoir appliquer cet algorithme, on doit définir :

- un critère permettant de réaliser la meilleure segmentation possible parmi toutes celles qui peuvent être faites sur les différentes variables ;

- une contrainte définissant un nœud terminal ;

- comment on affecte à chaque nœud terminal une classe ou une valeur.

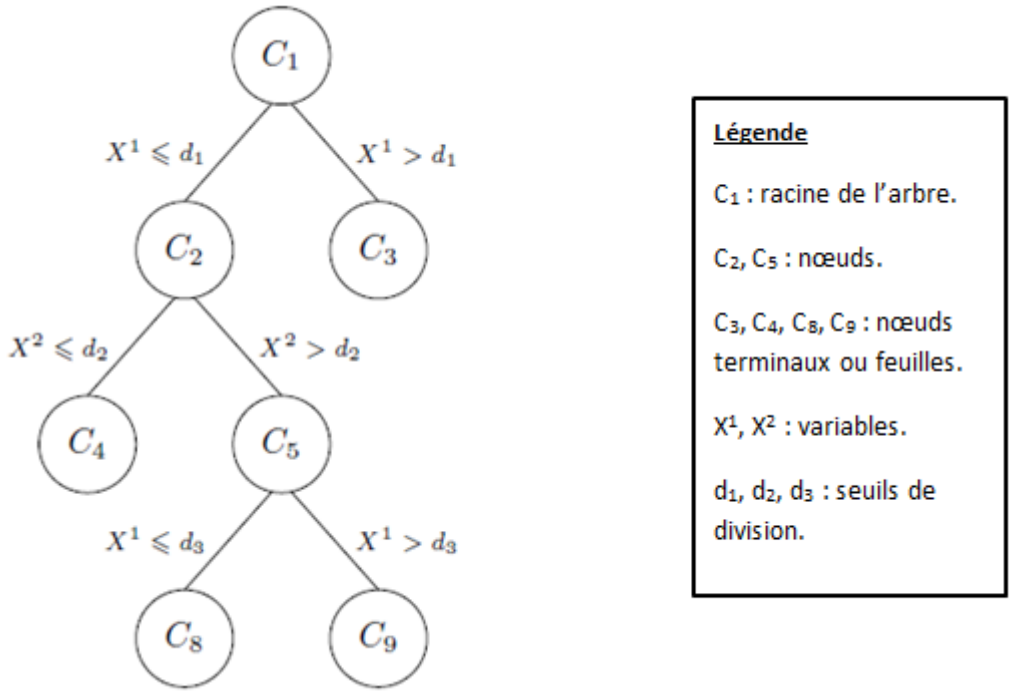


FIGURE 3.1 – Illustration d'un arbre CART

Principe de division

L'une des contraintes pour scinder un nœud est de s'assurer qu'aucun des deux nœuds fils n'est vide. Pour une variable qualitative ordinale ayant m modalités ou une variable quantitative, il existe $m-1$ divisions binaires possibles. Si la variable est nominale, on peut faire $2^{(m-1)}-1$ divisions. Il faut noter que l'algorithme favorise la sélection de variables ayant un grand nombre de modalités, car elles offrent une certaine souplesse dans la création des deux sous-ensembles. La segmentation d'un nœud s'appuie sur une fonction d'hétérogénéité servant à partager la population en des sous-groupes homogènes, au sens de la variable cible. Cette hétérogénéité est mesurée au sein d'un nœud par une fonction positive D qui est

- nulle, pour un nœud homogène : c'est-à-dire que les individus ont une valeur identique de Y ou appartiennent à une même modalité ;
- maximale si les valeurs de Y associées aux individus de la population sont assez dispersées.

Cette division crée pour un nœud K, deux nœuds fils qu'on notera K_d et K_g. De toutes les divisions possibles, celle qui est retenue est celle qui minimise la somme $D_{K_G} + D_{K_D}$. En pratique, cela revient à résoudre

$$\underset{\text{divisions de } X^j}{max} D_K - (D_{K_G} + D_{K_D})$$

Où D_{K_G} et D_{K_D} sont les mesures d'hétérogénéité des noeuds fils et D_K celle du noeud K.

Contrainte d'arrêt et affectation

L'allongement de l'arbre s'interrompt à un nœud donné, appelé nœud terminal ou feuille. Ceci se produit dans les cas suivants :

- Le nœud est homogène ;
- Aucune division n'est possible ;
- Le découpage est tellement fin que le nombre d'observations dans l'un des nœuds fils serait inférieur à une valeur seuil précisée au début de l'algorithme.

Si la variable cible est quantitative, à chaque feuille est associée une valeur correspondant à la moyenne des Yi des observations de la feuille. Si la variable cible est qualitative, on affecte à la feuille la modalité la plus représentée de Y.

Critère d'homogénéité

1. **Si Y est quantitative**
   Dans le cas d'une régression, l'hétérogénéité au sein du nœud K est mesurée par la variance :

   $$D_K = \frac{1}{|K|} \sum_{i \in K} (y_i - \overline{y}_k)^2$$

   Où $|K|$ est le nombre d'observations dans le nœud K.
   Lors de la segmentation d'un nœud, on cherche la variable et la division qui conduiront à la plus grande baisse d'hétérogénéité dans les nœuds fils. Ce qui revient à minimiser

   $$\frac{|K_G|}{n} \sum_{i \in K_G} (y_i - \overline{y}_{K_G})^2 + \frac{|K_D|}{n} \sum_{i \in K_D} (y_i - \overline{y}_{K_D})^2$$

   La meilleure segmentation est également celle qui présente le plus de significativité au sens du test d'analyse de variance de Fisher.

2. **Si Y est qualitative**
   Suppose que Y a m modalités. Il existe plusieurs mesures de l'hétérogénéité et les plus connues sont :

   - L'entropie :
     Pour le nœud K, l'entropie est donnée par

$$D_K = -2 \sum_{i=1}^{m} |K| p_K^l \log(p_K^l)$$

   Où p_K^l est la proportion de la modalité l de Y dans le nœud K.

   - La concentration de Gini

   Elle est définie par

$$D_K = p_K^l (1 - p_K^l)$$

   Comme pour une variable cible quantitative, on recherche lors de la segmentation, la possibilité de réduire au maximum l'hétérogénéité au sein des nœuds fils.

La notion d'élagage

Grâce à la démarche décrite jusqu'ici, on a construit ce qu'on appelle un arbre maximal. Celui-ci est très fin et dont susceptible d'être sujet au sur-ajustement. Afin d'éviter ce problème, on désire construire un arbre plus parcimonieux, qui sera plus robuste en termes de prédiction sur des données test. Cette procédure est appelée élagage (pruning en anglais) de l'arbre. Elle consiste à déterminer un sous-arbre de l'arbre maximal, pour lequel on aurait une bonne qualité de prédiction. A cet effet, Breiman et col. (1984) ont proposé une méthode reposant sur une suite emboitée de sous-arbres de l'arbre maximal, puis on choisit celui qui minimise l'erreur de généralisation. On aboutit à un optimum local, dont l'efficacité et la fiabilité sur un échantillon de validation seraient meilleures que celles de l'arbre maximal.

3.3.2 Random Forest

Le bagging

Suite aux remarques sur le manque de stabilité et le caractère sensible des arbres CART, Breiman introduit en 1996 une nouvelle méthode appelée le Bagging. Dans le cadre général, considérons une méthode de prédiction, qui sur un jeu de données X_n, construit un

prédicteur $\hat{h}(X_n, \theta)$. Le principe du bagging consiste à construire un grand nombre de prédicteurs, sur des échantillons tirés aléatoirement, en appliquant à chacun d'eux la même méthode de prédiction. Les prédicteurs sont ensuite agrégés, par exemple en faisant une moyenne ou un vote majoritaire. L'idée est qu'en appliquant la même règle de prédiction, les prédicteurs obtenus sur les échantillons bootstrap sont différents. On construit ainsi une collection variée de prédicteurs donc l'agrégation conduirait à un prédicteur performant. L'un des grands intérêts de cette méthode réside dans la réduction de la variance du prédicteur.

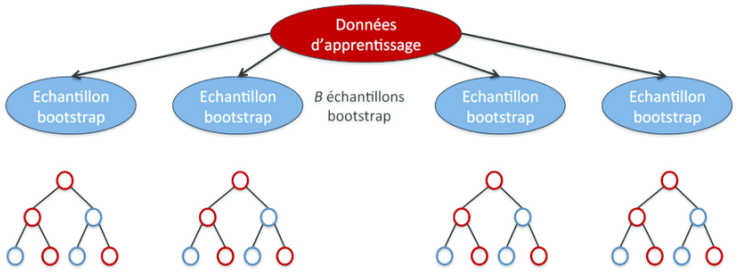


FIGURE 3.2 – Principe du bagging

Dans l'exemple présenté sur le figure ci-dessus, on construit B modèles à partir d'autant d'échantillons bootstrap. A chaque étape, un modèle est implémenté à partir de données tirées aléatoirement. Chaque échantillon est obtenu en faisant un tirage aléatoire avec remise sur les n observations dont on dispose. Lors de la construction d'un arbre, sont dites 'out-of-bag' (en français 'en dehors du bootstrap') les observations qui ne sont considérées. Dans la suite, nous verrons que celles-ci sont utilisées pour calculer un indicateur important pour ce type de modèle.

Construction de forêts aléatoires

Le principe de construction des forêts aléatoires (ou Random Forest), calqué sur le bagging, débute en générant des échantillons bootstrap $X_1, \ldots, X_B$ avec des colonnes non nécessairement identiques. Sur chacun des échantillons est ensuite appliqué un algorithme de type CART. De cette façon, on construit plusieurs arbres maximaux. L'ensemble des arbres obtenus sont ensuite élagués, pour aboutir au prédicteur de la forêt aléatoire. La figure récapitule ce processus de construction :

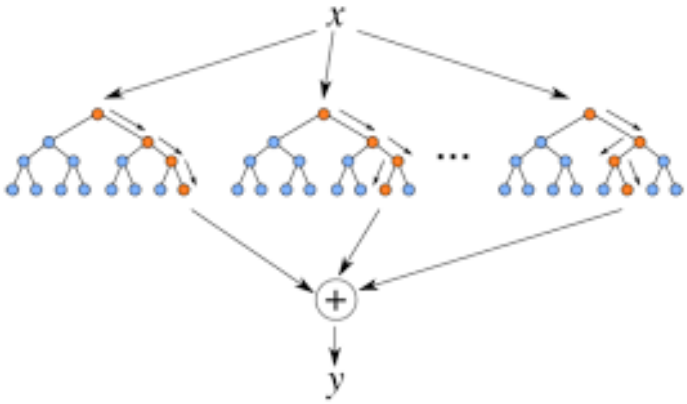


FIGURE 3.3 – Construction d'une forêt aléatoire

Ainsi, une forêt aléatoire est semblable à une procédure Bagging, la différence résidant dans la construction d'arbres individuels. Dans la construction des différents échantillons bootstrap, le nombre m de variables est fixé d'avance et est identique pour tous les échantillons. Prendre m = p (p étant le nombre total de variables explicatives) équivaut à du Bagging d'arbres CART non élagués. Ce tirage ajoute de l'aléa à l'algorithme, ce qui en fait un paramètre très important. Pour une forêt aléatoire, on distingue deux principales sources d'aléa :

- la première liée au bootstrap : tree sampling ;
- la seconde liée au choix du nombre de variables : feature sampling.

Ce qui revient à dire qu'on perturbe non seulement le jeu de données mais aussi le noyau de variables permettant la construction du prédicteur. D'après les travaux de Breiman (2001), la construction de forêts aléatoires (avec m convenablement choisi) donne de meilleurs résultats que le Bagging. Ceci vient du fait que l'introduction de l'aléa par le 'feature sampling' contribue à la robustesse du prédicteur, sans dégrader la performance de l'algorithme. Interpréter des résultats d'une forêt aléatoire est moins évident que ceux d'un arbre CART à cause du caractère aléatoire des arbres. Toutefois, l'algorithme offre la possibilité de déterminer l'importance des variables. Plus une variable est importante, plus grande est sa contribution dans la construction des arbres.

Erreur de généralisation

Outre le prédicteur, l'algorithme Random Forest détermine une estimation de son erreur de généralisation : l'erreur 'Out-Of-Bag' (OOB). L'idée sous-jacente est tirée du Bagging. Elle est calculée de la manière suivante :

Considérons une observation (X_i, Y_i) et les différents arbres pour lesquels elle est 'Out-Of-Bag'. En agrégeant les prédictions de ces arbres, on peut construire une estimation $(\hat{Y}_i)$ de Y_i. En répétant l'opération sur toutes les données, on peut déterminer la proportion d'observations mal classées $\frac{1}{n}\sum_{i=1}^{n} 1_{\hat{Y}_i \neq Y_i}$.. Cette proportion est l'erreur OOB de l'algorithme Random Forest. Au fil des itérations, elle permet de déterminer le nombre optimal d'arbres. L'un des avantages est qu'elle n'exige pas de découper les données d'apprentissage ; le découpage est déjà inclus dans la construction des échantillons bootstrap.

3.3.3 Gradient Boosting Machine

Le boosting

L'idée de base du boosting est proche de celle du bagging : on construit un ensemble d'estimateurs puis on les agrège par une moyenne pondérée ou un vote. La principale différence intervient dans la construction de ces estimateurs. En effet, dans le cas du boosting, ils sont construits de manière séquentielle, chacun étant une version adaptative du précédent. D'une itération à la suivante, on accorde plus de poids aux observations mal prédites, ce qui améliore progressivement la performance de prédiction. Corriger les poids au fil des itérations aide à mieux prédire les valeurs difficiles et l'agrégation atténue le risque de sur-apprentissage.

L'idée d'un tel modèle a été initiée en 1990 par Schapire, puis affinée par Freund et Schapire en 1996. Ils mettent sur pied un algorithme connu sous le nom d'AdaBoost (Adaptative Boosting), pour la prévision d'une variable binaire. Par la suite, de nombreuses études ont été menées pour améliorer cet algorithme, notamment pour la classification avec k classes (k>2), la régression... Cet algorithme présente des similitudes avec le bagging car il conduit lui aussi à une réduction de la variance des estimateurs. De plus, il apporte de la réduction de biais ce qui n'est pas le cas dans le bagging car les arbres sont y identiquement distribués. En conséquence, l'espérance de B arbres est proche de celle d'un arbre. Et donc le biais de ces arbres agrégés est sensiblement le même que celui d'un unique arbre ; ce qui n'est pas le cas avec le boosting.

Les différents algorithmes de boosting peuvent différer par :

- La méthode pondération : liée à la mise à jour des poids des observations mal prédites au fil des itérations ;
- L'objectif : classification, régression ;
- La méthode d'agrégation des modèles ;
- La fonction de perte : utilisée pour capter l'erreur d'ajustement.

Le bon comportement du boosting par rapport à d'autres techniques de discrimination est difficile à expliquer ou justifier par des arguments théoriques. À la suite d'une proposition de Breiman en 1999 (rapport technique) de considérer le boosting comme un algorithme global d'optimisation, Hastie et col. (2001) présentent le boosting dans le cas binaire sous la forme d'une approximation de la fonction f par un modèle additif construit pas à pas :

$$\hat{f}(x) = \sum_{m=1}^{M} c_m \delta(x; \gamma_m)$$

qui est une combinaison où c_m est un paramètre, δ le classifieur (faible) de base fonction de x et dépendant d'un paramètre m. Si l est une fonction perte, il s'agit, à chaque étape, de résoudre :

$$(c_m, \gamma_m) = \arg \min_{(c,\gamma)} \sum_{i=1}^{n} l(y_i, \hat{f}_{m-1}(x_i) + c * \delta(x_i; \gamma))$$

$\hat{f}_m(x) = \hat{f}_{m-1}(x) + c_m \delta(x; \gamma_m)$ est alors une amélioration de l'ajustement précédent. Dans le cas de l'algorithme adaboost pour l'ajustement d'une fonction binaire, la fonction de perte utilisée est $l(y, f(x)) = \exp -y * f(x)$. Il s'agit donc de résoudre

$$\begin{aligned}(c_m, \gamma_m) &= \arg \min_{(c,\gamma)} \sum_{i=1}^{n} \exp\{-y_i * (\hat{f}_{m-1}(x_i) + c * \delta(x_i; \gamma))\} \\ &= \arg \min_{(c,\gamma)} \sum_{i=1}^{n} w_i^m * \exp\{-c y_i \delta(x_i; \gamma)\}\end{aligned}$$

avec $w_i^m = \exp\{-y_i \hat{f}_{m-1}(x_i)\}\ w_i^m$
w_i^m ne dépendant ni de c ni de γ,, il joue le rôle de poids en fonction de la qualité de l'ajustement précédent. Quelques développements complémentaires montrent que la solution du problème de minimisation est obtenue en deux étapes :

- La recherche du classifieur optimal :

$$\gamma_m = \arg \min_{\gamma} \sum_{i=1}^{n} 1_{\{y_i \neq \delta(x_i;\gamma)\}}$$

- L'optimisation du paramètre c_m

$$c_m = \frac{1}{2} \log(\frac{1 - \hat{\epsilon}_p}{\epsilon_p})$$

Avec $\hat{\epsilon}_p$ l'erreur apparente de prévision, tandis que les poids w_i sont mis à jour avec :

$$w_i^m = w_i^{m-1} \exp\{-c_m\}$$

Gradient boosting machine

Dans le même esprit d'approximation adaptative, Friedman (2002) a proposé sous l'acronyme MART (multiple additive regression trees) puis sous celui de GBM (gradient boosting models) une famille d'algorithmes basés sur une fonction perte supposée convexe et différentiable notée l. Le principe de base est le même que pour adaBoost, construire une séquence de modèles de sorte que chaque étape, chaque modèle ajouté à la combinaison, apparaisse comme un pas vers une meilleure solution. La principale innovation est que ce pas est franchi dans la direction du gradient de la fonction perte, afin d'améliorer les propriétés de convergence. Une deuxième idée consiste à approcher le gradient par un arbre de régression afin d'éviter un sur-apprentissage. Le modèle adaptatif présenté précédemment est transformé en une descente de gradient :

$$\hat{f}_m = \hat{f}_{m-1} - \gamma_m \sum_{i=1}^{n} \nabla_{f_{m-1}} l(y_i, f_{m-1}(x_i))$$

Plutôt que de chercher un meilleur classifieur comme avec adaBoost, le problème se simplifie en la recherche d'un meilleur pas de descente γ :

$$\min_{\gamma} \sum_{i=1}^{n} [l(y_i, f_{m-1}(x_i) - \gamma \frac{\partial l(y_i, fm-1(x_i))}{\partial f_{m-1}(x_i)})]$$

3.4 Principe de l'analyse supervisée

En matière d'apprentissage statistique, on oppose fréquemment les méthodes dites d'apprentissage non supervisé aux méthodes d'apprentissage supervisé. Dans le premier cas, l'idée est de partitionner un ensemble d'éléments hétérogènes en des sous-groupes homogènes. Tandis que pour l'apprentissage supervisé on dispose déjà de groupes, et l'objectif est de ranger un nouvel élément dans les groupes existants. Ceci passe par la construction d'un prédicteur $f(.)$, qui à partir de données dites 'd'apprentissage', qui sera utilisé pour des besoins de prédiction.

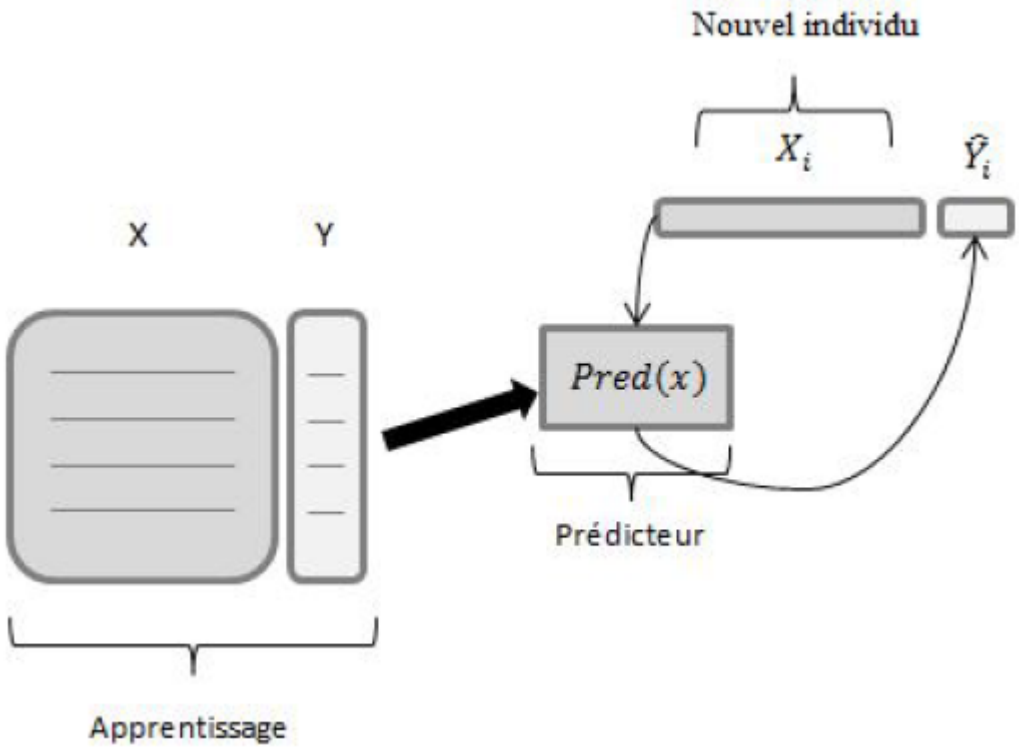


FIGURE 3.4 – Schéma du prédicteur

Les méthodes présentées dans les précédentes sections sont du domaine de l'apprentissage supervisé. Le prédicteur $f(.)$ détermine la probabilité de convertir un devis. Pour le construire et pouvoir l'utiliser, nous scindons aléatoirement la base d'étude en deux échantillons.

Le premier échantillon, qui représente la base d'apprentissage, est constitué de 80% des observations de la base d'étude. Il permet de calibrer le prédicteur. Le second échantillon, constitué de 20% de données, sera notre base test. Cette base, non utilisée pour construire $f(.)$, permet d'évaluer la qualité du prédicteur sur de nouvelles observations.

Dans chacun des sous-échantillons, il est nécessaire de vérifier que le taux de conversion est proche de celui observé sur la base d'étude, c'est-à-dire 39%.

Précisons aussi que dans le cas des modèles machine learning, on peut renforcer la performance du prédicteur sur les données d'apprentissage en procédant à une validation croisée (en anglais k-folds cross validation). Cette approche consiste à scinder les données d'apprentissage en deux : un échantillon d'entrainement et un échantillon de validation. On construit $f(.)$ sur les données d'entrainement et on évalue sa performance sur les données de validation. On répète ce processus k fois, afin de calibrer les hyper-paramètres constituant $f(.)$. La performance sur les données d'apprentissage sera déterminée en moyennant les performances obtenues sur les k itérations.

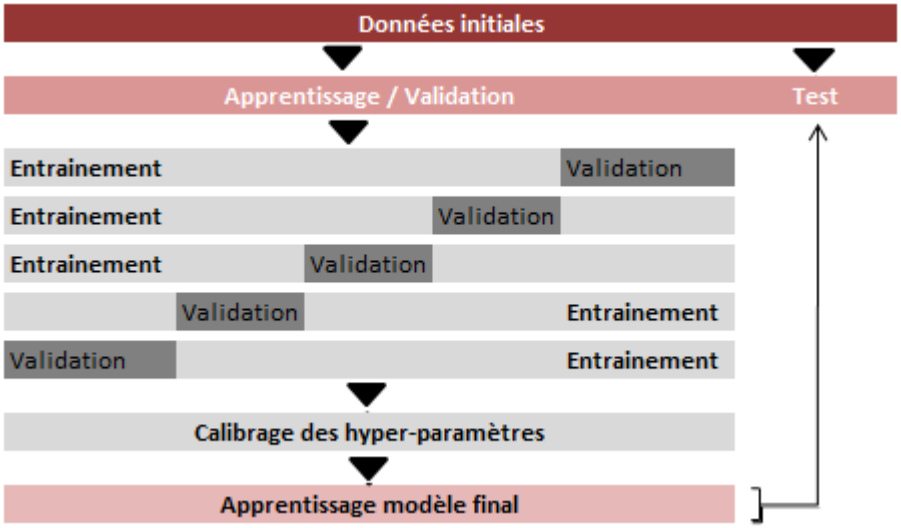


FIGURE 3.5 – Illustration du principe de validation croisée

3.5 Mesures de performance des modèles

3.5.1 Matrice de confusion

Une matrice de confusion, encore appelée tableau de contingence, est un outil utilisé dans le cadre de ce mémoire pour déterminer la performance d'un modèle supervisé. Elle croise la fréquence des prédictions avec celle des observations de la variable cible. Elle met en évidence les volumes des bonnes et des mauvaises prédictions. L'utilisateur définit un seuil s, afin d'attribuer une classe à chacune des prédictions. Un devis est converti si $\hat{P}(Y = 1|X = x) \geq s$.

Observations	Prédictions		Total
	$\hat{y}= 0$	$\hat{y}= 1$	
$Y = 0$	VN	FP	n_0
$Y = 1$	FN	VP	n_1
Total	$\hat{n}_0$	$\hat{n}_1$	n

TABLE 3.1 – Matrice de confusion

A partir de la matrice de confusion, on définit les quantités suivantes :

- Les **vrais positifs** (VP) : il s'agit du nombre VP d'observations bien classées telles que $\hat{y}_i = 1$ et $Y = 1$.
- Les **vrais négatifs** (VN) : il s'agit du nombre VN d'observations bien classées telles que $\hat{y}_i = 0$ et $Y = 0$.
- Les **faux positifs** (FP) : il s'agit du nombre FP d'observations bien classées telles que $\hat{y}_i = 1$ et $Y = 0$.
- Les **faux négatifs** (FN) : il s'agit du nombre FN d'observations bien classées telles que $\hat{y}_i = 0$ et $Y = 1$.
- le **taux d'erreur** : il s'agit de la proportion d'individus mal classés sur l'ensemble des données.

$$err = \frac{FP + FN}{n}$$

Cette mesure renseigne sur la qualité globale du modèle. Si err=10%, seul un devis sur 10 est mal classé. A partir du taux d'erreur, on peut définir une autre mesure

de qualité appelée l'*Accuracy*. Il s'agit du taux de bonnes prédictions : $Accuracy = 1 - err$. Notons que, le taux d'erreur ne permet pas de savoir si le modèle prédit mieux les 0 ou les 1. Les métriques suivantes sont définies dans ce sens.

- La **précision** : il s'agit de la proportion de 1 bien prédits sur l'ensemble des 1 prédits.

$$\text{précision } = \frac{VP}{\hat{n}_1}$$

- La **sensibilité** : il s'agit de la proportion de 1 prédits sur l'ensemble des 1 observés.

$$\text{sensibilité} = \frac{VP}{n_1}$$

Les deux indicateurs précédents permettent de mesurer la capacité de l'algorithme à identifier les devis convertis. Pour un modèle parfait, la précision et la sensibilité sont de 1 : l'ensemble des devis convertis sont bien prédits.

- La **spécificité** ou taux de vrais négatifs : il s'agit de la proportion de 0 prédits sur l'ensemble des 0 observés.

$$\text{spécificité} = \frac{VN}{n_0}$$

Elle donne une idée sur la capacité de l'algorithme à identifier le devis non convertis. Elle est de 1 pour un modèle parfait : tous les devis non convertis sont bien identifiés par le modèle.

- La **F_mesure** : il s'agit de la moyenne harmonique de la précision et de la sensibilité. Elle est définit par :

$$F_{mesure} = \frac{2}{\frac{1}{\text{précision}} + \frac{1}{\text{sensibilité}}}$$

Pour avoir une bonne valeur de F_{mesure}, il faudrait que FP et FN soient bas. Optimiser la F_{mesure} équivaut à trouver un bon compromis entre la sensibilité et la précision. Elle mesure la capacité de l'algorithme à proposer les solutions pertinentes.

Pour présenter les résultats des modèles, nous utiliserons une forme normalisée de la matrice de confusion. Celle-ci permet de met en évidence non pas les volumes des observations dans chaque classe, mais les proportions. Une matrice de confusion normalisée est de la forme :

Observations	Prédictions		Total
	$\hat{y}=0$	$\hat{y}=1$	
$Y=0$	a	b	n_0
$Y=1$	c	d	n_1

TABLE 3.2 – Matrice de confusion normalisée

Où :

$$a = \frac{VN}{n_0} \; ; \; b = \frac{FP}{n_0}$$
$$c = \frac{FN}{n_1} \; ; \; d = \frac{VP}{n_1}$$

Sur une matrice de confusion normalisée, on lit facilement la spécificité a et la sensibilité d qui sont utiles pour construire une courbe ROC.

3.5.2 Courbe ROC

Une **courbe ROC (Receiver Operating Characteristic)** est un outil graphique représentant les performances d'un modèle de classification. Elle est construite dans le plan donné par le taux de vrais positifs (la sensibilité) et le taux de faux positifs (1 – spécificité). La courbe permet de visualiser les valeurs du taux de vrais positifs en fonction du taux de faux positifs pour différents seuils de classification. Ce qui signifie que la courbe ROC relie des points (TFP(seuil) ; TVP(seuil)) en faisant varier le seuil. Pour de faibles valeurs de seuils, on augmente les chances de prédire des 1, qu'ils s'agissent de bonnes ou de mauvaises prédictions.

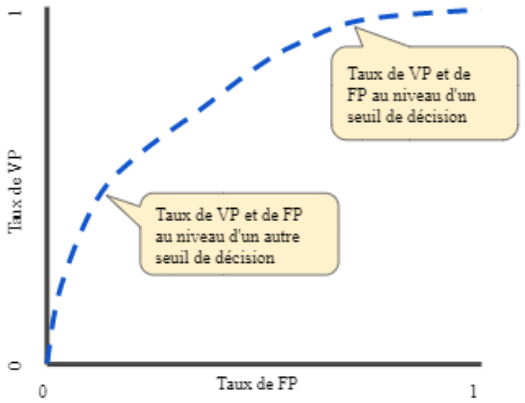


FIGURE 3.6 – Construction d'une courbe ROC

Après avoir tracé la courbe ROC, un indicateur pertinent à noter est l'aire sous celle-ci. Cet indicateur est connu sous le nom d'**AUC** qui signifie Area Under the Curve. Il varie entre 0 et 1. Il permet de mesurer la performance d'un modèle (régression logistique, algorithmes de machine learning ...). Dans le cas d'un modèle parfait, l'AUC est de 1. En général, lorsqu'on utilise l'AUC, on vérifie d'abord que celui du notre modèle est supérieur à 0.5, qui est l'AUC d'une classification aléatoire. La classification aléatoire ne fait aucune discrimination. Elle est représentée par la première bissectrice. Sur la figure suivante, nous représentons le tracé d'un courbe ROC et, l'AUC est matérialisée par une zone en gris. Plus un modèle est bon au sens de l'AUC, plus sa courbe ROC se rapproche de celle du modèle parfait, dans ce cas la zone couverte est maximale.

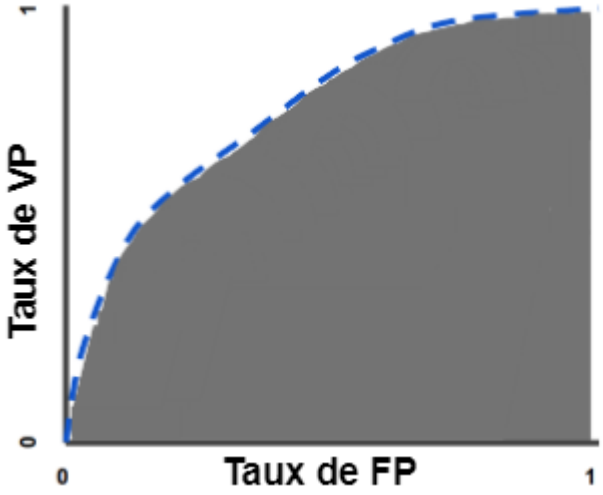


FIGURE 3.7 – Illustration d'une courbe ROC

Chapitre 4

Modélisation

Dans cette section, nous reviendrons sur les étapes de préparation de la base d'étude avant la phase de modélisation. Ensuite, nous présenterons les résultats de chaque méthode. Enfin, nous comparerons ces résultats pour déterminer quel est le modèle le plus performant.

4.1 Préparation de la base d'étude

Dans cette section, nous donnerons des détails sur l'analyse exploratoire de notre base d'étude, étape préalable à la modélisation.

Notre base d'étude est conçue comme la jointure de trois tables à savoir la base devis, les données externes et les indicateurs de compétitivité. Avant de nous lancer dans la modélisation, nous avons réalisé une analyse exploratoire de celle-ci, afin de repérer des valeurs manquantes et des variables non pertinentes.

- **Traitement des valeurs manquantes** : cette étape est rapidement réalisée car en parcourant l'ensemble des variables, nous n'avons pas retrouvé de valeurs manquantes. Pour nous, ce constat est logique car il est en accord avec les formats des variables que nous avons créés sur SAS. En effet, lors de la création de la base d'étude, nous avons défini des formats pour chaque variable. En définissant le format de chacune de nos variables, nous avons créé la modalité "Non Renseigné", qui permet de prendre en compte les cas où la variable n'est pas renseignée ou encore les cas où on observerait une valeur aberrante.
  Bien qu'elle traduise le fait que l'information n'a pas été fournie à l'établissement du devis, la modalité "Non Renseigné" peut nous apporter de l'information. Prenons le cas de la variable Nombre d'enfants. Lorsqu'aucune valeur n'est pas renseignée, on suppose que le prospect n'a pas d'enfant, ce qui peut avoir un impact sur le tarif proposé.

- **Traitement des variables non pertinentes** : par variable non pertinente, on désigne une variable pour laquelle les observations sont concentrées sur une modalité. Pour identifier les variables non pertinentes de notre base d'étude, nous avons fixé un seuil à 90%. Toute variable sur laquelle on observera, sur une modalité, une concentration d'au moins 90% des observations, sera considérée comme non pertinente. En procédant ainsi, nous avons pu identifier 7 variables. Elles ont toutes été supprimées de la base d'étude. C'est le cas de la variable code-fractionnement, qui regroupe près de 97% des observations sur la modalité mensuel.

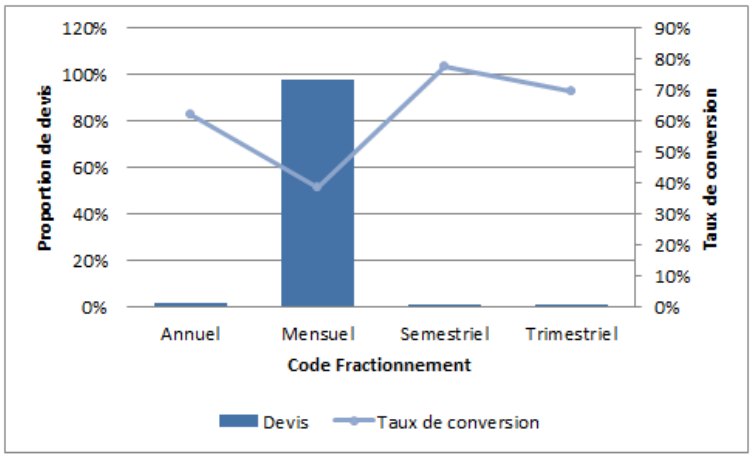


FIGURE 4.1 – Cas du code fractionnement

Comme précisé lors de la présentation des données, nous allons réaliser nos modèles de régression logistique sur le logiciel EMBLEM. Concernant le format des données, l'utilisation d'EMBLEM impose deux principales restrictions :

- toutes les variables explicatives doivent être transformées en des variables catégorielles ;
- le nombre maximal de modalités pour une variable est de 250.

Ces restrictions sont à l'origine de la création des classes pour nos indicateurs de compétitivité. Nous avons également changé le format des variables continues comme l'âge du conducteur, l'âge du véhicule.

Enfin, rappelons que nos travaux sur le modèle de conversion sont à intégrer dans un projet appelé **Scenario Testing**. Le cahier de charges définit par le Groupe impose d'utiliser le package python *scikit-learn* pour construire nos modèles machine learning. L'une des contraintes de l'utilisation de ce package est la « dichotomisation » des variables catégorielles de la base d'étude. **Dichotomiser** une variable catégorielle consiste à créer autant de variables binaires qu'il y a de modalités pour cette variable catégorielle. Prenons l'exemple de la variable enfant en âge de conduire (TP_ENF). Sur les 10 premières observations de notre base d'étude, on obtient :

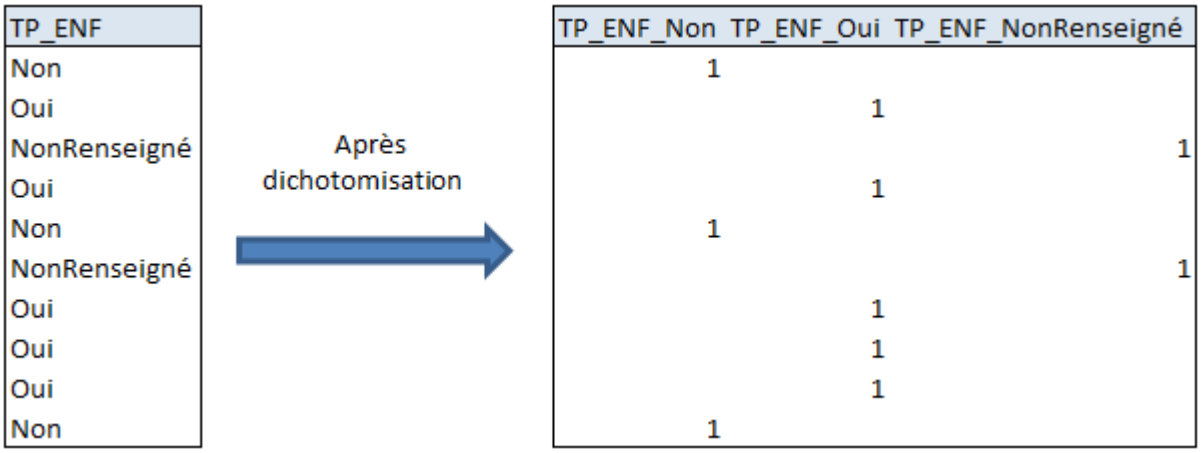


FIGURE 4.2 – Processus de dichotomisation

En appliquant la dichotomisation des variables catégorielles, on passe d'une base d'étude avec 76 colonnes à une base de modélisation avec 1500 colonnes.

4.2 Régression logistique

4.2.1 Corrélation entre les variables

Pour évaluer l'intensité du lien qu'il y a entre deux variables, nous nous appuyons sur le V de Cramer. L'un des avantages est qu'il peut être calculé sur Emblem. Il est donné par la formule :

$$V = \sqrt{\frac{\chi^2}{n(\min(l,c)-1)}}$$

Où :

- χ^2 est la statistique du test de Khi-deux ;
- l et c correspondent au nombre de modalités des deux variables ;
- n est la taille de l'échantillon.

Le V de Cramer est un indicateur qui prend ses valeurs dans l'intervalle [0 ; 1]. De faibles valeurs indiquent que les variables sont indépendantes. Sur le tableau suivant, vous verrez que lorsque le V de Cramer est supérieur à 0,6, les variables sont fortement correlées.

Valeur V_Cramer	Force du lien statistique
0	Absence de lien
$]0;0,1[$	Très faible
$[0,1;0,3[$	Faible
$[0,3;0,6[$	Modérée
$[0,6;0,8[$	Forte
$[0,8;1]$	Colinéarite

TABLE 4.1 – Seuils de V de Cramer

Pour les besoins du modèle de régression logistique, nous devons identifier les couples de variables fortement liées ; c'est-à-dire ayant un V de Cramer supérieur à 0,6. Le nombre de couples identifiés étant assez grand. Dans le tableau suivant nous avons répertorié quelques uns. Si deux variables corrélées sont significatives pour notre modèle, on retiendra celle qui minimise l'AIC.

Variable 1	Variable 2	V_Cramer
Alimentation	Boîte de vitesse	0.766
Alimentation	Energie	0.664
Formule	Franchise bris de glace	0.719
Franchise bris de glace	Franchise vol	0.67
Groupe SRA	Vitesse maximale	0.702
Région	Zone bris de glace	0.632

TABLE 4.2 – Exemple de variables fortement corrélées

4.2.2 La modélisation sous EMBLEM

Pour réaliser un modèle sur EMBLEM, nous avons précisé à la section 2.1 qu'il est nécessaire de créer au préalable deux fichiers : le fac et le bid. Sous SAS, nous avons réalisé l'ensemble des retraitements de données nécessaires pour pouvoir créer ces fichiers et les utiliser pour la régression logistique.

Pour appuyer le fait que notre choix se porte sur le logiciel EMBLEM pour cette modélisation, nous avons évoqué l'aspect pratique de celui-ci. En effet, après la lecture des données, il suffit de lui préciser quel type de modèle on veut réaliser, ainsi que les mesures de qualité de modèle. Un autre avantage est qu'on peut facilement observer la distribution des coefficients par modalité, sur chaque variable introduite dans le modèle. C'est aspect visuel est très parlant et, on s'appuie souvent dessus pour faire les regroupements de modalités. Nous illustrons nos propos par le graphique suivant, sur lequel vous observez d'une part un diagramme à bandes, représentant le nombre de devis par modalité, et d'autre part, les courbes d'évolution des coefficients (en orange) et des coefficients lissés (en vert). Pour lisser les coefficients, le logiciel propose un ajustement polynomial, dont le degré est à définir par l'utilisateur. Il fait ensuite le calcul des plus proches voisins et propose l'ajustement qui convient le mieux. Dans notre cas, il s'agit d'un polynôme de degré 4. Les fluctuations des coefficients sont mieux maitrisées.

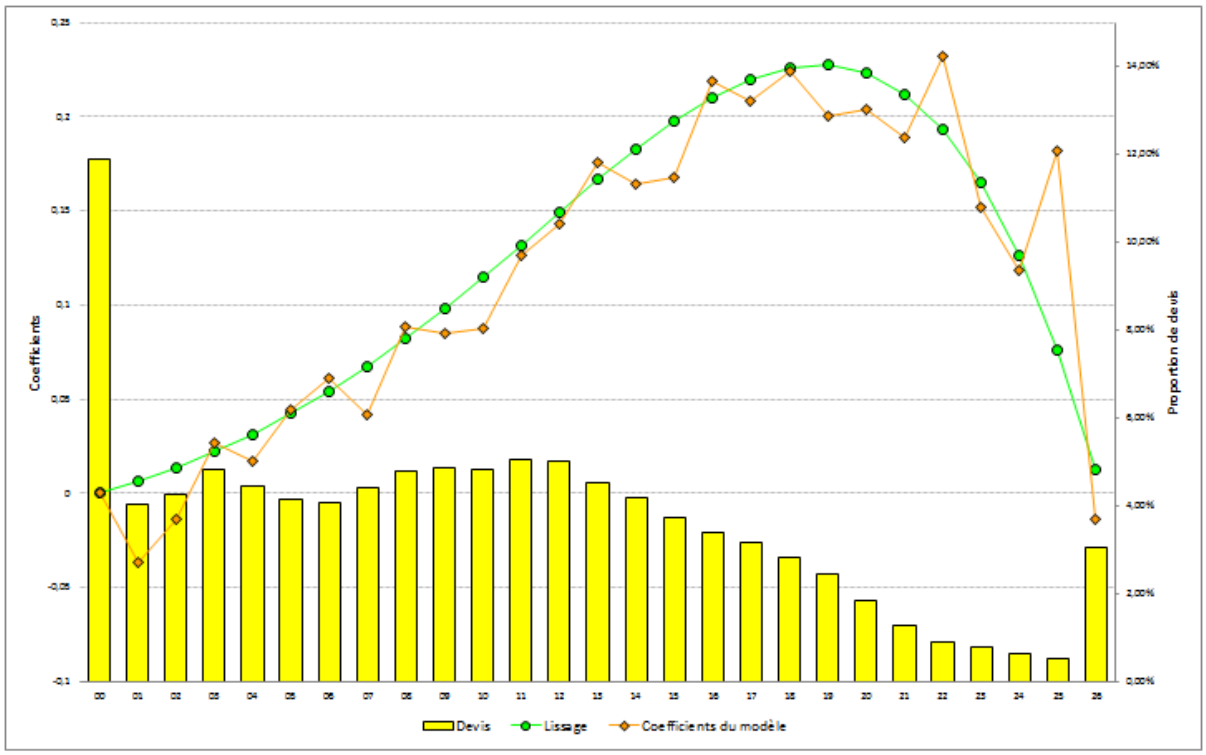


FIGURE 4.3 – Evolution des coefficients pour la variable ancienneté de permis

Premièrement, nous réalisons un modèle sans les indicateurs de compétitivité.[5] La méthode de sélection de variables que nous utilisons est la méthode backward. En introduisant l'ensemble des variables dans le modèle, nous sommes confrontés à une difficulté : le modèle ne converge pas. Ceci est dû au fait qu'il comporte un grand nombre de variables corrélées. Pour résoudre ce problème, nous identifions le couple de variables ayant le V de Cramer le plus élevé. Notre idée est d'exclure du modèle, une des variables de ce couple et de relancer le modèle. La variable conservée est celle qui minimise l'AIC. Tant que le modèle ne converge pas, on réitère l'opération.

Après avoir supprimé les perturbations liées aux variables corrélées, la déviance du modèle est de 779 603. Nous avons ensuite appliqué les étapes de la méthode backward, jusqu'à ce que la décroissance de l'AIC ne soit plus significative. 30 variables ont été retenues par le modèle, et il a une déviance de 775 367. L'analyse de la significativité des coefficients nous a conduits à réaliser des regroupements de modalités lorsque les coefficients estimés étaient non significatifs. Dans le tableau suivant, vous trouverez la liste des 10 variables les plus importantes du modèle.

5. Rappelons que les données externes ne sont pas prises en compe dans les modèles sans indicateurs de compétitivité.

Liste des variables significatives
Taux de réduction
Age du conducteur
Ancienneté de permis
Catégorie socioprofessionnelle
Région
Formule
Durée du relevé d'information
Zonier commercial Vol-incendie
Puissance du véhicule
Ancienneté d'acquisition du véhicule

TABLE 4.3 – Les variables les plus significatives du modèle sans les indicateurs de compétitivité

En parallèle, nous effectuons les mêmes opérations sur un jeu de données incluant les indicateurs de compétitivité. Nous avons remarqué qu'il y a des liens forts entre eux.

Premièrement, bien que les distributions soient différentes, l'information sous-jacente est la même. Prenons l'exemple de la moyenne marché (marché incluant Allianz). Lorsqu'on divise cette moyenne marché par la prime Allianz, on construit deux indicateurs : l'un suivant une loi normale et l'autre suivant une loi uniforme. On remarque que ces deux variables sont fortement corrélées. Dans le cadre de la régression logistique, on ne va pas les conserver pour faire un même modèle. Nous allons donc faire un modèle avec les indicateurs distribués de façon uniforme dans un premier temps. Puis, nous ferons un autre modèle avec les indicateurs distribués suivant une loi normale. Nous comparerons les deux modèles et, n'en retiendrons qu'un.

Deuxièmement, le fait de construire les indicateurs en les divisant par la prime Allianz ou par différence, n'empêche pas la forte corrélation entre eux. En effet, considérons l'exemple précédent de la moyenne marché. En la divisant par la prime Allianz, on crée deux indicateurs de compétitivité. On fait pareil lorsqu'on lui soustrait la prime Allianz. On se retrouve donc avec quatre indicateurs fortement corrélés. D'après le paragraphe précédent, on comprend qu'ils ne se retrouveront pas tous dans le modèle. Ensuite, il faudra décider si on conserve le modèle avec les indicateurs construits par division ou celui construit par différence. Pour cela, on se basera sur la déviance. C'est-à-dire que si les deux indicateurs sont significatifs pour le modèle, on retiendra celui qui concourt à réduire la déviance du modèle.

Pour appuyer les paragraphes précédents, nous présentons dans le tableau suivant, les V de Cramer de quelques indicateurs de compétitivité. Sur ce tableau, on peut aussi remarquer que la moyenne marché est fortement liée à la deuxième prime la plus élevée du marché, notée maximum2.

division \ différence		moyenne marché		maximum2*	
		uniforme	normale	uniforme	normale
moyenne marché	uniforme	0,86	0,677	0,753	0,48
	normale	0,874	0,548	0,801	0,604
maximum2*	uniforme	0,662	0,756	0,969	0,969
	normale	0,608	0,561	0,969	0,711

FIGURE 4.4 – Tableau des indicateurs de compétitivité avec V Cramer

Après implémentation, nous avons comparé les quatre modèles construits avec les différents types d'indicateurs (en croisant différence-division et distribution normale-distribution uniforme). Le modèle plus performant est celui qui utilise le croisement différence-normale. Parmi les quatre modèles, c'est celui qui donne la meilleure déviance. Par rapport au modèle sans les indicateurs, on constate que la déviance est meilleure. Le fait d'intégrer les indicateurs de compétitivité baisse la déviance de plus de 7500.

Modèles	Distribution		Forme		Gain d'optimisation	Déviance après optimisation
	normale	uniforme	différence	division		
Modèle 1	X		X		-1,31%	767 862
Modèle 2	X			X	-1,01%	769 854
Modèle 3		X	X		-0,83%	768 046
Modèle 4		X		X	-0,91%	769 836

TABLE 4.4 – Déviances des modèles

Pour notre modèle sans les indicateurs de compétitivité, vous trouverez ci-dessous la liste des 10 variables les plus significatives.

6. Pour un profil, maximum2 désigne la deuxième prime la plus élevée du marché.

Modèle sans les indicateurs	Modèle avec les indicateurs
Taux de réduction	Taux de réduction
Age du conducteur	Age du conducteur
Ancienneté de permis	Catégorie socioprofessionnelle
Catégorie socioprofessionnelle	Ancienneté de permis
Région	Potentiel Retail
Formule	Zonier commercial Vol-incendie
Durée du relevé d'information	Durée du relevé d'information
Zonier commercial Vol-incendie	Région
Puissance du véhicule	Prime moyenne du marché (normale-différence)
Ancienneté d'acquisition du véhicule	Puissance du véhicule

TABLE 4.5 – Récapitulatif des variables significatives des modèles avec et sans les indicateurs de compétitivité

Significativité des variables et coefficients

Pour vérifier la significativité des variables sous EMBLEM, on réalise un test de vraisemblance. Les valeurs des p-value nous indiquent si une variable est significative ou pas. Si la p-value est inférieure à 5%, on considère que la variable est significative. Concernant les coefficients, EMBLEM valide leur significativité si :

$$\text{Std Error}(\%) < 50\%$$

Avec Std Error $= |\frac{\text{Standard Error}}{\text{Valeur du coefficient estimé}}|$
Ceci revient à réaliser un Test de Wald. En effet, soit $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite.

- Les hypothèses du test sont H0 : $\beta_j = 0$ contre H1 : $\beta_j \neq 0$.
- Sous l'hypothèse H0, la statistique $T = \frac{\hat{\beta}_j}{\sqrt{v(\hat{\beta}_j)}}$ suit une loi normale centrée et réduite.
- Appelons T_{obs} la valeur observée de la statistique. Pour que H0 soit acceptée, il faut que $|T_{obs}| \geq q_{1-\frac{\alpha}{2}}$.
- Pour $\alpha = 4,56\%$, la table des quantiles de la loi normale nous donne $q_{1-\frac{\alpha}{2}} = 2$.
  A partir de ce seuil, on peut reconstruire la règle de décision d'EMBLEM :
  Si $|T_{obs}| \geq 2$, alors Std Error$(\%) \leq \frac{1}{2}$
  Par conséquent, H0 est acceptée.

Pour illustrer tout ceci, prenons l'exemple du coefficient bonus-malus. Il s'agit de l'une des variables retenues par le modèle. En effet, pour le test de vraisemblance, sa p-value est inférieure à 5%. Sur le graphique suivant sont représentés : la proportion de devis et le taux de conversion observé moyen par modalité. Le taux de conversion moyen observé correspond au taux de conversion réellement observé pondéré par la proportion de devis associée à chaque modalité.

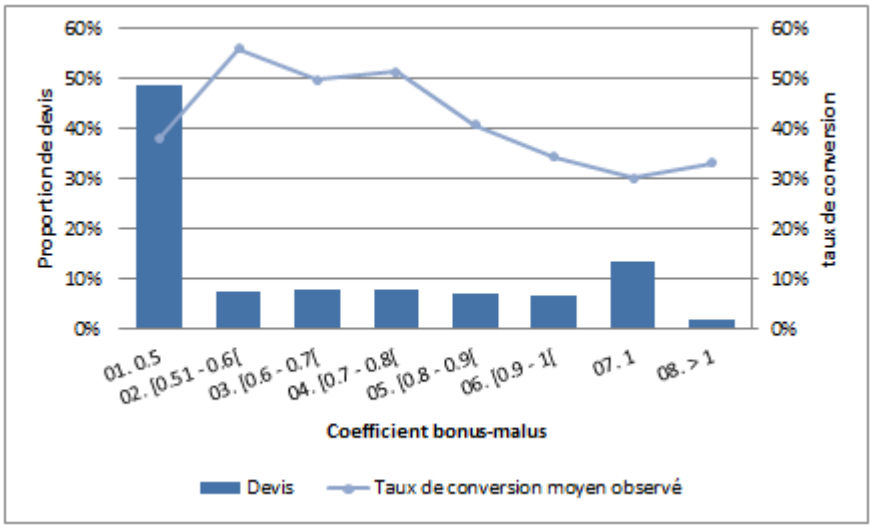


FIGURE 4.5 – Evolution du taux de conversion moyen obsersé en fonction du coefficient bonus-malus

Sur cette variable, le tableau suivant montre que les modalités sont toutes significatives : le test de Wald est concluant.

Modalités	coeff. estimés	StdErr(%)	Devis (%)	Odds-ratio
Ordonnée à l'origine	0,0851	0,0851	100	
0.5			48,70	
$[0,51; 0.6[$	0,18	9,50	7,30	1,20
$[0.6 - 0.8[$	0,08	20,00	15,50	1,08
≥ 0.8	- 0,03	0,4950	28,40	0,97

TABLE 4.6 – Les coefficients de la régression pour la variable coefficient bonus-malus

Interprétation des coefficients

Dans le cas de la régression logistique sur EMBLEM, l'interprétation des coefficients se fait par rapport à la modalité de référence. Ainsi, nous interprèterons les coefficients du modèle de régression comme un rapport de cotes (odds-ratio). Le rapport de chances est

calculé par rapport à la modalité de référence. Pour chaque variable, celle-ci correspond à la modalité ayant le plus grand effectif. Dans l'exemple du tableau précédent, la modalité de référence est 0.5. On constate que le rapport de chances est multiplié par 0,97 lorsque le coefficient bonus-malus est supérieur ou égal à 0,8 et par 1,2 lorsqu'il est dans l'intervalle [0,51 ; 0,6[. Ce qui signifie que relativement à la modalité de référence, on a moins de chances de convertir un devis lorsque le coefficient bonus-malus est supérieur ou égal 0,8. Par contre, on a plus de chances de le convertir si le coefficient bonus-malus est dans l'intervalle [0,51 ; 0,6 [.

Evaluation de la performance

Dans cette section, nous présenterons les résultats de la performance du modèle de régression logistique en nous appuyant sur les métriques introduites à la section 3.5. Nous présenterons d'abord les résultats sur le modèle sans les indicateurs de compétitivité, puis ceux du modèle incluant les indicateurs de compétitivité.

- **Modèle sans les indicateurs de compétitivité**

Le premier outil utilisé pour juger la qualité de notre modèle est la matrice de confusion normalisée. Pour la construire, nous avons fixé le seuil à 0,39, qui correspond au taux de conversion observé sur l'ensemble des données. La matrice de confusion normalisée ci-dessous est construite sur le jeu de données test.

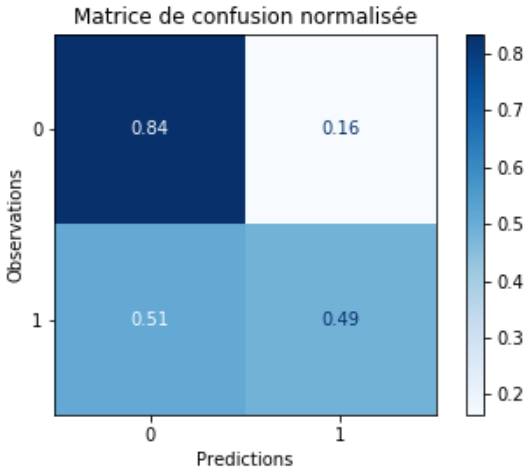


FIGURE 4.6 – Matrice de confusion du modèle sans les indicateurs de compétitivité

De cette matrice de confusion, on peut déduire :

- la sensibilité, qui est de 49% : 49% des devis réellement convertis ont pu être bien identifiés par l'algorithme.

- la spécificité, est de 84% : 84% de devis non convertis ont été bien identifiés par le modèle.

La précision du modèle est de 64,8% : ce qui signifie que sur l'ensemble des devis qui sont estimés convertis, 64,8% le sont réellement. Le taux d'erreur est de 30,1% : ce qui signifie que sur l'ensemble des devis de la base test, 30,1% ont été mal prédits. Soit un taux de bonnes prédictions à 69,9%. L'AUC sur le jeu de données test est de 73,7% et sur les données d'apprentissage, il est de 75%. Comparer l'AUC sur les données d'apprentissage et sur les données test nous permet de vérifier s'il y a eu sur-apprentissage. L'écart entre les deux valeurs étant faible, on conclut qu'il n'y pas eu de sur-apprentissage.

- **Modèle avec les indicateurs de compétitivité**

Le seuil de 0,39 est également retenu pour construire la matrice de confusion.

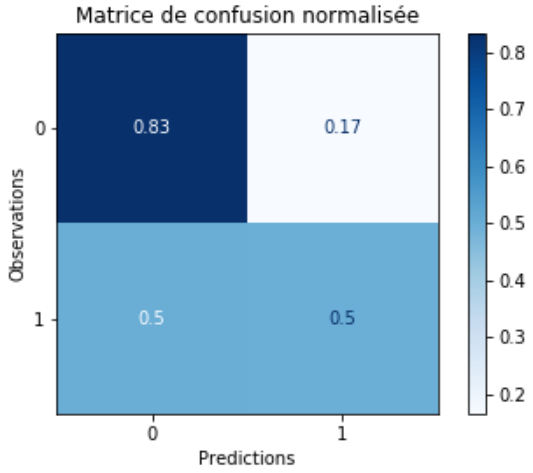


FIGURE 4.7 – Matrice de confusion du modèle avec les indicateurs de compétitivité

On en déduit que :

- la sensibilité est de 50% : la moitié des devis réellement convertis ont pu être identifiés par l'algorithme. Qu'il s'agisse du modèle avec ou de celui sans les indicateurs de compétitivité, la sensibilité est d'environ 0,5. Ce qui suppose que le modèle a une chance sur 2 de mal prédire un devis qui est réellement converti.
- la spécificité est de 83% : 83% de devis non convertis ont été bien identifiés par le modèle. La spécificité de ce modèle est légèrement inférieure à celle du précédent. En introduisant les indicateurs de compétitivité, on a légèrement amélioré la capacité du modèle à bien prédire des devis convertis, au détriment de sa capacité à prédire devis non convertis.

Le modèle présente une précision de 66% : c'est-à-dire que sur l'ensemble des devis qui sont supposés convertis par l'algorithme, 66% le sont réellement. C'est un point de plus que la précision du modèle précédent. Si l'algorithme estime qu'un devis est converti, les chances qu'il le soit réellement sont plus grandes que celles du modèle précédent. Le taux d'erreur est de 29,7% : ce qui signifie que sur l'ensemble des devis de la base test, 29,7% ont été mal prédits. Soit un taux de bonnes prédictions à 70,3%. Par rapport au modèle précédent, on a moins de chance de se tromper lors de la prédiction de l'acte de conversion. L'AUC sur le jeu de données test est de 74 ,6% et sur les données d'apprentissage, il est de 75,5%. Par conséquent, il n'y a pas eu de sur-apprentissage. En se basant sur l'AUC et le taux d'erreur, le modèle avec les indicateurs de compétitivité a un léger avantage, comme le montre la figure suivante.

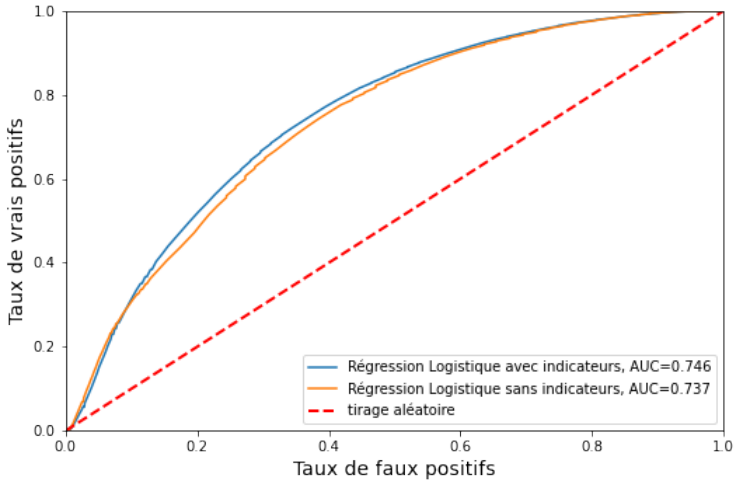


FIGURE 4.8 – Courbes ROC des modèles de régression logistique

4.3 Modèles machine learning

Rappelons que les méthodes de machine learning utilisées pour la modélisation du taux de conversion sont le grandient boosting et les forêts aléatoires (random forest). Elles ont toutes deux été implémentées sur Python, et le principe de modélisation est le même :

- on scinde les données en deux échantillons : données d'apprentissage et données test ;
- ensuite, pour implémenter le modèle sur les données d'apprentissage, on définit des valeurs initiales pour les hyperparamètres ;
- on observe les résultats sur les données test ;
- puis, on procède à une optimisation des hyperparamètres des modèles ;
- enfin, on retient les paramètres optimaux pour construire notre modèle final. Dans la suite, nous présenterons d'abord l'initialisation des modèles. Puis viendront le processus d'optimisation des paramètres et les résultats obtenus.

Il faut également noter que les modèles machine learning ne posent pas de restriction en cas de corrélation entre les variables. Par conséquent, les modèles incluant les indicateurs de compétitivité intègrent l'ensemble des indicateurs, ainsi que les données externes.

La significativité des variables sera matérialisée par leur contribution (ou importance) dans la construction des arbres qui constituent les modèles. Une variable ayant un fort pouvoir discriminant, sera très souvent utilisée dans la construction des arbres. Ce qui va accroitre sa contribution. Les variables ayant les contributions les plus élevées ont un rôle majeur sur la qualité du modèle.

Rappelons que nous avons procédé à une dichotomisation de notre base d'étude pour construire notre base de modélisation. Par conséquent, pour retrouver la contribution globale de chaque variable catégorielle X contenue dans la base d'étude, nous avons procédé à une agrégation des contributions des différentes variables dichotomiques créées à partir de X. C'est-à-dire :

Supposons que la variable X de notre base d'étude a trois modalités : mod1, mod2 et mod3. Suite à la dichotomisation, X engendrera 3 variables dans la base de modélisation. On les appellera X_mod1, X_mod2 et X_mod3. Dans la construction d'un modèle de machine learning, chacune de ces variables aura sa propre contribution. La contribution totale de la variable X sera la somme des contributions des variables X_mod1, X_mod2 et X_mod3.

4.3.1 Random Forest

Optimisation du modèle

Dans le cas du random forest, les hyperparamètres que nous comptons optimiser sont le nombre d'arbres à entrainer et la profondeur maximale de chaque arbre. Pour la robustesse des modèles, nous optons pour une validation croisée avec k=5.

Comme valeurs initiales, nous fixons le nombre d'arbres à 100 et la profondeur maximale à 500. Si le nombre d'arbres est grand, cela contribue à rendre le modèle plus robuste par contre avec des arbres trop profonds, on s'expose à du sur-ajustement.

Rappelons que pour les besoins de la modélisation, nous avons transformé nos variables catégorielles en variables dichotomiques. Par conséquent ce qui était une modalité pour une variable catégorielle de notre base d'étude, devient une variable explicative pour notre base de modélisation.

Dans le cas du modèle sans les indicateurs de compétitivité, nous avons un AUC de 79,5% et un taux d'erreur de 27,9%. Le modèle avec indicateurs de compétivité est légèrement meilleur, car il a un taux d'erreur de 27,6%. Vous trouverez dans le tableau ci-dessous la liste des facteurs les plus disciminants de nos deux modèles.

Variables	Modalités	Contribution
Taux de réduction	0%	4,42%
PSC	/	2,87%
Mois d'enregistrement du devis	/	1,86%
Taux client	<-10%	1,59%
Nombre d'enfants	Non Renseigné	1,53%
Statut marital	Inconnu	1,39%
Taux client	-10%	1,16%
Coefficient bonus-malus	/	0,93%
Enfants en âge de conduire	Non	0,88%
Ancienneté d'acquisition du véhicule	[1;12[	0,82%

Variables	Modalités	Contribution
Taux de réduction	0%	4,06%
PSC	/	2,20%
Potentiel Auto	/	1,68%
Potentiel Retail	/	1,67%
Mois d'enregistrement du devis	/	1,45%
Nombre d'enfants	Non Renseigné	1,35%
Taux client	<-10%	1,33%
Statut marital	Inconnu	1,17%
Taux client	-10%	1,03%
Enfants en âge de conduire	Non	0,77%

FIGURE 4.9 – Principaux facteus discriminants des modèles random forest sans les indicateurs (à gauche) et avec les indicateurs (à droite)

L'un des inconvénients de la dichotomisation est que la table utilisée pour la modélisation devient considérablement grande, par rapport à la base d'étude. Ce qui a pour conséquence de rallonger le temps d'exécution de l'algorithme. Pour les deux modèles (avec et sans les indicateurs de compétitivité), nous devons attendre plus de 4 heures pour avoir des résultats. Afin de réduire le temps d'exécution, nous avons décidé d'exclure les variables trop peu significatives et de reprendre la modélisation. Nous avons allégé notre base de modélisation de 15% des variables. Celles-ci ont une contribution totale de 2,9%. Exclure ces variables nous permet de réduire de moitié le temps d'exécution du programme.

Le premier modèle auquel nous nous intéressons est celui qui exclut les indicateurs de

compétitivité. Nous avons ensuite agrégé les variables de notre base de modélisation afin de retrouver à quelles variables elles correspondent dans notre base d'étude. Nous pouvons alors ordonner les variables en fonction de leur contribution :

Ordre d'imp.[7]	Variables	Contribution
1	Taux de réduction	7,388%
2	Novice[8]	4,494%
3	PSC	2,791%
4	Formule	2,788%
5	Zonier commercial Vol-Incendie	2,739%
6	Zonier commercial RC	2,545%
7	Nombre d'enfants[9]	2,540%
8	Zonier commercial Dommage	2,448%
9	Novice*Age_Conducteur*Age_Permis	2,339%
10	Région	2,134%

TABLE 4.7 – Les variables significatives du modèle sans les indicateurs de compétitivité

Le modèle est relativement bon, car il présente un AUC de 79,8% et un taux d'erreur de 27,7%. Ce qui signifie que globalement, plus de 7 devis sur 10 sont bien classés par l'algorithme ; une meilleure performance que les modèles de régression logistique.

Concernant le modèle avec les indicateurs de compétitivité, nous l'avons construit en utilisant toutes les variables du modèle précédent, auxquelles ont été rajoutés les indicateurs de compétitivité et les données externes. Les valeurs initiales des hyperparamètres sont les mêmes que celles du modèle précédent. Nous obtenons un AUC de 80,3% et un taux d'erreur de 26,9%. Après agrégation, nous observons les contributions des indicateurs de compétitivité et des données externes.

Bien qu'elles ne figurent pas parmi les variables les plus significatives, ces 24 variables ont une contribution totale de 19,418%. Pour la plupart, elles ont une contribution assez faible. Seules 6 d'entre elles ont des contributions de plus de 1%. Elles sont répertoriées dans le tableau suivant.

7. Ordre d'imp. : ordre d'importance des variables dans le modèle.
8. Novice : conducteur dont l'ancienneté de permis est inférieure à 3 ans.
9. Nombre d'enfants du conducteur principal.

Ordre d'imp.	Variables	Contribution
15	Potentiel Auto	1,683%
16	Potentiel Retail	1,668%
26	Prime la moins élevée (rapport - uniforme)	1,366%
32	Prime la moins élevée (différence - uniforme)	1,200%
38	Moyenne marché yc Allianz (différence - normale)	1,047%
39	Moyenne marché hors Allianz (rapport - normale)	1,030%

TABLE 4.8 – les indicateurs les plus significatifs du random forest

Maintenant, procédons à la phase d'optimisation des modèles :

Appelons n_{est} le nombre d'arbres à entrainer et max_d la profondeur maximale des arbres. Optimiser notre modèle random consiste à trouver le couple $(n_{est}\,;\,max_d)$ qui nous donne les meilleures performances sur les données test. Sur Python, nous utilisons la fonction RandomizedSearchCV, pour le faire. Pour utiliser cette fonction, nous devons définir des ensembles de valeurs qui seront utilisées pour construire des modèles et nous déterminerons ensuite quel est le couple $(n_{est}\,;\,max_d)$ optimal : celui qui maximise l'AUC. Concrètement, nous avons recherché :

- des valeurs dans l'intervalle $[50\,;\,500]$ pour le nombre d'arbres ;
- des valeurs dans l'intervalle $[50\,;\,1000]$ pour la profondeur.

On construit ainsi une multitude de modèles, en croisant les différentes possibilités. Ensuite on détermine le modèle le plus performant. On trouve alors que le nombre optimal d'arbres est 320 et la profondeur optimale est 150. Le couple de paramètres optimaux est le même, qu'il s'agisse du modèle avec les indicateurs de compétitivité ou de celui sans. Dans la section suivante, nous présentons les mesures de qualité de modèles, relatives aux modèles optimaux. Par la suite, nous avons ressorti les variables significatives de nos modèles (avec ou sans indicateurs de compétitivité) et nous les avons agrégées. Le tableau suivant donne les 10 variables les plus significatives pour nos modèles avec et sans indicateurs de compétitivité.

Variables	Contribution
Taux de réduction	8,388%
Novice	4,294%
PSC	2,87%
Formule	2,831%
Zonier commercial Vol-Incendie	2,759%
Zonier commercial RC	2,745%
Nombre d'enfants	2,654%
Zonier commercial Dommage	2,572%
Novice*Age_Conducteur*Age_Permis	2,558%
Région	2,543%

Variables	Contribution
Taux de réduction	7,421%
Novice	3,365%
Nombre d'enfants	2,307%
PSC	2,205%
Formule	2,186%
Statut marital	2,183%
Zonier commercial Vol-Incendie	2,101%
Zonier commercial RC	2,084%
Novice*Age_Conducteur*Age_Permis	2,034%
Zonier commercial Dommage	1,965%

FIGURE 4.10 – Les variables les plus significatives des modèles random forest sans les indicateurs (à gauche) et avec les indicateurs (à droite)

Sans surprise, les tableaux ci-dessus nous montrent que les modèles avec et sans les indicateurs de compétitivité ont en commun de nombreuses variables significatives. Bien que les potentiels auto et retail apparaissent comme des facteurs fortement discriminants de la base de modélisation, ils ont une importance moins grande après agrégation.

Toutefois, nous constatons que l'environnement géographique du risque a un impact important dans nos modèles. En effet, nous retrouvons, dans les deux cas, les zoniers commerciaux (Vol-incendie et Responsabilité Civile) et la région parmi les variables significatives. L'acte de conversion est donc fortement lié à la situation géographique du risque. Aussi, le fait de retrouver les potentiels auto et retail parmi les variables significatives de la base de modélisation, nous laisse penser que les habitudes de consommation des assurés dans une zone géographique, sont des facteurs que l'on pourrait exploiter pour mieux comprendre l'acte de conversion. De plus, le nombre d'agences du concurrent 3 et le nombre d'agences d'Allianz sont respectivement les 13è et 15è (sur 1273) facteurs les plus discriminants. Ce qui signifie que la représentativité d'un assureur dans une zone géographique influence la décision d'un prospect.

D'autre part, nous remarquons que le taux de réduction est la variable la plus significative des deux modèles. Elle compte pour plus de 7% dans chacun des modèles. Le taux de réduction est souvent appliqué dans le but de proposer un tarif attrayant pour le prospect. Compte tenu de la concurrence rude sur le marché, jouer sur les prix a une grande importance pour pouvoir attirer les prospects. En effet, pour des niveaux de garanties équivalents, le demandeur d'assurance aura tendance à se tourner vers l'assureur qui propose le prix le moins élevé. C'est potentiellement pour cette raison que le PSC et le mois sont également des facteurs discriminants. Le lien entre le taux de réduction et le PSC vient du fait que proposer une réduction de tarif a un effet immédiat sur la prime payée par l'assuré (actual price) et par extension sur le PSC.

Quant à la relation entre le taux de réduction et le mois, elle découle de la politique de commercialisation du produit. En effet, nous soulevons deux points : les mises à jour des tarifs chaque année (au moins une fois par an au mois d'avril) et les campagnes de marketing (proposer des mois d'assurance gratuits, proposer une réduction pour les étudiants, ...) dans le but d'attirer plus de clientèle. Donc en fonction du mois, des gestes commerciaux sont applicables sur les tarifs.

Enfin, il nous semblait cohérent de retrouver l'ancienneté de permis (reflétée par la variable novice) et la formule, comme variables significatives de nos modèles. En effet, les besoins d'assurance des prospects évoluent dans le temps. Par exemple, pour un conducteur peu expérimenté, il est recommandé de souscrire un contrat d'assurance tous risques. Ce qui implique que la demande d'assurance est influencée par ces deux facteurs. Donc, le taux de conversion est fonction non seulement des prix proposés par les acteurs de marché, mais également de la demande des prospects.

Evaluation de la performance

Dans la section précédente, nous avons montré comment ont été construits nos différents modèles random forest (avec et sans les indicateurs de compétitivité). Nous avons mis en pratique notre démarche pour optimiser les paramètres, et présenté les principaux facteurs discriminants. Dans cette section, nous voulons présenter l'ensemble des métriques que nous avons utilisées pour juger de la qualité des modèles.

- **Modèle sans les indicateurs de compétitivité**

Vous trouverez ci-dessous la matrice de confusion normalisée de notre modèle random forest sans les indicateurs de compétitivité :

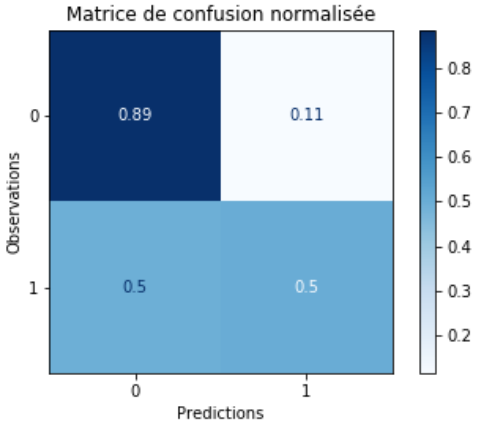


FIGURE 4.11 – Matrice de confusion du modèle sans les indicateurs de compétitivité

De cette matrice, on peut déduire :

- la spécificité du modèle qui est de 89% : c'est-à-dire que sur 10 devis non convertis observés, l'algorithme arrive à bien identifier à peu près 9.
- la sensibilité du modèle qui est de 50% : c'est-à-dire que sur l'ensemble des devis réellement convertis, 50% sont bien classés par l'algorithme.

La précision du modèle est de 75,1%. Ce qui signifie que sur l'ensemble des devis estimés convertis, 75,1% le sont réellement. Enfin, le taux d'erreur du modèle est de 25,2% (Accuracy de 74,8%). C'est-à-dire que 1 devis sur 4 est mal classé par l'algorithme.

Pour vérifier qu'il n'y a pas eu de sur-apprentissage, nous comparons l'AUC de la base test à celui de la base d'apprentissage. Si l'écart entre les deux valeurs est négligeable, on considère qu'il n'y a pas eu de sur-apprentissage. Dans le cas du modèle random forest sans les indicateurs de compétitivité, l'AUC de la base d'apprentissage 81,7%. C'est 1,2 point de plus que celui de la base test. On conclut alors qu'il n'y a pas eu de sur-apprentissage.

- **Modèle avec les indicateurs de compétitivité**

Vous trouverez ci-dessous la matrice de confusion normalisée de notre modèle random forest avec les indicateurs de compétitivité :

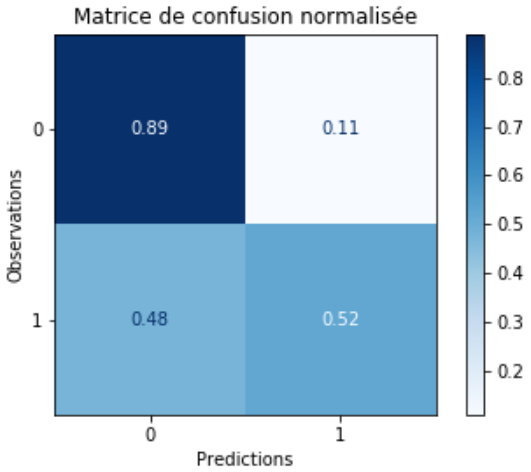


FIGURE 4.12 – Matrice de confusion du modèle avec les indicateurs de compétitivité

De cette matrice, on peut déduire :

- La spécificité du modèle qui est de 89% : c'est-à-dire que sur 10 devis non convertis observés, l'algorithme arrive à bien identifier à peu près 9. Cette métrique est identique à celle du modèle sans les indicateurs de compétitivité.

- La sensibilité du modèle qui est de 52% : c'est-à-dire que sur l'ensemble des devis réellement convertis, 52% sont bien classés par l'algorithme. Sur ce point, le modèle est meilleur que celui sans les indicateurs de compétitivité.

La précision du modèle est de 75,7%. Ce qui signifie que sur l'ensemble des devis estimés convertis par le modèle, 75,7% le sont réellement. Enfin, le taux d'erreur du modèle est de 24,8%. C'est-à-dire que 75,2% des devis (convertis ou non) sont bien classés par l'algorithme. Nous vérifions une fois de plus qu'il n'y a pas eu de sur-apprentissage, en comparant les AUC sur les jeux de données d'apprentissage et test. Sur les données d'apprentissage, l'AUC est de 82,8% tandis que sur les données test, il est de 81,2%. L'écart étant faible, nous en déduisons qu'il n'y pas eu de sur-apprentissage. Nous représentons ci-dessous les courbes ROC des modèles random forest avec et sans indicateurs de compétitivité. Cette figure permet de visualiser la légère différence en termes d'AUC (sur l'échantillon test) pour ces deux modèles. En effet, on remarque que les deux courbes sont quasiment superposées.

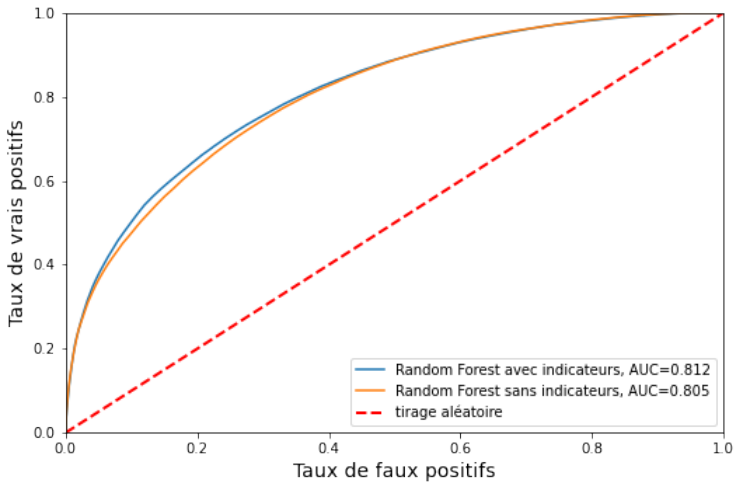


FIGURE 4.13 – Courbes ROC des modèles random forest

Bien que les différentes métriques présentées vont en faveur du modèle avec les indicateurs de compétitivité, nous constatons tout de même que le gain apporté par ces variables est faible. D'un point de vue opérationnel, ce modèle ne sera pas privilégié car il nécessite des moyens supplémentaires, pour un apport négligeable.

4.3.2 Gradient Boosting

Optimisation du modèle

Dans le cas du gradient boosting, les hyperparamètres que nous avons optimisés sont le nombre d'arbres à entrainer, la constante d'apprentissage (notée γ) et la profondeur maximale de chaque arbre. Pour la robustesse des modèles, nous optons pour une validation croisée, avec k=5. Comme pour le random forest, nous fixons initialement le nombre d'arbres à 100 et la profondeur maximale à 500. La valeur initiale du taux d'apprentissage est de 0,1. Le gradient boosting, comme le random forest, s'appuie sur la construction d'arbres élagués. Par conséquent, il faut faire attention aux valeurs des hyperparamètres, pour éviter le sur-ajustement.

Rappelons que le principe du boosting appliqué aux arbres CART consiste à construire plusieurs arbres les uns après les autres, en corrigeant sur un arbre l'erreur de prédiction commise sur le précédent. Ces arbres ne sont donc pas indépendants, contrairement au random forest. De ce fait, le temps d'exécution de l'algorithme est significativement plus long que celui d'un random forest. A cette complexité, s'ajoute celle liée à la dichotomisation des variables qui a tendance à augmenter le temps de calcul. Donc pour gagner en temps de calcul, nous reprendrons la même démarche que précédemment, c'est-à-dire :

1. construire un modèle initial, sans les indicateurs de compétitivité ;
2. agréger les variables et déterminer leur contribution ;
3. identifier les variables les moins significatives et les exclure de la modélisation ;
4. construire un nouveau modèle sans les indicateurs de compétitivité ;
5. construire un modèle avec l'ensemble des variables du modèle en 4., et y ajouter les indicateurs de compétitivité et les données externes.

Pour notre modèle initial sans les indicateurs de compétitivité, nous avons un AUC de 81,4% et un taux d'erreur de 24%. Les variables les plus significatives sont présentées dans le tableau suivant :

Variables	Modalités	Contribution
Taux de réduction	0%	11,118%
Nombre d'enfants en âge de conduire	Non Renseigné	8,315%
PSC	/	5,275%
Mois d'enregistrement du devis	/	3,369%
Ancienneté d'acquisition du véhicule[10]	[1 ; 12[	1,476%
Coefficient bonus-malus	/	1,264%
Taux de réduction	[-6% ; -5%]	0,769%
Age du véhicule à l'acquisition[10]	>=96	0,627%
Durée du relevé d'information[10]	<12	0,601%
Mode d'achat	Comptant	0,593%

TABLE 4.9 – Les facteurs les plus discriminants du modèle GBM sans indicateurs de compétitivité

On constate que tout comme pour le random forest, le taux de réduction, le PSC et le mois d'enregistrement du devis font partie des facteurs les plus discriminants dans l'explication de l'acte de conversion. Par rapport au random forest, quatre nouveaux facteurs se démarquent. Il s'agit de l'ancienneté d'acquisition du véhicule, son âge à l'acquisition, le mode d'achat du véhicule ainsi que la durée du relevé d'information. Pour ces deux dernières variables, nous ne saurons donner a priori une explication qui justifierait leur significativité. Par contre, il est plus facile pour nous de trouver un sens au fait que l'âge du véhicule influence la décision du prospect. En effet, l'âge du véhcule oriente souvent dans le choix de la formule souscrite et par extension sur la gamme de prix à laquelle l'assuré serait prêt à souscrire son contrat. Pour des véhicules plus anciens, les coûts de réparation sont souvent faibles. Par conséquent les prospects ont tendance à choisir une assurance autre que la "tous risques", qui est jugée trop coûteuse pour ce type de véhicules. Donc avoir un tarif compétitif sur ces formules peut avoir de l'importance lorsqu'on approche un prospect.

Après avoir agrégé nos variables, nous avons constaté que près de 20% des variables sont peu significatives. Elles ont une contribution totale de près de 4,6%. Nous décidons d'alléger le modèle, en retirant ces variables de la modélisation. Nous obtenons ainsi un nouveau modèle avec un AUC de 81% et un taux d'erreur de 24,7%. Contrairement au random forest, alléger la base de modélisation n'a pas amélioré la qualité du modèle. Toutefois, on gagne considérablement en temps d'exécution.

Avec une contribution de plus de 12%, le taux de réduction est la variable la plus significative de notre modèle gradient boosting (sans les indicateurs), comme c'était le cas pour le random forest. Ensuite viennent le nombre d'enfants et le PSC. Ces variables figuraient

10. L'ancienneté d'acquisition, l'âge du véhicule à l'acquistion et la durée du relevé d'information sont donnés en mois.

déjà parmi les variables les plus significatives du modèle random forest. Dans le tableau ci-dessous, sont listées les dix variables les plus significatives du modèle gradient boostiong (sans indicateurs). On y retrouve quasiment les mêmes que celles du modèle random forest (sans indicateurs) mais avec des niveaux différents de contribution. Ce constat nous ne semble pas illogique, car les algorithmes gradient boosting et random forest s'appuient tous les deux sur la construction d'arbres de classification.

Ordre d'imp.	Variables	Contribution
1	Taux de réduction	12,03%
2	Nombre d'enfants	8,701%
3	PSC	5,304%
4	Mois d'enregistrement du devis	3,370%
5	Novice	2,888%
6	Zonier commercial Vol-Incendie	2,527%
7	Région	2,400%
8	Zonier commercial RC	2,199%
9	Novice*Age_Conducteur*Age_Permis	2,197%
10	Coeffcient bonus-malus	2,158%

TABLE 4.10 – Les variables les plus significatives du modèle GBM sans indicateurs de compétitivité

Intéressons nous maintenant au modèle GBM incluant les indicateurs de compétitivité. Il est initialisé avec les mêmes hyperparamètres que le GBM sans indicateurs. Ce nouveau modèle a un AUC de 81,7% et un taux d'erreur de 24,3%. Les indicateurs de compétitivité et les données externes cumulent une contribution de 16,367%, Ce qui représente un peu plus de 3 points en moins par rapport au random forest. seulement 4 de ces variables ont une contribution supérieure à 1%. Toutefois, on note que les potentiels auto et retail font partie des variables les plus significatives du modèle. Ce sont respectivement les 7è et 8è variables avec les contributions les plus élevées.

Ordre d'imp.	Variables	Contribution
7	Potentiel Auto	2,185%
8	Potentiel Retail	2,054%
25	Prime la moins élevée du marché (rapport-uniforme)	1,102%
32	Prime la moins élevée du marché (différence-uniforme)	1,006%

TABLE 4.11 – Les indicateurs les plus significatifs du modèle gradient boosting

Maintenant, procédons à la phase d'optimisation des modèles :
Le principe d'optimisation est semblable à celui de la méthode random forest. Il s'agira,

à l'aide de la fonction RandomizedSearchCV, de déterminer le triplet (n_{est} ; max_d, γ) qui maximise l'AUC. La grille de recherche est la suivante :

- des valeurs dans l'intervalle [0,05 ; 0,5] pour la constante d'apprentissage ;
- des valeurs dans l'intervalle [50 ; 500] pour le nombre d'arbres ;
- des valeurs dans [50 ; 500] pour la profondeur des arbres.

On construit ainsi plusieurs modèles et, on détermine le plus performant. La valeur optimale pour le nombre d'arbres est 300, celle de la profondeur est de 150 et la constante d'apprentissage optimale est 0,09. Les paramètres optimaux sont les mêmes, qu'il s'agisse du modèle avec les indicateurs de compétitivité, ou de celui sans. Par la suite, nous avons ressorti les variables significatives de nos modèles (avec ou sans indicateurs de compétitivité) et nous les avons agrégées. Le tableau suivant donne les 10 variables les plus significatives pour nos modèles avec et sans indicateurs de compétitivité.

Variables	Contribution
Taux de réduction	12,406%
Nombre d'enfants	8,635%
PSC	5,275%
Mois d'enregistrement du devis	3,369%
Novice	2,883%
Zonier commercial Vol-Incendie	2,472%
Région	2,439%
Zonier commercial RC	2,313%
Novice*Age_Conducteur*Age_Permis	2,197%
Coefficient bonus-malus	2,161%

Variables	Contribution
Taux de réduction	11,834%
Nombre d'enfants	8,214%
PSC	4,205%
Novice	2,393%
Formule	2,235%
Potentiel Auto	2,095%
Potentiel Retail	2,073%
Ancienneté d'acquisition du véhicule	1,855%
Zonier commercial Vol-Incendie	1,855%
Région	1,831%

FIGURE 4.14 – Les variables les plus significatives des modèles gradient boosting sans les indicateurs (à gauche) et avec les indicateurs (à droite)

On remarque que ces variables ont un poids considérable dans la construction des modèles. Elles ont une contribution totale de plus de 44% pour le modèle sans indicateurs et de près de 39% pour le modèle incluant les indicateurs de compétitivité et les données externes. Dans les deux cas, les variables les plus importantes sont le taux de réduction, le nombre d'enfants du conducteur principal et le PSC. Elles apportent une contribution d'au moins 24% dans la construction des différents modèles.

Comme dans le cas des modèles random forest, le taux de réduction et le PSC se démarquent. Ces indicateurs sont liés, et reflètent le niveau de prix et la rentabilité estimée du contrat. Leur importance traduit la réponse du prospect à la stratégie commerciale de l'entreprise.

Certaines informations relatives au conducteur principal sont également influentes dans les deux modèles. Il s'agit ici du nombre d'enfants et du statut de novice. Comme le montre la figure 2.9, les novices représentent une part de la masse assurable sur laquelle nous observons les taux de conversion les plus bas, malgré une forte demande. De par leur manque d'expérience, ils représentent un gros risque pour l'assureur. De ce fait, les tarifs qui leur sont proposés sont souvent au dessus de la moyenne. Il nous semble raisonnable que cette variable soit significative. Quant au nombre d'enfants du conducteur principal, le sens de la significativité peut venir du faire qu'avoir au moins un enfant influence la prime à la hausse. En effet, du point de l'assureur, les enfants représentent des sources de dommages sur le véhicule de l'assuré. L'effet sur la prime étant variable en fonction de l'âge des enfants, ça a du sens de voir que cette information influence la prise de décision.

Dans le cas du modèle sans les indicateurs, une autre variable liée au conducteur se démarque : le coefficient bonus-malus. Cettte variable est le reflet de la fréquence observée de sinistres d'un conducteur. Le coeffcient bonus-malus peut grandement impacter la prime proposée à un potentiel client. Il s'agit d'un élément tellement important que certains assureurs proposent des produits spécifiques pour les clients ayant des coeffients bonus-malus élevés.

Sur le modèle avec indicateurs de compétitivité, on note que le zonier commercial (Vol-Incendie) et la région font partie des variables les plus significatives. Leur importance est toutefois moins grande comparée à ce qu'on obseve sur le modèle sans les indicteurs ou encore sur les modèles random forest. L'importance de ces variables géographiques dans les modèles est selon nous due à deux aspects. Le premier est relatif au fait que le lieu de garage habituel du véhicule a une influence sur la prime. En effet, le risque vol ou incendie n'est pas le même sur l'ensemble du territoire. Des analsyses de sinistralité permettent de classer différentes zones en fonction de leur exposition au risque de vol par exemple. C'est de cette façon que sont conçus les zoniers. Par conséquent, pour deux prospects qui ont en tous points un profil identique, la prime vol sera plus élevée pour celui qui déclare le garage habituel de son véhicule dans une zone plus risquée. Le second aspect est relatif aux habitudes de consommation en assurance dans les diférentes zones géographiques. Si on omet les considérations sur la représentativité des concurrents, nous supposons qu'il y a plus de chance de réaliser des affaires dans des zones à fort potentiel.

Evaluation de la performance

L'objet de la section précédente était de détailler le processus de construction de nos modèles gradient boosting (avec et sans les indicateurs de compétitivité). Nous avons présenté la démarche d'optimisation des paramètres, ainsi que les principaux facteurs discriminants. Dans cette section, nous présenterons l'ensemble des métriques que nous avons utilisées pour juger de la qualité des modèles.

- **Modèle sans les indicateurs de compétitivité**

Vous trouverez ci-dessous la matrice de confusion normalisée de notre modèle gradient boosting sans les indicateurs de compétitivité :

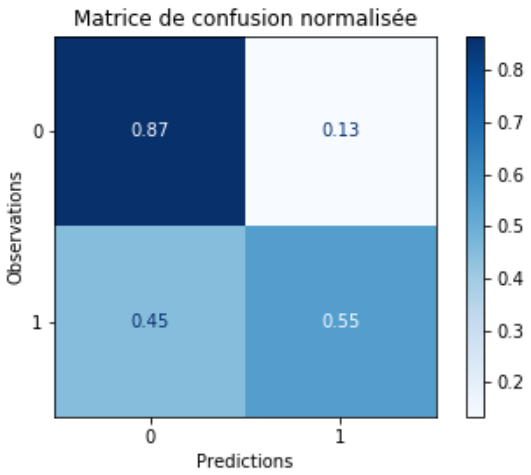


FIGURE 4.15 – Matrice de confusion du modèle sans les indicateurs de compétitivité

De cette matrice, on peut déduire :

- La spécificité du modèle, qui est de 87% : c'est-à-dire 87% des devis non convertis sont bien classés par le modèle.
- la sensibilité du modèle, qui est de 55% : c'est-à-dire que sur 10 devis convertis, plus de 5 sont bien classés par le modèle.

La précision du modèle est de 76%. Ce qui signifie sur l'ensemble des devis estimés convertis, 76% le sont réellement. Enfin, le taux d'erreur du modèle est de 24,1% (Accuracy de 75,9%). Ce qui signifie que lors de la prédiction, le modèle se trompe moins d'une fois sur quatre.

Evalué sur les données test, ce modèle a un AUC de 81,6%. Ce qui est légèrement mieux que les modèles random forest et nettement mieux que les résultats de la régression logistique. L'AUC sur les données d'apprentissage est de 82,9%. Il n'est pas significativement éloigné de l'AUC sur les données test. On en conclut qu'il n'y a pas de sur-apprentissage.

- **Modèle avec les indicateurs de compétitivité**

Vous trouverez ci-dessous la matrice de confusion normalisée de notre modèle gradient boosting avec les indicateurs de compétitivité :

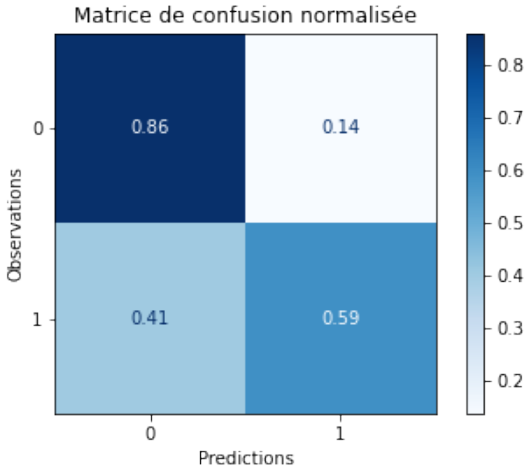


FIGURE 4.16 – Matrice de confusion du modèle avec les indicateurs de compétitivité

De cette matrice, on peut déduire :

- La spécificité du modèle, qui est de 86% : c'est-à-dire 86% des devis non convertis sont bien classés par le modèle.
- la sensibilité du modèle, qui est de 59% : c'est-à-dire que sur 5 devis convertis, à peu près 3 sont bien classés par le modèle.

La précision du modèle est de 78,8%. Ce qui signifie que sur l'ensemble des devis estimés convertis, 78,8% le sont réellement. Enfin, le taux d'erreur du modèle est de 23,2% (Accuracy de 76,8%). Evalué sur les données test, ce modèle a un AUC de 82,5%. Ce qui est légèrement mieux que les modèles random forest et nettement mieux que les résultats de la régression logistique. L'AUC sur les données d'apprentissage est de 84,1%. Il n'est pas significativement éloigné de l'AUC sur les données test. On en conclut qu'il n'y a pas de sur-apprentissage.

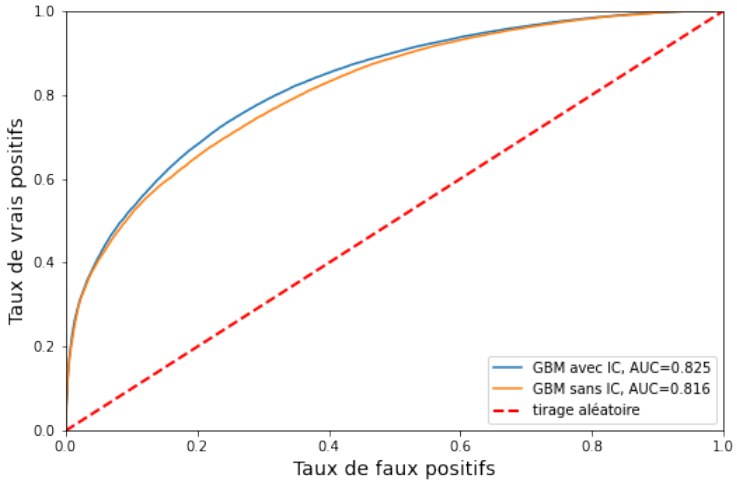


FIGURE 4.17 – Courbes ROC des modèles gradient boosting

Comparé au modèle précédent, ce modèle présente un meilleur AUC et un taux d'erreur plus faible. En observant la courbe ROC des deux modèles, le modèle incluant les indicateurs ne se démarque pas de façon nette. Tout comme pour le modèle random forest, les indiacteurs de compétitivité et les données externes n'ont pas apporté une amélioration significative sur la qualité globale du modèle. Cependant, le gain en terme de précision n'est pas à négliger. En effet, avec ce modèle, on a une meilleure estimation a priori du nombre de devis qui seront convertis : une information importante pour les souscripteurs.

4.4 Récapitulatif des résultats

Dans cette section, nous revenons sur l'ensemble des méthodes utilisées et présentons leurs résultats. Ceci nous permettra de comparer la qualité des résultats obtenus. Dans un premier temps, nous traiterons les modèles sans les indicateurs de compétitivité et ensuite, nous comparerons les résultats des modèles incluant les indicateurs de compétitivité.

4.4.1 Modèles sans les indicateurs de compétitivité

Dans le tableau suivant, sont présentées les dix variables les plus significatives de chaque méthode.

Régression Logistique	Random Forest	Gradient Boosting
Taux de réduction	Taux de réduction	Taux de réduction
Age du conducteur	Novice	Nombre d'enfants
Ancienneté de permis	PSC	PSC
Catégorie socioprofessionnelle	Formule	Mois d'enregistrement du devis
Région	Zonier Commercial Vol-Incendie	Novice
Formule	Zonier Commercial RC	Zonier Commercial Vol-Incendie
Durée du relevé d'information	Nombre d'enfants	Région
Zonier commercial Vol-Incendie	Zonier Commercial Dommage	Zonier Commercial RC
Puissance du véhicule	Novice*Age_Conducteur*Age_Permis	Novice*Age_Conducteur*Age_Permis
Ancienneté d'acquisition du véhicule	Région	Coefficient bonus-malus

FIGURE 4.18 – Les variables les plus significatives des trois méthodes (sans indicateurs)

D'après ce tableau, les modèles s'accordent tous sur l'importance du taux de réduction. Cette variable apparait comme la plus significative sur chacun des modèles. On remarque aussi que d'autres variables comme la région et les zoniers commerciaux figurent parmi les variables les plus significatives des trois modèles. L'ensemble des modèles s'accordent donc sur le fait que la situation géographique du risque est un facteur qui influence fortement la décision de concrétiser ou non un devis.

D'autre part, on remarque que les modèles de machine learning ont en commun un bon nombre de facteurs fortement discriminants. En effet, seulement deux des dix variables les plus significatives du modèle random forest ne figurent pas dans la liste des dix variables les plus significatives du gradient boosting. Ceci est certainement dû aux similitudes dans la conception des deux algorithmes.

Le modèle de régression logistique est le seul pour lequel la catégorie socioprofessionnelle, la durée du relevé d'information, la puissance du véhicule et l'ancienneté d'acquisition du véhicule sont parmi les variables les plus significatives. Bien que ces variables ne sont pas significatives pour les modèles machine learning, elles nous semblent tout de même

cohérentes. En effet, prenons le cas de la catégorie socioprofessionnelle. Deux de ses modalités sont "étudiant" et "cadre". Il y a du sens de considérer que le tarif proposé à un "cadre" puisse être nettement différent de celui qu'on proposerait à un "étudiant". Ce qui pourra donc influencer le prospect dans sa décision de concrétiser ou non le devis qui lui est proposé.

Comparons maintenant les différentes métriques calculées pour nos trois modèles. Elles sont renseignées dans le tableau ci-dessous. Le premier constat que nous faisons est que les méthodes de machine learning ont, quelle que soit la métrique considérée, de meilleurs résultats que la régression logistique. Cette remarque est assez cohérente selon nous, dans le sens où ces algorithmes ont été conçus de manière à ce que la prédiction soit plus robuste. Parmi les trois modèles, le random forest est celui qui a la meilleure spécificité : c'est le modèle qui permet le mieux d'identifier les devis non convertis. Il a deux points de plus que le gbm, qui est le second meilleur modèle en termes de spécificité. En ce qui concerne les mesures comme la sensibilié, l'accuracy (taux de bonnes prédictions) et l'AUC, les résultats sont en faveur du gradient boosting.

Au-delà de l'AUC et du taux de bonnes prédictions qui sont des métriques de qualité globale du modèle, nous nous intéressons particulièrement à la précision du modèle. En effet, du point de vue métier, cette métrique est très importante. Elle nous permet de connaitre a priori si un devis sera converti. Si un modèle a une précision de X% alors, il permet dans X% des cas de savoir si le devis sera a priori converti. Selon ce critère, les modèles machine learning donnent de meilleurs résultats, avec un léger avantage pour le gradient boosting.

Métriques	Régression Logistique	Random Forest	Gradient Boosting
Spécificité	84%	89%	87%
Sensibilité	49%	50%	55%
Précision	64,9%	75,1%	76%
Accuracy	69,9%	75%	75,9%
AUC	73,7%	80,5%	81,6%

TABLE 4.12 – Modèles sans les indicateurs : Comparaison des métriques

4.4.2 Modèles avec les indicateurs de compétitivité

Dans le tableau suivant, nous présentons pour chaque méthode, les variables plus significatives.

Régression Logistique	Random Forest	Gradient Boosting
Taux de réduction	Taux de réduction	Taux de réduction
Age du conducteur	Novice	Nombre d'enfants
Catégorie socioprofessionnelle	Nombre d'enfants	PSC
Ancienneté de permis	PSC	Novice
Potentiel Retail	Formule	Formule
Zonier commercial Vol-Incendie	Statut marital	Potentiel Auto
Durée du relevé d'information	Zonier Commercial Vol-Incendie	Potentiel Retail
Région	Zonier Commercial RC	Ancienneté d'acquisition du véhicule
Prime moyenne du marché (normale -	Novice*Age_Conducteur*Age_Permis	Zonier Commercial Vol-Incendie
Puissance du véhicule	Zonier Commercial Dommage	Région

FIGURE 4.19 – Les variables les plus significatives des trois modèles (avec indicateurs)

Tout comme sur les modèles sans les indicateurs de compétitivité, on remarque que le taux de réduction est la variable la plus significative de l'ensemble des modèles. On peut le considérer comme le premier facteur sur lequel on peut s'appuyer pour faire varier la probabilité de conversion de nos devis. C'est également un constat qui reflète l'environnement concurrentiel du marché de l'assurance automobile. Pour répondre à la forte demande en assurance et se démarquer des concurrents, l'un des éléments clés est de proposer des avantages tarifaires.

La région et les zoniers commerciaux restent des variables fortement significatives, preuve de l'importance de la situation géographique du risque pour la conversion. Un autre élément qui vient appuyer ce constat, est la présence du potentiel retail parmi les variables plus significatives dans les modèles de régression logistique et gradient boosting.

Concernant les indicateurs de compétitivité et les données externes, quelques unes de ces variables ont été significatives lors de la modélisation. C'est le cas de la prime moyenne pour la régression logistique et les potentiels auto et retail pour le gradient boosting. Ces variables ont contribué à l'amélioration de la qualité des modèles. En effet, en comparant les métriques des différents modèles, on constate que les modèles avec indicateurs de compétitivité sont légèrement meilleurs. Ces métriques sont renseignées dans le tableau suivant.

Métriques	Régression Logistique	Random Forest	Gradient Boosting
Spécificité	83%	89%	86%
Sensibilité	50%	52%	59%
Précision	66%	75,7%	77,9%
Accuracy	70,3%	75,2%	78,8%
AUC	74,6%	81,2%	82,5%

TABLE 4.13 – Modèles avec les indicateurs : Comparaison des métriques

Elles évoluent dans le même sens que celles des modèles sans les indicateurs de compéti-

tivité : les valeurs les plus faibles sont celles de la régression logistique et les plus fortes sont celles des modèles machine learning. Le modèle random forest est celui qui a la meilleure spécificité. L'introduction des indicateurs de compétitivité n'a pas d'effet sur celle-ci, elle reste à 89%. Par contre, la sensibilité du gradient boosting s'améliore de 4 points. Elle est désormais de 59%. Le modèle prédit mieux les devis réellement convertis. Le taux de bonnes prédictions et l'AUC de chaque modèle sont également améliorés.

Concernant la précision du modèle, nous faisons le même constat que sur les modèles sans les indicateurs de compétitivité : les modèles machine sont meilleurs que la régression logistque. Dans le cas du gradient boosting, notons que la précison a augmenté de quasiment trois points ; ce qui représente une réelle plus-value.

Conclusion

L'objectif de nos travaux était de construire un modèle dit de conversion, qui permettrait de prédire la probabilité qu'un prospect concrétise un devis en contrat. Pour ce faire, nous avons opté pour trois différentes approches d'apprentissage supervisé : la régression logistique, le random forest et le gradient boosting. L'idée était de comparer la performance de ces méthodes et de retenir celle qui fournit les meilleurs résultats. On pourra alors utiliser cette méthode afin d'alimenter l'outil d'aide à la décision dans le cadre de la mise à jour tarifaire : le Scenario Testing New Business.

Dans le souci de trouver des axes qui pourraient améliorer le modèle de conversion, nous avons eu recours à des sources de données externes. En effet, nous avons utilisé des indicateurs de compétitivité. Ils nous donnent une vision des tarifs proposés sur le marché et de notre positionnement tarifaire. Nous nous sommes également intéressés au maillage de notre réseau sur la France métropolitaine. Ceci nous a permis d'introduire de nouvelles informations : le nombre d'agences et les potentiels des INSEE sur l'ensemble de la France métropolitaine. Avec ces informations, nous avons une idée de la représentativité de nos points et de ceux des concurrents sur le territoire, ainsi que la consommation en assurance. D'autre part, utiliser le maillage du réseau de distribution dans l'explication de l'acte de conversion nous permet de nous démarquer de travaux existants sur le même thème.

Les principales métriques sur lesquelles nous nous sommes appuyés sont l'AUC et le taux de bonnes prédictions (Accuracy) des différents modèles. Il ressort de notre analyse que le modèle le moins performant est la régression logistique. Lorsqu'elle est implémentée sans les indicateurs de compétitivité, la régression logistique a un AUC de 73,7% et taux de bonnes prédictions de 69,9%. En ajoutant les indicateurs de compétitivité parmi les variables explicatives, on a un AUC de 74,6% et un taux de bonnes prédictions de 70,3%. Ensuite, vient le random forest. Dans ce cas, introduire les indicateurs de compétitivité conduit aussi à une légère amélioration de la qualité du modèle. En effet, sans les indicaters de compétitivité, le modèle a un AUC de 80,5% et un taux de bonnes prédictions de 75%. Tandis que le second a un AUC de 81,2% et un taux de bonnes prédictions de 75,3%. Enfin, la méthode la plus performante a été le gradient boosting. Pour le modèle sans les indicateurs de compétitivité, on a un AUC de 81,6% et un taux de bonnes prédictions de 75,9%. Pour le modèle avec les indicateurs de compétitivité, on a un AUC de 82,5% et un

taux de bonnes prédictions qui reste à 76,8%.

Certaines variables se sont démarquées sur nos différents modèles. C'est le cas du taux de réduction, qui est la variable la plus significative de nos six modèles. Ce qui traduit le fait que le prospect est sensible aux gestes commerciaux (offre promotionnelle, réduction de tarif, ...) qu'on peut lui proposer. Outre le taux réduction, nous avons également remarqué que les zoniers commerciaux figurent parmi les variables les plus significatives de l'ensemble des modèles. Constat jugé cohérent, car la localisation du véhicule est un facteur déterminant notamment pour les risques vol et incendie. Cette localisation impacte le prix proposé et par extension le choix du prospect.

Au terme de notre étude, nous constatons que les modèles avec indicateurs de compétitivité sont, de peu, meilleurs que les modèles sans. En termes d'AUC ou de taux de bonnes prédictions, les gains sont faibles. Toutefois, on note que les données externes, notamment les potentiels, ont montré qu'ils étaient significatifs pour le modèle gradient boosting et la régression logistique. Pour la suite, il serait intéressant selon nous, de croiser le modèle de conversion aux potentiels pour optimiser le maillage de notre réseau de distribution.

Bibliographie

A. CHARGERAUD. *Le taux de transformation en automobile : comparaison de différentes méthodes d'apprentissage.* Mémoire d'actuariat, 2016.

A. GERON. *Machine learning avec scikit-learn.* 2017.

Comité des Constructeurs Français d'Automobiles (CCFA). Evolution du parc automobile français. 2020.

Documentation allianz france sur le modèle de conversion.

Documentation allianz france sur le tarif technique.

Documentation emblem.

Fédération Française de l'Assurance (FFA). Le marché de l'assurance automobile des particuliers en 2018, etudes et chiffres clés. 2020.

Infostat Marketing. Implantation des assurances, analysez la concurrence.

J.R. QUINLAN. *Programs for machine learning.* M. Kaufmann, 1993.

J. ELITH, JR. LEATHWICK, and T. HASTIE. *A working guide to boosted regression trees.* 2008.

L. BREIMAN, J. FRIEDMAN, and OLSEN R. *Classification and regression trees.* Wadsworth Brooks, 1984.

L. ROUVIERE. *Sélection-"validation" de modèles.* 2015.

R. BELLINA. *Méthodes d'apprentissage appliquées à la tarification non-vie.* 2014.

R. GENUER and J. POGGI. *Arbres CART et Forêts aléatoires, Importance et sélection devariables.* 2017.

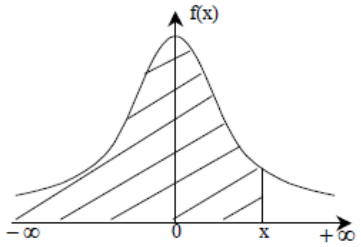
T. ROBERT. *Modélisation du taux de transformation et élasticité au prix.* Mémoire d'actuariat, 2019.

Annexes

A Table de la loi normale centrée réduite

Loi Normale centrée réduite

Probabilité de trouver une valeur inférieure à x.



$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

X	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

FIGURE 20 – Table de la loi normale centrée réduite

B Les variables de la base devis

Variables	Libellés	Variables	Libellés
AGECOND	Age du conducteur principal	TP_EMISSIONCO2	Emission de CO2
AGEPERM	Age d'obtentien du permis	TP_Energie	Type d'énergie
AGEVEH	Age du véhicule	TP_ENF	Enfant en âge de conduire
CFBMAPP	Coefficient bonus malus	TP_Formagev	Formule * Age_Véhicule
DTENRPSA	Date d'enregistement du devis	TP_FORMULE	Formule
INSEE	Code INSEE	TP_FRABDG	Franchise bris de glace
KM	Forfait kilométrique	TP_FRADOM	Franchise dommage
mois	Mois d'enregistrement du devis	TP_FRAVOL	Franchise vol
NOPSA	Numéro de police	TP_Garage	Garage
OPT_ASS	Assistance	TP_Gpsra	Groupe SRA
OPT_GC	Garantie conduteur	TP_Kros	Carosserie
REGION	Région	TP_MARITAL	Statut marital
TOP_CON	Option contenu	TP_MARQUE	Marque
TOP_EQU	Option équipements	TP_NBENF	Nombre d'enfants
TP_ABM50	Durée du bonus 50	TP_Notesecu	Note de sécurité
TP_ACHAT	Mode d'achat	TP_NOV	Novice
TP_AGEV_ACQ	Age_Véhicule * Ancienneté d'acquisition du véhicule	TP_NOV_Agec_Agep	Novice*Age_Conducteur*Age_Permis
TP_AIRBAGCOND	Airbag conducteur	TP_Nov_Agec_Agep_ENF	Novice*Age_Conducteur*Age_Permis*Nombre_d'enfants
TP_AIRBAGPASSAV	Airbag passager avant	TP_PDSAVIDE	Poids à vide
TP_ALIM	Alimentation	TP_Segment	Segment
TP_ANCACQ	Ancienneté d'acquisition du véhicule	TP_TRANSM	Transmission
TP_ANCP	Ancienneté de permis	TP_TXCLT	Taux de réduction
TP_BOITEVIT	Boite de vitesse	TP_VITMAXI	Vitesse max
TP_Cdfract	Code fractionnement	TP_Zone_BDG_com	Zonier commercial bris de glace
TP_CLPRIX	Classe de prix	TP_Zone_Dom_com	Zonier commercial dommage
TP_CLREP	Classe de réparation	TP_Zone_RC_com	Zonier commercial responsabilité civile
TP_COUPLE	Couple	TP_Zone_VI_com	Zonier commercial vol-incendie
TP_CSP	Catégorie socioprofessionnelle	TransAN	Conversion
TP_DureeRI	Durée du relevé d'information	Usage	Usage du véhicule

FIGURE 21 – Liste des variables de la base devis

C Découpage géographique

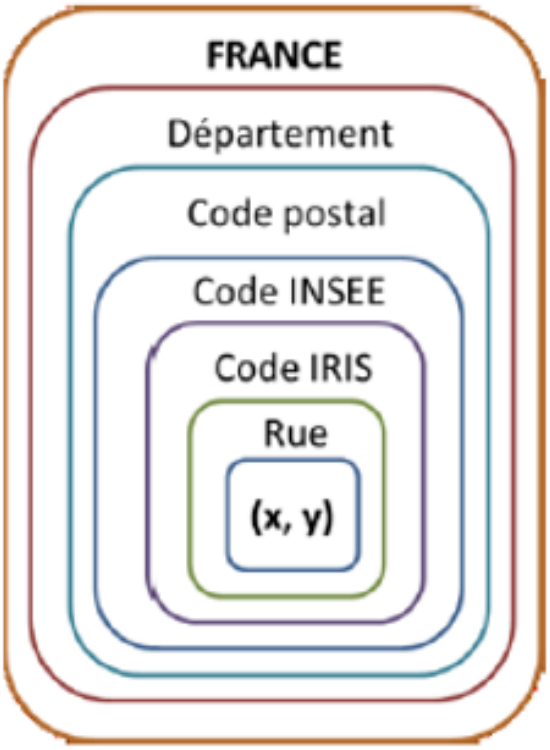


FIGURE 22 – Les différentes mailles géographiques

D Courbes ROC des modèles sans les indicateurs

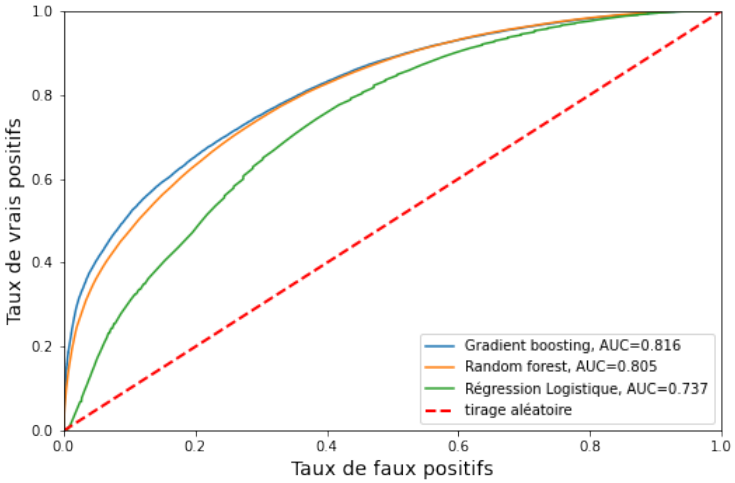


FIGURE 23 – Courbes ROC des trois modèles sans indicateurs

E Courbes ROC des modèles avec les indicateurs

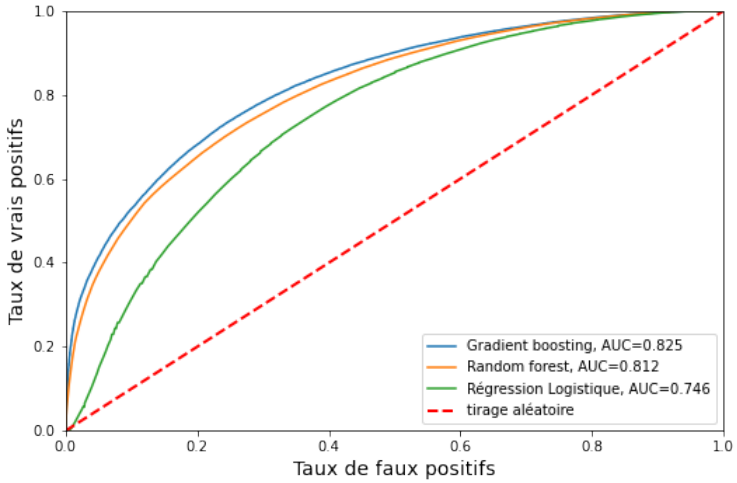


FIGURE 24 – Courbes ROC des trois modèles avec indicateurs