

Mémoire présenté devant l'ENSAE Paris  
pour l'obtention du diplôme de la filière Actuariat  
et l'admission à l'Institut des Actuares  
le 16/03/2022

Par : **Alice Bellot**

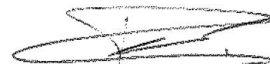
Titre : **Modélisation du changement de véhicule dans une optique de rétention**

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

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

*Membres présents du jury de la filière  
Nom :*

*Entreprise : Generali France  
Signature :*



*Membres présents du jury de l'Institut  
des Actuares*

*Directeur du mémoire en entreprise :*

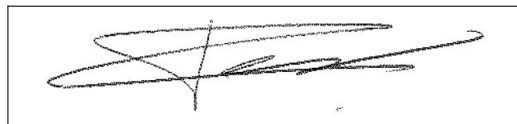
*Nom : Tanguy Carroussel  
Signature :*



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

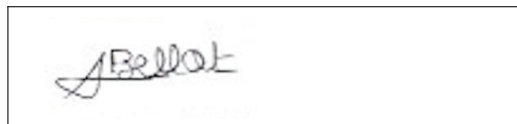
Secrétariat :

Signature du responsable entreprise



Bibliothèque :

Signature du candidat





## Résumé

Depuis de nombreuses années, l'assurance automobile occupe une place majeure dans le secteur de l'assurance IARD (Incendies, Accidents et Risques Divers). Exigée à un niveau minimal de couverture et produit d'appel par excellence, elle fait l'objet d'une concurrence toujours plus soutenue entre les différents acteurs du marché. Cette compétition est accentuée par la stagnation du parc automobile français, l'entrée en vigueur de nouvelles lois qui assouplissent les procédures de résiliation, l'apparition de nouveaux acteurs et canaux de distribution, ainsi que par les comparateurs d'assurance qui permettent aux clients d'avoir une vision plus large sur les différents produits, tarifs et prestations que proposent les assureurs.

Face à ce contexte de rétention qui fragilise la rentabilité des portefeuilles, les assureurs redoublent d'efforts pour proposer des prix attractifs, de meilleures garanties et s'adapter aux attentes des assurés.

La souscription d'un nouveau contrat étant un processus coûteux, il devient indispensable pour les assureurs de se recentrer autour des problématiques de rétention.

Parmi les principaux motifs de résiliation, le changement de véhicule représente une part très importante des départs et constitue ainsi l'une des principales fragilités des portefeuilles d'assurance automobile.

Il convient donc de s'interroger sur les facteurs influençant le choix des agents économiques de remplacer leur véhicule afin d'être en mesure d'anticiper cette décision et de mettre en place des actions adéquates – notamment au moyen de réductions tarifaires accordées sur les avenants - pour conserver les clients à valeur pour la compagnie.

Dans cette optique, la construction de scores de fragilité au changement de véhicule et à la résiliation pour ce motif constitue l'une des premières étapes. Pour cela, ce sont les méthodes d'arbres de classification qui retiennent notre attention. La comparaison d'un modèle de forêt aléatoire, utilisé pour sa simplicité et sa bonne lisibilité, et d'un modèle CatBoost, sélectionné pour sa robustesse et sa gestion optimisée des variables catégorielles, nous conduit à sélectionner la méthode qui discrimine le mieux les comportements étudiés.

Par la suite, une étude des résultats obtenus nous permet d'affiner la compréhension des profils de clients fragiles et des facteurs explicatifs.

Enfin, une intégration d'indicateurs de rentabilité des clients au sein du portefeuille détermine le périmètre des clients éligibles à des offres commerciales dans une optique de rétention.

**Mots clés** : assurance automobile, changement de véhicule, résiliation, rétention, ciblage, rabais commerciaux, scores de fragilité, modèles de classification, forêt aléatoire, CatBoost, valeur client.

## Abstract

For many years, car insurance has occupied a major place in the property and casualty insurance sector. Required at a minimum level of coverage and a prime loss leader, it is the subject of increasingly intense competition between the various market players. This competition is accentuated by the stagnation of the French car fleet, the entry into force of new laws that make cancellation procedures more flexible, the emergence of new players and distribution channels, as well as insurance comparison services that allow customers to have a broader view of the different products, rates and services offered by insurers.

Faced with this retention context that weakens the portfolios's profitability, insurers are redoubling their efforts to offer attractive prices and better guarantees and to adapt to the policyholders' expectations.

Since underwriting a new policy is a costly process, it is essential for insurers to refocus on retention issues.

Among the main reasons for termination, the change of vehicle represents a very large proportion of departures and is thus one of the main weaknesses of car insurance portfolios.

It is therefore necessary to examine the factors influencing the choice of economic agents to replace their vehicle in order to be able to anticipate this decision and to put in place adequate actions - in particular with tariff reductions on endorsements - in order to retain valuable customers.

From this point of view, the construction of fragility scores for vehicle change and termination for this reason is one of the first steps. For this purpose, classification trees' methods are the ones we focus on. The comparison of a Random Forest model, used for its simplicity and good readability, and a CatBoost model, selected for its robustness and its optimized management of categorical variables, leads us to select the method that best discriminates the behaviors studied.

Subsequently, an analysis of the results obtained allows us to refine our understanding of the profiles of fragile clients and the explanatory factors.

Finally, an integration of customer profitability indicators within the portfolio determines the scope of customers eligible for commercial offers with a view to retention.

**Key words** : car insurance, vehicle change, termination, retention, targeting, fragility scores, trade discounts, classification models, Random Forest, CatBoost, customer value.



## Note de synthèse

Depuis ces dernières années, le marché de l'assurance IARD a subi de nombreuses transformations. Les changements réglementaires au travers de la loi Hamon (2015), l'arrivée de nouveaux acteurs sur le marché et l'essor des comparateurs d'assurance exacerbent la concurrence entre les assureurs.

Afin d'être attractifs et s'adapter aux besoins des clients, les compagnies d'assurance développent de nouveaux produits, services et garanties et ajustent leurs tarifs. Néanmoins la souscription d'un contrat reste un processus coûteux puisqu'un effort financier est souvent réalisé pour proposer un tarif compétitif au nouveau venu; effort qui doit être compensé sur la durée de vie du contrat. Dans cette optique, les assureurs doivent se recentrer autour des problématiques de rétention.

Le changement de véhicule, à l'origine d'un nombre conséquent de résiliations, fragilise la durée de vie des contrats et menace ainsi la rentabilité des portefeuilles. Il devient alors primordial pour les assureurs d'être en mesure d'anticiper ce comportement afin de mettre en place des actions de rétention pour conserver les clients d'intérêt dans leur portefeuille.

Ces actions peuvent prendre plusieurs formes mais l'octroi de rabais sur le tarif commercial semble être l'une des meilleures solutions; la sensibilité au prix des individus étant un élément particulièrement déterminant.

L'objectif de ce mémoire consistera donc à mettre en place deux scores de fragilité permettant de prédire les probabilités qu'un client change de véhicule dans l'année et résilie son contrat par la suite. A la suite du calibrage et du choix des modèles, l'analyse des résultats obtenus permettra d'établir des profils de clients plus susceptibles d'adopter l'un des deux comportements étudiés.

La première étape de cette étude est la constitution de la base de données. Cette étape est primordiale car elle peut conditionner la qualité de l'ensemble des résultats qui seront obtenus par la suite. Nous avons à notre disposition des données internes pour caractériser le client, le contrat ou le bien assuré, mais aussi pour retracer le parcours client sur un historique étendu. L'ajout de données externes à partir de sources telles que l'INSEE ou l'organisme SRA (Sécurité et Réparation Automobile), complètent les informations recueillies, respectivement sur l'environnement géographique, économique et social de l'assuré à la maille de la commune et sur les caractéristiques détaillées de son véhicule.

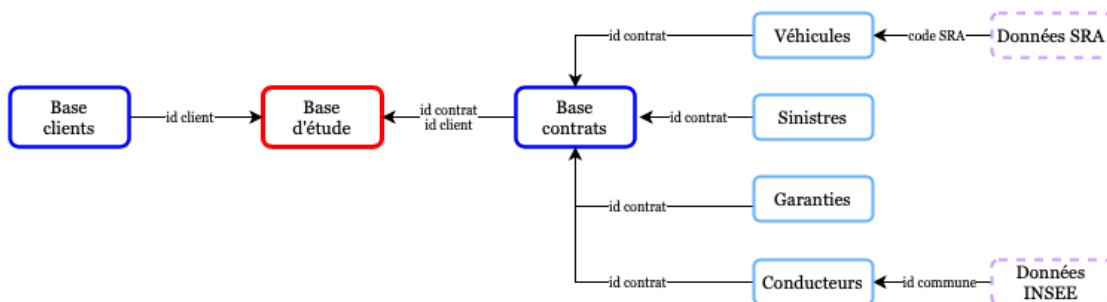


FIGURE 0.1 – Liaison entre les différentes sources de données

Suite à la récupération et au regroupement des différentes données, de nombreux traitements ont dû être appliqués afin de rendre la base de données représentative de la population étudiée et robuste à l'application de méthodes d'apprentissage automatique. De cette façon, les valeurs manquantes et aberrantes ont fait l'objet d'une analyse tandis que la répartition des classes des différentes variables catégorielles a été étudiée.

Sur l'ensemble des individus qui constituent notre base d'étude, environ 14% changent de véhicule au cours de chacune des trois années d'observation. En outre, il est particulièrement intéressant de constater que près d'un changement de véhicule sur deux conduit à une résiliation. Cela met en perspective l'importance pour Generali de mettre en place des actions de rétention sur son portefeuille automobile puisque le changement de véhicule est à l'origine d'une fuite de contrats non négligeable.

L'analyse des caractéristiques générales du portefeuille et une étude comportementale permettent d'identifier les caractéristiques jouant sur la fragilité au changement de véhicule et à la résiliation suite à la modification du bien assuré. Plusieurs tendances intéressantes se dégagent.

On remarque notamment que l'âge du véhicule n'a pas un effet linéaire sur la propension à acquérir un nouveau bien durant l'année. En effet, ce sont les véhicules les plus récents (moins de 4 ans) ou les plus anciens (plus de 10 ans) qui sont plus susceptibles d'être changés durant l'année. Si l'ancienneté du véhicule explique facilement la nécessité de le remplacer compte tenu de l'obsolescence de ses capacités techniques et technologiques, l'influence du caractère neuf est plus difficilement interprétable. Des recherches complémentaires ont permis d'associer le désir de changer de véhicule avec la dépréciation de son prix. En effet, une décote importante est observée dans les premières années après l'achat à neuf d'un véhicule. Ainsi, ce dernier perd en moyenne 15% de sa valeur la première année puis 10% les 3 années suivantes. C'est pour pallier cette dépréciation de la valeur et récupérer de la vente d'un véhicule un apport suffisant pour l'acquisition d'un nouveau bien, que l'on peut observer des reventes rapides après l'achat à neuf. Ce phénomène est d'autant plus vrai que la valeur des véhicules est élevée.

Par ailleurs, on peut noter que des variables telles que la durée de détention du véhicule, la formule souscrite, l'âge du conducteur, le prix à neuf du véhicule ou encore la nature du bien assuré sont plutôt révélatrices de la propension d'un client à changer d'automobile ou non durant l'année.

En outre, en s'intéressant spécifiquement à la résiliation pour cause de changement de véhicule, on constate que certaines variables corroborent l'hypothèse selon laquelle les individus sont sensibles au prix puisque des variables telles que la prime hors taxes annuelle ou la revalorisation de la prime par rapport à l'année précédente influencent également la décision de clôturer le contrat.

La détection des individus enclins à résilier suite à un changement de véhicule et l'analyse des facteurs déterminants de cette prise de décision sont capitales pour un assureur tant c'est une période charnière dans la vie d'un contrat automobile. En effet, comprendre les raisons qui poussent un assuré à clôturer son contrat permet à l'assureur d'anticiper ce moment et de mettre en place des actions appropriées afin de l'éviter.

Dans cette optique, la constitution de scores de fragilité va permettre d'attribuer une probabilité de changement de véhicule et de résiliation à chaque client de notre périmètre.

En tenant compte de la structure de notre base de données, composée d'un nombre conséquent de variables catégorielles, et de la distribution des classes de nos variables cibles, ce sont les méthodes de classification qui ont retenu notre attention pour modéliser les deux phénomènes étudiés.

Avant la modélisation, une pré-sélection des variables explicatives a été nécessaire afin d'optimiser les temps d'apprentissage et les performances prédictives. L'analyse des corrélations, le recours à des méthodes de régressions pénalisées et l'utilisation de l'importance des variables dans les méthodes utilisant des arbres de classification nous ont permis de définir la dimension optimale de la base de données pour la construction des modèles.

Par la suite, en s'assurant de la stabilité des résultats en ayant recours à des méthodes de validation croisée, nous avons comparé les résultats obtenus par un modèle de forêt aléatoire et un CatBoost.

Des métriques de scores adaptées à notre problématique, permettant de tenir compte du déséquilibre des classes de la variable cible, nous ont conduit à sélectionner le CatBoost. Si cet algorithme d'apprentissage supervisé est très puissant, robuste et permet un traitement optimal des variables catégorielles, l'interprétation des résultats obtenus reste plus difficile que dans le cadre de régressions linéaires ou de méthodes "simples" utilisant des arbres de décision.

Les seuils d'erreur de prédiction obtenus pour chacun des deux modèles construits sont satisfaisants.

Une analyse des métriques de scores en fonction des quantiles de probabilités prédites permet de refléter l'efficacité du ciblage que pourrait avoir la campagne de rétention selon le nombre de clients qui seraient contactés.

Outre ces considérations liées au profil des individus identifiés comme fragiles, il est nécessaire de tenir compte de l'intérêt qu'a la compagnie de conserver ces clients dans son portefeuille. En intégrant un indicateur de rentabilité des clients calculé indépendamment de cette étude - la valeur client -, nous pouvons exclure certains individus du périmètre des clients éligibles.

Enfin, l'estimation des coûts et des gains que pourraient générer la mise en place de la campagne de rétention est nécessaire afin de décider de l'intégration ou non des scores de fragilité dans les systèmes et de définir le nombre optimal de clients qui pourront en bénéficier.

Pour conclure, l'un des principaux défis pour l'activité d'assurance est d'anticiper les décisions prises par les assurés. L'étude menée lors de ce mémoire a tenté d'aller dans ce sens puisqu'elle apporte une connaissance des facteurs qui influencent la décision de changer de véhicule et de résilier son contrat à cet effet. La mise en place de scores de fragilité a permis d'identifier les clients susceptibles d'adopter les comportements étudiés, avec une grande importance accordée à la résiliation. En tenant compte de la possibilité de contacter les clients et en intégrant des indicateurs de rentabilité, il est possible de déterminer les profils éligibles aux offres commerciales.

Compte tenu du délai imparti pour construire ces modèles et des données disponibles au moment de l'étude, la qualité des modèles obtenus est satisfaisante. Néanmoins, des travaux supplémentaires pourraient être menés afin d'améliorer la précision des modèles en intégrant de nouvelles variables qui pourraient s'avérer significatives, ou encore de construire des modèles d'élasticité pour identifier les clients qui seraient sensibles aux offres de rabais sur le tarif.



## Executive summary

In recent years, the property and casualty insurance market has undergone many transformations. Regulatory changes through the Hamon law (2015), the arrival of new market players and the development of insurance comparators exacerbate the competition between insurers.

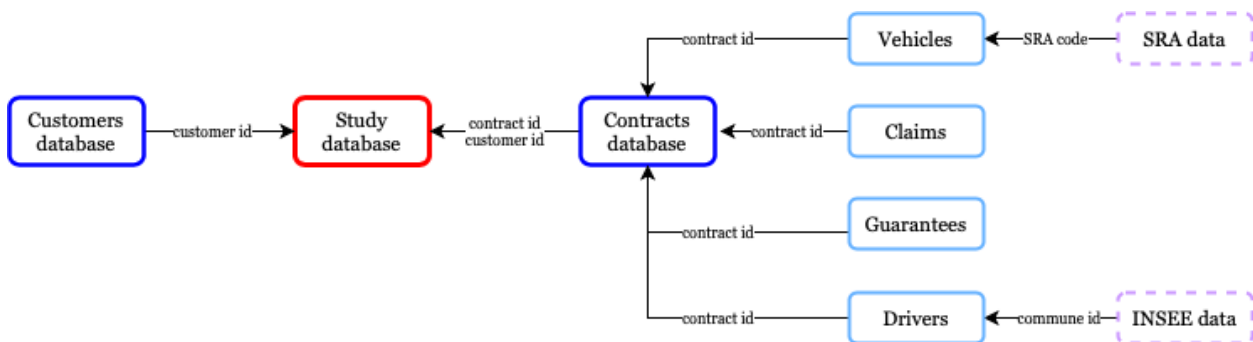
In order to be attractive and to adapt to the customers' needs, insurance companies develop new products, services and guarantees and adjust their tariffs. Nevertheless, underwriting a policy remains a costly process since a financial effort is often made to offer a competitive rate to the newcomer. This effort must be compensated for over the life of the contract. With this in mind, insurers need to refocus on retention issues.

Vehicle change, spearhead of a significant number of terminations, weakens the life of the contracts and thus threatens the profitability of portfolios. It is therefore essential for insurers to be able to anticipate this behavior in order to set up retention actions to keep the valuable policyholders in their portfolio.

These actions can take several forms but the granting of discounts on the commercial tariff seems to be one of the best solutions ; the price sensitivity of individuals being a particularly decisive element.

The objective of this memoir will be to set up two fragility scores allowing to predict the probabilities that a customer will change his vehicle within the year and cancel his contract afterwards. Following the calibration of the models, the analysis of the results obtained will allow us to establish customer profiles that are more likely to adopt one of the two behaviors studied.

The first step of this research is the constitution of the database. This step is crucial because it can condition the quality of all the results that will be obtained afterwards. We have at our disposal internal data to characterize the customer, the contract or the insured vehicle, but also to retrace the customer's path over an extended history. The inclusion of external data from sources such as INSEE or SRA (Automotive Safety and Repair), completes the data collected, respectively on the geographical, economic and social environment of the policyholder at the level of the commune and on the detailed characteristics of his vehicle.



Following the retrieval and aggregation of the different data, numerous treatments had to be applied in order to make the database representative of the population and robust to the application of machine learning methods. In this way, missing values and outliers were analyzed while the class distribution of the categorical variables was studied.

Of all the individuals in our database, about 14% change their vehicle in each of the three observation years. Moreover, it is particularly interesting to note that almost one out of two vehicle changes lead to termination. This puts into perspective the importance for Generali to set up retention actions on its automobile portfolio since the vehicle change leads to a significant leakage of contracts.

The analysis of the general characteristics of the portfolio and a behavioral study allow us to identify the factors that play a role in the fragility of vehicle change and the termination of the contract following the modification of the insured property. Several interesting trends emerge.

In particular, we note that the age of the vehicle does not have a linear effect on the propensity to acquire a new property. In fact, it is the most recent vehicles (less than 4 years old) or the oldest ones (more than 10 years old) that are more likely to be changed within the year. While the seniority easily explains the need to replace it due to the obsolescence of its technical and technological capabilities, the influence of the newness of the vehicle is more difficult to interpret. Additional research has made it possible to associate the desire to change a vehicle with the depreciation of its price. Indeed, a significant drop in the value is observed in the first years after buying new. On average, the vehicle loses 15% of its value in the first year and 10% in the following three years. It is to compensate for this effect and to get back from the sale a sufficient contribution for the acquisition of a new property, that one can observe fast resales after the purchase of a new vehicle. This phenomenon is all the more true as the value of the vehicle is high.

Furthermore, it can be noted that variables such as the holding period of the vehicle, the age of the driver, the guarantees covered by the contract, the price when new of the vehicle or the nature of the insured property are rather revealing of the propensity of a client to change his vehicle or not during the year.

Furthermore, when we look specifically at termination due to vehicle change, we find that some factors corroborate the hypothesis that individuals are price sensitive, since variables such as the annual pre-tax premium or the variation in the premium compared to the previous year strongly influence the decision to terminate the contract.

The detection of individuals inclined to terminate following a vehicle change and the analysis of the determining factors of this decision are crucial for an insurer as this is a decisive period in the life of a car policy. Indeed, understanding the reasons that push a policyholder to close his contract allows the insurer to anticipate this moment and to set up appropriate actions to avoid it.

With this in mind, the creation of fragility scores will make it possible to assign a probability of vehicle change and termination to each client in our scope.

Taking into account the structure of our database, composed of a large number of categorical variables, and the distribution of the classes of our target variables, classification methods were chosen to model the two phenomena under study.

Before modeling, a pre-selection of the explanatory variables was necessary in order to optimize the learning time and the predictive performance. The analysis of correlations, the use of penalized regression methods and the use of the features importances in the methods using classification trees allowed us to define the optimal size of the database for the construction of the models.

Subsequently, by ensuring the stability of the results through the use of cross-validation, we compared the results obtained by a Random Forest model and a CatBoost.

Score metrics adapted to our problem, allowing to take into account the imbalanced classes of the target variable, led us to select the CatBoost. If this supervised learning algorithm is very powerful, robust and allows an optimal treatment of categorical variables, the interpretation of the results obtained remains more difficult than in the framework of linear regressions or "simple" methods using decision trees.

The prediction error thresholds obtained for each of the two models constructed are satisfactory.

An analysis of the score metrics as a function of the quantiles of predicted probabilities makes it possible to reflect the effectiveness of the targeting that the retention campaign could have, depending on the number of customers who would be contacted.

In addition to these considerations related to the profile of individuals identified as fragile, it is necessary to take into account the company's interest in keeping these clients in its portfolio. By integrating a customers profitability indicator calculated independently of this study - the customer value - we can exclude some profiles from the scope of eligible customers.

Finally, estimating the costs and benefits of implementing the retention campaign is necessary in order to decide whether or not to integrate the fragility scores into the internal systems and to define the optimal number of customers who will benefit from an offer.

To conclude, one of the main challenges for the insurance business is to anticipate the decisions made by the policyholders. The study carried out in this memoir has attempted to go in this direction since it provides knowledge of the factors that influence the decision to change vehicle and to terminate one's contract for this purpose. The setting up of fragility scores allowed us to identify the customers likely to adopt these behaviors, with a great importance given to termination. By taking into account the possibility of contacting customers and by integrating profitability indicators, it is possible to determine the profiles eligible for commercial offers.

With regards to the time allowed to build these models and the available data at that time, the quality of the models obtained is satisfactory. Nevertheless, additional work could be done to improve the precision of the models by integrating new variables that could prove to be significant, or to build elasticity models to identify customers who would be sensitive to fare discount offers.



## Remerciements

Tout d'abord, je tiens à remercier Hamza El Hassani de m'avoir offert l'opportunité de travailler au sein de son équipe ainsi que mes collègues pour leur accueil.

Mes remerciements s'adressent spécialement à Tanguy Carroussel, mon tuteur en entreprise, pour l'intérêt du sujet sur lequel j'ai travaillé, sa disponibilité et son écoute ainsi que pour la qualité de son encadrement. Les nombreux points et échanges que nous avons partagé m'ont été d'une aide précieuse et m'ont permis de mener à bien mes travaux et de développer mes compétences professionnelles.

J'adresse également ma gratitude à l'équipe Client de la Direction de la Technique Assurance de Generali France dans laquelle j'ai évolué et aux autres collaborateurs avec lesquels j'ai pu être en contact pour leurs précieux conseils et le partage de leurs connaissances.

Par ailleurs, je remercie le corps enseignement ainsi que les intervenants de l'ENSAE ParisTech pour la qualité de la formation dispensée. Mes remerciements vont plus spécialement à Caroline Hillairet, ma tutrice pédagogique, pour son suivi, ses suggestions et son aide pour mener à bien ce mémoire.

Enfin, mes remerciements vont à toutes les personnes qui m'ont encouragée dans la réalisation de ce projet.

# Table des matières

<b>Résumé</b>	<b>1</b>
<b>Abstract</b>	<b>2</b>
<b>Note de synthèse</b>	<b>4</b>
<b>Executive summary</b>	<b>7</b>
<b>Remerciements</b>	<b>11</b>
<b>Introduction</b>	<b>14</b>
<b>1 Chapitre 1 : Assurance automobile et problématique de la rétention au changement de véhicule</b>	<b>15</b>
1.1 Le marché de l'assurance IARD en France . . . . .	15
1.2 L'assurance automobile : un marché hautement concurrentiel . . . . .	18
1.2.1 Le produit d'assurance automobile . . . . .	18
1.2.2 La stagnation du parc automobile français . . . . .	20
1.2.3 Une concurrence particulièrement marquée . . . . .	23
1.3 Le changement de véhicule au cœur des enjeux de rétention . . . . .	26
1.4 Le principe de tarification . . . . .	28
1.5 Déroulé de l'étude . . . . .	32
<b>2 Chapitre 2 : Éléments théoriques sur les méthodes de <i>machine learning</i></b>	<b>34</b>
2.1 Introduction aux arbres de classification avec l'algorithme CART . . . . .	34
2.1.1 Principe général . . . . .	34
2.1.2 Construction de l'arbre maximal . . . . .	35
2.1.3 Élagage de l'arbre maximal . . . . .	37
2.1.4 Avantages et inconvénients de la méthode CART . . . . .	39
2.2 La forêt aléatoire . . . . .	40
2.2.1 Description de la méthode . . . . .	40
2.2.2 Tree Bagging . . . . .	40
2.2.3 Feature sampling . . . . .	41
2.2.4 L'algorithme de la forêt aléatoire . . . . .	42
2.2.5 Avantages et inconvénients de la forêt aléatoire . . . . .	42
2.3 Le CatBoost . . . . .	43
2.3.1 Description de la méthode . . . . .	43
2.3.2 Introduction au Boosting . . . . .	43
2.3.2.1 Boosting . . . . .	43
2.3.2.2 Gradient Boosting . . . . .	44
2.3.2.3 Ordered Boosting . . . . .	46
2.3.3 Le Target Encoding . . . . .	46
2.3.4 L'algorithme du CatBoost . . . . .	48
2.3.5 Avantages et inconvénients du CatBoost . . . . .	49
2.4 Interprétabilité et explicabilité des modèles . . . . .	50
2.4.1 La réduction moyenne de l'impureté . . . . .	50
2.4.2 Les valeurs SHAP . . . . .	51
2.5 Les $K$ -moyennes . . . . .	54

<b>3</b>	<b>Chapitre 3 : Cadre de l'étude</b>	<b>56</b>
3.1	Périmètre de l'étude . . . . .	56
3.2	Constitution de la base de données . . . . .	57
3.2.1	Méthodologie de la création de la base . . . . .	57
3.2.2	Enrichissement de la base de données . . . . .	59
3.2.3	Traitement <i>a posteriori</i> sur la base de données . . . . .	68
3.2.3.1	Suppression des valeurs aberrantes . . . . .	68
3.2.3.2	Suppression des données manquantes . . . . .	76
3.2.3.3	Regroupement par classes . . . . .	78
3.2.3.4	Suppression de la temporalité des variables . . . . .	80
3.3	Analyse descriptive . . . . .	81
3.3.1	Caractéristiques générales du portefeuille . . . . .	81
3.3.1.1	Les assurés . . . . .	81
3.3.1.2	Les contrats . . . . .	83
3.3.1.3	Les véhicules . . . . .	84
3.3.2	Caractéristiques sur le changement de véhicule . . . . .	85
3.3.3	Caractéristiques sur la résiliation pour changement de véhicule . . . . .	87
<b>4</b>	<b>Chapitre 4 : Modélisation du changement de véhicule</b>	<b>90</b>
4.1	La validation croisée pour s'assurer de la stabilité des classifications . . . . .	91
4.1.1	Validation non croisée : échantillon d'apprentissage et échantillon test . . . . .	92
4.1.2	Validation croisée à $k$ -blocs . . . . .	92
4.2	Sélection des variables . . . . .	93
4.2.1	Analyse des dépendances . . . . .	95
4.2.2	Régressions linéaires pénalisées . . . . .	99
4.2.3	Sélection au moyen de l'importance des variables . . . . .	100
4.3	Optimisation des hyperparamètres des modèles de classification . . . . .	102
4.4	Analyse de la qualité des modèles . . . . .	105
4.5	Interprétation des résultats . . . . .	114
4.5.1	Interprétabilité du modèle . . . . .	115
4.5.2	Explicabilité du modèle . . . . .	117
<b>5</b>	<b>Chapitre 5 : Modélisation de la résiliation au changement de véhicule</b>	<b>121</b>
5.1	Sélection des variables . . . . .	121
5.2	Analyse de la qualité des modèles . . . . .	123
5.3	Interprétation des résultats . . . . .	127
<b>6</b>	<b>Chapitre 6 : Intégration du score dans les systèmes</b>	<b>132</b>
6.1	Identification des groupes de clients fragiles à la résiliation . . . . .	133
6.2	Intégration de la valeur client . . . . .	135
6.3	Détermination du ciblage des politiques de rétention . . . . .	137
6.4	Estimation des coûts et des gains d'une campagne de rétention . . . . .	138
6.5	Pistes d'amélioration . . . . .	140
	<b>Conclusion</b>	<b>142</b>
	<b>Bibliographie</b>	<b>144</b>
	<b>A Figures annexes</b>	<b>146</b>
	<b>B Éléments théoriques complémentaires</b>	<b>149</b>
	<b>Liste des figures</b>	<b>150</b>
	<b>Liste des tables</b>	<b>151</b>

## Introduction

Depuis de nombreuses années, l'assurance automobile occupe une place majeure dans le secteur de l'assurance IARD (Incendies, Accidents et Risques Divers). Exigée à un niveau minimal de couverture et produit d'appel par excellence, elle fait l'objet d'une concurrence toujours plus soutenue entre les différents acteurs du marché. Cette compétition est accentuée par la stagnation du parc automobile français, l'entrée en vigueur de nouvelles lois qui assouplissent les procédures de résiliation, l'apparition de nouveaux acteurs et canaux de distributions, ainsi que par les comparateurs d'assurance qui permettent aux clients d'avoir une vision plus large sur les différents produits, tarifs et prestations que proposent les assureurs.

Face à ce contexte de rétention fragilisé, les assureurs redoublent d'efforts pour proposer des prix attractifs, de meilleures garanties ou des offres plus adaptées aux besoins des assurés.

La souscription d'un nouveau contrat étant un processus coûteux, il est indispensable pour les compagnies d'assurance de se recentrer autour des problématiques de rétention. Dans cette optique, la connaissance client devient centrale afin d'appréhender les facteurs conduisant à la clôture des contrats et de mettre en place des politiques adaptées pour améliorer leur duration.

Parmi les principaux motifs de résiliation, le changement de véhicule représente une part très importante des départs et constitue ainsi l'une des principales fragilités des portefeuilles automobiles.

Lorsqu'un individu décide de remplacer son véhicule, ses attentes concernant le niveau de couverture et les dommages couverts par les garanties sont susceptibles de changer. Le défi des assureurs est alors d'anticiper l'instant auquel l'assuré adoptera ce comportement afin de répondre à ses nouveaux besoins et le fidéliser pour que ce dernier reste dans leur portefeuille.

L'objet de ce mémoire sera donc d'identifier les facteurs influençant la décision de changer de véhicule et celle de résilier suite à cela.

Une étude préliminaire menée sur le portefeuille permettra de relever des premiers éléments grâce à des analyses comportementales et d'éléments économiques ou sociaux.

Ensuite, pour améliorer la rétention au changement de véhicule, nous construirons un score de fragilité pour chacun des deux comportements mentionnés. La prédiction des probabilités sera faite au moyen de méthodes de classification telles que la forêt aléatoire ou le CatBoost.

Les modèles ainsi obtenus discriminent efficacement le changement de véhicule et la résiliation.

La mise en place de ces modèles de prédiction permet d'affiner l'analyse sur les facteurs influençant la résiliation et conduit à se questionner sur la pertinence de mettre en place des politiques de rétention au moyen de rabais commerciaux accordés sur le tarif. En effet, il convient de s'interroger sur la part des résiliations qui peuvent être imputées à la sensibilité au prix des individus afin d'optimiser les procédés tarifaires des avenants au changement de véhicule.

Par la suite, il sera nécessaire d'intégrer des indicateurs de rentabilité des clients au sein du portefeuille de Generali afin de déterminer le périmètre des individus éligibles à des offres commerciales dans une optique de rétention.



# 1 Chapitre 1 : Assurance automobile et problématique de la rétention au changement de véhicule

## 1.1 Le marché de l'assurance IARD en France

L'assurance peut se définir comme l'opération par laquelle une personne, l'assureur, s'engage à exécuter une prestation au profit d'une autre personne, l'assuré, en cas de réalisation d'un événement aléatoire, le risque, et en contrepartie du paiement d'une somme, la prime ou la cotisation.

Elle a pour vocation de couvrir des risques présents dans des secteurs très différents et de répondre à des besoins variés. Pour que le principe de mutualisation des risques puisse fonctionner, principe inhérent à l'activité d'assurance, cette dernière est divisée en plusieurs branches.

Ainsi, les 26 branches recensées dans le Code des Assurance peuvent être regroupées dans les deux grandes familles d'assurance qui ont été distinguées par la réglementation : l'assurance de personnes et l'assurance de biens et responsabilités.

Les assurances de personnes couvrent les risques qui portent atteinte à la personne. Cela concerne aussi bien les atteintes à son intégrité physique qu'à celles de sa vie. Ainsi, ces assurances couvrent les risques liés à la maladie, aux accidents corporels, à la dépendance, à l'incapacité, à l'invalidité, au décès ou à la retraite.

Les assurances de biens et de responsabilités couvrent quant à elles le patrimoine des personnes physiques et morales en indemnisant les sinistres ayant causé des pertes matérielles ou immatérielles, mais également des dommages corporels, matériels ou immatériels causés aux tiers.

Cette distinction aboutit à 18 branches classées dans la catégorie des assurances non-vie, 6 branches en vie, 1 branche particulière relative aux opérations tontinières et 1 branche non affectée.

L'assurance non-vie, aussi connue sous l'appellation d'assurance IARD (Incendies, Accidents et Risques Divers), regroupe ainsi l'ensemble des assurances hormis les contrats d'assurance vie. Autrement dit, elle concerne l'ensemble des conventions prudentielles qui ne sont pas liées ou conditionnées directement à la vie de l'assuré.

A titre d'exemple, les assurances retraite, décès ou épargne appartiennent à la catégorie des assurances vie tandis que les assurances santé, automobile, habitation ou encore multirisques professionnels appartiennent à la catégorie des assurances non-vie.

Les assurances de biens et de responsabilités ne protégeant pas les assurés et leur personne, elles sont souvent complétées par des garanties afin de couvrir ces dommages. De cette manière, si l'assurance vie est donc exclusivement une assurance de personnes, les garanties proposées dans les contrats d'assurance non-vie regroupent à la fois l'assurance de personnes et l'assurance de biens.

La cohabitation de l'assurance de biens et de responsabilités et l'assurance de personnes dans la branche non-vie permet d'établir une autre différenciation entre l'assurance vie et non-vie en ce qui concerne le règlement des sinistres. En effet, une distinction fondamentale existe dans les obligations de l'assureur lors de l'exécution du contrat entre ces deux catégories d'assurance.

Assurance de personnes	Prévoyance	Assurance vie
	Epargne / retraite	
	Assurance de biens et de responsabilités	Assurance non-vie
	Dommages corporels	
	Dommages incorporels	
	Responsabilités	

TABLE 1.1 – Les différents types d’assurance

En matière d’assurance de biens et de responsabilités, présente exclusivement en assurance non-vie, l’assureur doit indemniser le bénéficiaire du contrat des conséquences d’un sinistre, ce qui revient à procéder à une évaluation au cas par cas de la valeur des biens assurés : c’est le principe indemnitaire qui prévaut. Son fonctionnement se fonde sur le remplacement de l’assuré dans la situation qui aurait été la sienne en l’absence de sinistre. Le principe d’enrichissement de l’assuré étant interdit, le montant indemnisé ne peut dépasser la valeur du bien assuré au moment du sinistre.

A l’inverse, dans le cas de l’assurance de personnes, l’indemnisation qui sera versée est une somme déterminée préalablement, indépendamment des préjudices subis. Toutefois, on peut noter une atténuation à ce caractère exclusivement forfaitaire puisque certaines prestations en assurance de personnes (notamment pour les dommages corporels) présentent un caractère indemnitaire, par exemple les indemnités journalières de maladie.

La dualité entre l’assurance vie et non-vie s’exprime donc également au moment du traitement des sinistres puisque l’assurance vie présente une approche forfaitaire tandis que l’assurance non-vie peut proposer un recouvrement indemnitaire ou forfaitaire.

Cet antagonisme entre assurance vie et non-vie est marqué par la réglementation et le principe de spécialisation qui régit l’activité des entreprises d’assurance. Ce dernier stipule qu’un organisme n’est agréé que pour exercer des activités uniquement dans l’une de ces deux branches. La principale motivation est d’empêcher les assureurs d’utiliser l’épargne à long terme générée par les contrats d’assurance vie pour payer les sinistres rattachés aux assurances dommages. Ainsi, les grands groupes d’assurance présents dans les deux branches sont contraints d’isoler dans des sociétés distinctes les assurances vie d’une part et les assurances non-vie d’autre part.

Ce principe de spécialisation met en lumière une nouvelle classification qui permet de distinguer ces deux types d’assurance à savoir la gestion par capitalisation et la gestion par répartition. Comme mentionné précédemment, la réglementation française impose une séparation juridique entre les sociétés qui pratiquent des branches gérées par l’un ou l’autre de ces modes de gestion.

Dans les assurances gérées par répartition, l’assureur doit répartir entre les assurés sinistrés la masse des primes payées par l’ensemble des membres de la mutualité, la probabilité de réalisation du risque étant constante au cours du contrat. Les assurances qui utilisent ce mode de gestion sont les assurances dommages ainsi que deux types d’assurance de personnes : l’accident et la maladie.

Les assurances qui ne sont pas gérées par répartition sont donc gérées par capitalisation et sont généralement des assurances souscrites à long terme. Dans ce contexte, les primes vont être perçues selon la méthode des intérêts composés et le risque couvert n’est pas constant au cours du contrat. C’est notamment le cas des assurances sur la vie et la prévoyance collective.

Par la double différenciation entre l'assurance de personnes et celle de biens et de responsabilités - marquée par le mode de règlement des sinistres et le mode de gestion -, il est donc possible de classer chaque contrat d'assurance au sein de cette diversité que représente le secteur.

En outre, l'assurance IARD peut également se voir imposer un caractère obligatoire par la réglementation. Ainsi, l'assurance automobile pour les conducteurs et l'assurance habitation pour les locataires, colataires et copropriétaires sont exigées à un niveau minimum dans le but de couvrir les dommages causés à autrui. Au-delà de ces garanties obligatoires, chaque consommateur est libre de souscrire des garanties supplémentaires selon ses besoins.

L'assurance non-vie est donc une branche aux multiples caractéristiques et aux nombreux domaines d'application, ce qui la rend omniprésente mais implique également de nombreux défis pour les assureurs.

Afin de comprendre l'importance que représente ce secteur dans le marché français de l'assurance et ainsi mieux appréhender les enjeux que représentent les montants générés chaque année par cette activité, intéressons-nous aux chiffres de l'année 2020 présentés dans le rapport de la Fédération Française des Assurances [1].

En 2020, le marché français de l'assurance représente 201.9 milliards d'euros de cotisations sur les affaires directes. Si l'on regarde la décomposition de ces cotisations entre assurance de personnes et assurance de biens et responsabilités, la première catégorie mentionnée représente 141.7 milliards d'euros tandis que la seconde a permis de collecter 60.1 milliards d'euros sur l'année 2020, soit 29.7% de l'ensemble des cotisations.

En termes de croissance, le secteur assurantiel a subi directement les conséquences de la crise sanitaire du Covid-19 avec une baisse de l'ensemble des cotisations de 11.5% par rapport à 2019. Cette évolution provient du net repli de l'activité en assurance vie (-16.3%), légèrement atténuée par la faible hausse constatée sur le marché de l'assurance non-vie (+2.4%).

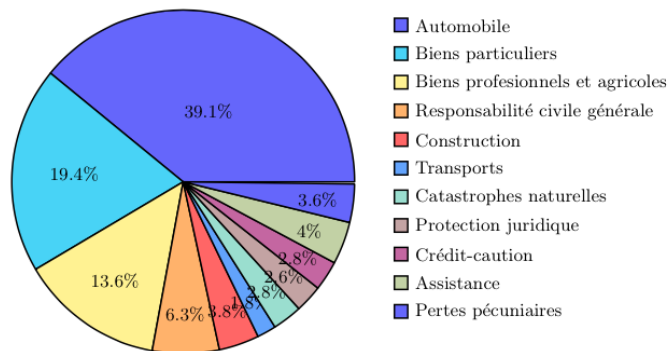


FIGURE 1.1 – Répartition des cotisations pour l'assurance de biens et de responsabilités

La France constitue ainsi le 3ème marché européen de l'assurance non-vie derrière l'Allemagne et le Royaume-Uni, marché largement porté par les particuliers qui représentent près de 63% des cotisations dommages.

En ce qui concerne le produit d'assurance IARD qui sera l'objet de ce mémoire, à savoir l'assurance automobile, ce dernier représente près de 40% des cotisations de l'assurance de biens et de responsabilités depuis de nombreuses années.

En raison de la place prépondérante qu'occupe cette branche d'assurance dans le portefeuille des assureurs, l'assurance automobile est donc un produit au cœur des défis que doivent relever les différents acteurs du marché pour rester compétitifs à la fois en termes de chiffres d'affaires et d'attractivité. Intéressons-nous donc de plus près à ce produit.

## **1.2 L'assurance automobile : un marché hautement concurrentiel**

### **1.2.1 Le produit d'assurance automobile**

Le contrat d'assurance automobile vise à proposer des garanties ayant pour objectif de couvrir les dommages matériels et corporels que le véhicule assuré pourrait occasionner, ainsi que les dommages que ce dernier pourrait subir. En outre, il a pour vocation de protéger contre les différents événements mettant en cause la responsabilité de l'assuré.

De principe indemnitaire et géré par répartition, il est destiné aux véhicules terrestres à moteurs circulant sur le territoire français, l'Espace Économique Européen ou dans la zone carte verte.

L'assurance automobile fait partie des assurances qui ont été rendues obligatoires par la réglementation. En France, c'est la loi du 27 février 1958 qui impose aux automobilistes de souscrire au minimum à une garantie "responsabilité civile", c'est-à-dire à une couverture des dommages causés aux tiers et aux passagers. Le non-respect de cette obligation d'assurance est puni pénalement.

De manière plus générale, les différentes garanties proposées par les contrats automobiles permettent de couvrir le propriétaire du véhicule, toute personne ayant la garde ou la conduite - même non autorisée - et les passagers du véhicule assuré. De la même façon, les opérations de chargement ou de déchargement peuvent être couvertes ainsi que les dommages causés par la chute des accessoires ou objets transportés - qu'ils surviennent au moment de la chute ou ultérieurement.

Les garanties proposées dans les contrats d'assurance automobile sont donc multiples et correspondent à différents niveaux de primes et de couverture. Globalement, on peut les regrouper en 3 principaux types de garanties : la garantie au tiers simple, au tiers étendu et la garantie tous risques.

L'assurance automobile "au tiers" se limite à la couverture de base. Elle comprend uniquement la garantie "responsabilité civile" qui est donc le minimum obligatoire auquel tout automobiliste doit souscrire pour rester dans la légalité s'il désire utiliser son véhicule. La garantie "au tiers étendu" reprend cette garantie et peut en inclure d'autres telles que : bris de glace, incendie, vol, attentat, catastrophes naturelles et technologiques. Enfin, la garantie "tous risques" reprend les garanties précédemment énoncées auxquelles s'ajoutent la garantie dommages tous accident qui couvre l'ensemble des dommages causés au véhicule en cas d'accident, responsable ou non, à l'arrêt ou en circulation. Elle permet à l'assuré d'être couvert quel que soit le contexte du sinistre et minimise ainsi ses pertes.

Bien évidemment, il existe des niveaux intermédiaires entre ces 3 niveaux principaux d'assurance ; les assureurs proposant de plus en plus aux consommateurs de personnaliser leur contrat afin de répondre à leurs demandes et s'adapter aux budgets que chacun est prêt à allouer à son produit d'assurance.

Ce caractère obligatoire et l'usage régulier des véhicules font de l'assurance automobile une dépense quasi incompressible pour les consommateurs et cela se traduit dans les chiffres.

Comme évoqué précédemment, l'assurance automobile concentre l'essentiel du marché des assurances dommages et représente un peu moins de la moitié des assurances souscrites par les particuliers en assurance de biens et de responsabilités.

En 2020, le chiffre d'affaires de l'assurance automobile s'établit à 23.5 milliards d'euros soit une hausse de 3% par rapport à 2019, croissance comparable à celle enregistrée l'année précédente.

En parallèle de cette hausse continue des cotisations, la fréquence des sinistres a également connu une nette amélioration, toutes garanties confondues, sous l'effet de la baisse du trafic routier durant la crise sanitaire.

Néanmoins, ce recul de la fréquence des sinistres est contrebalancé par l'augmentation des coûts moyens sur la même période. Si l'ensemble des prestations sur cette branche est en baisse de 6.95% par rapport à l'année précédente, en lien direct avec les conditions exceptionnelles de la crise sanitaire, elles ont tout de même atteint 17.4 milliards d'euros en 2020.

De nombreux facteurs peuvent expliquer cette hausse du coût moyen des sinistres que l'on observe depuis de nombreuses années.

Tout d'abord, on constate une forte dégradation de la sinistralité corporelle portée par une série de réformes défavorables aux assureurs. En effet, depuis 2013, les rentes versées aux accidentés de la route sont indexées sur l'inflation, ce qui représente un coût annuel de l'ordre de 2 à 4% de l'ensemble des cotisations.

Parmi les autres explications, nous avons également la hausse du coût de réparation des véhicules liée à la montée en puissance de la technologie embarquée, la généralisation des systèmes d'aide à la conduite et la sophistication des pièces automobiles, ou plus généralement de la montée en gamme des véhicules constituant le parc automobile français.

En outre, l'environnement de taux d'intérêts bas voire négatifs explique également cette hausse du coût des prestations versées. En effet, sur les branches longues telles que la Responsabilité Civile Automobile avec les rentes versées aux victimes de dommages corporels, la baisse des taux d'intérêts induit automatiquement une hausse des provisions techniques (provisions pour sinistres à payer, provisions mathématiques des rentes) à travers le phénomène d'actualisation.

En tenant compte de ces éléments, cela porte le ratio combiné de l'assurance automobile à 94.7% des primes en 2020. C'est la première fois depuis 2005 qu'il est à un niveau inférieur à 100!

En assurance, le ratio combiné permet de mesurer la rentabilité d'un produit en rapportant l'ensemble des frais (prestations versées, dotations aux provisions et frais généraux) aux ressources (total des cotisations et produits financiers).

Comparativement, la branche globale de l'assurance de biens et de responsabilités présente un ratio combiné de 96.7% en 2020.

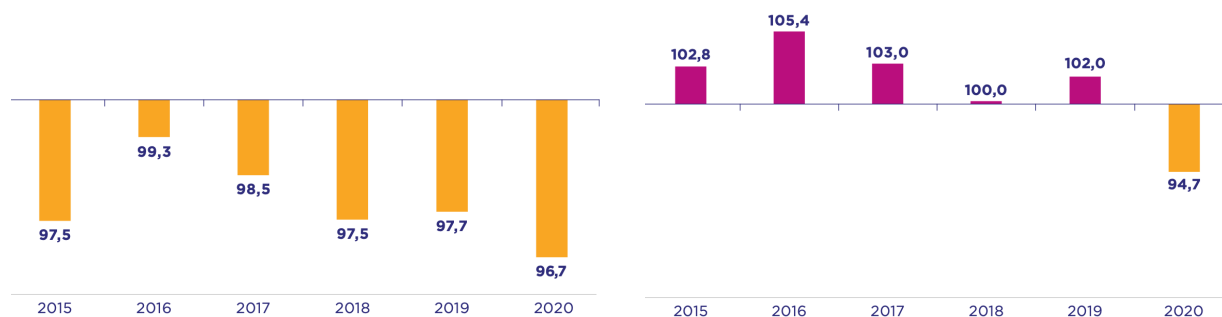


FIGURE 1.2 – Ratios combinés en assurance de biens et de responsabilités (à gauche) et en assurance automobile (à droite), nets de réassurance [1]

Si l'année 2020 présente des résultats singuliers liés à son contexte particulier, l'assurance automobile est globalement une branche aux résultats techniques négatifs et donc une branche plutôt déficitaire. Elle n'en reste pas moins un segment du marché de l'assurance non-vie particulièrement compétitif et très souvent utilisé comme un produit d'appel afin d'élargir la couverture du client (par exemple avec le produit habitation dans le cadre du multi-équipement auto-MRH).

En effet, c'est près d'un client sur deux qui rentre chez un agent général pour souscrire un contrat d'assurance automobile. Une fois la relation client amorcée, l'assureur a l'opportunité de lui proposer d'autres produits de sa gamme.

Produit d'ancrage de la relation avec le client en assurance non-vie, le produit automobile est un produit incontournable pour les assureurs car il génère un important chiffre d'affaires. Il occupe ainsi une place stratégique dans la politique de souscription des assureurs qui doivent de plus en plus se démarquer de leurs concurrents pour conserver leurs parts de marché.

Cette concurrence accrue observée sur le marché de l'assurance automobile est accentuée par deux principaux facteurs : la faible croissance du parc de véhicules et l'émergence de nouveaux acteurs.

### 1.2.2 La stagnation du parc automobile français

L'année 2020, lourdement marquée par la pandémie du Covid-19 et les restrictions de déplacement, a eu un fort impact sur le parc automobile français, autant sur son évolution que sur sa structure.

Au 1er janvier 2021, le parc automobile français en circulation, constitué de l'ensemble des véhicules immatriculés, a dépassé les 40 millions de véhicules, soit une hausse de près de 4% par rapport à l'année précédente. Détenu à 93% par des particuliers, il est constitué à 80% par des automobiles particulières, à 15% par des véhicules utilitaires légers (inférieurs à 5 tonnes) et à 2% par des autocars, autobus et véhicules industriels (supérieurs à 5 tonnes).

Dans le même temps, les ventes de véhicules ont chuté lourdement durant l'année 2020. Cela s'est traduit par une baisse des immatriculations de véhicules neufs de 25.5% et de 3.5% pour les véhicules d'occasion.

En mettant en perspective la hausse du nombre de véhicules en circulation et la baisse des ventes, on peut conclure qu'un nombre assez important de véhicules qui ne servaient plus ont été remis en circulation par leurs propriétaires. En mettant de côté les achats de véhicules neufs, on constate ainsi que l'âge moyen des véhicules nouvellement immatriculés en 2020 est de 16 ans. Ce comportement a vraisemblablement été motivé par la crise sanitaire qui a créé de nouveaux besoins de mobilité entre restrictions de circulation, évitement des transports en commun pour des raisons sanitaires, volonté de s'éloigner des grandes villes, etc. . .

Si l'on se concentre sur le marché d'occasion, la crise sanitaire a davantage pesé sur sa structure que sur son volume. En effet, comme nous venons de le mentionner, les immatriculations de véhicules de seconde main ont nettement moins diminué que celles des véhicules neufs, se maintenant à 5.6 millions de véhicules d'occasion vendus en 2020. Cependant, en regardant les différentes branches de ce marché, on constate que seules les immatriculations des véhicules de plus de 15 ans ont conclu l'année en hausse (+5.1%) tandis que les immatriculations des véhicules de plus de 10 ans (-0.6%), qui représentent la moitié des immatriculations, et de 2-5 ans (stable) ont plutôt bien résisté.

Par ailleurs, on relève une montée en puissance des offres locatives sur ce marché avec près de 3% des ventes, soit un volume avoisinant les 155 000 immatriculations.

Parallèlement, le marché des véhicules neufs a connu de grandes difficultés. Avec seulement 1.65 millions de véhicules neufs vendus, l'année 2020 marque un retour au niveau des ventes observé en 1972!

Malgré la remise en place du dispositif de la prime à la conversion par le gouvernement français en août 2020 pour stimuler les ventes et s'ancrer dans une démarche écologique, les ventes de véhicules neufs pâtissent largement des effets de la crise sanitaire sur le pouvoir d'achat couplés à une mise à l'arrêt des usines de fabrication automobile et des pénuries de matériaux.

Néanmoins, ce marché a connu une hausse très significative des ventes de véhicules neufs électriques et hybrides rechargeables qui ont quasiment doublé par rapport à 2019.

Ainsi, en 2020, le parc automobile français est constitué approximativement de 40% de véhicules achetés neufs et 60% de véhicules achetés d'occasion.

Ce ralentissement du marché des véhicules neufs et la résistance du marché de l'occasion se traduisent par le vieillissement du parc automobile français, en dépit des 200 000 primes à la conversion qui ont été accordées par le gouvernement.

Si l'on avait observé en 2020 une légère baisse de l'âge moyen du parc automobile français - alors en hausse continue depuis 2012 - grâce notamment au durcissement des règles de contrôle technique en 2019 qui avait accéléré la sortie du parc des véhicules les plus anciens, l'année 2021 a marqué une reprise à la hausse de cet âge moyen avec une valeur de 10.8 ans au 1er janvier.

Ce vieillissement du parc automobile français peut s'expliquer par le changement de comportement des Français en matière de consommation. En effet, les automobilistes conservent leur véhicule de plus en plus longtemps. Si l'on compare les habitudes de consommation entre 1990 et 2019, on constate ainsi que les automobilistes français renouvellent leur véhicule pratiquement 2 fois moins.

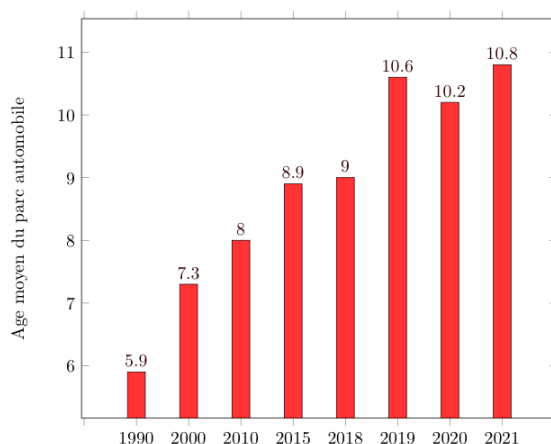


FIGURE 1.3 – Evolution de l’âge moyen du parc automobile français

Ce constat s’explique notamment par l’innovation constante dont font preuve les constructeurs pour rendre leurs véhicules plus fiables, améliorer leurs performances et augmenter leur durée de vie aussi bien mécaniquement, technologiquement qu’esthétiquement.

Ainsi, près d’un quart des véhicules circulants ont moins de 5 ans tandis que ce sont près de la moitié des véhicules roulants qui sont âgés de plus de 10 ans, avec près de 9.7 millions d’unités qui sont âgées entre 10 et 15 ans et 4.2 millions qui sont âgées de plus de 20 ans. Avec le succès de la prime à la conversion, dispositif régulièrement mis en place depuis plusieurs années, le poids des véhicules âgés entre 5 et 10 ans dans le parc automobile français a diminué pour atteindre 25% en 2021.

La chute des premières immatriculations, la résistance du marché de l’occasion et la remise en circulation d’anciens véhicules ont pour conséquence que le diesel demeure toujours largement majoritaire dans le parc roulant puisqu’il en représente près de 57%, néanmoins marqué par une baisse de 2 points par rapport au début de l’année 2020 et une baisse de 3.4% des ventes de voitures neuves (30.6% des ventes de 2020). Les motorisations alternatives, quant à elles, représentent 2% du parc avec en chef de file les hybrides non rechargeables (600 000 véhicules environ), suivies par les véhicules électriques (250 000) et les hybrides rechargeables (130 000). De manière globale, c’est plus de 97% du parc automobile qui fonctionnent à l’énergie thermique (essence ou diesel).

Un autre élément permettant d’analyser le parc automobile français sous un angle écologique est la vignette Crit’Air qui est un élément de régulation à des fins environnementales. Bien que toutes les automobiles n’en soient pas encore équipées, les éléments présents sur le certificat d’immatriculation permettent de déterminer à quelle vignette une voiture est éligible.

Ainsi au 1er janvier 2021, un tiers du parc roulant est éligible à une vignette Crit’Air 2 (véhicules à essence datant de 2006-2010 ou diesel depuis 2011), soit 2 points de moins que début 2020. Parallèlement, les vignettes Crit’Air 1 (essence après 2011, hybrides rechargeables) ont progressé de 1 point pour représenter 23% du parc automobile tandis que les vignettes Crit’Air 0 (électrique, hydrogène) en représentent 1%.

Au total, en 2022, c’est près de 43% du parc qui est menacé par les restrictions de circulation dans les ZFE (Zones à Faibles Émissions) en fonction des villes, soient les véhicules classés Crit’Air 3, Crit’Air 4, Crit’Air 5 et non classés. Les véhicules propres, quant à eux, représentent près de 24% du parc soient les 10 millions de véhicules catégorisés Crit’Air 0 et Crit’Air 1.



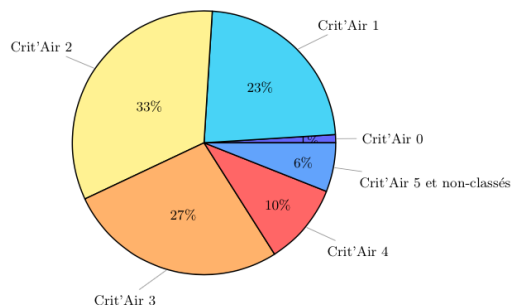


FIGURE 1.4 – Répartition du parc automobile français selon la vignette Crit'Air

Concernant la nature des véhicules constituant le parc automobile, la berline reste le modèle privilégié par les Français avec 58% des véhicules en circulation, suivie par les 17% de SUV qui devancent toujours les monospaces qui comptent pour 9% du parc. Subissant la pression mise sur les véhicules polluants, les véhicules de gamme supérieure tels que les coupés et les cabriolets ne représentent que 3% du parc et seulement 1% du marché des véhicules neufs en 2020. En outre, les marques françaises restent majoritaires avec 62% des automobiles.

Le vieillissement du parc automobile, qui s'explique en partie par une durée de détention des véhicules plus élevée, est donc directement lié à l'utilisation de moins en moins fréquente de ce moyen de transport. La prise de conscience de l'impact environnemental d'une utilisation régulière, la couverture des réseaux de transports publics ou encore la très forte hausse du prix des carburants sont autant de raisons qui expliquent que les Français font une utilisation beaucoup plus mesurée de leur véhicule et les conduit ainsi à les conserver plus longtemps.

Selon l'INSEE, le pourcentage des ménages disposant d'au moins un véhicule (taux de motorisation) est estimé à 86% en 2020. La variation de ce taux selon les communes est étroitement liée à la taille de l'agglomération d'habitation, le réseau de transports en commun ou à la distance domicile-travail, et donc directement associée à la nécessité de disposer d'un véhicule pour se déplacer.

Dans cet esprit, la part des véhicules utilisés quotidiennement a baissé tandis que le kilométrage moyen annuel a diminué de plus de 4 000 kilomètres par rapport à 2019 pour s'établir à 13 965 kilomètres parcourus en moyenne, tendance à mettre en perspective avec les effets de la crise sanitaire. Par ailleurs, le kilométrage moyen au compteur des véhicules a désormais dépassé les 100 000 kilomètres, confirmant la tendance des Français à conserver leur bien plus longtemps.

En définitive, la stagnation du parc automobile français est l'un des facteurs accentuant la concurrence déjà tendue entre les assureurs automobiles. En effet, si le nombre de véhicules à assurer n'évolue pas, la demande pour les produits d'assurance automobile stagne également dans un univers où les acteurs sont toujours plus nombreux avec une gamme de produits toujours plus variée.

### 1.2.3 Une concurrence particulièrement marquée

Tout d'abord, la concurrence sur la branche automobile de l'assurance non-vie est favorisée par le cadre réglementaire.

En effet, au fil des années, plusieurs lois ont imposé de nouvelles contraintes aux assureurs et accordé de nouveaux droits aux assurés.

A titre d'exemple, la loi Batinder du 5 juillet 1985 a accéléré les procédures d'indemnisation aux accidentés de la route, tout en permettant une harmonisation des procédures entre les différents assureurs. La loi Chatel appliquée en 2012 impose quant à elle aux assureurs d'informer les assurés de la durée de préavis pour résilier à l'échéance, avec pour objectif de protéger la tacite reconduction.

Mais c'est la loi relative à la consommation de mars 2014, dite loi Hamon, qui redistribue complètement les cartes sur le marché de l'assurance et « *rééquilibre les pouvoirs entre consommateurs et professionnels* ». Entrée en vigueur le 1er janvier 2015, cette loi vise à assouplir les procédures de résiliation des contrats d'assurance automobile, habitation et d'assurances affinitaires (objets et services) une fois la première année de souscription révolue.

En effet, jusqu'en 2014, les contrats d'assurance présentaient une durée annuelle avec tacite reconduction ; la résiliation à l'initiative de l'assuré pouvant intervenir uniquement dans les 2 mois précédant la date anniversaire du contrat. Désormais, les conditions de résiliation ont été largement assouplies avec la possibilité d'une résiliation infra-annuelle : une fois la première année de contrat écoulée, l'assuré a la possibilité de résilier son contrat à tout moment.

Par conséquent, cette loi a contribué à l'augmentation de la volatilité des clients et largement augmenté les taux de résiliations observés sur les portefeuilles des assureurs. Néanmoins, cela leur offre l'opportunité de capter de nouveaux clients.

Parallèlement à cette libération du marché imposée par une nouvelle réglementation, la concurrence entre les différents acteurs sur le marché de l'assurance automobile, et plus généralement sur le marché assurantiel, est d'autant plus importante avec l'émergence de nouveaux modes de distribution.

Classiquement, la distribution des produits d'assurance s'appuyait sur :

- **Des commerciaux salariés** de l'organisme assureur
- **Des agents généraux d'assurance** qui sont des travailleurs indépendants mandatés par un organisme d'assurance et qui distribuent ses produits, gèrent les contrats, conseillent et accompagnent les clients
- **Des courtiers en assurances** qui, à la différence des agents généraux, ne représentent pas la compagnie mais les clients et cherchent à leur proposer l'offre la mieux adaptée à leurs besoins. Ils peuvent donc travailler simultanément avec plusieurs organismes et faire jouer la concurrence
- **Des marques blanches** ou intermédiaires apporteurs d'affaires. Pour une entreprise, une marque blanche consiste à commercialiser sous une marque lui appartenant, des biens ou des services produits par une entreprise tierce mais qui demeure invisible aux yeux des clients, cette marque ne portant généralement pas son nom.

Plus récemment, deux nouveaux canaux de distribution ont émergé.

D'un côté, on a les bancassureurs qui s'appuient sur leurs réseaux bancaires pour développer leurs activités qui mêlent à la fois des activités bancaires classiques et des activités propres à l'assurance (perte de cartes, perte de bagages, assurance-vie, assurance automobile, etc. . .).

De l'autre, l'essor d'internet et des réseaux sociaux ont permis l'émergence de la distribution directe. Pour l'assuré, ce nouveau mode de distribution présente l'avantage de pouvoir souscrire directement via le site internet de l'organisme ou par téléphone sans avoir à passer par des intermédiaires (agents, courtiers, etc...). Pour l'assureur, cela le décharge de la contrainte du développement et de la gestion d'un réseau de proximité et lui permet ainsi de réduire ses coûts de fonctionnement et de gagner en transparence sur son positionnement. Néanmoins, si le canal direct facilite la dématérialisation des traitements et la délégation de certaines tâches au client, cette absence de contact physique complexifie la relation client et augmente le risque de fraude. De plus, ce mode de distribution engendre des coûts à l'entrée plus élevés de par la nécessité de développer des infrastructures digitales très perfectionnées et de mener de multiples campagnes marketing pour gagner en notoriété. Il existe ainsi une réelle barrière au développement sur ce type de distribution.

Par ailleurs, avec l'accès à l'information de plus en plus facilité par le développement à très grande échelle d'internet, les consommateurs ont désormais la possibilité de comparer les contrats proposés par les différents acteurs en fonction de multiples critères tels que les garanties, le tarif ou encore la nature et la qualité des prestations fournies.

Les nouvelles facilités juridiques pour changer de contrat d'assurance et l'émergence de nouveaux canaux de distributions qui simplifient la souscription d'un contrat et donnent la possibilité de déterminer la meilleure offre grâce aux comparateurs d'assurance, sont autant de raisons qui ont accru de manière phénoménale la concurrence existante sur le marché de l'assurance automobile.

Cette forte concurrence stimule les différents acteurs et les conduit à lancer des produits de plus en plus attrayants, autant au niveau des prestations et des garanties qu'au niveau des tarifs proposés, dans le but d'attirer davantage de clientèle.

Ainsi, on a vu apparaître de nouvelles armes de vente avec, par exemple, l'apparition des offres kilométriques pour s'adapter à l'utilisation que fait l'assuré de son véhicule, la création de nouvelles garanties telles que les garanties facultatives de services (assistance dépannage, protection juridique, etc...), de nouveaux services comme le prêt immédiat d'un véhicule lorsque le véhicule assuré est sinistré ou encore des réductions tarifaires sur mesure comme la création d'un bonus à vie.

En définitive, l'assurance automobile occupe une place majeure en assurance IARD puisqu'elle génère un important chiffre d'affaires. Sensible à la sinistralité et présentant un résultat technique négatif, elle n'en reste pas moins un produit central pour les assureurs car c'est le produit d'appel par excellence, renforcé par sa nature obligatoire. La concurrence sur cette branche est particulièrement rude, accentuée par la stagnation du parc automobile français et l'arrivée de nouveaux canaux de distribution qui poussent les acteurs du marché à être davantage compétitifs. Les tarifs, garanties et services proposés par le marché sont par ailleurs soumis aux comparateurs d'assurance qui permettent aux clients d'avoir une meilleure visibilité sur les offres disponibles et constituent un facteur de plus à la pression concurrentielle. Enfin, les assureurs allouent davantage de moyens pour séduire les assurés, notamment en investissant des sommes colossales dans les campagnes de publicité.

De ce fait, capter de nouveaux clients peut s'avérer être une opération bien plus coûteuse que de conserver les clients déjà présents en portefeuille.

### 1.3 Le changement de véhicule au cœur des enjeux de rétention

Le marché de l'assurance automobile étant un marché particulièrement concurrentiel, les acteurs présents sur ce marché se doivent de mettre en place des actions pour maintenir leur portefeuille de clients. Cela peut passer par deux principaux canaux : augmenter les souscriptions ou limiter les résiliations.

Si les politiques de souscriptions sont généralement favorisées par rapport aux politiques de rétention, notamment car cela permet de rester sur le devant de la scène en termes de visibilité et d'attractivité, elles représentent généralement un coût important pour l'assureur.

En effet, au moment même de la souscription d'un nouveau contrat, il réalise un effort financier afin de proposer un tarif intéressant au nouvel entrant et d'être attractif par rapport à ses concurrents. Cette concession permet une meilleure transformation et ainsi une meilleure pénétration du marché.

La prime est ensuite régulièrement réévaluée pour compenser cet effort sur toute la durée de vie du contrat.

Plusieurs types de revalorisations se distinguent :

- **Les revalorisations au terme** avec notamment l'actualisation du coefficient bonus-malus et l'indexation des coûts sur l'inflation
- **Les revalorisations liées à un avenant** sur le contrat et à une modification de ses termes ou du risque assuré
- **Les revalorisations décidées par l'assureur** dans le cadre des politiques de revalorisation

Les politiques de revalorisation des contrats sont un sujet à prendre avec précaution puisqu'elles pourraient entraîner des résiliations de la part de clients sensibles aux prix qui préféreraient alors se tourner vers la concurrence. C'est pourquoi ces dernières sont généralement associées à des modèles d'élasticité afin d'identifier les clients sensibles au prix pour lesquels une trop forte augmentation de la prime les conduiraient à clôturer leurs contrats.

En plus du coût financier à plus ou moins court terme que représente la souscription, celle-ci engendre également des coûts de publicité importants pour les compagnies d'assurance qui dépensent des montants très importants pour gagner en visibilité auprès de la clientèle.

Puisque la souscription d'un nouveau contrat représente un coût global assez élevé pour l'assureur, il est important pour lui de conserver le client dans son portefeuille le plus longtemps possible afin de rentabiliser cet investissement, à condition que ce dernier représente un "bon risque". Dans cette optique, la rétention des polices de l'assureur s'appuie sur une bonne compréhension des facteurs expliquant les résiliations.

Dans la plupart des assurances obligatoires, la résiliation est prise en charge par le nouvel assureur qui a le devoir de veiller à la continuité de la couverture de l'assuré entre l'ancienne et la nouvelle assurance.

Les résiliations peuvent se regrouper en 3 groupes distincts :

- **Les résiliations au terme** qui ont lieu à la date anniversaire du contrat et qui sont généralement liées à l'opposition de la tacite reconduction. En effet, les contrats d'assurance non-vie sont souscrits pour une durée déterminée, dans la plupart des cas pour un an, et une clause de tacite reconduction y est stipulée afin de permettre le renouvellement automatique du contrat, sauf en cas d'opposition de l'assureur ou de l'assuré. Le principal motif conduisant l'assuré à s'opposer au renouvellement de son contrat est généralement la revalorisation des primes ou bien l'attrait pour un tarif plus intéressant chez la concurrence.
- **Les résiliations hors échéance** c'est-à-dire toutes les résiliations qui ont lieu en dehors de la date anniversaire du contrat et qui sont désormais facilitées par la loi Hamon de 2015, comme évoqué précédemment.
- **Les résiliations imposées par la compagnie.** Ces dernières doivent être motivées, entre autres, par un non-paiement des cotisations, la survenance d'un sinistre, une déclaration inexacte, une omission de l'assuré, une fraude ou encore une aggravation du risque qui pousse l'assureur à rompre le contrat.

Le dernier type de résiliation étant décidé par l'entreprise, seuls les deux premiers sont d'intérêt pour les différentes politiques de rétention que peuvent mettre en place les compagnies d'assurance. Ils répondent tout deux à une déception de l'assuré qui recherche un équilibre entre le risque couvert, la relation client et le prix payé.

Les résiliations hors échéance sont un sujet particulièrement sensible puisqu'elles interviennent à des périodes aléatoires et sont donc difficiles à anticiper au contraire des résiliations au terme pour lesquelles l'entreprise peut facilement mettre en place des actions de rétention et proposer des offres à son client lorsque approche la date anniversaire du contrat.

En termes de volume, les résiliations hors échéance comptent pour plus de la moitié des départs, chiffres largement en hausse depuis la mise en vigueur de la loi Hamon. Elles représentent donc un véritable enjeu pour les organismes assureurs et illustrent parfaitement que la rétention au changement de véhicule est déterminante pour appréhender ce défi.

En effet, les raisons invoquées lors d'une résiliation hors échéance en assurance automobile sont multiples : changement de situation de l'assuré, perte du bien assuré, décès de l'assuré, etc. . . mais le changement de véhicule reste l'un des principaux motifs évoqués. Pour Generali, ce sont près d'un tiers des résiliations qui sont liées à un changement de véhicule.

Dans un objectif de rétention de leurs contrats automobiles, la question du changement de véhicule va donc représenter un enjeu particulièrement important pour les assureurs s'ils désirent garantir la stabilité et la rentabilité de leur portefeuille.

Mettre en place des modèles de scores de fragilité au changement de véhicule pourrait ainsi permettre de définir les clients susceptibles de changer de bien assuré dans un avenir proche pour ensuite implémenter des actions commerciales à leur égard afin de s'assurer qu'ils ne résilient pas à ce moment clé, d'autant plus qu'un tel changement de risque conduit généralement à une augmentation importante de la prime.

Cependant, le niveau tarifaire n'est pas la seule motivation de la résiliation ou non au changement de véhicule. En effet, si certains clients cherchent à maîtriser leurs coûts d'assurance, d'autres sont plus attachés à la stabilité de leur contrat, ce qui se remarque notamment lorsque les clients sont multi-équipés chez le même assureur.

En définitive, en raison de la concurrence accrue sur le marché de l'assurance automobile et des coûts importants que représentent les politiques de souscriptions, les actions de rétention apparaissent comme étant de première importance pour les assureurs. Ces dernières ont pour objectif de s'assurer un volume suffisant de contrats pour une bonne mutualisation mais également pour faire face à la concurrence. Ainsi, les compagnies d'assurance doivent mettre en place des politiques pour limiter les résiliations, et particulièrement celles qui sont liées à un changement de véhicule, véritable menace pour la rentabilité des portefeuilles. Cela peut se faire au moyen d'offres commerciales qui pourraient avoir un impact sur le niveau de tarification; la sensibilité au prix des clients restant assez importante.

Afin de comprendre dans quelle mesure les assureurs peuvent avoir un impact sur la prime proposée au client, notamment au moyen de réductions commerciales, il est important de comprendre comment sont tarifés les contrats en assurance automobile.

#### 1.4 Le principe de tarification

L'activité d'assurance est marquée par l'inversion du cycle de production. En effet, dans la plupart des autres industries et services, les coûts sont connus avant le montant des recettes. En revanche, en assurance, le prix de revient n'est connu qu'*a posteriori* et l'assureur doit fixer son prix de vente, la prime, avant de connaître ses coûts, les sinistres. Tout l'enjeu de la tarification est alors d'équilibrer ces deux flux, le montant des cotisations devant au moins compenser le montant estimé des flux sortants.

Pour rappel, une prime d'assurance et une cotisation désignent la même chose c'est-à-dire le coût de la police d'assurance. Cependant une légère nuance existe entre les deux : la prime d'assurance est définie comme la somme de l'ensemble des cotisations d'un assuré sur une année. Elle représente donc le coût global d'une police d'assurance pour une durée d'un an tandis qu'une cotisation représente son coût mensuel, trimestriel ou semestriel selon la fréquence des paiements choisie par l'assuré.

La prime payée par l'assuré, prime dite commerciale, se décompose comme la somme de la prime pure et des chargements moins les produits financiers :

$$\textit{prime commerciale} = \textit{prime pure} + \textit{chargements} - \textit{produits financiers} \quad (1)$$

La prime pure ou "prime de risque" est la prime permettant à l'assureur de régler les sinistres qui frappent l'ensemble des assurés. Autrement dit, c'est la prime strictement nécessaire à la compensation des risques au sein de la mutualité.

De manière simplifiée, la prime pure est calculée comme étant égale au produit de la fréquence du risque et du coût moyen d'un sinistre : c'est l'approche coût-fréquence (ou fréquence/sévérité).

On note :

- $N$  est une variable discrète à valeurs entières représentant le nombre de sinistres survenus au cours d'une période déterminée
- $X_i$ , pour  $1 \leq i \leq N$ , est une suite de variables aléatoires indépendantes et identiquement distribuées où  $X_i \geq 0$  désigne le coût du  $i$ -ème sinistre

Le montant cumulé des sinistres  $S$  peut alors se définir par :

$$S = X_1 + X_2 + \dots + X_N = \sum_{i=1}^N X_i \quad (2)$$

Ainsi, selon l'approche coût-fréquence, la prime pure peut se calculer comme :

$$\mathbb{E}[S] = \mathbb{E}[N] \times \mathbb{E}[X] \quad (3)$$

A ce montant, viennent s'ajouter divers chargements qui permettent aux assureurs de couvrir leurs coûts et dégager des bénéfices. Parmi eux, on trouve les chargements pour frais de gestion qui couvrent les frais de gestion de sinistres et les frais fixes (frais d'acquisition et d'administration, rémunération des apporteurs tels que les agents généraux et courtiers, ...); les chargements de sécurité qui permettent à l'assureur de résister à la volatilité naturelle des sinistres et de couvrir le risque de mauvaise tarification; ainsi que les chargements fiscaux qui permettent de couvrir les taxes auxquelles sont soumis les contrats d'assurance.

Les produits financiers, quant à eux, viennent alléger le montant payé par l'assuré et correspondent aux placements effectués par les assureurs sur les marchés et qui leur permettent de réaliser des plus-values.

De manière générale, la prime d'assurance peut être décomposée en 4 grandes parties :

1. **Le risque** qui représente le coût du sinistre à assurer s'il se réalisait. Il permet de déterminer le montant de la prime pure selon le profil de l'assuré
2. **Les frais** qui permettent à la compagnie de couvrir ses charges au moyen de l'application de divers chargements
3. **Le bénéfice** qui est la rémunération que s'accorde l'assureur sur le contrat. Il correspond à la marge brute qui est elle-même définie en fonction de la stratégie tarifaire et des politiques de souscription/rétention client.
4. **Les taxes** applicables sur les polices d'assurance qui sont fixées par le gouvernement, totalement indépendantes de l'assureur

La prime d'assurance automobile, comme la plupart des primes en assurance IARD, est fixée librement : l'assureur définit une prime de référence en fonction d'un tarif de base, du risque assuré, des garanties couvertes et des majorations prévues par la loi. Les risques sont ensuite réévalués régulièrement en fonction de plusieurs critères et entraînent ainsi une majoration ou non de la prime.

Aujourd'hui, la tarification en assurance automobile répond principalement au principe de segmentation du risque qui s'oppose au principe de mutualisation.

Dans l'approche par mutualisation, l'assureur raisonne en fonction de la globalité de son portefeuille pour déterminer une prime pure, ou prime de référence, unique.

A l'inverse, dans l'approche par segmentation, l'assureur répartit son portefeuille en groupes de risques homogènes et applique à chacun une prime pure ajustée en fonction du degré de risque qu'il représente. Ce principe permet de faire face au phénomène d'aléa moral en responsabilisant davantage les assurés et en les incitant à adopter des comportements prudents.

En outre, l'une des difficultés de l'activité d'assurance, particulièrement en assurance automobile, réside dans le phénomène d'anti-sélection. Ce dernier s'opère principalement lorsque la prime d'un assureur est moins segmentée que celle de ses concurrents. Dans ce cas, l'assureur en question offre la même prime aux plus ou moins bons conducteurs, tandis que ses concurrents proposent une prime différenciée. De cette façon, les bons conducteurs souscrivent chez la concurrence pour bénéficier d'un tarif adapté à leur profil tandis que les mauvais conducteurs souscrivent chez l'assureur à la prime unique car le risque qu'ils représentent est supérieur à la prime payée. Ce dernier enregistre alors une balance négative et se voit obligé d'augmenter ses prix, ce qui va l'exclure peu à peu du marché.

Ainsi, dans le but de proposer un tarif le plus adapté au risque des clients, l'assureur cherche à créer des sous-populations les plus homogènes possibles. Pour cela, il doit déterminer les facteurs d'hétérogénéité ( $\omega$ ) lui permettant de classer ses clients. En reprenant l'approche coût-fréquence présentée en (3), la prime est calculée selon le principe suivant :

$$\mathbb{E}[S|\omega] = \mathbb{E}[N|\omega] \times \mathbb{E}[X|\omega] \quad (4)$$

Pour déterminer ces facteurs, l'assureur doit utiliser des critères de tarifications c'est-à-dire qu'il doit déterminer les variables les plus significatives influençant le risque que représente la police d'assurance automobile.

Les principales variables utilisées comme critères tarifaires sont :

- **Le profil du ou des bénéficiaires du contrat** tels que l'âge du conducteur, l'ancienneté du permis, la situation matrimoniale, la catégorie socio-professionnelle, etc. . . Par exemple, la jeunesse du conducteur ou du permis sont des facteurs plutôt pénalisant dans la détermination de la prime. De plus, il est important de noter que certaines caractéristiques propres à la personne ne peuvent être prises en compte. C'est notamment le cas du sexe du client qu'il est interdit d'intégrer comme élément de tarification depuis 2012.
- **Les caractéristiques du véhicule** qui permettent à l'assureur d'estimer le risque intrinsèque du bien assuré. Les principaux critères retenus pour la tarification sont la puissance - les véhicules plus puissants étant à l'origine d'un nombre plus important d'accidents - et la classe de prix - le montant des réparations pour les garanties dommages étant directement liées à la valeur du bien.



- **L’usage du véhicule** pour lequel 5 principaux éléments sont retenus pour connaître son usage habituel ou non par le(s) conducteur(s) : l’usage socio-professionnel qui permet de connaître les conditions d’utilisation effectives, le mode de stationnement, le kilométrage du véhicule puisqu’un faible kilométrage diminue le risque de sinistre, la zone géographique car les zones à faible densité urbaine sont des zones où se produisent en moyenne moins d’accidents, et la conduite exclusive ou non.
- **Les antécédents du conducteur** jouent également un rôle important car ils participent à l’évaluation du risque de sinistre. Deux éléments font particulièrement augmenter le montant de la prime : l’utilisation du véhicule par un conducteur novice et la sinistralité aggravée pour laquelle un conducteur peut se voir appliquer une majoration sur la prime de référence dont la latitude est plafonnée par la loi :

Délits	Majorations
Alcoolémie, conduite sous emprise d’un état alcoolique	150%
Délit de fuite	100%
Multi-sinistralité (3 sinistres ou plus sur une période déterminée par l’assureur)	50%
Absence de déclaration des accidents ou des circonstances aggravantes	100%
Suspension du permis de conduire de 2 à 6 mois	50%
Suspension du permis de conduire de plus de 6 mois	100%
Plusieurs suspensions du permis de conduire au cours d’une période déterminée par l’assureur	200%
Annulation du permis de conduire	200%

D’autres critères permettent également d’adapter le tarif au profil du client et à ses attentes.

On a tout d’abord le niveau de garantie désiré par le client et les franchises qu’ils souhaitent appliquer au montant de ses indemnités. Bien entendu, plus les garanties sont étendues, plus la prime est importante tandis qu’une franchise élevée permet de réduire cette dernière.

De plus, depuis 1976, le coefficient de Réduction-Majoration (CRM) ou bonus-malus, est imposé à tous les assureurs français. Il est utilisé pour réduire ou majorer la cotisation que doit payer l’assuré selon la qualité de sa conduite et incite ainsi les automobilistes à adopter une conduite responsable. Ce coefficient s’applique sur la prime de référence pour la garantie des risques de responsabilité civile, de dommages au véhicule, d’incendies, de bris de glaces et de catastrophes naturelles. Un coefficient inférieur à 1 permet de diminuer la prime tandis qu’un coefficient supérieur à 1 l’augmente.

Comment est calculé ce coefficient ?

Un conducteur novice commence avec un coefficient égal à 1 (soit ni réduction ni majoration). Par la suite, pour chaque année sans sinistre responsable, la réduction du coefficient est de 5% avec une réduction maximale sur le tarif pouvant aller jusqu’à 50% (soit un coefficient bonus-malus de 0.5). En revanche, pour chaque sinistre responsable, la majoration est de 25% avec un plafond maximal de majoration de 250% (soit un coefficient bonus-malus de 3.5).

En plus de ces critères permettant de mieux cerner le profil de leurs assurés afin d'évaluer le risque de sinistre, et des critères imposés par la loi, les assureurs sont libres d'établir leurs tarifs en fonction de logiques qui leur sont propres. Ces logiques de tarification évoluent en fonction des objectifs commerciaux et stratégiques de l'entreprise. Par exemple, des réductions sur la prime peuvent être proposées au moment de la souscription d'un nouveau contrat dans le cadre d'un multi-équipement chez le même assureur.

Intéressons-nous aux ordres de grandeur des primes d'assurance automobile pour l'année 2020.

Selon le site de comparateur d'assurances LeLynx.fr, les Français auraient déboursé en moyenne 650€ en 2020 au titre de leur assurance automobile, toutes formules confondues. Cela marque une augmentation de 2.8% par rapport à 2019 et une hausse de 13% par rapport à 2015.

Comme nous l'avons évoqué précédemment, la zone géographique est un critère tarifant et on observe des différences assez marquées entre les régions. Ainsi, si les clients assurés en Pays de la Loire ou en Bretagne ont payé respectivement 563€ et 538€ en moyenne, les clients assurés dans les zones à risques plus élevés telles que l'Ile-de-France ou la Provence-Alpes-Côte-d'Azur ont payé respectivement 721€ et 713€ en moyenne.

Un autre critère tarifant important est l'âge du conducteur. On remarque que les Français âgés entre 18 et 25 ans payent en moyenne leur assurance automobile au prix de 1021€ par an, soit 67% de plus que la moyenne nationale. A l'inverse, les conducteurs considérés comme expérimentés et moins à risques - les seniors âgés entre 66 et 75 ans - payent quant à eux "seulement" 400€ en moyenne. Passé cet âge, le prix de l'assurance est revu à la hausse en raison des déficits physiques et cognitifs qui peuvent survenir avec la vieillesse et qui contribuent à augmenter le risque de sinistre.

## 1.5 Déroulé de l'étude

Le changement de véhicule représente un véritable défi pour les assureurs. Moment clé d'un contrat d'assurance automobile, il est à l'origine d'une part très importante des résiliations sur le portefeuille de Generali. Par ailleurs, lorsqu'un tel événement a lieu, c'est près d'un changement sur deux qui conduit à une résiliation.

Si le prix a un impact certain sur la décision d'un client de résilier son contrat d'assurance, le changement de véhicule n'est pas uniquement un défi tarifaire. En effet, au moment de modifier son contrat ou non par un avenant, de nombreux facteurs entrent en ligne de compte tels que l'expérience et la satisfaction du client avec son ancien véhicule ou encore l'attachement de l'assuré envers son assureur. Ainsi, on constate que certains clients ne sollicitent même pas leur compagnie d'assurance pour faire un devis sur un nouveau véhicule.

Les actions à mettre en place pour gérer le changement de véhicule consistent en des campagnes marketing pour inciter les clients à simuler un nouveau tarif mais également en une optimisation du tarif proposé lors du changement de leur véhicule.

Pour implémenter cela, l'assureur doit être en mesure d'identifier les clients les plus fragiles au changement de véhicule et à la résiliation afin de lancer des campagnes ciblées sur les clients qu'ils désirent conserver dans son portefeuille.

L'objet de ce mémoire sera de construire ces scores de fragilité. La méthodologie utilisée s'articule en 4 étapes principales :

1. Constitution et traitement de la base de données
2. Analyse descriptive du portefeuille et des profils de clients fragiles
3. Modélisation du changement de véhicule et de la résiliation sachant le changement de véhicule
4. Choix du score de fragilité et ciblage des clients

La première étape consiste à récupérer dans les systèmes d'information les différentes variables permettant d'identifier les changements de véhicule et les résiliations pour ce motif, ainsi que celles qui nous permettront d'établir les profils des individus les plus fragiles. Ces données peuvent être des caractéristiques classiques utilisées au moment de la tarification, mais également des variables permettant de suivre le parcours de l'individu ou des signaux clients.

La seconde étape permettra d'appréhender les caractéristiques des clients de notre périmètre d'étude. Cette analyse descriptive cherchera à souligner les tendances qui pourraient être significatives pour expliquer l'un ou l'autre des deux phénomènes étudiés.

La troisième phase consiste à mettre en place un procédé de scores quantifiant le risque de changement de véhicule ou le risque de résiliation suite à cette modification du bien assuré. Plusieurs modélisations seront testées et comparées afin de déterminer laquelle propose les meilleurs résultats et la meilleure adaptabilité à notre problématique.

Enfin, la dernière étape consistera à comprendre les caractéristiques des populations désignées comme sensibles pour ensuite proposer une stratégie de ciblage en fonction des profils des clients et de leur rentabilité.

## 2 Chapitre 2 : Éléments théoriques sur les méthodes de *machine learning*

Ce chapitre a pour objectif de présenter les différents modèles d'apprentissage statistique qui ont été utilisés au cours de ce mémoire. Après avoir expliqué le mode de construction des arbres de décision, nous nous intéressons à des méthodes ensemblistes telles que la forêt aléatoire ou le CatBoost. Cela nous permettra de comprendre comment fonctionnent ces algorithmes et mettra en lumière la nécessité de s'intéresser à des méthodes d'interprétabilité et d'explicabilité de ces modèles d'apprentissage supervisé, parfois considérés comme "boîtes noires". Enfin, la présentation d'un algorithme non supervisé de clustering permettra de disposer d'un outil statistique pour répartir une population en différents groupes homogènes en termes de caractéristiques.

### 2.1 Introduction aux arbres de classification avec l'algorithme CART

#### 2.1.1 Principe général

Les arbres de décision font partie de la famille des méthodes d'apprentissage supervisé et non paramétriques. Ils permettent d'expliquer une variable aussi bien quantitative (arbres de régression) que qualitative (arbres de classification).

L'un des algorithmes les plus répandus est l'algorithme CART (*Classification and Regression Trees*).

L'idée est de partitionner les observations de la variable cible en groupes homogènes en fonction d'une série de critères de segmentation basée sur les valeurs prises par les variables explicatives.

Pour expliquer les arbres de décision de manière simplifiée, nous nous placerons dans le cas d'arbres de classification binaire c'est-à-dire que le découpage d'un nœud est à l'origine de la création de deux branches. La méthode peut se généraliser à la régression ou à la classification à classes multiples.

Nous présentons l'algorithme le plus souvent utilisé, c'est-à-dire celui présenté dans les travaux de Breiman et al. en 1984 [3], auxquels pourra se référer le lecteur pour plus de détails.

Chaque nœud de l'arbre correspond à la segmentation d'une variable explicative tandis que les différentes branches qui découlent de cette scission représentent les différentes réponses possibles. Le principe est récursif jusqu'à l'obtention des nœuds terminaux.

A la fin de l'algorithme, lorsque toutes les questions ont été posées successivement, les différentes classes de la variable cible sont les différentes feuilles (nœuds terminaux) de l'arbre ainsi créé.

L'algorithme de l'arbre de décision fonctionne selon le principe suivant :

- On sépare les individus en deux sous-ensembles homogènes pour expliquer la variable de sortie. L'objectif est de trouver la variable d'entrée qui fournit la meilleure explication. Cette dernière est celle dont la division maximise le critère de segmentation sélectionné. Les sous-ensembles ainsi créés définissent les nœuds de l'arbre et à chacun d'entre eux est associée une mesure de proportion qui permet d'expliquer l'appartenance à une classe ou l'explicabilité de la variable de sortie.
- L'étape précédente est répétée de manière récursive : chaque sous-population est à nouveau divisée selon la variable explicative la plus pertinente

- Le processus s'arrête soit :
  - Lorsque le nœud est homogène, en d'autres termes il n'y a pas de division encore admissible
  - Lorsque le nombre d'observations présentes dans ce nœud est inférieur à la valeur du seuil qui a été fixé

On obtient alors les nœuds terminaux, appelés "feuilles" de l'arbre

La difficulté réside dans le fait de trouver le modèle optimal qui minimise les critères de segmentation et permet d'extrapoler les résultats obtenus sur une base d'apprentissage à une base de validation, c'est-à-dire un modèle qui se généralise bien à de nouvelles données.

La construction d'un arbre nécessite donc un processus de sélection des divisions successives, un critère de décision pour la réalisation d'une nouvelle coupe ou l'arrêt du processus et un mécanisme d'attribution des classes à chaque feuille de l'arbre.

Par la suite, on considérera une base de données composée de  $p$  variables explicatives ( $X \in \mathbb{R}^p$ ,  $p \in \mathbb{N}$ ) et d'une variable catégorielle  $Y \in \{1, \dots, K\}$  que l'on cherche à expliquer ( $K \in \mathbb{N}$ ). On suppose également que l'on dispose d'un échantillon  $y : (y_1, \dots, y_n) \in \mathbb{R}^n$  correspondant aux valeurs cibles et  $x = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$  correspondant aux variables explicatives (avec  $n \in \mathbb{N}$  le nombre d'individus).

On note :

- $I$  l'ensemble des individus et  $i \in I$  un individu
- $C$  une partition de l'échantillon et  $t \in C$  un nœud
- $N(t)$  le nombre d'individus dans le nœud  $t$
- $N_k(t)$  le nombre d'individus de la classe  $k$  dans le nœud  $t$
- $p(k|t) = N_k(t)/N(t)$  la proportion d'individus de classe  $k$  dans le nœud  $t$
- $y(t)$  la prédiction faite pour les individus du nœud  $t$

La construction de l'arbre CART se fait en suivant la procédure suivante :

- Détermination de l'arbre maximal ou arbre saturé
- Élagage et construction de l'arbre optimal

### 2.1.2 Construction de l'arbre maximal

La première partie de l'algorithme CART consiste à créer un arbre dit maximal, ou arbre saturé.

Préalablement, définissons l'impureté. Un nœud est dit pur si une seule classe  $y$  est représentée. Il faut définir une fonction qui permet de mesurer le degré d'impureté ou d'hétérogénéité d'un nœud. Ainsi, pour un nœud  $t$  donné, la fonction positive  $Imp$  renvoie un réel positif telle que :

$$\phi : [0; 1]^k \rightarrow \mathbb{R}, \quad Imp(t) = \phi(\{p(k|t), k \in K\}) \quad (5)$$

Cette fonction  $\phi$  vérifie les conditions suivantes :

- Son maximum est atteint uniquement lorsque les classes sont équi-réparties c'est-à-dire au point  $(1/K, \dots, 1/K)$ . Dans ce cas, les valeurs que peuvent prendre la variable cible ont la même probabilité

- Son minimum est atteint lorsque le nœud est pur, c'est-à-dire lorsque la valeur de la variable cible est similaire pour tous les individus : la population est donc homogène.
- Elle est symétrique, c'est-à-dire qu'elle ne privilégie pas une classe plutôt qu'une autre

Les mesures les plus utilisées pour mesurer l'impureté sont l'entropie ou l'indice de Gini, qui sont définies pour un nœud  $t$  par :

$$Ent(t) = \sum_{k \in K} p(k|t) \log(p(k|t)) \quad (6)$$

$$Gini(t) = \sum_{k=1}^K p(k|t)[1 - p(k|t)] \quad (7)$$

En utilisant ces critères, comment est divisé un nœud ?

Soit une variable explicative  $x_j$  ( $j \in \{1, \dots, p\}$ ) que l'on souhaite utiliser pour diviser un nœud donné (appelé "nœud mère") en deux sous-nœuds (appelés "nœuds fils"), que l'on notera  $t_G$  et  $t_D$  (résultant des branches gauche et droite de la division du nœud).

Dans le cas où cette variable est quantitative, on chercherait un réel  $z$  qui sépare au mieux le nœud  $t$  en deux sous-nœuds :

$$t_G(z) = \{w \in t, x_j(w) \leq z\} \quad (8)$$

$$t_D(z) = \{w \in t, x_j(w) > z\} \quad (9)$$

En notant  $d_z(t)$  la division du nœud  $t$  associé au réel  $z$ , on cherche donc  $z$  qui maximise la réduction de l'impureté, c'est-à-dire la différence entre l'impureté du nœud  $t$  et la somme de l'impureté des deux sous-nœud obtenus :

$$z \longrightarrow \Delta Imp(d_z(t), t) = Imp(t) - \left( \frac{N(t_G(z))}{N(t)} Imp(t_G(z)) + \frac{N(t_D(z))}{N(t)} Imp(t_D(z)) \right) \quad (10)$$

Autrement dit, on cherche à maximiser l'homogénéité des nœuds obtenus, un nœud étant parfaitement homogène s'il ne contient que des individus de la même classe de la variable cible.

Pour créer l'arbre maximal, on part de la racine de l'arbre (un arbre initialement vide) et on divise les données en sélectionnant la variable explicative qui permet de maximiser la réduction de l'impureté de ce nœud. Cette règle de division permet d'obtenir à chaque étape des groupes de données les plus homogènes possibles.

Sur chacun des sous-nœuds ainsi formés on remet en place le même processus, et ainsi de suite. Il faut fixer un critère pour fixer l'arrêt de l'algorithme. On en retient deux :

- Ne pas découper un nœud qui contient moins d'un certain nombre d'individus, nombre fixé au moment du paramétrage du modèle
- Ne pas découper un nœud pur, c'est-à-dire qu'un nœud pur est nécessairement une feuille de l'arbre.

A la fin de l'algorithme, on a ainsi créé l'arbre maximal  $T_{max}$ . Néanmoins, l'arbre qui sera utilisé pour la prédiction est généralement une version élaguée de cet arbre saturé.

### 2.1.3 Élagage de l'arbre maximal

L'arbre saturé obtenu a tendance au sur-ajustement : il reproduit avec une bonne précision les valeurs de la base d'apprentissage mais présente une fragilité au moment de prédire la variable cible sur la base de validation.

L'objectif de l'élagage de l'arbre est d'aboutir à un modèle certes moins précis et complexe sur la base d'apprentissage mais dont les performances prédictives sur de nouvelles données sont plus stables et robustes.

Cette technique consiste, en partant des feuilles de l'arbre vers les racines, à supprimer certaines séparations (nœuds) afin de regrouper des données. Ces dernières sont des scissions qui n'améliorent pas de manière significative l'arbre sur la base de validation.

De manière générale, les algorithmes supervisés visent à obtenir la meilleure performance possible lors de leur construction, évaluée selon une fonction objectif. Lors de la construction de l'arbre maximal, la fonction qui est maximisée au moment de la scission des nœuds est la réduction de l'impureté calculée au moyen de l'indice de Gini. Cependant, cette mesure ne tient pas compte de la complexité du modèle.

Il est alors nécessaire de définir une mesure d'erreur qui sera une notion fondamentale pour comprendre l'élagage de l'arbre saturé : l'erreur de complexité.

Tout d'abord, le taux d'erreur  $R(t)$  au sein d'un nœud  $t$  peut se définir ainsi :

$$R(t) = \sum_{k \in K, k \neq y(t)} p(k|t) \quad (11)$$

Autrement dit, il s'agit de la proportion des individus qui sont présents dans le nœud  $t$  et qui n'ont pas la même valeur de la variable cible que la prédiction faite par le nœud.

On peut généraliser cette notion à l'ensemble de l'arbre et non plus à un unique nœud. Si l'on note  $\tau$  l'ensemble des nœuds terminaux (i.e. feuilles) de l'arbre  $T$ , le taux d'erreur  $R(T)$  de cet arbre est :

$$R(T) = \sum_{t \in \tau} \frac{N(t)R(t)}{n} = 1 - \frac{1}{n} \sum_{t \in \tau} N_{y(t)}(t) \quad (12)$$

Autrement dit, il s'agit de la somme des taux d'erreur de chacun des nœuds terminaux, pondérés par la proportion des individus qui sont présents au sein de chaque nœud.

Néanmoins, ce taux d'erreur ne prend pas en compte la complexité de l'arbre, qui peut notamment être mesurée par sa profondeur. En effet, pour éviter le sur-apprentissage, on cherche en général à ne pas avoir un arbre trop profond. Pour y remédier, on introduit un terme de pénalité pour favoriser les arbres peu complexes.

En introduisant un réel positif  $\alpha$  pour représenter ce paramètre de complexité, le taux d'erreur pénalisé d'ordre  $\alpha$  de l'arbre  $T$  ( $R_\alpha$ ) peut se définir comme :

$$R_\alpha(T) = R(T) + \alpha|T| \quad (13)$$

Cette mesure pénalise les arbres avec beaucoup de nœuds terminaux (feuilles). En effet, le paramètre  $\alpha$  permet de quantifier la complexité du modèle :

- Si  $\alpha$  est nul, l'arbre maximal  $T_{max}$  minimise l'erreur de complexité
- Un  $\alpha$  élevé va favoriser des arbres "simples" contenant peu de subdivisions et donc de feuilles

En prenant en compte ces mesures, comment l'arbre maximal est-il élagué ?

En reprenant les notations précédentes, c'est-à-dire que  $T$  désigne un arbre et  $t$  un nœud de cet arbre, on peut noter  $T(t)$  le sous-arbre issu du nœud  $t$ .

Étant donné le paramètre de complexité  $\alpha$ , l'élagage de l'arbre est jugé utile si le taux d'erreur pénalisé du sous-arbre  $T(t)$  est plus grand que celui du nœud  $t$ , c'est-à-dire si  $R_\alpha(T(t)) > R_\alpha(t)$ . Autrement dit, si la division d'un nœud fait augmenter le taux d'erreur du modèle.

En outre, l'objectif est d'optimiser les paramètres du modèle qui minimisent la fonction objectif, ici le taux d'erreur pénalisé d'ordre  $\alpha$ . Il est donc nécessaire de trouver la valeur optimale du paramètre de complexité.

$$R_\alpha(T(t)) \geq R_\alpha(t) \iff R(T(t)) + \alpha|T(t)| \geq R(t) + \alpha \iff \alpha \geq \frac{R(t) - R(T(t))}{|T(t)| - 1} \quad (14)$$

Si on note :

$$g(t, T) = \frac{R(t) - R(T(t))}{|T(t)| - 1}$$

On appelle alors maillon faible  $t_{min}$  le nœud qui minimise cette fonction :

$$t_{min} = \arg \min_{t \in T} g(t, T) \quad (15)$$

Le paramètre de complexité choisi est alors :  $\alpha = g(t_{min}, T)$

De manière générale, l'objectif de l'élagage de l'arbre est donc de déterminer une suite de sous-arbres optimaux qui minimisent l'erreur de complexité d'ordre  $\alpha$ , avec  $\alpha$  qui varie en partant de 0. Dans cette suite d'arbres optimaux, l'algorithme détermine l'arbre optimal en le testant sur une base de validation. C'est l'arbre qui minimise l'erreur sur la base de validation qui sera retenu.

On peut résumer de manière simplifiée cet algorithme d'élagage qui part des  $k$  feuilles de l'arbre maximal et qui renvoie une liste d'arbres issus de  $T_{max}$ .

- Entrée : arbre maximal  $T_{max}$
- Initialisation :  $k = 0$ ,  $\alpha_0 = 0$  et  $T_0 = T_{max}$



- Itération : Tant que  $|T_{max}| > 1$  :
  - On calcule le maillon faible de l'arbre  $T_k : t_{min}^{k+1} = \arg \min_{t \in T_k} g(t, T_k)$
  - On calcule le paramètre de complexité correspondant  $\alpha_{k+1} = g(t_{min}^{k+1}, T_k)$
  - On construit l'arbre  $T_{k+1}$  à partir de  $T_k$ , en élaguant toutes les branches qui sont issues de  $t_{min}^{k+1}$
  - On répète l'opération avec  $k = k + 1$
- Outputs :  $\{\alpha_1, \dots, \alpha_N\}$  et  $\{T_{max}, T_1, \dots, T_N\}$  avec  $N \in \mathbb{N}$  le nombre d'itérations de l'algorithme

L'arbre que l'on retient est l'arbre  $T_k$  qui minimise le taux erreur pénalisé  $R_{\alpha_k}(T_k)$ . Si l'on représente la courbe des taux d'erreur pénalisés d'ordre  $\alpha$  ( $R_{\alpha_i}(T_i)$ ) en fonction de  $\alpha_i$  pour  $i = 1, \dots, N$ , l'arbre optimal est celui dont le coefficient de complexité permet à la courbe  $(R_{\alpha_i}(T_i), \alpha_i)$  d'atteindre son minimum.

#### 2.1.4 Avantages et inconvénients de la méthode CART

Les algorithmes CART présentent de nombreux atouts :

- Le traitement des données devient plus simple. Contrairement aux méthodes traditionnelles, les arbres sont en capacité de segmenter les variables qui sont continues pour chaque nœud. Ainsi, une base de données comportant à la fois des variables quantitatives et qualitatives pourra être très bien gérée par ces méthodes.
- Ils sont non-paramétriques : aucune hypothèse de linéarité n'est formulée en amont. Les modèles sont donc capables de capter les effets non linéaires de la variable cible et des variables explicatives.
- La présentation des résultats sous forme d'arbre permet une représentation visuelle et une bonne interprétation des résultats

Cependant, ils présentent également quelques limites :

- Les arbres peuvent avoir une forte dépendance aux données d'apprentissage et donc une variance élevée. De plus, ils peuvent également dépendre de la profondeur de la base d'apprentissage et une modification de cette dernière pourrait modifier leur structure et impacter leurs prédictions. On est donc confronté à un risque de sur-apprentissage qui peut empêcher les modèles de se généraliser.
- Le critère de segmentation a tendance à sélectionner les variables possédant un grand nombre de modalités. De ce fait, l'exploitation d'une base de données dont les classes sont asymétriques conduit à un risque de prédictions fausses.
- L'élagage peut être délicat

## 2.2 La forêt aléatoire

### 2.2.1 Description de la méthode

Comme nous venons de le mentionner, les arbres de décisions CART souffrent de plusieurs faiblesses et notamment d'une instabilité intrinsèque : une légère modification de la base d'apprentissage peut modifier toute la structure de l'arbre.

Pour pallier ce problème et améliorer la précision et la stabilité des modèles obtenus, des méthodes ensemblistes basées sur des techniques d'échantillonnage ont été proposées. Elles consistent à développer de nombreux arbres sur des sous-ensembles de données légèrement différents, pour ensuite les agréger et donner une prédiction de la variable cible.

Les forêts aléatoires (*random forests*, en anglais), introduites par Breiman en 2001 [5], figurent parmi les méthodes d'ensemble d'arbres les plus connues. L'idée est de conserver les avantages de la méthode CART tout en remédiant à ses limites, principalement la dépendance à la base d'apprentissage, l'effet de sur-ajustement du modèle et de ce fait, la complexité de la phase d'élagage.

Plutôt que d'utiliser un unique arbre relativement complexe pour réaliser entièrement la tâche prédictive, elles permettent de construire de nombreux estimateurs individuels (arbres), qui une fois regroupés (d'où le terme "forêt"), permettront de fournir une réponse globale à la même problématique à laquelle chacun tentait de fournir une solution.

Bien que plus difficiles à interpréter que les arbres décisionnels uniques, les forêts aléatoires sont plus performantes que ces derniers. De plus, leur rapidité d'apprentissage, leur robustesse et leur flexibilité en font une méthode largement utilisée.

Cette méthode correspond en fait à la réunion de deux algorithmes : le "Tree Bagging" et le "Feature sampling".

### 2.2.2 Tree Bagging

Le "Bagging" ou "Bootstrap AGGREGatING" est un concept introduit par Breiman en 1994 [4]. C'est un méta-algorithme d'apprentissage ensembliste conçu pour améliorer la stabilité et la précision des modèles de *machine learning*, tout en réduisant la variance et en évitant le sur-apprentissage.

La technique consiste à créer différents échantillons bootstrap, à entraîner des arbres de décision sur chaque échantillon, puis à combiner les prédictions directement à l'aide d'une statistique.

La création des échantillons à l'aide de la méthode bootstrap consiste à sélectionner des exemples au hasard avec remise dans l'ensemble des données d'apprentissage. La remise permet à une observation d'être potentiellement sélectionnée dans plusieurs échantillons.

Un arbre de décision est ensuite ajusté sur chaque échantillon de données. Chacun d'entre eux sera un peu différent étant donné les différences dans l'ensemble des données d'entraînement. Ces différences entre les arbres sont souhaitables car elles augmentent la "diversité" de l'ensemble, ce qui signifie que les membres de cet ensemble ont une corrélation plus faible dans leurs prédictions ou leurs erreurs de prédiction.

L'un des avantages de la mise en place du Bagging est qu'il ne surajuste généralement pas l'ensemble de données d'apprentissage et que le nombre de membres de l'ensemble peut continuer à être augmenté jusqu'à ce que les performances sur l'ensemble de données de validation cessent de s'améliorer.

Le Tree Bagging utilise cette notion et permet d'assembler des arbres de décision construits à partir d'observations tirées aléatoirement.

Si l'on reprend les notations précédentes, la base d'apprentissage est constituée des couples entrée/sortie  $(X, Y)$  avec  $X$  une matrice de taille  $n \times p$  ( $n$  observations et  $p$  variables explicatives) et  $Y$  un vecteur de taille  $n$ .

$$X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

$$Y = \begin{pmatrix} 0 \\ 1 \\ \dots \\ 1 \\ 0 \end{pmatrix}$$

L'algorithme consiste à générer un nombre  $B$  d'arbres de décision de la manière suivante :

- Tirer aléatoirement avec remise  $B$  échantillon de  $(X, Y)$ , que l'on note  $(X_b, Y_b)$
- Un arbre de décision est entraîné sur chacun des couples  $(X_b, Y_b)$

On obtient ainsi  $B$  arbres qui sont à priori différents les uns des autres car ils ont été entraînés sur des données différentes. Pour prédire la variable de sortie sur de nouvelles données, la majorité est prise parmi l'ensemble des réponses (cas d'une classification).

### 2.2.3 Feature sampling

Dans l'expression "Random Forest", "Forest" renvoie à l'assemblage des arbres de décisions tandis que "Random" renvoie au double échantillonnage sur les données d'entrée et sur les variables utilisées.

Ainsi, le Feature Sampling correspond au tirage aléatoire des variables explicatives et se rapproche également de la notion de Bagging. Néanmoins, au lieu de tirer des sous-échantillons de la base d'apprentissage pour entraîner les arbres, il propose des variables explicatives sélectionnées aléatoirement.

Cette méthode consiste donc à construire différents arbres de décisions en ne donnant à chacun d'entre eux qu'un nombre restreint de variables explicatives, tirées de manière aléatoire. Elle a pour but de réduire la variance du modèle final.

### 2.2.4 L'algorithme de la forêt aléatoire

Pour résumer, la forêt aléatoire peut être vue comme une méthode combinant les principes du Tree Bagging et du Feature Sampling. L'objectif est de rajouter de l'aléa dans la construction des arbres de décision CART.

La proposition de Breiman [5] peut se résumer de la manière suivante :

- On crée  $B$  nouveaux ensembles d'apprentissage par un double processus d'échantillonnage :
  - Sur les observations : on procède à la création de sous-échantillons obtenus par tirage avec remise d'un nombre  $N$  d'observations, de même taille que l'échantillon initial (bootstrap)
  - Sur les  $p$  variables explicatives : on détermine le nombre de variables utilisées ( $m$ ) pour la segmentation de chaque arbre :
    - ◇ En classification :  $m \leq \sqrt{p}$
    - ◇ En régression :  $m \leq p/3$
- Sur chacun des  $B$  échantillons, on construit un arbre de décision sans élagage, en limitant sa croissance par validation croisée. Le but est de trouver la meilleure partition issue des  $m$  variables explicatives qui ont été sélectionnées.
- On conserve en mémoire les  $B$  prédictions de la variable cible pour chaque observation d'origine
- La prédiction finale de la forêt aléatoire est alors un simple "vote à la majorité" :
  - ◇ En classification : classe la plus représentée parmi les  $B$  variables arbres
  - ◇ En régression : moyenne empirique des valeurs prédites avec les  $B$  arbres

De cette manière, chaque arbre généré par cette méthode a une vision parcellaire de la problématique, avec des variables explicatives et des observations différentes.

### 2.2.5 Avantages et inconvénients de la forêt aléatoire

L'algorithme de la forêt aléatoire présente de nombreux avantages, parmi lesquels :

- Il permet d'éviter en partie le sur-ajustement
- Pour l'analyse et l'interprétation des résultats, il donne un aperçu du pouvoir prédictif du modèle et permet d'avoir une première idée sur les variables importantes pour expliquer le phénomène étudié
- Il est relativement efficace en termes de précision

Néanmoins, il présente également certaines limites :

- Les temps d'apprentissage peuvent être longs si les bases de données sont importantes et composées de multiples variables. Un traitement préalable est généralement nécessaire pour éviter de fournir au modèle des variables non pertinentes.
- Pour un problème de régression, la prédiction des valeurs extrêmes peut s'avérer instable
- L'agrégation de multiples arbres en forêts fait perdre l'aspect visuel d'un arbre unique

## 2.3 Le CatBoost

### 2.3.1 Description de la méthode

CatBoost est un algorithme d'apprentissage automatique développé par Yandex en 2017 [13]. Démontrant ses performances au fil de compétitions kaggle, il est très efficace lorsque les bases de données contiennent des variables catégorielles qui jouent un rôle prépondérant.

Il s'appuie notamment sur l'amplification du gradient qui est un algorithme dans lequel sont construits des prédicteurs simples qui améliorent la fonction objectif. Autrement dit, plutôt que de construire immédiatement un modèle complexe, de nombreux petits modèles sont implémentés.

Par ailleurs, la différence avec les autres algorithmes d'amplification du gradient (Gradient Boosting) est qu'il traite avec succès les variables catégorielles et utilise un nouveau schéma pour calculer les valeurs des feuilles lors de la sélection de la structure de l'arbre, ce qui permet de réduire le sur-ajustement.

En effet, la plupart des implémentations d'amplification du gradient utilisent des arbres de décision comme prédicteurs de base. Cependant, si ces derniers sont utiles pour les variables numériques, dans la pratique, de nombreux jeux de données contiennent également des variables catégorielles importantes pour la compréhension du phénomène étudié. Ces dernières n'étant pas nécessairement comparables les unes aux autres, il faut souvent effectuer un traitement sur les données pour convertir les valeurs catégorielles en valeurs numériques avant l'étape d'apprentissage. Avec le modèle CatBoost, cette étape n'est plus nécessaire car l'algorithme effectue automatiquement ce traitement.

Comparativement aux autres modèles, Catboost introduit deux avancées uniques :

- La mise en œuvre du Boosting ordonné (Ordered Boosting) qui constitue une alternative à l'approche du Boosting classique
- Un processus de traitement des variables catégorielles, notamment au moyen de l'encodage cible (Target Encoding)

Ces deux techniques permettent de lutter contre le décalage de prédiction causé par un type particulier de "fuites de cible" (*target leakage* en anglais) présent dans toutes les implémentations existantes des algorithmes d'amplification du gradient.

### 2.3.2 Introduction au Boosting

#### 2.3.2.1 Boosting

Les algorithmes de Boosting figurent parmi les algorithmes les plus populaires et les plus utilisés. Ils peuvent être considérés comme l'une des techniques les plus puissantes pour construire des modèles prédictifs.

Se basant sur une agrégation d'arbres – tout comme la forêt aléatoire -, la méthode consiste à générer des arbres en séries, c'est-à-dire que chaque arbre généré (à l'exception du premier) prend en compte les résultats de son prédécesseur et notamment l'erreur commise. Le nouvel arbre généré (apprenant faible) aura alors pour objectif de corriger et combler les lacunes de son prédécesseur en donnant plus de poids aux données mal ajustées.

Le Boosting concentre ainsi ses efforts sur les observations qui sont les plus difficiles à prédire tandis que l'agrégation de l'ensemble des modèles permet d'éviter le sur-apprentissage.

Différents algorithmes de Boosting existent et diffèrent selon plusieurs caractéristiques : la manière de pondérer les données mal-ajustées pour renforcer l'apprentissage, l'objectif (régression/classification), la fonction de perte pour mesurer l'ajustement du modèle ou encore la façon d'agrèger les modèles successifs. Les deux principaux sont le Boosting adaptatif (AdaBoost) et l'amplification du gradient (Gradient Boosting). Les modèles XGBoost, LightGBM et CatBoost sont fondamentalement des implémentations différentes de ce dernier.

### 2.3.2.2 Gradient Boosting

Introduit par Jerome Friedman en 2001 [6], le Gradient Boosting, comme son nom l'indique, combine la méthode du Boosting et de la descente du gradient.

Autrement dit, cette méthode redéfinit le Boosting comme un problème d'optimisation numérique où l'objectif est de minimiser la fonction de perte du modèle en ajoutant des apprenants faibles et en utilisant une procédure de type descente du gradient. Plusieurs types de fonctions de perte pouvant être utilisées, cette technique peut aussi bien être appliquée à une régression qu'à une classification.

La différence entre le Gradient Boosting et AdaBoost – algorithme que nous ne développerons pas dans ce mémoire - réside dans la gestion des valeurs sous-ajustées de son prédécesseur. Contrairement à AdaBoost qui ajuste les poids d'instance à chaque interaction, le Gradient Boosting essaie d'adapter le nouveau prédicteur aux erreurs résiduelles commises par le prédicteur précédent.

Fondamentalement, cet algorithme implique donc trois éléments : une fonction de perte à minimiser, un apprenant faible pour faire des prédictions et un modèle additif pour ajouter des apprenants faibles afin de minimiser la fonction de perte.

On note  $B_k$  l'arbre construit à l'étape  $k$  et  $f_k = \sum_{j=1}^k B_j$  le modèle construit à l'étape  $k$ . La base d'apprentissage est toujours composée des couples entrée/sortie  $(X, Y)$ . Pour simplifier, on notera  $X_i$  toutes les données de la ligne  $i$  de la matrice  $X$ . A l'étape  $k$ , le modèle nous donne la prédiction  $f_k(X)$  de  $Y$ .

Tout d'abord, nous allons présenter le cas de la régression pour lequel la forme de la fonction de coût donne des résultats facilement interprétables. Ainsi, nous prenons comme fonction de coût la fonction des moindres carrés définie de la manière suivante :

$$J = \sum_{i=1}^p j(Y_i, f(X_i)) = \sum_{i=1}^p \frac{[Y_i - f(X_i)]^2}{2} \quad (16)$$

Le gradient de  $J$  par rapport à  $f(X_i)$  est :

$$\nabla J = \frac{\partial J}{\partial f(X_i)} = \frac{\partial \sum_{l=1}^p j(Y_l, f(X_l))}{\partial f(X_i)} = \frac{\partial \sum_{l=1}^p \frac{[Y_l - f(X_l)]^2}{2}}{\partial f(X_i)} = f(X_i) - Y_i \quad (17)$$

L'algorithme du Gradient Boosting opère de la manière suivante :

- Tout d'abord, il construit un arbre de décision  $B_1$ , visant à prédire  $Y$  à partir de  $X$ . On a alors  $B_1(X_i) = Y_i + [B_1(X_i) - Y_i]$ . La valeur observée des insuffisances du modèle sont appelés les résidus. Pour  $B_1$  on a  $res_1 = -[\partial J / \partial f(X_i)] = Y_i - f_1(X_i)$
- Pour combler les lacunes du premier arbre, il construit  $B_2$  qui vise à prédire les résidus de l'étape précédente c'est à dire  $-\partial J / \partial f(X_i)$ . Une fois le deuxième arbre construit, le modèle est donc égal à  $f_2(X) = B_1(X) + B_2(X)$ . Nous avons  $B_2(X_i) = [Y_i - f_1(X_i)] + [f_2(X_i) - Y_i]$ . Les résidus de cette seconde étape sont  $res_2 = -[\partial J / \partial f(X_i)] = Y_i - f_2(X_i)$
- Pour combler les lacunes du second arbre, il construit  $B_3$ , qui vise à prédire  $res_2$
- Cette procédure est itérée, en suivant la descente du gradient. A l'étape  $k$ , on a  $res_{k-1} = -[\partial J / \partial f(X_i)] = Y_i - f_{k-1}(X_i)$

Dans le cas où la variable cible est catégorielle avec  $K$  modalités (de 1 à  $K$ ), l'algorithme reste globalement identique mais il faut définir une fonction de coût adaptée à la classification et en dériver le gradient.

$$y_i^k = \begin{cases} 1 & \text{si } Y_i = k \\ 0 & \text{sinon} \end{cases} \quad (18)$$

$$j(y_i, f(x_i)) = - \sum_{k=1}^K y_i^k \log(p^k(x_i)) \quad (19)$$

Où  $p^k(x_i)$  correspond à la probabilité conditionnelle d'appartenir à la classe  $k$ , c'est-à-dire :

$$p^k(x_i) = \frac{\exp(f^k(x_i))}{\sum_{k=1}^K \exp(f^k(x_i))} \quad (20)$$

Dans ce cas, le gradient pour la classe  $k$  correspond à l'écart entre l'indicatrice correspondante et la probabilité d'appartenance à cette classe :

$$\nabla J = y_i^k - p^k(x_i) = y_i^k - \frac{\exp(f^k(x_i))}{\sum_{k=1}^K \exp(f^k(x_i))} \quad (21)$$

L'amplification du gradient est une technique populaire en raison de sa précision et de sa rapidité, en particulier pour les données complexes et volumineuses. Néanmoins, c'est un algorithme glouton qui peut facilement surcharger un ensemble de données d'apprentissage. Il peut alors bénéficier de méthodes de régularisation qui pénalisent diverses parties de l'algorithme et améliorent généralement les performances de l'algorithme en réduisant le sur-apprentissage.

### 2.3.2.3 Ordered Boosting

Les gradients utilisés à chaque étape sont estimés en utilisant les valeurs cibles des mêmes observations que celles sur lesquelles le modèle actuel est construit. Cela entraîne un décalage de la distribution des gradients estimés dans n'importe quel domaine de l'espace des caractéristiques par rapport à la distribution réelle des gradients dans ce même domaine, ce qui entraîne un sur-ajustement (décalage de prédiction).

Pour y remédier et obtenir des résidus "non décalés", CatBoost échantillonne un nouveau jeu de données indépendamment à chaque étape du Gradient Boosting et applique le modèle actuel aux nouveaux exemples d'apprentissage.

Supposons que nous entraînons un modèle avec  $I$  arbres. Pour que les résidus  $res_{I-1}$  ne soient pas décalés, nous devons avoir entraîné  $f_{I-1}$  sans les instances originales.

Comme des résidus non biaisés sont nécessaires pour tous les échantillons d'entraînement, aucun exemple ne peut être utilisé pour l'entraînement de  $f_{I-1}$ , ce qui, à première vue, rend le processus d'entraînement impossible.

Cependant, CatBoost maintient un ensemble de modèles qui diffèrent par les observations utilisées pour l'apprentissage. Ensuite, pour calculer les résidus sur un exemple, on utilise un modèle entraîné sans celui-ci.

Pour illustrer l'idée, supposons que nous prenons une permutation aléatoire  $\sigma$  des exemples d'apprentissage et que nous maintenons  $n$  modèles "de soutien" différents  $M_1, \dots, M_n$  de sorte que le modèle  $M_i$  soit appris en utilisant uniquement les  $i$  premiers exemples de la permutation. À chaque étape, afin d'obtenir les résidus pour le  $j$ -ème échantillon, CatBoost utilise le modèle  $M_{j-1}$ .

On parle alors de Boosting ordonné (Ordered Boosting).

### 2.3.3 Le Target Encoding

De nombreux algorithmes d'apprentissage automatique nécessitent que les données soient numériques. Ainsi, avant la phase d'apprentissage, il faut convertir les données catégorielles sous forme numérique. Il existe plusieurs méthodes d'encodage catégoriel.

CatBoost se distingue des autres algorithmes, et notamment de ceux utilisant aussi l'amplification du gradient, par sa solution originale pour l'encodage des caractéristiques catégorielles.

Soit  $\rho$  une constante, définie au moment du paramétrage du modèle (*one hot max size*). Si le nombre de modalités d'une variable est inférieur ou égal à  $\rho$  alors on applique un encodage one-hot (*one-hot encoding* en anglais) à la variable catégorielle. Cependant, si le nombre de modalités est supérieur, CatBoost applique l'encodage cible (*target encoding* en anglais) qui est une solution d'encodage qui réduit le sur-apprentissage.

L'encodage one-hot consiste à encoder une variable à  $n$  modalités sur  $n$  bits dont un seul prend la valeur 1, le numéro du bit valant 1 étant le numéro de l'état pris par la variable. Cet encodage est courant en apprentissage automatique où l'on représente usuellement une variable catégorielle à  $n$  modalités par  $n$  variables binaires, la  $i$ -ème variable binaire représentant la  $i$ -ème catégorie.



L'encodage cible consiste à remplacer les valeurs de la variable catégorielle par l'estimation de la cible attendue  $y$  conditionnée par la catégorie. L'approche la plus simple consiste à utiliser la valeur moyenne de la cible sur les observations appartenant à la modalité.

Dans CatBoost, l'encodage cible n'est réalisé qu'avec des variables cibles binaires. Dans le cas où le problème n'est pas une classification binaire, la variable cible est transformée en plusieurs variables binaires par encodage one-hot et, pour chacune de ces variables, un encodage cible est réalisé.

Néanmoins, il s'avère que l'application de cette méthode sans précaution peut entraîner des "fuites de cible" puisque la cible est utilisée pour prédire la cible. Cela peut ainsi conduire au sur-apprentissage.

Pour pallier ce problème, une solution possible est le *holdout target encoding*. Avec cette approche, une partie de la base d'entraînement est utilisée pour calculer les valeurs cibles pour chaque modalité, et l'apprentissage est effectué sur le reste des données. Cela résout le problème de la fuite de cible mais contraint à sacrifier une partie des données d'entraînement.

Pour cette raison, les solutions les plus utilisées dans la pratique sont l'encodage cible  $K - folds$  et l'encodage cible *Leave-one-out*.

L'idée derrière l'encodage cible  $K - folds$  est très similaire à la validation croisée  $K - folds$  : les données d'entraînement sont divisées en plusieurs échantillons et dans chacun d'entre eux, les valeurs des caractéristiques catégorielles sont remplacées par les statistiques cibles calculées sur les autres échantillons. L'encodage cible *Leave-one-out* est un cas particulier d'encodage  $K - folds$  où  $K$  est égal à la longueur des données d'apprentissage.

Cependant, elles n'empêchent pas complètement le sur-apprentissage.

CatBoost utilise une stratégie plus efficace qui s'appuie sur le principe d'ordonnement et s'inspire des algorithmes d'apprentissage en ligne qui obtiennent des exemples d'entraînement de manière séquentielle dans le temps. Ainsi, les valeurs de la variable cible pour chaque observation ne dépendent que de l'historique observé.

Autrement dit, il introduit un "temps" artificiel : une permutation aléatoire  $\sigma_1$  de l'ensemble des données d'apprentissage est effectuée. Ensuite, un encodage cible d'un certain type est effectué sur chaque observation en utilisant uniquement les objets qui sont placés avant l'objet courant.

Il est intéressant de remarquer qu'en utilisant une seule permutation aléatoire, les observations précédentes ont une plus grande variance dans la statistique cible que les suivantes. À cette fin, CatBoost utilise différentes permutations pour les différentes étapes de l'amplification du gradient.

L'encodage cible ordonné fonctionne de la façon suivante :

- On génère plusieurs permutations aléatoires sur l'ensemble des observations d'apprentissage
- On convertit la valeur de la variable catégorielle par un entier
- Toutes les modalités des variables catégorielles sont transformées en valeur numérique par l'intermédiaire de la fonction ci-dessous :

$$avg\_target = \frac{countInClass + prior}{totalCount + 1} \quad (22)$$

On note :

- *countInClass* est le nombre de fois la modalité de la variable catégorielle est associée à une valeur de la variable cible
- *prior* est la valeur préliminaire du numérateur
- *totalCount* est le nombre de fois où la variable catégorielle est équivalente à la modalité

Un autre particularité du CatBoost est l'utilisation de combinaisons de variables catégorielles comme variables catégorielles supplémentaires qui capturent les dépendances d'ordre élevé. Le nombre de combinaisons possibles croît de manière exponentielle avec le nombre de caractéristiques catégorielles dans l'ensemble de données, et il est impossible de toutes les traiter.

CatBoost construit les combinaisons de manière avide : pour chaque division d'un arbre, il combine (concatène) toutes les caractéristiques catégorielles (et leurs combinaisons) déjà utilisées pour les divisions précédentes de l'arbre actuel avec toutes les caractéristiques catégorielles du jeu de données. Les combinaisons sont ensuite converties avec l'encodage cible.

### 2.3.4 L'algorithme du CatBoost

Pour lutter contre le décalage des prédictions, problème présent dans la plupart des algorithmes d'amplification du gradient, CatBoost utilise une stratégie efficace. Elle s'appuie sur le principe du Ordered Boosting et s'inspire en parallèle des algorithmes d'apprentissage en ligne.

L'algorithme CatBoost fonctionne de la façon décrite dans la figure 2.1. Il a 2 modes pour déterminer la structure des arbres : Plain et Ordered.

Le mode Plain correspond à la combinaison entre l'algorithme classique du Gradient Boosting pour les arbres de décision avec une variable cible ordonnée.

Dans le mode Ordered, une permutation aléatoire, notée  $\sigma_2$ , est effectuée sur le jeu d'entraînement. On applique  $M_1, \dots, M_n$  modèles qui sont entraînés sur les  $i$  premières données dans les échantillons permutés par  $\sigma_2$ . A chaque étape, afin d'obtenir le résidu pour le  $j$ -ème échantillon, CatBoost utilise le modèle  $M_{j-1}$ .

Malheureusement, cet algorithme n'est pas réalisable dans la plupart des tâches pratiques en raison de la nécessité de maintenir  $n$  modèles différents, ce qui augmente la complexité et les besoins en mémoire par  $n$  fois. Catboost implémente une modification de cet algorithme, sur la base de l'algorithme d'amplification du gradient, en utilisant une structure arborescente partagée par tous les modèles à construire.

Afin d'éviter le décalage de prédiction, Catboost utilise des permutations telles que  $\sigma_1 = \sigma_2$ . Cela garantit que la cible  $y_i$  n'est pas utilisée pour la formation  $M_i$  ni pour le calcul de la statistique de la cible ni pour l'estimation du gradient.

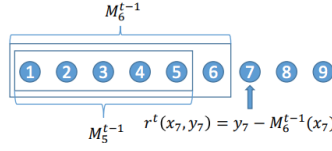


Figure 1: Ordered boosting principle.

---

**Algorithm 1: Ordered boosting**

---

**input** :  $\{(\mathbf{x}_k, y_k)\}_{k=1}^n, I;$   
 $\sigma \leftarrow$  random permutation of  $[1, n];$   
 $M_i \leftarrow 0$  for  $i = 1..n;$   
**for**  $t \leftarrow 1$  **to**  $I$  **do**  
    **for**  $i \leftarrow 1$  **to**  $n$  **do**  
         $r_i \leftarrow y_i - M_{\sigma(i)-1}(i);$   
    **for**  $i \leftarrow 1$  **to**  $n$  **do**  
         $\Delta M \leftarrow$   
             $LearnModel((\mathbf{x}_j, r_j) :$   
                 $\sigma(j) \leq i);$   
         $M_i \leftarrow M_i + \Delta M;$   
**return**  $M_n$

---



---

**Algorithm 2: Building a tree in CatBoost**

---

**input** :  $M, \{y_i\}_{i=1}^n, \alpha, L, \{\sigma_i\}_{i=1}^s, Mode$   
 $grad \leftarrow$   $CatGradient(L, M, y);$   
 $r \leftarrow$   $random(1, s);$   
 $G \leftarrow (grad_r(1), \dots, grad_r(n))$  for *Plain*;  
 $G \leftarrow (grad_{r, \sigma_r(1)-1}(i) \text{ for } i = 1 \text{ to } n)$  for *Ordered*;  
 $T \leftarrow$  empty tree;  
**foreach** *step of top-down procedure* **do**  
    **foreach** *candidate split c* **do**  
         $T_c \leftarrow$  add split  $c$  to  $T;$   
        **if**  $Mode == Plain$  **then**  
             $\Delta(i) \leftarrow$  avg( $grad_r(p)$  for  
                 $p : leaf(p) = leaf(i)$ ) for all  $i;$   
        **if**  $Mode == Ordered$  **then**  
             $\Delta(i) \leftarrow$  avg( $grad_{r, \sigma_r(i)-1}(p)$  for  
                 $p : leaf(p) = leaf(i), \sigma_r(p) < \sigma_r(i)$ )  $\forall i;$   
         $loss(T_c) \leftarrow ||\Delta - G||_2$   
         $T \leftarrow$  argmin $_{T_c}(loss(T_c))$   
    **if**  $Mode == Plain$  **then**  
         $M_{r'}(i) \leftarrow M_{r'}(i) - \alpha$  avg( $grad_{r'}(p)$  for  
             $p : leaf(p) = leaf(i)$ ) for all  $r', i;$   
    **if**  $Mode == Ordered$  **then**  
         $M_{r', j}(i) \leftarrow M_{r', j}(i) - \alpha$  avg( $grad_{r', j}(p)$  for  
             $p : leaf(p) = leaf(i), \sigma_{r'}(p) \leq j$  for all  $r', j, i;$   
**return**  $T, M$

---

FIGURE 2.1 – Algorithme CatBoost

### 2.3.5 Avantages et inconvénients du CatBoost

Le modèle CatBoost présente plusieurs avantages qui justifient son utilisation :

- Un algorithme innovant pour le prétraitement des variables catégorielles. Ainsi, il n'est plus nécessaire d'encoder les variables catégorielles par soi-même avant la phase d'apprentissage – l'algorithme s'en charge d'emblée. Pour les bases de données contenant des variables catégorielles, la précision obtenue sera généralement meilleure que celle des autres algorithmes
- L'implémentation de l'Ordered Boosting, une alternative à l'algorithme du Gradient Boosting classique basée sur des permutations aléatoires. Sur les petits ensembles de données, le Gradient Boosting tend rapidement au sur-apprentissage. Dans le cas du CatBoost, il y a une modification spéciale pour de tels cas qui permet d'éviter ou de réduire le sur-apprentissage.
- La modélisation de l'interaction entre les variables de structures non-linéaires
- La gestion des valeurs manquantes en interne
- Grâce à l'implémentation d'arbres symétriques, il est plus rapide que certains de ses concurrents

Néanmoins, il présente également certaines faiblesses :

- La calibration des hyperparamètres peut s'avérer difficile
- L'interprétabilité des résultats est plus difficile que dans le cas de modèles plus "simples"
- Il ne possède pas d'autres boosters que les arbres de disponibles

## 2.4 Interprétabilité et explicabilité des modèles

### 2.4.1 La réduction moyenne de l'impureté

En apprentissage automatique, l'une des étapes clés est de choisir judicieusement les variables explicatives pour expliquer le phénomène étudié car elles contribuent fortement aux performances des modèles. Afin de sélectionner les variables pertinentes et réduire la dimension des bases d'apprentissage et de validation, nous pouvons utiliser, pour les méthodes utilisant des arbres de décision, l'importance des variables (*feature importance*, en anglais).

Comme nous l'avons vu précédemment, chaque nœud d'un arbre de décision correspond à la division des valeurs d'une seule caractéristique, de sorte que des valeurs similaires de la variable dépendante se retrouvent dans le même ensemble (nœud ou feuille) après la division. Dans le contexte d'une classification, la sélection de la variable de division est généralement basée sur le critère de l'impureté de Gini, tandis que dans le contexte d'une régression, c'est la variance intra-classe.

En restant dans le contexte d'une variable cible catégorielle (classification), le critère de Gini mesure le gain moyen de pureté par fractionnement d'une variable donnée. Si la variable est utile, elle a tendance à scinder les nœuds étiquetés mixtes en nœuds purs à une seule classe. De ce fait, il est possible de calculer dans quelle proportion chaque variable contribue à minimiser l'impureté.

Ainsi, l'importance d'une variable (ou importance de Gini) est calculée comme la diminution de l'impureté totale d'un nœud, pondérée par la probabilité d'atteindre ce nœud, et moyennée sur tous les arbres de l'ensemble. La probabilité du nœud peut être calculée par le nombre d'observations qui atteignent le nœud, divisé par le nombre total d'observations.

A chaque fractionnement dans chaque arbre, l'amélioration du critère de division (ici, l'impureté de Gini) est la mesure d'importance attribuée à la variable de fractionnement, ensuite cumulée sur tous les arbres de la forêt séparément pour chaque variable.

Détaillons comment cette mesure est construite.

Pour chaque arbre de décision, l'importance du nœud est calculée en utilisant l'impureté de Gini. Si l'on suppose que la division d'un nœud conduit à la création de deux "nœuds fils" (arbre binaire), l'importance du nœud  $j$  est donné par :

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (23)$$

On note :

- $w_j$  est le nombre pondéré d'observations atteignant le nœud  $j$
- $C_j$  est l'impureté du nœud  $j$  au sens de Gini
- $left(j)$  et  $right(j)$  les nœuds fils de la division gauche et droite du nœud  $j$

Autrement dit, soit un nœud  $j$  pour lequel on peut calculer l'impureté de Gini ( $C_j$ ). La division de ce nœud sur une variable donnée entraîne la création de deux sous-nœuds : un nœud de gauche et un nœud de droite. Pour ces derniers, l'impureté de Gini est respectivement  $C_{left(j)}$  et  $C_{right(j)}$ . Si l'on soustrait l'impureté de ces deux sous-nœuds de l'impureté du nœud  $j$ , on peut estimer de combien le fractionnement effectué a permis de diminuer l'impureté de Gini.

Cependant, pour l'importance d'une variable, nous sommes intéressés par son importance globale et non par son importance pour un seul nœud. En effet, une variable donnée peut être présente dans différentes branches d'un arbre.

L'importance de chaque variable  $i$  pour un arbre de décision est donc calculée en agréant l'importance de cette variable sur chacune des scissions dont elle est à l'origine. Cette valeur peut ensuite être normalisée à une valeur comprise entre 0 et 1 en la divisant par la somme de toutes les valeurs d'importance de toutes les variables.

$$normfi_i = \frac{\sum_{j: \text{division de } j \text{ sur } i} ni_{ij}}{\sum_{k \in \text{ensemble des noeuds}} ni_k} \quad (24)$$

$$RFfi_i = \frac{\sum_{\text{ensemble des arbres}} normfi_i}{T} \quad (25)$$

Avec  $normfi_i$  l'importance normalisée de la variable  $i$  dans un arbre,  $RFfi_i$  l'importance de la variable  $i$  calculée à partir de tous les arbres du modèle et  $T$  est le nombre total d'arbres.

Ainsi, plus cette valeur est élevée, plus la variable est importante pour expliquer et prédire le phénomène étudié.

Les avantages de cette méthode sont que les calculs sont relativement rapides et qu'elle est facile à mettre en place. Néanmoins, elle a tendance à attribuer une importance élevée aux variables continues et aux variables catégorielles avec de nombreuses modalités.

## 2.4.2 Les valeurs SHAP

Pour mesurer l'importance des variables dans le cadre de modèles complexes tels que le CatBoost, plusieurs mesures sont à notre disposition. Celle qui retiendra notre attention sera la valeur SHAP.

Cette valeur décompose une valeur de prédiction en contributions de chaque variable explicative. Elle mesure ainsi l'impact d'une caractéristique sur une valeur de prédiction unique par rapport à la prédiction de base.

La valeur de SHAP (SHapley Additive exPlanation) proposée par Lundberg et Lee en 2017 [9] est donc une méthode pour expliquer les prédictions individuelles. Son but est d'expliquer la prédiction d'une instance  $x$  en calculant la contribution de chaque caractéristique à la prédiction et permet ainsi de quantifier le rôle de chaque variable dans la décision finale du modèle. L'idée est de moyenner l'impact qu'une variable a pour toutes les combinaisons de variables possibles.

Le principe est issu de la théorie des jeux puisque nous considérerons la prédiction comme le gain d'une coalition dans lequel chaque modalité d'une variable est un joueur. SHAP sera donc basée sur les valeurs de Shapley qui permettent d'indiquer comment répartir équitablement le gain entre les joueurs.

Tout d'abord, introduisons la valeur de Shapley.

Lloyd Shapley a proposé ce concept de solution pour un jeu coopératif en 1953 [17]. L'objectif est de décrire comment la contribution à un gain total généré par une coalition peut être répartie entre les joueurs. Sa définition est assez simple : la contribution individuelle d'un joueur est la valeur attendue, parmi toutes les coalitions possibles ne contenant pas ce joueur, du changement de prédiction causé par l'ajout de ce joueur à la coalition.

On suppose qu'il y a  $N$  joueurs et que  $S$  soit un sous-ensemble de ces  $N$  joueurs. Soit  $c(S)$  la valeur totale de ces  $S$  joueurs. Lorsque le joueur  $i$  rejoint ce sous-ensemble de joueurs, sa contribution marginale est de  $[c(S \cup i) - c(S)]$ . Si l'on prend la moyenne de sa contribution sur les différentes permutations possibles dans lesquelles la coalition peut être formée, on obtient la contribution du joueur  $i$  ( $\varphi_i(c)$ ) :

$$\varphi_i(c) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (n - |S| - 1)!}{n!} [c(S \cup \{i\}) - c(S)] \quad (26)$$

On précise que ce résultat ne nécessite pas d'hypothèses sur la fonction caractéristique  $c$ .

Pour obtenir  $\varphi_i$  il faut donc calculer pour chaque coalition dans laquelle le joueur  $i$  n'apparaît pas, la différence de gain  $[c(S \cup i) - c(S)]$ . Cela permet de comparer le gain obtenu de la coalition avec et sans ce joueur, afin de mesurer son impact lorsqu'il collabore avec cet ensemble des joueurs. Si cette différence est positive, cela signifie que le joueur  $i$  contribue positivement à cette coalition. À l'inverse, si la différence est négative, cela signifie que le joueur  $i$  pénalise le groupe. Enfin, si la différence est nulle, cela indique que le joueur  $i$  n'apporte rien à ce groupe.

On calcule ensuite la moyenne de ces écarts sur toutes les coalitions dans lesquelles le joueur  $i$  apparaît.

Pour mieux comprendre le calcul du nombre de coalitions dans lequel le joueur  $i$  apparaît, on peut réécrire la formule précédente :

$$\varphi_i(c) = \frac{1}{n} \sum_{k=0}^{n-1} \binom{n-1}{k}^{-1} \sum_{S \subseteq N \setminus \{i\}, |S|=k} [c(S \cup \{i\}) - c(S)] \quad (27)$$

Cela montre bien que l'on a fait la moyenne des différences.

Le dénombrement est le suivant : en ayant préalablement classé les coalitions par cardinal, on fait la moyenne des écarts pour toutes les coalitions possibles. Pour une coalition de taille  $k$ , le nombre de combinaisons de joueurs (coalitions) possibles est de  $C_{n-1}^k$ .

On moyenne ensuite ces résultats intermédiaires : il y a  $n$  tailles de coalitions différentes, l'ensemble vide étant la coalition de cardinal 0.

Pour parvenir à une contribution équitable, Shapley a établi les quatre axiomes suivants :

- **Efficacité** : la somme des valeurs de Shapley de tous les joueurs est égale à la valeur de la coalition totale
- **Symétrie** : tous les joueurs ont une chance équitable de participer au jeu.

- **Contribution nulle sans effet** : si le joueur  $i$  ne contribue pas à une quelconque coalition  $S$ , alors sa contribution est nulle :  $\varphi_i(c) = 0$ .
- **Additivité** : pour toute paire de jeux  $c_1$  et  $c_2$  :  $\varphi(c_1 + c_2) = \varphi(c_1) + \varphi(c_2)$ , où  $(c_1 + c_2)(S) = c_1(S) + c_2(S)$  pour tout  $S$ . Cette propriété nous permet de faire la simple somme arithmétique.

Le principal intérêt de la valeur de Shapley est la simplicité de son expression ( $n$  coefficients, sommables) tout en tenant compte de la contribution d'un joueur aux coalitions auxquelles il appartient. Elle permet ainsi de tenir compte de l'interaction entre les joueurs.

Pour appliquer cette méthode à l'apprentissage automatique, les joueurs deviennent les modalités  $x_i$  prises par  $x$  sur chaque variable explicative. On souhaite désormais expliquer la prédiction  $f(x)$  associée à une observation  $x$ . Ainsi, le gain à répartir devient la différence entre la prédiction  $f(x)$  et la moyenne des prédictions  $E[f(x)]$ .

Pour une coalition  $u$  de valeur  $x_u$  de  $x$ , la fonction caractéristique du jeu est :

$$c(u) = E(f(X)|X_u = x_u) \quad (28)$$

Ainsi, la valeur de Shapley associée à une modalité  $x_i$  est définie par :

$$\varphi_i(f, x) = \sum_{u \subseteq \{1, \dots, n\} \setminus i} \frac{|u|! (n - |u| - 1)!}{n!} [E(f(X)|X_{u \cup \{i\}} = x_{u \cup \{i\}}) - E(f(X))] \quad (29)$$

De cette façon, le coefficient  $\varphi_i$  explique comment les valeurs  $x_i$  contribuent à décaler la prédiction  $f(x)$  de la moyenne  $E[f(x)]$  des prédictions.

La somme des valeurs de Shapley d'une observation  $x$  est égale à l'écart entre la prévision  $f(x)$  et la moyenne des prévisions  $E[f(x)]$  :

$$f(x) - E[f(x)] = \sum_{i=1}^p \varphi_i \quad (30)$$

L'intérêt de l'approche proposée par Lundberg et Lee [9], basée sur le théorème de Shapley, est de tenir compte des effets d'interactions. En effet, il est possible que deux modalités  $x_i$  et  $x_j$  prises isolément n'aient aucun pouvoir prédictif, alors qu'elles peuvent être très informatives une fois couplées ensemble.

Par ailleurs, les valeurs SHAP reprennent toutes les propriétés des valeurs de Shapley en théorie des jeux coopératifs.

En moyennant les valeurs absolues des valeurs SHAP obtenues pour chaque modalité et chaque observation, nous pouvons remonter à l'importance globale des variables.

Dans la pratique, l'estimation des valeurs SHAP n'est pas aisée. Tout d'abord, l'estimation des espérances conditionnelles peut d'avérer difficile. Ensuite, la mise en place de cette approche est très

coûteuse car le processus doit être répété pour toutes les coalitions possibles. Le temps de calcul augmente de façon exponentielle avec le nombre de caractéristiques, c'est pourquoi il est courant que seules les contributions de quelques échantillons des coalitions possibles soient calculées.

La bibliothèque Python SHAP propose un algorithme d'estimation optimisé pour les modèles basés sur les arbres de décision (XGBoost, CatBoost, LigthGBM) : Tree Explainer. Il reprend les travaux proposés par Lundberg, Erion et Lee en 2019 [11].

Cet algorithme récursif permet d'accéder, avec un temps de calcul raisonnable, à une approximation des valeurs de Shapley dans le cas de modèles d'arbre de décision ou de modèles ensemblistes. Par cette approche, nous pouvons calculer l'importance globale des variables mais surtout les effets des variables pour chaque observation du jeu de données.

En définitive, les valeurs de Shapley appliquées à l'apprentissage automatique sont une technique d'explicabilité locale qui permet de comprendre la prédiction associée à une observation.

Pour une observation donnée, elles donnent la possibilité de quantifier l'impact de chaque variable sur la prédiction qui lui est attribuée par le modèle.

L'avantage est qu'elles prennent en compte les interactions entre les variables explicatives pour déterminer leur impact dans la prédiction du modèle. Néanmoins, une limite persiste en termes d'intelligibilité : l'absence, en général, de valeurs cibles pour valider les résultats.

## 2.5 Les $K$ -moyennes

L'algorithme des " $K$ -moyennes" (aussi connu sous le nom de " $K$ -means"), proposé dans les années 1960, est un algorithme non supervisé de clustering non hiérarchique. Il permet d'identifier  $K$  groupes d'observations d'une base de données ayant des caractéristiques similaires.

Le principe est de considérer les observations dans un environnement multidimensionnel représentant les différentes caractéristiques. Pour pouvoir les regrouper en  $K$  clusters distincts, l'algorithme a besoin d'un moyen de comparer le degré de similarité entre les différentes observations. De cette façon, deux données qui se ressemblent auront une distance de dissimilarité réduite tandis que deux données plutôt différentes auront une distance plus grande.

Les littératures mathématiques et statistiques proposent plusieurs mesures de distance pour évaluer cette "proximité", mais la plus connue reste la distance euclidienne.

Soit  $x_i^k$ , le  $k$ -ème élément d'observation de la variable  $x_i$ . La distance euclidienne est donnée par :

$$d(x_i, x_{i'}) = \sqrt{(x_i^1 - x_{i'}^1)^2 + \dots + (x_i^p - x_{i'}^p)^2} \quad (31)$$

Il est important de mentionner que les échelles de mesure des variables explicatives peuvent affecter les métriques de distance. Il est ainsi nécessaire de standardiser les variables explicatives avant de mettre en œuvre l'algorithme.



Pour la construction de l'algorithme, il est nécessaire de déterminer *a priori* un nombre de clusters  $K$ . Ce choix n'est pas forcément intuitif, notamment lorsque le jeu de données est important. Un nombre élevé pourrait conduire à un partitionnement trop fragmenté des données, ce qui pourrait empêcher de relever des tendances intéressantes dans les observations. À l'inverse, un nombre trop petit pourrait conduire à avoir des clusters trop généralistes contenant beaucoup de données.

La méthode la plus usuelle pour choisir le nombre de clusters est de lancer l'algorithme avec différentes valeurs de  $K$  et calculer la variance des différents clusters.

Si l'on note  $c_j$  le centre du cluster  $j$ ,  $x_i$  la  $i$ -ème observation dans le cluster ayant pour centroïde  $c_j$  et  $D(c_j, x_i)$  la distance entre le centre du cluster et le point  $x_i$ , la variance des clusters se calcule comme suit :

$$V = \sum_j^K \sum_{x_i \rightarrow c_j} d^2(c_j, x_i) \quad (32)$$

Autrement dit, la variance est la somme des distances entre chaque centroïde d'un cluster et les différentes observations incluses dans le même cluster.

De cette façon, on cherche à trouver le nombre de clusters qui minimise la distance entre les observations des différents clusters et leur centroïde associé. On parle de minimisation de l'inertie intra-classe.

Par ailleurs, on peut remarquer que minimiser l'inertie intra-classe revient à maximiser l'inertie inter-classe c'est-à-dire la distance entre les individus appartenant à des groupes différents.

Selon le théorème de Huygens, l'inertie totale est égale à la somme de l'inertie intra-classe (indicateur de compacité des classes) et de l'inertie inter-classe (indicateur de séparabilité des classes).

Résumons de manière simplifiée comment fonctionne l'algorithme des  $K$ -moyennes.

Durant la phase d'initialisation, les barycentres des clusters sont générés aléatoirement et les individus les plus proches sont affectés à chacun d'entre eux, en fonction d'une mesure de proximité telle que la distance euclidienne.

Les barycentres sont ensuite recalculés et l'affectation des individus change tant que la variance intra-classe décroît significativement.

La classe  $k$  la plus proche de l'observation ayant une information manquante est également déterminée. Si cette dernière est qualitative alors la valeur imputée correspond à la valeur majoritaire dans la classe  $k$ , alors que si elle est quantitative, la valeur imputée est la moyenne des valeurs de la classe.

Le processus s'arrête une fois que l'algorithme a convergé ou que l'inertie totale de la population s'est stabilisée.

## 3 Chapitre 3 : Cadre de l'étude

### 3.1 Périmètre de l'étude

L'objectif de ce mémoire est d'appréhender les moments clés dans la vie d'un contrat d'assurance automobile. Dans un premier temps, il s'agit de détecter le moment où l'assuré est plus sensible à changer de véhicule. Ensuite, il s'agit de déterminer si ce dernier est susceptible de résilier sa police suite à cet évènement.

Le changement du bien assuré étant un motif fréquemment avancé pour justifier la résiliation, il est particulièrement important de capter l'instant précis auquel le client est le plus fragile dans le but de mener des actions de rétention, et ce d'autant plus que la résiliation d'un contrat d'assurance pourrait entraîner des résiliations en cascade des autres contrats en portefeuille du client.

En vue de modéliser de façon robuste le changement de véhicule et la résiliation au changement de véhicule, il est indispensable de réunir des données fiables. La première étape consiste donc à définir le périmètre de la base d'étude en tenant compte des données à notre disposition dans les différents systèmes d'information et de leur robustesse au fil des années d'étude.

En effet, il faut s'assurer de la continuité des données c'est-à-dire garantir qu'à la fois les procédés tarifaires et les systèmes d'information (SI) n'aient pas été fondamentalement modifiés au cours de la période d'étude et que l'ensemble des variables soient disponibles. A titre d'exemple, il ne serait pas pertinent de prendre en compte des données antérieures et postérieures à 2015 pour analyser le comportement de résiliation au vu de la modification réglementaire entraînée par la loi Hamon.

En tenant compte de ces considérations, notre base d'étude sera constituée de sorte à analyser les occurrences de changements de véhicule et/ou de résiliations pour ce motif dans le portefeuille de Generali entre 2018 et 2020.

Par ailleurs, afin d'avoir une photographie du contrat avant et après le phénomène étudié et observer le parcours client sur plus d'une année, le périmètre des données collectées pourra être étendu afin de disposer d'une profondeur d'historique suffisante. Ainsi, les données récupérées s'étalent de 2016 à 2020, nous permettant d'observer les deux phénomènes pour les 3 années d'étude et de disposer d'une profondeur d'historique d'au moins un an.

Une fois, le périmètre temporel de la base d'étude clairement défini, il est nécessaire de cadrer les contrats qui feront partie de notre analyse ou non. Notre étude se concentrera donc uniquement sur les véhicules 4 roues inférieurs à 3,5 tonnes. Les produits couvrant les 2 roues, les tracteurs ou les flottes de véhicules (contrats multi-véhicules) ne seront pas intégrés dans notre analyse.

De plus, pour des raisons de disponibilité et de fiabilité des données, nous ne retiendrons que les contrats faisant partie du réseau des agents de Generali, excluant ainsi les contrats souscrits via les canaux de distribution des courtiers, du réseau salarié et des marques blanches.

Le périmètre d'étude étant clairement défini, il s'agit désormais de constituer la base d'étude à partir des différentes données disponibles, de générer de l'information à partir de certaines d'entre elles, ou encore d'en ajouter à partir de sources externes ; tout en faisant attention à ce que l'information soit la plus pertinente et complète possible pour répondre à notre problématique.

## 3.2 Constitution de la base de données

### 3.2.1 Méthodologie de la création de la base

Les données internes sont regroupées dans les systèmes d'information de Generali en plusieurs tables, catégorisées selon la nature de l'information qu'elles apportent. Il peut survenir que les informations soient présentes dans plusieurs tables voire différentes variables, et il nous appartient donc de sélectionner celles nous semblant les plus fiables et les mieux construites pour notre analyse.

Parmi l'ensemble des tables à notre disposition, nous avons conservé :

- **La table relative au conducteur** regroupe les informations détaillées sur le conducteur qui sont collectées au moment de la souscription du contrat : âge du conducteur principal, ancienneté du permis de conduire, sexe, nature (personne physique ou morale), catégorie socio-professionnelle, situation maritale, nombre d'enfants, code commune du domicile, ...
- **La table relative au véhicule** regroupe les informations sur le bien assuré : âge du véhicule, date d'acquisition, carrosserie, marque, modèle, usage, mode de parking, code SRA, ...
- **La table relative au contrat** regroupe les informations sur la police d'assurance : état du contrat (actif / résilié), date de souscription, date d'effet, date et motif de résiliation, date du dernier avenant, date de contentieux ou de mise en demeure, fractionnement des paiements (annuel, semestriel, trimestriel, mensuel), montant annuel des primes hors taxes, montant des commissions acquises, montant du rabais accordé, écart au tarif, formule souscrite, offre kilométrique, présence d'un conducteur occasionnel, coefficient bonus-malus, ...

Cette table avec l'ensemble des informations sur la police d'assurance est regroupée par numéro de contrat. Chaque contrat bénéficie ainsi d'un numéro unique. De ce fait, un client détenant plusieurs contrats chez Generali disposera d'une ligne pour chacun de ses contrats dans cette base.

- **La table relative aux clients** permet d'avoir une vue d'ensemble sur le portefeuille de contrats que détient l'assuré chez Generali : ancienneté du client chez Generali, nombre total de contrats détenus par le client, nombre de contrats détenus par branche d'assurance, nombre d'affaires nouvelles et de résiliations sur une période pré-définie, date du dernier mouvement effectué sur le portefeuille (souscription / résiliation), prime totale du portefeuille, ...

Cette table avec les informations sur le portefeuille de contrats détenu par un client est regroupée par numéro de client. Chaque client bénéficie ainsi d'un numéro unique. Ainsi, contrairement à la table relative aux contrats, un client détenant plusieurs contrats ne sera représenté que par une unique ligne dans cette base de données.

- **La table relative aux garanties** recense les différents types de garanties que l'assuré a souscrit dans sa police d'assurance et retrace la décomposition de la prime en fonction de chacune d'entre elles.
- **La table relative aux sinistres** retrace l'historique des sinistres survenus sur la police d'assurance : date de survenance du sinistre, caractère responsable ou non du sinistre, date d'ouverture et de clôture du dossier, charge nette et brute du sinistre, ...

Ces différentes tables sont interconnectées entre elles par les deux clés primaires que nous avons évoqué : le numéro de contrat et le numéro client. Ces numéros étant uniques, ils nous permettront de lier les différentes informations entre elles.

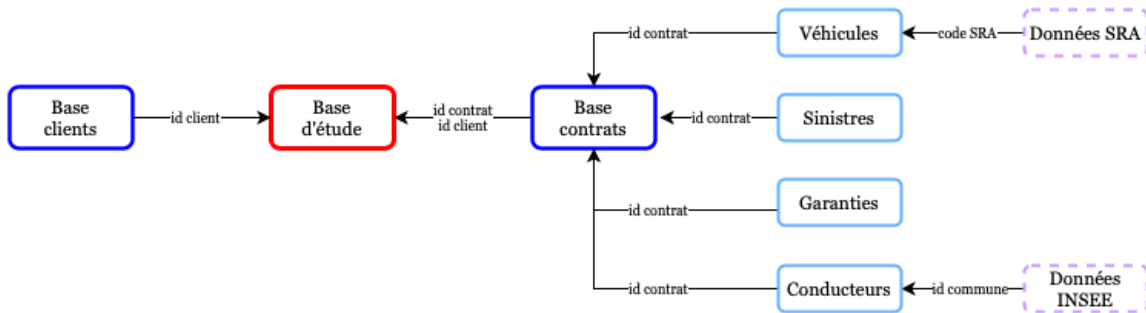


FIGURE 3.1 – Schéma sur la méthode de regroupement des différentes bases

Notre période d’observation s’étalant sur 3 années et le périmètre de collecte des données sur 5 ans, il serait trop fastidieux d’essayer d’obtenir une photographie mensuelle des contrats et ainsi d’observer le changement de véhicule ou la résiliation entre deux mois consécutifs.

En contrepartie d’une légère perte d’information, notamment sur la temporalité de la survenance du phénomène, nous avons fait le choix de collecter les données à la date de la dernière mensuelle des années considérées. Ainsi, nous obtenons une image annuelle des contrats et nous comparerons les informations recueillies à la fin de chaque année afin d’identifier un potentiel changement.

Une fois l’ensemble des tables extraites des systèmes pour chaque année entre 2016 et 2020, il s’agit de les rassembler et de les lier entre elles afin de créer notre base d’étude.

La première étape consiste à définir le périmètre des contrats de cette base, le périmètre d’analyse s’étendant uniquement de 2018 à 2020.

La première année d’analyse sur l’occurrence ou non d’un changement de véhicule ou d’une résiliation étant 2018, le point de départ de notre base est donc l’ensemble des contrats répondant au périmètre défini dans la partie précédente (contrats 4 roues, réseau agent, mono-véhicule) et actifs l’année précédente, c’est-à-dire 2017.

A partir de cette liste de contrats, nous rattachons les contrats de l’année suivante, 2018, à partir de la clé primaire qu’est le numéro d’identification du contrat. Le type de jointure utilisée est ce qu’on appelle une liaison "*outer join*" : elle permet de relier les données présentes dans les deux bases à partir de la clé primaire, tout en conservant les informations présentes uniquement dans l’une ou l’autre des deux bases.

Autrement dit, pour les contrats actifs à la fois en 2017 et 2018, nous disposerons d’une photographie des contrats pour ces deux années. En revanche, pour les contrats actifs en 2017 mais résiliés en 2018, nous ne pourrions récupérer que le motif de résiliation pour 2018, les autres données n’étant plus renseignées dans les systèmes. A l’inverse, pour les contrats qui n’étaient pas présents en 2017 mais qui le sont en 2018, c’est-à-dire des affaires nouvelles, nous récupérerons une image des contrats uniquement pour cette dernière année.

La même opération est répétée avec les contrats actifs en 2018 puis ceux de 2019, ce qui nous permet de réunir les informations pour les contrats pour lesquels nous souhaitons observer la survenance ou non d'un changement de véhicule ou d'une résiliation pour ce motif.

L'ajout des différentes informations pour chaque contrat en fonction des années se fait par colonne, en prenant soin de labelliser les variables par l'année correspondante, de sorte à pouvoir observer l'évolution des contrats au cours du temps.

Année	Numéro contrat	Immatriculation du véhicule	Age du véhicule
2017	AR230987	AB-123-RE	9
2018	AR230987	DE-345-TR	0

Numéro contrat 2017	Immatriculation du véhicule 2017	Age du véhicule 2017	Numéro contrat 2018	Immatriculation du véhicule 2018	Age du véhicule 2018
AR230987	AB-123-RE	9	AR230987	DE-345-TR	0

TABLE 3.1 – Méthodologie de la constitution de la base de données par année

La seconde étape consiste à enrichir notre base à partir des différentes informations contenues dans les autres tables (conducteur, véhicule, client, sinistres et garanties) entre 2016 et 2020 pour le périmètre des contrats précédemment obtenu, au moyen de l'une des deux clés primaires.

La manière dont est constituée la base d'étude nous permet ainsi d'avoir une image du contrat, du conducteur, du véhicule assuré, du profil client ou des sinistres avant le changement de véhicule, que ce dernier entraîne une résiliation ou non, et d'avoir cette même vision après avoir observé ce phénomène, à condition que le client soit resté en portefeuille.

### 3.2.2 Enrichissement de la base de données

#### 3.2.2.1 Création des variables cibles

Dans le cas du changement de véhicule, on peut distinguer trois cas distincts au sein du portefeuille :

1. Les assurés ne changeant pas de véhicule sur la période d'observation
2. Les assurés changeant de véhicule au cours de la période d'observation et ayant modifié leur police d'assurance par un avenant
3. Les assurés changeant de véhicule au cours de la période d'observation mais ayant résilié leur contrat d'assurance automobile chez Generali

Dans le premier cas, nous disposons d'une image du contrat pour l'année N et l'année N+1. Seules des informations relatives au client, au contrat, aux sinistres ou au parcours client sont susceptibles d'avoir changé.

Dans le second cas, nous disposons également d'une photographie du contrat pour l'année N et l'année N+1. En revanche, en plus des informations client ou parcours client qui pourraient avoir changé, les données sur le véhicule et le contrat ont nécessairement été modifiées par un avenant.

Enfin, dans le dernier cas, nous ne disposons que des informations pour l'année N, les informations pour l'année N+1 n'étant plus disponibles à l'exception du motif de résiliation du contrat.

Pour notre étude, nous avons besoin d'identifier les contrats ayant rencontré la seconde ou la troisième situation ; la première pouvant se déduire des deux précédentes puisqu'elle correspond à une absence d'évènement.

La première variable cible de notre étude, correspondant à la seconde situation, permet de capter les contrats ayant changé de véhicule assuré entre deux années d'étude et qui sont restés actifs dans le portefeuille, avec un avenant de modification du contrat. Cette variable binaire vaut 1 si un tel changement a été observé et 0 sinon.

Pour identifier les changements de véhicule, la méthode choisie est la comparaison des numéros d'immatriculation des véhicules assurés entre deux années d'observations, à savoir 2017 - 2018, 2018 - 2019 et 2019 - 2020.

Une méthode alternative aurait pu être la comparaison des numéros d'identification SRA. Cependant, elle nous aurait exposé à omettre les changements de véhicule s'effectuant pour le même modèle de véhicule (disposant donc du même numéro SRA). En outre, la connaissance des bases de données nous permet de savoir que cette donnée n'est pas systématiquement renseignée dans les systèmes, à la différence du numéro d'immatriculation, et que sa fiabilité peut parfois être remise en question.

En conservant la première méthode évoquée, il faut prendre en compte plusieurs paramètres afin de s'assurer que l'identification des changements de véhicule soit la plus fiable possible.

Le cas le plus simple consiste à comparer les numéros d'immatriculation étant déjà au format imposé par la réglementation en vigueur, tout en s'assurant que l'évolution de la durée de détention du véhicule suive son cours normal.

<b>Année</b>	<b>Numéro client</b>	<b>Numéro contrat</b>	<b>Immatriculation du véhicule</b>	<b>Durée détention du véhicule</b>	<b>Changement de véhicule</b>
N	543	AR230987	AB-123-RE	8	0
N+1	543	AR230987	AB-123-RE	9	1
N+2	543	AR230987	DE-345-TR	0	0
N+3	543	AR230987	DE-345-TR	1	

TABLE 3.2 – Méthode d'identification des changements de véhicule

Dans cet exemple, on identifie le contrat comme ayant changé de véhicule entre l'année N+1 et l'année N+2. L'objectif de notre étude étant d'identifier les différentes caractéristiques amenant le changement de véhicule, nous attribuons la variable du changement de véhicule à l'année précédant sa survenance (ici, N+1) afin d'avoir une image du contrat avant l'observation du phénomène.

Se limiter à la comparaison des immatriculations conduirait à introduire un biais assez significatif dans l'identification des changements de véhicule. En effet, plusieurs éléments peuvent venir entraver cette étape : la mise en conformité des immatriculations avec le nouveau système d'immatriculation, les immatriculations provisoires et les erreurs de saisie commises au moment de leur renseignement dans les systèmes.

Le cas le plus complexe consiste à prendre en compte les changements d'immatriculation qui sont liés à une application du nouveau système d'immatriculation des véhicules (SIV).

En effet, en 2009, un nouveau système d'immatriculation des véhicules est entré en vigueur. Jusqu'en 2020, les véhicules pouvaient encore présenter des plaques au format de l'ancien système d'immatriculation datant de 1950, à condition qu'aucune opération ayant nécessité l'édition d'un nouveau certificat n'ait eu lieu (changement de propriétaire, changement de domicile, modification de l'état civil, édition d'un duplicata à la suite d'une perte ou d'un vol, etc...). Cependant, à partir de 2020, ce nouveau système d'immatriculation reposant sur le format "AA-000-AA" régit tous les véhicules immatriculés, prenant ainsi la place de l'ancien système géographique qui reposait sur le format "00-AAA-00".

De ce fait, il est nécessaire de ne pas confondre les changements d'immatriculation qui sont liés à un changement de véhicule et ceux qui sont liés à une actualisation du numéro d'immatriculation. Afin de prendre en compte cette distinction, l'étude de l'immatriculation couplée au suivi de l'évolution de la durée de détention du véhicule semble nécessaire.

Année	Numéro client	Numéro contrat	Immatriculation du véhicule	Durée détention du véhicule	Changement de véhicule
N	635	AR230988	12-AHD-23	8	0
N+1	635	AR230988	12-AHD-23	9	0
N+2	635	AR230988	AA-229-AB	10	

TABLE 3.3 – Méthode d'identification des mises en conformité des numéros d'immatriculation

Dans cet exemple, on peut constater que l'immatriculation du bien assuré a été modifiée entre l'année N+1 et N+2. Cependant, en regardant de plus près la durée de détention du véhicule, on constate que cette dernière suit son cours normal, sans qu'un changement de véhicule ne semble avoir été opéré entre ces deux dates. Or, on remarque également que l'immatriculation du véhicule a changé de format, passant de l'ancien au nouveau format. On parvient donc à la conclusion que le véhicule assuré est resté le même sur toute la période et qu'aucun changement n'est survenu.

Dans le cas inverse où l'immatriculation du véhicule passerait de l'ancien au nouveau format mais que la durée de détention du véhicule serait modifiée, on pourrait conclure à un changement effectif du véhicule assuré.

Pour identifier ces cas, il s'agit de récupérer les contrats pour lesquels la durée de détention du véhicule suit son cours normal d'une année à l'autre et pour lesquels l'immatriculation du véhicule assuré change, commençant par deux chiffres une année et par deux lettres l'année suivante.

Un autre cas à considérer concerne les immatriculations provisoires. Ces dernières sont délivrées par la préfecture lorsqu'il n'est pas possible d'immatriculer un véhicule de façon définitive, en raison de documents ou pièces justificatives manquantes. Elles ont la particularité de commencer par "WW".

Ainsi, il est aisé d'identifier les changements d'immatriculation liés au passage d'une immatriculation provisoire à une immatriculation définitive. En effet, les véhicules pour lesquels l'immatriculation commence par "WW" une année puis par deux lettres l'année suivante, conformément au nouveau SIV, et pour lesquels la durée de détention suit son cours normal, ne seront pas identifiés comme des changements de bien assuré.

Enfin, le dernier cas à prendre en compte pour ne pas biaiser notre variable cible en comptabilisant des "faux" changements de véhicule est la robustesse des données. La qualité des données reste un problème majeur dans de nombreux secteurs d'activités et les données saisies pour les numéros d'immatriculation n'échappent pas à la règle. Ainsi, il se peut que ce numéro change d'une année à l'autre en raison d'une correction de la donnée dans les systèmes. Le cas le plus fréquemment rencontré est l'inversion de deux caractères alphanumériques.

Pour exclure les changements liés à cette mauvaise qualité de la donnée, une analyse de la correspondance des immatriculations a été utilisée. Les numéros d'immatriculation présentant 7 caractères alphanumériques, les contrats pour lesquels la correspondance entre deux années est de 5 ou 6 caractères, sachant que la détention du véhicule augmente normalement, peuvent être considérés comme une mise à jour des données et non comme un changement de véhicule. La volumétrie de ces cas n'étant pas importante, il peut être utile d'effectuer une vérification manuelle supplémentaire.

La seconde variable cible de notre étude consiste à identifier les contrats qui ont été résiliés pour cause de changement de véhicule. L'identification de ces contrats est bien plus aisée que les changements de véhicule restant actifs dans le portefeuille puisqu'il suffit de récupérer le motif de résiliation du contrat.

Ainsi, les contrats pour lesquels nous considérerons qu'ils ont été résiliés pour cause de changement de véhicule sont les contrats ayant pour motif de résiliation "Disparition du risque" ou "Vente de l'objet assuré".

<b>Année</b>	<b>Numéro client</b>	<b>Numéro contrat</b>	<b>Immatriculation du véhicule</b>	<b>Motif de résiliation</b>	<b>Résiliation pour chgmt</b>
N	543	AR230989	BH-345-FV		0
N+1	543	AR230989	BH-345-FV		1
N+2	543	AR230989		Vente de l'objet assuré	

TABLE 3.4 – Méthode d'identification des résiliations pour changement de véhicule

Une fois les deux variables cibles créées, nous créons une dernière variable synthétisant l'information recueillie par chacune d'entre elles. Elle est simplement la somme des deux variables précédemment créées et permet de capter si une forme de changement de véhicule est survenue d'une année à l'autre, que le contrat reste en portefeuille ou non.



Pour notre étude, nous utiliserons dans un premier temps cette variable synthétique afin de prédire si l'individu va changer de véhicule dans l'année. Par la suite, nous modéliserons uniquement la résiliation pour changement de véhicule afin de cibler, parmi les clients susceptibles de changer de véhicule, ceux qui sont fragiles à la résiliation.

Maintenant que nous avons créé les variables qui constitueront les variables exogènes des différentes modélisations, il est intéressant d'extraire de l'information des variables endogènes que nous avons récupéré dans les différentes sources de données.

### **3.2.2.2 Création de nouvelles variables explicatives**

Les variables descriptives que nous avons à disposition dans notre base d'étude sont nombreuses et peuvent être regroupées en deux groupes distincts. D'une part, nous avons les informations techniques et de sinistralité qui regroupent les critères pris en compte au moment de la tarification ainsi que l'historique des sinistres. D'autre part, nous avons les informations commerciales et marketing qui permettent de connaître le profil du client et son parcours depuis sa première souscription.

Extraire de l'information des données techniques et de sinistralité peut nous permettre d'identifier les clients pour lesquels le risque est mal évalué ou pour lesquels la sinistralité a été peu importante au cours des dernières années, etc...

Extraire de l'information des données commerciales et marketing peut, en revanche, nous permettre de baliser le parcours du client et mesurer son attachement avec son assureur.

Avec la profondeur d'historique dont nous disposons, il nous sera possible d'observer l'évolution de ces différents types de données au cours du temps. L'extraction d'informations à partir des données brutes sera particulièrement intéressante pour l'étude du comportement de résiliation qui peut être assez sensible aux mouvements effectués sur les contrats. Par exemple, si connaître le nombre de contrats détenus par un assuré est une donnée importante puisqu'elle permet de mesurer à quel point il est engagé avec Generali, savoir si le client est dans une démarche de résiliation ou de souscription est d'autant plus intéressant car observer des signaux de désengagement est un indicateur de fragilité à la résiliation pour les contrats restants en portefeuille.

Concernant les données techniques, nous allons essayer de capter des changements de situation dans la vie de l'assuré. De tels changements pourraient entraîner une revalorisation de la prime si ces derniers ont trait avec les variables entrant en jeu dans la tarification ; mais peuvent également être des facteurs poussant l'individu à changer de véhicule ou résilier son contrat.

Ainsi, pour une année d'étude considérée, nous allons déterminer si un changement dans la vie de l'assuré a eu lieu par rapport à l'année précédente : changement de catégorie socio-professionnelle, arrivée d'un enfant à charge, changement de lieu d'habitation. Les variables permettant d'identifier ces changements sont créées sous le format binaire et sont construites en comparant la valeur de la variable pour l'année N et celle pour l'année N-1.

Le changement de catégorie socio-professionnelle peut avoir un impact sur l'utilisation qu'un individu peut faire de son véhicule. En effet, on sait qu'une part assez importante des trajets est réalisée à des fins professionnelles. Ainsi, changer d'activités, passer à la retraite ou perdre son emploi peuvent considérablement modifier les habitudes d'utilisation du véhicule.

L'arrivée d'un enfant, quant à lui, peut avoir un impact sur le type de véhicule nécessaire pour se déplacer et ainsi créer le besoin d'acquérir un nouveau bien plus adapté.

Enfin, le département ou la région d'habitation ont également un fort impact sur l'utilisation du véhicule. En effet, un individu passant d'une région plutôt urbaine à une région plus rurale, ou inversement, ne fera pas la même utilisation de son véhicule. En outre, la localisation du lieu de garage a un impact sur la tarification et pourrait donc conduire à une revalorisation de la prime qui pourrait pousser l'assuré à se tourner vers la concurrence. Enfin, l'implantation de Generali en France varie beaucoup selon les régions et un individu se déplaçant vers une région moins distribuée pourrait préférer se tourner vers un assureur avec lequel il aurait plus de proximité.

Concernant les données de sinistralité, nous pouvons retracer un historique sur les 36, 24 et 12 derniers mois, en utilisant également la distinction entre les sinistres responsables et non-responsables. La sinistralité passée peut avoir un fort impact sur le changement de véhicule en lui-même, un véhicule ayant subi de multiples sinistres étant plus susceptible d'être remplacé car plus ou moins endommagé ; mais aussi sur la résiliation puisque la manière dont ont été gérés les sinistres influence la satisfaction du client et impacte donc sa volonté de rester ou non chez son assureur.

De la même façon, savoir si un contentieux a eu lieu entre l'assuré et l'assureur au cours des dernières années peut être un signal fort pour la résiliation.

Enfin, directement en lien avec la sinistralité et les données techniques, la variation du coefficient de Réduction-Majoration (CRM) par rapport à l'année précédente peut apporter de l'information supplémentaire sur les sinistres récemment survenus et sur la variation de la prime en découlant.

Concernant les données marketing et commerciales, il est particulièrement important de cibler des variables permettant d'avoir une idée sur la fidélisation du client, son parcours ou encore les évolutions tarifaires du contrat.

Ainsi, concernant le parcours du client chez Generali, nous créons plusieurs variables à la fois à partir des données brutes mais aussi grâce à l'historique dont nous disposons.

Comme nous l'avons évoqué, savoir si le client a été dans une démarche récente de souscription ou de résiliation est un signal clé quant à son attachement avec la compagnie. A partir du nombre de contrats détenus par le client l'année précédente, nous créons des variables permettant de retracer l'évolution du nombre de contrats détenus au total et par branche d'activité (IARD / vie, automobile, multirisques habitation, etc...).

Par ailleurs, à partir des données permettant de retracer les dates des derniers mouvements effectués sur le portefeuille (dernière souscription, dernière souscription en IARD / vie, dernière résiliation, etc...), nous créons des variables permettant de calculer le laps de temps s'étant écoulé depuis. Cela vient compléter l'information apportée par l'évolution du nombre de contrats.

En outre, une autre variable que nous pouvons construire à partir du nombre de contrats détenus par famille de produits est le multi-équipement, ou plus particulièrement le multi-équipement en auto-MRH qui sont deux produits faisant l'objet d'une offre de la part de Generali. En effet, un client disposant de plusieurs familles de produits au sein de la même compagnie est un client qui

est plus lié à son assureur et qui pourrait donc être moins susceptible de résilier un contrat dans l'une des familles de produits détenus. L'analyse de l'évolution du multi-équipement par rapport à l'année précédente est également un point d'intérêt.

Par expérience, on sait que le prix est un facteur déterminant dans la résiliation. Le pouvoir d'achat des Français étant sous tension depuis de nombreuses années et le secteur automobile représentant un poste de dépense assez conséquent, que ce soit pour l'achat du véhicule, les dépenses liées au carburant, l'entretien ou l'assurance, une augmentation significative de la prime d'assurance pourrait conduire à la résiliation du contrat. Cela est d'autant plus vrai au moment du changement de véhicule où la prime est généralement fortement revalorisée. Ainsi, étudier l'évolution de la prime par rapport à l'année précédente, en montant comme en variation relative peut être particulièrement intéressant.

Enfin, pour un sujet aussi comportemental que le changement de véhicule, il est nécessaire d'essayer de capter les habitudes de consommation des assurés. En effet, tandis que certains individus ne changent de véhicule que lorsque la "nécessité" se présente, d'autres ont tendance à changer plus régulièrement de véhicule, ce qui s'observe notamment lors de l'acquisition de véhicules onéreux pour lesquels la revente rapide permet de pallier la dépréciation de la valeur qui survient très rapidement après l'achat à neuf. C'est pourquoi, nous avons décidé de créer une variable croisée permettant de retracer l'historique des changements de véhicule du client sur 10 ans. On croise ainsi le nombre de changements de véhicule sur les 10 dernières années avec le nombre d'années durant lesquelles le contrat était présent sur cette période. Cela permet de capter la fréquence de changement de véhicule tout en ne pénalisant pas l'information pour les clients dont la présence en portefeuille est plus récente.

### **3.2.2.3 Apport de données externes**

Outre les informations qui ont été apportées sur la base d'étude en utilisant les données brutes présentes dans les systèmes d'information et la profondeur d'historique dont nous disposons, il peut être judicieux de compléter l'information à partir de données provenant de sources externes.

Nous allons ainsi apporter deux types d'informations : des données socio-démographiques relatives au lieu de domiciliation de l'assuré et des données relatives aux caractéristiques des véhicules assurés.

#### → Données géographiques

Tout d'abord, nous allons récupérer des données publiées par l'INSEE. Nous raisonnerons à deux niveaux différents : au niveau de la commune et au niveau de la Zone Économique (ZE) qui nous permettra d'avoir une vision plus large sur l'environnement de vie de l'assuré.

La clé nous permettant de rattacher les données présentes dans les bases de données de l'INSEE à notre base d'étude sera le code commune INSEE. Par la suite, une table de correspondance entre les codes communes et les codes ZE nous permettra de rattacher les données relatives aux ZE.

L'environnement de vie d'un individu influence fortement l'usage qu'il devra faire de son véhicule. Par exemple, dans une commune à forte densité de population et une concentration importante d'équipements (commerces, éducation, culture, transports, etc...), la nécessité d'utiliser un véhicule pour se déplacer est moindre, autant en raison des distances à parcourir que pour la couverture des

réseaux de transports en commun. Pour caractériser cela, nous récupérons les informations relatives au nombre d'équipements de la commune, la densité de population (échelle de 1 à 4 ; 1 étant associé à une forte densité) ainsi que le taux de motorisation.

Par ailleurs, les trajets à des fins professionnels représentant une part assez importante des déplacements, on peut également s'intéresser à l'environnement socio-économique entourant le domicile. Ainsi, nous récupérons les informations relatives au pourcentage de la population de la commune étant âgé de respectivement 15-24 ans, 25-54 ans et plus de 55 ans, ainsi que le taux de chômage de la commune et de la ZE, et plus spécifiquement le taux de chômage des 25-54 ans.

Enfin, les informations relatives aux revenus individuels des assurés n'étant pas disponibles lors de la tarification d'un contrat automobile – cette information étant requise pour certains produits d'assurance vie par exemple –, l'apport du revenu médian par commune est un élément qui pourrait permettre d'estimer le pouvoir d'achat dont dispose le client.

### → Données SRA

Dans un second temps, la nature des véhicules jouant un rôle prépondérant dans la tarification, il est important de pouvoir caractériser de manière précise le bien assuré. En outre, certains types de véhicules font l'objet de comportements différents en matière d'achat et de revente.

La table de correspondance SRA (Sécurité et Réparation Automobiles) est une table qui recense les caractéristiques techniques et commerciales de la plupart des modèles de véhicules commercialisés sur le marché automobile français, à l'exception près de certains modèles haut de gamme et des véhicules de collection. Chaque modèle de véhicule dispose ainsi d'un numéro unique d'identification SRA. Cette information étant présente dans notre base d'étude, c'est par cette clé d'identification que nous pourrions rattacher les informations de cette base externe.

Parmi l'ensemble des caractéristiques disponibles, nous allons sélectionner celles qui nous semblent déterminantes pour expliquer le changement de véhicule. Ainsi, nous ne tiendrons pas compte des caractéristiques physiques du véhicule (poids, longueur, largeur, hauteur, etc...) qui sont trop spécifiques et diverses et qui compliqueraient la modélisation.

Parmi les variables retenues, les variables les plus importantes sont le groupe et la classe SRA qui sont deux facteurs rentrant en ligne de compte au moment de la tarification.

Noté de 20 à 50, le groupe SRA représente la puissance du véhicule mais également sa dangerosité – 20 étant attribué à une faible puissance et 50 à une forte puissance. Cette donnée est notamment utilisée pour la tarification de la garantie responsabilité civile puisque la puissance est intimement liée aux dommages que peut causer le véhicule.

La classe SRA du véhicule, quant à elle, peut être séparée en deux valeurs :

- **La classe de prix**, notée de A à V (A représentant une faible valeur et V une forte valeur) permet d'établir la valeur à neuf TTC du véhicule.
- **La classe de réparation**, notée de A à ZE, permet d'établir la valeur des pièces pour réparation. Cette valeur est utile à l'assureur pour estimer les coûts en cas de sinistres.

La classe de réparation n'étant pas systématiquement renseignée dans la table de correspondance SRA, nous récupérerons la classe de prix.

Outre ces variables permettent de caractériser la puissance du véhicule et sa valeur, nous compléterons les données relatives aux véhicules assurés par plusieurs variables :

- **Le nombre de places**
- **L'appartenance du modèle à une série limitée** ou non
- **Le nombre de cylindres**
- **La cylindrée en cm<sup>3</sup>**
- **La puissance en DIN** qui représente la puissance réelle au niveau des roues du véhicule selon l'unité de mesure des chevaux dits vapeurs. Elle peut entrer en ligne de compte dans le calcul de la prime automobile
- **La vitesse maximale** du véhicule
- **La puissance administrative** qui permet de mesurer la puissance du véhicule selon l'unité de mesure des chevaux dits fiscaux. La puissance fiscale est utilisée dans le calcul de la prime d'assurance automobile mais également pour déterminer le montant des taxes sur les certificats d'immatriculation, calculées par région, dues lors de l'immatriculation du véhicule
- **Le dernier prix de vente à neuf connu** (en euros)

Lors du rattachement de ces données externes depuis la base de correspondance SRA vers notre base d'étude, nous avons rencontré deux problèmes majeurs : la présence de modèles non identifiés dans la base d'étude et des données manquantes.

Tout d'abord, certains modèles de véhicules présents dans la base d'étude, identifiés par leur numéro SRA, ne font pas partie de la liste des modèles décrits dans la base de correspondance SRA. Cela peut s'expliquer par deux raisons.

La première est que certains modèles manquants font partie du périmètre des modèles n'étant pas décrits dans la table SRA, c'est-à-dire certains véhicules haut de gamme, les véhicules de collection ou les véhicules importés de l'étranger mais non commercialisés en France. Pour ces véhicules-là, nous ne pourrions récupérer aucune information supplémentaire.

La seconde raison tient dans la qualité de renseignement des numéros SRA dans la base d'étude. En effet, pour certains contrats, la clé d'identification des modèles est non ou mal renseignée. Pour ces véhicules, nous allons récupérer les caractéristiques des modèles les plus proches en utilisant une analyse de correspondance avec la distance de Jaro-Winkler.

Autrement dit, cela revient à comparer le modèle du véhicule pour lequel le code SRA est manquant avec l'ensemble des modèles de la même marque. Le modèle et ses caractéristiques qui sera affilié au modèle manquant est celui qui maximise la distance de Jaro-Winkler, le score obtenu devant dépasser 0.5 afin de s'assurer qu'un modèle trop éloigné ne soit pas rattaché par défaut. Cette méthode nous permet de corriger les effets des erreurs de saisie dans le nom des modèles ou du manque d'informations issue d'un code de correspondance mal renseigné.

Numéro contrat	Marque	Modèle	Code SRA
AR543654	FORD	RANGER 2.2 TDCI 150 SIMPLE	

Marque	Modèle	Code SRA	Distance de Jaro-Winkler
FORD	RANGER 2.2 TDCI 125 SIMPLE CABINE XL	FO30029	0.7386
FORD	RANGER 2.2 TDCI 150 SIMPLE CABINE XL	FO30028	0.7663
FORD	RANGER 2.2 TDCI 125 DOUBLE CABINE XL	FO30023	0.7107
FORD	RANGER 2.2 TDCI 143 SIMPLE CABINE XL	FO30011	0.7243
...	...	...	...
FORD	ESCORT 2.0T COSWORTH 4x4	FO07353	0.5705

Numéro contrat	Marque	Modèle	Code SRA
AR543654	FORD	RANGER 2.2 TDCI 150 SIMPLE	FO30028

TABLE 3.5 – Analyse de correspondance des modèles de véhicules avec la distance de Jaro-Winkler

En revanche, les contrats pour lesquels le modèle n'est pas renseigné tandis que le code SRA ne correspond à aucun numéro connu seront des contrats pour lesquels nous ne pourrons pas non plus récupérer de l'information concernant les caractéristiques détaillées du véhicule assuré.

Ensuite, la seconde problématique rencontrée est un nombre assez important de données manquantes concernant le dernier prix de vente à neuf connu du modèle. Cette donnée, bien qu'en partie contenue dans l'information de la classe SRA, paraît particulièrement importante au moment de prédire un changement de véhicule puisque les véhicules de différentes classes de prix ne font pas l'objet des mêmes types de comportement. Ainsi, nous avons estimé le prix médian des véhicules par marque, classe, groupe et nombre de places. Pour les modèles pour lesquels cette donnée n'était pas disponible, nous avons rattaché le prix médian correspond à sa marque, classe, groupe et nombre de places.

### 3.2.3 Traitement *a posteriori* sur la base de données

Une fois la base d'étude construite, il est nécessaire d'y effectuer quelques traitements *a posteriori*. Ils s'attèlent à la fois à la gestion des valeurs manquantes et aberrantes ainsi qu'à la segmentation de certaines variables quantitatives.

#### 3.2.3.1 Suppression des valeurs aberrantes

En statistiques, les valeurs aberrantes sont des observations qui n'appartiennent pas à une population en particulier et qui s'écartent des autres données, par ailleurs bien structurées. Détecter de telles valeurs est particulièrement important puisque ces dernières pourraient adopter un comportement sous-jacent totalement différent du reste des données et ainsi biaiser le pouvoir prédictif des modèles que nous mettrons en place par la suite.

En effet, les algorithmes d'apprentissage automatique (*machine learning*) sont sensibles à l'intervalle et à la distribution des valeurs. Les données aberrantes pourraient induire une erreur dans le processus d'apprentissage, ce qui se traduirait par des temps d'apprentissage plus longs, des modèles moins précis et par conséquent de moins bons résultats.

Comment les valeurs aberrantes sont introduites dans l'ensemble des données ? Cela intervient par deux moyens : au moment de la collecte des données ou au moment de leur renseignement dans les systèmes d'information.

La première raison évoquée fait sens en assurance puisque les populations assurées en automobile peuvent être très hétérogènes. Ainsi, les valeurs aberrantes ne correspondent pas nécessairement à des "erreurs" dans les systèmes mais font partie de sous-population dont les caractéristiques semblent anormales en comparaison des caractéristiques moyennes observées sur le portefeuille. Néanmoins, elles peuvent également correspondre à des déclarations erronées de la part des assurés.

La seconde raison évoquée est un problème récurrent dans les métiers faisant usage de systèmes d'information : la qualité des données. Ces valeurs extrêmes peuvent ainsi provenir d'erreur de saisie dans les systèmes, d'erreurs d'unité de mesure, etc. . .

Si possible, les valeurs exceptionnelles doivent être exclues de l'ensemble des données. Cependant, cette suppression ne doit pas être effectuée sans une analyse préalable car ces dernières pourraient également apporter une information non négligeable. Par ailleurs, exclure certaines caractéristiques de notre base d'étude pourrait nous mener à construire un modèle qui ne serait pas apte à gérer ces valeurs au moment de prédire le changement de véhicule pour une nouvelle année. Indirectement, cela impliquerait donc que l'on décide d'exclure certains clients des campagnes de rétention. Il faut donc mettre dans la balance le coût opérationnel de la conservation de telles valeurs et leur impact sur l'analyse et la prédiction.

Outre la connaissance métier qui permet de détecter les valeurs ne faisant pas partie des échelles de grandeurs possibles pour une variable, diverses méthodes statistiques ont été développées pour identifier les valeurs qui paraissent exceptionnelles au regard de la distribution d'une variable donnée.

Nous utiliserons principalement les histogrammes qui permettent d'analyser la répartition des effectifs d'une variable en fonction des différentes valeurs prises, ainsi que les "boxplot", plus connu en français sous le nom de "boite à moustaches".

Une fois les valeurs aberrantes identifiées au moyen de différentes méthodes statistiques, il existe plusieurs façons de les traiter dans notre base de données :

- **Par élagage** : supprimer de notre analyse les contrats pour lesquels la valeur d'une variable donnée dépasse un certain seuil. En opérant de la sorte, la distribution des données de la variable considérée devient moins épaisse qu'en présence des valeurs extrêmes. Cela permet notamment d'accélérer les temps d'apprentissage des modèles.
- **Par plafonnement** : on établit une limite puis on plafonne les valeurs aberrantes à cette valeur. Autrement dit, pour une variable donnée, on remplace la valeur observée par le seuil défini pour les contrats présentant une valeur en dessous du seuil minimal ou au-dessus du seuil maximal que l'on s'est fixé. Il est également possible d'utiliser la médiane du portefeuille comme valeur de remplacement.

- **Traiter les valeurs aberrantes comme des valeurs manquantes** : les contrats pour lesquels la valeur d'une variable considérée dépasse les seuils fixés seront considérés comme ne disposant pas de l'information pour cette variable donnée.
- **Par discrétisation** : on inclut les valeurs extrêmes dans un groupe particulier et on les force à se comporter de la même manière que les autres points de ce groupe.

Pour le traitement *a posteriori* de la base de données, nous combinerons une analyse au moyen des boxplot, des histogrammes et de l'analyse des comportements en matière de changement de véhicule (toute forme, contrats actifs, contrats résiliés) pour identifier les valeurs aberrantes qu'il serait jugé utile de supprimer de notre base d'étude afin d'optimiser le pouvoir prédictif des modèles que nous mettrons en place.

Nous appliquerons ces méthodes d'analyse statistiques aux variables quantitatives tels que l'âge de l'assuré, l'âge du véhicule, le prix à neuf du véhicule, la prime du contrat, la revalorisation de la prime par rapport à l'année précédente, l'écart au tarif, l'ancienneté contrat et client, etc...

Tout d'abord, nous allons analyser les valeurs relatives à la prime payée par le client dans le cadre de sa police d'assurance automobile : le CET, la revalorisation et le montant de la prime totale hors taxes. Ces dernières jouent un rôle prépondérant dans la résiliation et dans le choix de se tourner vers la concurrence puisque la plupart des individus sont sensibles au prix.

### → Le CET

Le premier point d'intérêt est le coefficient d'écart au tarif (CET). Il est calculé comme le rapport entre ce que le client paye réellement et ce qu'il devrait payer s'il ne faisait pas l'objet de majoration ou de réduction commerciale. Il permet ainsi d'avoir une idée sur les politiques tarifaires mises en place par Generali à l'encontre d'un client. A noter que le coefficient bonus-malus n'est pas pris en compte dans le calcul de l'écart au tarif.

Sur l'ensemble du portefeuille, la distribution des valeurs du CET est concentrée entre 0.8 et 1.2. Autrement dit, cela correspond à un intervalle de 20% de réduction commerciale et 20% de majoration, ce qui est cohérent avec la politique tarifaire de Generali.

En revanche, on remarque que certains contrats présentent des CET inférieurs à 0.5 voire des valeurs négatives, ce qui est techniquement impossible puisque cela correspondrait à des niveaux de réductions commerciales supérieurs à 50% du tarif initial, qui est la réduction maximale accordée pour un client.

Par ailleurs, on constate qu'une faible part des contrats présente un CET supérieur à 2, signifiant que les clients en question paieraient plus du double du tarif initial. L'observation du boxplot nous permet de confirmer que ces valeurs semblent "anormales" au regard de la distribution du portefeuille. De plus, on peut constater que les comportements en matière de changement de véhicule (contrats actifs, résiliations, toute forme) commencent à devenir volatiles pour des valeurs aux alentours de 2, en raison de la faible volumétrie des contrats.

Ces analyses nous conduisent à considérer des valeurs du CET inférieures à 0.5 ou supérieures à 2 comme des valeurs extrêmes. L'étape suivante consiste à choisir la manière de gérer ces données.



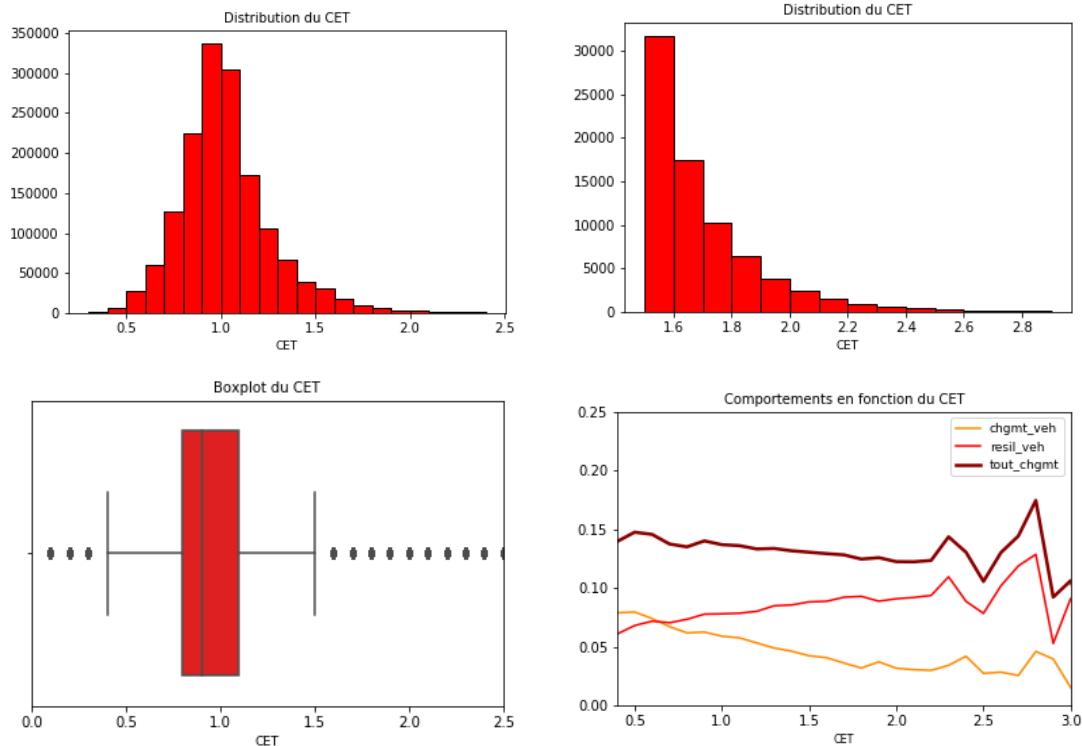


TABLE 3.6 – Analyse des valeurs extrêmes pour le CET du contrat

Supprimer de tels contrats de notre base d'étude auraient-ils du sens ? Pour les valeurs supérieures à 2, la réponse est à l'affirmative puisque les clients présentant de tels niveaux de revalorisation sont des clients que la compagnie ne souhaite pas "protéger" en lui proposant des conditions tarifaires avantageuses et pour lesquels une résiliation ne serait pas nécessairement dommageable. Il serait alors tout à fait acceptable de ne pas effectuer de prédictions sur ces contrats et ainsi les exclure des campagnes de rétention au changement de véhicule.

Cependant, il faut également prendre en considération le fait que de telles valeurs peuvent être le résultat d'erreurs commises dans les systèmes d'information. En effet, le montant de rabais accordé et ainsi la prime de base ne sont pas des données dont la fiabilité est facilement vérifiable. Des valeurs erronées peuvent conduire automatiquement à des valeurs du CET déconnectées de la réalité.

En tenant compte de l'ensemble de ces considérations, le choix a été fait de traiter les valeurs aberrantes du CET par plafonnement c'est-à-dire d'attribuer la valeur du seuil maximal (2) ou du seuil minimal (0.5) aux valeurs considérées comme extrêmes.

### → La prime totale hors taxes

Le second point d'intérêt concerne la valeur de la prime en elle-même. Nous disposons de cette valeur sous la forme du montant total hors taxes annualisé payé par l'assuré.

L'analyse graphique au moyen d'un histogramme, d'un boxplot et de l'analyse des comportements en fonction de la valeur de la prime annuelle hors taxe se trouve dans l'annexe A.1.

La distribution des valeurs suggère qu'une grande majorité des clients payent une prime comprise entre 300€ et 600€. La volumétrie des contrats décroît très nettement à partir d'un montant supérieur à 1000€, qui apparaît donc comme une valeur extrême au regard de la distribution du portefeuille de Generali. En reprenant les ordres de grandeurs énoncés dans le chapitre 1, la prime moyenne sur le marché français se situe aux alentours des 600€. Il n'est donc pas étonnant qu'une telle valeur puisse être considérée comme disproportionnée. Un tel montant peut aussi bien se référer à un véhicule assuré à forte valeur et/ou une forte couverture, qu'à un client à forte sinistralité que la compagnie a décidé de fortement majorer.

Exclure de notre base d'étude de tels contrats reviendrait à écarter de nos politiques de rétention les contrats avec des primes élevées, contrats qu'il pourrait être intéressant de garder en portefeuille pour s'assurer un volume de chiffres d'affaires, à condition que le risque ne soit pas trop élevé.

Le choix des contrats pour lesquels nous déciderons de mettre en place des politiques de rétention n'intervenant qu'à la fin du processus de modélisation, nous décidons de conserver l'intégralité des contrats dans notre base d'étude, quel que soit leur niveau de prime.

### —> **La revalorisation de la prime**

Enfin, le dernier élément relatif au tarif à analyser est la revalorisation de la prime. Chaque année, Generali revalorise les primes de ses contrats en tenant compte de la valeur qu'elle attribue à chaque client. La sensibilité au prix étant un facteur à prendre en compte pour expliquer la résiliation, il est nécessaire que la revalorisation de la prime soit une donnée suffisamment fiable.

Cette dernière a été calculée en comparant les primes de l'année N et de l'année N-1. Il est important de distinguer les revalorisations qui sont liées à une modification du risque telle qu'un changement de véhicule ou un changement de formule, de celles qui découlent des politiques de la compagnie.

L'analyse graphique au moyen d'un histogramme, d'un boxplot et de l'analyse des comportements en fonction de la revalorisation de la prime se trouve dans l'annexe A.2.

Les revalorisations mises en place par la compagnie se situent dans un ordre de grandeur compris entre 1 et 10% chaque année. En ayant connaissance de cette politique, on peut considérer que les valeurs supérieures à 25%, sans survenance d'un sinistre, correspondent à des contrats pour lesquels une erreur a été introduite dans les systèmes de données, pour au moins l'une des années servant de base à la création de la variable.

De la même façon, les contrats pour lesquels la revalorisation de la prime est inférieure à -50% d'une année à l'autre paraissent suspects. En effet, comme nous l'avons mentionné, la réduction maximale accordée dans le cadre de la souscription de plusieurs produits d'assurance est de l'ordre de 50%.

Supprimer les contrats présentant des échelles de revalorisations anormales nous paraît cohérent pour notre étude puisqu'il est primordial que le montant des primes payées par le client soit une donnée fiable. Une valeur aberrante laisse supposer que la donnée sur la prime est mal renseignée pour au moins l'une des deux années utilisées pour calculer la revalorisation.

Ainsi, sont supprimées les revalorisations inférieures à -50% et celles supérieures à 25%, pour les contrats n'ayant pas subi de sinistres entre les deux années ni une modification du risque assuré.

Ensuite, nous allons analyser les variables en lien avec les caractéristiques du véhicule. Le changement de véhicule étant principalement lié au bien en lui-même, il est important de pouvoir appréhender les caractéristiques des véhicules qui constituent notre portefeuille.

→ **L'âge du véhicule**

La première caractéristique du véhicule qui nous intéresse est son âge, qui est à première vue un facteur déterminant dans la prise de décision de son remplacement.

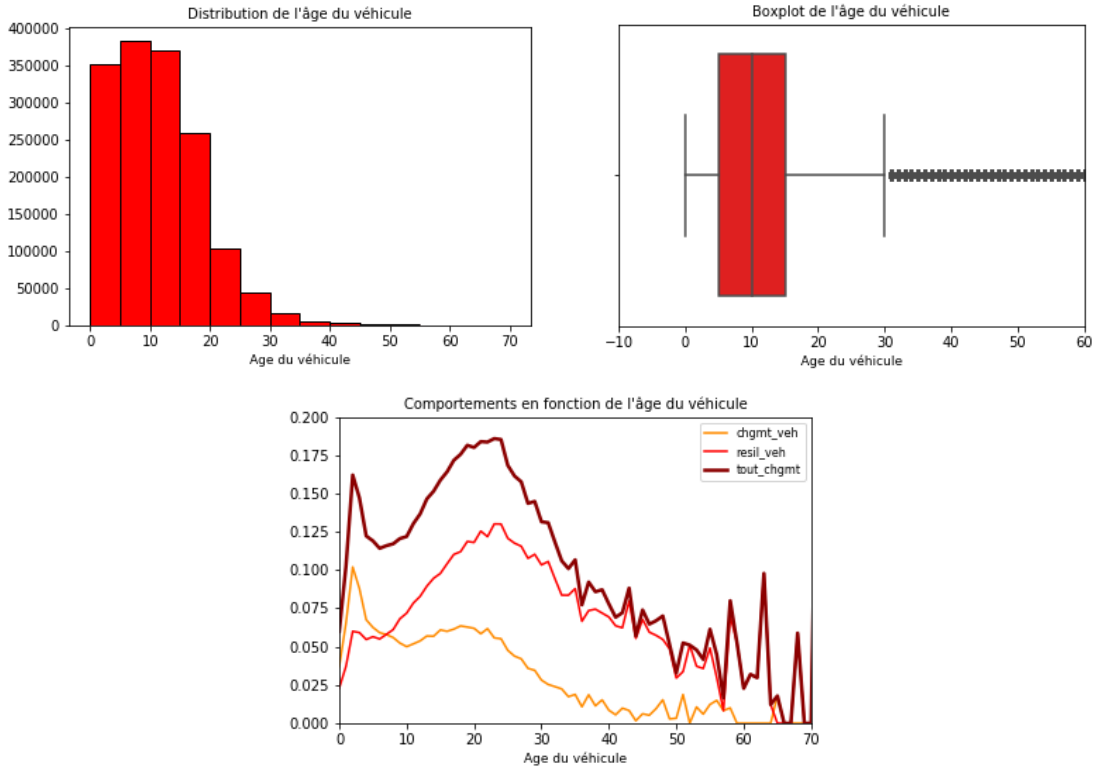


TABLE 3.7 – Analyse des valeurs extrêmes pour l'âge du véhicule

En observant la distribution des valeurs du portefeuille, on constate que la grande majorité de la flotte de véhicules assurés est âgée entre 0 et 20 ans.

Les valeurs exceptionnelles concernent des véhicules âgés de plus de 30 ans, ce qui se confirme par les comportements en matière de changement de véhicule qui deviennent extrêmement volatiles à partir de cette ancienneté, en raison de la faible volumétrie de ces véhicules.

Ces véhicules considérés comme "exceptionnels" sont donc des véhicules datant des années 1990 dont la technologie, l'ergonomie ou les caractéristiques purement esthétiques sont largement dépassées par rapport aux véhicules commercialisés de nos jours. Leur détention pourrait alors ne pas correspondre à des motivations rationnelles. Prédire le changement de véhicule sur cette population pourrait s'avérer difficile puisque les facteurs à prendre en considération pourraient être totalement différents des facteurs pris pour les véhicules plus récents.

La volumétrie de ces véhicules étant relativement faible (moins de 30 000 contrats sur l'ensemble des 3 années d'étude), il nous paraît judicieux de supprimer ces contrats de notre base de données.

#### —> **La durée de détention du véhicule**

Directement en lien avec l'âge du véhicule, la durée de détention du véhicule peut également faire l'objet d'une attention particulière.

Si la durée de détention d'un véhicule est généralement comprise entre 0 et 10 ans, les détentions supérieures à 15 ans peuvent être considérées comme des valeurs exceptionnelles au regard de la distribution du portefeuille.

Dans la réalité, une telle durée de détention ne nous paraît pas être particulièrement élevée au regard du coût que représente l'achat d'un véhicule, de l'essor de la technologie qui permet de prolonger la durée de vie des véhicules et de l'utilisation plus ou moins fréquente qu'un individu peut faire de son bien. En parallèle, les valeurs relativement basses de la détention d'un véhicule, qui se différencie de son âge, peuvent s'expliquer par la prépondérance du marché de l'occasion qui permet aux propriétaires de véhicules d'en changer plus régulièrement grâce à la réduction de la dépense qu'il engendre.

Parce que certaines valeurs seront déjà éliminées de notre base de données lors de la suppression des valeurs aberrantes de l'âge du véhicule, que la détention d'un véhicule est liée à l'environnement de vie d'un individu et que les seuils déterminés par les différentes analyses graphiques ne nous semblent pas satisfaisants, nous décidons de ne pas effectuer de traitement particulier sur les valeurs extrêmes de la durée de détention d'un véhicule.

#### —> **Le dernier tarif à neuf connu du véhicule**

Enfin, la dernière caractéristique sur laquelle se portera notre attention est la valeur du dernier tarif connu à neuf, en euros, du bien assuré. Cette donnée est particulièrement intéressante parce qu'elle apporte de l'information à la fois sur la catégorie du véhicule assuré, sur la dépense qu'a pu engendrer l'acquisition de ce véhicule, et donc indirectement sur le niveau de vie de l'assuré, mais également sur les comportements de revente du véhicule.

En effet, à l'exception des véhicules de collection, le prix d'un véhicule neuf subit en moyenne une dépréciation de 15% la première année, de 10% les 3 années suivantes puis une décote de 6% pendant les 4 prochaines années. Un véhicule dont la valeur d'achat à neuf est relativement élevée pourrait ainsi être sujet à une revente rapide afin de remédier à cette dépréciation de la valeur, le gain tiré de la revente servant d'apport pour l'achat d'un nouveau véhicule.

L'analyse graphique au moyen d'un histogramme, d'un boxplot et de l'analyse des comportements en fonction de la valeur du véhicule à neuf se trouve dans l'annexe A.3.

Le portefeuille de Generali sur le périmètre que nous avons défini est principalement constitué de véhicules de classe moyenne avec une valeur comprise entre 10 000€ et 30 000€. La volumétrie de ces véhicules étant très conséquente, les véhicules dont la valeur d'achat à neuf excède 45 000€ sont considérés comme exceptionnels.

Le changement de véhicule, quelle que soit sa forme, en fonction de la valeur d'achat à neuf du véhicule présente une tendance stable jusqu'à une valeur d'achat de 100 000€ pour ensuite présenter des comportements beaucoup plus volatiles.

S'ils représentent une faible part de la flotte des véhicules assurés, supprimer ces véhicules de notre périmètre serait-il cohérent relativement aux politiques de rétention à mettre en place ? Les exclure signifierait que nous ne souhaitons pas mettre en place des actions commerciales pour les véhicules haut de gamme. Or, en termes de primes et de potentiels contrats souscrits, les clients concernés sont intéressants pour la compagnie. Les modèles de classification que nous utiliserons par la suite sont en capacité de gérer ces valeurs en séparant les véhicules haut de gamme des autres véhicules et en leur affectant des comportements particuliers. Pour ces raisons, la valeur du véhicule ne sera pas un critère restrictif pour notre base de données.

Enfin, le dernier angle d'analyse concerne les caractéristiques des clients eux-mêmes.

→ **L'âge du conducteur**

Concernant l'âge du conducteur, les premières valeurs aberrantes sont relatives à d'éventuelles valeurs qui seraient inférieures à 18 ans. En effet, en France, il n'est pas possible de souscrire une assurance automobile avant cet âge-là puisque l'obtention du permis est conditionnée par la majorité. De la même façon, des valeurs supérieures à 120 ans sont considérées comme biologiquement impossibles. Ces données peuvent être remplacées par une valeur manquante.

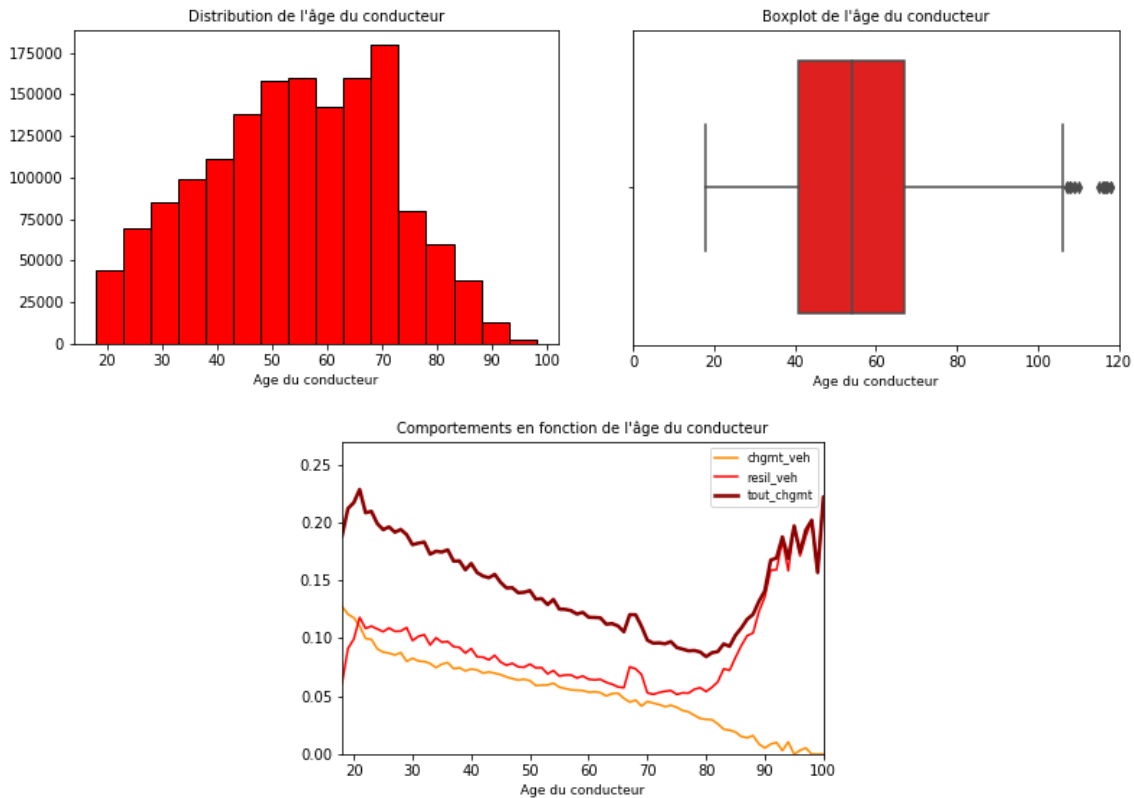


TABLE 3.8 – Analyse des valeurs extrêmes pour l'âge du conducteur

Au regard de la distribution d'âge de notre portefeuille d'assurés, les valeurs extrêmes sont des clients âgés de plus de 90 ans. La probabilité de changer son véhicule passé cette maturité est quasiment nulle. Dans le sens commun cela fait sens puisque ces personnes sont généralement des individus ne conduisant plus et qui peuvent conserver leur véhicule pour une valeur sentimentale ou pour en faire bénéficier des tierces personnes.

Ne pas prendre en compte ces individus dans notre base d'étude ferait sens car ces assurés peuvent constituer un risque très élevé. Par ailleurs, il n'est pas certain qu'ils soient sensibles aux offres de rétention. Pour ces raisons, nous supprimons de notre base les clients âgés de plus de 90 ans.

#### → **L'ancienneté du client**

Enfin, la dernière variable que nous analysons est l'ancienneté du client en portefeuille.

De la même façon que l'âge du véhicule et sa durée de détention, l'ancienneté du client et son âge sont directement liés. C'est pourquoi, nous n'opérerons pas de traitements particuliers sur cette variable, nous assurant seulement que cette dernière reste dans des échelles de grandeurs réalistes - remplaçant ces valeurs aberrantes par des valeurs manquantes.

### **3.2.3.2 Suppression des données manquantes**

Outre la gestion des valeurs aberrantes, les valeurs manquantes doivent également faire l'objet d'un traitement dans le but de préparer la base de données pour la modélisation. En effet, de nombreux algorithmes d'apprentissage ne supportent pas les informations non renseignées.

Au regard de la physionomie de notre base de données, les valeurs manquantes sont bien entendues à gérer uniquement pour les contrats étant actifs et non pour les contrats résiliés pour lesquels nous ne disposerons jamais des informations, à l'exception du motif de résiliation.

De la même façon que pour les valeurs extrêmes, il existe plusieurs façons de gérer ces valeurs :

- **Supprimer les lignes** pour lesquelles les valeurs d'une variable donnée sont manquantes. Dans le cas où la proportion de valeurs manquantes serait trop importante, il est plus judicieux de supprimer la variable plutôt que de perdre une masse conséquente de contrats avec l'information qu'ils apportent pour les autres variables.
- **Imputer les valeurs manquantes avec la médiane ou la moyenne** : pour les variables continues de notre base de données, nous pouvons combler les valeurs manquantes par la médiane ou la moyenne du portefeuille pour la variable considérée. Par rapport à la méthode précédente, cela permet de limiter la perte de données. En revanche, elle peut provoquer des fuites de données et ne prend pas en compte la covariance entre les caractéristiques.
- **Imputer les valeurs manquantes avec la classe majoritaire** : pour les variables qui sont catégorielles, nous pouvons combler les valeurs manquantes par la catégorie la plus représentée dans notre portefeuille. Dans le cas où le nombre de données manquantes serait important, il est également possible de les regrouper dans une classe à part entière. Cette méthode présente les mêmes avantages et inconvénients que la méthode précédente

pour les variables continues. Néanmoins, ajouter une nouvelle catégorie pour les valeurs manquantes peut complexifier la tâche prédictive puisque le nombre de classes à encoder devient plus important.

- **Utiliser la méthode de la dernière observation reportée** (LOCF en anglais) : pour une variable ayant un comportement longitudinal, il peut être judicieux d'utiliser la dernière observation pour combler la valeur manquante.
- **Remplacer la valeur manquante par une valeur par défaut**

En fonction de la nature des variables qui constituent notre portefeuille et de l'importance de l'information qu'elles apportent, nous adapterons le traitement des valeurs manquantes.

Ainsi, pour les variables relatives à l'âge de l'assuré, l'âge du véhicule, l'ancienneté du contrat et du client, la prime totale payée hors taxes, la formule souscrite, etc. . . , nous supprimons de notre base d'études les contrats pour lesquels une de ces valeurs n'est pas renseignée. En effet, l'information apportée par ces variables est singulière pour le changement de véhicule et il ne serait pas pertinent de combler les valeurs non renseignées par les valeurs moyennes ou médianes du portefeuille. Imputer une valeur arbitraire et unique à l'ensemble des valeurs manquantes pourrait ainsi induire un biais dans nos modèles prédictifs, coût d'autant plus lourd à supporter que la volumétrie des valeurs manquantes n'est pas très conséquente.

Pour les variables relatives aux derniers mouvements effectués sur le contrat (date de dernière souscription, dernière résiliation, nombre d'affaires nouvelles ou de résiliations sur les 12 derniers mois, nombre de sinistres sur les 12, 24 ou 36 derniers mois, etc. . . ), l'absence de valeur renseignée est une information en elle-même. Une valeur manquante est donc remplacée pour une valeur nulle ou bien par une catégorie indiquant que la valeur n'est pas renseignée, signifiant l'absence d'actions.

Par ailleurs, certaines variables de notre base d'étude ont des valeurs manquantes qui peuvent être facilement comblées par une valeur par défaut. Par exemple, l'absence de renseignements sur le nombre de conducteurs assurés peut être remplacée par une valeur de 1, l'absence d'informations sur le nombre d'enfants peut être remplacée par 0, etc. . .

Enfin, une variable utilisée pour la tarification automobile s'avère être mal renseignée avec parfois une valeur par défaut ne correspondant à l'intervalle de grandeur défini par la réglementation : le coefficient de Réduction-Majoration (CRM). Permettant de se faire une idée sur la qualité de la conduite et la sinistralité du conducteur, cette variable est d'une grande importance et il n'est pas concevable de ne pas la prendre en compte ou de lui imputer une valeur par défaut.

Pour traiter cette variable, nous allons utiliser la méthode de la dernière observation reportée à laquelle nous apporterons une variante. En effet, comme nous l'avons expliqué dans le premier chapitre de ce mémoire, le CRM est une variable dont la valeur évolue d'année en année. Il est donc possible, en observant les valeurs renseignées l'année suivante ou précédente – si ces dernières sont renseignées –, de déduire la valeur manquante pour l'année d'intérêt.

Ainsi, pour les valeurs manquantes du CRM pour lesquelles nous disposons au moins d'une valeur antérieure ou postérieure à l'année considérée, nous pouvons croiser les informations relatives aux sinistres et appliquer les traitements suivants :

- En l'absence de sinistre, nous pouvons appliquer l'évolution normale du CRM c'est-à-dire une réduction de 5%. Nous multiplions donc le CRM par 0.95 par rapport à l'année précédente ou le divisons par 0.95 par rapport à l'année suivante.
- Dans le cas de la survenance d'un sinistre à 50% responsable, le CRM subit une augmentation de 12.5%. Si un tel sinistre a eu lieu, ce que nous pouvons retracer avec les variables relatives aux sinistres, nous multiplions le CRM par 1.125 par rapport à l'année précédente.
- Dans le cas de la survenance d'un sinistre à 100% responsable, la même méthode est utilisée que pour un sinistre à 50% responsable, à l'exception près que la majoration appliquée est de 25% soit un coefficient multiplicateur de 1.25 par rapport à l'année précédente.

Par ailleurs, pour les nouveaux conducteurs (ancienneté du permis inférieure à 1 an) pour lesquels la valeur du CRM serait manquante, nous pouvons leur attribuer la valeur arbitraire de 1 qui est la valeur par laquelle un conducteur novice débute.

Au moment d'appliquer ces différents traitements, il faut également prendre en considération les bornes inférieures et supérieures du CRM qui sont imposées par la réglementation, à savoir une valeur minimale de 0.5 et une valeur maximale de 3.5.

Cette méthodologie permet de compléter une grande partie des observations manquantes. Les contrats restants sans informations sont des contrats pour lesquels nous ne disposons pas non plus des valeurs pour les années précédente et suivante. Pour combler ces dernières observations, nous leur attribuerons la valeur de 1 qui correspond à aucune réduction ni majoration de la prime.

Pour finir, pour les autres variables présentant des valeurs manquantes, nous faisons le choix de ne pas combler ces absences de renseignement par les caractéristiques moyennes ou médianes du portefeuille ou bien par la classe la plus représentée. En effet, le changement de véhicule est un sujet comportemental dont la tâche prédictive représente un défi pour les assureurs. Opérer de la sorte pourrait induire un biais dans nos données et pourrait nous empêcher de capter les caractéristiques fondamentales qui conduisent un individu à changer son véhicule puis faire le choix de rester chez son assureur ou de se tourner vers un concurrent.

De ce fait, si la volumétrie des valeurs manquantes pour ces variables n'est pas trop importante, alors ces dernières seront supprimées; autrement elles ne seront pas prises en compte car elles semblent être de second plan et les rendre robustes se ferait au détriment de la taille de notre base d'étude.

### 3.2.3.3 Regroupement par classes

Une fois la gestion des valeurs manquantes et aberrantes effectuée, il peut être intéressant, notamment pour l'interprétabilité, de constituer des classes à partir de certaines variables continues.

Par ailleurs, il faut également s'assurer que les classes des variables catégorielles automatiquement créées dans les systèmes d'information soient suffisamment représentatives pour avoir un pouvoir explicatif non négligeable au moment de la modélisation du changement de véhicule. Dans le cas contraire, il sera nécessaire de regrouper certaines classes entre elles pour que ces dernières aient des effectifs suffisants.



Ainsi, pour la situation maritale du client, nous pouvons regrouper les clients pacsés et mariés ainsi que les clients divorcés et séparés, qui apportent la même information.

De la même façon, les contrats en leasing représentant une minorité des contrats, nous pouvons effacer la distinction des contrats en leasing avec et sans perte financière afin de réunir un effectif plus conséquent.

Enfin, le type de garage utilisé pour parquer son véhicule est une donnée utilisée dans la tarification des contrats d'assurance automobile et pourrait être une information intéressante car ce dernier peut avoir un lien avec le lieu d'habitation ou le nombre de sinistres subis. Les catégories présentes dans les systèmes d'information sont multiples : sans garages, garages collectifs couverts et non couverts, garages individuels couverts et non couverts. Chacune des informations apportées par ces différentes modalités est singulière. Néanmoins, la faible volumétrie de contrats dans certaines classes pourrait induire de la volatilité dans les comportements étudiés. C'est pourquoi, au détriment d'une information plus spécifique, nous avons décidé de regrouper les garages collectifs couverts et non couverts entre eux ainsi que les garages individuels couverts et non couverts.

Outre les variables catégorielles déjà constituées dans les systèmes d'information, nous pouvons en créer de nouvelles, ne serait-ce que pour analyser de manière descriptive notre portefeuille. On pense notamment à constituer des classes d'âge du conducteur, d'âge du véhicule, d'ancienneté du contrat ou encore d'ancienneté du véhicule. Cela permet de répartir la population en tranches homogènes et d'analyser les comportements de changement de véhicule pour les différentes sous-populations.

Pour ce faire, pour chacune des variables continues citées, nous allons créer un arbre de décision prenant uniquement en compte cette variable et la variable cible du changement de véhicule (toute forme), en spécifiant le nombre minimal d'observations que nous souhaitons dans chaque classe. En fonction des seuils proposés par cette modélisation, nous pourrions constituer des classes dans lesquelles les comportements en matière de changement de véhicule seront homogènes.

A titre d'exemple, pour l'âge du véhicule, l'arbre de décision donne les résultats présentés dans la figure 3.2.

Nous pouvons donc créer les classes suivantes : les véhicules âgés de moins d'un an, ceux dont l'âge est compris entre 1 et 2 ans, 2 et 4 ans, 4 et 10 ans, 10 et 15 ans, 15 et 25 ans et ceux plus anciens.

L'ensemble des traitements que nous avons effectué sur la base de données nous permettent d'avoir des données "propres" avec un minimum de valeurs manquantes, des échelles de grandeur contrôlées, des classes pour les variables catégorielles avec des effectifs suffisants et des comportements intra-classe homogènes. Le dernier traitement *a posteriori* que nous devons effectuer sur notre base d'étude consiste à supprimer l'indice temporel des variables.

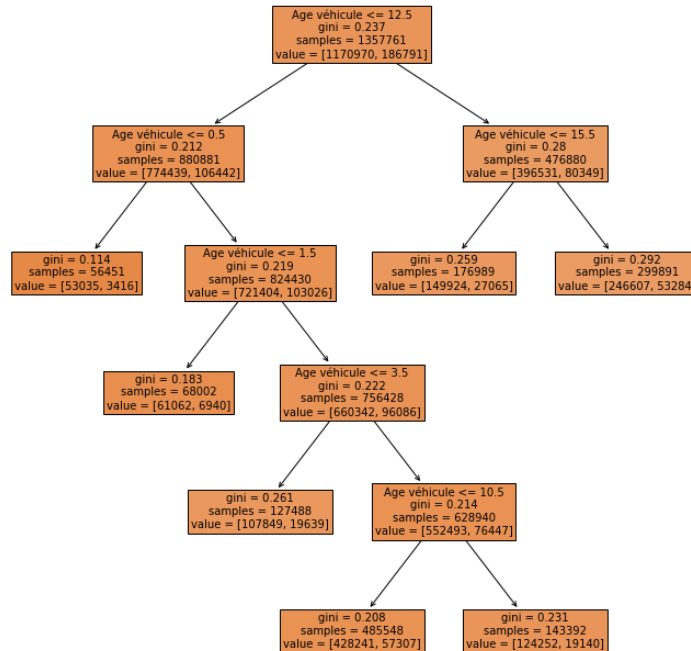


FIGURE 3.2 – Arbre de décision pour la segmentation de l’âge du véhicule

### 3.2.3.4 Suppression de la temporalité des variables

Pour notre modélisation, nous ne souhaitons pas intégrer la temporalité des données collectées : la date à laquelle les caractéristiques du changement de véhicule ont été observées ne devra pas être prise en compte. En effet, le modèle que nous souhaitons implémenter doit pouvoir s’adapter aux années qui suivront la modélisation. De ce fait, il ne nous intéresse pas qu’un changement de véhicule ait eu lieu entre 2017 et 2018 ou entre 2019 et 2020, mais ce sont les caractéristiques qui ont conduit à ce changement qui seront l’objet de notre étude.

Notre base de données étant constituée de telle sorte que les variables sont ajoutées en colonne par année pour un contrat donné, nous devons la transformer afin d’avoir ces informations en lignes tout en supprimant l’indice temporel des différentes variables. Par ailleurs, nous avons exploité la profondeur d’historique de notre base d’étude en créant différentes variables permettant de retracer les évolutions des caractéristiques du portefeuille et d’identifier la survenance des deux phénomènes étudiés. Nous pouvons donc définitivement supprimer les informations de 2016 et 2020.

Ainsi, pour chacune des années d’étude, nous récupérons la liste des contrats actifs avec les variables cibles permettant d’identifier si une forme de changement de véhicule est survenu l’année suivante, les caractéristiques du contrat, conducteur, véhicule, client, sinistres, etc... ainsi que les variables permettant de retracer leurs évolutions et le parcours client. Une fois ces données récupérées, nous effaçons les indicateurs temporels de chaque variable et agrégeons les contrats par ligne.

De cette façon, un contrat présent plusieurs années de suite sur la période d’étude sera représenté par plusieurs lignes dans la base de données finale. Cependant, certaines variables évoluant au cours du temps, nous n’aurons pas plusieurs fois la même photographie du contrat.

Numéro contrat 2017	Age du véhicule 2017	Changement de véhicule 2017	...	Numéro contrat 2018	Age du véhicule 2018	Changement de véhicule 2018	...
AR230987	12	1	...	AR230987	3	0	...

Année	Numéro contrat	Age du véhicule	Changement de véhicule	...
2017	AR230987	12	1	...
2018	AR230987	3	0	...

TABLE 3.9 – Méthodologie de la constitution de la base de données sans indice temporel

La base de données ainsi constituée comporte près de 1.35 millions de ligne, ce qui en fait une base relativement robuste pour la modélisation.

Avant de modéliser le comportement du changement de véhicule, il est nécessaire de faire une analyse descriptive de notre base de données afin d’avoir une meilleure compréhension de notre portefeuille et des individus qui le constituent. Une première analyse permettra d’avoir une visibilité sur les caractéristiques générales de notre portefeuille d’assurés. Par la suite, nous pourrons établir un premier profil des assurés sensibles à changer de véhicule et/ou à résilier pour ce motif.

### 3.3 Analyse descriptive

#### 3.3.1 Caractéristiques générales du portefeuille

##### 3.3.1.1 Les assurés

Sur le portefeuille, la moyenne d’âge des assurés est plutôt élevée et s’établit à 53.3 ans. Ainsi, la population âgée entre 45 et 60 est largement représentée avec près de 28% de notre base d’étude tandis que les individus âgés de moins de 30 ans en représentent moins de 10%.

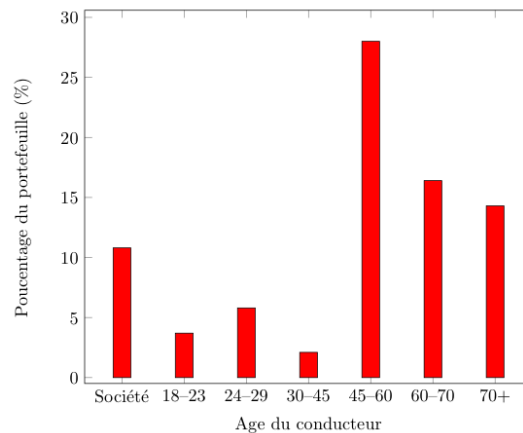


FIGURE 3.3 – Distribution de l’âge des assurés dans le portefeuille

En lien direct avec cet ancienneté relative de la population, la durée moyenne depuis l'obtention du permis de conduire s'élève à près de 31.8 ans. Dans la lignée, il est alors tout à fait cohérent que le coefficient de réduction-majoration soit à son niveau le plus bas, c'est-à-dire 0.5, pour près de 73% du portefeuille. En effet, au moment de l'obtention du permis, ce coefficient s'établit à 1. Une année sans sinistre responsable ou partiellement responsable permettant de diminuer ce coefficient de 5%, il faut donc près de 14 ans sans accident pour bénéficier de la réduction maximale.

Par ailleurs, en accord avec l'âge relativement avancé de la population assurée, près d'un cinquième du portefeuille sont des personnes retraitées. La catégorie socio-professionnelle (CSP) majoritaire est les employés avec près de 40% du portefeuille, suivie par les artisans, commerçants et chefs d'entreprises qui en représentent 10%.

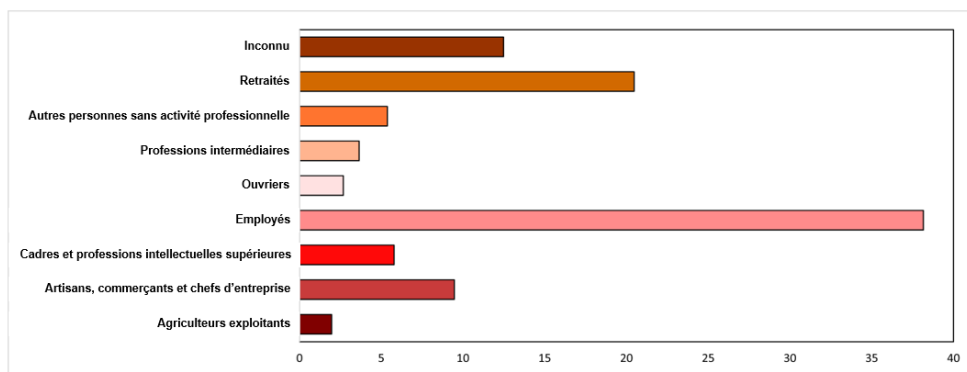
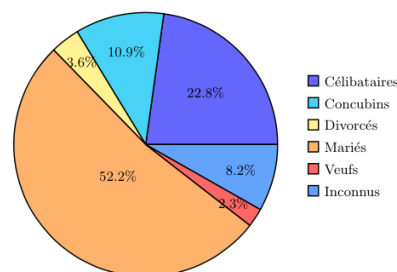


FIGURE 3.4 – Distribution de la catégorie socio-professionnelle dans le portefeuille

Outre l'âge et la catégorie socio-professionnelle, la situation maritale d'un individu est également une caractéristique à prendre en considération puisqu'elle peut influencer le nombre d'individus déclarés sur le contrat. En effet, au moment d'assurer son véhicule, le nombre de conducteurs en faisant usage doit être déclaré, ainsi que leur fréquence d'utilisation, afin d'être couvert en cas d'accident. Ainsi, les compagnies d'assurance définissent un conducteur principal, dont nous analyserons les caractéristiques pour prédire le changement de véhicule, un conducteur secondaire, généralement un conjoint ou un parent proche, mais proposent également des garanties autorisant la présence d'un conducteur occasionnel ou d'un enfant de l'assuré.

Sur notre périmètre d'étude, ce sont près de 52% des assurés qui sont mariés ou pacsés, 11% qui vivent en concubinage et 23% qui se déclarent comme célibataires. Le nombre moyen d'enfants à charge est de 0.58, ce qui s'explique par l'âge avancé de la population assurée. De plus, le nombre moyen de conducteurs assurés sur les contrats est de 1.4 soit seulement un tiers des contrats qui comptent un conducteur secondaire. La plupart des ménages semblent alors posséder un véhicule par personne ou bien une seule une personne du ménage utilise l'automobile comme moyen de transport. Par ailleurs, seulement 5% des contrats ont souscrits à une garantie autorisant un conducteur occasionnel ou la conduite par un enfant du ménage.

FIGURE 3.5 – Distribution de la situation maritale des assurés dans le portefeuille



Si près de 60% du portefeuille sont des particuliers, les professionnels et les petites entreprises représentent respectivement 16% et 9.5% de notre base d'étude.

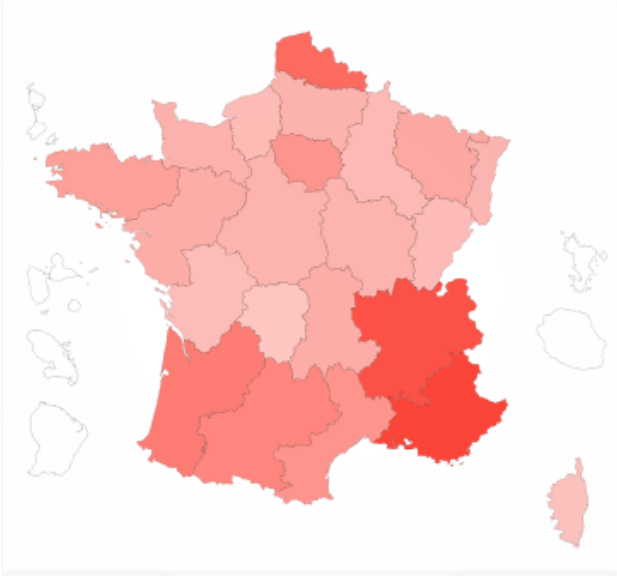


FIGURE 3.6 – Répartition de la population d'étude sur les régions de France (*l'intensité de la couleur étant proportionnelle au nombre de clients*)

du sud du pays, par la proximité avec l'Italie qui est le premier marché de Generali et où la compagnie a donc une implantation plus forte.

### 3.3.1.2 Les contrats

Une fois le profil moyen du client assuré en automobile établi, nous allons nous concentrer sur les caractéristiques des contrats en eux-mêmes puisque ces dernières peuvent donner des indications sur l'attachement de l'assuré à son assureur.

En moyenne, les contrats de notre périmètre ont été souscrits il y a 7 ans tandis que les assurés sont en moyenne présents dans le portefeuille de Generali depuis plus de 14 ans. Ces données mettent en avant la fidélisation du client qui semble bien présente au sein de ce portefeuille, aspect renforcé par les chiffres du portefeuille moyen avec une moyenne de 2 contrats détenus en automobile et 3.2 contrats détenus toutes familles confondues.

Nécessairement, cela signifie que les clients détiennent plus d'une famille de contrats dans leur portefeuille. La tarification des contrats étant propre à chaque famille de produits et des réductions commerciales étant accordées lorsque certaines d'entre elles sont détenues simultanément, l'indicateur du multi-équipement peut être un signal important sur l'attachement du client à son assureur. Sur notre population d'étude, ce sont ainsi près de 65% des clients qui sont multi-équipés et près de la moitié des clients qui sont multi-équipés en auto-MRH, deux produits d'assurance faisant l'objet d'une offre lorsqu'ils sont détenus simultanément dans le portefeuille.

Enfin, il peut être intéressant de s'intéresser à la répartition du portefeuille sur le territoire français. En effet, le réseau d'agents de Generali n'est pas uniformément implanté dans toutes les régions.

Cela se traduit par une présence marquée dans les régions des Hauts-de-France, Bouches-du-Rhône, Provence-Alpes-Côte-D'azur ou, dans une moindre mesure, en Ile-de-France et en Nouvelle Aquitaine.

La surexposition de ces régions par rapport aux autres peut s'expliquer en partie par une densité de population plus forte et donc un potentiel de clients plus important, par un niveau de vie plus élevé que la moyenne nationale (à l'exception de la région du nord de la France) qui en fait une clientèle cible pour Generali et enfin, pour les deux régions

En outre, les garanties souscrites peuvent être de bons indicateurs sur l'importance qu'un client attache à son véhicule. En effet, comme nous l'avons évoqué dans le premier chapitre, le niveau minimal de garantie requis pour circuler ne permet pas de couvrir toutes les formes de dommages. Des véhicules assurés sous cette formule peuvent donc être des véhicules pour lesquels l'assuré ne souhaite pas allouer une part importante de son budget car l'usage qu'il en fait n'est pas régulier, la valeur du véhicule est devenue trop faible, etc... Ainsi, près des deux tiers du portefeuille sont assurés avec la formule la plus complète, ce qu'on appellera la formule "L3", tandis que 12% du portefeuille bénéficient d'un niveau de couverture minimal, que l'on appellera la formule "L1". Par ailleurs, près de 13% du portefeuille ont souscrit une assurance kilométrique dont 9% des assurés qui prévoient d'effectuer moins de 8 000 kilomètres sur l'année.

Enfin, une grande partie des individus étant sensibles aux prix, le montant de la prime annuelle hors taxes est un élément qui sera particulièrement intéressant, notamment pour prédire la résiliation. La moyenne annuelle du portefeuille s'élève à 480 € tandis que la moitié du portefeuille paye une prime comprise entre 304 € et 591 €. Ces cotisations sont payées pour les deux tiers mensuellement et à 28% annuellement. En moyenne, le niveau de cotisation augmente de 1.95% d'une année à l'autre, ce qui représente pour notre portefeuille une hausse de 9.3 €. Par ailleurs, les clients ont en moyenne bénéficié d'une réduction de 2% par rapport au tarif initial, en atteste l'écart au tarif moyen qui s'établit à 0.98.

### 3.3.1.3 Les véhicules

Maintenant que nous connaissons les principales caractéristiques des assurés, celles de leurs contrats et les habitudes moyennes de détention, nous allons présenter le profil des biens assurés.

En adéquation avec le vieillissement du parc automobile français, l'âge moyen des véhicules assurés est de 10.2 ans. Près de 30% des véhicules sont âgés entre 4 et 10 ans, 25% ont entre 10 et 15 ans tandis que 25% des véhicules sont âgés de moins de 4 ans.

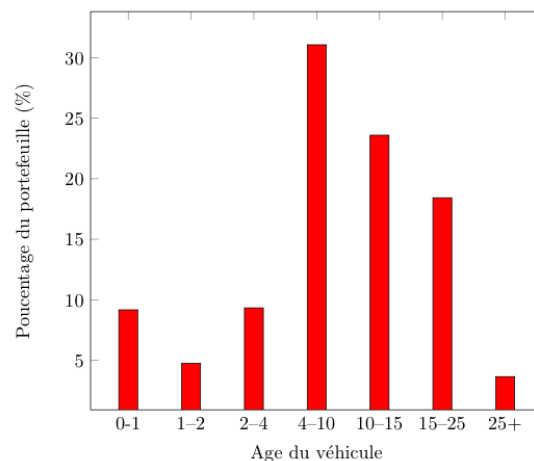


FIGURE 3.7 – Répartition de l'âge des véhicules dans le portefeuille

La durée de détention moyenne des véhicules est de 4.5 ans, ce qui est relativement peu élevé en comparaison à l'âge des véhicules. Cela s'explique par la prépondérance du marché de l'occasion avec

près de 70% des véhicules qui sont des véhicules de seconde main. Même s'ils participent à diminuer la durée de détention des véhicules du portefeuille, les contrats de leasing représentent seulement 5.5% des véhicules assurés.

Comme nous l'avons mentionné précédemment, la classe et le groupe SRA sont deux variables particulièrement importantes pour caractériser le véhicule assuré. Respectivement, elles permettent de donner une information sur le prix à neuf du véhicule et sur sa puissance.

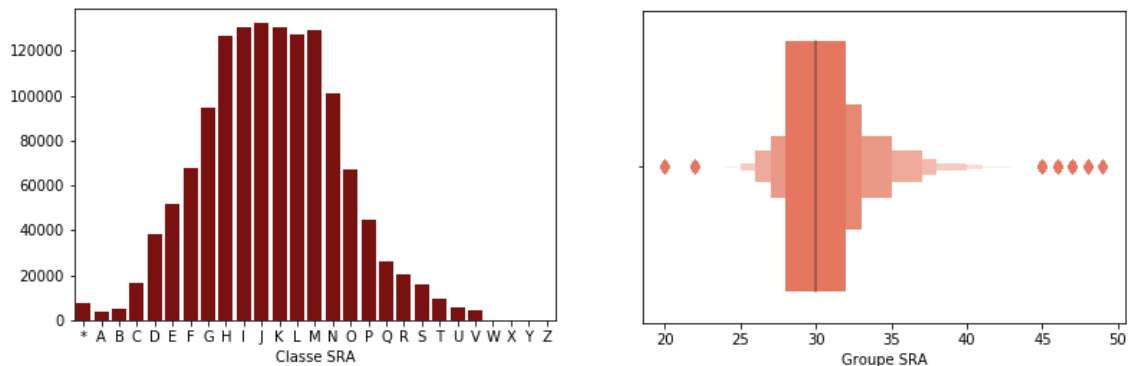


TABLE 3.10 – Répartition du groupe et de la classe SRA dans le portefeuille

Tout d'abord, concernant la classe SRA, plus elle avance dans l'alphabet, plus la voiture est haut de gamme. Sur la population d'étude, on remarque une répartition logique selon les différentes classes avec une surreprésentation des niveaux intermédiaires par rapport aux niveaux extrêmes.

De la même façon, concernant le groupe SRA, dont la valeur est comprise entre 20 et 50, la valeur médiane du portefeuille s'élève à 30, ce qui correspond à une puissance intermédiaire.

### 3.3.2 Caractéristiques sur le changement de véhicule

Parmi les 1.35 millions d'individus de notre base d'étude (1 357 761 pour être précis), 186 791 ont changé de véhicule au cours de l'une des années d'analyse considérées (2018 à 2020). Cela inclut les individus ayant changé de véhicule et qui sont restés chez Generali mais également ceux qui ont résilié leur contrat automobile à la suite de ce changement de véhicule. En analysant séparément les 3 années d'observation, les chiffres restent stables d'une année à l'autre et s'établissent à environ 13.7% du portefeuille qui effectuent un changement de véhicule chaque année.

Intéressons-nous de manière exhaustive au profil des individus qui changent ou non de véhicule.

Si l'on observe l'âge des assurés changeant de véhicule, on remarque une population globalement plus jeune que la population moyenne du portefeuille. En effet, les individus changeant de véhicule ont en moyenne 49.5 ans contre 53.9 ans pour ceux n'en changeant pas. Plus spécifiquement, on remarque que la jeune génération est plus fragile au changement de véhicule avec respectivement 21.4% et 19.6% des individus âgés de 18-23 ans et 24-29 ans qui ont changé de véhicule. A l'inverse, parmi les personnes âgées de 60-70 ans ou de plus de 70 ans, ce sont seulement 10% des individus qui ont adopté ce comportement.

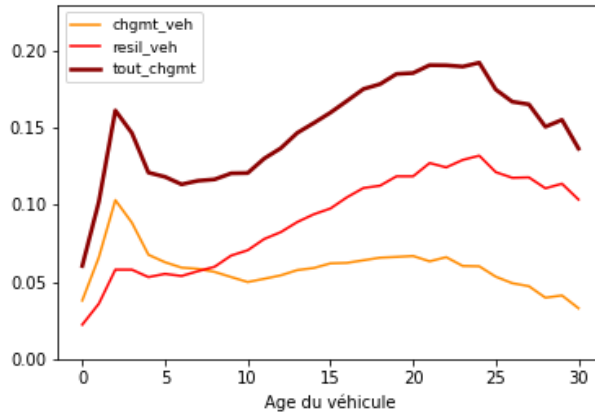


FIGURE 3.8 – Propension à changer de véhicule avec l'âge du véhicule

De la même façon, on constate que les véhicules plus âgés sont plus exposés au remplacement. En effet, l'âge moyen des véhicules qui ont été remplacés est de 11.3 ans contre 10 ans pour les véhicules qui ont été conservés. Cependant, ce constat est à nuancer puisque le changement de véhicule ne présente pas une tendance linéaire avec l'âge du véhicule. Cette tendance peut s'expliquer par la diversité des véhicules qui constituent notre base d'étude, comprenant aussi bien des véhicules d'entrée de gamme que des véhicules haut de gamme, et par le phénomène de décote qui affecte la valeur des véhicules. De manière simplifiée, la dépréciation de la valeur d'un véhicule après son achat à neuf peut conduire les propriétaires à le revendre rapidement après l'achat pour remédier à ce phénomène ou à le conserver sur une plus longue durée pour rentabiliser son coût d'achat qui ne pourra plus être compensé par la revente du bien.

Si les contrats en leasing sont peu représentés dans notre portefeuille, le fait que le véhicule fasse l'objet d'un tel contrat de location avec option d'achat semble être un élément fortement prédictif du changement de véhicule. En effet, un client détenteur de ce contrat est locataire du véhicule pendant une période déterminée, généralement comprise entre 24 et 72 mois, et peut, à son terme, décider de racheter ou non le véhicule. Le changement de véhicule est donc un mécanisme largement simplifié par rapport à un individu propriétaire de son bien.

De manière moins marquée, un véhicule qui a été acheté d'occasion est plus susceptible d'être changé qu'un véhicule qui a été acheté à neuf, sûrement en raison de son ancienneté ou des habitudes de consommation des individus ayant déjà eu recours à ce mode d'achat.

Si l'on regarde les caractéristiques intrinsèques du véhicule, on constate que les véhicules haut de gamme, c'est-à-dire ceux avec une classe SRA et/ou un groupe SRA élevé, sont plus sujets au changement de véhicule. Outre la volumétrie relativement faible sur cette catégorie de véhicule, la nature et la valeur de ces biens peuvent expliquer ces comportements puisque la revente rapide pour éviter la dépréciation de la valeur, la volonté d'acquérir les derniers modèles mis sur le marché ou le simple désir d'effectuer un nouvel achat peuvent entrer en ligne de compte dans la décision de changer cette catégorie d'automobile.

Concernant les caractéristiques du contrat, plus de la moitié des changements de véhicule ont lieu sur des polices relativement récentes (moins de 4 ans d'ancienneté) tandis que les polices plus anciennes (plus de 14 ans d'ancienneté) présentent des taux de changement moins élevés.



Si l'ancienneté du contrat semble jouer un rôle dans le changement de véhicule, les formules souscrites peuvent également être de bons indicateurs pour anticiper cette décision. En effet, on constate que les véhicules assurés avec la formule minimale sont plus susceptibles d'être remplacés au cours de l'année. Cela est directement lié à leur âge puisque les véhicules plus anciens sont généralement assurés avec des garanties moins étendues que ceux plus récents en raison d'une valeur plus faible à protéger. Ainsi, la moyenne d'âge des véhicules assurés avec la formule minimale est de 17.5 ans tandis que celle des biens assurés avec la formule maximale est de 7.1 ans. En outre, les véhicules pour lesquels le contrat n'impose pas une restriction kilométrique font proportionnellement plus l'objet d'un changement de véhicule que ceux n'étant pas restreints. Cela peut s'expliquer par le fait que ces véhicules sont moins utilisés pendant l'année et sont ainsi moins sujets à d'éventuels sinistres.

Enfin, parmi les éléments permettant de démarquer les individus ayant tendance à plus changer de véhicule que les autres, le coefficient de bonus-malus est également un marqueur. On constate ainsi que les conducteurs avec un CRM supérieur à 0.9, c'est-à-dire des conducteurs ayant obtenu leur permis depuis peu ou des conducteurs sinistrés, sont plus fragiles au changement de véhicule. On peut recouper cette analyse avec l'âge du conducteur ou avec le fait que les véhicules qui ont été endommagés ont peut-être plus la nécessité d'être remplacés que les autres.

Outre le profil des individus, il peut être intéressant d'analyser les différences de caractéristiques existantes entre les anciens et les nouveaux véhicules pour appréhender les préférences des individus. Ainsi, sur l'ensemble des clients ayant changé de véhicule et restés en portefeuille – il n'étant pas possible d'observer les caractéristiques des nouveaux véhicules pour les clients ayant résilié – la moitié ont acquis un véhicule d'une classe de prix supérieure. Par ailleurs, 30% des changements de véhicule conduisent à l'achat d'un véhicule neuf, le reste étant dominé par les achats d'occasion qui tirent la moyenne d'âge des véhicules acquis à 6 ans. Enfin, si les trois quarts des changements ne conduisent pas le client à modifier sa formule, le portefeuille étant largement dominé par la formule la plus complète, 20% entraînent un rehaussement de la formule souscrite.

### **3.3.3 Caractéristiques sur la résiliation pour changement de véhicule**

Parmi les 186 971 contrats du portefeuille qui ont changé de véhicule au cours de l'une des années d'observation, 102 925 contrats ont été résiliés suite à ce changement. Cela signifie que près d'un changement de véhicule sur deux conduit à une résiliation. Plus globalement, ce sont près d'un tiers des résiliations d'une année donnée qui le sont pour cause de changement de véhicule. Ces chiffres mettent en relief la problématique de rétention à laquelle est confrontée Generali puisque le changement de véhicule, moment clé dans la vie d'un contrat automobile, est un événement entraînant une fuite de contrats assez importante.

En termes de primes, entre 2018 et 2020, le changement de véhicule représente annuellement 30 millions d'euros, dont 16 millions sortent du portefeuille. En tenant compte du fait qu'un changement de véhicule entraîne en moyenne une augmentation de la prime de l'ordre de 10.5%, le manque à gagner est donc d'autant plus important. Il est donc primordial de comprendre les facteurs conduisant un assuré à résilier son contrat au moment où il décide de changer de bien assuré.

La résiliation pour changement de véhicule étant une donnée intégrée dans la variable permettant d'identifier toute forme de changement de véhicule, les caractéristiques conduisant à la résiliation seront sensiblement les mêmes que celles du changement de véhicule en lui-même.

Ainsi, on constate que lorsqu'un individu relativement jeune change son véhicule, il a plus tendance à résilier qu'un individu plus âgé. De plus, l'acte de résiliation est davantage marqué chez les personnes sans activités professionnelles ou les petites entreprises, tandis que les exploitants agricoles sont les clients les moins sujets à résilier leur contrat.

Comme pour le changement de véhicule dans toutes ses formes, les contrats assurés avec le minimum de garanties, sans restriction kilométriques, les véhicules faisant l'objet d'un contrat de leasing ou achetés d'occasion ou encore les conducteurs avec des coefficients bonus-malus élevés sont autant de profils qui sont sensibles à la résiliation suite à un changement de véhicule.

L'âge du véhicule est de nouveau un facteur qui va retenir notre attention puisque les différentes formes de changements de véhicule, c'est-à-dire ceux restants en portefeuille et ceux résiliés, présentent des tendances inverses en fonction de l'âge du véhicule. Comme on peut le remarquer sur la figure 3.8, plus un véhicule est âgé, plus la probabilité que le client change ce véhicule et résilie son contrat automobile chez Generali est forte. À l'inverse, plus un véhicule est âgé, plus la probabilité que le client change son véhicule et modifie son contrat par un avenant est faible.

Lorsque l'on observe le niveau de résiliation en fonction de la classe SRA, on observe des taux de résiliation beaucoup plus importants sur les véhicules haut de gamme, laissant entendre une sous-compétitivité du produit sur cette catégorie de véhicule. Cependant, cela est à mettre en relief avec la faible volumétrie de contrats. De la même façon, le faible volume de véhicules d'entrée de gamme entraîne une forte volatilité des taux de résiliation.

Par ailleurs, l'ancienneté du contrat joue en la faveur de l'assureur puisqu'on observe une tendance décroissante de la résiliation avec la durée du contrat. Pour les 4 premières années de vie d'un contrat, le taux de résiliation est relativement important et se situe à 9%. Au bout de 14 ans, il n'est plus que de 5%. Ainsi, plus de la moitié des résiliations pour changement de véhicule correspondent à des contrats antérieurs à 4 ans. Cette relative fragilité renvoie aux procédés de souscription et aux politiques de revalorisation car il est surprenant que les clients les plus récemment en portefeuille soient ceux qui quittent, proportionnellement, le plus rapidement la compagnie à ce moment-clé.

En dehors des caractéristiques que nous avons déjà ciblées pour établir les profils du changement de véhicule, d'autres paramètres entrent en jeu pour expliquer la résiliation.

L'attachement du client à sa compagnie d'assurance est l'un d'entre eux. En effet, comme nous l'avons vu précédemment, la plupart des assurés du portefeuille détiennent plus d'un contrat chez Generali et pour la plupart, de familles d'assurance différentes. De tels clients sont alors moins sujets à résilier leur contrat que des clients mono-détenteurs pour des raisons qui peuvent être multiples : volonté de détenir l'ensemble de ses contrats chez un même assureur, obtention de réduction commerciale grâce à sa fidélité client, relation de confiance avec son assureur, etc... Un client mono-équipé chez Generali a ainsi 11% plus de chance de résilier son contrat pour changement de véhicule qu'un client qui est multi-équipé.

Outre le nombre de contrats détenus par un client, le montant de la prime payée ou encore la formule souscrite, une analyse de l'évolution de ces caractéristiques dans le temps pourrait apporter une information supplémentaire. En effet, un individu qui serait dans un processus de résiliation de ses contrats chez Generali ou encore un client dont la prime a été fortement revalorisée par rapport à l'année précédente pourrait être beaucoup plus fragile à la résiliation.

Ainsi, on constate que les clients qui sont dans un processus de résiliation de leurs contrats, c'est-à-dire ceux dont le nombre de contrats total détenu en portefeuille a baissé par rapport à l'année précédente, ont 1.12 fois plus de chance de résilier leur contrat automobile que les clients dont le portefeuille est resté stable ou s'est élargi. De la même façon, 10% des clients dont la dernière résiliation remonte à moins d'un an résilient leur contrat au moment du changement de véhicule.

Dans le cadre d'une étude sur la résiliation, la prime payée au titre de la police d'assurance est un facteur primordial. Lors de la revalorisation annuelle des primes, l'assuré est alors plus enclin à résilier. Ceci est également vrai au moment du changement de véhicule qui entraîne en moyenne une augmentation importante de la prime en raison de la nouveauté relative du véhicule assuré et de l'élargissement des garanties souscrites qui l'accompagne généralement. De ce fait, l'assuré pourrait estimer que ce nouveau tarif est trop élevé pour son niveau de risque et se tourner vers la concurrence pour trouver un tarif plus compétitif. On remarque ainsi que plus la prime d'un contrat a été revalorisée par rapport à l'année précédente, plus il y a de chances que ce contrat soit résilié au moment du changement de véhicule, comme nous pouvons le voir dans l'annexe A.2.

Enfin, un dernier élément marquant dans la vie d'un contrat d'assurance est la survenance d'un sinistre. Dans notre portefeuille, la proportion des assurés ayant été sinistrés lors des 36 derniers mois est faible avec seulement 15% des assurés qui ont subi un sinistre et 2.2% qui en ont subi plus d'un. Cette sinistralité est légèrement marquée par la résiliation, car si 12% des assurés n'ayant subi aucun sinistre résilient pour un changement de véhicule, ce sont 14% et 16% des individus ayant subi un ou plus d'un sinistre qui clôturent leur contrat, chiffres à mettre en perspective avec la volumétrie des polices sinistrées. L'explication peut provenir à la fois d'un sinistre important qui conduit à la destruction du véhicule entraînant sa disparition ou encore du mécontentement du client suite à la gestion d'un sinistre qui le pousse à changer d'assureur.

En définitive, cette section nous aura ainsi permis d'avoir une meilleure connaissance de notre base de modélisation et d'identifier les profils et les caractéristiques influençant la décision de changer de véhicule et celle de résilier son contrat d'assurance suite à cela. L'étape suivante consiste désormais à modéliser ces deux comportements qui sont des "moments clés" dans la vie d'un contrat automobile.

## 4 Chapitre 4 : Modélisation du changement de véhicule

L'objectif de ce chapitre est de présenter la mise en place des modèles utilisés pour prédire la probabilité qu'un individu de notre portefeuille change de véhicule dans l'année en fonction de ses caractéristiques, celles de son véhicule, des spécificités de son contrat, son parcours client ou encore de son environnement.

Estimer cette probabilité revient à calibrer un score de fragilité au changement de véhicule qui, une fois intégré dans les systèmes d'information, pourra permettre d'identifier les clients qui pourraient faire l'objet d'une offre commerciale dans le cadre des politiques de rétention et d'améliorer le processus de renouvellement tarifaire au changement de véhicule.

Nos données comportant des étiquettes, la modélisation du changement de véhicule fera donc appel à des algorithmes d'apprentissage supervisé. Par ailleurs, l'analyse d'un tel comportement, c'est-à-dire celui de changer ou non de véhicule pendant l'année, relève d'un phénomène de type binaire. En effet, la variable cible que nous avons créé dans le chapitre 3, vaut 1 si l'individu a changé de véhicule durant l'année, qu'il ait résilié ou non son contrat automobile, et 0 dans le cas contraire.

Les modélisations qui seront implémentées devront être en mesure de gérer deux problématiques majeures qui sont inhérentes à notre base de données.

La première est relative au déséquilibre des classes de notre variable cible avec une répartition de l'ordre de 86% - 14% pour le changement de véhicule. Comme nous le verrons plus tard, une telle distribution de la variable à prédire peut conduire les algorithmes d'apprentissage supervisé au sur-apprentissage, affectant de manière significative la qualité des modèles et leur pouvoir prédictif.

Le second défi tient dans la prise en compte des variables catégorielles dans la modélisation. En effet, elles sont nombreuses dans notre base de données et le nombre de classes pour chacune d'entre elles peut être suffisamment conséquent pour rendre leur encodage difficile.

Pratiques de référence en actuariat, les régressions linéaires sont largement utilisées puisqu'elles donnent la possibilité, entre autres, de prédire et expliquer les valeurs prises par une variable cible qualitative, souvent binaire (fraudes, résiliations, nouvelles souscriptions, etc...). Cependant, puisqu'elles ne permettent pas de répondre facilement aux deux problématiques que nous venons d'énoncer, nous ne les retiendrons pas pour modéliser le changement de véhicule.

Pour y répondre, ce sont les méthodes de classification qui nous ont donc semblé les plus appropriées. Ces méthodes de traitement plus souples ont été largement développées avec l'explosion des volumes de données et de leur complexité. Contrairement aux méthodes de régressions linéaires, elles ne cherchent pas à expliquer l'influence de chaque variable explicative sur la variable cible mais plutôt le comportement de groupes homogènes.

Portées par leur caractère non paramétrique, qui leur confère généralement une bonne qualité prédictive, ces méthodes d'apprentissage se fondent sur la constitution d'arbres de décision. On peut distinguer deux types d'arbres : les arbres dits de régression et les arbres dits de classification. Les premiers cherchent à expliquer le niveau d'une réponse quantitative tandis que les seconds cherchent à prédire une variable qualitative. Notre étude cherchant à expliquer la présence ou l'absence d'un changement de véhicule, nous utiliserons donc les arbres de classification.

Parmi l'ensemble des modèles appartenant à cette catégorie d'apprentissage, deux ont retenu notre attention.

Le premier modèle mis en place sera la forêt aléatoire (Random Forest), algorithme fonctionnant bien dans une variété de situations et de cas d'utilisation. L'approche générale consiste à ajuster un certain nombre d'arbres de décisions sur des sous-échantillons puis d'en faire la moyenne. Dans un certain sens, c'est un "vote à la majorité". Il permet d'augmenter la précision et d'éviter le sur-ajustement. L'avantage de ce modèle est sa facilité d'utilisation, sa polyvalence et son bon fonctionnement en présence de valeurs aberrantes ou de données déséquilibrées.

Le second modèle utilisé est l'un des algorithmes les plus récents : le CatBoost. Créé par la société russe Yandex en 2017, cet algorithme particulièrement puissant, robuste et précis se distingue par sa gestion des variables catégorielles. La forme particulière d'encodage cible qu'il implémente lui permet d'accélérer le temps nécessaire pour entraîner et prédire les données mais améliore également la précision. Les avantages de cette méthode sont sa rapidité de calcul et de prédiction, sa gestion optimisée des variables catégorielles et la gestion de ses paramètres permettant de gérer les classes déséquilibrées et d'éviter le sur-apprentissage. En outre, si la lisibilité des résultats des méthodes de classification est parfois remise en cause avec le phénomène de "boite noire", le CatBoost dispose d'une bibliothèque permettant une interprétation visuelle de l'importance des différentes variables.

Un autre algorithme particulièrement utilisé, notamment dans les compétitions, est le XGBoost. Ce puissant algorithme d'apprentissage automatique s'appuyant sur une bibliothèque d'amplification du gradient permet d'obtenir des résultats rapides et précis. Cependant, nous n'aurons pas recours à ce modèle pour notre étude car c'est un algorithme aussi complexe que le CatBoost, dont la lisibilité des résultats peut être difficile et qui ne permet pas une gestion optimisée des variables catégorielles contrairement au modèle précédent.

Les éléments théoriques des deux modèles utilisés sont présentés dans le chapitre 2 de ce mémoire.

Indépendamment du modèle utilisé, le processus de modélisation consistera dans premier temps à sélectionner les variables explicatives pour expliquer le changement de véhicule. Par la suite, l'analyse de la qualité des différentes modélisations et l'optimisation des hyperparamètres nous permettront de maximiser le pouvoir prédictif des modèles. Enfin, l'interprétation des résultats nous permettra d'approfondir notre connaissance sur les facteurs influençant la prise de décision.

#### **4.1 La validation croisée pour s'assurer de la stabilité des classifications**

L'évaluation des modèles de classification, toute comme celle des modèles de régression, est un défi puisqu'il n'y a aucun moyen de savoir si un modèle est adapté pour prédire un comportement avant qu'il ne soit utilisé sur un nouveau jeu de données.

Si la grande force des modèles de classification réside dans leur forme non paramétrique qui leur alloue une plus grande flexibilité et leur permet de mieux capter le phénomène étudié, ils accusent en revanche d'un fort risque de sur-apprentissage. Résultat direct d'un mauvais dimensionnement de la structure du modèle, les modèles n'expliquent plus que le phénomène étudié mais poussent la compréhension jusqu'au bruit. Le modèle devient alors moins performant lors de ses prévisions car il n'est plus en mesure de gérer l'apparition de nouveaux scénarios.

L'estimation des performances du modèle sur les données déjà disponibles est donc un véritable enjeu pour déterminer si le modèle est bien généralisé, sous-ajusté ou sur-ajusté.

Pour ce faire, il est nécessaire de valider le modèle. Le processus de validation consiste à décider si les résultats numériques qui quantifient les relations hypothétiques entre les variables sont acceptables pour décrire les données.

En apprentissage automatique, la validation croisée (*cross validation* en anglais) est une méthode d'estimation de la fiabilité d'un modèle fondée sur une technique d'échantillonnage. De manière générale, il s'agit de construire notre modèle sur une partie des observations puis de vérifier que l'on observe la même qualité sur le reste des observations.

Il existe de nombreuses variantes de validation mais on peut distinguer principalement la validation non-croisée et la validation croisée à  $k$ -blocs (*k-folds validation* en anglais).

#### 4.1.1 Validation non croisée : échantillon d'apprentissage et échantillon test

Tout d'abord, la validation non-croisée consiste à diviser la base d'étude en deux sous-ensembles :

- Le premier, dit **base d'apprentissage** ou base de modélisation (communément supérieur à 60% de l'échantillon initial), permet de définir la structure du modèle et le calibrer.
- Le second, dit **base de validation**, est utilisé une fois la modélisation figée pour valider le modèle et éventuellement comparer différents modèles entre eux.

Avec cette approche, le modèle est bâti sur l'échantillon d'apprentissage puis validé sur l'échantillon test avec le(s) score(s) de performance de notre choix.

Disposant d'une base de données suffisamment large, nous répartissons cette dernière à hauteur de 70% pour la base de modélisation tandis que les 30% restants seront exploités pour la validation.

Néanmoins, il est important de tenir compte des spécificités de notre base de données dont la variable à expliquer présente une distribution des classes déséquilibrée. Pour éviter que la performance d'apprentissage et de validation ne soit biaisée par une répartition changeante des classes d'un ensemble d'apprentissage ou de validation à un autre, il est nécessaire d'utiliser la version stratifiée de la validation non croisée. Cette méthode veille à diviser les données de manière aléatoire tout en conservant la même distribution de classes déséquilibrée pour chaque sous-ensemble.

#### 4.1.2 Validation croisée à $k$ -blocs

Comme nous l'avons mentionné, la principale faiblesse des arbres de décision se situe dans le risque de sur-apprentissage. Pour y remédier, des procédés de Bagging peuvent être utilisés. Ils consistent à construire plusieurs arbres sur des échantillons différents d'une même base, tirés aléatoirement, avant de les agréger.

La validation croisée à  $k$ -blocs reprend cette méthodologie et permet de s'assurer que toutes les observations de l'ensemble de données initial aient une chance d'apparaître dans l'ensemble d'apprentissage et dans l'ensemble de validation.

Pour commencer, il s'agit de séparer de manière aléatoire la base de données initiale en  $k$  échantillons. Le choix du nombre de sous-échantillons est important puisqu'un nombre trop élevé pourrait conduire à un modèle certes moins biaisé mais dont la forte variance pourrait conduire au sur-apprentissage. A l'inverse, un nombre insuffisant reviendrait à implémenter la méthode évoquée.

Par la suite, on sélectionne tour à tour un des  $k$  échantillons qui constituera l'échantillon de validation ; le reste, c'est-à-dire l'union des  $k - 1$  échantillons, constituera l'ensemble d'apprentissage.

Le procédé de validation du modèle est sensiblement le même que la validation non croisée - le modèle étant calibré sur l'ensemble d'apprentissage puis évalué sur l'ensemble de validation - à l'exception près que l'apprentissage puis la validation se font de manière itérative et indépendante sur les  $k$  échantillons de l'ensemble d'apprentissage. A l'issue de la procédure, chaque observation a ainsi servi une fois dans un jeu de test et  $k - 1$  fois dans un jeu d'entraînement.

La performance de la modélisation peut alors être évaluée de deux manière :

- En évaluant les prédictions faites sur l'ensemble des données ; une prédiction ayant été effectuée pour chaque point de l'ensemble du jeu de données
- En moyennant les performances obtenues sur les  $k$  échantillons. L'analyse de l'écart-type peut également être intéressante pour estimer le biais et quantifier la variation des performances sur chacun des échantillons.

De la même façon que la validation non croisée, il sera nécessaire pour notre étude d'utiliser la version stratifiée de la validation croisée à  $k$ -blocs pour s'assurer que la répartition des classes de la variable cible soit identique d'un échantillon à l'autre et que cette dernière corresponde à la distribution des classes de l'ensemble complet de données.

Dans notre processus de modélisation, nous pourrons utiliser ces deux méthodes de validation. La seconde méthode étant plus longue à implémenter, nous pourrons l'utiliser au moment de valider définitivement le calibrage des modèles ou pour comparer les modèles finaux obtenus pour le Random Forest et le CatBoost. A l'inverse, la validation non croisée, moins consommatrice en temps de calcul, pourra être utilisée lors du calibrage des modèles qui est un processus traditionnellement long.

## 4.2 Sélection des variables

La sélection des variables est le processus qui consiste à réduire le nombre de variables prédictives (*features* en anglais) lors du développement d'un modèle.

Cette réduction de la dimensionnalité permet à la fois d'optimiser les temps d'apprentissage et de développement des modèles et de faciliter leur intégration dans les systèmes opérationnels puisque la récupération des données nécessaires à leur mise en place en sera plus aisée.

Par ailleurs, réduire la dimension de la base d'étude permet d'éviter le risque de sur-apprentissage. En effet, à mesure que le nombre de variables explicatives augmente, les modèles deviennent plus complexes pour tenter de capter le maximum de profils et d'avoir une prédiction précise. Cette complexification entraîne un risque de sur-apprentissage qui dégrade la performance des modèles. Un nombre limité de variables explicatives tend donc à favoriser la robustesse des résultats.

Enfin, l'interprétation des résultats sera d'autant plus aisée que le nombre de facteurs explicatifs est restreint et que la structure du modèle est simple. En effet, l'augmentation de la dimension a tendance à rendre les données éparses et éloignées et donc à fausser les méthodes d'analyse de données traditionnelles. Ce manque de densité des données dans l'espace impacte les méthodes nécessitant le principe de significativité statistique.

A l'heure du Big Data, les flux de données disponibles ont explosé et le concept introduit par Bellman en 1961 [2], le fléau de la dimension, devient un défi majeur à relever au moment de répondre à des problématiques par la mise en place d'algorithmes d'apprentissage automatique. Ainsi, de nombreuses techniques de réduction de dimension ont été proposées afin de représenter les données dans un espace adéquat et plus facilement interprétable par les méthodes d'analyse classiques.

Ces méthodes consistent à évaluer la relation entre chaque variable d'entrée et la variable cible à l'aide de métriques et à sélectionner celles ayant la relation la plus forte avec la variable à expliquer. Elles reposent principalement sur un algorithme de recherche et un critère d'évaluation utile à l'analyse de la pertinence des sous-ensembles potentiels de variables.

Ces méthodes peuvent s'envisager en termes de méthodes supervisées et non supervisées. La différence tient dans la sélection ou non des caractéristiques en fonction de la variable cible. Les techniques de sélection des variables non supervisées ignorent la variable cible - telles que les méthodes éliminant les variables redondantes en utilisant la corrélation -, tandis que les techniques supervisées en tiennent compte - telles que les méthodes éliminant les variables non pertinentes.

Par ailleurs, on distingue les méthodes par enveloppement (*wrapper*) ou par filtrage (*filter*). Ces méthodes sont presque toujours supervisées et sont évaluées sur la base de la performance d'un modèle. Les méthodes de sélection des caractéristiques enveloppantes créent de nombreux modèles avec différents sous-ensembles de variables explicatives puis sélectionnent les caractéristiques qui donnent le modèle le plus performant selon une mesure d'évaluation prédéfinie. Les méthodes de sélection filtrante utilisent quant à elles des techniques statistiques pour évaluer la relation entre chaque variable d'entrée et la variable cible, et les scores de notation sont utilisés comme base pour choisir (et donc filtrer) les variables qui seront utilisées dans le modèle.

Enfin, il existe certains algorithmes d'apprentissage automatique qui effectuent la sélection des caractéristiques automatiquement dans le cadre de l'apprentissage du modèle. Avec ces méthodes de sélection intrinsèques, le modèle n'inclura que les prédicteurs qui aident à maximiser la précision. Cette sélection est présente pour les modèles de régressions pénalisées comme LASSO et RIDGE mais également dans les ensembles d'arbres de décision.

Dans le but de sélectionner les variables d'entrée de notre modélisation, nous mettrons en application certaines des méthodes présentées ci-dessus.

Avant toute chose, il est possible d'éliminer certaines variables de notre base sans analyse préalable. Ces variables sont relatives à l'identification du client dans les systèmes d'information (numéro de contrat, numéro client et code du véhicule) ou des variables catégorielles dont le nombre de modalités est trop important (code commune INSEE, modèle et marque du véhicule, etc. . .)

En outre, dans le chapitre précédent, nous avons construit plusieurs variables catégorielles à partir de variables quantitatives afin d'obtenir une segmentation par facteurs de risques qui facilitait l'analyse



descriptive de notre base d'étude. Le propre des arbres de classification étant de déterminer des sous-groupes adoptant des comportements homogènes, nous retiendrons les variables continues plutôt que les variables segmentées pour la modélisation.

Pour aborder la modélisation, le choix a donc été fait de conserver l'information le plus large possible afin de capter au mieux les signaux clients pouvant influencer le changement de véhicule. Certaines des méthodes évoquées nous permettront de pré-sélectionner un certain nombre de variables explicatives qui serviront de périmètre d'entrée aux modèles testés. Par la suite, l'analyse avancée de la qualité de ces modèles en fonction des différentes variables sélectionnées nous permettra de restreindre une fois de plus les facteurs explicatifs du changement de véhicule.

#### 4.2.1 Analyse des dépendances

Les modèles d'apprentissage automatique (*machine learning* en anglais) sont sensibles aux valeurs aberrantes ou manquantes et à la colinéarité entre les facteurs explicatifs. Les deux premiers points ont été gérés au moment du traitement de la base de données. Le dernier fera l'objet de cette section.

En effet, pour assurer la qualité du modèle, les variables explicatives ne doivent pas être fortement corrélées. En effet, la redondance des informations pourrait rendre les résultats instables et compliquer leur interprétation, tout en augmentant inutilement la dimensionnalité de la base de données.

Les variables constituant notre base d'étude pouvant être quantitatives ou qualitatives et l'étude des dépendances différant selon le type de variables, il sera nécessaire de mettre en place différentes méthodes d'analyse statistiques.

Intéressons-nous dans un premier temps aux dépendances entre les variables quantitatives. Il existe plusieurs méthodes statistiques destinées à quantifier et tester la liaison entre deux variables quantitatives : on parle d'analyse de corrélation [15]. L'une d'entre elles est la matrice de corrélation qui permet une représentation visuelle des interdépendances.

Plusieurs groupes de variables semblent fortement liées entre elles, marqués par des zones plus claires ou plus foncées sur la matrice de corrélation présentée dans la figure 4.1.

Tout d'abord, l'âge du conducteur et l'ancienneté du permis sont extrêmement corrélés (0.95). L'obtention du permis étant conditionnée par la majorité et la plupart des individus l'obtenant au même stade de leur vie, l'information apportée par cette variable ne se distingue pas de celle apportée par l'âge du conducteur.

Ensuite, plusieurs caractéristiques du bien assuré semblent interdépendantes. Ainsi, les variables décrivant la puissance du véhicule et sa dangerosité sont fortement liées, telles que la puissance administrative, la puissance mécanique, le nombre de cylindres, la vitesse maximale ou encore le groupe SRA. De la même façon, le dernier tarif connu du véhicule à neuf et le groupe SRA sont corrélés puisque la puissance d'un véhicule détermine également son prix. En sélectionnant les variables qui nous semblent d'expérience significatives pour expliquer le changement de véhicule et en nous autorisant un seuil maximal de corrélation de 0.75, les variables que nous retiendrons pour caractériser un véhicule sont le nombre de places, le dernier tarif connu du véhicule à neuf, la puissance administrative et le groupe SRA. Ces dernières feront par la suite l'objet d'autres étapes de sélection des variables, ce qui nous permettra de restreindre davantage la dimensionnalité.

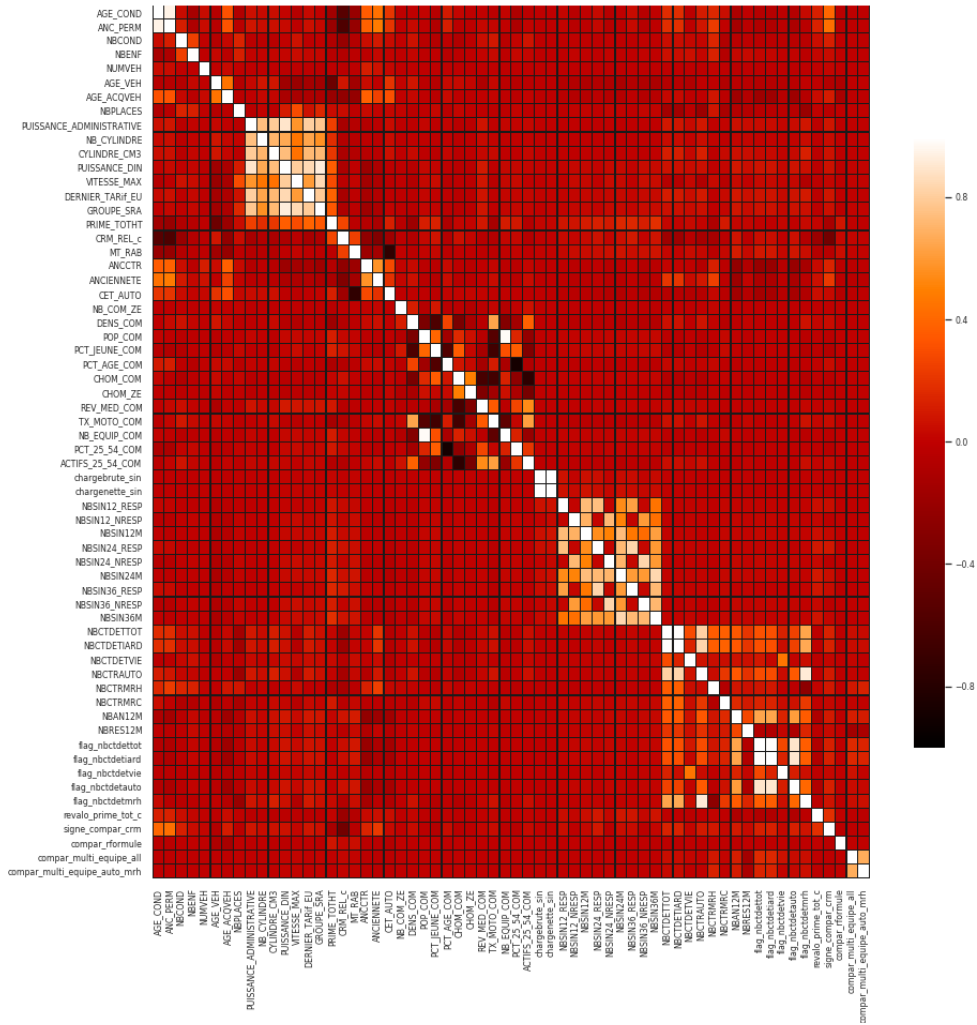


FIGURE 4.1 – Matrice de corrélation des variables quantitatives avant sélection des variables

On remarque ensuite que certaines données décrivant l’environnement géographique de l’individu apportent la même information. Issues des tables de l’INSEE, ces données sont définies au niveau de la commune d’habitation du client et il est donc tout à fait logique que les informations se recoupent. Le choix des variables à conserver s’appuie sur la pertinence des données pour expliquer le changement de véhicule et sur la sélection de variables ayant le nombre minimal de liaisons avec d’autres données. Ainsi, au niveau de la commune, nous conservons le nombre d’équipements, la densité, le taux de chômage et le revenu médian.

La quatrième zone de corrélation que l’on observe, relative au nombre de sinistres, provient de la méthodologie de constitution de notre base de données. En effet, dans les différents systèmes d’informations, nous avons récupéré le nombre et le montant des sinistres pour différentes périodes, à savoir les derniers 12, 24 et 36 mois. Les historiques plus anciens contiennent ainsi nécessairement les informations des historiques plus récents. L’intuition nous pousse à penser que l’accumulation de sinistres peut être un facteur favorisant le changement de véhicule, c’est pourquoi nous faisons le choix de conserver l’historique le plus ancien.

Cette décision arbitraire pourra être challengée au moment de la modélisation en remplaçant cet historique et en observant l'évolution de la qualité du modèle. Outre la date de survenance des sinistres, nous supprimons la distinction entre sinistres responsables et non responsables dont l'information est déjà contenue dans le niveau du coefficient bonus-malus et sa variation.

Enfin, le dernier groupe de variables présentant une forte corrélation sont les variables relatives au nombre de contrats détenus et les variations de leur détention en portefeuille. De la même façon que pour les sinistres, la décomposition des variables est à l'origine de cette interdépendance. En effet, le nombre de contrats détenus au total regroupe l'information du nombre de contrats détenus en IARD ou en vie, tout comme le nombre de contrats détenus au total ou en IARD regroupe l'information du nombre de contrats détenus en automobile, MRH ou MRC, etc... La variation de la détention en portefeuille suit la même logique. Pour notre étude, nous ne retiendrons que les variables relatives au niveau de détention en automobile ou le niveau de détention globale, le choix entre les deux étant à effectuer au moment de la modélisation.

Après suppression des différentes variables, la matrice de corrélation arbore désormais la forme présentée dans la figure 4.2. Outre le nombre de caractéristiques qui a été considérablement réduit, on constate également que les variables conservées affichent des niveaux de colinéarité plus faibles.

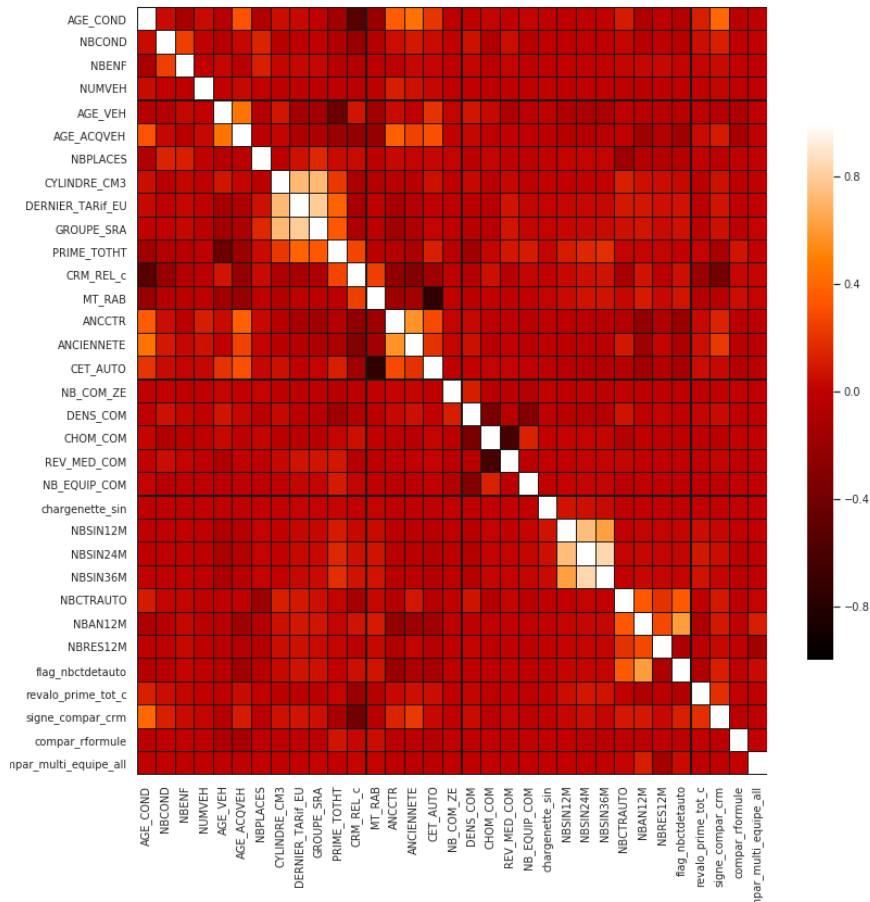


FIGURE 4.2 – Matrice de corrélation des variables quantitatives après sélection des variables

Intéressons-nous désormais aux dépendances entre les variables catégorielles [16]. Nous avons donc besoin d'une méthode statistique permettant de mesurer l'association entre deux caractéristiques catégorielles : le V de Cramer. Ce dernier est basé sur une variation nominale du test du Xhi-deux.

Le principe de ce test est de comparer les effectifs réels des croisements des modalités entre deux variables catégorielles avec les effectifs théoriques, obtenus si les deux variables étaient indépendantes. Mesurant la force de la liaison entre deux variables qualitatives, il a l'avantage de compenser à la fois l'effet du nombre de modalités et du nombre d'observations mais également d'être symétrique et donc insensible à la permutation des deux variables étudiées. Le résultat est compris dans l'intervalle  $[0; 1]$  où 0 signifie une absence d'association, c'est-à-dire l'indépendance des deux variables, et 1 une association complète, c'est-à-dire la colinéarité. Contrairement à la corrélation, il n'y a pas de valeurs négatives car une association négative ne peut exister.

Afin de détecter les interactions, nous avons décidé de croiser toutes les variables catégorielles entre elles, en limitant la dimension à 2, puis de calculer le V de Cramer. La représentation visuelle de la matrice des dépendances est présentée dans la figure 4.3.

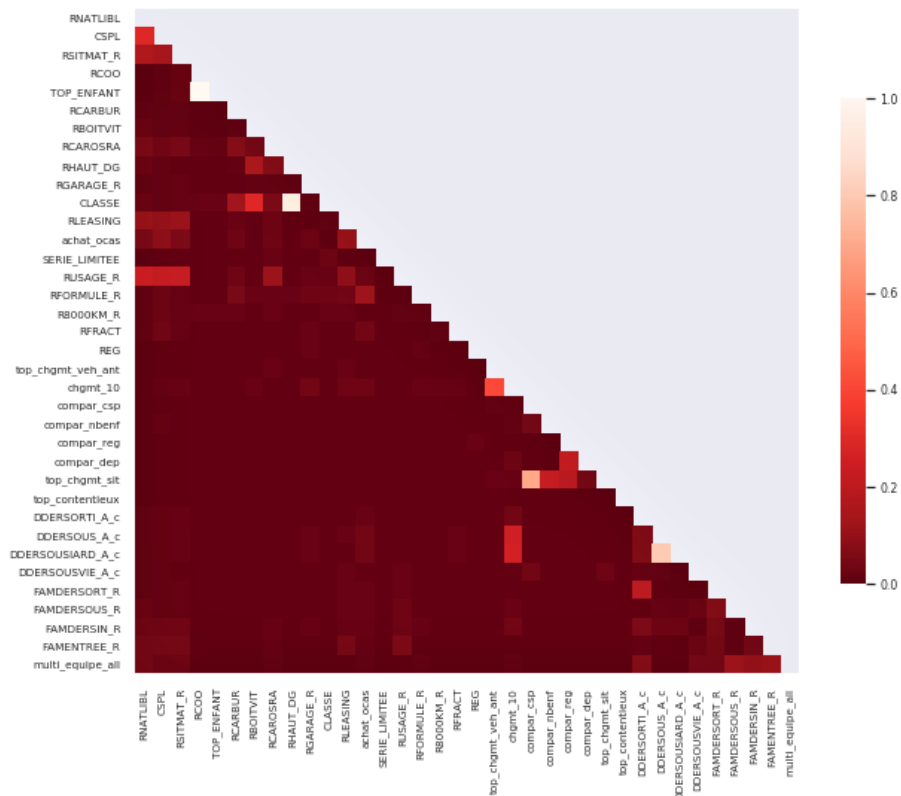


FIGURE 4.3 – V de Cramer pour les variables catégorielles avant sélection des variables

De manière générale, on remarque que les variables qualitatives de notre base d'étude ne sont pas dépendantes les unes des autres.

Comme nous l'avons expliqué, la présence de facteurs fortement colinéaires empêchent la convergence des modèles. C'est pourquoi, nous devons supprimer quelques variables supplémentaires.

Tout d'abord, on observe que la classe du SRA du véhicule et la variable binaire identifiant si le véhicule est haut de gamme ont une forte association. L'information apportée par la classe SRA étant plus détaillée, nous gardons ce niveau d'information.

Par ailleurs, la présence d'un enfant conducteur sur le contrat et la présence d'un conducteur occasionnel sont également interdépendantes. La présence d'un enfant conducteur étant contenue dans l'information sur la présence d'un conducteur occasionnel, il est plus judicieux de garder le niveau de détail le plus large.

Enfin, de la même manière que le nombre de contrats détenus par famille de produits, les dates de dernière souscription en assurance vie ou non-vie sont interconnectées avec celles toutes familles de produits confondues. Comme auparavant, nous conservons le niveau le plus englobant car ce dernier nous permettra de capter l'ensemble des mouvements sur le portefeuille.

Le V de Cramer obtenu après la suppression des variables est présenté en annexe dans la figure A.1.

Au moyen de la matrice de corrélation et du V de Cramer, nous avons pu analyser de manière séparée la dépendance entre les variables quantitatives d'une part et entre les variables catégorielles d'autre part, nous permettant de diminuer de manière significative la dimension de notre base de données. Mais qu'en est-il de l'interaction entre une variable catégorielle et une variable quantitative ?

Pour mesurer l'intensité de cette relation, on peut calculer un paramètre appelé rapport de corrélation (souvent indiqué par la lettre  $\eta$ ). Il se définit comme la variance pondérée de la moyenne de chaque catégorie, divisée par la variance de tous les échantillons. Autrement dit, étant donné un nombre continu, il permet d'évaluer dans quelle mesure il est possible de déterminer à quelle catégorie il appartient. Le résultat est, ici aussi, compris dans l'intervalle  $[0,1]$ .

L'analyse du rapport de corrélation pour chacune des combinaisons entre les variables continues et catégorielles, non présentée ici, permet de conclure que les variables sont suffisamment indépendantes pour servir de base à notre modélisation.

#### 4.2.2 Régressions linéaires pénalisées

Dans l'optique d'affiner la pré-sélection des variables qui serviront de point de départ à la modélisation, nous utiliserons des régressions pénalisées telles que RIDGE [7] et LASSO [18]. Ces deux techniques de régularisation ont pour but de limiter les problèmes d'instabilité des prédictions des régressions linéaires. Elles vont permettre de "distordre" l'espace des solutions afin d'empêcher l'apparition de valeurs trop élevées : on parle de rétrécissement. Pour ce faire, on introduit un terme de pénalité à la fonction de coût du problème de régression linéaire.

Si la qualité des régressions obtenue n'est pas l'objet de notre intérêt, elles permettent de faire une première sélection sur les variables qui pourraient avoir un impact, négatif ou positif, sur la probabilité de changer de véhicule. En effet, ces types d'algorithmes supervisés permettent d'identifier les variables qui sont fortement associées avec la variable cible.

La principale différence entre ces deux régressions se fait au niveau de la parcimonie, c'est-à-dire la tendance à sélectionner des modèles moins complexes. Ainsi, plus la pénalisation augmente via le coefficient  $\lambda$ , plus la pénalisation  $L^2$  (RIDGE) aura tendance à réduire les coefficients sans les annuler.

A l'inverse, une pénalisation  $L^1$  (LASSO) annulera beaucoup plus de coefficients, sélectionnant des modèles plus parcimonieux.

L'utilisation de ces deux méthodes de régressions linéaires nécessite un traitement préalable sur la base de données. L'encodage des variables catégorielles est ainsi requis et nous conduit à transformer chacune des classes sous le format binaire (encodage one-hot). La base de données obtenue comporte alors plus de 250 variables, justifiant une fois de plus le recours à des méthodes de classification qui autorisent un encodage de ces variables plus adapté.

A partir des variables retenues par l'analyse des dépendances, on met tout d'abord en place une régression LASSO. Après avoir sélectionné les variables ayant un impact sur la variable cible, l'algorithme force les coefficients vers 0 lors du processus de rétrécissement (*shrinkage*). Cela permet de rendre les modèles moins sensibles à l'ensemble de données et d'atténuer les limites de la cognition humaine puisque moins de variables sont sélectionnées.

Ce sont justement ces variables dont les coefficients sont significativement différents de 0 que nous souhaitons observer. Nous classons leurs coefficients par ordre de grandeur puis analysons les effets des variables pour lesquels les coefficients sont les plus élevés (négatifs ou positifs).

Ainsi, parmi l'ensemble des variables quantitatives et des différentes classes des variables qualitatives, 55 semblent significatives pour expliquer le changement de véhicule selon la régression LASSO. De manière exhaustive, il semble être favorisé par l'âge du véhicule, la souscription de la formule minimale, l'absence de restrictions kilométriques, l'option de leasing, la présence d'un conducteur occasionnel, etc... A l'inverse, les facteurs incitant un individu à conserver son véhicule sont l'âge du conducteur et la durée de détention du véhicule, l'ancienneté du client en portefeuille, la densité de la commune, le nombre de conducteurs assurés, etc...

De la même façon, la régression RIDGE permet d'avoir un aperçu des variables influençant le changement de véhicule. A la différence de la régression LASSO, elle ne force pas les coefficients vers 0. De cette façon, tous les coefficients obtenus ont une valeur négative ou positive. L'analyse des variables significatives se fait donc en classant les coefficients obtenus par ordre de grandeur puis en analysant ceux avec les valeurs les plus fortes (négatives ou positives).

Les facteurs explicatifs retenus sont sensiblement similaires à ceux de la régression LASSO. Seul l'ordre d'importance des variables est amené à changer, sans modification importante à notifier.

L'utilisation des méthodes de régressions pénalisées, couplée à l'analyse descriptive des facteurs influençant le changement de véhicule que nous avons développé dans le chapitre précédent, nous donnent une première idée sur les variables pouvant jouer sur la prédiction du changement de véhicule. Ces analyses préalables nous permettent de réduire la dimension de la base de données et ainsi de diminuer le nombre de variables que nous allons soumettre lors des premières modélisations.

#### **4.2.3 Sélection au moyen de l'importance des variables**

Les différentes méthodes de sélection des variables ont restreint la base à une soixantaine de variables. Sur cette base, nous appliquons la méthode de validation non croisée afin de la séparer en un échantillon d'apprentissage et un échantillon de validation.

Sur l'échantillon d'apprentissage, nous allons modéliser le changement de véhicule au moyen des deux méthodes que nous avons choisies, à savoir le Random Forest et le CatBoost. Pour chacun de ces modèles, nous procéderons de la même manière pour parvenir à sélectionner le jeu final de variables explicatives.

Dans les méthodes de classification, il est courant d'observer l'importance des variables dans le modèle (*feature importance* en anglais) afin de comprendre quelles variables sont les plus discriminantes pour prédire le comportement étudié. L'importance des facteurs est une mesure du gain qu'ils apportent, et dans le cas des méthodes de classification, elle se calcule à partir du gain apporté sur l'indice de Gini. Le gain moyen est ensuite calculé en prenant en compte le gain apporté sur l'ensemble des segmentations effectuées par l'arbre sur cette variable. C'est cette mesure que nous pouvons ensuite utiliser pour comparer les différentes variables.

En prenant l'exemple de la classification au moyen du CatBoost sur la base réduite à 62 variables explicatives, nous obtenons les poids relatifs suivants :

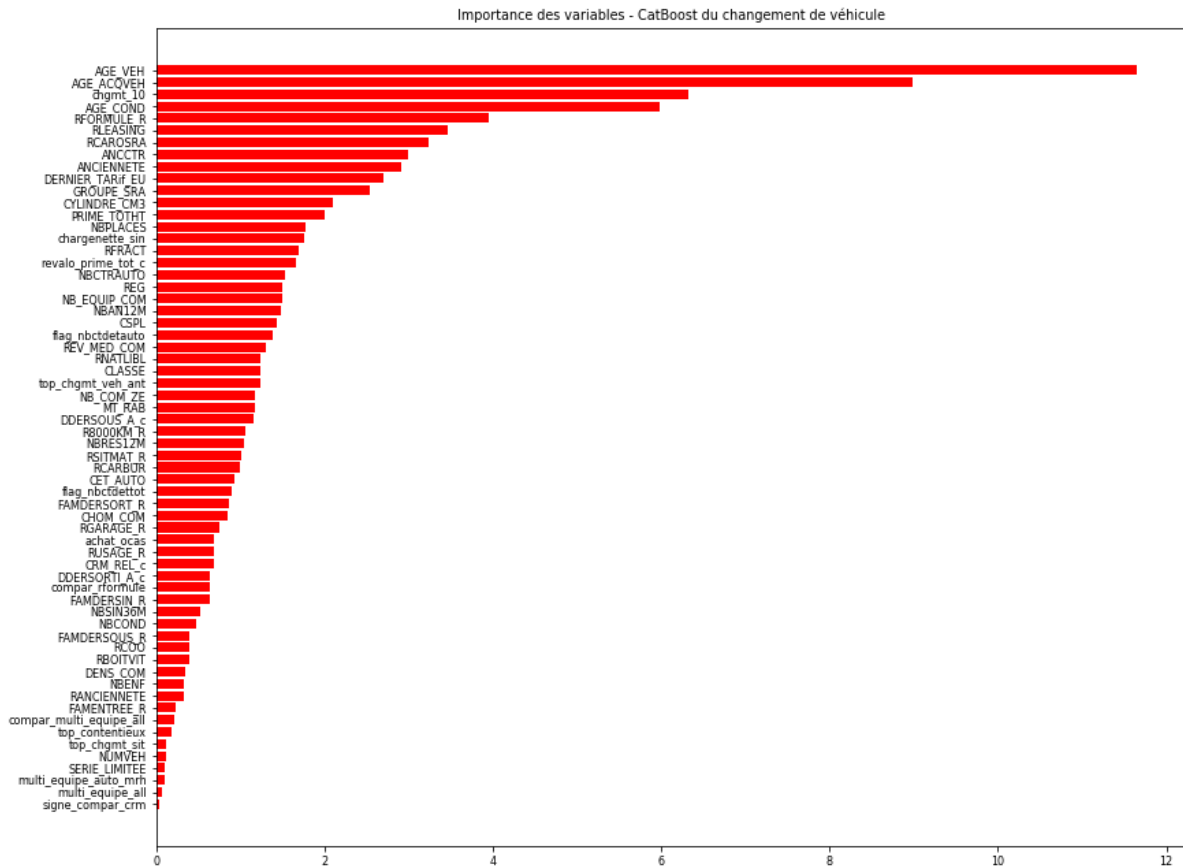


FIGURE 4.4 – Importance des variables dans la modélisation CatBoost du changement de véhicule avant sélection des variables

On constate que plusieurs variables semblent apporter un faible gain au modèle. Au-delà de ce faible apport, conserver ces variables complexifient les arbres construits, et donc les modèles, et rendent d'autant plus difficile leur interprétation.

De manière itérative, on élimine donc successivement les variables contribuant le moins au modèle. Nous utilisons une méthode par sélection descendante (*backward*) c'est-à-dire que nous supprimons successivement les variables les moins explicatives jusqu'à ce que le gain apporté par chacune des variables soit jugé satisfaisant. Cette manière de procéder permet d'éviter la redondance des informations mais cette dernière est sensible au choix du point de départ et aux données.

La conservation d'une variable est donc décidée en regardant sa contribution au modèle et en analysant si sa suppression n'altère pas la qualité du modèle – ou inversement si son ajout ne permet pas d'augmenter de manière significative sa qualité (méthode de sélection ascendante (*forward*)) alors qu'il réduit le nombre de degrés de liberté.

Cependant, pour une vision opérationnelle, ces critères statistiques ne sont pas suffisants. Parmi les variables non sélectionnées, il faut tenter de comprendre celles dont les effets paraissent peu captés alors que l'information qu'elles apportent semblait au premier abord intéressante. Il est en effet possible que la construction de certaines variables ne soit pas adaptée à notre problématique et empêche l'information d'être exploitée correctement.

Par exemple, lors des premières modélisations que nous avons effectuées, la variable binaire identifiant si un changement de véhicule a eu lieu l'année précédente n'avait pas un impact significatif sur la probabilité de changer de véhicule l'année suivante, positivement ou négativement. Convaincus que le comportement en termes de changement de véhicule observé dans le passé pouvait avoir un impact sur les prédictions, nous avons alors construit une variable retraçant cet historique sur les dix dernières années, variable qui s'est avérée très significative.

Cette méthode de sélection descendante des variables nous mène ainsi à sélectionner uniquement 27 variables pour chacune des modélisations, nombre acceptable lorsque les modèles devront être intégrés dans les processus opérationnels. Les variables retenues pour modéliser le changement de véhicule sont présentées en annexe dans la table A.4.

En définitive, l'ensemble des méthodes développées dans cette section nous ont permis de réduire considérablement la dimension de notre base de données, passant de près de 150 variables explicatives à une trentaine. Cette sélection nous permet ainsi d'optimiser les temps de calcul et la qualité des modélisations mises en place grâce à la suppression de la colinéarité entre les variables et l'élimination des variables peu influentes, mais également de faciliter l'intégration des scores dans les process.

Si la sélection des variables est une étape importante dans la construction de modèles prédictifs et permet d'améliorer leurs performances, l'ajustement des hyperparamètres est également une étape nécessaire pour adapter le mieux possible les différentes méthodes de classification à nos données.

### 4.3 Optimisation des hyperparamètres des modèles de classification

L'un des avantages des modèles d'apprentissage automatique, et qui justifient leur popularité, est qu'ils sont facilement adaptables à toutes les situations. La contrepartie de cette flexibilité sont les difficultés qui peuvent être rencontrées lors du paramétrage de certains modèles.

Les paramètres des modèles de *machine learning* qui peuvent être modifiés sont appelés les hyperparamètres. Le nombre de paramètres à optimiser varie d'une méthode à une autre et ce dernier peut vite devenir conséquent pour les méthodes avec des arbres de décision.



Afin de tirer les meilleurs résultats possibles d'une méthode d'apprentissage, il est donc nécessaire de trouver la meilleure combinaison possible de ces hyperparamètres. Une recherche manuelle présente rapidement des limites opérationnelles. Par ailleurs, procéder de la sorte conduirait à optimiser indépendamment les différents paramètres d'un modèle, laissant de côté l'interaction entre les hyperparamètres pour un ensemble de données.

Heureusement, plusieurs méthodes d'optimisation vont nous permettre de tester une série de paramètres et de comparer les performances de chaque combinaison pour en tirer le meilleur paramétrage.

La première méthode est le GridSearch qui, comme son nom l'indique, va considérer le problème d'optimisation comme un problème de recherche dans une grille. Pour chaque hyperparamètre, nous allons fixer un ensemble de valeurs possibles. Par la suite, pour chaque combinaison, le modèle sera entraîné sur l'échantillon d'apprentissage puis les résultats des performances du modèle sur l'échantillon de validation seront conservés en mémoire. Il suffira ensuite de sélectionner la combinaison d'hyperparamètres optimisant les métriques de scores que nous aurons défini.

Si cette méthode est très efficace lorsque le nombre d'hyperparamètres à optimiser est restreint, elle sera très vite limitée en termes de capacités de calculs si ce nombre devient plus important.

Une des options pour pallier cette limite est d'utiliser la méthode RandomSearch. Le système de grille est conservé mais la manière de sélectionner les combinaisons de paramètres diffère. En effet, plutôt que de fixer un ensemble de valeurs possibles pour un hyperparamètre, on définit un intervalle de valeurs et on lui associe une distribution de probabilités. On entraîne ensuite le modèle sur la base d'apprentissage et on conserve les résultats des performances sur la base de validation.

L'avantage de cette technique, comparée à la précédente, est qu'elle permet de prendre en considération des plages de valeurs qui n'auraient pas été explorées par le GridSearch et donc de tester un nombre plus important de valeurs.

Pour notre étude et pour chacune des deux méthodes de classification retenues, nous utiliserons la méthode GridSearch, couplée à la validation non croisée, pour déterminer dans un premier temps les ordres de grandeur de chacun des hyperparamètres à optimiser. Les métriques de scores sur lesquelles nous nous sommes basées pour élire les meilleures combinaisons seront présentées dans la partie suivante. Cette première sélection nous permettra par la suite d'utiliser la seconde méthode dont les temps de calcul seront nettement accélérés grâce à la réduction des intervalles de valeurs possibles. Le choix final de la meilleure combinaison d'hyperparamètres pour chacun des modèles sera ensuite validé au moyen de la validation croisée par  $k$ -blocs afin de s'assurer que les performances obtenues soient stables quel que soit l'échantillon utilisé pour entraîner le modèle.

La méthodologie pour optimiser les paramètres étant désormais clairement définie, il faut maintenant s'interroger sur la liste de ces hyperparamètres dont la valeur est à modifier.

Tout d'abord, l'un des problèmes majeurs que nous avons soulevé, et qui s'est confirmé au moment des premiers tests de modélisation, est le déséquilibre des classes de la variable cible qui conduit les modèles au sur-apprentissage. Il est donc primordial d'y remédier.

En effet, la plupart des algorithmes d'apprentissage automatique supposent que les données sont réparties de manière égale entre les classes. Si tel n'est pas le cas, le modèle peut être tenté de prédire trop fréquemment la classe majoritaire. En procédant de la sorte, il se focalise sur la minimisation

du risque d'erreur plutôt que de se concentrer sur la classe minoritaire, ce qui conduit souvent à une classification biaisée.

Les algorithmes que nous avons sélectionné permettent d'agir sur l'échantillonnage des observations pour que les différentes classes de la variable cible deviennent équi-représentées. Pour cela, nous pouvons attribuer des poids différents aux classes majoritaire et minoritaire. L'idée sous-jacente est de pénaliser les erreurs de classification commises sur la classe minoritaire en lui attribuant un poids plus élevé et en réduisant dans le même temps celui de la classe majoritaire. Cette différence de poids influencera la classification de la variable cible durant la phase d'apprentissage.

Par ailleurs, l'utilisation du Random Forest et du CatBoost se justifie également en présence de données déséquilibrées puisque leurs algorithmes entraînent des modélisations plus solides et robustes grâce à l'agrégation de plusieurs arbres de classification qui limite la possibilité de sur-apprentissage.

L'enjeu de l'optimisation des paramètres consiste donc à déterminer les meilleurs poids à attribuer à la classe majoritaire et à la classe minoritaire (ici, le changement de véhicule).

Ainsi, pour chaque classe  $j$  de la variable cible, on peut lui attribuer le poids ( $\omega_j$ ) calculé selon la formule suivante (celui-ci pouvant faire l'objet d'ajustements au moment du paramétrage) :

$$\omega_j = \frac{\text{nombre total d'observations}}{\text{nombre de classes} \times \text{nombre d'observations de la classe } j}$$

Parmi les autres paramètres à optimiser pour le Random Forest, on retient :

- **Le nombre d'arbres dans la forêt** (*n estimators*)
- **La profondeur maximale de l'arbre** (*max depth*), pour que l'algorithme ne segmente pas trop une partie de la population au détriment d'autres individus
- **Le nombre minimal d'individus par feuille** (*min sample leaf*), pour éviter que l'arbre creuse trop dans une direction et définisse des classes trop réduites.
- **Le nombre minimal d'individus placés dans un noeud** (*min sample split*) avant que le noeud ne soit divisé

Pour le CatBoost, on retient :

- **Le nombre d'itérations** (*iterations*) détermine le nombre d'arbres
- **La profondeur des arbres** (*depth*) détermine le nombre d'interactions entre les variables que chacun des arbres peuvent prendre en compte.
- **Le taux d'apprentissage** (*learning rate*) : plus sa valeur est faible, plus le nombre d'itérations demandé pour arriver à de bonnes performances est élevé.
- **Le nombre maximal de modalités** (*one hot max size*) d'une variable catégorielle pour que l'encodage one-hot s'applique. Au-delà, l'algorithme utilisera l'encodage cible.

Les hyperparamètres qui ne sont pas optimisés adoptent les valeurs définies par défaut dans les algorithmes d'apprentissage automatique.

## 4.4 Analyse de la qualité des modèles

Un modèle peut être considérée comme de "bonne qualité" s'il décrit correctement les valeurs observées. L'évaluation de ses performances se fait au moyen des méthodes de validation qui consistent à entraîner le modèle sur une base d'apprentissage puis à comparer les résultats obtenus sur la base de validation avec les réelles observations.

Les métriques de score utilisées pour cette analyse diffèrent selon que les méthodes utilisées soient des méthodes de régression ou de classification.

Pour les problèmes de classification, les mesures consistent à comparer l'étiquette de classe attendue, c'est-à-dire la classe réellement observée, à l'étiquette prédite.

Les métriques standards fonctionnent en général bien pour la plupart des problématiques et c'est la raison pour laquelle elles sont largement répandues. Cependant, les utiliser sans avoir fait un raisonnement préliminaire sur les résultats attendus pourrait conduire à mal comprendre les performances du modèle. C'est ce qui peut rendre complexe le choix de ces métriques.

Cette sélection est rendue d'autant plus difficile par la distribution asymétrique des classes de notre variable cible. En effet, il en résulte que certaines des métriques pourtant largement utilisées comme l'AUC ou encore l'exactitude (*accuracy* en anglais) deviennent moins fiables et donnent une fausse idée de la qualité d'apprentissage. Par exemple, dans le cas de classes déséquilibrées où la classe majoritaire représenterait 90% des données, un classificateur qui prédirait systématiquement cette classe aurait une exactitude de 90%, mais il serait dans la pratique inutile. En se penchant sur d'autres métriques, on se rendrait alors compte que ce dernier n'est absolument pas discriminant.

En présence de classes déséquilibrées, il est donc nécessaire de trouver des mesures d'évaluation plus pertinentes. Contrairement aux mesures standards qui traitent toutes les classes comme étant d'importance égale, ces mesures devront traiter les erreurs de classification de la classe minoritaire comme étant plus importantes que celles faites sur la classe majoritaire.

Ainsi, pour évaluer la performance de nos modèles, nous utiliserons des métriques telles que la matrice de confusion, la précision, le rappel, le  $F_\beta$  - score ou encore la courbe de précision-rappel.

Tout d'abord, la matrice de confusion est un tableau croisé permettant de mesurer la qualité d'un système de classification. Chaque ligne correspond à une classe réellement observée tandis que chaque colonne correspond à une classe prédite par le modèle. Toutes les observations qui se situent sur la diagonale de la matrice ont été correctement prédites, tandis que les autres correspondent à des erreurs du modèle. Elle permet donc de donner une représentation visuelle des classes qui sont correctement prédites, celles qui sont erronées et le type d'erreurs commises.

Dans le cas d'une variable cible binaire, la matrice de confusion présente la forme suivante :

		Prédictions	
		0	1
Classes réelles	0	TN	FP
	1	FN	TP

Les matrices de confusion obtenues pour la forêt aléatoire ou le CatBoost sont présentées ci-dessous :

Random Forest			CatBoost		
	0	1		0	1
0	286 998	64 294	0	280 615	70 677
1	35 586	20 451	1	33 058	22 979

TABLE 4.1 – Matrices de confusion des modèles du changement de véhicule

Si l'on compare les résultats des deux modélisations, on constate que la forêt aléatoire semble mieux discerner la classe majoritaire (absence de changement) en commettant moins d'erreurs sur sa prédiction tandis que le CatBoost prédit correctement un nombre plus important de changements de véhicule et commet moins d'erreurs sur cette classe.

L'exploitation de cette matrice de confusion permet de créer plusieurs indicateurs synthétiques :

- **L'exactitude** correspond à l'ensemble des prédictions qui sont correctes :

$$Exactitude = \frac{TP + TN}{TP + TN + FP + FN} \quad (33)$$

- **Le taux d'erreur** correspond à l'ensemble des erreurs de classification. Il peut s'interpréter comme la probabilité de faire une mauvaise prédiction à l'aide de l'arbre de décision :

$$Taux\ d'erreur = \frac{FP + FN}{TP + TN + FP + FN} = 1 - exactitude \quad (34)$$

- **La spécificité** (ou taux de vrais négatifs) correspond à la fraction de prédictions négatives qui sont correctes parmi toutes les classes négatives de l'ensemble de données :

$$Spécificité = \frac{TN}{FP + TN} \quad (35)$$

- **Le rappel** (ou sensibilité ou taux de vrais positifs) correspond à la fraction de prédictions positives qui sont correctes parmi toutes les classes positives de l'ensemble de données :

$$Rappel = \frac{TP}{TP + FN} \quad (36)$$

- **La précision** correspond à la fraction des prédictions positives qui sont correctes parmi l'ensemble des prédictions positives :

$$Précision = \frac{TP}{TP + FP} \quad (37)$$

De manière plus spécifique, la précision permet de caractériser le degré de certitude du résultat lorsque le modèle prédit un changement de véhicule. Optimiser cette métrique permet de minimiser le taux d’erreur parmi les changements de véhicule prédits par le modèle.

Le rappel, quant à lui, indique dans quelle mesure le modèle permet de distinguer cette classe. Optimiser cette métrique revient à tenter de détecter un maximum de changements de véhicule.

Les métriques de scores obtenues pour les différentes modélisations sont présentées dans la table 4.2.

Métriques	Random Forest	CatBoost
Base d’apprentissage		
Exactitude	77.46%	75.18%
Base de validation		
Exactitude	75.47%	74.53%
Précision	24.15%	24.59%
Rappel	36.49%	41.01%
$F_1$ - score	29.05%	30.70%
$F_{0.5}$ - score	25.88%	26.67%

TABLE 4.2 – Métriques de scores des modèles du changement de véhicule

Si l’on compare l’exactitude obtenue sur l’échantillon d’apprentissage et l’échantillon de validation, aucun écart important n’est à notifier, laissant penser que les modèles ne sur-apprennent pas et sont capables de gérer l’apparition de nouveaux scénarios.

Ainsi, sur l’ensemble de validation, 75.5% des prédictions sont correctes pour la forêt aléatoire tandis que 74.5% le sont pour le CatBoost.

Ces scores très proches se retrouvent également au niveau de la précision avec environ 24% des prédictions de changement de véhicule qui sont correctes pour les deux modèles. Autrement dit, cela signifie que seulement un quart des changements de véhicule prédits par les modèles en sont réellement. Le niveau de ce score met en relief la difficulté de prédire un tel comportement qui prend en compte une multitude de facteurs aussi bien économiques, financiers, géographiques, liés au véhicule détenu et aux préférences des individus.

Par ailleurs, le CatBoost performe mieux pour capter le comportement étudié puisqu’il parvient à identifier 41% de l’ensemble des changements de véhicule, contre 36.5% pour le Random Forest.

La précision et le rappel étant deux métriques qui fournissent des informations précises sur la qualité de la classification, notamment sur la classe minoritaire, il peut être intéressant de les combiner en un seul score : le  $F_1$ -score. Il se définit comme la moyenne harmonieuse entre la précision et le rappel :

$$F_1 - score = 2 \times \frac{precision \times rappel}{precision + rappel} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (38)$$

Cette métrique tend à favoriser les classificateurs qui sont à la fois forts en précision et en rappel plutôt que les classifieurs qui mettent l'accent sur l'un au détriment de l'autre. Attribuer un poids identique à la précision et au rappel suppose ainsi qu'il est aussi coûteux de manquer un changement de véhicule (vrai positif) que de déclarer un changement de véhicule qui n'en est pas un (faux positif).

A l'inverse, le  $F_\beta$  - score permet de calculer une moyenne pondérée entre la précision et le rappel. Le facteur réel positif  $\beta$  est choisi de manière à ce que le rappel soit considéré  $\beta$  fois plus important que la précision :

$$F_\beta - score = (1 + \beta^2) \times \frac{precision \times rappel}{(\beta^2 \times precision) + rappel} = \frac{(1 + \beta^2) \times TP}{[(1 + \beta^2) \times TP] + (\beta^2 \times FN) + FP} \quad (39)$$

Dans le cadre de notre problématique, l'étude de la précision apporte des informations qui nous seront plus utiles que celles du rappel. En effet, cette mesure permet indirectement de mesurer l'efficacité du ciblage des politiques de rétention puisqu'elle indique, parmi le nombre de clients qui seraient contactés, le pourcentage des individus qui seraient concernés par les actions déployées puisque réellement susceptibles de changer de véhicule. Le rappel, quant à lui, donne un ordre d'idée sur la capacité du modèle à identifier l'ensemble des changements de véhicule. Le portefeuille automobile de Generali étant important, il ne sera pas possible de mener des actions sur l'ensemble des clients identifiés comme fragiles aussi bien pour des raisons économiques qu'opérationnelles. De ce fait, il est plus pertinent de se focaliser sur l'efficacité du ciblage et donc la précision.

Dans cette optique, nous calculerons le  $F_{0.5}$  - score pour évaluer et comparer les modèles, accordant ainsi 2 fois plus de poids à la précision qu'au rappel.

Pour les deux modèles, les scores obtenus sont très similaires, confirmant les résultats proches en termes de qualité de prédiction sur le changement de véhicule.

Outre la mesure du  $F_{0.5}$  - score, un arbitrage entre précision et rappel peut être fait au moment de la modélisation. Augmenter l'une suppose nécessairement de diminuer l'autre métrique.

Ce choix peut être fait au moment du calibrage du modèle et la recherche de la meilleure combinaison d'hyperparamètres ou au moment de déterminer le seuil dont la valeur permet de déterminer si les probabilités prédites appartiennent à l'une ou l'autre des classes de la variable binaire à expliquer. Arbitrairement, il est fixé à 0.5. Ce seuil de décision crée un compromis entre le nombre de négatifs et de positifs prédits car, tautologiquement, l'augmentation du seuil diminuera le nombre de positifs prédits et augmentera le nombre de négatifs prédits. Il ne peut être considéré comme un hyperparamètre au sens de l'ajustement du modèle car il ne modifie pas sa flexibilité.

La courbe de précision-rappel permet de représenter graphiquement cette dualité entre ces deux métriques de score. Elle permet de visualiser comment le choix du seuil affecte les performances du classificateur en termes de précision et de rappel, et peut aider à sélectionner le meilleur seuil pour une problématique en particulier.

Cette courbe se concentre principalement sur la performance de la classe positive qui est cruciale lorsque les classes sont déséquilibrées.

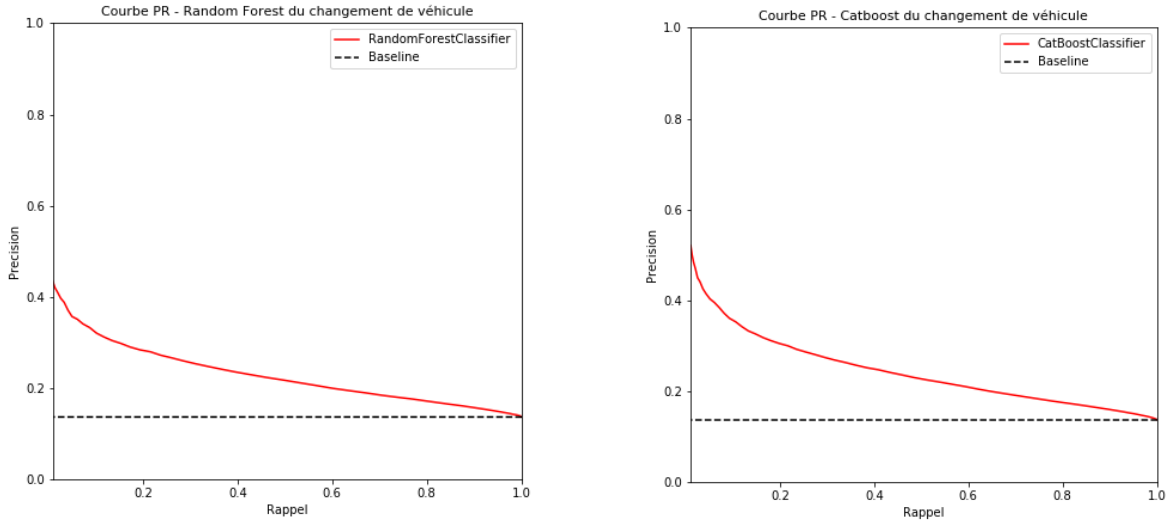


FIGURE 4.5 – Courbes de précision-rappel des modèles du changement de véhicule

Graphiquement, la ligne en pointillé représente un classificateur qui prédirait systématique le changement de véhicule. La proportion de changement de véhicule étant aux alentours des 14% dans notre base de données, il est logique que la précision d’un tel modèle se situe à ce niveau.

Dans l’espace PR (précision-rappel), l’objectif est d’être dans le coin supérieur droit (1,1). Cette situation correspond à un modèle avec une précision de 1, signifiant que tous les changements de véhicule prédits en étaient réellement, et un rappel de 1, signifiant que l’ensemble des changements de véhicule ont été correctement identifiés. Ainsi, le nombre de faux positifs et de faux négatifs est nul et le modèle identifie parfaitement toutes les classes.

Dans la pratique, les modèles construits se situent entre ces deux lignes : ils ne sont pas parfaits mais fournissent de meilleures prédictions qu’un modèle totalement arbitraire. Un bon classificateur maintiendra à la fois une précision et un rappel élevés.

La figure 4.5 illustre bien le fait qu’une fois le modèle calibré, il n’est plus possible d’améliorer la performance de l’une de ces deux métriques sans détériorer celle de l’autre.

Dans notre cas, augmenter le seuil au-delà de 0.5 améliore la précision du modèle puisqu’en devenant plus exigeant sur le niveau de fragilité requis pour être classé comme un changement de véhicule, on réduit le nombre de faux positifs. A l’inverse, ce rehaussement du seuil augmente le nombre de faux négatifs et détériore ainsi le rappel.

Le seul moyen d’améliorer simultanément la précision et le rappel, qui sont les métriques qui nous servent de base pour analyser la qualité de nos modèles, est de calibrer les hyperparamètres des modèles en les prenant comme mesures d’optimisation.

Ainsi, en suivant la méthodologie détaillée dans la section précédente, nous recherchons la combinaison d’hyperparamètres permettant de maximiser le  $F_{0.5}$  - score. La calibration des modèles permettant de tenir compte du déséquilibre des classes, le seuil des probabilités n’aura ainsi pas besoin d’être modifié.

Une autre représentation graphique traditionnellement utilisée pour évaluer la performance d'un modèle de classification est la courbe ROC et le score AUC associé.

La courbe ROC (de l'anglais, *Receiving Operator Characteristics*), construite à partir de la sensibilité (rappel) et de la spécificité, est une courbe de probabilité qui permet de visualiser les performances des modèles de classification. En fonction de toutes les valeurs de seuils possibles de la variable cible étudiée, elle exprime le taux de vrais positifs, ou sensibilité (fraction des positifs qui sont effectivement détectés) en fonction du taux de faux positifs, ou  $1 - \text{spécificité}$  (fraction des négatifs qui sont incorrectement détectés).

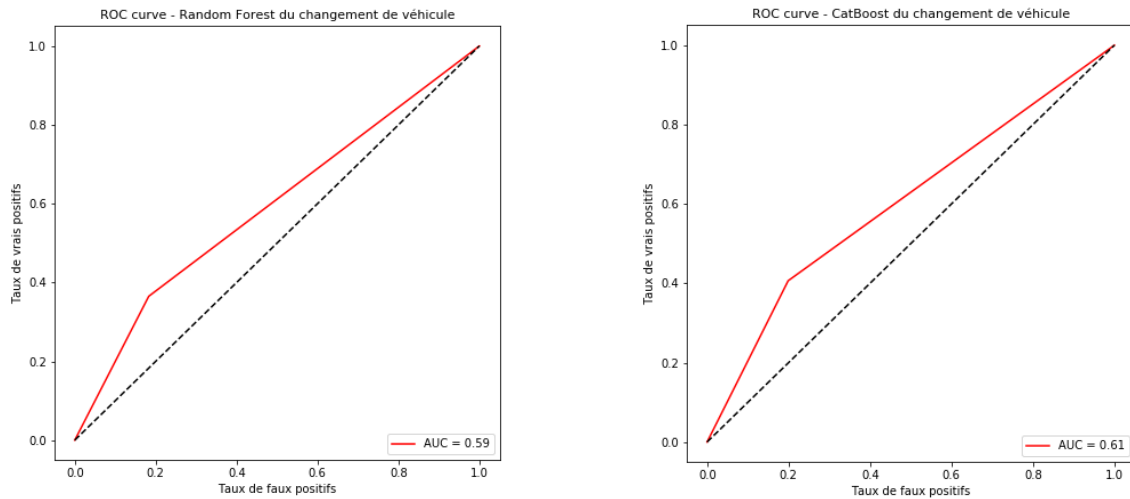


FIGURE 4.6 – Courbes ROC et AUC des modèles du changement de véhicule

La ligne pointillée en diagonale correspond à la performance que nous obtenons en moyenne lorsque nous tirons au hasard des scores sur l'intervalle  $[0; 1]$  ou si le modèle prédit systématiquement la classe majoritaire ou la classe minoritaire.

Dans l'espace ROC, le but est d'être situé dans le coin supérieur gauche  $(0, 1)$ . Cette situation correspond à un modèle avec un rappel de 1, signifiant que tous les changements de véhicule ont été correctement identifiés, et aucun faux positif, signifiant qu'aucun changement de véhicule prédit n'en était pas un dans la réalité.

Une mesure associée est l'AUC (*Area Under Curve*) qui mesure l'aire sous la courbe ROC et représente ainsi le degré de la séparabilité. Elle indique dans quelle mesure le modèle est capable de faire la distinction entre les classes. Plus l'AUC est élevé et proche de 1, plus le modèle est capable de prédire correctement les différentes classes : ses capacités discriminatoires sont bonnes. Lorsque l'AUC du modèle est de 0.5, cela signifie que le modèle n'a aucune capacité de séparation des classes. En dessous de cette valeur, le modèle a tendance à les inverser.

En analysant ces métriques à partir de la figure 4.6, on constate que les courbes ROC de chacun des modèles ne s'éloignent pas fortement de la première diagonale tandis que l'AUC est respectivement de 0.59 et 0.61 pour le Random Forest et le CatBoost. Au sens de ces mesures, les modèles présenteraient donc des performances de discrimination médiocres.



Néanmoins, si la courbe ROC et l'AUC sont des mesures populaires pour analyser et comparer les modèles, elles sont peu adaptées lorsque les classes de la variable cible sont déséquilibrées. En effet, la courbe ROC n'est pas sensible au taux de déséquilibre des classes puisque le taux de faux positifs, en abscisse, est stable quand le nombre de négatifs dans la distribution de la variable cible est lui élevé. De même, le taux de vrais positifs, en ordonnée, ne prend pas en compte ce déséquilibre.

Dans ce cadre, la courbe de précision rappel et l'aire sous cette courbe (AUC-PR) sont les métriques à privilégier puisqu'elles permettent d'intégrer la notion de déséquilibre et seront ainsi plus informatives.

L'AUC-PR mesure l'aire sous la courbe de précision-rappel. Une façon de l'obtenir est de calculer la précision moyenne comme la moyenne pondérée de la précision obtenue pour chaque seuil de classification et de la variation du rappel par rapport au seuil précédent. Ainsi, c'est une sorte de précision moyenne pondérée sur tous les seuils.

Dans un classificateur de base, cette métrique dépendra de la fraction des observations appartenant à la classe minoritaire. Ainsi, dans un jeu de données équilibré, le classificateur de base aura un AUC-PR de 0.5. Comparativement à cette valeur de base, plus le score AUC-PR est élevé, plus le classificateur est performant pour la tâche donnée.

Pour nos deux modèles, les scores d'AUC-PR obtenus sont respectivement de 0.25 et 0.23, pour le CatBoost et le Random Forest. La valeur de base étant de 0.14 (14% de changements de véhicule), on peut donc estimer que la qualité de discrimination des modèles est plutôt bonne.

Afin de tenir compte des spécificités de notre base d'étude, c'est-à-dire la présence d'une distribution asymétrique de la variable cible, les métriques les plus informatives pour analyser la performance des classificateurs sont donc la matrice de confusion, la précision et le rappel ainsi que d'autres métriques construites à partir de ces dernières, à savoir la courbe de précision-rappel et l'AUC-PR.

L'analyse de ces différentes métriques nous ont permis de constater que les deux modèles présentent des performances relativement similaires.

Pour pousser l'analyse un peu plus loin et déterminer laquelle des deux modélisations semble la plus appropriée pour notre problématique, nous pouvons nous intéresser à ces scores sur les différents quantiles de probabilités prédites.

En effet, la finalité de notre étude est d'identifier les clients sur lesquels il serait intéressant de mener des politiques de rétention au changement de véhicule. Le nombre de clients dans le portefeuille étant important et les actions marketing pouvant s'avérer coûteuses, il ne sera pas possible de mettre en place des actions pour l'ensemble des clients identifiés comme fragiles. Avoir connaissance des performances de notre modèle en fonction du périmètre des clients que nous souhaiterons toucher peut donc apporter une réelle information.

Ainsi, il est intéressant d'analyser l'évolution de la précision et du rappel en fonction des quantiles de probabilités prédites de changement de véhicule. Ces évolutions pour chacun des deux modèles sont présentées dans la figure 4.7 (Random Forest) et la figure 4.8 (CatBoost).

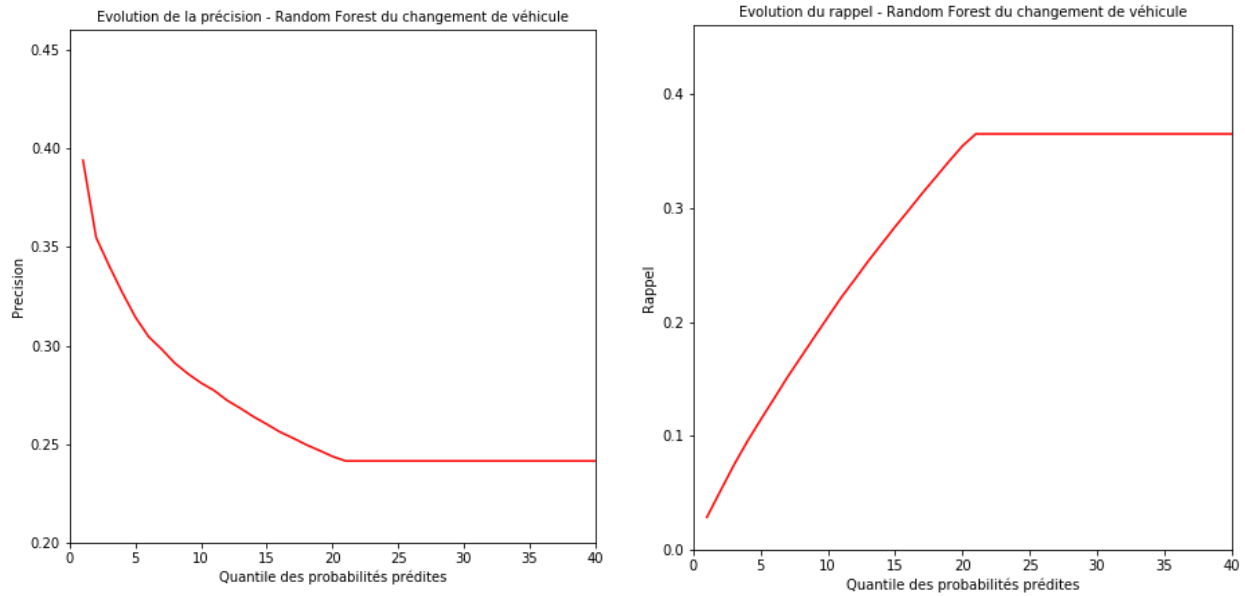


FIGURE 4.7 – Évolution de la précision et du rappel pour le Random Forest du changement de véhicule en fonction des quantiles de probabilités prédites

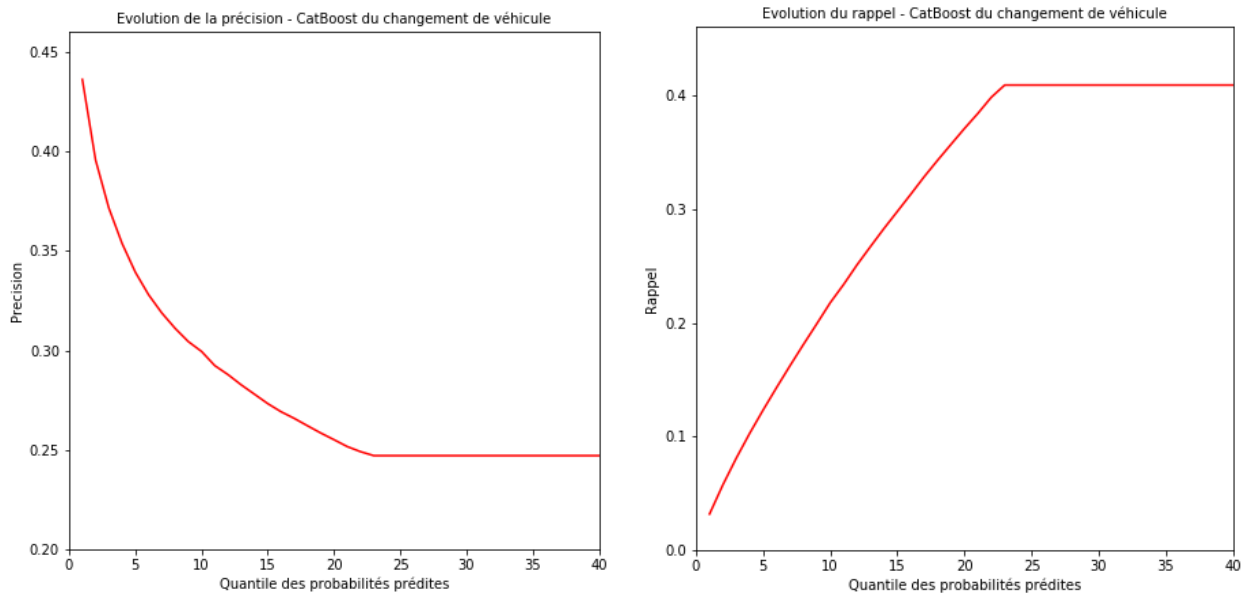


FIGURE 4.8 – Évolution de la précision et du rappel pour le CatBoost du changement de véhicule en fonction des quantiles de probabilités prédites

De ces graphiques, l'information la plus marquante que nous pouvons retirer est que la performance des modèles plafonne au-delà des 25% des probabilités prédites les plus élevées.

Métriques	Random Forest	CatBoost
<b>1%</b>		
Précision	39.6%	44.09%
Rappel	2.86%	3.2%
<b>5%</b>		
Précision	31.44%	34.02%
Rappel	11.42%	12.36%
<b>10%</b>		
Précision	28.11%	29.81%
Rappel	20.43%	21.67%
<b>15%</b>		
Précision	26.01%	27.38%
Rappel	28.36%	29.86%
<b>20%</b>		
Précision	24.38%	25.48%
Rappel	35.44%	37.04%
<b>25%</b>		
Précision	24.15%	24.59%
Rappel	36.49%	41.01%
<b>50%</b>		
Précision	24.15%	24.59%
Rappel	36.49%	41.01%

TABLE 4.3 – Métriques de scores sur les quantiles de probabilités prédites des modèles du changement de véhicule

Si la précision sur l'ensemble de l'échantillon pour les deux modèles était sensiblement similaire pour les deux modèles, on constate que le CatBoost performe mieux sur les clients les plus fragiles. En effet, sur respectivement 1% et 5% des clients avec les probabilités prédites les plus fortes, 44.1% et 34% des prédictions de changement de véhicule sont correctes. Pour le RandomForest, ces scores s'établissent à 39.6% et 31.4%.

Par ailleurs, si la modélisation CatBoost présentait des scores plus favorables concernant le pourcentage de changements de véhicule qu'elle parvenait à capter sur l'ensemble de l'échantillon, cette tendance se confirme lorsque l'on regarde les différents quantiles de prédictions. En effet, 1% et 5% des probabilités prédites les plus élevées représentent respectivement 3,2% et 12,4% de l'ensemble des changements de véhicule de la base de données. Pour le Random Forest, ces scores s'établissent à 2,9% et 11,4%.

Au regard de ces métriques, le CatBoost dont les hyperparamètres ont été optimisés semblent ainsi proposer les meilleures qualités prédictives.

En définitive, la sélection des métriques à utiliser pour évaluer la performance d'un modèle est une étape particulièrement importante puisqu'elle permet de déterminer si la modélisation s'adapte bien aux données et est suffisamment discriminante pour distinguer les profils fragiles au changement de véhicule. La faible proportion des changements de véhicule dans notre base d'étude a été un aspect à appréhender avec précaution puisqu'il conduit de nombreuses métriques traditionnellement utilisées à être peu informatives, voire trompeuses. Les indicateurs retenus sont ainsi des scores accordant une forte importance aux performances du modèle sur la classe minoritaire, objet de notre intérêt. Le choix des métriques d'optimisation nous a ainsi permis de disposer de points de repère pour comparer les deux modèles mais également pour évaluer les différentes versions d'un même modèle lors de la recherche de la meilleure combinaison d'hyperparamètres. La conclusion de cette partie est que le modèle offrant les meilleures garanties au niveau de la fiabilité des prédictions est le CatBoost.

## 4.5 Interprétation des résultats

Après avoir pu juger de la qualité prédictive des modèles, nous allons à présent nous intéresser à l'interprétation des résultats obtenus du modèle CatBoost. Un point d'attention particulier sera porté sur l'importance des variables dans la classification ainsi que leur impact à la hausse ou à la baisse sur la probabilité de changer de véhicule.

Si dans de nombreux cas il est nécessaire d'avoir des modèles précis, il est tout autant important d'avoir des modèles interprétables. Effectivement, en dehors de vouloir connaître les prédictions faites par le modèle, nous voulons comprendre ce qui les conduit à être plus ou moins élevées et quelles sont les variables qui sont importantes pour déterminer les résultats.

La compréhension des modèles permet de remédier à l'effet "boite noire" souvent associé aux méthodes de classification.

En effet, d'un point de vue opérationnel, les arbres construits à travers ces méthodes utilisent parfois des segmentations arbitraires. De ce fait, les seuils de classification déterminés pour les variables continues, bien que mathématiquement justes, ne sont pas toujours adaptés à la vision des équipes.

Par ailleurs, il en découle également des "effets de saut", résultat de l'absence de propriété de continuité. Concrètement, cela signifie que deux individus ayant des caractéristiques très proches pourront avoir des scores significativement différents. Justifier cette prédiction pourrait s'avérer difficile d'un point de vue métier. Si dans le cadre de notre étude, des scores très éloignés résulteraient dans la possibilité ou non de faire l'objet d'une politique de rétention et donc de bénéficier d'offres commerciales, ce point pourrait être plus difficile à expliquer si ces méthodes étaient utilisées pour des problématiques de tarification.

Pour terminer, une dernière fragilité des arbres tient dans le traitement des interactions entre les différentes variables. En effet, au moment de la détermination d'un nœud, le découpage de la variable continue est indépendant de celui-ci fait sur une variable différente à l'étape précédente puisque les différents pas sont indépendants les uns des autres. Plusieurs auteurs, comme Rakotomalala [14] parlent de la "myopie" des arbres de décision.

Pour faire face à ces limites, il est donc nécessaire d'analyser avec minutie les résultats du modèle. En effet, outre la qualité prédictive des modèles, il faut chercher à connaître les variables influentes afin de vérifier la cohérence avec la connaissance métier et maîtriser le niveau de risque du modèle.

Pour éviter que les individus ne subissent des décisions arbitraires émanant uniquement des modèles, des dispositions légales ont été prises avec l'article 22 du Règlement Européen sur la Protection des Données (RGPD) qui a pour vocation de poser des règles applicables au profilage et aux décisions entièrement automatisées. Ainsi, les modèles sans explication risquent d'entraîner des sanctions pouvant aller jusqu'à 20 millions d'euros ou, dans le cas d'une entreprise, à 4% du chiffre d'affaires mondial total de l'exercice précédent (le montant maximal étant retenu).

Les besoins de transparence et de confiance dans les algorithmes d'apprentissage automatique ont ainsi fait émerger deux concepts : l'interprétabilité et l'explicabilité.

- Un modèle est dit **interprétable** s'il est possible d'identifier les caractéristiques ou variables qui participent le plus à la décision, voire même d'en identifier l'importance.
- Un modèle est dit **explicable** s'il est possible d'en rendre compte à partir des données connues de la situation. En d'autres termes, s'il est possible de mettre en relation les valeurs prises par certaines variables et leurs impacts sur la prévision du score et ainsi sur la décision.

Un modèle explicable est interprétable mais l'inverse n'est pas automatique. C'est pourquoi l'interprétabilité est la première notion à analyser avant de s'intéresser à l'explicabilité.

#### 4.5.1 Interprétabilité du modèle

L'interprétabilité d'un modèle permet de comprendre comment un algorithme prend sa décision, c'est-à-dire quelles données sont prises en compte pour établir le score ou la classification.

Pour extraire des informations du modèle, la première étape consiste à définir l'importance des variables du modèle de manière globale. Pour ce faire, nous utiliserons le poids affecté à chaque variable dans la prise de décision de l'algorithme. Ce poids peut être mesuré au moyen de l'importance des variables calculée au moment de la phase d'apprentissage ou de validation des modèles. La figure 4.9 hiérarchise les variables selon le niveau d'explication qu'elles apportent à la variable cible.

On constate ainsi que 5 variables participent grandement à la détermination du changement de véhicule : l'âge et la durée de détention du véhicule, le comportement observé sur les 10 dernières années, l'âge du conducteur ainsi que la formule souscrite sur le contrat.

L'âge du véhicule intervient en première position. Plusieurs phénomènes peuvent entrer en ligne de compte pour déterminer le moment le plus opportun à la revente d'un véhicule et à son remplacement. Tout d'abord, la dépréciation de la valeur du véhicule après son achat à neuf peut être un facteur pouvant expliquer que certains biens récemment acquis soient mis à la vente afin que cette dernière constitue un apport confortable pour l'acquisition d'un nouveau bien. Par ailleurs, on peut également imaginer que l'âge du véhicule est lié à son kilométrage. Ce paramètre entre en considération pour définir le meilleur moment de revente d'un véhicule mais impacte également son usure et ainsi sa nécessité d'être remplacé. D'autres éléments comme l'obsolescence du véhicule due à son âge, l'arrivée sur le marché de nouveaux modèles plus performants, la sécurité de remplacer un véhicule devenu moins fiable, etc. . . sont autant de facteurs qui expliquent le fort impact de l'âge sur la prédiction du changement de véhicule.

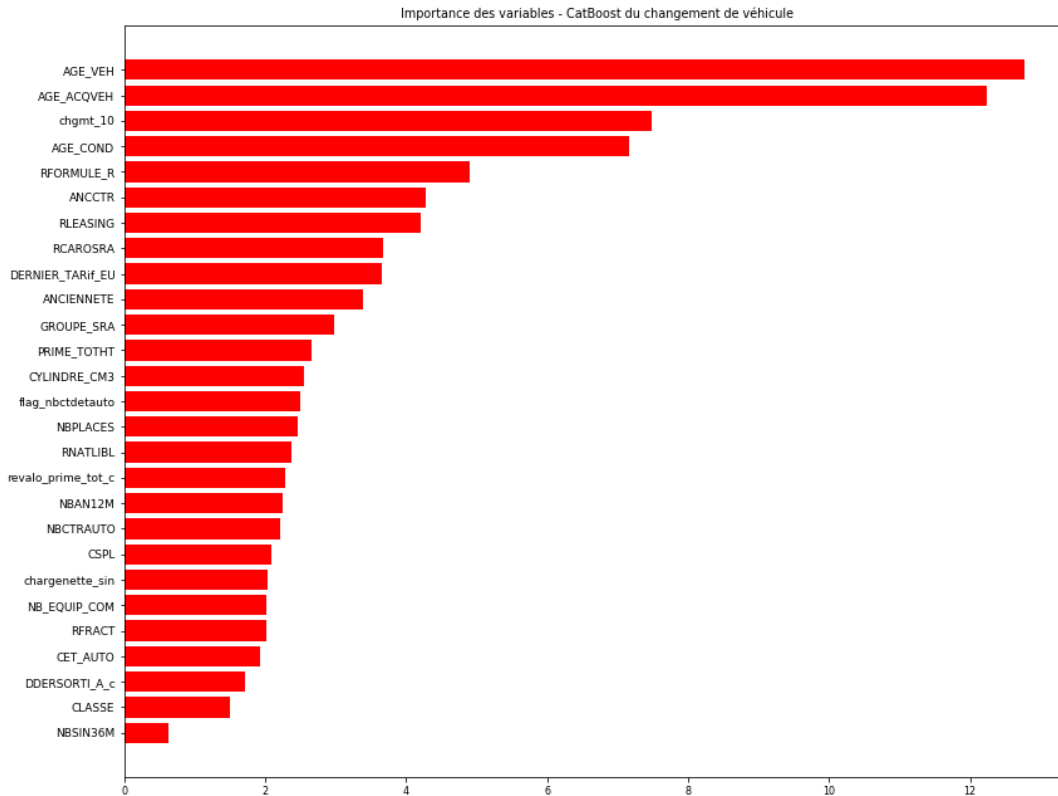


FIGURE 4.9 – Importance des variables pour le CatBoost du changement de véhicule

Indirectement liée à l'âge du véhicule, bien que ces deux variables ne soient pas colinéaires, la durée de détention du véhicule intervient en seconde position. Sachant que près de deux tiers de notre portefeuille sont des véhicules achetés d'occasion, la durée de détention est nécessairement un facteur important puisque ces derniers n'ont pas été acquis au moment où la valeur du véhicule et ses caractéristiques techniques étaient à leur optimum. Elle présente donc un angle d'analyse différent. La durée de détention peut ainsi se référer à la fois aux habitudes de consommation des individus avec la durée qu'ils estiment individuellement adéquate pour conserver un véhicule, à l'essor du marché de seconde main qui leur permet de changer plus régulièrement de véhicule grâce à l'offre de véhicules à moindres coûts, ou encore à l'intérêt porté au secteur automobile, etc. . .

Ensuite, le comportement observé en matière de changement de véhicule sur les 10 dernières années rapporté à la présence en portefeuille est une variable particulièrement intéressante pour prédire un changement de véhicule dans l'année qui suit. Cette variable permet de capter le comportement moyen des individus, pour ceux dont l'ancienneté du contrat est suffisamment importante pour disposer d'un historique informatif. Ainsi, il est tout à fait logique que des individus ayant changé de véhicule 3 fois en 10 ans ou 2 fois en 5 ans soient considérés comme potentiellement plus fragiles que des individus n'ayant pas ou qu'une seule fois changé de véhicule sur les mêmes périodes.

D'un impact tout aussi marqué, l'âge du conducteur explique sa volonté de changer ou non d'automobile. Le patrimoine, les revenus et le pouvoir d'achat évoluant au cours du cycle de vie d'un individu, ce dernier est plus enclin à investir dans l'achat d'un nouveau bien à un instant donné de sa vie plutôt qu'à un autre.

Pour les mêmes raisons que nous avons détaillé précédemment, la formule souscrite sur le contrat, l'option de leasing, le dernier tarif connu à neuf du véhicule, etc. . . ont un impact assez significatif sur la construction du score de fragilité.

#### 4.5.2 Explicabilité du modèle

Comparée à l'étude de l'interprétabilité du modèle, l'analyse sur l'explicabilité consiste à changer d'échelle afin d'extraire des informations locales pour des cas spécifiques de notre base de données.

Pour cela, nous utiliserons l'approche proposée par Lundberg et Lee en 2017 [9] : SHAP (Shapley Addition exPlanations). Extension de la valeur de Shapley, cette approche permet d'expliquer les résultats de n'importe quel modèle d'apprentissage automatique.

L'un de ses avantages est l'explicabilité globale – les valeurs SHAP calculées collectivement permettent de montrer à quel point une variable contribue, positivement ou négativement, à la variable cible. Cependant, son réel apport concerne l'analyse locale. En effet, chaque observation reçoit son propre ensemble de valeurs SHAP de sorte qu'il est possible d'expliquer au cas par cas la prédiction et la contribution des différents facteurs. Cet angle d'analyse nous permet de mettre en évidence et de contraster les impacts des différentes variables explicatives.

Grâce au fait que les valeurs soient calculées pour chaque observation de la base de données, il est possible de représenter chaque cas par un point et ainsi avoir une information supplémentaire sur l'impact d'une caractéristique sur la variable cible en fonction de sa valeur.

Sur la figure 4.10, les points rouges représentent ainsi des valeurs élevées de la variable considérée tandis que les points bleus des valeurs basses. Leur positionnement par rapport à l'axe permet de visualiser si ces dernières ont un impact négatif (à gauche) ou positif (à droite) sur la variable cible.

Tout d'abord, on remarque que les effets de l'âge du véhicule ne sont pas linéaires, ce qui conforte ce que nous avons observé lors de l'analyse comportementale en fonction des différentes segmentations de cette variable. En effet, les valeurs élevées et basses de l'âge du véhicule semblent avoir un impact à la fois négatif et positif sur la probabilité de changer de véhicule. Cependant, on peut souligner que les véhicules très âgés ont tendance à plus impacter positivement ce comportement, comme nous l'avons observé pour les véhicules âgés de plus de 15 ans.

En revanche, la tendance de la durée de détention du véhicule est plus nette. Ainsi, on observe que les détentions longues ont tendance à favoriser le changement de véhicule tandis que les acquisitions plus récentes réduisent cette probabilité.

De la même façon, l'âge du conducteur semble présenter un impact assez linéaire. Les jeunes individus impactent positivement cette probabilité tandis que les personnes plus âgées la réduisent.

On peut se référer à la théorie microéconomique du cycle de vie développée par Franco Modigliani et Albert Ando [12] pour expliquer comment un agent économique choisit son niveau de consommation et d'épargne au cours de sa vie. L'idée est que les individus rationnels accumulent de l'épargne pendant leur période d'activité pour ensuite désépargner une fois le passage à la retraite pour maintenir leur niveau de consommation. L'âge détermine donc à la fois les revenus de l'individu et son patrimoine.

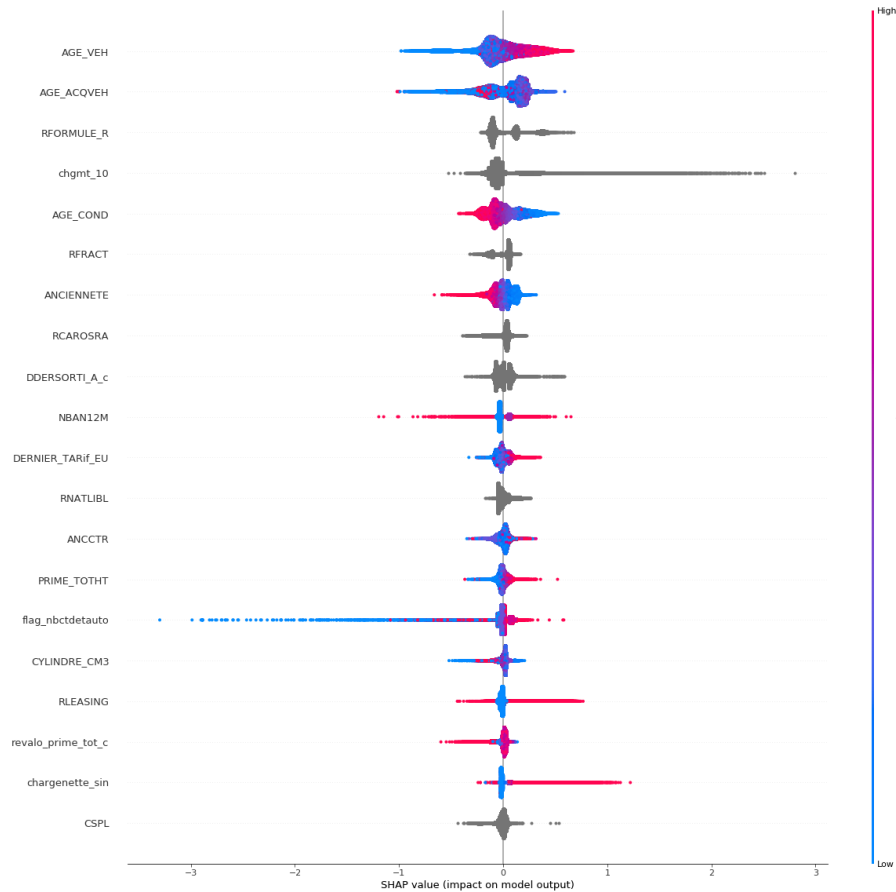


FIGURE 4.10 – Importance des variables au moyen des valeurs SHAP pour le CatBoost du changement de véhicule

Selon cette théorie, trois périodes se distinguent :

1. **La jeunesse** : la consommation est supérieure au revenu et le patrimoine est négatif. Les agents sont dans une phase d'investissement et empruntent pour financer la vie courante. Durant cette période, ils sont donc plus susceptibles d'acquérir un bien immobilier ou un véhicule.
2. **La période d'activité** : les revenus deviennent supérieurs à la consommation car les individus font le choix d'épargner et de constituer un patrimoine pour financer la consommation de la période suivante. Ainsi, ils peuvent être moins enclins à investir dans des biens matériels, tels que des véhicules, en prévision d'une baisse des revenus ; ou bien ces derniers ont déjà été acquis durant la période précédente.
3. **La période de retraite** : les revenus diminuent mais les agents utilisent leur patrimoine pour maintenir leur consommation. Cette phase correspond à une période de désépargne durant laquelle l'achat d'un nouveau véhicule pourrait ainsi paraître comme secondaire. Outre cet aspect financier, les contraintes biologiques liées à l'avancement de l'âge sont également des raisons qui limitent l'utilisation d'un véhicule et diminuent ainsi l'intérêt d'en changer.



Par ailleurs, l'ancienneté du contrat et l'ancienneté en portefeuille ont des impacts opposés. Si le caractère récent du contrat atténue la probabilité de changer de véhicule, un client récemment entré dans le portefeuille de Generali est plus fragile au changement qu'un client plus ancien. Comme nous l'avons vu, le changement de véhicule est à l'origine de nombreuses résiliations. De ce fait, les contrats plus récents peuvent correspondre à des véhicules ayant été récemment acquis. En suivant le même raisonnement, les contrats plus anciens sont ainsi plus exposés à un changement de bien assuré puisque l'âge avancé du véhicule a tendance à augmenter la probabilité de changement. L'effet de l'ancienneté du client est plus difficile à analyser de manière indépendante puisqu'un client plus ancien pourrait avoir déjà effectué des changements de véhicule, ce qui est capté dans la variable croisée retraçant l'historique des changements de véhicules en fonction de la présence du contrat sur les 10 dernières années.

Si l'âge du véhicule a un impact non linéaire, une autre variable caractérisant le véhicule présente un impact plus lisible : le dernier prix connu à neuf du véhicule. On remarque ainsi que les véhicules haut de gamme, dont les prix sont élevés, ont un impact positif sur la fragilité au changement de véhicule. A l'inverse, les véhicules moins onéreux ont un impact négatif sur la valeur de ce score. Comme évoqué à de multiples reprises, la dépréciation du prix est un phénomène dont il faut tenir compte, et celui-ci est d'autant plus important pour les véhicules à forte valeur pour lesquels la perte de valeur en montant absolu est beaucoup plus importante et stimule ainsi des comportements hétérogènes par rapport à la population moyenne. De plus, certains véhicules de luxe ou de collection ne répondent pas à une acquisition "rationnelle" et sont donc plus difficiles à prédire.

Cette méthode d'analyse nous a ainsi permis d'estimer et d'analyser le sens des relations existantes entre certaines variables explicatives et la variable cible.

Néanmoins, certaines variables catégorielles aux multiples modalités, telles que le comportement du changement de véhicule sur les 10 dernières années, la carrosserie du véhicule, la CSP, etc... ne seront pas interprétables avec cette approche. En effet, elles correspondent à des variables non ordinales dont l'encodage des catégories ne pourra donner lieu à un ordonnancement possible des valeurs et pour lesquelles une valeur faible ou élevée ne serait pas représentative.

C'est l'une des limites de l'interprétation des modèles de classification : la lecture des résultats des modèles est rendue plus difficile par la transformation de l'ensemble des variables catégorielles en valeur numérique au travers de la procédure d'encodage, lecture qui est plus aisée dans le cas des modèles de régression linéaires où les différentes classes sont transformées en variables binaires et où l'interprétation est facilitée par le recours aux odd-ratios. La visualisation des arbres de classification obtenus pourrait permettre de récupérer cette information mais la multiplicité des nœuds et des branches détériore la lisibilité des résultats. Pour remédier à cela, des analyses descriptives sont donc à réaliser en dehors de la modélisation pour comprendre l'impact de ce type de variables, même si ces dernières pourraient être prises en compte différemment au moment de la modélisation.

Pour conclure ce chapitre sur la modélisation du changement de véhicule, les méthodes de classification justifient d'une qualité de discrimination tout à fait acceptable et leur forme non paramétrique, la gestion des classes déséquilibrées et l'encodage des variables catégorielles permettent de bien s'adapter à notre problématique. En tenant compte des spécificités des différentes métriques de score pour évaluer la performance des modèles, les résultats des différentes modélisations nous poussent à choisir le modèle CatBoost. Les méthodes de validation croisée permettent quant à elles de s'assurer de la stabilité des classifications, quelles que soient les observations retenues pour entraîner et valider le modèle.

Si l'analyse des résultats des méthodes de classification est généralement moins aisée que celle des méthodes de régression, facilitée par l'utilisation des odds-ratios et des coefficients associés à chaque variable explicative, plusieurs outils sont disponibles pour établir la liste des facteurs incitant les individus à changer de véhicule ou bien à le conserver.

Couplés à l'analyse descriptive que nous avons fait de la base d'étude, les résultats de ce modèle nous permettent d'établir les profils des individus considérés comme fragiles au changement de véhicule dans l'année à venir. En fonction du niveau de risque que représente un individu et ce qu'il rapporte à la compagnie, l'assureur déterminera s'il souhaite que ce client reste dans son portefeuille et pourra alors mettre en place des stratégies commerciales pour s'adapter à ses nouvelles attentes.

Cela est d'autant plus vrai s'il estime que le client est susceptible de résilier son contrat à ce moment clé qu'est le changement de véhicule. La volumétrie des assurés résiliant chaque année pour ce motif étant importante, il pourrait être intéressant, en plus de la construction d'un score de fragilité au changement de véhicule, de construire un modèle permettant de prédire la probabilité de résiliation d'un client. C'est à partir de ce score que des actions commerciales particulières, que l'on regroupe sous l'appellation "d'actions de rétention", pourront être mises en place pour inciter le client à rester chez Generali. Cette modélisation fera l'objet du chapitre suivant.

## 5 Chapitre 5 : Modélisation de la résiliation au changement de véhicule

L'objectif de ce chapitre est de modéliser la résiliation ayant pour motif le changement de véhicule. Pour un assureur, il est capital de comprendre l'acte de résiliation et ses motivations afin d'avoir la possibilité de proposer des offres commerciales pour inciter les clients à rester dans son portefeuille. Cette mise en place de politiques de rétention est d'autant plus importante à l'heure actuelle que la concurrence sur le marché de l'assurance automobile est élevée et que la souscription est un procédé coûteux.

Dans le chapitre précédent, nous avons construit un score de fragilité au changement de véhicule à partir d'un modèle CatBoost. La variable expliquée regroupait à la fois les changements de véhicule étant restés dans le portefeuille de Generali, avec un avenant de modification du risque, et les changements de véhicule ayant conduit à la résiliation du contrat automobile. Pour la modélisation mise en place dans ce chapitre, nous reprendrons cette dernière variable qui nous donne la probabilité conditionnelle de résilier sachant qu'un changement de véhicule a eu lieu. Elle permet de cibler non plus les clients fragiles au changement de véhicule mais les clients qui sont susceptibles de quitter le portefeuille à ce moment clé.

Comme nous l'avons décrit dans le chapitre 3, c'est près d'un changement de véhicule sur deux qui entraîne une résiliation. Le suivi et la fidélisation du client à cette période charnière du contrat d'assurance automobile doivent donc faire l'objet d'une attention particulière.

L'étude et la modélisation de la résiliation suivront la même méthodologie utilisée pour prédire le changement de véhicule. Les caractéristiques de la base d'étude sont sensiblement similaires, à l'exception près que les classes de la variable cible sont encore plus déséquilibrées avec seulement 7% du portefeuille qui résilie leur contrat d'assurance suite à un changement de véhicule.

Cette asymétrie marquée de la distribution de la variable à expliquer et la multiplicité des variables catégorielles nous conduisent, une nouvelle fois, à nous tourner vers des méthodes de classification dont l'utilité et la qualité prédictive ont été démontrées dans le chapitre précédent.

Ainsi, nous comparerons une fois de plus les performances d'une forêt aléatoire et d'un CatBoost.

### 5.1 Sélection des variables

Le processus de sélection des variables, étape primordiale pour optimiser le pouvoir prédictif des modèles construits, suivra le même procédé que la modélisation précédente.

Tout d'abord, l'analyse des dépendances au moyen de matrices de corrélation et de  $V$  de Cramer nous avait permis de constater que plusieurs variables étaient extrêmement dépendantes les unes des autres. Plusieurs variables parmi les caractéristiques des véhicules, les sinistres, la détention de différentes familles de produits ou encore les données INSEE sur l'environnement d'habitation sont colinéaires. La variable à expliquer n'étant pas prise en compte dans cette analyse, la suppression de certaines variables sera donc identique à ce qui avait été effectué auparavant.

Ensuite, la réduction de la dimension de notre base de données sera favorisée par l'analyse des résultats donnés par une première modélisation au moyen des régressions pénalisées.

L'étude des coefficients significativement différents de 0 de la régression LASSO nous permet d'avoir une idée sur les variables ayant un impact, positif ou négatif, sur la résiliation au changement de véhicule. De plus, l'analyse des coefficients les plus élevés de la régression RIDGE nous permet de faire la même analyse.

Ainsi, parmi l'ensemble des variables quantitatives et des différentes classes des variables qualitatives, 40 semblent significatives pour expliquer la résiliation au changement de véhicule. De manière exhaustive, elle semble être favorisée par l'ancienneté du véhicule, la souscription de la formule minimale, la prime totale hors taxes, la nature de l'assuré, le nombre de résiliations sur les 12 derniers mois, etc... A l'inverse, les facteurs incitant un individu à changer de véhicule et rester chez le même assureur sont son âge, la durée de détention du véhicule, l'ancienneté du contrat, le nombre de contrats détenus en portefeuille, le montant de rabais accordé ou encore le nombre de personnes assurées sur le contrat, etc...

Cette pré-sélection des variables nous permet ainsi de limiter le nombre de variables d'entrée pour la modélisation qui sera faite au moyen d'une forêt aléatoire ou d'un CatBoost.

Une sélection descendante, selon le critère d'importance de la variable dans la classification, permet de réduire successivement le nombre de variables utilisées pour parvenir à un périmètre d'une trentaine de variables.

Au moment de l'analyse des dépendances, nous avons identifié plusieurs variables apportant sensiblement la même information mais pour un niveau de renseignement différent. C'était notamment le cas pour le nombre de sinistres survenus sur les 36, 24 ou 12 derniers mois ou encore pour le nombre de contrats automobile ou total détenus en portefeuille. Nous avons alors testé plusieurs modélisations en intégrant les différents niveaux d'information et en conservant ceux permettant de maximiser le  $F_{0.5}$  - score.

Au final, les variables utilisées pour modéliser la résiliation au changement de véhicule sont listées dans le tableau ci-dessous :

<b>Véhicule</b>	<b>Conducteur</b>	<b>Contrat</b>	<b>Parcours client</b>
Age véhicule	Age conducteur	Ancienneté contrat	Nb contrats auto
Durée détention	Ancienneté client	Prime HT	Variation nb contrats auto
Carrosserie	Nature	Leasing	Revalorisation prime
Groupe SRA	CSP	Fract. paiements	Changements véhicule 10 ans
Classe SRA	Situation maritale	Nb conducteurs	Multi-équipement auto-MRH
Cylindrée (cm3)	Nb équip. commune	CET	Nb résiliations 12M
Nb sinistres 36M			Nb AN 12M
Charge nette sin.			Date dernière souscription IARD
			Date dernière sortie
			Famille dernier sinistre

TABLE 5.1 – Listes des variables explicatives pour la modélisation de la résiliation au changement de véhicule

## 5.2 Analyse de la qualité des modèles

Au regard des métriques que nous avons utilisé dans le chapitre précédent, nous pouvons effectuer une optimisation des hyperparamètres pour chacun des modèles de classification. Cette recherche de la meilleure combinaison permet d’adapter le mieux possible les modèles à nos données et ainsi tirer le plus d’informations des différentes variables pour maximiser la qualité des modèles.

Par exemple, une recherche faite au moyen de la méthode de grilles de paramètres (GridSearch) permet de choisir la combinaison d’hyperparamètres maximisant le  $F_{0.5}$  - score. Cette recherche, qui peut s’avérer très coûteuse en temps de calcul, s’arrête une fois que les gains en termes de scores ne sont plus assez significatifs pour justifier l’apprentissage de la recherche.

Ainsi, sur la table 5.2, on constate que la huitième combinaison d’hyperparamètres pour le CatBoost permet d’obtenir de très bons résultats en termes de rappel tandis que la soixantième combinaison permet d’en obtenir de meilleurs en termes de précision. Pour les raisons que nous avons évoquées, la précision est la métrique de score sur laquelle se focalise notre intérêt. Les hyperparamètres sélectionnés devront permettre de l’optimiser, sans pour autant trop affaiblir les scores en termes de rappel. Dans ce cadre, nous essayons de maximiser le  $F_{0.5}$  - score tout en maintenant des scores élevés en termes de précision et de rappel.

N°	Iterations	Depth	One-hot max	Learning rate	Poids $\omega_j$	Précision	Rappel	$F_1$	$F_{0.5}$
1	1000	8	8	0.05	[1 ; 6]	0.191	0.215	0.202	0.195
2	1000	8	8	0.05	[1 ; 7]	0.173	0.306	0.221	0.189
3	1000	8	8	0.05	[1 ; 8.5]	0.153	0.429	0.226	0.175
...	...	...	...	...	...	...	...	...	...
8	1000	8	8	0.05	[1 ; 12]	0.126	0.643	0.211	0.151
...	...	...	...	...	...	...	...	...	...
60	2000	8	10	0.075	[1 ; 8.75]	0.149	0.42	0.22	0.171
...	...	...	...	...	...	...	...	...	...

TABLE 5.2 – GridSearch des hyperparamètres du Catboost de la résiliation au changement de véhicule (*one-hot max = one hot max size*)

La problématique à laquelle nous tentons de répondre présente les mêmes difficultés que la modélisation du changement de véhicule, à savoir le déséquilibre des classes, ainsi que les mêmes objectifs, à savoir la volonté de trouver le modèle permettant de cibler de manière optimale les clients pour des campagnes de rétention. De ce fait, les différentes métriques de score utilisées pour analyser la qualité des modèles seront également la matrice de confusion, la précision, le rappel, le  $F_\beta$  - score, la courbe de précision-rappel et l’analyse de ces métriques sur les quantiles de probabilités.

Ainsi, si l’on observe les matrices de confusion pour un modèle de forêt aléatoire et d’un CatBoost dont les hyperparamètres ont été optimisés, on remarque que le premier modèle cité semble mieux prédire l’absence de résiliation en identifiant un nombre plus important de cas et en commettant moins d’erreurs sur cette classe, tandis que le second modèle réalise de meilleures prédictions sur la classe positive c’est-à-dire la résiliation.

Random Forest			CatBoost		
	0	1		0	1
0	321 099	55 352	0	302 393	74 058
1	20 880	9 998	1	17 908	12 970

TABLE 5.3 – Matrices de confusion des modèles de la résiliation au changement de véhicule

Métriques	Random Forest	CatBoost
Base d'apprentissage		
Exactitude	83.32%	77.66%
Base de validation		
Exactitude	81.28%	77.42%
Précision	15.29%	14.90%
Rappel	32.38%	42.00%
$F_1$ - score	20.78%	22.00%
$F_{0.5}$ - score	17.10%	17.11%

TABLE 5.4 – Métriques de scores des modèles de la résiliation au changement de véhicule

Si les scores obtenus en termes de précision sont proches, respectivement 15.3% et 14.9% des prédictions de résiliations qui sont correctes pour le Random Forest et le Catboost, ce dernier a des scores bien supérieurs en termes de rappel avec 42% de l'ensemble des résiliations qui ont été identifiées contre seulement 32% pour l'autre modèle.

Si les niveaux des scores de rappel sont similaires à ceux obtenus pour le changement de véhicule, les scores de précision semblent bien inférieurs. En effet, pour le meilleur des deux modèles au sens de cette métrique, seulement 1 prédiction de résiliation sur 6 est correcte. Néanmoins, ce résultat est à mettre en perspective avec la répartition de cette classe dans le portefeuille car cette dernière représente seulement 7.5% de l'ensemble des données. Un modèle qui prédirait aléatoirement la résiliation aurait une précision de 7.5%. Autrement dit, cela signifie que les modèles obtenus sont déjà deux fois plus performants que le hasard, tout comme l'étaient les modèles du changement de véhicule.

Par ailleurs, on remarque que les  $F_{0.5}$  - scores sont sensiblement les mêmes pour les deux modèles, ces derniers ayant des scores de précisions quasi identiques et cette métrique étant prépondérante dans le calcul de cet indicateur. A l'inverse, poussé par un meilleur rappel, le CatBoost dispose d'un  $F_1$  - score plus élevé.

L'analyse de la courbe de précision-rappel à partir de la figure 5.1 permet une nouvelle fois d'observer la dualité entre ces deux scores en fonction du seuil choisi pour classer les probabilités prédites dans l'une ou l'autre des deux classes de la variable cible. Le pourcentage de l'ensemble des résiliations capté par la forêt aléatoire étant moindre que le CatBoost, sa courbe est plus tassée vers le bas.

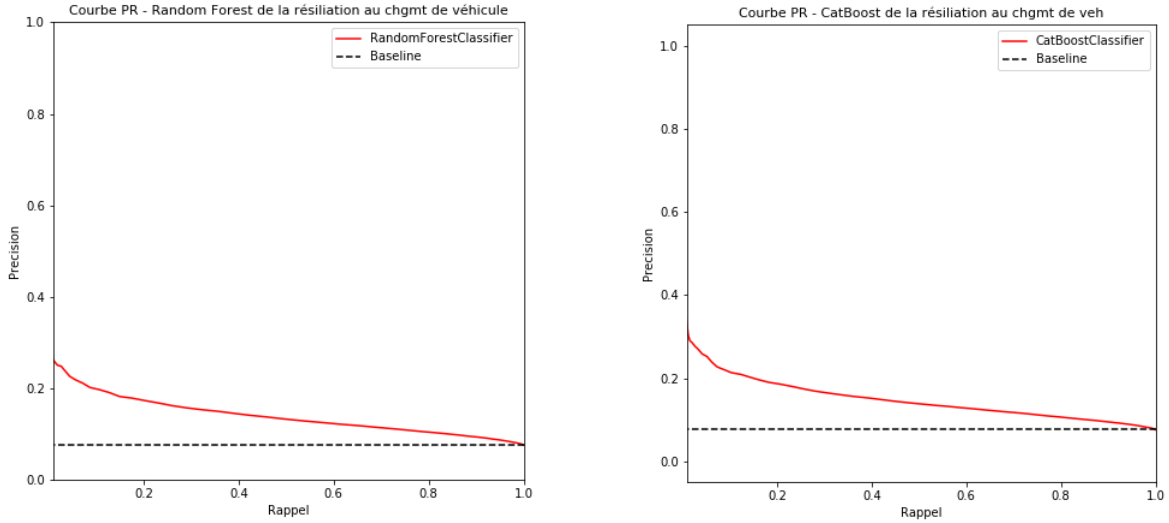


FIGURE 5.1 – Courbes de précision-rappel des modèles de résiliation au changement de véhicule

L'étude de la précision et du rappel sur les quantiles de probabilités sera celle qui sera la plus révélatrice pour choisir la modélisation utilisée puisqu'elle permet d'avoir un ordre d'idée sur l'efficacité potentielle d'une campagne de rétention en fonction du ciblage effectué.

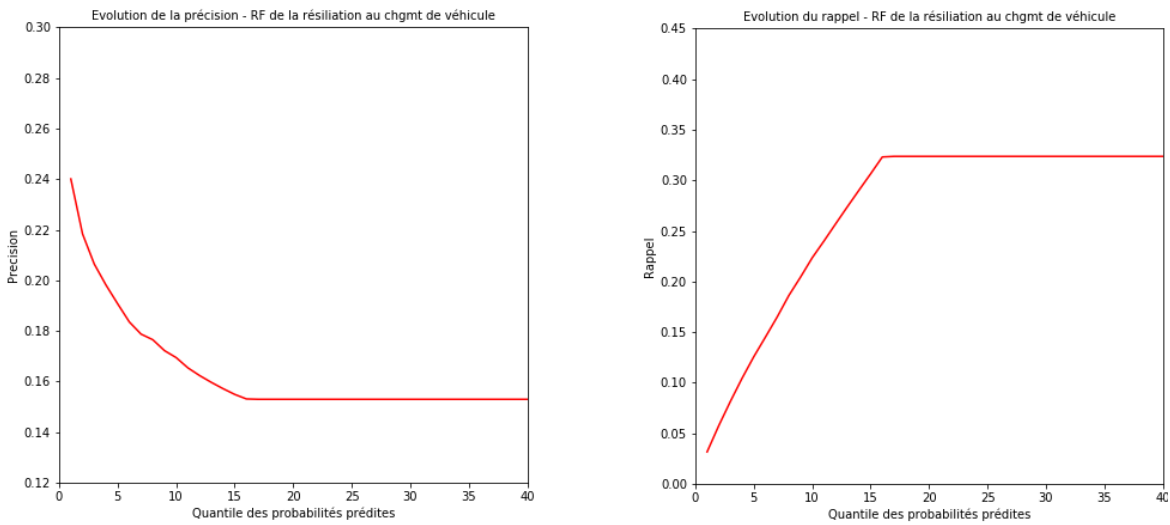


FIGURE 5.2 – Évolution de la précision et du rappel pour le Random Forest de la résiliation au changement de véhicule en fonction des quantiles de probabilités

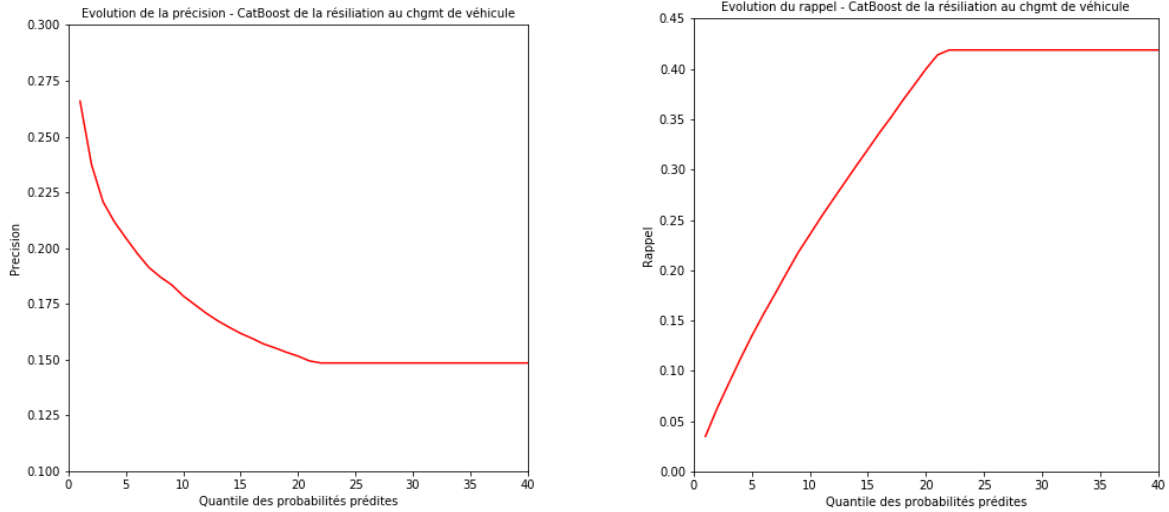


FIGURE 5.3 – Évolution de la précision et du rappel pour le CatBoost de la résiliation au changement de véhicule en fonction des quantiles de probabilités

Métriques	Random Forest	CatBoost
<b>1%</b>		
Précision	24.01%	26.58%
Rappel	3.16%	3.51%
<b>5%</b>		
Précision	19.07%	20.41%
Rappel	12.58%	13.45%
<b>10%</b>		
Précision	16.94%	17.82%
Rappel	22.35%	23.51%
<b>15%</b>		
Précision	15.49%	16.26%
Rappel	30.65%	32.18%
<b>20%</b>		
Précision	15.29%	15.17%
Rappel	32.38%	40.02%
<b>25%</b>		
Précision	15.29%	14.90%
Rappel	32.38%	42.00%
<b>50%</b>		
Précision	15.29%	14.90%
Rappel	32.38%	42.00%

TABLE 5.5 – Métriques de scores sur les quantiles de probabilités prédites des modèles de la résiliation du changement de véhicule



Ainsi, pour le modèle CatBoost, une campagne qui ciblerait les 1% des clients dont les probabilités prédites de résiliation suite à un changement de véhicule sont les plus fortes aurait 26.6% de chance de s'être adressée à un client susceptible d'adopter ce comportement. En ciblant ce périmètre de clients, la compagnie aurait par ailleurs contacté 3.5% de l'ensemble des clients qui résilieraient pour ce motif. Pour le modèle de forêt aléatoire, ces scores sont de 24.1% pour l'efficacité du ciblage de la campagne et de 3.1% pour la part de l'ensemble des résiliations.

L'étude sur les différents quantiles de probabilités nous montre que le Catboost propose des meilleurs scores jusqu'aux 15% des clients les plus fragiles à la résiliation. Pour le reste, les scores en termes de précision restent très proches pour les deux modèles, bien que la proportion des résiliations identifiées (le rappel) reste nettement supérieure pour le CatBoost.

Au regard de ces scores et de leur signification pour les campagnes de rétention, le CatBoost semble, une nouvelle fois, être le meilleur modèle pour prédire la résiliation au changement de véhicule.

Les résultats ainsi obtenus semblent satisfaisants lorsqu'on utilise des métriques adaptées à notre base de données et à la problématique à laquelle nous souhaitons répondre.

En effet, le modèle mis en place permettra d'établir un score de fragilité au changement de véhicule suivi d'une résiliation. A partir de ces prédictions, l'assureur identifiera les clients qu'il souhaite contacter, en fonction de critères que nous aborderons dans le chapitre suivant. La qualité du modèle déterminera ainsi le pourcentage de clients contactés qui seraient réellement concernés par les offres de la compagnie.

Les chiffres à retenir pour ce modèle sont que sur les 1%, 5% et 10% des probabilités prédites les plus fortes, la précision est respectivement de 26.6%, 20.4% et 17.8%. Autrement dit, cela signifie que sur ces quantiles de probabilités, le modèle performe près de 4 fois, 3 fois et 2.5 fois mieux que le hasard. Par ailleurs, des campagnes qui cibleraient ce périmètre de clients permettraient de toucher respectivement 3.5%, 13.5% et 23.5% de l'ensemble des clients fragiles à la résiliation suite à un changement de véhicule.

Après avoir pu juger de la qualité prédictive des modèles, nous allons nous intéresser à l'interprétation des résultats obtenus. En effet, pour un assureur, il est nécessaire de comprendre les facteurs poussant un client à résilier son contrat afin de pouvoir agir sur certaines variables qui sont à sa main, et plus particulièrement sur les caractéristiques tarifaires du contrat. L'analyse de l'interprétabilité et de l'explicabilité du modèle nous permettra de comprendre les caractéristiques d'un individu, de son bien assuré, de son contrat ou encore son parcours client qui le conduisent à se tourner vers la concurrence au moment de changer de véhicule.

### 5.3 Interprétation des résultats

Pour l'interprétabilité du modèle, nous reprendrons l'analyse faite au moyen des poids affectés à chaque variable dans la prise de décision de l'algorithme. Ce poids peut être mesuré au moyen de l'importance des variables calculée au moment de la phase d'apprentissage ou de validation du modèle. Ces derniers sont représentés dans la figure 5.4.

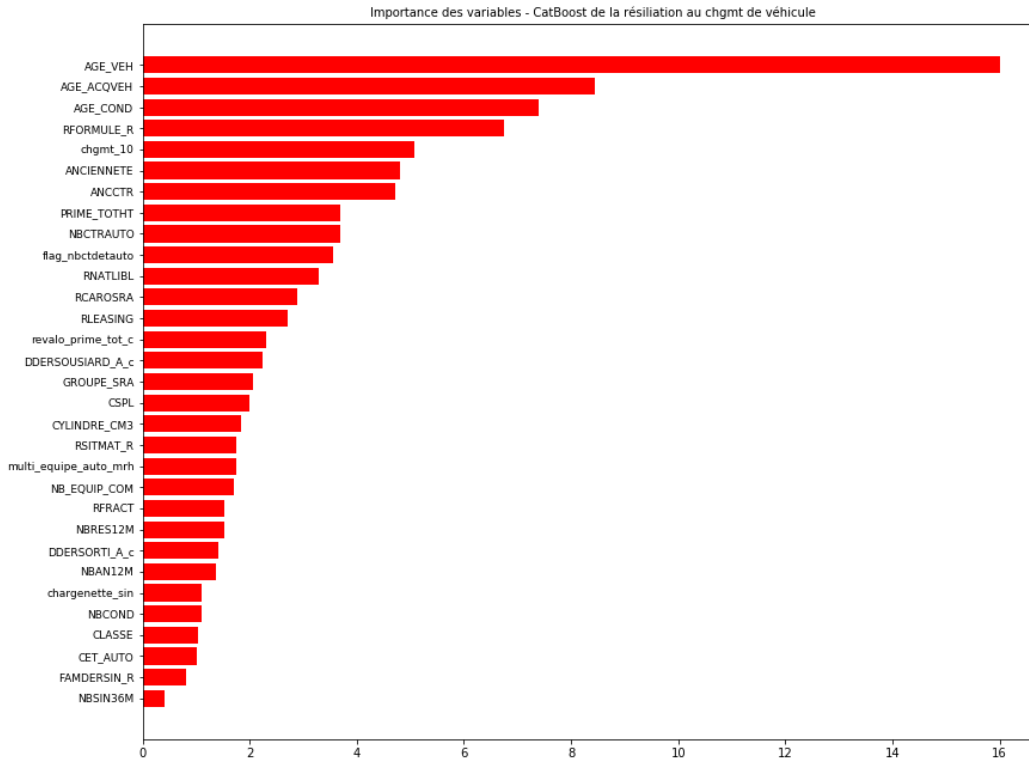


FIGURE 5.4 – Importance des variables pour le CatBoost de la résiliation au changement de véhicule

On constate ainsi que les 5 variables qui jouaient un rôle prépondérant dans la détermination du changement de véhicule jouent un rôle tout aussi important dans celle de la résiliation au changement de véhicule. Cela semble logique puisque les changements de véhicule ayant résilié sont intégrés dans la variable identifiant l'ensemble des changements de véhicule. Les facteurs influençant la décision de remplacer le bien assuré puis de clôturer son contrat n'ont donc aucune raison de fondamentalement changer par rapport aux facteurs expliquant uniquement le changement de véhicule.

La variable qui explique le mieux la résiliation pour ce motif reste l'âge du véhicule. Outre l'importance de cette variable dans la détermination du moment auquel il est opportun de changer son véhicule, cette dernière a un rôle prépondérant dans la tarification de la police d'assurance. En effet, le niveau de prime est fortement lié à la valeur du véhicule. On remarque ainsi une diminution de la prime à mesure que le véhicule vieillit et perd de sa valeur, baisse qui n'est pas proportionnelle à la décote puisque le risque de sinistre augmente parallèlement.

Si la durée de détention du véhicule et l'âge du conducteur sont des variables jouant un rôle tout aussi important que pour les modèles du changement de véhicule, on remarque que la formule souscrite pèse davantage dans la prédiction de la résiliation. En effet, c'est une composante majeure du niveau de la prime payée par le détenteur du contrat. Ainsi, on peut imaginer que les clients assurés selon les formules les moins complètes - généralement associées à des véhicules plutôt anciens -, verront leur prime fortement revalorisée par l'avenant de modification du risque lorsqu'ils décideront de changer de véhicule et de passer éventuellement à une formule plus couvrante. De même, les individus assurés avec la formule la plus complète et payant une prime déjà élevée pourraient vouloir se tourner vers la concurrence pour bénéficier de réductions sur le tarif au moment d'acquérir un nouveau bien.

En lien direct, on remarque que le niveau de la prime payée et le taux de revalorisation de cette dernière par rapport à l'année précédente sont également des facteurs impactant la décision de résilier son contrat.

Ensuite, pour l'explicabilité du modèle, on peut analyser l'effet de chaque variable prise individuellement sur la prédiction de la résiliation. Cette analyse au moyen des valeurs SHAP est présentée dans la figure 5.5.

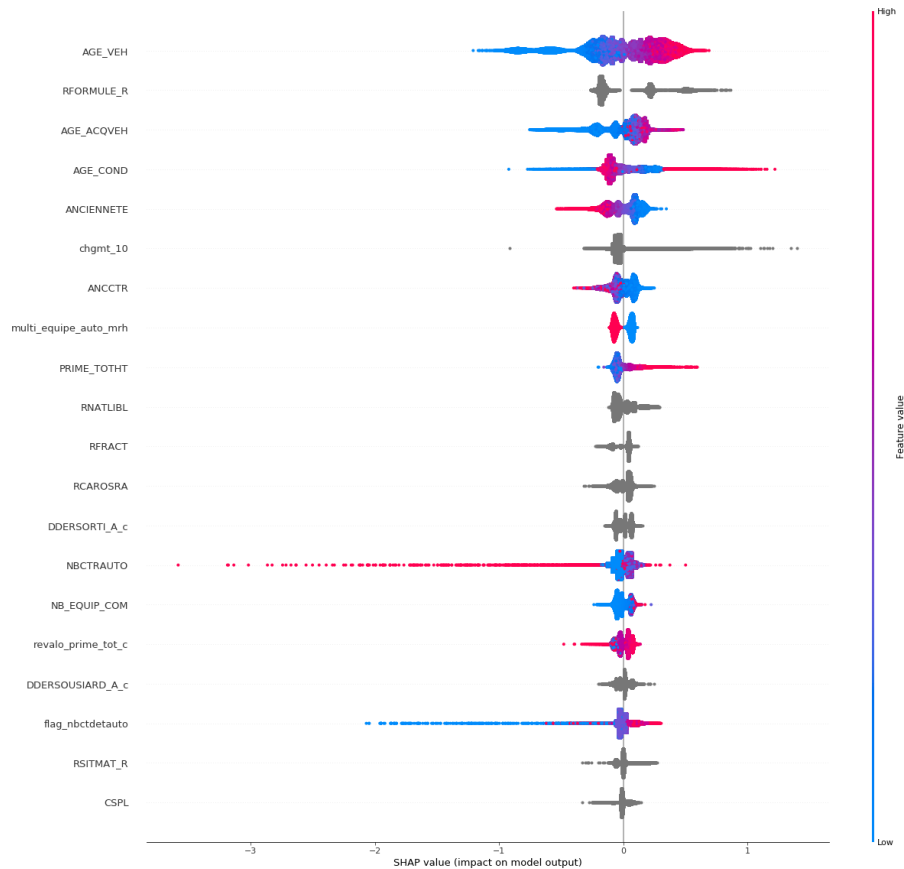


FIGURE 5.5 – Importance des variables au moyen des valeurs SHAP pour le CatBoost de la résiliation au changement de véhicule

Comme nous l'avons évoqué précédemment, on constate que la probabilité de changer de véhicule puis de résilier est croissante avec l'âge du véhicule. Mis à part les véhicules de collection ou les véhicules de luxe, les véhicules d'un âge avancé sont des véhicules pour lesquels l'utilité de les protéger avec des formules complètes est moindre. Ainsi, au moment du remplacement du bien assuré, il est fréquent d'observer une augmentation importante de la prime à la fois liée à la valeur du nouveau véhicule assuré et la montée en gamme de la couverture souscrite.

Parmi les autres éléments liés à la sensibilité au prix des individus et à la recherche d'un tarif plus intéressant comme motif de résiliation, on remarque que le niveau de la prime payée et sa revalorisation par rapport à l'année précédente jouent un rôle important.

Outre le changement de véhicule qui entraîne une augmentation moyenne de la prime de l'ordre de 10%, la manière dont le client perçoit le montant qu'il paye chaque année au titre de sa police d'assurance automobile influence sa volonté de rester chez le même assureur au moment de changer de véhicule. En effet, la résiliation des contrats automobiles ayant été facilitée par la loi Hamon, un client dispose d'une grande liberté pour comparer les tarifs proposés et faire jouer la concurrence entre les assureurs.

De cette façon, un client qui paye un montant élevé ou dont la revalorisation de la prime a été importante durant les années passées aura plus tendance à rechercher un tarif plus compétitif.

Par ailleurs, le niveau de revalorisation de la prime est un indicateur de la protection tarifaire dont bénéficie le client au sein de Generali. En effet, chaque année, des modèles d'élasticité au prix sont mis en place afin d'évaluer la probabilité que les clients résilient leur contrat en fonction des niveaux de revalorisation de leur prime. De cette façon, les clients à forte valeur (notion que nous expliquerons dans le chapitre suivant) peuvent être protégés lors de la mise en place de ces politiques de revalorisation puisque la compagnie veut s'assurer de les conserver en portefeuille. Ainsi, les clients dont la prime a été fortement revalorisée sont des clients qui sont moins intéressants pour l'assureur et qui seront moins susceptibles de se voir offrir des avantages : ils sont ainsi plus enclins à se tourner vers la concurrence afin de bénéficier de réductions à la souscription.

En combinant l'effet de la revalorisation de la prime et du montant payé par l'assuré, on peut donc supposer qu'une hausse des primes l'incite à comparer les offres des autres assureurs, le montant de référence servant pour la comparaison étant la prime de la police d'assurance actuelle.

Néanmoins, on remarque également qu'une revalorisation importante peut diminuer le score de résiliation au changement de véhicule. On peut attribuer cet effet à des polices ayant récemment souscrit de nouvelles garanties, augmentant ainsi le niveau de cotisation. Le moment où des garanties sont rajoutées au contrat peut être vu comme la souscription d'un nouveau contrat puisqu'avant cet avenant, le client a certainement demandé un devis auprès de Generali pour le comparer à ceux de ses concurrents. Ainsi, si le client est resté chez son assureur initial, c'est qu'il estime que les offres sont satisfaisantes, constat qu'il pourra réitérer au moment du changement de véhicule.

En outre, l'âge du conducteur a des effets contrastés sur la résiliation. Si la probabilité de changer de véhicule présentait une tendance plutôt décroissante avec l'âge, la résiliation au changement est différemment impactée par la jeunesse ou l'âge avancé du conducteur.

De manière générale, on constate que la jeunesse de l'assuré semble augmenter cette probabilité. Chez les particuliers, on sait que l'âge est primordial pour estimer le risque assurantiel sur les produits automobiles. Résultat d'un coefficient de réduction majoration plus élevé - en raison de la plus faible ancienneté du permis de conduire - et d'un risque technique plus fort, les jeunes individus payent généralement des primes plus élevées que la moyenne. De ce fait, de par leurs moyens financiers qui sont généralement moins élevés à cette période de leur vie - comme expliqué dans la théorie du cycle de vie de Modigliani et Ando [12] - et leur présence sur les plateformes digitales, ces derniers peuvent avoir plus tendance à rechercher les meilleurs tarifs pour assurer leur nouveau véhicule.

Pour les clients plus âgés, la résiliation peut également s'expliquer par la théorie du cycle de vie qui implique que ces individus disposent de moyens moins importants que durant leur période active.

De ce fait, ils pourraient être plus intéressés par la réduction de la dépense assurantielle grâce à des offres commerciales avantageuses proposées par des assureurs adverses. De la même façon, les clients plus âgés sont considérés comme plus risqués et pourraient ainsi faire l'objet de revalorisation de leur prime plus importante, faisant entrer en jeu la sensibilité au prix.

Concernant les caractéristiques de l'assuré, on constate que sa nature, sa situation maritale ou encore sa CSP jouent un rôle important dans la probabilité de résilier son contrat suite à un changement de véhicule. Néanmoins, la nature de ces variables - non ordinales - ne nous permet pas de pouvoir analyser explicitement l'effet de l'appartenance à une classe sur ce score de fragilité. Pour cela, il serait nécessaire d'analyser les classifications effectuées par le modèle et récupérer les scores affectés à chacune des classes dans les différentes branches ; ce qui semble peu réalisable au vu de la taille des arbres générés.

En définitive, les différentes modélisations testées ont abouti à la création d'un score de fragilité conçu de façon à détecter les résiliations parmi les changements de véhicule.

Tout comme la mise en place du score de fragilité au changement de véhicule, les problématiques inhérentes à la base de données nous ont conduit à recourir à des méthodes de classification pour expliquer le phénomène étudié. L'analyse de leurs performances nous ont alors poussé à sélectionner une fois de plus un modèle CatBoost qui semble donc particulièrement approprié pour prédire les différentes formes de changement de véhicule.

Les qualités de discrimination du modèle retenu semblent tout à fait acceptables pour la mise en place d'une politique de rétention car ce dernier performe bien sur un périmètre plutôt large des clients les plus fragiles. Par ailleurs, l'implémentation des méthodes de validation croisée nous a permis de s'assurer que les résultats obtenus étaient robustes.

En recoupant l'analyse descriptive que nous avons fait des facteurs influençant la résiliation au changement de véhicule et les résultats du modèle en termes d'interprétabilité et d'explicabilité, nous avons pu établir les profils des individus fragiles à la clôture de leur contrat. Les caractéristiques qui influencent la résiliation sont sensiblement les mêmes que celles que nous avons relevé pour expliquer toutes les formes de changement de véhicule, auxquelles s'ajoutent des éléments liés au niveau de tarification et à la sensibilité au prix des agents économiques.

Tout l'enjeu de l'intégration de ce score dans les systèmes réside désormais dans la définition du ciblage des politiques de rétention. En effet, la mise en place de ces campagnes commerciales présente des limites à la fois opérationnelles et financières. Il est donc nécessaire de déterminer le périmètre des clients auxquels des offres seront transmises.

L'intégration du score dans les systèmes et le choix des clients ciblés, en fonction de leurs caractéristiques et ce qu'ils apportent à l'entreprise, feront l'objet du chapitre suivant.

## 6 Chapitre 6 : Intégration du score dans les systèmes

Ce chapitre vise à décrire les procédures d'intégration du score de résiliation au changement de véhicule dans les systèmes dans l'optique de mettre en place des politiques de rétention pour les clients qui seront éligibles à ces offres.

Tout d'abord, la construction du score pour une nouvelle année est une étape relativement aisée. En effet, pour l'année pour laquelle nous souhaitons identifier les clients susceptibles de changer de véhicule et de résilier leur contrat d'assurance, nous récupérons les variables retenues lors de la construction du modèle de résiliation dans le chapitre précédent. Les caractéristiques ainsi récupérées correspondent à la dernière photographie du client, de son véhicule et de son contrat qui sont à notre disposition dans les différents systèmes d'information.

Bien entendu, le périmètre des clients doit être le même que celui que nous avons défini au début de ce mémoire, à savoir les contrats mono-véhicules, 4 roues et appartenant au réseau des agents de Generali. A défaut, nous nous exposerions à des prédictions biaisées puisque les caractéristiques prises en compte n'auraient pas été intégrées lors de la calibration des modèles.

Une fois ces données récupérées, nous appliquons les mêmes traitements qui avaient été effectués sur la base d'étude, comme décrits dans le chapitre 3 : ajout de données externes, création de variables pour retracer le parcours client, suppression et traitement des valeurs aberrantes ou manquantes, regroupement des classes, etc. . .

Cette étape justifie le choix de restreindre la dimension de la base de modélisation au moment de la phase d'apprentissage puisqu'en plus d'améliorer la qualité des modèles construits, un nombre réduit de variables facilite la création de la base pour la prédiction.

Une fois la base de données pour l'année d'intérêt constituée, il est alors possible de calculer la probabilité que chaque client change de véhicule et donne un terme à sa police d'assurance automobile chez Generali. C'est en appliquant les poids du modèle de résiliation sur les nouvelles données qu'on aboutit ainsi à notre score de fragilité.

Dans le chapitre précédent, nous avons identifié les caractéristiques conduisant les clients du portefeuille à être plus ou moins sensibles à la résiliation suite à un changement de véhicule. Il peut également être intéressant de regarder si des groupes de clients homogènes en termes de fragilité se dégagent afin d'avoir un premier aperçu sur les profils qui pourraient être touchés par des actions de rétention.

Pour la suite de ce chapitre, nous effectuerons nos analyses sur la base de validation que nous avons séparé par validation non croisée lors de la construction des modèles. Plutôt que de prédire la résiliation pour une nouvelle année et d'effectuer nos analyses sur les résultats obtenus, le recours à la base de validation se justifie par la possibilité de comparer les prédictions du modèle avec les comportements observés dans la réalité et ainsi d'avoir un ordre de grandeur sur le nombre de clients qui auraient pu être intéressés par une offre commerciale et les primes associées.

## 6.1 Identification des groupes de clients fragiles à la résiliation

Afin d'identifier des groupes de clients homogènes en fonction de leur fragilité à la résiliation au changement de véhicule et de leurs profils, nous allons utiliser un algorithme non-supervisé de clustering relativement répandu : les  $K$ -moyennes ( $K$ -means en anglais).

En apprentissage automatique, le clustering est une discipline ayant pour objectif de séparer les données en groupes homogènes ayant des caractéristiques communes. Il est particulièrement utilisé en marketing où l'on cherche souvent à segmenter les bases de clients pour détecter des comportements particuliers.

Étant donné des points et un entier  $K$  dont nous définissons la valeur, l'algorithme vise à diviser les données en  $K$  groupes, appelés clusters, homogènes et compacts. Pour plus de détails, les explications théoriques de cet algorithme sont données dans le chapitre 2.

Pour appliquer cette méthode à nos prédictions et ne pas surcharger l'analyse, nous allons sélectionner seulement quelques variables de notre base de validation pour ensuite construire différents clusters et analyser leurs caractéristiques.

Ainsi, nous retiendrons les variables qui étaient les plus importantes pour la construction des arbres de classification du modèle CatBoost de la résiliation au changement de véhicule, et en priorité les variables continues qui sont plus facilement interprétables. De cette façon, en plus du score de fragilité à la résiliation, l'âge et le groupe SRA du véhicule, le nombre de contrats automobile détenus, le nombre de sinistres survenus au cours des 36 derniers mois, l'âge du conducteur, l'ancienneté du contrat, la prime totale hors taxes ou encore la souscription d'une formule tous risques seront utilisés pour créer des sous-populations homogènes.

Pour déterminer le nombre de clusters idéal, la méthode du coude (*elbow method* en anglais) est une méthode largement utilisée. Elle s'appuie sur la notion d'inertie intra-classe qui se définit ici comme la somme des distances euclidiennes entre chaque point et son centroïde associé. Plus le nombre initial de clusters est élevé, plus on réduit l'inertie : les points ont plus de chance d'être proche d'un des multiples centroïdes.

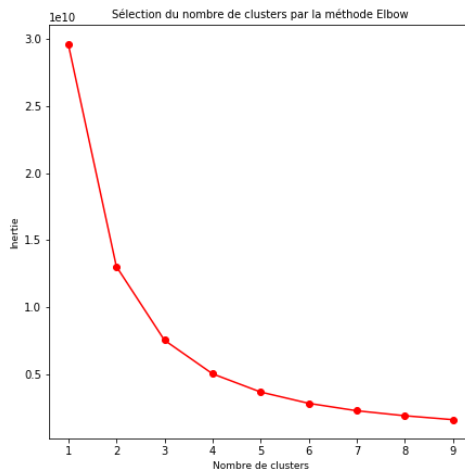


FIGURE 6.1 – Sélection du nombre de clusters idéal par la méthode du coude

En analysant la figure 6.1, on constate que le nombre de clusters optimal se situe entre 3 et 4 puisque la diminution de la variance (inertie) intra-classe commence à stagner à partir de ce moment-là.

En retenant 4 clusters et en appliquant cet algorithme à notre base de validation avec les variables sélectionnées, on obtient les résultats suivants :

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Score de fragilité	0.4061	0.3891	0.4383	0.3353
Age du véhicule	13.59	6.51	5.01	8.21
Age du conducteur	56.35	48.77	47.93	52.12
Formule L3	0.3908	0.9075	0.9548	0.8167
Nombre de sinistres 36M	0.126	0.329	0.461	0.218
Ancienneté du contrat	7.36	6.03	4.99	7.27
Prime HT	280.44	831.44	1493.15	514.43
Nombre contrats auto	2.03	2.09	2.29	1.91
Groupe SRA	29.65	31.91	33.39	30.74
Leasing	0.0103	0.1494	0.2774	0.056

TABLE 6.1 – Résultats de l'algorithme  $K$ -means

Deux groupes en particulier attirent notre attention, de par le niveau de leur score de fragilité à la résiliation au changement de véhicule et leurs caractéristiques qui sont très distinctes :

1. Le premier groupe (cluster 3) présente une appétence moyenne à la résiliation pour changement de véhicule plutôt élevée par rapport aux autres clusters. Il regroupe des véhicules relativement récents et haut de gamme, en témoigne la moyenne du groupe SRA associé qui indique que les véhicules sont puissants et donc avec un prix plus élevé que la moyenne. Les clients qui composent ce groupe sont plus jeunes que la population moyenne du portefeuille et ont souscrit un contrat depuis quelques années seulement. Les biens assurés étant des biens à forte valeur, la quasi-totalité de cette sous-population est assurée avec la formule maximale (95.5%). La combinaison d'un bien assuré à forte valeur et de la souscription d'une assurance tous risques établit la prime moyenne à 1493 €, ce qui représente des clients qui pourraient s'avérer être intéressants pour Generali.
2. Le second groupe (cluster 1) présente également une fragilité à la résiliation plutôt élevée par rapport aux autres groupes mais présente des caractéristiques presque inverses au groupe précédent. En effet, on constate tout d'abord que la prime moyenne de ce cluster est largement inférieure à celle du premier groupe tandis qu'une faible part de ces contrats, en moyenne beaucoup plus anciens, est assurée avec la formule maximale (39.1%). Les biens assurés par ces clients sont de plus faible valeur, en atteste l'âge moyen des véhicules autour des 13.5 ans et leur moindre puissance correspondant à un groupe SRA moins élevé. En outre, les individus sont en moyenne beaucoup plus âgés. Si ces clients semblent moins intéressants en termes de chiffres d'affaires (primes), leur forte probabilité de changer de véhicule dans l'année et l'augmentation de la prime qui pourrait résulter de l'acquisition d'un véhicule plus récent et plus onéreux, en font des clients qui pourraient également être d'intérêt.



En termes de répartition sur la base de validation, le premier groupe (cluster 3) ne représente que 2.4% des individus tandis que le second groupe (cluster 1) en représente 43.7%. Les deux autres clusters identifiés (clusters 2 et 4) qui présentent des tendances moins marquées, mais pas moins intéressantes, comptent pour respectivement 15% et 38.9% des clients de la base.

Cette analyse par clustering nous a permis de cerner les différents profils qui pourraient être concernés par des campagnes de rétention au changement de véhicule.

Cependant, si l'attribution d'un score de fragilité à chacun des clients permet d'estimer leur probabilité de résilier, celui-ci n'est pas révélateur de l'intérêt pour Generali de les conserver. En effet, si un client risqué et peu rentable se voyait attribué un score de fragilité élevé, il ne serait pas pertinent que ce dernier soit éligible à une offre commerciale.

Dans cette optique, afin de cibler les clients fragiles auxquels la compagnie devrait soumettre une offre, nous allons intégrer un indicateur calculé au sein de l'équipe dans laquelle j'ai évolué au cours de ce stage : la valeur client.

## 6.2 Intégration de la valeur client

La valeur client est un indicateur synthétisant la profitabilité actuelle et future de l'ensemble des contrats appartenant à la sphère d'un client, qui est définie comme le foyer fiscal du client (conjoint, enfants, parents) et/ou l'entreprise s'il est assuré en tant que professionnel.

Cet indicateur intègre deux valeurs différentes :

1. **La valeur actuelle** qui est fonction de la rentabilité technique du contrat, des trajectoires de majorations et des commissions.
2. **La valeur future** qui est basée sur la durée en portefeuille, l'appétence à se multi-équiper ou encore sur la proactivité des réseaux de distribution à multi-équiper leurs clients.

Par exemple, la valeur des contrats automobiles va dépendre de certaines caractéristiques du contrat (formule souscrite), de l'offre proposée (écart au tarif) ou du client (âge, coefficient de réduction-majoration). L'ensemble de ces éléments pris en compte conjointement va permettre de calculer la valeur. Sur les autres produits, les méthodes de valorisation et les variables prises en compte diffèrent.

Chaque contrat est ainsi valorisé par année, en tenant compte d'hypothèses économiques et de lois de chute. Les valeurs obtenues pour chaque année de détention sont ensuite agrégées pour obtenir une valeur unique par contrat. Par la suite, la valeur de chacun des contrats est agrégée au niveau individuel pour chaque client, en tenant compte de l'ensemble de ses liens : on obtient la valeur du client au niveau de sa sphère.

Dans la construction de cet indicateur, la survenance d'un sinistre n'est pas un fait rédibitoire. En effet, un client à valeur peut avoir des sinistres ; autrement un client sinistré serait quasiment toujours à faible valeur. Un client à valeur est un client qui va rester longtemps et se multi-équiper, quelle que soit sa sinistralité actuelle. De la même façon, un client à faible valeur peut aussi être un client sans sinistre : sa faible valeur provient souvent d'un ou plusieurs contrats mal tarifés ayant un tarif inférieur au tarif technique.

L'avantage de cette méthode de valorisation est qu'elle permet de refléter la durée des clients. En effet, c'est la composante principale de la valeur client : projeter une marge sur plusieurs années a beaucoup plus de poids que la marge technique à un instant donné. C'est ce qui explique les cibles de la compagnie dont l'appétence au multi-équipement et la durée sont plus élevées : propriétaires de leur logement, familles installées, formule maximale, coefficient de réduction-majoration à 0.5, ...

Les clients sont ensuite classés par un système d'étoiles en fonction de la segmentation de la valeur de la sphère client. Cette classification permet de décomposer le portefeuille en fonction de la rentabilité estimée des assurés.

De cette façon, les clients à forte valeur (5 étoiles) sont souvent des clients multi-équipés ou avec un potentiel de multi-équipement, avec de nombreux liens et payant un prix au-dessus du tarif technique. A l'inverse, les clients à faible valeur (1 étoile) sont souvent mono-équipés et payent un tarif qui ne couvre pas leur risque.

Il existe un écart de rentabilité et de résiliation significatif entre les clients à valeur et les autres. Ces derniers concilient les intérêts de la compagnie et des réseaux de distribution car ils sont aussi vecteur de stabilité des commissions.

L'intégration de ce score dans les systèmes a plusieurs vocations :

1. **La gestion des rabais commerciaux** : permettre une déformation du portefeuille vers les clients cibles tout en limitant les rabais commerciaux sur les clients moins rentables.
2. **L'optimisation des processus de revalorisation** :
  - Pour les clients à forte valeur : protection face à des majorations excessives sur l'ensemble de leurs contrats
  - Pour les clients à valeur intermédiaire : analyse de la sensibilité au prix via des données comportementales (modèles d'élasticité)
  - Pour les clients à faible valeur : sur-majoration sur les contrats et interdiction aux intermédiaires (agents, courtiers, ...) de protéger ces clients
3. **La surveillance du portefeuille** : identification des clients avec des contrats déficitaires ou des intermédiaires "destructeurs" de valeur. Cela va permettre d'améliorer la rentabilité globale du portefeuille, diminuer le besoin de revalorisation et augmenter le rabais moyen sur les clients à forte valeur. Cela conduit principalement à la limitation des actions commerciales envers les clients à faible valeur et à la mise en place d'actions dans le but de stimuler les transferts des clients à moyenne valeur (3 étoiles) vers les clients à forte valeur (4 et 5 étoiles). Ces dernières ont trait au multi-équipement du client et de sa sphère, à la limite de la défense de ces clients au renouvellement afin de permettre un rehaussement du tarif payé, etc...
4. **Fidélisation des clients à valeur** avec l'utilisation de tous les leviers de rétention possibles :
  - Création de services "premiums"
  - Parcours de gestion de sinistres accéléré
  - Personnalisation du service client (par exemple, fluidification des lignes téléphoniques)
  - etc...

Dans ce cadre, il est clair que clients classifiés comme "à faible valeur" ne pourront faire l'objet d'offres commerciales pour la rétention au changement de véhicule. En effet, ces clients ne sont pas profitables pour Generali et leur proposer des rabais ne ferait que diminuer leur valeur puisque le tarif s'éloignerait davantage du tarif technique. De plus, cela pourrait également augmenter leur durée, élément pénalisant pour l'assureur puisqu'il serait avantageux pour la compagnie que ce type de client quitte son portefeuille.

Parce que la valeur est un indicateur particulièrement révélateur de la rentabilité d'un client pour son assureur, nous intégrons cette donnée à notre base de validation, ainsi qu'aux bases qui serviront pour la prédiction du score de fragilité pour une nouvelle année.

Nous allons désormais déterminer les critères que nous serons susceptibles d'utiliser pour cibler les clients éligibles aux campagnes de rétention.

### 6.3 Détermination du ciblage des politiques de rétention

Pour qu'une opération commerciale de type rétention soit profitable à la compagnie, il est nécessaire que les clients soient correctement ciblés afin que le retour sur investissement soit positif et que l'offre soit suffisamment attractive pour intéresser et fidéliser les clients.

En reprenant le principe de la valorisation client que nous avons détaillé dans la partie précédente, faire bénéficier des clients "destructeurs de valeur" pourraient être préjudiciable à long terme pour la compagnie autant par la réduction sur le tarif qui serait accordée que par le gain de durée qui conduirait les clients à rester plus longtemps dans le portefeuille.

Sur notre base d'étude, la répartition du portefeuille est largement orientée vers les clients 3 et 4 étoiles. Les clients 5 étoiles, clients très profitables et à fort potentiel, ne représentent ainsi que 4% du portefeuille.

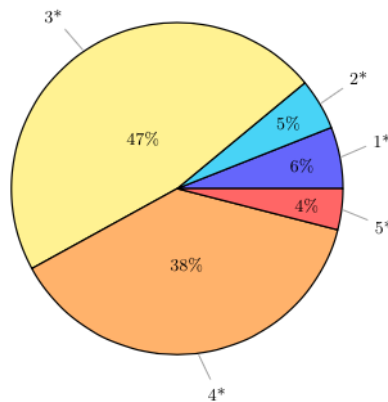


FIGURE 6.2 – Répartition de la segmentation de la valeur client dans le portefeuille

De cette façon, les clients classés 1 et 2 étoiles seront exclus des campagnes de rétention au changement de véhicule car ils sont identifiés comme "destructeurs de valeur" et donc peu rentables pour l'assureur. Subsistent les clients 3, 4 et 5 étoiles.

Un autre élément à prendre en considération pour déterminer la cible des offres commerciales est la disponibilité des moyens de contact. En effet, lors de la souscription d'un contrat, il est toujours demandé aux clients s'ils acceptent d'être contactés par la compagnie et selon quels canaux : courrier, e-mail, téléphone, SMS, etc. . . . Dans notre cas, le refus d'être contacté sera un élément discriminant pour le bénéfice de la campagne de rétention puisqu'il ne serait pas possible d'entrer en contact avec le client pour lui soumettre des propositions.

La prise en compte conjointe du score de fragilité, de la valeur client et de l'accord d'être contacté nous permet alors de définir le ciblage des actions de rétention. Ainsi, en fonction du nombre total de clients auquel la compagnie décidera de s'adresser, les deux tiers d'entre eux seront des clients classés entre 3 et 5 étoiles, contactables par au moins un canal et présentant les plus forts scores de fragilité. Le tiers restant des clients à contacter concernera les clients classés entre 4 et 5 étoiles, également contactables par au moins un canal et présentant les scores de fragilité les plus élevés.

Une fois la définition des clients éligibles à une offre de rétention établie, il est nécessaire de définir le nombre de clients qui pourront en faire l'objet. Le choix du périmètre de la campagne peut se faire au moyen d'une estimation des coûts et des gains.

#### **6.4 Estimation des coûts et des gains d'une campagne de rétention**

L'estimation des coûts et des gains d'une campagne commerciale, sous certaines hypothèses, permet d'avoir une appréciation de sa rentabilité.

Pour ce faire, il sera nécessaire de formuler certaines hypothèses sur le coût d'un contact, le taux de succès de contact des clients, le taux de succès de la campagne de rétention ou encore sur la durée additionnelle générée.

Afin de tenir uniquement compte de l'effet de rétention des primes sur le portefeuille, nous ne tiendrons pas compte de l'augmentation de la prime induite par un changement de bien assuré ou du rabais qui pourrait être accordé sur le tarif par l'offre commerciale.

Pour commencer notre analyse, le premier élément à considérer est le taux de précision du modèle de résiliation en fonction des quantiles de probabilités prédites. En effet, cette métrique nous indique le nombre de clients qui seraient concernés par la proposition parmi l'ensemble des clients contactés. Autrement dit, c'est un indicateur d'efficacité du ciblage effectué.

Notre base de validation comportant environ 400 000 clients, une précision de 26.5% sur les 1% des clients les plus fragiles à la résiliation au changement de véhicule signifie que si 1% des clients du portefeuille étaient contactés, soit 4 000 personnes, environ 1 060 seraient réellement susceptibles de changer de véhicule dans l'année et de clôturer leur contrat pour ce motif.

En fonction du nombre de clients contactés, il faut également chiffrer le coût opérationnel de la mise en place de la campagne de rétention. Ainsi, des hypothèses sont à formuler sur le coût journalier d'un chargé d'opérations et sur le nombre journalier de dossiers qui pourraient être traités. Ces deux hypothèses permettent de calculer le coût de la campagne.

Il est alors nécessaire d'en estimer les gains.

En fonction du nombre de clients contactés et des clients réellement concernés par les offres, on formule une hypothèse sur le taux de réussite du contact. En effet, bien que les clients faisant partie de notre périmètre aient tous accepté d'être contactés par au moins un canal, rien n'assure qu'ils répondent aux différentes sollicitations. Cette hypothèse réduit ainsi le nombre de clients qui pourraient être sensibles à notre offre.

Par ailleurs, les critères poussant un client à résilier son contrat d'assurance automobile au moment de changer de véhicule sont multiples : tarif trop élevé, tarif plus attractif chez la concurrence, expérience client avec Generali, insatisfaction du traitement des sinistres, etc. . . De ce fait, il est clair que les rabais proposés sur le tarif ne permettront pas de retenir l'ensemble des clients auxquels une offre a été soumise. Une hypothèse supplémentaire doit alors être formulée sur le taux d'efficacité de la campagne, taux qu'il est plus judicieux de fixer bas afin de ne pas surestimer les gains. Autrement dit, ce taux donne le pourcentage des clients qui seront sensibles à l'offre formulée et qui décideront de modifier leur contrat par un avenant de changement de risque plutôt que de souscrire un nouveau contrat chez la concurrence.

La combinaison de toutes ces hypothèses nous donne le nombre final estimé de clients qui seront retenus grâce à cette campagne de rétention. A partir de ce chiffre, nous estimons le montant des primes qui ont ainsi pu être conservées. Cette estimation donne le gain, ou du moins l'absence de pertes, en termes de primes sur une année.

En outre, il est acceptable d'admettre qu'un client changeant de véhicule et bénéficiant de réductions commerciales sur son tarif restera plus d'une année en portefeuille. Une hypothèse supplémentaire sur la durée additionnelle générée est donc à formuler. Elle indique le nombre moyen d'années durant lesquelles le client va rester en portefeuille suite à la modification de son contrat pour changement de bien assuré. A la différence des contrats en assurance-vie, la durée sur la branche automobile n'est pas particulièrement élevée. De ce fait, nous pourrions formuler des hypothèses de gains de durée situés entre 2 et 4 ans afin de ne pas être trop optimistes. Cela nous permet alors d'estimer le gain total de primes sur les années où le client restera supposément en portefeuille.

Enfin, le gain de primes ainsi estimé n'est pas un gain net puisqu'il faut prendre en compte le ratio combiné qui s'établit aux alentours des 95% sur ce portefeuille. Ainsi, sur l'ensemble des primes conservées, seulement 5% du montant représentera un bénéfice pour la compagnie. Ce calcul peut s'effectuer pour une année et pour le nombre d'années de durée supplémentaire.

Ainsi, en comparant la marge technique dégagée en fonction de la durée additionnelle et des coûts de la campagne, on peut estimer les bénéfices que pourraient engendrer la mise en place d'une campagne de rétention au changement de véhicule.

Pour déterminer le nombre de clients à contacter permettant d'optimiser les bénéfices, il est donc nécessaire d'évaluer les coûts et les gains de la compagnie en fonction du nombre de clients contactés, de la précision du modèle sur le périmètre considéré tout en tenant compte des différentes hypothèses que nous avons formulées.

Par ailleurs, une étude de sensibilité doit également être menée afin d'identifier les valeurs associées à chacune des hypothèses qui conduirait la campagne à ne plus être rentable. Les principales variables dont nous pouvons faire varier les valeurs, notamment parce qu'elles sont difficiles à estimer, sont le taux de succès de la campagne et le gain de durée.

## 6.5 Pistes d'amélioration

La quantification et l'estimation des coûts et des gains des campagnes commerciales s'appuient sur des hypothèses qui peuvent être discutables tant que ces dernières n'ont pas été mises en place et que leurs effets sur la rétention n'ont pas été constatés.

Cependant, l'un des paramètres majeurs de cette estimation est la précision du modèle de résiliation. Améliorer la qualité de prédiction de ce dernier permettrait d'affiner le ciblage en étant plus précis sur la fragilité des clients à la clôture de leur contrat pour changement de véhicule. De cette façon, les coûts de mise en place des opérations commerciales s'en trouveraient réduits puisqu'un pourcentage plus important de clients contactés seraient réceptifs aux offres formulées.

Pour améliorer la qualité des modèles, plusieurs pistes pourraient être explorées. Ces dernières nécessiteraient d'apporter des informations supplémentaires par rapport à ce qu'il avait été possible de récupérer au cours des différentes analyses et modélisations.

Concernant les caractéristiques des biens assurés, une approximation du nombre de kilomètres parcourus par le véhicule permettrait d'avoir une vision plus fine sur son état général.

Par ailleurs, comme nous l'avons vu, le type du bien assuré est un élément jouant sur la probabilité de le remplacer. A partir des données SRA, nous avons pu récupérer le dernier tarif connu à neuf du véhicule. Cependant, cette donnée ne nous donne aucune indication sur son prix de revente éventuel et sur l'apport pécuniaire qu'elle pourrait engendrer. A cette fin, il pourrait être intéressant d'utiliser des méthodes de *web-scraping* pour récupérer ces informations sur des sites de ventes de véhicules d'occasion afin d'avoir un ordre d'idée sur la valeur actuelle du véhicule.

De plus, les données relatives aux signaux clients pourraient être exploitées afin d'être intégrées dans les modèles de prévision. On pense notamment à la demande du relevé d'informations qui était une donnée insuffisamment identifiée au moment de la construction des modèles. Cette dernière permettrait d'identifier les clients qui souhaiteraient résilier leur contrat pour changer d'assureur puisque le relevé d'information est l'une des pièces requises lors de la souscription d'un contrat d'assurance automobile.

Enfin, une autre donnée dont la fiabilité n'était pas suffisante pour être intégrée dans les modélisations mais qui pourrait faire l'objet d'un traitement dans les systèmes est la simulation d'un tarif pour un nouveau véhicule. Dans les bases internes, il serait intéressant d'identifier les clients simulant un avenant de changement de véhicule en récupérant le code SRA du véhicule testé et en le comparant avec le code SRA du véhicule assuré ; une différence entre les deux indiquant nécessairement que le client a simulé des tarifs pour un véhicule qui n'est pas (encore) le sien.

La récupération ou la mise à disposition de telles variables seraient intéressantes pour mieux cibler les individus susceptibles de changer de véhicule ou de résilier pour ce motif. L'intégration de ces données dans les différentes modélisations réduirait ainsi le seuil d'erreur des modèles construits et rendrait les campagnes commerciales moins coûteuses car il serait nécessaire de contacter un nombre inférieur de clients pour le même résultat en termes de rétention (en supposant que le taux de succès de la campagne reste fixe).

Par ailleurs, l'un des reproches qui est généralement émis à l'encontre des méthodes de classification, notamment aux nouveaux algorithmes tels que le CatBoost ou le XGBoost, est la difficulté de l'interprétation de leurs résultats. En effet, pour certaines variables, et plus particulièrement les variables catégorielles, il reste difficile d'expliquer leurs effets sur la construction du score de fragilité. De la même façon, l'interaction entre les variables est difficile à capter lors de l'analyse des résultats obtenus. Approfondir l'analyse ou trouver des modélisations plus facilement interprétables et s'adaptant aussi bien aux données pourrait également être une des pistes d'amélioration.

## Conclusion

Pour conclure, ce mémoire a eu pour objectif d'analyser et prédire deux comportements observables au cours de la durée de vie d'une police automobile : le changement de véhicule et la résiliation qui peut en résulter.

Moment clé dans la vie d'un contrat d'assurance automobile, le changement de véhicule entraîne une augmentation de la prime payée puisqu'il engendre généralement une augmentation de la valeur du bien assuré et une hausse du taux de couverture. Parce que la sensibilité au prix des individus est un élément non négligeable, ce comportement fragilise la durée de vie des contrats et est à l'origine d'une part importante des résiliations.

A l'heure où la concurrence sur le marché assurantiel se fait de plus en plus pressante et que les clients ont la possibilité de comparer les produits proposés par les différents acteurs, il est nécessaire de pouvoir anticiper cette décision et d'être proactif afin de s'adapter aux nouvelles attentes des assurés et les fidéliser.

Cela constitue l'un des principaux défis actuariels que doivent relever les assureurs pour maintenir la rentabilité de leur portefeuille et amortir les coûts de la souscription des contrats, processus coûteux qui doit être compensé par leur duration.

Dans cette optique, prédire le changement de véhicule et mener des actions de rétention appropriées pourrait leur permettre de diminuer leur taux de résiliation sur cette branche de leur portefeuille.

Ce sujet comportemental est particulièrement dur à anticiper puisqu'il correspond à des prises de décision relativement hétérogènes, variant en fonction de la nature du bien assuré, de l'environnement de vie de l'individu, de ses caractéristiques individuelles, etc... La prédiction de la résiliation au changement de véhicule reprend ces difficultés auxquelles s'ajoutent des facteurs liés à la sensibilité au prix des individus, la satisfaction client, etc...

Le dimensionnement de deux scores de fragilité au changement de véhicule et à la résiliation correspondante s'est fait au moyen de méthodes de classification, méthodes qui s'adaptaient le mieux à notre problématique et à la structure de notre base de données. En utilisant des métriques de scores permettant de refléter réellement le pouvoir prédictif des modèles, nous avons abouti à la conclusion que le CatBoost proposait la meilleure modélisation pour les deux sujets.

La qualité de discrimination obtenue permet d'atteindre des seuils d'erreur convenables tout en restant fidèle à la réalité de la distribution des classes sur les échantillons d'apprentissage et de validation. L'analyse des résultats permet de mettre en avant des variables significatives cohérentes pour expliquer l'un ou l'autre des phénomènes étudiés et d'identifier des profils de clients fragiles.

Les modèles ainsi calibrés pourraient faire l'objet d'améliorations afin d'augmenter leur précision. En effet, bien que le maximum d'informations ait été recueilli pour caractériser le client, le bien assuré, le contrat ou suivre le parcours client, certaines informations qui auraient été d'intérêt n'étaient pas disponibles au moment de l'étude. Résoudre ce problème de qualité de données et étoffer la base de modélisation pourrait par conséquent réduire les seuils d'erreur de prédiction.



Intégrer le score de fragilité à la résiliation au changement de véhicule dans les systèmes permettrait, pour une période donnée, d'identifier les clients susceptibles de quitter le portefeuille à la suite de l'acquisition d'un nouveau bien et de mettre en place des actions de rétention.

Dans ce cadre, améliorer la précision des modèles serait également un moyen d'améliorer le rendement des campagnes commerciales mises en place. En effet, le ciblage des clients sensibles à la résiliation serait affiné et les moyens déployés pour soumettre une offre s'en trouveraient réduits puisqu'il serait nécessaire de contacter un nombre inférieur de personnes pour toucher un nombre identique de clients réellement concernés.

Par ailleurs, la définition des profils de clients éligibles à ces offres commerciales s'appuie sur des indicateurs permettant d'évaluer la rentabilité des assurés pour la compagnie et sur la possibilité de les contacter. La détermination du nombre de clients qui pourront bénéficier de cette offre peut se faire au moyen d'une optimisation des coûts et des gains de la campagne ; le nombre de clients retenu permettant de maximiser le résultat net.

Outre ces critères d'éligibilité, il serait intéressant de croiser le score de fragilité avec une analyse d'élasticité au prix des clients lors des avenants afin d'identifier plus précisément les clients qui seraient susceptibles de rester chez leur assureur s'ils se voyaient proposer un rabais commercial sur le tarif. Ce complément d'analyse pourrait également permettre de limiter le nombre de clients à contacter et augmenter le taux de succès des actions de rétention tout en réduisant leurs coûts.

Finalement, ce mémoire s'inscrit dans un contexte professionnel où il s'avère indispensable de pouvoir expliquer et justifier clairement les résultats obtenus. L'analyse du comportement des clients et les différents profils qui en ressortent doivent être facilement communicables aux différents services impliqués dans la mise en place des actions de rétention.

La réalisation de cette étude sur un sujet comportemental tel que le changement de véhicule ouvre également des perspectives sur d'autres domaines d'application. Par exemple, une étude de la rétention client au déménagement en assurance multirisques habitation (MRH) pourrait reprendre la même méthodologie utilisée tandis que des actions de rétention seront à mener sur la branche santé suite à l'entrée en vigueur de la loi du 1er décembre 2020 qui étend les modalités de résiliation de la loi Hamon aux complémentaires santé.

## Références

- [1] **Fédération Française de l'Assurance**, *Rapport annuel 2020*, 2021
- [2] **BELLMAN** Richard Ernest, *Adaptive Control Processes*, Princeton University Press, 1961
- [3] **BREIMAN** Leo, **FRIEDMAN** Jerome H., **OLSHEN** Richard A. et **STONE** Charles J., *Classification and Regression Trees*, Chapman and Hall, 1984
- [4] **BREIMAN** Leo, *Bagging predictors*, Machine learning, 1994
- [5] **BREIMAN** Leo, *Random forests*, Machine learning, 2001
- [6] **FRIEDMAN** Jerome H., *Greedy function approximation : A gradient boosting machine*, The Annals of Statistics, 2001
- [7] **HOERL** Arthur E. et **KENNARD** Robert W., *Ridge Regression : Applications to Non-orthogonal Problems*, Technometrics, 1970
- [8] **JARO** Matthew A., *Advances in record linking methodology as applied to the 1985 census of Tampa Florida*, Journal of the American Statistical Society, 1989
- [9] **LUNDBERG** Scott M. et **LEE** Su-In, *A unified approach to interpreting model predictions*, Advances in Neural Information Processing Systems, 2017
- [10] **LUNDBERG** Scott M., *Consistent feature attribution for tree ensembles*, 2017
- [11] **LUNDBERG** Scott M., **ERION** Gabriel G. et **LEE** Su-In, *Consistent individualized feature attribution for tree ensembles*, 2019
- [12] **MODIGLIANI** Franco et **ANDO** Albert, *The 'life-cycle' hypothesis of saving : aggregate implications and tests*, American Economic Review, 1963
- [13] **PROKHORENKOVA** Liudmila, **GUSEV** Gleb, **VOROBEB** Aleksandr, **DOROGUSH** Anna Veronika, **GULIN** Andrey, *CatBoost : unbiased boosting with categorical features*, Yandex, 2019
- [14] **RAKOTOMALALA** Ricco, *Arbres de décision*, revue MODULAD, 2005
- [15] **RAKOTOMALALA** Ricco, *Analyse de corrélation, Etude des dépendances - variables quantitatives*, Université Lyon Lumière 2, 2017
- [16] **RAKOTOMALALA** Ricco, *Etude des dépendances - Variables qualitatives, Tableaux de contingences et mesures d'association*, Université Lyon Lumière 2, 2020
- [17] **SHAPLEY** Lloyd S., *A value for n-person games*, Contribution to the Theory of Games, 1953
- [18] **TIBSHIRANI** Robert, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, 1996
- [19] **WINKLER** William E., *The state of record linkage and current research problems*, Statistics of Income Division, 1999

## Mémoires d'actuariat

**DELALANDE** Jean, *Rétention au changement de véhicule*, ENSAE, 2015

**DELCAILLAU** Dimitri, *Contrôle et Transparence des modèles complexes en actuariat*, EURIA, 2019

**FRANQUET** Sophie, *Modélisation de la fréquence des sinistres graves en assurance automobile : apports et interprétabilité des méthodes d'apprentissage statistique*, EURIA, 2018

**LAMON** Claire, *Modélisation et analyse de comportements clients en assurance dommage - application au changement de véhicules et à la résiliation de contrats*, Université de Paris Dauphine, 2019

## A Figures annexes

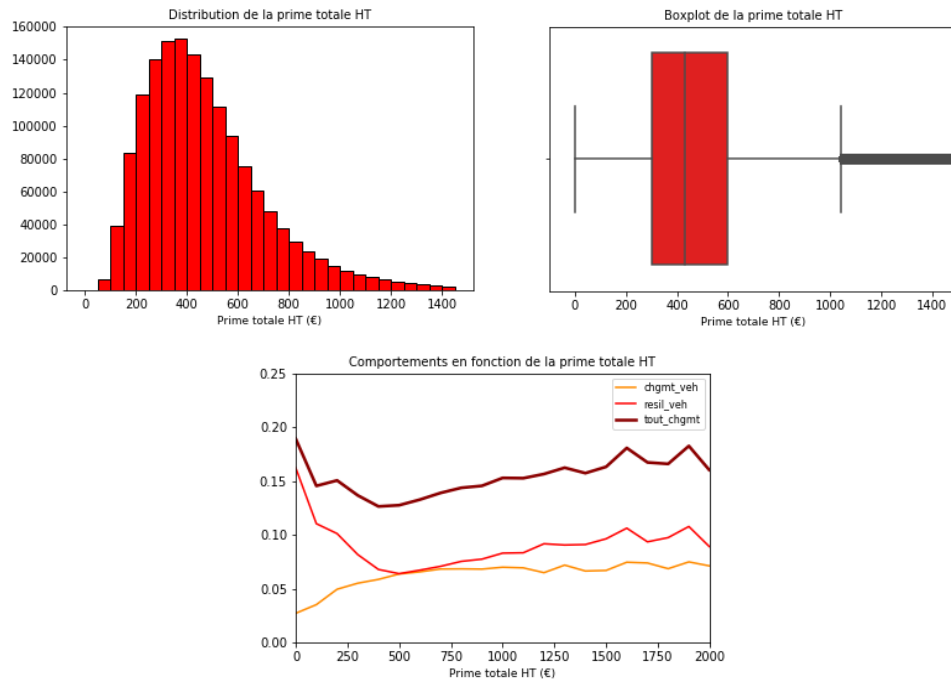


TABLE A.1 – Analyse des valeurs extrêmes pour la prime totale annuelle hors taxes

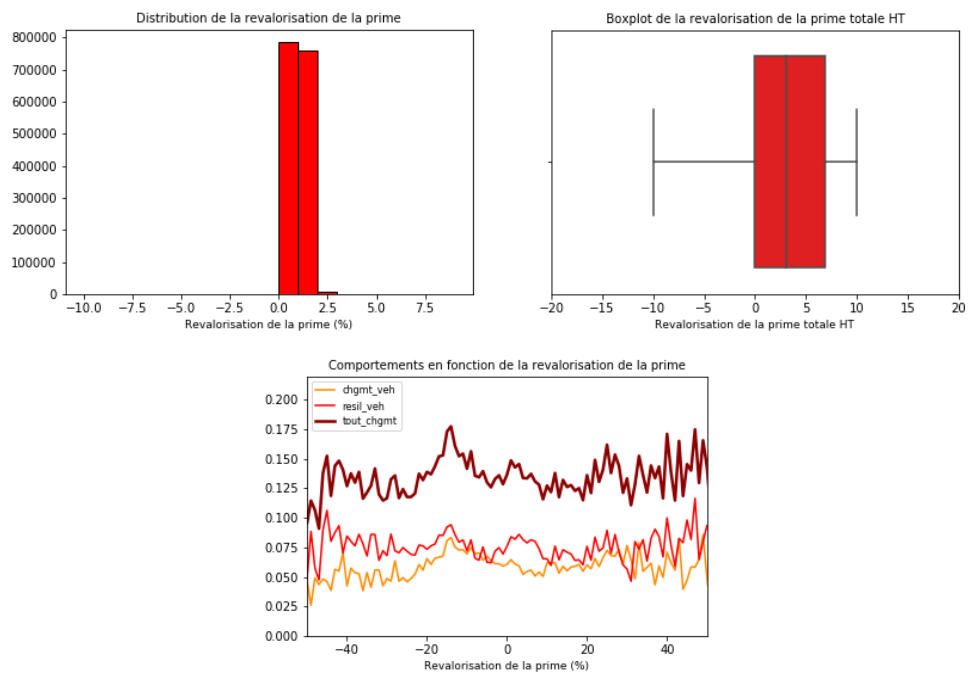


TABLE A.2 – Analyse des valeurs extrêmes pour la revalorisation de la prime totale annuelle hors taxes

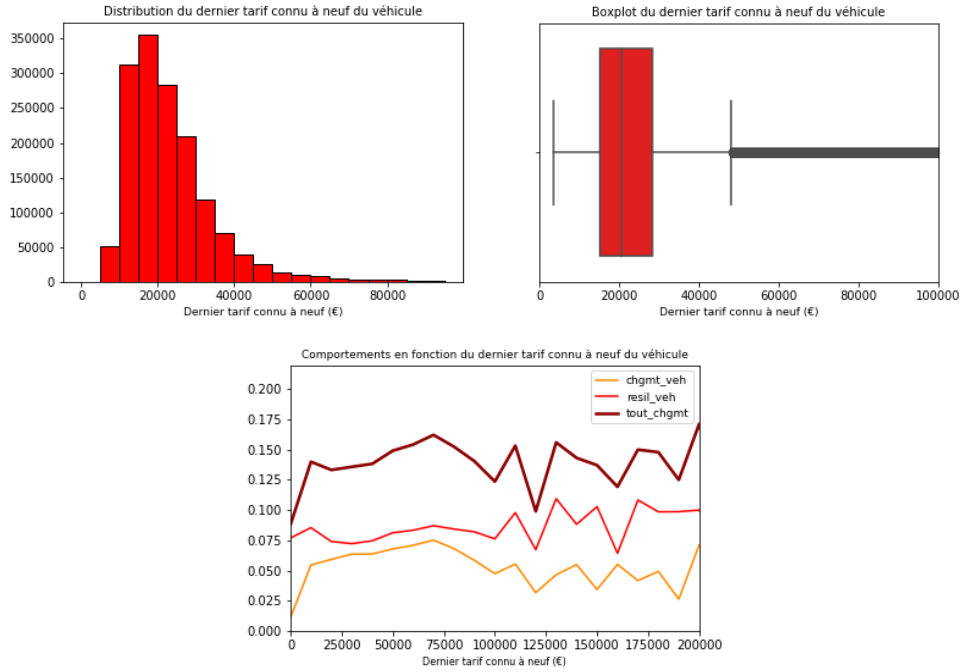


TABLE A.3 – Analyse des valeurs extrêmes pour le dernier tarif à neuf connu du véhicule

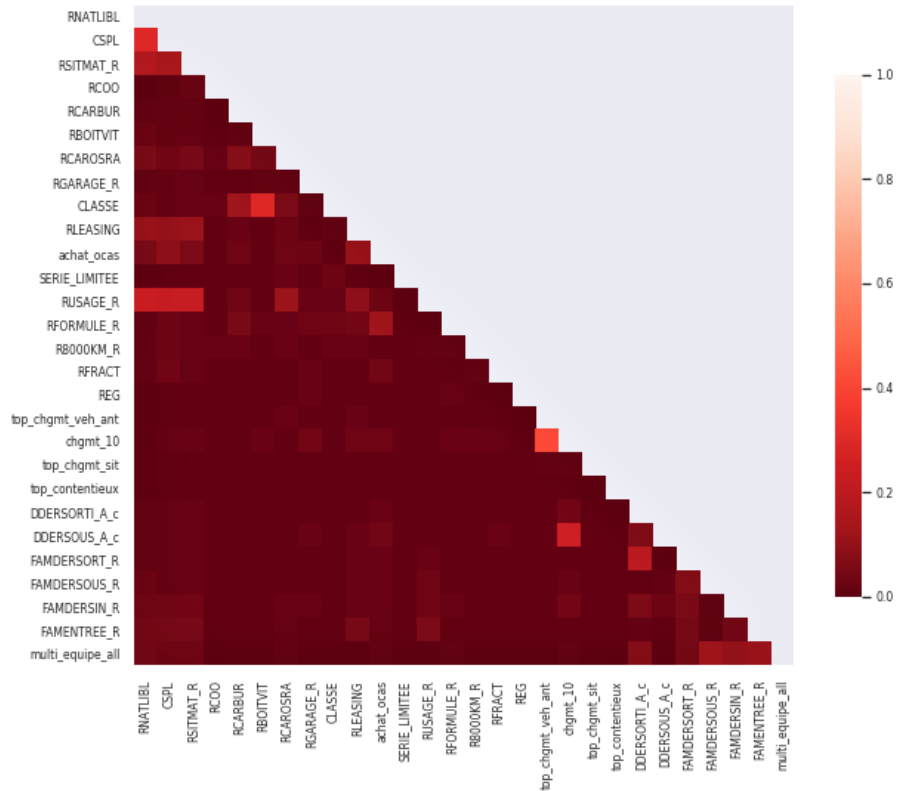


FIGURE A.1 – V de Cramer pour les variables catégorielles après sélection des variables

<b>Véhicule</b>	<b>Conducteur</b>	<b>Contrat</b>	<b>Parcours client</b>
Age véhicule	Age conducteur	Formule	Nb contrats auto
Durée détention	Ancienneté client	Ancienneté contrat	Variation nb contrats auto
Carrosserie	Nature	Prime HT	Revalorisation prime
Groupe SRA	CSP	Leasing	Changements véhicule 10 ans
Classe SRA	Nb équip. commune	Fract. paiements	Nb AN 12M
Cylindrée (cm3)		CET	Date dernière sortie
Prix à neuf			
Nb places			
Nb sinistres 36M			
Charge nette sin.			

TABLE A.4 – Listes des variables explicatives pour la modélisation du changement de véhicule

## B Éléments théoriques complémentaires

### B.1 La distance de Jaro-Winkler : une mesure de similarité entre les chaînes de caractères

La distance de Jaro-Winkler est une variante de la distance de Jaro (Matthew A. Jaro, 1989 [8]) proposée par William E. Winkler en 1999 [19]. Ce score de similarité compris entre 0 et 1, permet de mesurer la similarité entre deux chaînes de caractères – 0 représentant l’absence de similarité et 1 une similarité parfaite. Plus sa valeur est élevée, plus les chaînes de caractères sont similaires.

On note :

- $|s_i|$  est la longueur de la  $i$ -ème chaîne de caractères, pour  $i = 1, 2$
- $m$  est le nombre de caractères correspondants entre les chaînes
- $t$  est le nombre de transpositions

Deux caractères sont considérés comme correspondants si leur éloignement, c’est-à-dire la différence entre leurs positions dans leurs chaînes respectives, ne dépasse pas :

$$\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1. \quad (40)$$

Le nombre de transpositions  $t$  est obtenu en comparant le  $i$ -ème caractère correspondant de  $s_1$  avec le  $i$ -ème caractère correspondant de  $s_2$ . Le nombre de fois où ces caractères sont différents, divisé par deux, donne le nombre de transpositions.

La distance de Jaro est ainsi calculée de la façon suivante :

$$\begin{cases} 0 & \text{si } m = 0 \\ d_j = \frac{1}{3} \left( \frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & \text{si } m \neq 0 \end{cases} \quad (41)$$

La méthode introduite par Winkler utilise un coefficient de préfixe  $p$  (pour lequel il propose une valeur  $p = 0.1$ ) qui favorise les chaînes commençant par un préfixe commun de longueur  $\ell$  (avec  $\ell \leq 4$ ). La distance de Jaro-Winkler est ainsi calculée :

$$d_w = d_j + \ell p (1 - d_j) \quad (42)$$

## Table des figures

0.1	Liaison entre les différentes sources de données . . . . .	4
1.1	Répartition des cotisations pour l'assurance de biens et de responsabilités . . . . .	17
1.2	Ratios combinés en assurance de biens et de responsabilités et en assurance automobile . . . . .	20
1.3	Evolution de l'âge moyen du parc automobile français . . . . .	22
1.4	Répartition du parc automobile français selon la vignette Crit'Air . . . . .	23
2.1	Algorithme CatBoost . . . . .	49
3.1	Schéma sur la méthode de regroupement des différentes bases . . . . .	58
3.2	Arbre de décision pour la segmentation de l'âge du véhicule . . . . .	80
3.3	Distribution de l'âge des assurés dans le portefeuille . . . . .	81
3.4	Distribution de la catégorie socio-professionnelle dans le portefeuille . . . . .	82
3.5	Distribution de la situation maritale des assurés dans le portefeuille . . . . .	82
3.6	Répartition de la population d'étude sur les régions de France . . . . .	83
3.7	Répartition de l'âge des véhicules dans le portefeuille . . . . .	84
3.8	Propension à changer de véhicule avec l'âge du véhicule . . . . .	86
4.1	Matrice de corrélation des variables quantitatives avant sélection des variables . . . . .	96
4.2	Matrice de corrélation des variables quantitatives après sélection des variables . . . . .	97
4.3	V de Cramer pour les variables catégorielles avant sélection des variables . . . . .	98
4.4	Importance des variables dans la modélisation CatBoost du changement de véhicule avant sélection des variables . . . . .	101
4.5	Courbes de précision-rappel des modèles du changement de véhicule . . . . .	109
4.6	Courbes ROC et AUC des modèles du changement de véhicule . . . . .	110
4.7	Évolution de la précision et du rappel pour le Random Forest du changement de véhicule en fonction des quantiles de probabilités prédites . . . . .	112
4.8	Évolution de la précision et du rappel pour le CatBoost du changement de véhicule en fonction des quantiles de probabilités prédites . . . . .	112
4.9	Importance des variables pour le CatBoost du changement de véhicule . . . . .	116
4.10	Importance des variables au moyen des valeurs SHAP pour le CatBoost du changement de véhicule . . . . .	118
5.1	Courbes de précision-rappel des modèles de résiliation au changement de véhicule . . . . .	125
5.2	Évolution de la précision et du rappel pour le Random Forest de la résiliation au changement de véhicule en fonction des quantiles de probabilités . . . . .	125
5.3	Évolution de la précision et du rappel pour le CatBoost de la résiliation au changement de véhicule en fonction des quantiles de probabilités . . . . .	126
5.4	Importance des variables pour le CatBoost de la résiliation au changement de véhicule . . . . .	128
5.5	Importance des variables au moyen des valeurs SHAP pour le CatBoost de la résiliation au changement de véhicule . . . . .	129
6.1	Sélection du nombre de clusters idéal par la méthode du coude . . . . .	133
6.2	Répartition de la segmentation de la valeur client dans le portefeuille . . . . .	137
A.1	V de Cramer pour les variables catégorielles après sélection des variables . . . . .	147



## Liste des tableaux

1.1	Les différents types d'assurance . . . . .	16
3.1	Méthodologie de la constitution de la base de données par année . . . . .	59
3.2	Méthode d'identification des changements de véhicule . . . . .	60
3.3	Méthode d'identification des mises en conformité des numéros d'immatriculation . . . . .	61
3.4	Méthode d'identification des résiliations pour changement de véhicule . . . . .	62
3.5	Analyse de correspondance des modèles de véhicules avec la distance de Jaro-Winkler . . . . .	68
3.6	Analyse des valeurs extrêmes pour le CET du contrat . . . . .	71
3.7	Analyse des valeurs extrêmes pour l'âge du véhicule . . . . .	73
3.8	Analyse des valeurs extrêmes pour l'âge du conducteur . . . . .	75
3.9	Méthodologie de la constitution de la base de données sans indice temporel . . . . .	81
3.10	Répartition du groupe et de la classe SRA dans le portefeuille . . . . .	85
4.1	Matrices de confusion des modèles du changement de véhicule . . . . .	106
4.2	Métriques de scores des modèles du changement de véhicule . . . . .	107
4.3	Métriques de scores sur les quantiles de probabilités prédites des modèles du changement de véhicule . . . . .	113
5.1	Listes des variables explicatives pour la modélisation de la résiliation au changement de véhicule . . . . .	122
5.2	GridSearch des hyperparamètres du Catboost de la résiliation au changement de véhicule . . . . .	123
5.3	Matrices de confusion des modèles de la résiliation au changement de véhicule . . . . .	124
5.4	Métriques de scores des modèles de la résiliation au changement de véhicule . . . . .	124
5.5	Métriques de scores sur les quantiles de probabilités prédites des modèles de la résiliation du changement de véhicule . . . . .	126
6.1	Résultats de l'algorithme $K$ -means . . . . .	134
A.1	Analyse des valeurs extrêmes pour la prime totale annuelle hors taxes . . . . .	146
A.2	Analyse des valeurs extrêmes pour la revalorisation de la prime totale annuelle hors taxes . . . . .	146
A.3	Analyse des valeurs extrêmes pour le dernier tarif à neuf connu du véhicule . . . . .	147
A.4	Listes des variables explicatives pour la modélisation du changement de véhicule . . . . .	148