Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 17/11/2021

Par : **Sénévo Léonard AGBEDJINOU**

Titre : **Détection des profils à fort risque de résiliation en assurance Auto avec des méthodes d'apprentissage automatique**

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

M. Christian ROBERT

Membres présents du jury de l'Institut des Actuaires

Mme Florence PICARD
M. Olivier CONSTANTIN

Entreprise :
Nom : Vincent SERGENT
Signature :

Directeur du mémoire en entreprise :

Nom : Aubin CHAUVEAUX
Signature :

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

Secrétariat :

Signature du candidat

Bibliothèque :

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

Résumé

Le marché de l'assurance auto en France est très concurrentiel avec une cohabitation de plusieurs fournisseurs d'assurance couplé d'une stagnation du parc des véhicules assurés. On distingue notamment les bancassureurs, les assureurs directs ou encore les sociétés avec intermédiaires. La conséquence directe est qu'il est non seulement difficile de gagner de nouveaux clients (affaires nouvelles) mais aussi de garder plus longtemps un client en portefeuille. C'est pourquoi l'étude des résiliations est devenue un sujet crucial dans la plupart des compagnies d'assurance. Elle permet de connaître le profil de risque des assurés qui résilient le plus afin de prendre des mesures adéquates pour les fidéliser ou non.

Ce mémoire traite de la résiliation de l'assurance auto du portefeuille GAN AUTO. L'objectif étant de garder les clients le plus longtemps, il s'intéresse uniquement aux résiliations qui émanent de la volonté de l'assuré. Une première statistique descriptive après l'étape de prospection des données est réalisée. Après cela, un modèle de régression logistique est mis en place. Ce modèle a permis de décorréler les effets de chaque modalité des différentes variables et de mieux expliquer le profil de risque de chaque assuré selon certains critères "métiers". Le pouvoir prédictif de ce modèle étant limité notamment à cause du déséquilibre entre les classes des contrats résiliés et non résiliés, nous avons mis en place deux modèles de machine learning : les forêts aléatoires -Random Forest et le XGBoost. Des techniques de rééchantillonnage sont ensuite appliquées pour améliorer les résultats obtenus notamment en termes de AUC, courbe ROC, courbe Rappel- Précision. Enfin, une méthodologie innovante basée sur la construction d'un scoring de sinistralité avec l'algorithme BiRank qui utilise un graphe biparti valué est mise en oeuvre pour mieux cibler les contrats résiliés.

Mots clés : *Résiliation*, *Régression logistique*, *Random Forest*, *XGBoost*, *BiRank*, *Graphe biparti.*

Abstract

The car insurance market in France is very competitive, with several insurance providers cohabiting and the number of insured vehicles stagnating. There are bank insurers, direct insurers and companies with intermediaries. The direct consequence is that it is not only difficult to win new customers but also to keep a customer in the portfolio for longer. This is why the study of cancellations has become an import issue in most insurance companies. It allows to know the risk profile of the policyholders who cancel the most in order to take adequate measures to keep them.

This thesis deals with the cancellation of car insurance in the "GAN AUTO" portfolio. As the objective is to keep customers as long as possible, it focuses only on cancellations that emanate from the will of the insured. A first descriptive statistic after the data collection stage is carried out. After that, a logistic regression model is set up. This model made it possible to capture the effects of each modality of the different variables and to better explain the risk profile of each insured according to certain business criteria. Finally, as the predictive power of this model was limited, mainly because of the imbalance between the classes of cancelled and non-cancelled contracts, we implemented two machine learning models: random forests and XGBoost. Resampling techniques are applied to improve the results obtained, particularly in terms of AUC, ROC curve and Recall-Accuracy curve. Finally, an innovative methodology based on the construction of a claims scoring with the BiRank algorithm which uses a valuated bipartite graph network, is implemented to better target the terminated contracts.

Key words : *Cancellations*, *Logistic Regression*, *Random Forest*, *XGBoost*, *BiRank*, *Bipartite graph.*

Note de synthèse

Cadre de l'étude

Le portefeuille GAN Auto subit depuis des années, des mouvements techniques dont les résiliations de contrats. Ces résiliations peuvent émaner de l'assuré ou de la compagnie. Le premier type de résiliation se trouve renforcé par un contexte réglementaire qui protège les consommateurs. En effet, avec les lois Chatel et Hamon, les assurés peuvent résilier leurs contrats d'assurance auto après les 12 premiers mois, à tout moment sans aucune pénalité. A cela, s'ajoute un marché de plus en plus étroit avec de nombreux fournisseurs (la bancassurance, les assureurs directs et les mutuelles) et un parc automobile assuré en faible progression.

Il en découle comme corollaire pour un assureur, la difficulté de, non seulement gagner de nouveaux clients, mais les garder le plus longtemps possible. Cette étude s'inscrit donc dans la logique de mieux comprendre les déterminants de la résiliation de contrat émanant du vouloir de l'assuré. *In fine*, cela permettrait de mieux cibler les contrats qui ont une probabilité de résiliation élevée et de mieux orienter les leviers de fidélisation tels que les gestes commerciaux, les chèques franchises ou encore les rabais tarifaires.

Données de travail et plan

Pour modéliser le risque de résiliation, nous avons à notre disposition deux types de données : les données internes et les données externes. Les données disponibles vont du 1er Janvier 2017 au 31 décembre 2020.

Les données internes sont relatives aux contrats et aux différents mouvements qu'ils subissent : les sinistres, les résiliations. Elles portent aussi sur les caractéristiques du client, sur sa situation socio-professionnelle ou encore tous les contrats qu'il détient chez GAN Assurances.

Les données externes utilisées dans cette étude sont les bases des unités urbaines de 2020 de l'INSEE.

Le plan du travail se décline en quatre parties. Dans un premier temps, nous allons faire une statistique descriptive pour identifier les variables les plus discriminantes pour la prédiction de la résiliation d'un contrat. La seconde partie sera consacrée à la régression logistique pour modéliser la probabilité de résilier son contrat. Quant à la troisième partie, il sera question de la mise en oeuvre des méthodes de machine learning : Random Forest et XGBoost. L'ultime partie concerne la mise en oeuvre d'un scoring de sinistralité avec l'algorithme BiRank.

Le contenu du travail

L'avant première : statistique univariée

L'objectif de ce premier jet dans le sujet est de déceler si possible une liste potentielle des variables les plus discriminantes en la propension à résilier. De ce fait, dans un premier temps, un rappel sur la structure du portefeuille sera effectué. Il en ressort qu'entre 2017 et 2020, le portefeuille GAN AUTO a perdu environ 3% de clients.

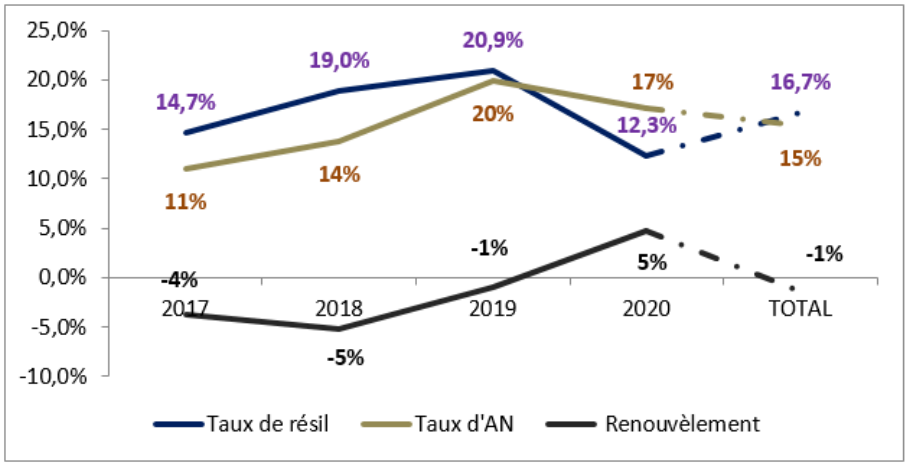


FIGURE 1 – Évolution de la structure du portefeuille

La sinistralité des contrats résiliés est presque le double de la sinistralité des contrats actifs. Cette sinistralité a été mesurée par la fréquence de sinistre hors nulle et hors sans suite et le coût moyen. Toutefois, les résultats de 2020, comme tout le reste des statistiques descriptives effectuées, sont différents de ceux des autres années. On enregistre une baisse de résiliation de plus de 9 points de pourcentage.

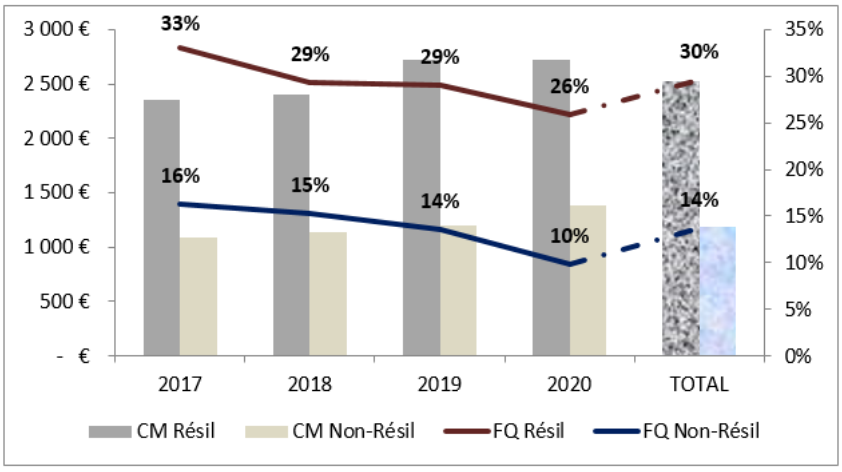


FIGURE 2 – Sinistralité en fonction de la résiliation

Cette baisse est expliquée par le taux de circulation qui suit la même tendance que la fréquence mensuelle de sinistres en 2020. La variable "antécédent de sinistre" est alors pressentie pour expliquer les résiliations de contrats.

Les variables les plus discriminantes sont notamment le CRM (Coefficient de Réduction et de Majoration), l'antécédent de sinistres et l'évolution tarifaire à l'issue de cette étape.

Une première modélisation du risque de résiliation avec la régression logistique

La régression logistique a permis de détecter le profil des contrats les plus risqués. Ils sont définis par les variables suivantes :

— le CRM ;
— antécédent de sinistre sur les 24 derniers mois ;
— ancienneté du contrat ;
— la variation de prime en % .

Les contrats qu'il faut surveiller sont ceux ayant un CRM supérieur à 1 avec une ancienneté comprise entre 1 et 3 ans, au moins 2 sinistres lors des 24 derniers mois, pour une revalorisation tarifaire de plus de 3%. Ces contrats sont ceux qui ont le plus de chance d'être résiliés lors de la version en cours du contrat.

Toutefois, la régression logistique a un pouvoir prédictif faible avec un $AUC = 0,59$. En effet, l'AUC (*Area Under the Curve*) exprime la probabilité pour un modèle de placer un positif (un contrat résilié) devant un négatif (un contrat non résilié). Sa valeur varie de $0,5$ à 1.

Machine learning : Random Forest et XGBoost

Avec les données déséquilibrées, les prédicteurs se trouvent de manière générale avec une variance élevée. En d'autres termes, ils sont très sensibles aux nouvelles données. Pour palier à cela, des méthodes ensemblistes seront mises en oeuvre. Leur caractéristique principale est de réduire la variance.

La mise en oeuvre du Random Forest et du XGBoost sur l'échantillon d'entrainement a donné des performances similaires. En effet, le Random Forest donne un $AUC = 0,705$ alors que XGBoost fournit un $AUC = 0,71$. La légère différence entre les performances des deux classifieurs est justifiable. En effet, XGBoost agit non seulement sur la variance du modèle, mais aussi sur son biais qui mesure l'erreur de prédiction du modèle. Il est le modèle qui est choisi pour détecter les contrats résiliés.

Détection des contrats résiliés avec XGBoost à l'aide d'un scoring de sinistralité avec BiRank

Pour améliorer le pouvoir prédictif du XGBoost, nous avons construit un scoring de sinistralité sur les contrats sinistrés. La construction du scoring a nécessité l'utilisation de l'algorithme BiRank de He et al. [18], qui est lui-même une version personnalisée de l'algorithme PageRank de Google. L'algorithme fonctionne avec un graphe biparti construit entre les contrats (représentés en nombres) et les types de sinistres (DOM pour dommages, RC pour Responsabilité Civile, VOL pour Vol, BDG pour Bris de Glaces et AUG pour Autres garanties). Un exemple de ce graphe est représenté sur la figure ci-dessous.

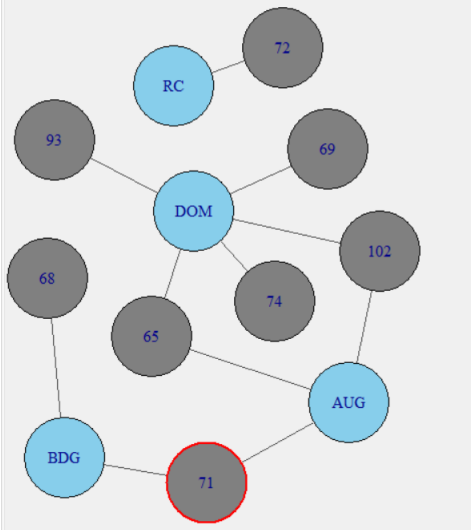


FIGURE 3 – Un exemple de réseau de graphe biparti entre les contrats et les types de sinistres

Le scoring obtenu a permis d'améliorer les résultats avec un AUC qui passe de $0,71$ à $0,76$.

Pour une meilleure explicabilité des résultats, nous avons mis en oeuvre la SHAP Value qui mesure l'influence des variables sur la probabilité de résiliation.

Executive Summary

Study framework

For years, the GAN Auto portfolio has been subject to technical changes, including contract cancellations. These cancellations can be initiated by the insured or by the company. The first type of cancellation is reinforced by a regulatory context that protects consumers. Indeed, with the Chatel and Hamon laws, policyholders can cancel their car insurance contracts after the first 12 months at any time without any penalty. In addition to this, the market is becoming increasingly narrow, with numerous suppliers (bank insurers, direct insurers and mutual insurance companies) and the number of cars insured is growing slowly.

The consequence is that it is difficult for an insurer not only to win new customers but also to keep the customers they already have for as long as possible. This study is therefore part of the logic of better understanding the determinants of contract cancellation due to the policyholder's wishes. This would make it possible to better target contracts that have a high probability of cancellation and to better orient loyalty-building measures such as commercial gestures, franchise vouchers or rate discounts.

The Working data and plan

To model the risk of termination, we have two types of data at our disposal : internal data and external data. The available data goes from January 1st, 2017 to December 31, 2020.

The internal data concerns in particular the data on contracts and the various movements they have had : claims, cancellations. The internal data also concern data on the characteristics of the client, their professional situation and all the contracts they hold with GAN Assurances.

The external data used in this study are the INSEE urban unit databases for 2020.

The work plan is divided into four parts. In the first part, we will carry out a descriptive statistic to identify the most discriminating variables for the prediction of the cancellation

of a contract. In the second part, we will use logistic regression to model the probability of terminating a contract. In the third part, we implemented machine learning methods : Random Forest and XGBoost. The last part concerns the implementation of a claims scoring with the BiRank algorithm.

Main work

Uni-variate statistic

The objective of this first draft in the subject is to identify, if possible, a potential list of the most discriminating variables in the propensity to terminate. Therefore, as a first step, a reminder of the portfolio structure will be given. This shows that between 2017 and 2020, the GAN AUTO portfolio lost 3% of customers.

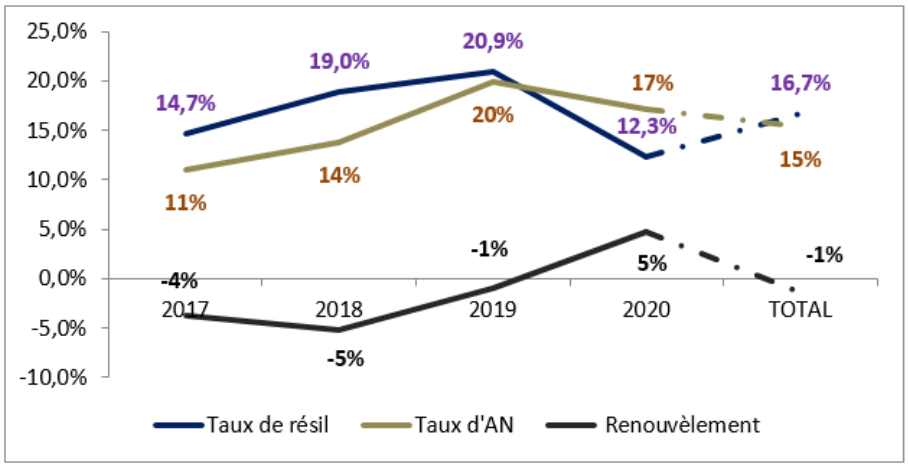


FIGURE 4 – Evolution of the portfolio structure

The loss experience of terminated contracts is almost double the loss experience of active contracts. This loss experience was measured by the frequency of claims excluding nil and no claims and the average cost. However, the results for 2020, like all other descriptive statistics, are different from those of other years. There is a drop in cancellations of more than 9 percentage points.

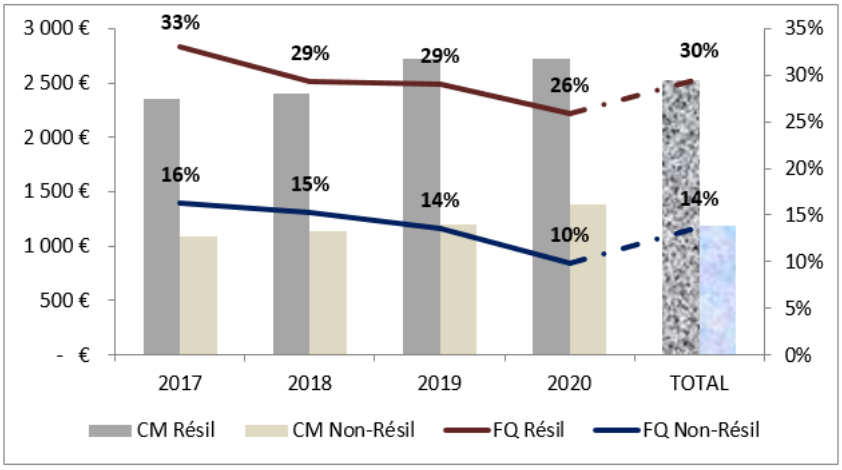


FIGURE 5 – Claims experience as a function of termination

This decrease is explained by the circulation rate, which follows the same trend as the monthly claim frequency in 2020. The claims history variable is then suggested as an explanation for policy cancellations. The effect of the covid-19 crisis will be zoomed in and analysed in detail.

The most discriminating variables are the CRM (Coefficient of Bonus Malus), the claims history and the tariff evolution at the end of this stage.

A first modeling of the risk of termination with the logistic regression

Logistic regression has allowed us to detect the profile of the riskiest contracts. These are :

- CRM (Bonus Malus) ;
- history of claims over the last 24 months ;
- Contract age ;
- Premium variation in %.

The contracts that should be monitored are those with a CRM greater than 1 and a length of service of between 1 and 3 years with at least 2 claims in the last 24 months for a rate increase of more than 3%. These contracts are the most likely to be terminated in the current version of the contract.

However, the logistic regression has a low predictive power with a $AUC = 0.59$.

Machine learning : Random Forest and XGBoost

With unbalanced data, predictors are generally found to have high variance. In other words, they are very sensitive to new data. To overcome this, set methods will be imple-

mented. Their main characteristic is to reduce the variance.

The implementation of the Random forest and the XGBoost on training data gave different performances. Indeed, the random forest gives a $AUC = 0,705$ whereas XGBoost gives a $AUC = 0,71$. These results are justifiable because XGBoost acts not only on the variance but also on the bias which measures the prediction error of the model. It is the model that is chosen to detect terminated contracts.

Detection of terminated contracts with XGBoost using claims scoring with BiRank

To improve the predictive power of the XGBoost, we constructed a claims score on the claims contracts. The construction of the scoring required the use of the BiRank algorithm of He et al. [18], which is itself a customized version of Google's Page-Rank algorithm. The algorithm works with a bipartite graph constructed between contracts (represented as numbers) and claim types (DOM for Damage, RC for Third Party Liability, VOL for Theft, BDG for Glass Breakage and AUG for Other Coverage). An example of this graph is shown in the figure below.

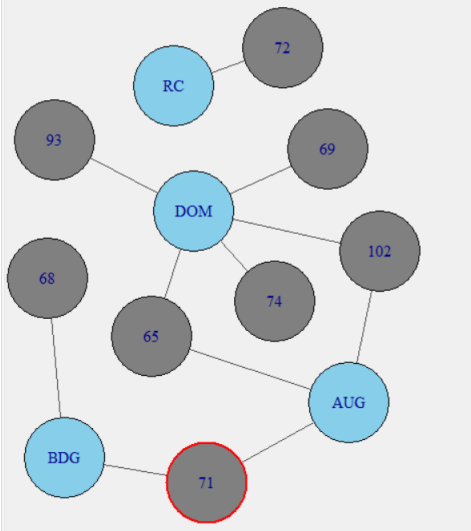


FIGURE 6 – An example of a bipartite graph network between contracts and types of claims

The scoring obtained has improved the results with an AUC that goes from 0.71 to 0.76.

For a better explanation of the results, we implemented the SHAP Value which measures the influence of the variables on the probability of cancellation.

Remerciements

Je tiens à remercier de vives voix M. CHAUVEAUX Aubin, qui m'a encadré tout au long de la réalisation de ce mémoire. Ses multiples conseils, tant dans le cadre de ce travail que dans la vie professionnelle en général m'ont été d'une grande aide. Je ne saurais exprimer ma profonde gratitude et mes remerciements à M. SERGENT Vincent et M. HOUNSOUNON Cédric ainsi qu'à toute l'équipe du Pôle Actuariat de la Direction Technique de GAN ASSURANCES. Ils ont été des collègues dévoués au travail bien fait, au partage et à l'entraide.

Je voudrais remercier particulièrement Mlle MBIA NDI Marie Thérèse qui m'a beaucoup soutenu, tant sur le plan technique qu'émotionnel durant ce travail.

Mes remerciements vont aussi à l'endroit du corps administratif et professoral de l'Ecole Nationale de la Statistique et de l'Administration Economique ENSAE Paris pour leur dévouement à la cause estudiantine, dévouement qui leur donne la force de continuer à forger les futurs ingénieurs et décideurs du monde. Je voudrais saluer particulièrement mon référent M. ROBERT Christian pour ses suggestions et pour son encadrement.

Enfin, j'aimerais remercier tous mes amis m'ayant apporté leurs aides et remarques afin que ce travail soit finalisé tel qu'il l'est actuellement.

Table des matières

Introduction générale

L'assurance auto a évolué depuis des années en France. C'est un marché très concurrentiel avec un parc automobile en stagnation. A cela, s'ajoute un environnement réglementaire de plus en plus exigeant. Dans ces contextes, il est aujourd'hui difficile pour un assureur de gagner un nouveau client ou mieux de conserver un client le plus longtemps possible dans son portefeuille.

C'est pourquoi l'étude des résiliations est un sujet crucial chez les distributeurs d'assurance auto. Elle leur permet de comprendre les déterminants de ce mouvement technique et de proposer des solutions en vue d'une meilleure politique de fidélisation du client. Couramment, ils utilisent les gestes commerciaux, les chèques franchises ou encore les modulations tarifaires.

Plusieurs mémoires de l'Institut des actuaires ont traité le sujet de la résiliation, chacun à sa manière. Par exemple, en 2016, Hajar MARKAOUI [1] a étudié la probabilité de résiliation dans un portefeuille d'assurance auto dans le but de stabiliser la structure du portefeuille. Il a cherché à quantifier l'impact de ce mouvement technique sur la tarification des affaires nouvelles.

Dans ce mémoire, nous allons analyser le phénomène de résiliation des contrats d'assurance Auto chez GAN Assurances sous un autre angle. Contrairement à l'étude susmentionnée, notre étude se focalisera principalement sur la détection des résiliations. Ainsi, ce travail peut servir d'aide à la décision pour la direction marketing pour la fidélisation des clients.

Un contrat d'assurance auto peut être résilié suite à une demande de l'assuré ou par une volonté motivée de l'assureur. L'objectif étant de fidéliser les clients, nous nous intéresserons uniquement aux résiliations qui émanent de la volonté de l'assuré. Pour ce faire, une première modélisation basée sur la régression logistique sera effectuée. Ensuite, nous mettrons en oeuvre deux modèles dits de Machine Learning : Random Forest et XGBoost. Ainsi, quitte à refaire le modèle sur les bases ultérieures, nous pourrons alors détecter les contrats susceptibles d'être résiliés dans les 12 prochains mois ; y compris l'échéance principale.

1. Le titre du mémoire : *« Analyse de la probabilité de résiliation en assurance automobile : comparatif de deux méthodologies d'estimation et conséquences sur la tarification»* à consulter ici Hajar MARKAOUI

Enfin, nous mettrons en oeuvre une approche innovante pour améliorer les résultats du meilleur modèle de détection de résiliation. Cette approche combine un réseau biparti et un scoring de sinistralité construit avec l'algorithme BiRank.

Généralités et contexte de l'étude

Les fournisseurs d'assurance automobile ne cessent d'augmenter avec les années, rendant la concurrence encore plus rude qu'avant. Aussi, les *start ups*, les grandes compagnies ou encore les bancassureurs se livrent une "bataille" pour garder leurs clients en portefeuille ou mieux, proposer des offres plus alléchantes pour ravir les clients à leurs concurrents. Il est admis que les contrats qui restent le plus longtemps en portefeuille rapportent plus que les nouveaux contrats (affaires nouvelles). C'est pourquoi les assureurs mettent en place de nombreuses stratégies pour fidéliser les assurés. Parmi ces stratégies, nous pouvons citer : *des gestes commerciaux* et *des avantages clients*. De ce fait, l'étude des résiliations devient un sujet central chez les fournisseurs d'assurances.

Mais **comment mieux cibler les populations les plus à même de résilier** ? Une question à laquelle certains éléments de réponses seront apportés tout au long de ce mémoire. Mais bien avant, il est judicieux de mieux comprendre dans un premier temps le fonctionnement du marché de l'assurance auto en France.

Ainsi, dans ce chapitre, nous allons dans un premier temps rappeler les caractéristiques de l'assurance auto en France. Notre objectif est de mettre l'accent sur le caractère obligatoire de cette assurance et d'énumérer quelques unes des garanties proposées par GAN Assurances. Dans un second temps, nous mettrons la lumière sur le marché de l'assurance auto en France. L'objectif dans ce cas sera de faire ressortir l'étroitesse de ce marché et de justifier le qualificatif de marché concurrentiel de ce dernier. Enfin, nous rappellerons la vie d'un contrat d'assurance avant de finir par la déclinaison des objectifs de ce mémoire.

1.1 L'assurance auto

1.1.1 Une assurance obligatoire

Depuis 1958, l'assurance auto est obligatoire en France. En effet, d'après la loi du 27 février 1958, reprise dans le code des assurances dans son article L. 211-1 et dans le code de la route dans son article L. 324-1, il est obligatoire de souscrire une assurance automobile car *« la responsabilité civile - d'un conducteur - [...] peut être engagée en raison de dommages subis par des tiers »* [1]. De ce fait, pour circuler, les véhicules doivent être couverts par une assurance garantissant cette responsabilité.

Toutefois, le conducteur ne peut souscrire à l'assurance auto chez n'importe quel assureur. La loi stipule qu'elle doit être souscrite chez un assureur qui est agréé par la Fédération Française d'Assurance - FFA - pour pratiquer les opérations d'assurances sur le sol français.

1. Code des Assurances , Version du 28/07/2021

1.1.2 Présentation du produit GAN Auto

GAN Auto est le nom commercial donné à l'assurance automobile proposée par GAN Assurances. Ce produit couvre une large gamme de risques liés aux véhicules terrestres à moteurs. Il propose plusieurs garanties de sorte à satisfaire les besoins de ses différents clients. Nous pouvons répartir ces garanties en trois (03) groupes : *les garanties indispensables*, *les garanties liées aux dommages causés au véhicule* et *les autres garanties optionnelles*.

Les garanties indispensables sont les garanties de base. Elles sont offertes dans tous les contrats d'assurance auto à l'exception des véhicules au repos. Par exemple, dans ces garanties de base, nous pouvons trouver la garantie *Responsabilité Civile (RC) Circulation* qui est le minimum nécessaire pour répondre à l'obligation d'assurance auto en France.

Les garanties liées aux dommages causés au véhicule peuvent être vues comme des garanties qui permettent aux assurés de se prémunir contre les dommages que peuvent subir leurs véhicules. On peut citer par exemple les garanties *Bris de glaces* ou encore *Dommages tous accidents et vandalisme*.

Les autres garanties optionnelles sont des garanties qui sont pour la plupart proposées en option aux assurés pour répondre à des besoins spécifiques. On y trouve par exemple la garantie *Assistance 0 Km*. Rappelons qu'en assurance, l'assistance peut être définie comme une opération par laquelle l'assureur s'engage à porter secours à un assuré en cas de difficultés dont la nature à été précisée préalablement dans le contrat. Le tableau 1.1 résume l'ensemble des garanties par groupe.

TABLE 1.1 – Les différentes garanties offertes par GAN Auto

Les garanties indispensables	Les dommages au véhicule	Les autres garanties optionnelles
RC circulation	Bris de glaces	Assistance 0 Km
RC hors circulation	Catastrophes naturelles	Protection Juridique Automobile
Défense pénale et recours suite à un accident	Attentats et actes de terrorisme	Location de véhicules de remplacement
Assistance au véhicule	Incendie	Marchandises et Animaux Transportés
Assistance aux personnes en déplacement	Evènement climatiques	Contenu du véhicule
Accidents corporels du conducteur	Vol	Aménagements du véhicule
	Dommages tous accidents et vandalisme	Valeur d'achat 3 ans
	Pannes mécaniques Complète	Pertes financières (LOA-LDD-Crédit)
	Auto Presto Privilège	

1.1.3 La structure du produit GAN Auto

Le produit GAN Auto est commercialisé par formule et par option selon le profil et les besoins de l'assuré. En effet, on distingue quatre formules. Chaque formule est composée des garanties de base auxquelles l'assuré peut ajouter des garanties optionnelles. A ces formules, s'ajoute une formule exceptionnelle pour les véhicules au repos.

- **Formule Tiers :** Cette formule est constituée par ***les garanties indispensables*** ;

- **Formule Tiers + :** Cette formule permet de bénéficier de *la Formule Tiers* à laquelle viennent s'ajouter toutes les garanties des dommages au véhicule sauf les trois (03) dernières mentionnées dans le tableau 1.1 ;
- **Formule Tous Risques :** Elle intègre la garantie Dommages tous accidents et vandalisme en ajout de la *Formule Tiers+* ;
- **Formule Tous risques + :** Elle permet de souscrire aux garanties Pannes mécaniques Complète et assistance 0 km en addition à la *Formule Tous Risques* ;
- **Formule Véhicule au Repos :** Elle contient essentiellement 5 garanties dont la RC hors circulation, les catastrophes naturelles, attentats et actes de terrorisme, incendie et Vol.

Il faut noter qu'à l'exception de la formule Véhicule au Repos, des garanties optionnelles sont accessibles en fonction du genre de véhicule. Ainsi, le produit GAN Auto est un produit complet, offrant différentes possibilités à l'assuré selon ses besoins.

1.2 Le marché de l'assurance auto en France

Avec ses 67, 4 millions d'habitants au 1er janvier 2021 (INSEE) [5] , la France est l'un des plus vastes marchés de l'assurance automobile de l'Union Européenne. Ce marché compte de nombreux fournisseurs et devient depuis quelques années, le terrain de quelques phénomènes qui le caractérisent. Ainsi, dans cette sous section, nous allons énumérer les différents types de forunisseurs d'assurance présents sur ce marché avant de regarder les évolutions du parc national des véhicules assurés.

1.2.1 Les modes de distribution de l'assurance auto

La FFA distingue cinq types de fournisseurs d'assurances auto selon leurs modes de distribution : *les assurances directes*, *la bancassurance*, *les sociétés avec intermédiaires*, *les mutuelles sans intermédiaires*, et *les autres sociétés sans intermédiaires*.

Les assurances directes

On parle d'assurance directe lorsqu'une société d'assurance vend ses produits par le biais de ses salariés. Ces salariés peuvent être par exemple des agents généraux qui les représentent exclusivement.

La bancassurance

Apparut en France dans les années 70 [2], la bancassurance peut être assimilée à la distribution des produits d'assurance à travers les guichets des banques. Elle fonctionnerait donc comme les assureurs classiques à la seule différence qu'elle se base sur le réseau des banques ; ce qui lui donne l'avantage d'une meilleure affinité avec les clients.

Les sociétés avec intermédiaires

Ce sont les assureurs qui distribuent leurs produits à travers des intermédiaires non salariés. Ces intermédiaires sont les agents généraux, les courtiers ou les mandataires d'assurances. Ils sont rémunérés par des commissions qu'ils perçoivent sur les contrats.

Les Mutuelles sans intermédiaires

Ce sont des mutuelles qui vendent leurs produits d'assurances sans intermédiaire. Elles travaillent uniquement avec leurs salariés. Le plus souvent, ces mutuelles sont spécialisées dans les assurances pour des voitures professionnelles. Elles regroupent ainsi des sociétaires dans les catégories socioprofessionnelles suivantes : fonctionnaires, artisans, commerçants.

Autres sociétés sans intermédiaires

C'est l'ensemble des autres forunisseurs d'assurance automobile qui n'appartiennent pas à l'ensemble des classes susmentionnées. Leur caractérisqtique principale est qu'ils fonctionnent sans intermédiaires.

Ces différents fournisseurs d'assurance auto ont collecté 58.2 milliards d'euros de cotisations sur l'année 2019. Ce chiffre est en constante progression mais reste limité par une évolution de plus en plus faible du parc des véhicules assurés en France.

1.2.2 Parc des véhicules assurés

Le parc national de véhicules assurés est en constante progression depuis 2015 ; atteignant la barre des 43,07 millions en 2019. D'après le tableau 1.2, nous notons une évolution moyenne de 1,3% par année entre 2015 et 2019. Néanmoins, cette hausse est de plus en plus faible ; passant de 1,7% entre 2015 et 2016 à 1,1% entre 2018 et 2019.

2. « Au début des années 70, les ACM (Assurances du Crédit Mutuel) Vie et IARD obtiennent leur agrément, marquant ainsi l'histoire de l'assurance. L'idée leur est venue de se passer d'intermédiaire pour l'assurance des crédits emprunteurs, et de devenir eux-mêmes assureur de leurs propres clients de banque »[9].

TABLE 1.2 – Parc des véhicules assurés en France

Année	2015	2016	2017	2018	2019
Parc de vehicules (en Million)	40,90	41,60	42,20	42,60	43,07
Evolution annuelle en %	-	1,7	1,4	0,9	1,1

Note de lecture : En 2019, 43,07 millions de véhicules sont assurés en France avec une hausse moyenne de 1,3% par année entrre 2015 et 2019 .
Source : Les chiffres clés de l'assurance (2015-2019) - FFA

1.2.3 Un marché marqué par la concurrence

Le marché d'assurance auto en France est marqué par une forte concurrence entre les différents forunisseurs ; ce qui est observable dans la répartition des cotisations en fonction du mode de distribution.

En effet, d'après le tableau 1.3, nous remarquons une montée en puissance des bancassureurs qui gagnent chaque année des parts supplémentaires du marché (5% en moyenne). Dans son article intitulé «*Classement Auto-MRH 2019 : les assureurs en (re)conquête* », l'ARGUS de l'assurance, révèle que les 5 premiers assureurs ayant les plus fortes progressions de chiffres d'affaire entre 2017 et 2018 en assurance auto sont des bancassureurs. Dans l'ordre de mérite, ce sont Natixis Assurances, Société Générale Assurances, Suravenir Assurances, Crédit Agricole Assurances et Groupe des Assurances du Crédit Mutuel. La même tendance continue en 2019 car le Crédit Agricole Assurances et le Crédit Mutuel ont gagné respectivement 119 000 et 117 000 assurés en 2019 d'après le classement publié par l'ARGUS de l'assurance.

Dans le même temps, les assureurs directs ne cessent de perdre du marché. En 2019, ces derniers ont perdu, selon les données de la FFA, environ 92% de cotisations par rapport au niveau de leurs collectes de primes en 2018 [3].

La lecture des tableaux 1.2 et 1.3 nous emmène à faire une double conclusion. Premièrement, il est à noter **une stagnation du parc automobile national**. Deuxièmement, il y a **une forte concurrence sur le marché de l'assurance auto**. Ces deux conclusions ont des conséquences directes sur le parcours client chez les différents fournisseurs : il est non seulement difficile de réaliser une affaire nouvelle - gain d'un nouveau client - mais aussi de garder un client chez soi. Plus encore, une succession d'évènements sur le marché de l'assurance auto renforce ces conclusions :

— Un nombre important de fournisseurs d'assurance auto sur le marché français - 92

3. Cela peut être expliqué par le retrait de 2 assureurs sur le marché d'assurance auto en 2019. Selon les données de la FFA, le nombre d'assureurs sur ce marché est passé de 94 en 2018 à 92 en 2019. Il s'agirait probablement des assureurs direct. (Je cherche encore les noms)

TABLE 1.3 – Evolution des cotisations par mode de distribution par année

Année	2016/2015	2017/2016	2018/2017	2019/2018
Assurances directes	4%	4%	-7%	-92%
Autres sociétés sans intermédiaires	-1%	-5%	-2%	2%
Bancassurance	5%	5%	5%	4%
Mutuelles sans intermédiaires	-1%	-1%	-2%	0%
Sociétés avec intermédiaires	-1%	0%	0%	4%

Note de lecture : Depuis 2015, on note une montée en puissance de la bancassurance alors que les assurances directes et les autres sociétés sans intermédiaires perdent des parts du marché.
Source : Données résultant des calculs sur les chiffres clés de l'assurance (2015-2019) - FFA

en 2019 - avec une prolifération de nouveaux acteurs notamment les bancassureurs et les start-ups ;

— Un marché dynamique avec de plus en plus de propositions d'offre d'assurance auto « à la carte » pour des segments de marché de plus en plus fins : professionnels, étudiants, retraités et personnes morales ;

— Des politiques de marketing agressif marquées par des budgets publicitaires de plus en plus élevés notamment sur les plateformes de *streaming* (YouTube, télévisions) et des pages internet ;

— Une modernisation du système de vente avec une adaptation aux nouveaux canaux de distribution tels que les comparateurs d'assurances en ligne, les appels téléphoniques pour proposer directement aux clients des produits qui pourraient les intéresser ;

— La digitalisation du processus de souscription, l'optimisation du questionnaire de souscription ou encore des méthodes plus courantes chez les bancassureurs caractérisées par une proposition de produits à des consommateurs déjà clients de la banque filiale.

1.3 Vie d'un contrat en assurance auto

Le contrat est un accord de volontés entre deux ou plusieurs personnes destiné à créer, modifier, transmettre ou éteindre des obligations selon l'article 1101 du Code civil [4]. En assurance, cette définition reste valable. En effet, par un contrat d'assurance, l'assureur s'engage à faire des prestations vis-à-vis de l'assuré lors de la réalisation d'un risque contre le paiement d'une prime. Et comme nous pouvons le voir dans la définition du code civil, un contrat peut subir des modifications au cours de sa vie, lesquelles constituent d'ailleurs un autre contrat. Ainsi, un contrat d'assurance auto a une « vie » que nous pourrons décliner en des termes suivants :

— **La souscription :** Elle marque le début du contrat. L'acte de souscription d'un

4. Le code civil français, la version du 28/07/2021

contrat d'assurance est un acte d'engagement du souscripteur après une apposition de sa signature sur la police d'assurance pour ainsi approuver les termes qui y sont inscrits. La durée du contrat d'assurance auto est d'une année avec tacite reconduction ;

— **Des modifications :** Un contrat d'assurance auto peut subir des modifications au cours de sa vie. Ces modifications peuvent advenir suite à un changement de régime matrimonial, de véhicule ou encore suite à un changement de formule par exemple. Certains changements génèrent des avenants alors que d'autres conduisent à la résiliation du contrat d'assurance ;

— **La résiliation :** La fin d'un contrat d'assurance auto intervient avec la résiliation. Cette résiliation du contrat peut advenir sur demande de l'assuré ou sur l'initiative de l'assureur. Nous irons plus en détail des différents motifs dans la section 2.3.2.

En somme, un contrat d'assurance naît (l'acte de souscription), peut subir des modifications au cours de sa vie et puis meurt soit par une résiliation ou soit par une arrivée à terme du contrat : l'échéance principale. Toutefois, il faut mentionner que l'acte de résiliation, s'inscrivant dans les droits du consommateur, est renforcé depuis quelques années dans le milieu assurantiel grâce à deux lois notamment : les lois Chatel et Hamon.

Dans la suite de cette partie, nous allons rappeler ces lois avant de conclure par les objectifs de l'étude.

1.4 Contexte et objectifs de l'étude

1.4.1 Contexte réglementaire : les lois Hamon et Chatel

Deux lois ont marqué les droits des consommateurs d'assurance en France, impactant par la même occasion les mouvements de résiliation des contrats d'assurance. Il s'agit des lois Hamon et Chatel.

Loi Chatel

A partir de 1989, en France, les particuliers qui ont des contrats d'assurance non vie avec tacite reconduction, peuvent le résilier annuellement avec un préavis de 2 mois avant la date d'échéance principale du contrat.

Mais, depuis précisément le 28 janvier 2005, une nouvelle loi a été promulguée, introduisant de nouvelles dispositions qui visent à faciliter la résiliation des contrats à tacite reconduction : il s'agit de la loi 2005-67, plus connue sous la dénomination « **Loi Chatel** ». L'assurance auto étant un contrat par tacite reconduction, est concernée par ces dispositions. Le 28 juillet 2008, 6 mois après la promulgation, ces dispositions seront reprises par le code des assurances.

Plus concrètement, les assureurs ont l'obligation d'informer annuellement(à l'échéance principale) la date limite à laquelle l'assuré a la possibilité de dénoncer la reconduction automatique de son contrat. Ce rappel doit parvenir à l'assuré au moins 15 jours avant la date limite, lequel pourrait alors décider de ne plus reconduire son contrat.

Loi Hamon

Publiée au Journal Officiel le 17 mars 2014, la loi Hamon s'inscrit dans le renforcement de la protection des assurés en complément de la Loi Chatel. La principale mesure de cette loi est qu'elle permet à un assuré de quitter plus facilement son assureur. Il peut désormais, après la première année de souscription, résilier son contrat quand il le souhaite, à certaines conditions, et ce, sans préjudice financier. Par ailleurs, dans son article 59, la loi Hamon oblige les assureurs à motiver les résiliations de contrats venant de ces derniers [5]. En effet, ils doivent indiquer à l'assuré les motifs de leurs décisions dans la lettre de résiliation. Ces motifs peuvent être par exemple *« résiliation suite à la surveillance de portefeuille »* ou encore *« résiliation pour aggravation de risque »*.

La conséquence directe de ces lois est qu'aujourd'hui, si un assuré n'est pas satisfait des prestations de son assureur, il peut résilier plus facilement son contrat et ce, sans aucune pénalité. Peut-être a-t-il trouvé un contrat couvrant les mêmes risques avec une meilleure prime ? Ou mieux a-t-il eu une mauvaise expérience avec son assureur dans la gestion d'un sinistre qu'il a subi ? Toutes les raisons sont bonnes pour résilier son contrat.

1.4.2 Objectifs de l'étude

Comme nous avons pu le constater à la section 1.2.3, le marché de l'assurance auto en France est très étroit et très concurrentiel. Par ailleurs, les différents fournisseurs, selon leur mode de distribution, ne subissent pas les modifications de portefeuille pareillement. Pour rappel, les bancassureurs gagnent des parts supplémentaires du marché et les assureurs directs sont en récession. A cela, s'ajoutent les différentes pratiques sur le marché d'assurance que nous avons énumérées dans la même section.

GAN Assurances, faisant partie des sociétés avec intermédiaires, n'échappe pas aux conséquences de ce marché concurrentiel. En effet, contrairement à la plupart de ses concurrents sur le marché qui utilisent plusieurs modes de distribution à la fois, GAN distribue ses produits exclusivement par le biais des intermédiaires. Ces intermédiaires sont constitués par son réseau d'agents qui sont des entrepreneurs indépendants et qui ne sont pas salariés de GAN. Ce choix du modèle d'affaires - *business model en anglais* - a

5. L113-12-1 : « La résiliation unilatérale du contrat d'assurance couvrant une personne physique en dehors de son activité professionnelle par l'assureur, dans les cas prévus au présent livre ou en application du premier alinéa de l'article L. 113-12, doit être motivée »

divers avantages :

— **Proximité avec le client :** Les agents vivent dans les quartiers et connaissent mieux les besoins des clients. Avec une étroite collaboration avec la compagnie GAN, ces agents promeuvent la marque GAN et permettent à cette dernière de proposer des produits plus flexibles afin de répondre aux besoins des clients ;
— **Une meilleure écoute du client :** Les agents sont disponibles et sont à l'écoute des besoins des clients. De ce fait, ils nouent une relation de confiance avec ces derniers. Aussi, faut-il rappeler que les agents sont rémunérés par des commissions et la performance de leur portefeuille (le paquet de contrats qu'ils ont et leur duration). Ils ont donc intérêt à satisfaire les clients et les garder dans leur portefeuille le plus longtemps possible ;
— **Un meilleur acte de gestion :** Lors de la survenance d'un sinistre par exemple, il est plus facile de se rapprocher de son agent et laisser ce dernier faire le nécessaire pour l'indemnisation. Cela devrait, si le postulat nous est permis, de réduire les délais de traitement des sinistres et par conséquent, un meilleur acte de gestion.

Malgré ce modèle d'affaires qui offre plusieurs avantages, le portefeuille GAN auto subit depuis des années, les mouvements techniques à l'instar de la résiliation dont il faut étudier les déterminants. En effet, la résiliation peut advenir par l'initiative de l'assuré ou par l'assureur suite à la surveillance. La surveillance de portefeuille est ce que nous pouvons définir par le fait de "surveiller" certains contrats à cause de certaines de leurs caractéristiques. Ces caractéristiques sont généralement l'antécédent de sinistre ou encore le nombre de sinistres graves.

Dans ce mémoire, nous nous focaliserons sur les résiliations qui dépendent exclusivement du vouloir de l'assuré. De ce fait, notre objectif est double. Le premier objectif est de **connaître le profil de risque vis-à-vis de la résiliation de chaque contrat actif sur 12 mois, y compris la résiliation à l'échéance principale**. Cela revient donc à affecter une probabilité de résiliation à tous les contrats vus en début de date d'effet de version jusqu'à la prochaine reconduction. Les modèles que nous étudierons par la suite nous permettront de détecter les profils à fort risque de résiliation et de comprendre leurs caractéristiques propres. Ensuite, le second objectif de ce mémoire sera de **faire des recommandations pour mieux anticiper les contrats susceptibles d'être résiliés pour par exemple prendre des actes commerciaux qu'il faut pour les fidéliser**. La conséquence lointaine serait, s'il faut le mentionner, une amélioration du résultat technique.

1.5 Conclusion

Dans cette première partie, nous avons pu comprendre ce qu'est un produit d'assurance auto, les lois qui entourent sa distribution et les différents mouvements auxquels il peut être sujet. Nous avons conclu que le marché d'assurance auto en France est très concurrentiel avec un parc national de véhicules assurés en faible progression. Par conséquent, les assureurs se battent tous de leurs côtés pour fidéliser leurs clients car, avec les lois Hamon et Chatel, ces derniers peuvent résilier leurs contrats à tout moment.

Pour arriver à leurs fins, les assureurs ont besoin de comprendre les déterminants de ce phénomène et de savoir l'ensemble des paramètres sur lesquels ils peuvent agir pour conserver leurs clients le plus longuement possible. C'est à ce jeu que ce mémoire se prête. Les assureurs s'investissent aussi dans leur quotidien pour gagner de nouveaux clients. Pour ce faire, ils adoptent plusieurs stratégies. Nous pouvons citer par exemple des programmes de publicité, une simplification du questionnaire de souscription et sa digitalisation, une segmentation de plus en plus fine des profils de risques ou encore une guerre de prix pour gagner des affaires nouvelles.

Afin d'étudier la propension à résilier son contrat d'assurance auto, il est important de constituer une base de données renfermant les informations qui pourraient expliquer ce phénomène. C'est ce qui fera l'objet du chapitre suivant.

Bases de données et statistique univariée

L'étude des résiliations nécessite des données variées sur les contrats. Autant les caractéristiques du contrat sont importantes, il n'en est pas moins pour des données liées à l'acte de gestion, à l'évolution tarifaire d'année en année. Pour constituer une base de données de travail avec ces différentes informations, nous avons à notre disposition plusieurs bases de données. La description de ces données et le processus de la constitution de la base de données finale de travail seront les contenus de la première partie de ce chapitre. Ensuite, s'en suivront les premiers résultats sur la dynamique de résiliation dans le portefeuille. Il sera question notamment de faire une première étude statistique univariée.

2.1 Les différentes données disponibles pour l'étude

Nous avons essentiellement deux sources de données : les données internes et les données externes. Les données internes sont récupérées directement avec SAS via l'accès à une bibliothèque. Les données utilisées vont du 1er janvier 2017 au 31 décembre 2020.

2.1.1 Les données internes

Nous avons 5 types de données internes que sont les bases image, risque, résiliation, sinistre et client.

Base image

La base image contient tous les mouvements techniques des différents contrats : les contrats actifs - AC - dont les affaires nouvelles - AN , les contrats résiliés - AR, les contrats remplaçants - RM, ou encore les contrats remplacés RME.

Elle contient n observations et m variables. Elle renferme uniquement des informations qui servent à identifier les contrats mais ne contient pas de données sur les critères tarifants [1]. Il y a également toutes les versions de tous les contrats. La première version d'un contrat serait le tout premier contrat souscrit par un nouveau client ; une seconde version serait le même contrat après reconduction et ainsi de suite pour les versions 3 et plus. Cela permet d'avoir des informations sur l'ancienneté du contrat et mesurer le degré de fidélité des assurés.

La variable IDDHK est l'identifiant de chaque contrat avec distinction de la version (num_image dans l'aperçu dans le tableau 2.1). Dans la pratique, on utilise plutôt l'identifiant du contrat sans distinction de la version IDDHL couplé avec les dates de début et de fin de contrat pour identifier de façon unique un contrat.

1. Les critères tarifants sont notamment les caractéristiques du véhicule et de l'assuré - le conducteur

Dans l'exemple de la base de données affichée ci-dessous, nous avons les contrats visibles de 2 clients différents. Le premier client **IDDHL=00002** a 3 versions de contrat. La version 3 du contrat sera résiliée le 30/06/2020. Dans l'écriture dans la base image, on remet le même contrat et on corrige la date de fin de version pour la version en cours et on reporte la résiliation en mettant comme date de début de version, la date d'effet de la résiliation.

IDDHL	Num_image	date début de version	date fin de version	date fin de version corrigée	statut du contrat	....
00002	2	01/01/2019	31/12/2019	31/12/2019	AC	...
00002	1	01/01/2018	31/12/2018	31/12/2019	AC	...
00002	3	01/01/2020	31/12/2020	30/06/2020	AC	....
00002	1	01/07/2020	31/12/2020	31/12/2020	AR	...
...	...	...	...	....	...	..
00021	1	01/02/2019	31/01/2020	31/01/2020	AC	...

TABLE 2.1 – Aperçu de la base image

Base résiliation

Cette base de données contient exclusivement des données sur les résiliations et surtout sur ses motifs. Les motifs nous permettront de délimiter notre périmètre de travail et de regarder de près la répartition du taux de résiliation par motif. La clé IDDHL avec les dates de début et de fin de version permettent de faire la jointure avec les autres données.

Base Risque

Elle renferme des informations sur les caractéristiques des véhicules et de l'assuré. C'est cette base qui contient les critères tarifant de tous les contrats actifs. On y trouve les identifiants IDDHL et les dates de début et de fin de version des contrats.

Base Sinistre

Cette base contient des informations sur les sinistres survenus par contrat et par garantie ainsi que les charges de garantie sinistrée de tous les produits de GAN. Chaque ligne correspond à une garantie sinistrée. Nous avons extrait le produit qui nous intéresse : GAN AUTO, pour avoir les informations sur les sinistres survenus en assurance Auto.

Il faut noter qu'il y a une différence entre charge de sinistre et coût réel de sinistre. En effet, la charge de sinistre correspond au montant réellement déboursé par l'assureur alors que le coût réel du sinistre est la charge majorée par les recours éventuels. Nous ne disposons dans notre base sinistre que des données sur la charge de sinistre. De ce fait, il n'est pas inopportun de rencontrer des charges négatives. On y trouve également des informations sur le degré de responsabilité de l'assuré sur un sinistre du type RC ou Dommages. IDDHL permet d'identifier le contrat, NODOSIN (numéro de dossier) permet d'identifier le sinistre et la date de survenance du sinistre permet de rattacher le sinistre à la bonne version du contrat. En effet, du fait que les dates de début et de fin de version

du même contrat ne chevauchent jamais, il suffit de vérifier que la date de survenance du sinistre est entre ces deux dates pour lier le sinistre au bon contrat.

Base Client

La base client est une base commune à tous les produits GAN. Elle renferme les informations sur l'ensemble des différents contrats détenus par un assuré chez GAN ASSURANCES. Par exemple, si un assuré à 1 contrat en assurance auto et 1 contrat en assurance multirisque habitation -MRH, ces informations y figurent. Cette base nous apporte une information supplémentaire sur le poids du client dans notre portefeuille. On y trouve aussi des informations sur la situation socio-professionnelle du client. Peut-être un étudiant résilierait plus qu'un haut cadre ?

2.1.2 Les données externes

Après une prospection, nous avons soupçonné qu'il peut y avoir un effet géographique sur la dynamique des résiliations. Pour inclure ces données, nous avons récupéré les données sur les unités urbaines de l'INSEE - **Base des unités urbaines 2020 au 1er janvier 2020** - que vous pouvez trouver en cliquant ici. L'objectif est de localiser le garage du conducteur et de savoir s'il se trouve dans une zone urbaine, rurale ou dans une banlieue.

Nous avons exploré la piste des données sur la concurrence notamment une prime du marché par profil de risque mais la tâche s'avère compliquée. En effet, il faudrait faire du web-scraping pour récupérer cela ; ce qui n'entre pas dans la pratique de GAN ASSURANCES et de sa Direction Technique. En effet, cela peut engendrer des poursuites judiciaires car le web-scraping peut ralentir le site de souscription par exemple des concurrents, lesquels pourront perdre de nouveaux clients potentiels.

2.2 Constitution de la base de données

Le pré-traitement des données est une étape importante dans toute étude faisant intervenir les données. C'est sans doute l'étape la plus fastidieuse et déterminante pour l'obtention des résultats reflétant l'observation. Dans cette section, nous allons décrire pas à pas les différentes transformations et travaux que nous avons réalisés pour obtenir la base de données finale qui nous servira par la suite dans l'implémentation des modèles.

2.2.1 Jointure des tables

La jointure des tables constitue la première étape de notre processus de constitution de la base de données de travail. Elle va nous permettre de mettre en un seul bloc les différentes données susmentionnées. Les différentes jointures effectuées sont résumées sur

la figure 2.1. Nous rappelons que la constitution de la base de données s'est faite sur SAS avec la méthode ***left join*** de la procédure ***proc sql***.

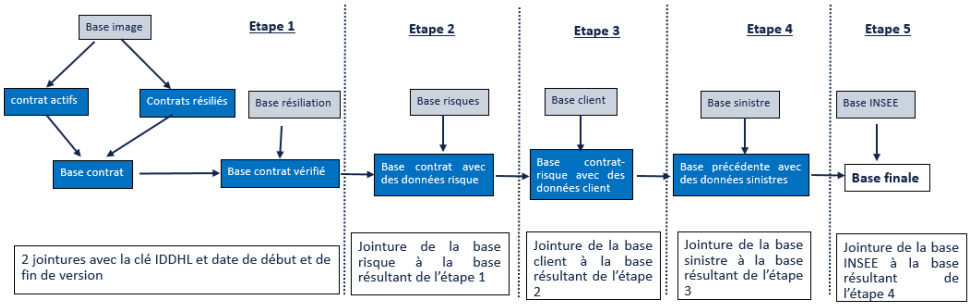


FIGURE 2.1 – Les différentes étapes de jointures des bases de données

- **Étape 1 :** Cette première étape consiste à identifier les contrats résiliés. De ce fait, nous avons d'abord extrait de la base image deux sous groupes d'observations : les contrats résiliés (AR) et les contrats actifs -AC. Ensuite à l'aide de IDDHL et de la relation existante entre les dates de début de version et de fin de version corrigée respectivement de l'observation (AR) et celle (AC), nous avons indexé les contrats résiliés dans notre base des contrats actifs. En clair, nous indexons la version du contrat actif qui est résilié par la suite.
  En prenant le même exemple que sur le tableau 2.1, on obtient un tableau suivant avec la création de la variable « **Resiliation** » qui prend la valeur 1 si le contrat est résilié par la suite et 0 sinon :

IDDHL	Num_image	date début de version	date fin de version	date fin de version corrigée	statut du contrat	....	Resiliation
00002	2	01/01/2019	31/12/2019	31/12/2019	AC	...	0
00002	1	01/01/2018	31/12/2018	31/12/2019	AC	...	0
00002	3	01/01/2020	31/12/2020	30/06/2020	AC	....	1
1 ...	...	...	...	....	...	..	...
00021	1	01/02/2019	31/01/2020	31/01/2020	AC	...	0

TABLE 2.2 – Aperçu de la base contrat

  Par la suite, nous avons réalisé une seconde jointure avec la **base résiliation** pour rattacher les informations sur les motifs de résiliation. Cette jointure a aussi permis de contrôler si les contrats résiliés sont bien indexés. On obtient à la fin de cette étape la **base contrat vérifié** comme sur la figure 2.1 ;

- **Étape 2 :** A la base obtenue à l'étape 1, nous avons joint la **base risque** pour avoir les caractéristiques de chaque version de contrat. La clé de jointure ici est l'identifiant IDDHL et les dates de début de version et de fin de version corrigé. On obtient à cette étape la **base contrat avec des données risque ;**

- **Étape 3 :** Pour avoir des informations sur le client (poids du client chez GAN et aussi sa situation socio-professionnelle), nous avons joint à la base précédente la **base client**. La clé est IDDHL et les dates ;

- **Etape 4 :** L'antécédent de sinistralité des différentes versions de contrat s'avère importante d'après un *benchmark* dans les mémoires traitant la résiliation. De ce fait, un premier traitement a été réalisé sur la base sinistre. Après le calcul des charges des sinistres, du nombre de sinistres pour chaque contrat et par garantie, du délai de traitement du dossier de sinistre, du degré de responsabilité, nous avons joint ces informations à la base obtenue à la fin de l'étape précédente. Nous avons réalisé le même exercice mais sur les antécédents de sinistre pour chaque contrat sur les 12 mois, 24 mois, 36 mois, 48 mois et 60 mois vu à partir de la date de début de version de contrat ;

- **Étape 5 :** Enfin, pour avoir des données géographiques, nous avons rattaché à chaque contrat, à l'aide de l'adresse du garage du véhicule, des données de l'INSEE sur les unités urbaines. On obtient la base de données brute finale.

Ce processus de jointure a permis de construire une base de données unique qui contient des informations sur les critères tarifant, les données clients ou encore les antécédents de sinistres.

2.2.2 Les catégories de variables

La constitution de cette base de données finale s'est faite suite à une recherche des variables potentielles qui peuvent expliquer les résiliations et aussi suivant la pratique métier. L'ensemble des variables peuvent être catégorisées en plusieurs groupes :

Les caractéristiques du contrat

Les caractéristiques du contrat sont les critères tarifants c'est-à-dire des variables qui ont permis de calculer la prime du contrat. On distingue entre les caractéristiques du véhicule (type de véhicule, usage de véhicule, cylindré du véhicule, marque ...), les caractéristiques du conducteur (âge du conducteur, situation socio-professionnelle, âge du permis du conducteur, type de permis de conducteur) et l'ancienneté du contrat ou encore l'ancienneté de souscription d'un contrat d'assurance auto. Un récapitulatif est sur le tableau 2.3.

Acte de gestion du contrat

Nous désignons par acte de gestion du contrat, toutes les variables qui donnent des informations sur les environnements sinistre et client du contrat. L'environnement sinistre

Variables	**Libellé**
Caractéristiques du véhicule	
energie_veh	Énergie utilisée par le véhicule
classe	La classe du véhicule
Genre_vehicule	Le genre de véhicule
Marque	La marque du véhicule
groupe	groupe du véhicule
usage	usage du véhicule
cylindre	Le cylindré du véhicule
Caractéristiques du conducteur	
situation_pro	La situation professionnelle du conducteur
tr_age_assure	La tranche d'âge de l'assuré
anc_permis	L'ancienneté de permis de conduire
Caractéristiques propre du contrat	
anc_contrat	Ancienneté du contrat - mesure la fidélité chez GAN
anc_souscription	ancienneté de souscription d'une assurance auto

TABLE 2.3 – Quelques variables - caractéristiques du contrat

renferme les informations sur la sinistralité du contrat et la façon dont ce dernier est traité. Le délai de traitement des sinistres, le nombre de sinistres, type de garantie affectée (Bris de Glaces, Dommages etc...), les charges de sinistres sont quelques unes des variables de cette sous-catégorie.

Par l'environnement client, nous désignons les variables qui décrivent le type de client détenteur du contrat. On y trouve le nombre de contrats que le client détient chez GAN, s'il possède uniquement des contrats dans un seul produit ou non.

Prime et évolution tarifaire

Cette catégorie décrit les variations de primes en euro et en pourcentage par année, les avantages et les gestes commerciaux.

- *tx_derog :* c'est le taux de dérogation qui est la somme des gestes commerciaux et les avantages du contrat. C'est un instrument couramment utilisé par les assureurs pour fidéliser leurs clients. Nous verrons donc par la suite si cette variable est très déterminante pour la propension à résilier ;
- *Cat_Var_prime :* C'est la variation de la prime commerciale d'assurance d'une version sur une autre regroupée par intervalle. Cela reflète d'une part l'évolution tarifaire et d'autre part, la politique tarifaire ;
- *Cat_delta_prime :* C'est comme la variable précédente à la seule différence qu'elle est en %. L'objectif est de voir *in-fine* si l'assuré raisonne en euros ou en variation absolue (en %).

Variables	Libellé
Environnement sinistre	
delai_trait_meean	délai moyen de traitrement du sinstre du contrat en cours
delai_trait_12_mean	délai moyen de traitement des antécédents de sinistres 12 mois
delai_trait_24_mean	délai moyen de traitement des antécédents de sinistres 24 mois
nb_sin	Nombre de sinistres du contrat en cours
nb_sin_resp	nombre de sinistres responsables du contrat en cours
cout_sin	La charge de sinistres du contrat en cours
nb_sin_12_an	antécédent de sinistre 12 mois
nb_sin_24_an	antécédent de sinistre 24 mois
nb_sin_an	antécédent de sinistre globale
sin_dom	nombre de sinistres dommages du contrat en cours
sin_rc	nombre de sinistres RC du contrat en cours
sin_bdg	nombre de sinistres bris de glaces du contrat en cours
sin_dom_an	nombre de sinistres dommages du contrat en cours
sin_bdg_an	nombre de sinistres bris de glaces du contrat en cours
sin_rc_an	nombre de sinistres RC du contrat en cours
On fait la même chose en terme de charge	de sinistres des contrats en cours et pour antécédents
Environnement client	
MONOCONT	Indicateur qui indique si le client n'a souscrit qu'à un seul contrat
MONOPROD	Indicateur qui indique si le client n'a souscrit qu'à un seul produit
nb_contrat	nombre totale de contrats qu'a le client chez GAN -tout produit confondu
prime_totale	La prime totale des contrats détenus par l'assuré à GAN -Tout produit confondu

TABLE 2.4 – Quelques variables - acte de gestion

Typologie de résiliation

Il s'agit du motif de résiliation - *motif*. Cela nous permettra de limiter notre périmètre de travail en se concentrant uniquement sur les résiliations qui émanent de l'initiative de l'assuré.

Positionnement géographique

Il s'agit des variables qui permettent de situer le lieu de garage du véhicule de l'assuré. On distingue les variables suivantes :

- *departement :* Il renseigne le département dans lequel se situe le garage. Cela nous permettra de faire une étude géographique pour voir s'il y a une disparité entre les départements en termes de risques de résiliation du contrat ;
- *TYPE_COM :* Une variable qui apporte plus d'informations sur le lieu. Elle permet de savoir si la localité est une zone urbaine ou rurale ;
- *STATUT_COM :* Cette variable reprend les mêmes informations que *TYPE_COM* avec une ramification plus fine. Elle renseigne si la localité est un centre-ville, une ville isolée ou une banlieue.

2.2.3 Traitement des valeurs manquantes

Le traitement des valeurs manquantes diffère d'une variable à une autre. En premier lieu, les contrats actifs dans la base image qui n'ont pas pu être rattachés à des informations de la base risque sont tout simplement et purement supprimés. Ensuite, nous observons les variables une à une et selon leur signification, nous essayons de comprendre

ce qui pourrait signifier un non remplissage de cette dernière ; suite à quoi nous décidons du traitement approprié. Par exemple :

— *anc_permis :* Le non renseignement du permis de conduire et de son ancienneté suppose soit que le gestionnaire a oublié de remplir ces informations, ou soit que le contrat ne nécessiterait pas ces informations pour la détermination de la prime. Or, sachant que l'ancienneté du permis de conduire est un critère tarifant (d'après le tarificateur GAN Auto) pour les particuliers, les deux hypothèses ne semblent pas être les bonnes. Il s'agirait donc d'un contrat détenu par une personne morale (entreprise). Les valeurs manquantes sont donc remplacées par « *Personne morale*» ;
— *tr_age_assure :* Le traitement des valeurs manquantes de l'âge de l'assuré est le même que celui de l'ancienneté du permis de conduire ;
— *le nombre de sinistres, les charges de sinistres, les délai de traitements :* L'absence de ces informations signifie juste que le contrat n'est pas sinistré. Les valeurs manquantes sont remplacées par le chiffre 0.
— *Autres variables :* Pour un certain nombre de variables, nous avons préféré remplacer les valeurs manquantes par « *Non renseigné* ». L'objectif est de ne pas complètement supprimer ces contrats, ce qui induirait une perte d'information et surtout du fait que nous ne sommes pas sûr de l'importance de ces variables et du pourquoi certaines modalités ne sont pas renseignées.

2.3 Statistique descriptive

Après cette première partie de préparation de données, nous allons maintenant entamer une première analyse univariée. Pour ce faire, nous allons dans un premier temps regarder la structure du portefeuille de 2017 en 2020. Ensuite, l'axe central de ce mémoire étant l'étude de la résiliation, nous allons décliner les statistiques univariées de quelques variables explicatives et les différents motifs de résiliations.

2.3.1 Structure du portefeuille et la crise sanitaire 2020

Notre point de départ est de regarder les statistiques générales sur les mouvements techniques. Nous nous intéresserons ici à trois types de mouvements : les affaires nouvelles [2] (AN), les résiliations (AR) et les contrats en portefeuille (PTF). Les PTF sont les contrats qui sont actifs en fin d'exercice et les AR sont les résiliations de contrats enregistrés au cours de l'année civile. Les résiliations sont prises à leurs dates d'effet. L'objectif est de voir l'état de santé du produit GAN Auto.

2. Une affaire nouvelle est un contrat qui n'est rentré en portefeuille que dans l'année d'exercice considéré.

... Contrats en PTF, AR, AN et taux de renouvellement

Sur les définitions données aux PTF, AN et AR plus haut, nous nous baserons sur les formules suivantes pour calculer les taux de résiliation (*Taux de AR*), d'affaires nouvelles (*Taux de AN*) et de renouvellement de portefeuille (*Taux de renouvellement*).

$$\boxed{Taux\ de\ AR = \frac{AR}{PTF + AR}} \tag{2.1}$$

$$\boxed{Taux\ de\ AN = \frac{AN}{PTF}} \tag{2.2}$$

$$\boxed{Taux\ de\ renouvellement = taux\ de\ AN - taux\ de\ AR} \tag{2.3}$$

Nous affichons la taille du PTF en base 2017 dans le tableau 2.5. En d'autres termes, nous rapportons les tailles du portefeuille des autres années sur sa taille en 2017. On remarque de manière générale qu'il y a une diminution du nombre de contrats en portefeuille entre 2017 et 2019. En 2020, on observe une légère augmentation par rapport à 2019 mais son niveau n'atteint pas celui de 2017 ; ce qui est la conséquence d'une diminution du taux de résiliation en 2020.

Le portefeuille affiche un taux de résiliation, tout motif confondu, de 16,7% entre 2017 et 2020. Sur la même période, le PTF de GAN est constitué de 15.3% de nouveaux contrats. Ce qui fait qu'il enregistre une perte de 1,4% par rapport à 2017. Par ailleurs, comme nous pouvons le remarquer sur le tableau 2.5 et la figure 2.2, le taux de résiliation, tout motif confondu, évolue entre 2017 et 2019 avant de chuter drastiquement en 2020. La raison de cette évolution est la migration des anciens contrats vers le nouveau produit GAN Auto et la mise en place des outils de surveillance de contrats plus poussés. Ainsi, des anciens contrats sont résiliés par GAN et remplacés par de nouveaux.

Année	PTF	Taux de AR	Taux de AN	Taux de renouvellement
2017	100%	14,7%	11,0%	-4%
2018	93%	19%	13,8%	-5%
2019	91%	20.9%	19,9%	-1%
2020	97%	12,3%	17, 1%	5%
TOTAL	-	16,7%	15,3%	-1%

TABLE 2.5 – Structure du portefeuille entre 2017 et 2020

... Zoom sur l'année 2020 : Effet de la crise sanitaire liée au Covid-19

La structure du portefeuille semble différente en 2020 par rapport aux autres années. Nous notons une perte d'environ 9 points de pourcentage du taux de résiliation, engendrant un gain de contrats (taux de renouvellement positif) de 5%. La raison principale de

cette observation est la crise sanitaire liée au Coronavirus. La crise sanitaire a engendré une diminution de la sinistralité et vraisemblablement, une diminution de la circulation de véhicules assurés. Cette observation conforte nos soupçons du rôle de l'acte de gestion (sinistralité et gestion de sinistres) dans les résiliations des contrats par les assurés. Suite à ces remarques, nous nous sommes posés une question importante dans notre processus de modélisation : **faut-il considérer les observations de 2020 pour la modélisation ? Se faisant, ne minimiserons-nous pas la probabilité de résiliation globale de notre portefeuille** ? Nous apporterons des éléments de réponses en fin de ce chapitre.

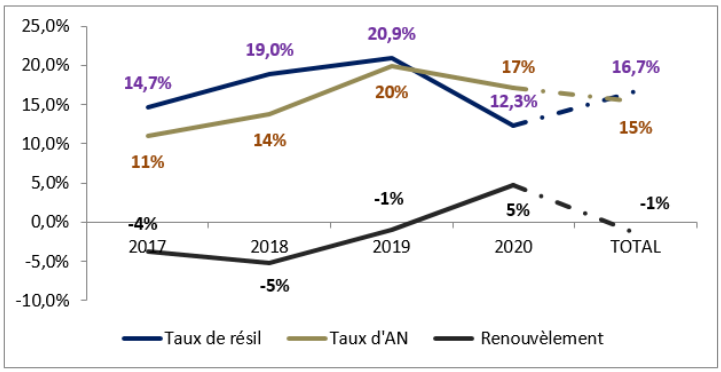


FIGURE 2.2 – Évolution de la structure du portefeuille

... Comparaison de la sinistralité des contrats en PTF Vs les contrats résiliés

La sinistralité en assurance IARD sur une période donnée se mesure à travers la fréquence (FQ) et le coût moyen (CM). De manière générale, ces indicateurs sont calculés par exercice - année civile. Le coût moyen permet de quantifier la charge moyenne nette de recours d'un sinistre. Il se calcule par :

$$\boxed{CM = \frac{Charge\ totale}{Nombre\ de\ sinistres}} \tag{2.4}$$

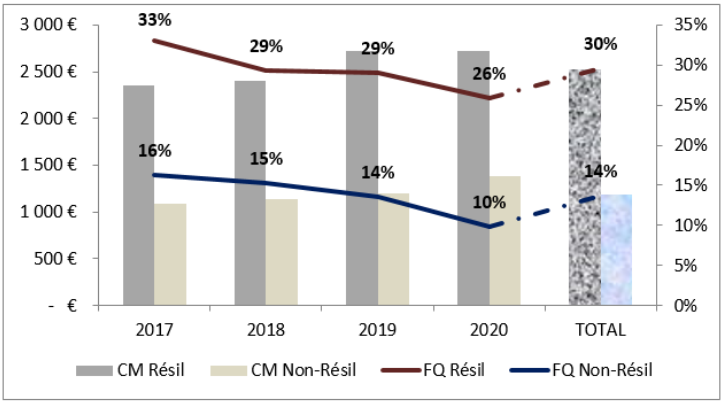
La fréquence de sinistre mesure la probabilité de survenance d'un sinistre par période d'exposition donnée - typiquement sur un an. Elle se calcule de la manière suivante :

$$\boxed{FQ = \frac{Nombre\ de\ sinistres}{Exposition\ au\ risque}} \tag{2.5}$$

Les sinistres considérés sont les sinistres hors sans suites et hors nuls.

L'exposition au risque mesure le temps réel pendant lequel l'assuré est couvert par l'engagement pris par l'assureur vis-à-vis de ce dernier. Elle est donnée. C'est le nombre de jours objet d'une garantie sur 360. Par exemple, un contrat d'assurance qui couvre la période du 01/06/2017 au 31/05/2018 a une exposition de 0,5 pour l'année 2017 et une exposition de 0,5 pour l'année 2018.

FIGURE 2.3 – Fréquence et coût moyen des contrats résiliés Vs Non résiliés



Note de lecture : La sinistralité des contrats résiliés est presque le double de celle des contrats en portefeuille. De manière globale, on note une fréquence de sinistre moyenne de 30% et un coût moyen d'environ 2500 euros pour les contrats résiliés. Ces indicateurs en 2020 sont un peu plus à la baisse ; une observation qui est due à la crise sanitaire liée au coronavirus.

Nous rappelons que la fréquence est calculée au niveau contrat et non par garantie. Prenons un contrat qui a eu un sinistre affectant 2 garanties (RC et Dommages par exemple). Puisque c'est le même contrat, nous considérons que c'est un et un seul sinistre. Une considération des garanties reviendrait à dire qu'il y a 2 sinistres.

La charge du sinistre est la charge globale nette de recours de toutes les garanties affectées pour un sinistre donné. La figure 2.3 montre que les contrats résiliés ont une plus forte sinistralité par rapport aux contrats en portefeuille. Par ailleurs, cette sinistralité a un niveau plus bas en 2020 avec une perte de fréquence d'environ 4 points de pourcentages par rapport à 2019. Le coût moyen est resté presque stable sur les 4 années avec une légère inflation chaque année.

Les observations en 2020 sont attendues et sont justifiables par la crise covid. Avec la crise sanitaire, il y a eu des confinements ; lesquels ont eu comme conséquence la diminution de la circulation de véhicules, minimisant ainsi l'exposition réelle au risque de sinistres du type accidents de circulation et dommages principalement. La confrontation des données sur la circulation des véhicules en France en 2020 et les fréquences mensuelles de sinistres en assurance auto toute garantie comprise permettent de mieux déceler l'effet covid.

Pour mieux se rendre compte visuellement de nos affirmations, nous avons calculé l'évolution de la fréquence mensuelle au cours de l'année 2020 par rapport à celle du mois de janvier ainsi que l'évolution de la circulation par rapport à celle enregistrée en janvier 2020.

Un premier résultat justificatif :
L'observation du nuage de points de la variation de la fréquence en fonction de la variation de la circulation de véhicule semble s'ajuster à une droite comme nous pouvons le voir sur la figure 2.3.1. Pour confirmer cela, nous avons réalisé une régression linéaire simple. Le modèle est bien ajusté avec une p-valeur de $7,77e-5$, le coefficient associé à la circulation avec la même p-valeur et un coefficient de détermination de $R = 85,9\%$.

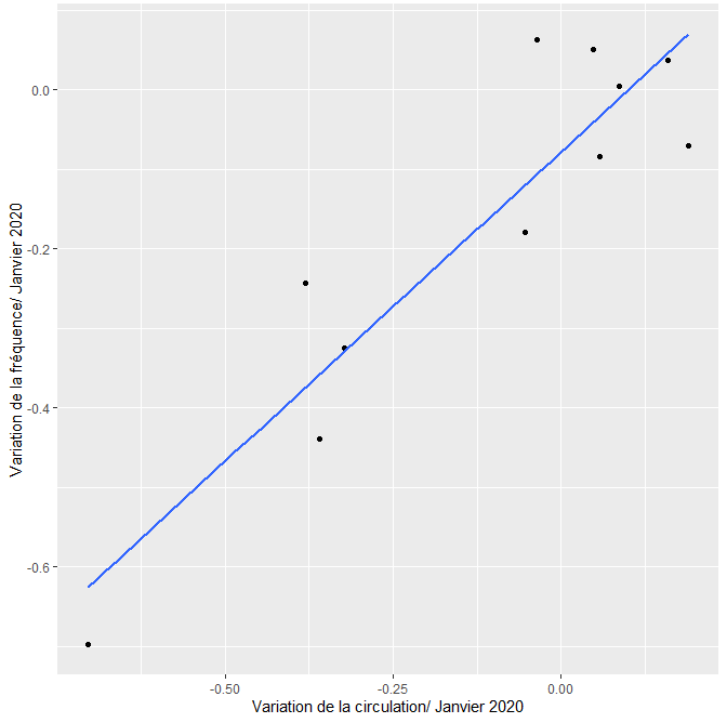


FIGURE 2.4 – Nuage de points des évolutions de circulation de véhicules et de la fréquence en 2020

Un second résultat révélateur :

Le second résultat est l'allure jointe des 2 courbes sur la figure 2.5. Ceci s'apparente à la justification précédente mais il est intéressant de regarder de prêt les effets des confinements. Nous remarquons une diminution de la circulation générale par rapport à son niveau de janvier 2020 sur toute l'année. La baisse de circulation a atteint son pic (-70%) au mois d'avril, marquant le début du premier confinement. Nous pouvons voir aussi clairement le début du deuxième confinement qui enregistre la seconde baisse la plus importante sur l'année ; soit au mois de novembre (-32%). La fréquence suit la même tendance sur l'année. Les gains de fréquence et de circulation enregistrés entre le mois juillet et septembre seraient plutôt dus à des effets saisonniers et au déconfinement.

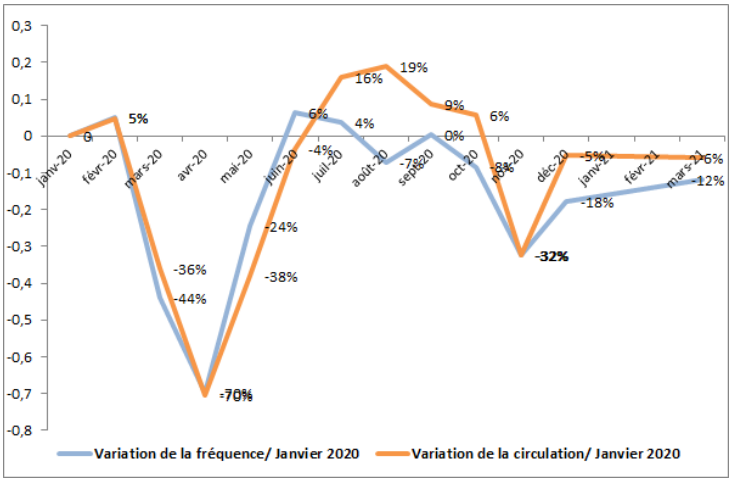


FIGURE 2.5 – Évolution de circulation de véhicules et de la fréquence en 2020

Au vu de ces deux résultats ci-dessus, il est clair que la fréquence est fortement liée à la circulation des véhicules en France. Par conséquent, c'est l'explication plausible qu'il faut retenir sur les observations mitigées sur la sinistralité en 2020.

Dans cette sous-section, deux conclusions peuvent être retenues. Premièrement, pour tout motif de résiliation confondu, le taux de résiliation augmente d'année en année, à un niveau plus élevé que les affaires nouvelles. La conséquence directe en est que GAN a perdu sur les 4 dernières années, 3% de marché comparativement au niveau de son portefeuille en 2017. Deuxièmement, la sinistralité des contrats résiliés fait presque le double de celle des contrats non résiliés. De ce fait, il est trivial de soupçonner la sinistralité comme variable déterminante pour l'explication de la résiliation dans les modèles que nous mettrons en oeuvre dans les chapitres suivants. Dans la suite de ce chapitre, nous allons voir plus en détail les résiliations par motifs et par modalités de quelques variables qui se sont avérées discriminantes suite à une première analyse univariée.

2.3.2 Un aperçu des résiliations par motif

Comme nous l'avons rappelé dans la section 1.4.1, les résiliations émanant de l'assureur doivent être motivées. De même, les résiliations venant de l'initiative de l'assuré sont enregistrées en interne avec un motif. Quels sont ces motifs ? Quelle est la répartition de la résiliation par rapport aux différents motifs ? Nous apporterons dans les lignes suivantes des éléments de réponses à ces interrogations.

Les différents motifs de résiliation

Pour une étude plus succincte, nous avons regroupé les différents motifs en des groupes. La liste exhaustive peut être consultée en annexe A.2.1. On distingue après regroupement :

- **Vente de véhicule :** Ce motif concerne les résiliations qui se font suite à une vente du véhicule assuré par le client. N'ayant plus de support de risque - absence du risque (le véhicule), le contrat est résilié ;
- **Initiative de l'assuré :** Ce motif, qui peut s'appliquer à la plupart des autres motifs dans lesquels la demande de résiliation vient de l'assuré, renferme dans notre cas des résiliations à échéance (Loi Chatel) et résiliation suite à refus tarifaire ;
- **Concurrence :** Il concerne des résiliations qui interviennent sur demande de l'assuré pour une souscription chez un concurrent (les autres distributeurs d'assurance auto en France - AXA, Allianz, MACIF ... - etc) ;
- **Remplacement :** La résiliation par remplacement concerne pour la plupart des contrats que nous pouvons surnommer "résiliation-affaire nouvelle". En effet, c'est de la ressouscription de contrat d'assurance. Lorsqu'il y a par exemple un ajout de risque, migration de contrat, l'ancien contrat est tout simplement résilié et un nouveau contrat est acté pour le remplacer. Il faut noter que ces types d'affaires nouvelles ne sont pas considérées comme telles et par conséquent, prennent la valeur 0 pour la variable *top_an* qui identifie les affaires nouvelles qui sont véritablement un gain d'un nouveau client ;
- **Hamon :** Ce motif concerne les résiliations dont le motif renseigné par les gestionnaires dans l'outil de gestion est principalement "Hamon date classique" avec l'assuré comme demandeur de résiliation du contrat ;
- **Changement de situation :** Ce motif intervient lorsque l'on résilie son contrat suite à un changement de type de contrat, un changement de situation socio-professionnelle ou encore un départ à l'étranger ;
- **Destruction de véhicule :** Ce motif concerne les résiliations selon les dispositions indiquées dans l'article L121-9 du Code des Assurances, suite à une destruction de véhicule qui peut intervenir après un grave accident ou encore pour un véhicule hors usage ;
- **Compagnie :** Il regroupe l'ensemble des motifs valables dans le cadre de la loi Hamon dans son article 59 qui stipule la motivation nécessaire d'une résiliation venant de l'assureur. On y trouve par exemple la suspension de contrat suite à un non paiement, la résiliation suite à une déclaration inexacte, la résiliation à cause d'un contentieux manuel entre les deux parties ou encore la résiliation suite à la surveillance de portefeuille ;
- **Autres motifs :** Nous avons regroupés le reste des motifs possibles, avec des poids mineurs dans "autres" motifs de résiliation. On peut y trouver par exemple le décès de l'assuré (article L121-10 du Code des assurances).

Pour rappel, notre objectif est de comprendre les **résiliations qui interviennent suite à une demande de l'assuré.** De ce fait, nous verrons par la suite les motifs que nous allons conserver et définir le périmètre final de notre étude.

Répartition des résiliation par motif et par année

Le tableau 2.6 représente la proportion des résiliations d'un motif sur l'ensemble des résiliations dans l'année de tous les motifs confondus. Nous observons qu'il y a des motifs plus récurrents que d'autres et la répartition n'est pas très stable dans le temps. Entre 2017 et 2018, *la vente du véhicule* et *initiative de l'assuré* sont les motifs les plus prépondérants dans notre population de contrats résiliés. A partir de 2019, on observe un changement structurel dans les données. En effet, le motif *remplacement* devient plus récurrent, passant de 13% à 30% en moyenne par année entre 2019 et 2020. A contrario, le motif *initiative de l'assuré* chute de presque 50% pour 10% en moyenne sur les 2 dernières années. Ce changement structurel peut être justifié par la migration des contrats, ce qui induit le remplacement de plusieurs contrats. La plupart de ces contrats sont, comme nous avons mentionné plus haut, des « *résiliation-affaire nouvelle* » ou encore de la ressouscription.

Motif	2017	2018	2019	2020	Total 2017-2020
Vente de véhicule	25	28	34	34	30
Initiative de l'assuré	63	47	12	8	33
Concurrence	0	4	12	13	7
Remplacement	5	13	30	31	20
Hamon	3	3	4	5	4
Changement de situation	0	1	4	5	2
Destruction de véhiucle	0	1	3	2	2
Compagnie	2	2	1	1	2
Autres	0	0	0	0	0

TABLE 2.6 – Répartition des résiliation par motif et par année en % des Résiliations

Fiabilité des données

La question de la fiabilité des données est centrale dans une étude basée sur l'analyse des données. En effet, à travers ces analyses, nous tentons de comprendre un phénomène - problème d'interprétation - ou encore, de prédire l'avènement du phénomène - problème de prédiction. Ainsi, si les études se font sur des données qui ne sont pas fiables, la conclusion est triviale : les résultats obtenus sont caduques et ne seront à coup sûr d'aucune utilité.

C'est pourquoi nous nous sommes posés la question sur la fiabilité des données que nous utilisons dans cette étude. Les données étant internes et la plupart des études en internes sont basées sur ces données, nous ne doutons pas de leur fiabilité. Néanmoins, en regardant de plus près les observations en termes de motifs de résiliation, le doute s'est installé. En effet, nous avons remarqué par exemple que dans les motifs qui constituent la concurrence, les concurrents les plus renseignés sont ceux dont le nom commence par la lettre "A" ; simple coïncidence ou un effet de la liste déroulante des motifs de résiliation

ordonnés par ordre alphabétique ? De plus, la répartition des motifs qui est très hétérogène dans le temps nous fait croire que les gestionnaires pourraient faire des erreurs dans le renseignement du bon motif de résiliation.

Toutefois, nous faisons l'hypothèse que le gestionnaire distingue bien entre une résiliation qui viendrait selon le vouloir de l'assuré et celui qui vient de l'assureur. Sous cette hypothèse, 2 grands groupes de motifs sont constitués : **COMPAGNIE** et **INITIATIVE DE L'ASSURÉ**. Nous rappelons que ces deux grands motifs sont ainsi dénommés par abus de langage et que l'on ne doit pas les confondre avec les motifs du même nom mentionné dans les lignes précédentes. Quels sont les motifs constitutifs de chaque groupe ?

Les motifs retenus pour l'étude

Notre objectif étant de comprendre les déterminants des résiliations en vue de garder le plus longtemps possible les clients en portefeuille, le premier grand groupe des motifs - COMPAGNIE - semble ne pas être la population des résiliations sur laquelle il faut agir. Ainsi, notre périmètre de travail concernera uniquement des résiliations appartenant au second groupe - INITIATIVE DE L'ASSURÉ. Ce dernier est constitué de : **Destruction de véhicule**, **Changement de situation**, **Hamon**, **Concurrence**, **Vente de véhicule** et **initiative de l'assuré.** Pour tous ces motifs, il est facile de justifier que l'acte de résiliation intervient suite à la demande de l'assuré. Dans notre base de données, les contrats résiliés ne correspondant pas à ces motifs sont tout simplement supprimés. Des statistiques univariées seront réalisées par la suite sur les données résultantes afin de déceler le pouvoir discriminant possible des variables.

La nouvelle répartition des résiliations sur la période 2017-2020 se trouve sur la figure 2.7.

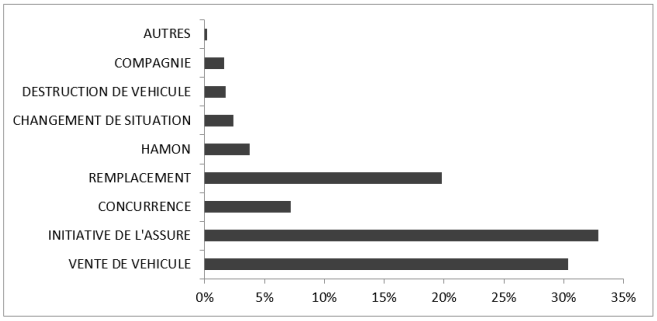


FIGURE 2.6 – Répartition des résiliations par motifs - tous les motifs confondus

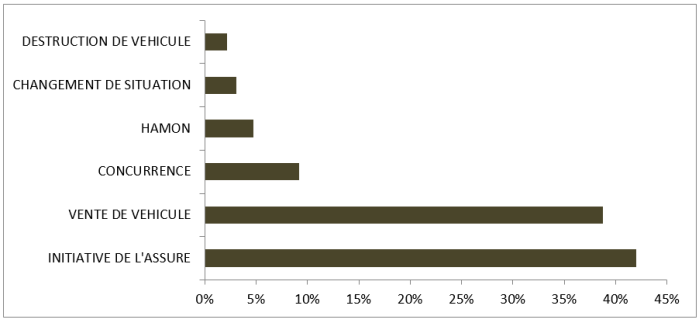


FIGURE 2.7 – Répartition des résiliations par motifs retenus pour l'étude

2.3.3 Statistique univariée

L'objectif de cette partie de statistique univariée est de savoir si nous pouvons décliner le pouvoir discriminant de certaines variables quant à leurs capacités à distinguer entre la population des contrats résiliés et celle des contrats non résiliés.

Un premier aperçu des résiliations selon les départements et le lieu de résidence sera vu par la suite. Nous souhaitons voir s'il y a par exemple plus de résiliations dans certains départements par rapport à d'autres ou encore si les assurés des banlieues ont une propension à résilier plus élevée que ceux vivant dans des villes isolées. Nous rapporterons également le taux de résiliation en fonction des antécédents de sinistres et par le coefficient bonus-malus du conducteur qui se sont avérés discriminants après cette première étude univariée.

Notations :

- *Resiliation* est la variable qui identifie si un contrat est résilié (Resiliation=1) ou 0 sinon ;
- n est le nombre total d'observations qui peut être selon le cas sur l'année considérée ou toute année confondue ;
- n_i l'effectif de la modalité i d'une variable qui peut être selon le cas, sur une année considérée ou sur toute la période d'étude ;
- r_i l'effectif des contrats résiliés ayant la modalité i .

Dans la suite donc de ce chapitre, on considérera les formules suivantes :

- $Poids_i(2017-2020) = \frac{n_i}{n}$: C'est la part du portefeuille ayant la modalité i d'une variable ;
- $taux_{resiliation} = \frac{r_i}{ni}$: le taux de résiliation associé à la modalité i.

Les résiliations et les indicateurs géographiques

L'exploration des données sur le plan géographique est très intéressante dans le sens où les observations pourraient nous dévoiler s'il y a des disparités entre les régions et les départements. Par ailleurs, rappelons-le, GAN ASSURANCES travaille exclusivement avec son réseau d'agents généraux. De ce fait, les agents étant basés dans des zones géographiques bien précises, ces observations pourront nous permettre de cibler dans un premier temps les zones et les agents en vue d'une amélioration de la rétention des clients en portefeuille.

D'ailleurs, une tentative d'inclusion des variables sur les agents a été couronnée par un échec. En effet, nous avons essayé d'utiliser des données internes sur les « les agents 360 » qui est un indicateur qui prend la valeur 1 si l'agent a de très mauvais résultats en termes de performances de son portefeuille, et 0 sinon. Toutefois, l'indicateur n'a été mis

en place qu'à partir de janvier 2020. Ce qui fait que nous n'avons pas pu l'utiliser car ne couvrant pas toute la période d'étude.

... La carte choroplèthe de France métropolitaine avec les taux de résiliation par département

Une carte choroplèthe ou encore une carte thématique peut être définie comme une carte où les zones géographiques sont colorées ou remplies d'un motif, le plus souvent, d'une couleur selon la valeur prise par une mesure de statistique. Dans notre cas, cette mesure statistique est le taux de résiliation et le niveau géographique considéré est le département. Ainsi, à l'aide de la librairie *raster* de R, nous avons réalisé une carte de France en colorant les départements par des couleurs dépendant du niveau du taux de résiliation de ces derniers, allant du blanc au bleu. Plus la couleur tend vers le bleu, plus le taux de résiliation dans ce département est élevé.

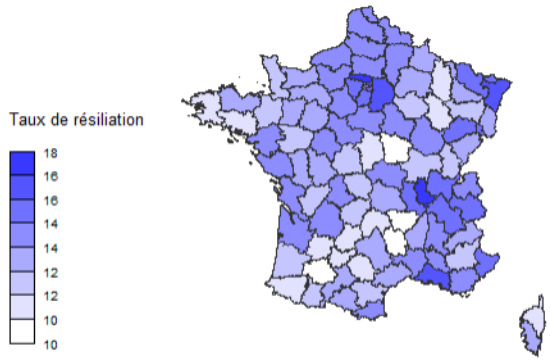


FIGURE 2.8 – Taux de résiliation par département en 2017

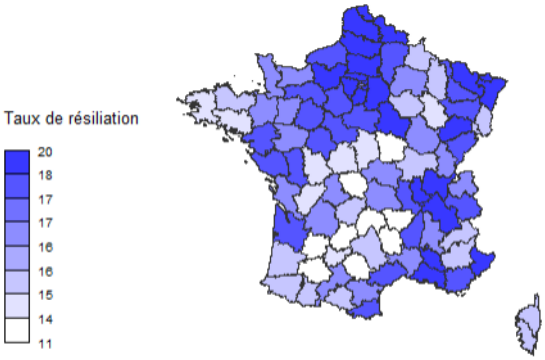


FIGURE 2.9 – Taux de résiliation par département en 2018

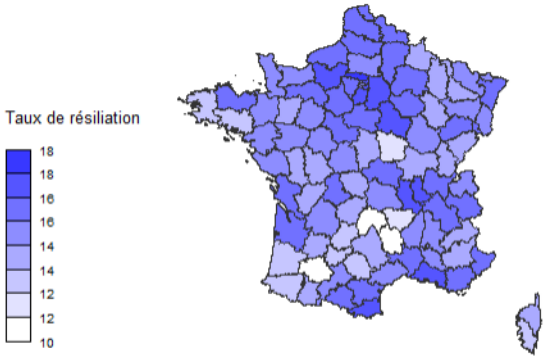


FIGURE 2.10 – Taux de résiliation par département en 2019

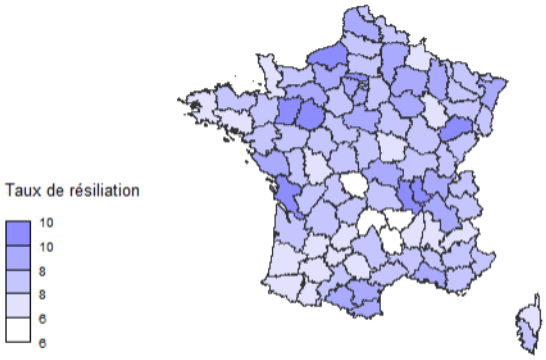


FIGURE 2.11 – Taux de résiliation par département en 2020

Comme nous pouvons le voir sur les figures 2.8, 2.9, 2.10 et 2.11 , les départements se situant au nord de la France (les régions d'Île de France, Picardie, Nord-Pas-de-Calais) et le long de la frontière Est semblent avoir un fort taux de résiliation. A l'inverse, de manière générale, les départements des régions Auvergnes, Limousin ou Midi-pyrénées semblent être des clients fidèles. Nous remarquons que la carte de 2020 est plus blanchâtre que les autres. La raison est celle que nous avons évoquée plus haut : la crise sanitaire liée au coronavirus.

Pour tenter d'expliquer ces disparités entre les départements, nous avons réalisé les mêmes cartes sur l'âge moyen de la population assurée par département. Les résultats en annexe A.1.1 montrent que plus l'âge moyen est faible, plus élevée est la propension à résilier du département. De ce fait, il apparaît intéressant d'utiliser l'âge de l'assuré pour les modèles de prédiction par la suite.

... Les résiliations et les unités urbaines

Le taux de résiliation est de niveau différent selon l'unité urbaine considérée. Comme le montre la figure 2.12, les zones urbaines ont un plus grand taux de résiliation que les zones rurales. Avec une segmentation plus fine, nous constatons des disparités au sein même des zones urbaines. L'observation de la figure 2.13 suggère que les assurés habitant dans les banlieues ont le taux de résiliation le plus élevé, ensuite viennent les clients dont les garages sont respectivement dans les centre-villes et les villes isolées. Les assurés des zones rurales ont une faible propension à résilier. Le faible taux de résiliation dans les zones rurales pourrait être expliqué par l'âge. En effet, dans les zones rurales, nous avons un âge moyen plus élevé dans notre population de contrats que dans les autres zones comme nous pouvons le percevoir en annexe A.1.2.

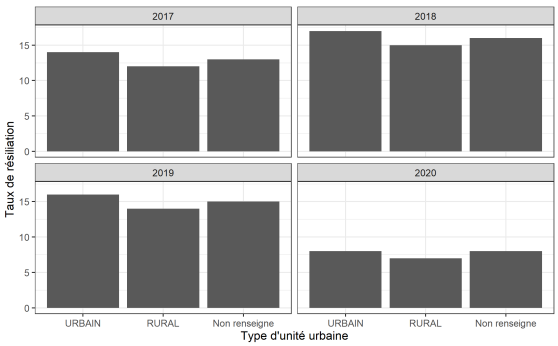


FIGURE 2.12 – Taux de résiliation par unité urbaine et par année

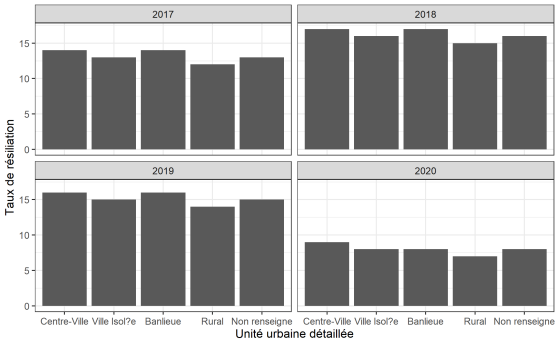


FIGURE 2.13 – Taux de résiliation par unité urbaine et par année - segmentation fine

Quelques variables intéressantes

Pour ne pas étaler toute la panoplie de variables explicatives potentielles de l'objet de ce mémoire, nous allons nous limiter dans cette présentation de statistique univariée aux variables portant sur les antécédents de sinistres, les variations de primes d'assurance et le CRM (coefficient bonus-malus).

... Les antécédents de sinistres 12, 24 et 36 mois

La lecture du tableau 2.3 montre que les antécédents de sinistres ont un fort pouvoir discriminant sur le taux de résiliation. De manière générale, on observe que les contrats qui ont au moins eu un sinistre dans le passé ont plus de chance d'être résilié par la suite. Toutefois, il est à noter que plus l'horizon considéré (12 mois, 24 mois, 36 mois) est éloigné, plus le pouvoir discriminant de l'antécédent de sinistre est faible. Ces observations sont bien visibles sur les figures 2.14 et 2.15. La pente de la courbe représentant le taux de résiliation en fonction du nombre de sinistres dans le passé est plus élevée en considérant les 12 derniers mois que dans le cas des 24 derniers mois. On peut aussi remarquer que le taux de résiliation par antécédent de sinistre, quelque soit l'horizon considéré, de 2017 varie peu. La raison principale réside en la profondeur de la base sinistre utilisée.

Antécédent - Modalités	2017	2018	2019	2020	TOTAL	Poids 2017-2020
Antécédent 12 mois						
00.Pas de sinistre	13	16	15	8	13	90
01.1 sinistre	16	19	18	11	16	9
02.2 sinistres	24	28	27	16	24	1
03.3 sinistres	28	40	44	24	37	0
04.Plus de 3 sinistres	36	44	59	22	44	0
Antécédent 24 mois						
00.Pas de sinistre	13	16	15	8	13	85
01.1 sinistre	16	19	17	10	16	12
02.2 sinistres	24	25	23	13	21	2
03.3 sinistres	28	35	31	19	29	0
04.Plus de 3 sinistres	36	42	41	26	37	0
Antécédent 36 mois						
00.Pas de sinistre	13	16	15	8	13	84
01.1 sinistre	16	19	17	10	15	13
02.2 sinistres	24	25	22	12	19	2
03.3 sinistres	28	35	29	16	24	0
04.Plus de 3 sinistres	36	42	35	20	29	0

TABLE 2.7 – Taux de résiliations par antécédent de sinistre (en annexe)

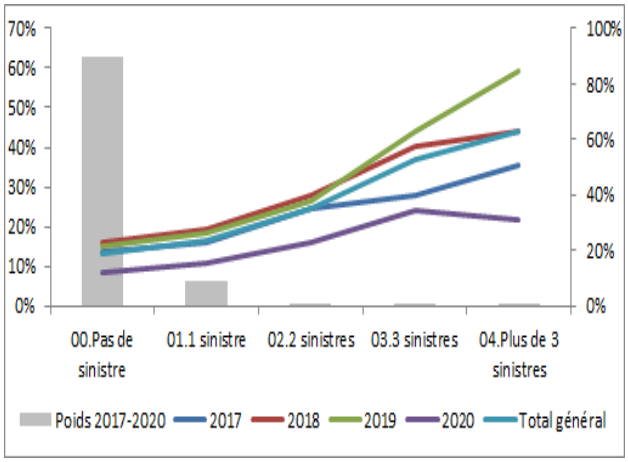


FIGURE 2.14 – Taux de résiliation par antécédent de sinistres 12 mois

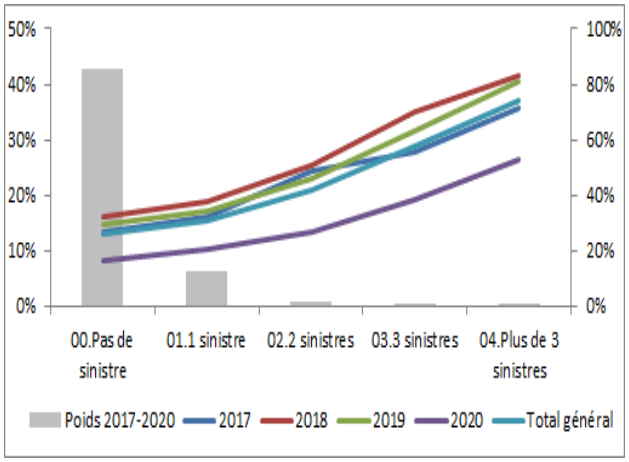


FIGURE 2.15 – Taux de résiliation par antécédent de sinistres 24 mois

... Les variations de prime commerciale en euros et en %

L'élasticité-prix (prime d'assurance) peut expliquer le taux de résiliation selon la plupart des mémoires d'actuariat de l'Institut des Actuaires - IA - qui ont traités le sujet. Les derniers en date sont *« assurance automobile : analyse de l'impact d'une variation du tarif sur le comportement des assurés lors de l'acte de souscription et de résiliation »* de M. Mehdi BOUEDDINE en 2013 ou encore *« la modélisation de la valeur contrat PNPV par la refonte du modèle de résiliation »* de M. Léonard FONTAINE en 2011.

Dans ce mémoire, nous avons décidé de simplifier les calculs en créant deux variables à partir de la prime. La prime est hors taxes. Étant consommateur de produit d'assurance nous-même, nous avons estimé que l'assuré serait très sensible à la variation de sa prime d'assurance entre la version $n-1$ et la version n en euros vu que le niveau des primes en assurance auto n'est pas très élevé ; la prime moyenne de l'assurance Auto en France en 2019 est de 243 euros (formule tiers - RC, vol, incendie et bris de glaces) selon les données de la FFA. Nous avons aussi calculé cette variation en évolution absolue (en %). Nous rappelons les formules des deux variables var_prime et $delta\_prime$ de la version n du contrat :

$$\boxed{var\_prime_n = prime_n - prime_{n-1}} \tag{2.6}$$

$$\boxed{delta\_prime_n = \frac{prime_n - prime_{n-1}}{prime_{n-1}} \times 100} \tag{2.7}$$

La répartition de l'évolution tarifaire est telle que les primes évoluant entre 1 et 3 euros sont majoritaires. En évolution en %, les contrats les plus prépondérants sont ceux ayant une augmentation de prime de 2 à 3% entre deux versions. Ce résultat n'est pas surprenant puisque la majoration tarifaire normale appliquée sur les contrats est de $2,75\%$.

Les résultats obtenus sur les figures 2.16 et 2.17 sont ceux à quoi l'on pouvait s'attendre intuitivement. En effet, la tendance globale est telle que plus la variation positive de prime est grande, plus le taux de résiliation est élevé.

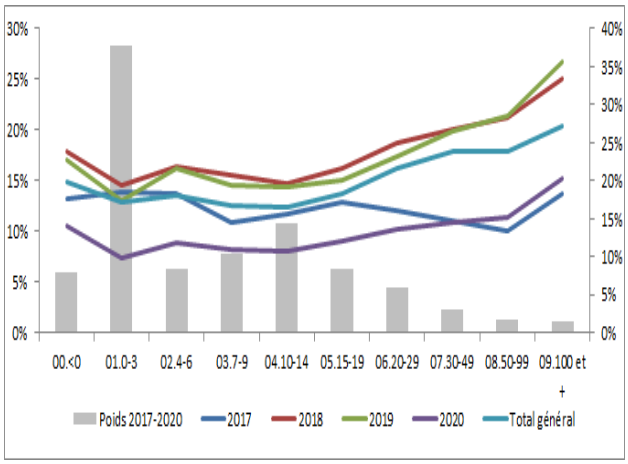


FIGURE 2.16 – Taux de résiliation en fonction de la variation de prime en euros

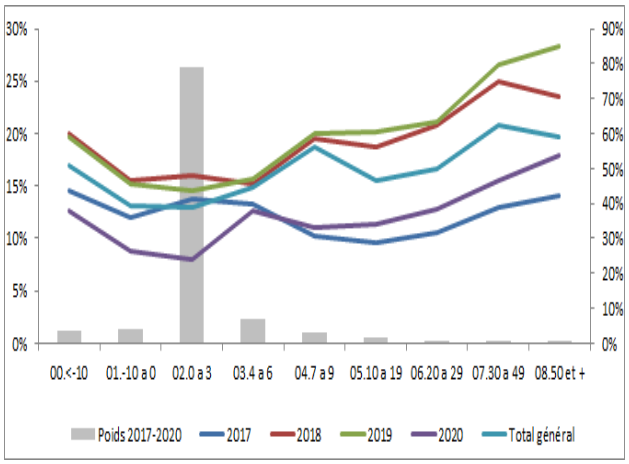


FIGURE 2.17 – Taux de résiliation en fonction de la variation de prime en %

Un résultat surprenant concerne les variations négatives. En effet, les contrats ayant une diminution de prime d'une année par rapport à une autre semblent avoir un taux de résiliation moyenne légèrement plus élevé qu'une faible augmentation de primes. Ces résultats trouvent leur explication dans la modulation tarifaire appliquée par les agents. En effet, chaque année, une enveloppe est allouée à chaque agent pour que celui-ci décide pour certains contrats de ne pas appliquer la majoration tarifaire. Ainsi, puisque leurs rémunérations dépendent de la taille et les performances de leurs portefeuilles, les agents ont tendance à faire bénéficier ces enveloppes aux clients qui sont susceptibles de résilier.

... Le coefficient réduction-majoration : CRM

Rappel sur le CRM

Légalement, il faut attendre la fin des années 80 pour voir l'apparition légale du système de coefficient de réduction-majoration plus connu sous la dénomination Bonus-Malus. Ce système est légalisé dans l'article A121-1 du code des assurances : *« les contrats d'assurance [...] concernant des véhicules terrestres à moteur doivent comporter la clause de réduction ou de majoration des primes ou cotisations »* [3].

L'objectif de ce système est de répercuter la sinistralité antérieure de chaque contrat sur la prime d'assurance sur une période de 12 mois consécutifs. Quelques véhicules terrestres à moteurs font exception selon le même article notamment :

3. Article A121-1 du Code des assurances, la version en vigueur le 05 août 2021 consultable ici

— les cyclos, motocyclette légère, quadricycle léger à moteur ;
— les véhicules de collection, véhicule d'intérêt général prioritaire, véhicule et matériel agricoles, matériel forestier, matériel de travaux publics.

Le principe est simple. Tous les assurés commencent avec un CRM=1.00 qui constitue la classe initiale. Ensuite, lorsque l'assuré n'a pas de sinistre responsable sur la période de 12 mois consécutifs, un coefficient bonus est appliqué (< 1). Dans le cas contraire, c'est un coefficient malus qui est appliqué (> 1) selon que la responsabilité soit totale ou partiel. Le CRM varie entre 0.5 et 3.5. La borne inférieur est atteinte lorsque l'assuré n'a pas de sinistres responsables pendant 13 années consécutives. Aussi devons-nous nous attendre à une forte corrélation entre l'âge de l'assuré et le CRM.

La résiliation en fonction du CRM

Le taux de résiliation semble évoluer positivement avec le CRM. En effet, plus le CRM est grand, plus le taux de résiliation est élevé. Néanmoins, il faut remarquer que notre portefeuille comporte assez de risque stable (CRM= 0.5) comme nous pouvons l'observer sur la figure 2.18. En termes d'exposition, il y a très peu de contrats avec un CRM supérieur à 1 (0.181%) du portefeuille. Ce nettoyage de portefeuille est surtout observé à partir de 2019 probablement grâce à la surveillance de portefeuille. Les assurés ayant un CRM supérieur à 1 résilient 2 fois plus que le niveau moyen de résiliation comme nous pouvons le voir sur la figure 2.19.

Les contrats ayant un CRM supérieur à 2 ont un taux de résiliation faible. En effet, ce sont des contrats qui devraient normalement passés en surveillance. Ce sont des mauvais risques et aucun autre assureur n'accepterait de les prendre normalement. De ce fait, intuitivement, on s'attend à ce que ces assurés résilient par eux mêmes le contrat. Par ailleurs, nous disposons de très peu de contrats avec ce CRM (7 observations au total). Pour la suite, nous avons décidé de les supprimer.

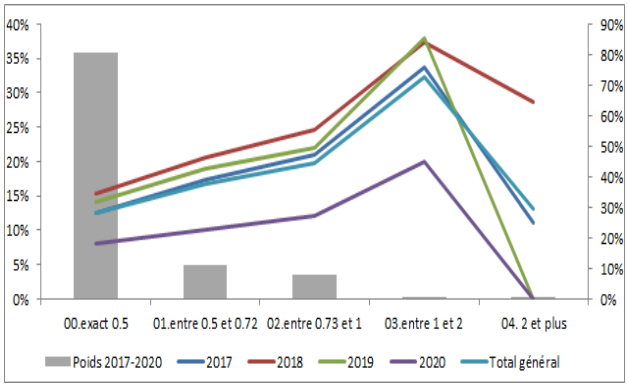


FIGURE 2.18 – Taux de résiliation en fonction du CRM et par année

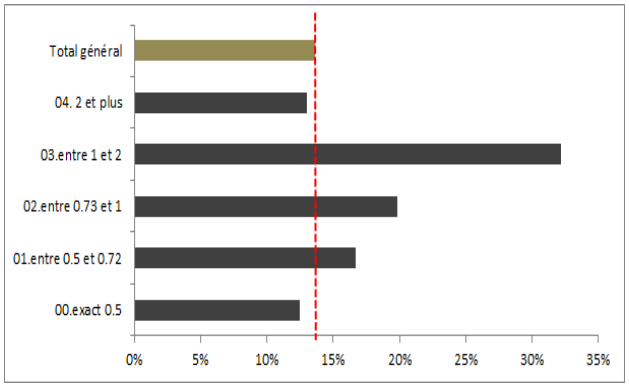


FIGURE 2.19 – Taux de résiliation et CRM - Zoom sur Total général

Après ce premier jet sur l'étude statistique de nos données, nous allons maintenant dans la suite de ce chapitre étudier la corrélation entre les différentes variables.

2.4 Étude la corrélation et base de données finale

2.4.1 Étude de la corrélation

L'analyse de la corrélation est une étape fondamentale dans le processus de modélisation. En effet, lorsque l'on introduit deux ou plusieurs variables explicatives qui sont fortement corrélées entre elles, cela crée des biais de mesure car elles apportent la même information. Nous allons donc étudier dans ce paragraphe les liaisons entre les potentielles variables explicatives. Ce qui nous permettra d'éliminer un certain nombre de variables et ainsi, obtenir la base de données finale de travail.

Dans la pratique courante, selon les types de variables (qualitatives ou quantitatives), plusieurs indicateurs sont utilisés parmi lesquels nous pouvons citer : ρ **de Pearson**, **le V de Cramer**, **le coefficient Phi** ou encore **le rapport de corrélation** η^2.

Dans notre étude, nous allons nous limiter à l'utilisation de 3 de ces mesures de liaisons entre 2 variables :

- le ρ de Pearson pour la corrélation entre 2 variables quantitatives ;
- le V de Cramer pour la mesure de l'intensité de liaison entre 2 variables qualitatives ;
- le carré du rapport de corrélation η^2 pour mesurer la corrélation entre une variable quantitative et une variable qualitative.

Les variables quantitatives discrètes seront considérées comme des variables qualitatives ordinales ayant autant de classes que de valeurs différentes.

Avant de passer à l'application sur les différentes variables dont nous disposons dans notre base de données, il convient de rappeler la théorie qui gouverne ces mesures.

ρ de Pearson

Considérons deux variables quantitatives X et Y qui prennent des valeurs respectives $(x_i)_{i=1...n}$ et $(y_i)_{i=1...n}$. Le ρ de Pearson s'obtient de la manière suivante :

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.8}$$

Avec $Cov(X,Y)$ la covariance entre X et Y et ;
σ_X et σ_X respectivement les écart-types de X et de Y.

Dans la pratique, c'est l'estimateur non biaisé $\hat{\rho}_{XY}$ qui est utilisé. Il s'écrit :

$$\hat{\rho}_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} \tag{2.9}$$

avec $\hat{\sigma}_X$ et $\hat{\sigma}_X$ les écart-types empiriques respectifs de X et Y ; et $\hat{\sigma}_{XY}$ la covariance empirique de X et Y. Leurs formules estimatives sont les suivantes :

$$\hat{\rho}_{XY} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}$$

et

$$\hat{\rho}_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}; \hat{\rho}_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

V de Cramer

Très utilisé dans les études actuarielles nécessitant une étude de corrélation, le V de Cramer mesure l'intensité de la liaison entre deux variables. Il est basé sur le test d'indépendance khi-deux χ^2 de karl Pearson.

Le test d'indépendance χ^2 de Karl Pearson

Développé par Karl PEARSON (1857-1936) en 1900, le test d'indépendance χ^2 permet une appréciation de l'existence ou non d'une relation entre deux variables qualitatives ou entre une variable quantitative et l'autre qualitative, ou bien encore entre deux variables quantitatives dont les valeurs ont été regroupées. Toutefois, ce test ne permet pas, de manière générale, de connaître le sens de cette dépendance.

Le test permet de vérifier l'hypothèse *« H_0 : X et Y sont indépendants »* contre l'hypothèse alternative *« H_1 : X et Y ne sont pas indépendants »*. Considérons donc deux variables qualitatives X et Y tel que le nombre de modalités soit respectivement I et J ($I, J \geq 2$). La mise en oeuvre du test s'appuie par le tableau de contingence (confère tableau 2.8) à partir duquel on calcule la statistique χ^2_{calcul} selon la formule 2.10.

	Y_1	Y_2	...	Y_J	TOTAL
X_1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,J}$	$n_{1,.}$
X_2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,J}$	$n_{2,.}$
...	...	...	...	...	..
X_I	$n_{I,1}$	$n_{I,2}$	...	$n_{I,J}$	$n_{I,.}$
TOTAL	$n_{.,1}$	$n_{.,2}$	...	$n_{.,J}$	n

TABLE 2.8 – Tableau de contingence

On a :

$$\chi^2{}_{calcul} = \sum_{i,j} \frac{(n_{i,j} - \frac{n_{i,.}n_{.,j}}{n})^2}{\frac{n_{i,.}n_{.,j}}{n}} \tag{2.10}$$

Sous condition que la statistique du test de $\chi^2{}_{calcul}$ suit une distribution de χ^2 avec une probabilité α, l'issue du test revient à comparer le résultat $\chi^2{}_{calcul}$ au fractile d'ordre $1-\alpha$ de la loi Khi-2 à $(I-1)(J-1)$ degrés de liberté ($\chi^2{}_{1-\alpha}[(I-1)(J-1)]$).

— Si $\chi^2{}_{calcul} > \chi^2{}_{1-\alpha}[(I-1)(J-1)]$, alors on rejette l'hypothèse H_0 ;

— Sinon, on ne rejette pas l'hypothèse H_0 ;

— De manière général, un niveau $\alpha = 5\%$ est fixé.

Normalisation du test d'indépendance χ^2 : V de Cramer

Le test d'indépendance χ^2 ne permet pas de connaître le sens de la dépendance entre X et Y. En effet, théoriquement, la valeur de $\chi^2{}_{calcul}$ varie entre 0 et ∞. L'objectif est de ramener la valeur prise par χ^2 entre 0 et 1. Pour ce faire, sous les mêmes notations que précédemment, le V de Cramer est donné par la formule 2.11 :

$$V = \sqrt{\frac{\chi^2{}_{calcul}}{\chi^2{}_{max}}} \tag{2.11}$$

Avec

$$\chi^2{}_{max} = n \times min\,(I-1, J-1) \tag{2.12}$$

D'après les formules 2.10 et 2.12, il est trivial de démontrer que $\chi^2{}_{max}$ est supérieur ou égal à $\chi^2{}_{calcul}$. Par conséquent, le V de Cramer est compris entre 0 et 1. Lorsqu'il prend la valeur 0, alors il y a indépendance entre X et Y alors que la valeur 1 s'interprétera comme un cas de dépendance parfaite entre X et Y.

Le carré du rapport de corrélation η^2

Le carré du rapport de corrélation mesure l'intensité de la liaison entre une variable quantitative et une variable qualitative. Pour mieux illustrer l'aspect théorique de cet indicateur, nous allons nous placer dans le cadre de deux variables X (qualitative) et Y (quantitative). Nous supposons que la variable X a J modalités ($J \geq 2$) et nous noterons y les valeurs prises par Y.

Soit :

— $y_{i,j}$ la valeur prise par Y pour la $i^{\text{ème}}$ observation de la classe j (de la variable X) ;

— $\bar{y}_{.,j}$ la moyenne de Y des observations de la classe j (de la variable X) ;

— $\bar{y}$ la moyenne de Y de toutes les observations ;

— J_j le nombre d'observations dont la valeur de X appartient à la classe j.

Le carré du rapport de corrélation η^2 se calcule suivant la formule :

$$\eta^2{}_{X,Y} = \frac{SCE_{inter}}{SCE_{totale}} = \frac{\sum_i \sum_{i \in J_j} J(\bar{y}_{.,j} - \bar{y})^2}{\sum_i \sum_{i \in J_j} (\bar{y}_{i,j} - \bar{y})^2} \tag{2.13}$$

Ce rapport mesure le pourcentage de la variabilité de la variable Y dû aux différences entre les différentes modalités de X. Tout comme le V de Cramer, il varie entre 0 et 1 et les interprétations sont quasi-identiques. Lorsqu'il prend la valeur 0, la non liaison entre X et Y est la conclusion car autrement, cela signifierait que toutes les modalités (les classes) de X admettent la même moyenne pour la variable Y. Lorsque $\eta^2 = 1$, alors les variables X et Y sont parfaitement corrélées.

Règle de décision

Nous considérons qu'au-dessus d'une valeur de V de Cramer supérieur à 0.7, les variables sont corrélées. Le seuil fixé est le même pour η^2 et pour la valeur absolue de ρ de Pearson. Lorsque deux variables sont corrélées, nous ne gardons que l'une d'entre elles. Toutefois, lorsque nous estimons qu'une variable a une importance d'ordre "métier", nous la gardons même si elle est corrélée avec des variables retenues dans le modèle.

Application

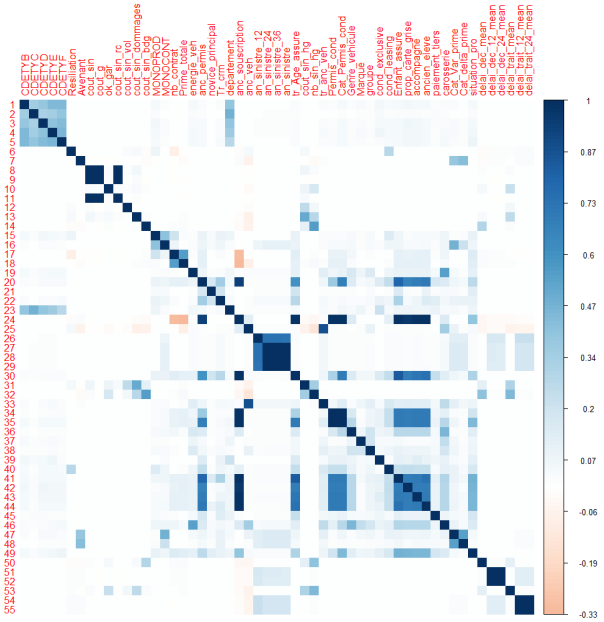
L'application des indicateurs de corrélation nous a emmené à mettre de côté les variables suivantes :

- Les zoniers : ils sont très peu corrélés avec les résiliations (coef $<$ 0,001) et les zoniers sont eux mêmes très corrélés avec le département que nous avons décidé de garder ;
- Le coût de sinistre grave : Le coût des sinistres graves est très corrélé avec le coût des sinistres sans prise en compte des graves. Pour garder l'effet de sinistre grave ou non, nous avons décidé de ne garder que le nombre de sinistres graves et le coût de sinistre sans prise en compte du grave ;
- Les antécédents de sinistres de plus de 36 mois et les coûts afférents : Les antécédents de sinistres sont très corrélés entre eux. Nous avons décidé de ne garder que les antécédents de sinistres des 12 derniers mois et des 24 derniers mois ;
- les délais de déclaration : D'un côté, ils sont très peu corrélés avec la résiliation et semblent ne pas avoir un intérêt d'ordre métier. De plus, ils sont très corrélés entre eux donc nous avons décidé de les mettre tous de côté ;
- conducteur avec leasing : Cette variable est très intéressante car elle indique si le véhicule, sous-jacent du risque est une location de longue durée : en moyenne 3 ans. Cette variable est très corrélée avec la résiliation. Mais nous l'avons supprimée et supprimée tous les contrats dont cette variable prend la valeur 1. En effet, tous ces

contrats sont résiliés au bout de 3 ans ;
— les variables novice de permis des conducteurs désigné et principal. Elles sont très corrélées avec l'ancienneté du permis de conduire que nous avons décidé de garder.

Notre matrice de corrélation étant très grande (104 variables), il est très difficile d'observer visuellement les variables corrélées entre elles. Le traitement nous a valu une exportation en fichier .Excel et une macro sur VBA pour colorier directement les variables corrélées entre elles à fonction du seuil fixé. C'est pourquoi, sur la figure ci-dessous, nous avons juste zoomé sur une liste de variables. La matrice de corrélation est en annexe A.1.5.

FIGURE 2.20 – Corrélation des variables - une sélection des variables



2.4.2 Base de données de travail

A l'étape de l'étude de corrélation, nous avons obtenu une base de données à n observations et m variables, couvrant la période de 2017 à 2020. L'objectif de ce mémoire étant de comprendre les résiliations et de mettre en place un modèle prédictif du risque de résiliation pour les futurs contrats, il est important de travailler sur des données ne reflétant que ce risque. Or comme nous l'avions constaté dans les statistiques descriptives, les données de 2020 sont fortement affectées par la crise sanitaire. De ce fait, afin d'avoir un modèle robuste et qui mesure le risque de résiliation en période normale, nous avons décidé de limiter les données sur la période 2017 à 2019. Ainsi, notre base de données finale contient n variables et m observations. C'est sur cette base de données finale que nous réaliserons les différents modèles dans les chapitres suivants.

Dans ce chapitre, nous avons fait état des données disponibles. Un processus de travail a été réalisé sur ces dernières pour aboutir à une base de données prête pour la modélisation. Ce chapitre a été l'occasion d'explorer les premiers résultats empiriques sur le taux de résiliation. Globalement, certaines variables semblent discriminer l'indicateur "taux de résiliation" parmi lesquelles les antécédents de sinistre, le CRM, les variables indiquant le lieu géographique. Ceci n'étant qu'une impression toute naïve, nous trouverons par la suite, à l'aide des méthodes de sélection de variables plus sophistiquées, les variables qui expliquent mieux le risque résiliant. Ces variables sortiront-elles du lot ?

Modélisation classique du risque de résiliation : la régression logistique

Dans ce chapitre, nous allons faire une première modélisation de la probabilité qu'un contrat soit résilié dans les 12 prochains mois - y compris la résiliation à l'échéance principale avec la régression logistique. En effet, c'est un modèle qui s'inscrit bien dans le cadre bayésien pour l'apprentissage supervisé. Notre objectif est de mettre en évidence la présence d'une liaison sous-jacente entre une variable catégorielle Y (Résiliation qui prend 2 valeurs (0 ou 1)) et un ensemble de variables explicatives $X = (X_1, X_2, ...X_p)$. La régression logistique appartient à un groupe de modèles plus vaste : la régression linéaire généralisée - *Generalized Linear Models GLM* - en anglais. Les hypothèses du GLM sont en annexe A.3 (Encadré 1).

3.1 Les motivations du choix de la régression logistique

La première raison du choix de la régression logistique est que la variable d'intérêt est binaire (0 ou 1) et donc nous sommes dans le cadre d'un problème de classification. La régression logistique peut nous permettre d'affecter une probabilité à chaque contrat selon ses caractéristiques. Ainsi, nous pourrons avoir plus de libertés sur le seuil de séparation des populations des résiliés et des non résiliés.

De plus, le nombre de variables explicatives étant élevé, la régression logistique nous permettra d'estimer les effets associés à chacune d'entre elles individuellement. Ce qui apportera plus de robustesse dans le modèle.

Enfin, une autre raison et pas des moindres, c'est un modèle très prisé par les actuaires pour sa simplicité et son interprétabilité.

3.2 Théorie sur la régression logistique

Nous nous plaçons dans le cadre de notre étude. Notons Y la variable à prédire - à expliquer - (*Resiliation*) qui est binaire. $Y = 1$ si le contrat est résilié et 0 sinon. Notons $X = (X_1, X_2, ..., X_p)$, une collection de p variables explicatives, lesquelles pouvant être quantitatives et qualitatives .

Dans une régression logistique, nous cherchons à modéliser l'espérance conditionnelle de Y sachant X. Pour des raisons de simplification, nous notons $\pi(x) = \mathbb{E}(Y|X = x)$. Y ne pouvant prendre que 2 valeurs, cette espérance conditionnelle n'est autre que **la probabilité que Y prenne la valeur 1 sachant X** ; avec Y qui suit une loi de Bernoulli.

En résumé, nous avons :

$$\mathbb{E}(Y|X=x) = \mathbb{P}(Y=1|X=x) = \pi(x)$$

3.2.1 Le modèle

La forme spécifique de la régression logistique que nous allons utiliser dans ce mémoire est :

$$\pi(x) = \frac{\exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)} \implies \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

avec $\beta = (\beta_0, \beta_1, ..., \beta_p)$ les paramètres du modèle.

La seconde égalité ci-dessus suggère que la transformation avec une fonction de lien *logit* g permet de définir plus facilement la probabilité de résiliation π. Cette transformation est très importante car elle permet d'avoir des propriétés intéressantes de la régression linéaire. En effet, $g(\pi)$ est linéaire en les paramètres du modèle.

$$g(z) = \log\left(\frac{z}{1-z}\right)$$

3.2.2 Estimation des paramètres

L'estimation de la probabilité π par la régression logistique sur notre base de données nécessite une estimation des valeurs des paramètres inconnus β. La méthode générale d'estimation des paramètres d'une régression linéaire généralisée est **le maximum de vraisemblance**. Nous allons donc l'utiliser pour estimer les coefficients de notre régression logistique.

Pour ce faire, considérons n observations (y_i, x_i) ; $i = 1, 2, ..., n$, où y_i est la valeur de la $i^{\text{ème}}$ observation de Y et $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ la $i^{\text{ème}}$ observation des variables indépendantes $X = (X_1, X_2, ..., X_p)$.

Le maximum de vraisemblance permet de déterminer les paramètres qui maximisent la probabilité d'obtenir les valeurs observées de Y. Sa mise en oeuvre est basée sur la vraisemblance $\mathcal{L}(\beta)$. La vraisemblance exprime la probabilité des données observées en fonction des paramètres inconnus β. Ainsi, pour une observation (y_i, x_i), nous pouvons facilement justifier que la vraisemblance s'écrit :

$$\mathcal{L}(\beta, y_i) = \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

Les observations étant indépendantes, la vraisemblance basée sur les n observations

est le produit des vraisemblances des observations individuelles. Elle s'écrit :

$$\mathcal{L}(\beta, y) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i} \tag{3.1}$$

Les estimations de β de notre modèle sont celles qui maximisent l'équation 3.1. Toutefois, pour effectuer les calculs, il est plus facile de travailler avec la log-vraisemblance $\mathcal{L}_n(\beta)$ (confère l'équation 3.2).

$$\begin{aligned}
\mathcal{L}_n(\beta, y) &= \log(\mathcal{L}(\beta, y)) = \log\left(\prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}\right) \\
&= \sum_{i=1}^{n} \left\{ y_i \log\left[\pi(x_i)\right] + (1 - y_i) \log\left[1 - \pi(x_i)\right] \right\} \\
&= \sum_{i=1}^{n} \left\{ y_i \log\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] + \log\left[1 - \pi(x_i)\right] \right\}
\end{aligned} \tag{3.2}$$

La fonction $x \mapsto \log(x)$ étant strictement croissante, maximiser $\mathcal{L}(\beta, y)$ revient à maximiser $\mathcal{L}_n(\beta, y)$. De ce fait, tout le problème revient à trouver l'estimateur $\hat{\beta}$ qui maximise la log-vraisemblance. Or d'après l'équation 3.2, nous voyons que cette dernière dépend de la probabilité $\pi(x_i)$ qui est inconnue, laquelle dépend elle-même des paramètres $\beta = (\beta_0, \beta_1, ..., \beta_p)$. Par conséquent, $\hat{\beta}$ est obtenu en déterminant les zéros de $\mathcal{L}_n(\beta, y)$. En dérivant $\mathcal{L}_n(\beta, y)$ par rapport à β_j, on obtient un système de deux équations dont la résolution se fait numériquement à l'aide des méthodes itératives comme la méthode de Newton-Raphson.

... Zoom sur l'algorithme de Newton Raphson

Le principe de cette méthode est simple. On démarre à partir d'une initialisation du coefficient β et ensuite on estime le même coefficient β à l'étape suivante suivant la formule :

$$\beta^{\mathrm{i}+1} = \beta^{\mathrm{i}} - \left(\frac{\partial^2 \mathcal{L}_n}{\partial\beta . \partial\beta^{\mathrm{t}}}\right) \times \left(\frac{\partial \mathcal{L}_n}{\partial\beta}\right) \tag{3.3}$$

Avec :

- β^{t} est la transposée du vecteur β ;
- $\frac{\partial^2 \mathcal{L}_n}{\partial\beta . \partial\beta^{\mathrm{t}}}$ la matrice des dérivées partielles secondes de la log-vraisemblance dite **matrice hessienne**. Son inverse correspond à la matrice de variance-covariance des coefficients. Ce dernier est très utile lors des tests de significativité des coefficients.

La formule itérative de l'équation 3.3 ne s'exécute pas à l'infini. Pour converger vers $\hat{\beta}$, de nombreuses règles d'arrêt permettent de stopper les calculs. Parmi elles, nous pouvons

citer :

— La fixation du nombre maximum d'itérations. Cela permet de limiter considérablement le temps de calcul et de plus, permet d'éviter les boucles qui tournent à l'infini lorsqu'il n y a pas de convergence ;

— La fixation d'un seuil ε pour lequel, la variation de la valeur de la log-vraisemblance d'une étape à une autre doit être supérieure. Dans le cas contraire, on arrête le processus.

Ainsi, puisqu'il convient de le mentionner, les différents logiciels de statistique n'utilisent pas forcément les mêmes règles d'arrêt. Ce qui peut avoir une incidence sur le résultat. Il n'est donc pas "bizarre" de retrouver des valeurs légèrement différentes des coefficients β d'un logiciel à un autre.

L'obtention des coefficients β_j ne nous assure pas une bonne adéquation du modèle. Des tests de significativité des coefficients β_j sont alors réalisés.

3.2.3 Test de significativité du modèle

Tester la significativité du modèle revient à tester si globalement, l'ensemble des variables X sont significativement liées à notre variable à expliquer Y. Autrement, il revient à tester si les p coefficients ne sont pas simultanément nuls contre l'hypothèse alternative qu'au moins l'un de ces coefficients soit non nul. Ce test est encore appelé le test **du rapport de vraisemblance**. Mathématiquement, nous pouvons écrire le test d'hypothèse de la façon suivante :

$$H_o : \beta_1 = \beta_2 = ... = \beta_p = 0 \quad Vs \quad H_1 : \exists i, \beta_i \neq 0$$

La statistique de ce test d'hypothèse est le rapport des maximums de vraisemblance S qui s'écrit :

$$S = -2 \log \left(\frac{\mathcal{L}(\hat{\beta}_{H_0}, y)}{\mathcal{L}(\hat{\beta}, y)} \right)$$

Avec :

— $\hat{\beta}_{H_0}$: l'estimateur des coefficients β sous l'hypothèse H_0 ;

— $\hat{\beta}$: l'estimateur des coefficients β sous l'hypothèse alternative H_1 ;

— $\mathcal{L}(\hat{\beta}_{H_0}, y)$ la vraisemblance du modèle ne contenant aucune variable (sous l'hypothèse H_0).

La statistique S suit sous H_0 une loi de χ^2 à p degrés de liberté. La règle de décision est telle que si $\mathcal{P}\left(\chi^2(p) \geq S\right) \leq \alpha$, alors l'on rejette l'hypothèse H_0 et le modèle est significatif. Dans la pratique, $\alpha = 5\%$.

Une fois le modèle validé comme étant significatif, il est important de regarder si les

prédicteurs sont significatifs au seuil de 5% par défaut.

3.2.4 Test de significativité des prédicteurs

Le test de significativité de chaque variable est une étape importante dans une modélisation. En effet, si l'on n'est pas convaincu que la variable X_i explique de manière significative la variable indépendante Y ; il serait imprudent de commenter le modèle et décliner les effets de ces variables sur le comportement de la variable à prédire. De manière générale, 3 tests peuvent être réalisés : **le test de Wald**, **le test de score** et **le rapport de vraisemblance**.

... Le rapport de vraisemblance

Tout comme le test de significativité du modèle dans le paragraphe précédent, le test basé sur le rapport des vraisemblances consiste ici à comparer les vraisemblances de deux modèles emboîtés ; l'un comportant la variable dont on veut juger son rôle, l'autre n'en comportant pas. L'on peut formuler le test d'hypothèse pour la variable X_i de la façon suivante :

$$H_o : \beta_i = 0 \quad Vs \quad H_1 : \beta_i \neq 0 \tag{3.4}$$

Comme mentionnée plus haut, la statistique du test est le rapport de vraisemblance S_i qui s'écrit :

$$S_i = -2 \log \left(\frac{\mathcal{L}(\hat{\beta}_{.|X_i}, y)}{\mathcal{L}(\hat{\beta}, y)} \right)$$

Avec :

- $\mathcal{L}(\hat{\beta}_{.|X_i}, y)$ la vraisemblance du modèle ne contenant pas la variable X_i (sous l'hypothèse H_0) ;
- $\mathcal{L}(\hat{\beta}, y)$ la vraisemblance du modèle contenant la variable X_i (sous l'hypothèse alternative H_1).

Sous H_0, S_i suit une loi de χ^2 à 1 degré de liberté. Ainsi, pour conclure que la variable X_i contribue significativement à expliquer la variable Y, il suffit que $\mathcal{P}\left(\chi^2(1) \geq S_i\right) \leq \alpha$.

..... Le test de score

Basé sur le même test d'hypothèse mentionné à l'équation 3.4, le test de score cherche à vérifier si le gradient au point $\hat{\beta}_{H0}$ − coefficients estimés des variables $X_{j,j\neq i}$ - de la *vraisemblance* est proche de 0. La statistique nommée **Score** dans le cadre de ce test d'hypothèse est la suivante :

$$Score_i = \left[\frac{\partial \mathcal{L}}{\partial \beta}\right]^t_{\hat{\beta}_{H_0}} \left[\mathbb{I}(\hat{\beta}_{H_0})\right]^{-1} \left[\frac{\partial \mathcal{L}}{\partial \beta}\right]_{\hat{\beta}_{H_0}}$$

Avec la notation A^t désignant la transposée de la matrice A.

Sous l'hypothèse H_0, $Score_i$ suit une loi de χ^2 à 1 degré de liberté. De ce fait, la variable X_i est significative au niveau $(1-\alpha)\%$ si $\mathcal{P}\left(\chi^2(1) \geq Score_i\right) \leq \alpha$.

..... Le test de Wald

La statistique de Wald mesure l'apport en terme de variabilité d'information d'un régresseur X_i. Elle est basée sur la normalité asymptotique de l'estimateur $\hat{\beta}$. La statistique permettant de tester si la variable X_i explique significativement Y s'écrit :

$$W_i = \frac{\hat{\beta}_i^2}{\hat{\sigma}^2(\hat{\beta}_i)}$$

Avec $\hat{\sigma}^2(\hat{\beta}_i)$ la variance du coefficient $\hat{\beta}_i$. Cette variance est lue directement sur la diagonale de la matrice de variance-covariance ; laquelle est l'inverse de la matrice hessienne $\frac{\partial^2 \mathcal{L}_n}{\partial \beta . \partial \beta^{\mathrm{t}}}$ que nous avons mentionné dans l'algorithme de Newton Raphson.

Après avoir testé les coefficients β_i, il est important de regarder si le modèle est adéquat. En effet, pour utiliser ce modèle pour affecter une probabilité de résiliation à nos futurs contrats d'assurance automobile sur 12 mois (y compris la résiliation à l'échéance), nous devons vérifier que le modèle est adéquat à nos données.

3.2.5 Adéquation du modèle

Il existe plusieurs manières de vérifier l'efficacité du modèle et par là, valider son adéquation aux données. Dans le cadre d'une régression logistique ayant pour objectif de faire une classification binaire, deux grandes catégories d'indicateurs peuvent être utilisées : **les indices de mesures de l'efficacité globale du modèle que sont les critères AIC et BIC, et l'analyse des résidus.** Dans le cadre de la classification, d'autres indicateurs permettent de valider le modèle. Il s'agit notamment de comparer les prédictions du modèle aux observations. Nous développerons les outils d'appréciation de la prédiction faite par le modèle dans la section 3.3.3.

.... Les critères AIC et BIC

Les critères AIC - *Akaike Information Criterion* et BIC - *Bayesian Information Criterion* permettent de sélectionner les modèles dans le sens où ils réalisent un compromis entre nombre de variables explicatives dans le modèle et son pouvoir explicatif. Contrairement au coefficient de détermination R^2, ces critères pénalisent la vraisemblance en tenant compte du nombre de paramètres p à estimer et la taille n de l'échantillon de données. En gardant les mêmes notations que précédemment, ces deux critères sont calculés de la manière suivante :

$$AIC = -2\mathcal{L}_n + 2p \tag{3.5}$$

et :

$$BIC = -2\mathcal{L}_n + p\log(n) \tag{3.6}$$

S'il faut faire un choix entre plusieurs modèles sur ces critères, le meilleur modèle est celui ayant la valeur la plus petite. Par exemple, considérons 2 modèles qui ont la même valeurs de $\mathcal{L}_n$ mais avec un nombre de variables p différent. Selon l'objectif poursuivi, un des critères est préféré pour faire le choix du modèle :

— Dans notre exemple, le modèle avec le moins de variables sera sélectionné avec le critère AIC. En effet, d'après la formule 3.5, AIC pénalise le nombre de variables utilisées ;

— Le choix basé sur le BIC dépendra de la taille de l'échantillon. En effet, ce critère pénalise beaucoup plus les modèles complexes (avec beaucoup de variables) car la pénalité est en $log(n)$ sur le nombre de variables. De manière générale, il est préféré à l'AIC dès lors que que l'on est avec des données de grandes tailles. En supposant que $log(n) \geq 2 \implies n \geq 8$, ce qui est généralement le cas, le modèle avec le moins de variables sera privilégié.

.... Analyse des résidus : résidu de Pearson et la déviance

La déviance notée $\mathcal{D}$ est utilisée pour choisir le modèle le plus adéquat et ajusté aux données. Elle mesure l'écart entre la vraisemblance du modèle $\mathcal{L}$ et la vraisemblance du modèle saturé $\mathcal{L}_s$. Le modèle saturé est le modèle qui contient autant de paramètres que d'observations distinctes et par conséquent, décrit de manière exacte les données. Elle est calculée avec la formule :

$$\mathcal{D} = -2\log\left(\frac{\mathcal{L}_s}{\mathcal{L}}\right) = 2\log(\mathcal{L}) - 2\log(\mathcal{L}_s) \tag{3.7}$$

Soit ε_i le résidu brut, issu du modèle, associé à la $i^{\text{ème}}$ observation. Il s'écrit :

$$\varepsilon_i = y_i - \hat{\pi}_i$$

Nous rappelons que $\hat{\pi}_i$ est la probabilité que y_i prenne la valeur 1 et suit une loi de Bernoulli. De là, on peut démontrer que ε_i suive une loi de Bernoulli de paramètre $p_\varepsilon = \pi_i$. Il s'en suit d'après les moments d'ordres 1 et 2 de la loi de Bernoulli que $\mathbb{E}(\varepsilon_i) = 0$ et $\mathbb{V}(\varepsilon_i) = \pi_i(1-\pi_i)$. Les résidus r_i de Pearson s'écrivent alors :

$$r_i = \frac{\varepsilon_i}{\sqrt{\mathbb{V}(\varepsilon_i)}} = \frac{y_i - \hat{\pi}_i}{\sqrt{\pi_i(1-\pi_i)}}$$

L'étude des résidus de Pearson reviens à vérifier que ces derniers sont distribués normalement et centrés sur 0. La statistique construite avec les résidus r_i est la somme des carrés des résidus, laquelle suit une loi de χ^2 à $n-p-1$ degrés de liberté lorsque le modèle ajusté est correct.

3.3 Application aux données

3.3.1 Division des données en échantillon d'apprentissage et test

Pour construire un modèle prédictif, nous avons besoin des données. La pratique usuelle consiste à diviser la base de données initiale en deux parties : une pour construire le modèle prédictif - la régression logistique - et l'autre pour l'évaluer.

Dans le cadre de notre modélisation, nous avons effectué deux types d'échantillonnage que nous avons dénommés *échantillonnage 1* et *échantillonnage 2*. Les régressions logistiques entraînées sur ces échantillons seront respectivement *modèle 1* et *modèle 2*.

Échantillonnage 1 : échantillonnage basé sur la chronologie

Le premier échantillonnage que nous réalisons est celui naïf. En effet, les données s'étendant de 2017 à 2019, nous avons séparé les données en 2 parties sur la base chronologique. Notre échantillon d'apprentissage est constitué des données de 2017 et 2018 et la validation du modèle est faite sur l'échantillon test constitué des données de 2019. L'objectif ici est de voir si le modèle est robuste sur les années. Se faisant, notre échantillon d'apprentissage contient 1 908 965 observations (68,40%) et notre échantillon test contient 882 212 observations (31,60%). Il faut rappeler le caractère non aléatoire de l'échantillonnage dans ce premier cas. De ce fait, nous sommes conscients des biais que cela peut introduire dans le modèle.

Échantillonnage 2 : échantillonnage aléatoire stratifié

Le second échantillonnage, le plus scientifique est un échantillonnage aléatoire stratifié. Les observations ont été tirées aléatoirement dans la base de données initiale pour s'assurer que toutes les caractéristiques observées au niveau de l'ensemble des contrats se retrouvent dans les échantillons. Nous avons veillé à ce que les taux de contrats résiliés et non résiliés soient les mêmes dans les échantillons comme dans la base initiale. La manoeuvre a été effectuée avec la fonction *createDataPartition* de la librairie *caret*. Nous avons dans ce processus 70% des données à l'échantillon d'apprentissage et 30% de données à l'échantillon test. [1]

1. Il nous faut suffisamment de données pour construire un modèle consistant et également suffisamment pour l'évaluer. Le choix de ces proportions est aussi d'avoir des proportions comparables à l'échantillonnage 1 pour des fins de comparaison.

Maintenant que nous avons les échantillons test et d'apprentissage, nous passons à l'étape de modélisation. Nous rappelons que les modèles sont entraînés sur les échantillons d'entrainement et nous faisons la prédiction sur les échantillons test.

3.3.2 Les résultats de la modélisation

La phase de modélisation s'est faite en plusieurs étapes. Une première tentative de sélection de variables avec l'algorithme *Stepwise* n'a pas donné des résultats concluants. En effet, la régression logistique est très "gourmande" en variables. Toutes les variables introduites sont sélectionnées et sont toutes significatives. De ce fait, nous avons sélectionné un nombre de variables nous-mêmes en nous basant sur le pouvoir discriminant des variables lors de la phase de la statistique descriptive et aussi par intérêt métier.

De ce fait, 10 variables ont été identifiés et nous avons procédé à la modélisation.

.... Le regroupement des modalités

Pour obtenir un modèle qui est significatif globalement et dont la p-value associée à la statistique de Wald soit inférieure à 5% pour toutes les modalités, nous avons procédé comme suit :

- Nous laissons tourner dans un premier temps le modèle et observons l'ensemble des variables - t modalités des variables - qui sont significatives à 5% ;
- En se basant sur les odds-ratios et les intervalles de confiance, nous regroupons les modalités de certaines variables ; surtout des modalités qui ne sont pas significatives ;
- Enfin, nous relançons le modèle et procédons au regroupement jusqu'à obtenir un modèle dont tous les coefficients sont significatifs.

Les résultats obtenus après cette modélisation ne sont pas concluants comme nous pouvons le voir sur le tableau 3.4. Néanmoins, ces modèles nous permettent de dé-corréler les effets de toutes les modalités des variables. Ainsi, toutes choses étant égales par ailleurs, nous pouvons alors comparer le profil de risque des différents assurés selon leurs caractéristiques. A cet effet, nous allons plutôt comparer les odds ratios plutôt que les coefficients β_i directement. Ces odds ratios sont calculés de la manière suivante :

$$Odd\_ratio_{\beta_i} = \exp(\beta_i)$$

.... Les tests d'adéquation et présentation des modèles

Les tableaux 3.1 et 3.2 présentent les odds-ratios , les intervalles de confiance associés et la p-valeur associée au test de Wald sur les différents coefficients.

D'après ces résultats, tous les coefficients sont significatifs à 5% car ayant tous des p-valeurs inférieures à 0,05. Les sorties .R des modèles sont en annexe A.4.

Intercept mesure la probabilité de résiliation de notre profil de référence. Les interprétations qui suivront seront faites comparativement à cette modalité de base.

Modalités	Odd_Ratios	2.5 %	97.5 %	P-valeur
(Intercept)	0.20	0.20	0.20	0.00
anc_contrat01.<1 an	Ref*	-	-	-
anc_contrat02.entre 1 et 3 ans	1.17	1.15	1.18	0.00
anc_contrat03.entre 3 et 5 ans	1.15	1.14	1.17	0.00
anc_contrat04.entre 5 et 9 ans	1.05	1.04	1.06	0.00
anc_contrat05.entre 9 et 20 ans	0.87	0.86	0.88	0.00
anc_contrat06.20 ans et plus	0.79	0.77	0.81	0.00
tr_Age_assure01.<= 34 ans	1.05	1.03	1.06	0.00
tr_Age_assure02.35-49 an	Ref*	-	-	-
tr_Age_assure03.50-59 ans	0.94	0.93	0.95	0.00
tr_Age_assure04.60-69 ans	0.85	0.84	0.86	0.00
tr_Age_assure05.70-79 ans	0.68	0.67	0.70	0.00
tr_Age_assure06.80 ans et plus	0.79	0.77	0.81	0.00
tr_Age_assure07.Personne morale	1.30	1.28	1.33	0.00
tr_anc_veh00.<=1 ans	0.48	0.47	0.49	0.00
tr_anc_veh01.1 ans	0.72	0.70	0.74	0.00
tr_anc_veh02.2 ans	0.90	0.88	0.92	0.00
tr_anc_veh03.3 ans	0.87	0.86	0.89	0.00
tr_anc_veh04.4-6 ans	0.83	0.82	0.84	0.00
tr_anc_veh07.7-8 ans	0.85	0.84	0.86	0.00
tr_anc_veh08.9-14 ans	Ref*	-	-	-
tr_anc_veh09.15-19 ans	1.23	1.22	1.25	0.00
tr_anc_veh10.>=20 ans	1.25	1.23	1.26	0.00
paiement_tiersMensuel	Ref*	-	-	-
paiement_tiersTrimestriel	1.11	1.08	1.15	0.00
paiement_tiersSemestriel	0.94	0.93	0.95	0.00
paiement_tiersAnnuel	0.84	0.83	0.85	0.00
Tr_crm01.exact 0.5	Ref*	-	-	-
Tr_crm01.entre 0.5 et 0.72	1.23	1.21	1.24	0.00
Tr_crm02.entre 0.73 et 1	1.44	1.42	1.47	0.00
Tr_crm03.entre 1 et 2	2.22	2.07	2.38	0.00
Cat_delta_prime00.<-10	0.98	0.96	1.00	0.02
Cat_delta_prime01.-10 a 0	0.80	0.79	0.82	0.00
Cat_delta_prime02.0 a 3	Ref*	-	-	-
Cat_delta_prime03.4 a 6	1.31	1.29	1.33	0.00
Cat_delta_prime04.7 a 19	1.28	1.26	1.31	0.00
Cat_delta_prime05.20 et +	1.09	1.06	1.12	0.00
energie_vehdiesel	Ref*	-	-	-
energie_vehelectrique	1.36	1.21	1.53	0.00
energie_vehessence	0.95	0.94	0.96	0.00
energie_vehGPL - Non renseigne	0.95	0.92	0.97	0.00
Genre_vehicule20	Ref*	-	-	-
Genre_vehicule21 - 22	1.10	1.08	1.12	0.00
Genre_vehicule23	1.36	1.28	1.44	0.00
Genre_vehicule30 à 32	0.84	0.82	0.85	0.00
Genre_vehicule33 et +	0.63	0.61	0.66	0.00
tx_derog	0.02	0.02	0.02	0.00
an_sinistre_2400.Pas de sinistre	Ref*	-	-	-
an_sinistre_2401.1 sinistre	1.33	1.32	1.35	0.00
an_sinistre_2402.2 sinistres et +	2.15	2.09	2.22	0.00

TABLE 3.1 – Régression logistique - Modèle 1
Ref*= Modalité de référence

Modalités	Odd_Ratios	2.5 %	97.5 %	P-valeur
(Intercept)	0.19	0.19	0.20	0.00
anc_contrat01.<1 an	Ref*	-	-	-
anc_contrat02.entre 1 et 3 ans	1.21	1.20	1.23	0.00
anc_contrat03.entre 3 et 5 ans	1.21	1.19	1.23	0.00
anc_contrat04.entre 5 et 9 ans	1.11	1.10	1.13	0.00
anc_contrat05.entre 9 et 20 ans	0.92	0.91	0.93	0.00
anc_contrat06.20 ans et plus	0.84	0.82	0.86	0.00
tr_Age_assure01.<= 34 ans	1.06	1.05	1.08	0.00
tr_Age_assure02.35-49 ans	Ref*	-	-	-
tr_Age_assure03.50-59 ans	0.94	0.93	0.95	0.00
tr_Age_assure04.60-69 ans	0.85	0.83	0.86	0.00
tr_Age_assure05.70-79 ans	0.68	0.67	0.70	0.00
tr_Age_assure06.80 ans et plus	0.80	0.78	0.82	0.00
tr_Age_assure07.Personne morale	1.32	1.30	1.35	0.00
tr_anc_veh00.<=1 ans	0.47	0.46	0.49	0.00
tr_anc_veh01.1 ans	0.69	0.67	0.71	0.00
tr_anc_veh02.2 ans	0.89	0.87	0.91	0.00
tr_anc_veh03.3 ans	0.87	0.85	0.89	0.00
tr_anc_veh04.4-6 ans	0.82	0.81	0.83	0.00
tr_anc_veh07.7-8 ans	0.84	0.83	0.86	0.00
tr_anc_veh08.9-14 ans	Ref*	-	-	-
tr_anc_veh09.15-19 ans	1.26	1.24	1.27	0.00
tr_anc_veh10.>=20 ans	1.29	1.27	1.30	0.00
paiement_tiersMensuel	Ref*	-	-	-
paiement_tiersTrimestriel	1.15	1.11	1.18	0.00
paiement_tiersSemestriel	0.91	0.90	0.93	0.00
paiement_tiersAnnuel	0.83	0.82	0.84	0.00
Tr_crm01.exact 0.5	Ref*	-	-	-
Tr_crm01.entre 0.5 et 0.72	1.19	1.18	1.21	0.00
Tr_crm02.entre 0.73 et 1	1.41	1.39	1.43	0.00
Tr_crm03.entre 1 et 2	2.14	2.00	2.30	0.00
Cat_delta_prime00.<-10	1.06	1.04	1.08	0.00
Cat_delta_prime01.-10 a 0	0.87	0.85	0.88	0.00
Cat_delta_prime02.0 a 3	Ref*	-	-	-
Cat_delta_prime03.4 a 6	1.15	1.13	1.17	0.00
Cat_delta_prime04.7 a 19	1.21	1.19	1.23	0.00
Cat_delta_prime05.20 et +	1.14	1.11	1.17	0.00
energie_vehdiesel	Ref*	-	-	-
energie_vehelectrique	1.38	1.24	1.54	0.00
energie_vehessence	0.94	0.94	0.95	0.00
energie_vehGPL - Non renseigne	0.94	0.92	0.96	0.00
Genre_vehicule20	Ref*	-	-	-
Genre_vehicule21 - 22	1.10	1.08	1.12	0.00
Genre_vehicule23	1.38	1.30	1.47	0.00
Genre_vehicule30 à 32	0.85	0.84	0.87	0.00
Genre_vehicule33 et +	0.77	0.75	0.79	0.00
tx_derog	0.15	0.14	0.16	0.00
an_sinistre_2400.Pas de sinistre	Ref*	-	-	-
an_sinistre_2401.1 sinistre	1.27	1.25	1.28	0.00
an_sinistre_2402.2 sinistres et +	1.94	1.90	1.99	0.00

TABLE 3.2 – Régression logistique - Modèle 2
Ref*= Modalité de référence

.... Le profil de référence

Notre profil de référence est un nouveau assuré chez GAN (ancienneté de contrat de moins d'un an - affaire nouvelle), âgé de 35 à 50 ans non inclus, détenant un véhicule vieux de 9 à 15 ans, payant mensuellement sa prime d'assurance, ayant un CRM de 0,5

(donc ayant eu au moins 13 ans d'ancienneté de souscription), ayant un véhicule de genre faible (20) et utilisant le carburant diesel.
Les écarts de risque de résiliation se feront par rapport à cet individu référent de notre portefeuille GAN Auto.

.... Les odds-ratios et les profils de risques de résiliation

Dans cette partie consacrée à la discussion autour des profil de risque de résiliation, nous avons remarqué que les deux modèles procurent des conclusions similaires. De ce fait, nous allons nous baser sur le modèle 1 pour les commentaires.

Ancienneté du contrat - anc_ contrat **:** Toutes choses étant égales par ailleurs, les contrats ayant une ancienneté variant entre 1 et 5 ans sont les plus risqués. En effet, d'après les résultats des deux modèles, la conclusion est la même. Par exemple, selon le modèle 1 - confère le tableau 3.1, les contrats ayant entre 1 et 3 ans d'ancienneté ont 17% plus de chance de résilier que les affaires nouvelles. Néanmoins, plus l'ancienneté augmente (à partir de 10 ans), les clients deviennent plus fidèles et le risque de résiliation est plus faible.

Âge de l'assuré - tr_ Age_ assure **:** Les jeunes sont les plus risqués. En effet, les jeunes de moins de 35 ans ont 5% de plus de chances de résilier leurs contrats par rapport à ceux ayant entre 35 et 50 ans. Le profil le moins risqué est les assurés âgés de 80 ans et plus. En effet, ils ont 21% de moins de chances de résilier leurs contrats que leurs cadets de 35 à 50 ans.

Il faut noter que les personnes morales sont très risquées quant à leur propension de résilier un contrat d'assurance auto. En effet, ils ont 30% de chances supplémentaires de résilier leurs contrats par rapport au contrat de référence.

Âge du véhicule - tr_ anc_ veh **:** Sans surprise, la tendance des résultats est celle attendue : plus le véhicule est vieux, plus le contrat dont il est le sous-jacent est susceptible d'être résilié. En effet, les véhicules mis en circulation il y a plus de 20 ans sont 2,6 (1,25/0,48) fois plus risqués que les véhicules nouvellement en circulation.

La variation de prime en % - Cat_ delta_ prime **:** Toutes choses étant égales par ailleurs, les contrats qui résilient le plus sont ceux ayant une variation de prime entre 4% et 6%. Néanmoins, nous nous attendions à ce que plus cette variation augmente, plus la propension de résilier augmente. Ce qui n'est pas le cas car, nous voyons que bien que le risque soit plus élevé de manière générale pour les contrats ayant une revalorisation tarifaire de plus de 3%, il semble diminuer légèrement à partir de 20%. Ce qui suggère qu'il y a probablement un autre effet que nous n'arrivons pas à capter.

Il serait intéressant de regarder le croisement de cette évolution tarifaire et le niveau de prime par exemple. Une revalorisation de 2% sur une prime de 500 euros ; soit 10 euros

de plus (prime suivante = 510 euros) a-t-il le même effet qu'une revalorisation de 2% sur une prime initiale de 250 euros (255 euros de prime pour la version suivante) ?

Le CRM - Tr_ crm **:** Le coefficient de réduction et de majoration joue beaucoup dans la détermination de la prime, laquelle semble-t-il avoir un effet sur la propension à résilier d'après le paragraphe précédent. Comme attendu, plus le CRM est élevé, plus la propension à résilier est élevée. En effet, les contrats les plus risqués sont ceux ayant un CRM strictement supérieur à 1. Ces contrats sont 2,22 fois plus risqués que les contrats ayant un CRM de 0,5 et 1,54 fois plus risqués que ceux ayant un CRM compris entre 0,73 et 1.

L'antécédent de sinistre 24 mois - an_ sinistre_ 24 **:** L'antécédent de sinistre s'est révélé très discriminant pour la probabilité de résilier son contrat d'assurance auto. Les contrats qui ont 2 sinistres au moins dans les 24 mois précédents, ont 215% de chances supplémentaires d'être résiliés que ceux n'ayant pas de sinistre sur la même période. Ces contrats sont 1,62 fois plus risqués que ceux ayant uniquement un sinistre sur les deux dernières années.

3.3.3 Évaluation de la performance du modèle

Après application des modèles sur nos données, il convient de mesurer leur performance et efficacité. Dans le cadre de la régression logistique, elles se mesurent à travers plusieurs indicateurs. Dans cette section, nous allons notamment nous baser sur **la matrice de confusion**, **la courbe ROC** et **la courbe Rappel- Précision**. Nous rappelons que ces indicateurs sont valables pour tout classifieur pouvant fournir les probabilités π et les classes. De ce fait, nous utiliserons les mêmes indicateurs dans les chapitres suivants.

La matrice de confusion

La matrice de confusion est très utilisée dans l'évaluation des modèles de classification. Dans la suite de cette sous-section, nous allons décliner comment elle est construite et les différents indicateurs que l'on peut calculer à partir de cette matrice.

.... Construction de la matrice de confusion

La construction de la matrice de confusion permet de confronter les observations et les prédictions faites par le modèle. Par conséquent, elle nous permettra de voir les bonnes et les mauvaises prédictions, leurs structures entre les différentes classes et de calculer quelques indicateurs permettant d'apprécier la qualité de nos prédictions.

Comme son nom l'indique, la matrice de confusion est une matrice (un tableau) qui a en colonnes les classes prédites et en lignes les classes réellement observées dans la base de données. Elle peut être représentée dans notre cas de la manière suivante :

$Y \times \hat{Y}$	Classe 1	Classe 0	Total
Classe 1	a	b	$a+b$
Classe 0	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

TABLE 3.3 – La matrice de confusion - forme simplifiée

La classe objet de notre étude est la classe 1. De ce fait, elle est appelée *classe positive*. Ainsi, la lecture du tableau 3.3 se fait de la manière suivante :

- a représente le nombre des observations qui sont positives et qui sont prédites par le modèle comme telles. Ce sont les *vrais positifs*. Dans notre cas, ce sont les contrats résiliés qui sont bien prédits par le modèle ;
- c désigne l'effectif des observations de la classe 0 (classe négative) qui sont classées comme faisant partie de la classe 1 (positive) par notre modèle. Ce sont les *faux positifs*. Dans notre cas, c'est les contrats non résiliés qui sont classés comme étant des contrats résiliés par notre modèle ;
- b représente l'effectif des observations de la classe 1 prédites comme étant de la classe 0. Ce sont les *faux négatifs*. Typiquement dans notre cas, c'est l'effectif des contrats résiliés qui sont prédits comme non résiliés ;
- d représente le taux des *vrais négatifs* c'est-à-dire les contrats non résiliés bien détectés par le modèle.

Sur la base des éléments ci-dessus, nous allons dans la suite identifier les différents indicateurs qui en découlent et comment ces derniers sont utilisés pour évaluer les modèles de classification.

.... Les différents indicateurs

Les indicateurs que nous allons évoquer sont notamment *le taux d'erreur*, *le taux de succès*, *le rappel*, *la précision*, *la spécificité*.

Le taux d'erreur :

Le taux d'erreur mesure le taux de mauvais classement du modèle. Il est égal au nombre de mauvais classement rapporté au nombre total d'observations. D'après les notations du tableau ci-dessus, il se calcule par :

$$\boxed{\varepsilon = \frac{b+c}{n} = 1 - \frac{a+d}{n}}$$

Le taux de succès :

Le taux de succès - *accuracy en anglais* - mesure la proportion de bonnes prédictions faites par le modèle. C'est l'un des indicateurs que le statisticien regarde en premier. En effet, il permet de juger de la performance globale du modèle. Son calcul se fait en rapportant les bonnes prédictions sur l'effectif global.

$$\boxed{\theta = \frac{a+d}{n} = 1 - \varepsilon}$$

Le rappel :

Le rappel ou encore La sensitivité ou encore *le taux de vrais positifs - TVP* mesure la faculté du modèle à prédire les individus de la classe positive. De ce fait, c'est un indicateur très important lorsque l'objectif est de détecter la classe positive. Elle se calcule avec la formule :

$$\boxed{Rappel = TVP = \frac{a}{a+b}}$$

La précision :

La précision du modèle est un indicateur qui mesure la probabilité qu'un indicateur classé positif soit effectivement de la classe positive. Par exemple, dans notre cas, c'est très important car celle nous permet de connaître l'erreur que le modèle fait en prédisant qu'un contrat sera résilié dans les 12 mois suivants. Elle est déterminée comme suit :

$$\boxed{Précision = \frac{a}{a+c}}$$

La spécificité

La spécificité mesure la proportion des négatifs détectés par le modèle.

$$\boxed{Spécificité = \frac{d}{c+d}}$$

De manière générale, tous les indicateurs ne sont pas utilisés. Selon les thématiques, certains d'entre eux apparaissent plus pertinents que d'autres. Par exemple, notre objectif étant de prédire la classe positive donc maximiser le rappel et de minimiser le taux des faux positifs c'est-à-dire $1 -$ Précision.

.... Les résultats des modèles

Nous avons prédit la probabilité $\hat{\pi}$ de résiliation liée à chaque contrat dans nos échantillons tests avec les modèles 1 et 2. Ensuite, nous avons fixé le seuil de classification à 0.5. De ce fait, lorsque $\hat{\pi} >= 0.5$, alors le contrat est résilié et lorsque $\hat{\pi} < 0.5$, la conclusion inverse est actée. Pour évaluer les performances des deux modèles, nous avons calculé la matrice de confusion et les résultats obtenus sont consignés dans le tableau 3.4.

Modèles	*Accuracy*	*Taux d'erreur*	*Rappel*	*Précision*
Modèle 1	84,46%	15,54 %	0,0016	0,252
Modèle 2	84,62 %	15,38 %	0,0004	0,337

TABLE 3.4 – Quelques indicateurs des modèles logistiques

Les deux modèles présentent des bons *accuracy* : 85%. Toutefois, les deux modèles ont des précisions trop faibles (de l'ordre de 30%). En effet, un bon modèle de classification doit avoir des valeurs de précision, de rappel et de spécificité très élevées (proche de 1) et des valeurs faibles du taux des faux positifs (1 − Précision) et du taux d'erreur (proche de 0).

Le taux d'erreur (taux de mal classés) du modèle aussi permet de juger un modèle de classification. Ce taux est en effet dépendant de la distribution de la variable à prédire dans l'échantillon test. Dans notre cas, si par exemple le modèle classe systématiquement résilié un contrat, nous aurons un taux d'erreur de 15,24% (la proportion des positifs désignés négatifs). Ce chiffre nous indique que le modèle construit est "bon". Pourtant, il n'arrive pas à détecter les positifs, c'est-à-dire à trouver les potentiels contrats qui pourraient être résiliés dans les 12 prochains mois. Pour cette raison, nous utiliserons plutôt la courbe ROC pour évaluer nos modèles. C'est un outil graphique d'évaluation et de comparaison des modèles.

La courbe ROC

La courbe ROC - *Receiver Operating Characteristics curve en anglais* - est un outil graphique très utilisé en apprentissage supervisée pour comparer les classifieurs. Elle présente des caractéristiques très intéressantes qui font d'elle, l'un des indicateurs les plus prisés par les data scientistes. Parmi ses caractéristiques, on trouve notamment :

— Une indépendance de la courbe ROC vis-à-vis des matrices de coûts de mauvaise affectation. En effet, indépendamment de la combinaison des coûts de mauvaise affectation utilisée pour deux classifieurs, elle permet de voir si l'un dépasse l'autre ;
— Une opérationnalité dans des situations de données déséquilibrées. Comme nous pouvons le constater dans les résultats ci-dessus, le fait que les données soient déséquilibrées impactent fortement les performances de nos modèles en ces indicateurs mentionnés dans le tableau 3.4. La courbe ROC permet de s'affranchir de ces impacts de la probabilité *a priori* des différentes classes. De ce fait, il en découle le fait que les résultats et les interprétations qui découlent de la courbe ROC sont valables même si l'échantillon test n'est pas représentatif ;
— Enfin, elle permet de calculer l'indicateur AUC - *Area Under Curve en anglais* - ou mieux l'aire sous la courbe ROC. Cet indicateur est très facile à calculer et permet de synthétiser les informations de la courbe ROC. *In fine*, une comparaison de l' AUC permet d'avoir des informations sur le pouvoir prédictif des classifieurs.

La deuxième caractéristique mentionnée plus haut justifie notre choix d'utiliser la courbe ROC pour évaluer nos modèles. L'avant dernière caractéristique, quant à elle, nous réconforte dans la construction de la courbe ROC sur nos données d'échantillon test déséquilibré.

.... Construction de la courbe ROC

La construction de la courbe ROC est axée sur deux indicateurs que nous avons définis à partir de la matrice de confusion. Il s'agit du taux de vrais positifs - TVP ou encore le rappel et le taux de faux positifs TFP qui n'est autre chose que : $1 - \textit{Spécificité}$. La courbe ROC n'est finalement que le nuage de points de TVP en fonction des TFP.

Comme nous l'avons mentionné dans la partie des résultats des modèles, nous fixons dans un premier temps le seuil de probabilité à 0.5 auquel nous comparons la probabilité prédite $\hat{\pi}$ pour enfin déduire la classe prédite $\hat{y}$ et calculer les indicateurs issus de la matrice de confusion et plus particulièrement les TVP et TFP. L'obtention de ce nuage se fait de la même façon. En effet, on fait varier le seuil dans l'intervalle $[0, 1]$ et l'on récupère TVP et TFP qui varient eux-mêmes entre le continuum d'intervalle de valeurs possibles $[0, 1]$.

Théoriquement, deux situations extrêmes peuvent se présenter pour un modèle de classification : soit le modèle discrimine parfaitement les classes ou soit à l'inverse, il alloue aléatoirement les valeurs. Dans le premier cas, tous les positifs ont des valeurs probabilistes supérieures aux négatifs et dans ce cas, on aura la courbe ROC collée à l'axe des ordonnées et à l'extrémité nord du repère. A l'inverse dans le second cas, l'allocation aléatoire fera que les positifs et les négatifs ne sont pas distinguables et de ce fait, nous aurons la courbe ROC qui va se confondre avec la première bissectrice.

Enfin, le TVP et la *Spécificité* étant des indicateurs asymétriques, il est trivial de déduire que plus le TVP est élevé, plus la *Spécificité* est faible. Ce qui justifie la forme concave de la courbe ROC dans le cadran $[0, 1] \times [0, 1]$. En pratique dans les logiciels, on représente le couple TVP x *Spécificité*.

La lecture de la courbe ROC n'est pas systématique. L'on peut avoir deux classifieurs dont les courbes ROC se ressemblent fortement. Il est alors difficile de les comparer visuellement. C'est pourquoi une première conclusion basée sur la courbe est l'aire sous cette dernière. Elle est dénommée par le sigle AUC.

.... L'indicateur AUC

L'AUC est un indicateur qui synthétise les conclusions de la courbe ROC. C'est l'aire de la surface délimitée par l'axe des abscisses et la courbe ROC. Cette aire « **exprime la probabilité de placer un individu positif devant un négatif** »[13]. Sa valeur varie entre $0,5$ et 1. En effet, lorsque le classifieur classe aléatoirement les individus, la courbe ROC est confondue avec la première bissectrice et $AUC = 0,5$ - le modèle de référence. A contrario, dans le cadre d'une discrimination parfaite, $AUC = 1$. Différentes

interprétations s'imposent selon la valeur de l'AUC d'un modèle. Nous distinguons :

— Si $AUC = 0,5$: alors le classifieur ne permet pas de discriminer les classes. Il ne fait pas mieux qu'une allocation aléatoire des classes - le modèle de référence ;
— Si $0,5 < AUC < 0,7$: le classifieur discrimine passablement les classes ;
— Si $0,7 \leq AUC < 0,8$: le classifieur discrimine acceptablement les classes ;
— Si $0,8 \leq AUC < 0,9$: le modèle a un pouvoir discriminant très excellent ;
— Si $AUC \geq 0,9$: le classifieur est presque parfait ; il a un pouvoir discriminant exceptionnel.

Comment calculer l'AUC en pratique ? La plupart des logiciels de statistiques fournissent directement la valeur de l'AUC. Néanmoins, nous nous permettons de rappeler quelques méthodes classiques permettant de calculer l'AUC. Parmi elles, il est à noter la méthode d'intégrale numérique ou encore la méthode des trapèzes.

.... La courbe ROC, AUC et nos modèles logistiques

Les deux modèles présentent des AUC respectifs de 0,596 et 0,615. Ce qui est très faible. On voit bien que la courbe ROC n'est pas très proche du haut du cadran. Toutefois, le modèle basé sur les données aléatoires présente de meilleurs résultats et tous les deux font mieux que le modèle de référence qui est celui qui affecte aléatoirement les classes.

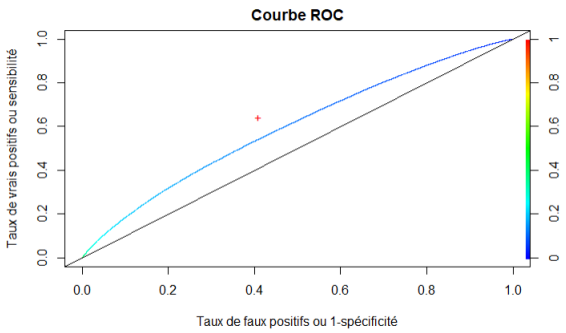


FIGURE 3.1 – La courbe ROC du modèle 1 - AUC= 0.596

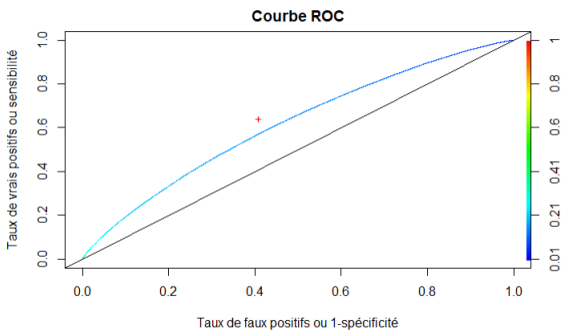


FIGURE 3.2 – La courbe ROC du modèle 2 - AUC =0.615

La courbe Rappel-Précision

Dans la section précédente, nous avons évalué le pouvoir discriminant de nos différents modèles avec la courbe ROC et son indicateur associée : AUC.

Toutefois, dans le cadre de ce mémoire, nous sommes face à deux exigences contradictoires. Dans un premier temps, nous voulons détecter les profils à fort risque de résiliation dans les 12 prochains mois. Ainsi, il faut que notre modèle nous permette de retrouver une proportion élevée des contrats résiliés potentiels et par conséquent, un TVP (*Rappel*) élevé. Dans un second temps, nous voudrions que notre population de contrats détectés

comme étant de potentielles futures résiliations soit constituée uniquement des contrats résiliés. En d'autres termes, la probabilité qu'un contrat choisi dans l'ensemble des contrats détectés comme de potentielles résiliations soit effectivement un contrat qui sera résilié soit élevée. Il s'agit donc d'avoir une valeur élevée de la *Précision*.

Conceptuellement, sa construction est analogue à celle de la courbe ROC. Il suffit donc de faire varier le seuil de probabilité prédite entre 0 et 1 ; construire la matrice de confusion et calculer les indicateurs *Rappel* - *Précision*. La suite logique est de tracer le nuage de points dans le plan $Rappel \times Prcision$.

Lorsque le seuil est faible (proche de 0), il est facile de voir que de plus en plus de négatifs vont se confondre aux positifs. De ce fait, le modèle perd en *Précision*. Toutefois, le seuil étant faible, le modèle réussira à mieux capter l'ensemble des positifs ; il y a donc un gain de *Rappel*. En résumé, plus le seuil est faible, plus le *Rappel* est élevé alors que le modèle perdra de plus en plus en *Précision*.

A contrario, lorsque le seuil est élevé, le modèle gagne en *Précision* car seuls les positifs sont de plus en plus détectés mais perd en *Rappel* car seule une faible proportion des positifs sera inclue dans les contrats détectés.

La courbe Rappel- Précision permet donc de visualiser comment le modèle arbitre entre ces deux indicateurs.

.... La courbe Rappel - Précision et les modèles 1 & 2

Comme l'on peut s'y attendre, les courbes rappel-précision montrent que nos régressions logistiques ont un faible pouvoir prédictif. Cela est notamment dû au fait que les données soient déséquilibrées.

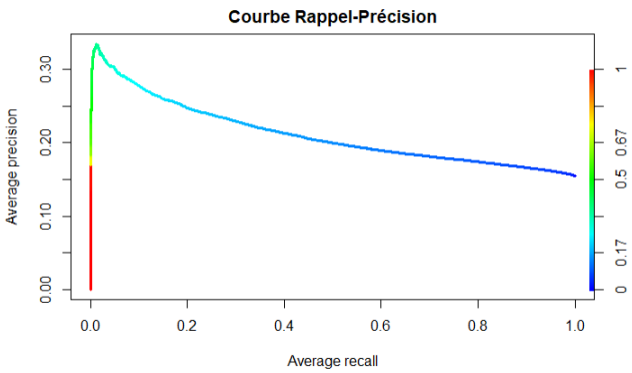


FIGURE 3.3 – La courbe Rappel - Précision du modèle 1

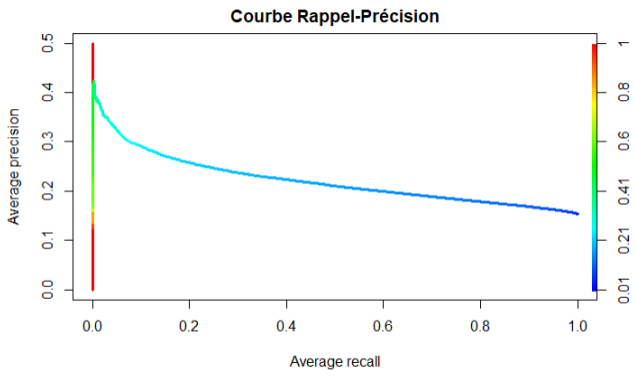


FIGURE 3.4 – La courbe Rappel - Précision du modèle 2

3.4 Une première conclusion sur le risque de résiliation

A la fin de cette modélisation, plusieurs conclusions sont à tirer :

— Les modèles ont un faible pouvoir prédictif. Toutefois, cela n'empêche pas qu'ils nous ont permis de mieux expliquer les déterminants du taux de résiliation et de connaître le profil des assurés à forte propension à résilier ;

— Les variables les plus discriminantes sont : le CRM, l'ancienneté du contrat et l'antécédent de sinistres.

La lecture des odds-ratios nous emmène à définir la règle suivante : les clients ayant souscrit à une assurance auto chez GAN avec une ancienneté comprise entre 1 et 3 ans, avec un CRM de plus de 1 et ayant au moins 2 sinistres lors des 24 derniers derniers mois sont les plus à même de résilier leurs contrats.

Détection des contrats à fort risque de résiliation avec les méthodes d'apprentissage automatique

Dans le chapitre précédent, nous avons modélisé la probabilité de résiliation d'un contrat avec la régression logistique. C'est un modèle paramétrique qui suppose une loi *a priori* du comportement des assurés en termes de résiliation.

Mais les données ne se comportent pas aussi bien souvent qu'on le souhaite *a priori* ; supposer que le logit de la probabilité de résilier un contrat est linéaire en ses paramètres, est une grande hypothèse. C'est pourquoi, nous allons dans ce chapitre challenger les modèles précédents par des méthodes d'apprentissage automatique. Le credo de ces méthodes est qu'elles essayent de rendre compte le plus fidèlement possible du comportement des données, sans aucune supposition de lois *a priori*. Les amateurs de ces méthodes diront qu'ils font parler les données et uniquement les données.

Pour ce faire, nous allons dans un premier temps annoncer le canevas méthodologique que nous suivrons tout au long de ce chapitre. Ensuite, nous procéderons à la mise en oeuvre des différents modèles candidats selon les conclusions que nous pouvons tirer du comportement des données.

Dans ce chapitre, sauf mention contraire, nous nous baserons sur les deux types d'échantillonnages effectués dans la section 3.3.1.

4.1 Le canevas méthodologique

Par *canevas méthodologique*, nous entendons tout le processus du raisonnement et du choix des modèles que nous mettrons en oeuvre dans ce chapitre. En effet, tous les problèmes de classification ne s'abordent pas de la même manière. Entre les observations, les statistiques univariées et la distribution des classes, l'actuaire est amené à se diriger vers un ensemble précis de méthodes d'apprentissage automatique.

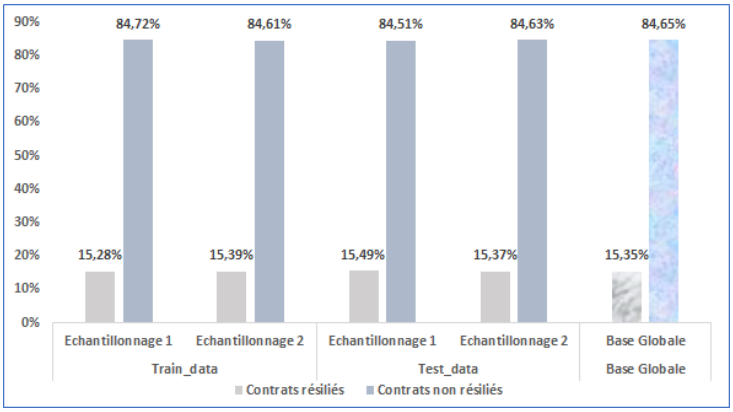
4.1.1 Constat de données déséquilibrées

Comme c'est souvent le cas dans la pratique, l'observation de la figure 4.1 montre le caractère déséquilibré des deux classes de notre variable d'intérêt. De manière générale dans nos différents échantillons, il y a environ 15% de contrats résiliés - "1" contre 85% de contrats non résiliés - "0".

Ce déséquilibre suggère une sous-représentation des contrats résiliés. Fort de ce constat,

des traitements s'imposent selon les cas pour mieux construire un modèle robuste qui reflète le vrai comportement des assurés. Selon les pratiques et les cas, la prise ne compte de ce déséquilibre se fait soit en agissant directement sur les paramètres du modèle afin de pénaliser la classe la moins représentée ou soit en agissant dans un premier temps sur les données d'entrainement.

FIGURE 4.1 – Répartition des contrats résiliés et non résiliés par échantillon



Note de lecture : Répartition déséquilibrée des contrats dans les différents échantillons. De manière générale, il y a 85% de contrats non résiliés et 15% de résiliation.

4.1.2 Méthodologie

Comment est-ce qu'il faut aborder un problème de qualification ? Que faire dans un environnement de données déséquilibrées ? Des prospections sur le sujet nous ont emmenés à adopter une démarche progressiste dans laquelle les conclusions d'une étape orientent le choix du traitement que nous adoptons à l'étape suivante.

Étape 1 : Analyse Factorielle de Données Mixtes - AFDM

Dans notre processus de détection des profils à fort risque de résiliation de contrat d'assurance automobile, nous débutons par une analyse factorielle de données mixtes. Elle nous permettra de conclure si les deux classes se comportent différemment et distinctement. Pour observer cela, il nous suffira de projeter les individus - les contrats - sur les deux premiers axes factoriels par exemple en fonction de la variable cible - Résiliation - et observer si les deux populations se distinguent clairement sur le graphe obtenu.

Étape 2 : Un réseau de neurone à l'entrainement

Si cette première étape se révèle concluante c'est-à-dire que l'AFDM permet de distinguer les deux classes correctement, alors l'étape 2 de notre processus de modélisation se résumera à la construction d'un réseau de neurones qui permettrait de bien discriminer les deux classes. Par exemple, nous pouvons entraîner un réseau de neurones sur l'une

des deux classes. Typiquement, ce dernier va apprendre par exemple à connaître uniquement le comportement des contrats non résiliés. Puisque les deux classes sont distinctes en termes de comportement dans les données, alors dans l'échantillon test, le réseau peut facilement reconnaître ce qu'il a appris de ce qu'il ne connaît pas. Le plus populaire dans ce cas de figure est le réseau GAN - *Generative Adversarial Networks*. Et dans ce cas, ce sera la fin de notre processus de modélisation ; nul besoin d'aller chercher d'autres modèles.

Sinon, on passe à l'étape 3.

Étape 3 : Application des modèles assemblistes - Random Forest ou XGBoost sans ré-échantillonnage

Lorsque l'étape 1 ne permet pas de distinguer les classes, nous passons à l'application de quelques modèles de classification notamment le Random Forest - la forêt aléatoire en français - ou encore le XGBoost - *eXtreme Gradient Boosting*. Selon les modèles, nous aurons la possibilité de pénaliser la classe minoritaire dans le souci d'améliorer le pouvoir discriminant des classifieurs. Particulièrement, comme nous le verrons par la suite, le XGBoost permet d'allouer des poids différents aux différentes classes pour tenir compte du déséquilibre existant entre ces dernières. Si les résultats obtenus ne sont pas concluant, on passe à l'étape 4.

Étape 4 : Agir sur les données d'entrainement - le ré-échantillonnage

Lorsque l'étape 3 ne permet pas d'obtenir des résultats robustes, il faut agir sur les données qui ont servi à entraîner les modèles. Il s'agit notamment des méthodes de ré-échantillonnage qui permettent de rééquilibrer les deux classes dans la base d'entrainement. On distingue trois (03) méthodes : ***l'oversampling***, ***l'undersampling*** et ***le smote***. L'oversampling consiste à créer de nouveaux individus de la classe minoritaire , l'undersampling, à supprimer les individus de la classe majoritaire. Quant à la méthode smote, elle mixe les deux premières.

Après cette étape de ré-échantillonnage en vue de rééquilibrer l'échantillon d'entrainement, on passe à l'étape 5.

Étape 5 : Demi-tour à "2 vitesses" : retour à l'étape 3

Cette dernière étape consiste à appliquer nos modèles assemblistes sur les données d'entrainement ré-échantillonnées comme à l'étape 3. Seulement, il faudra faire attention dans ce cas de ne pas appliquer des pénalisations selon les classes par exemple pour l'algorithme XGBoost. De manière générale, cette étape permet d'obtenir de meilleurs résultats par rapport à l'étape 3 avant ré-échantillonnage.

Dans la suite de ce chapitre, nous suivrons ces différentes étapes et justifierons nos choix à chaque étape. Les étapes décrites ci-dessus peuvent être résumées sur la figure ci-dessous :

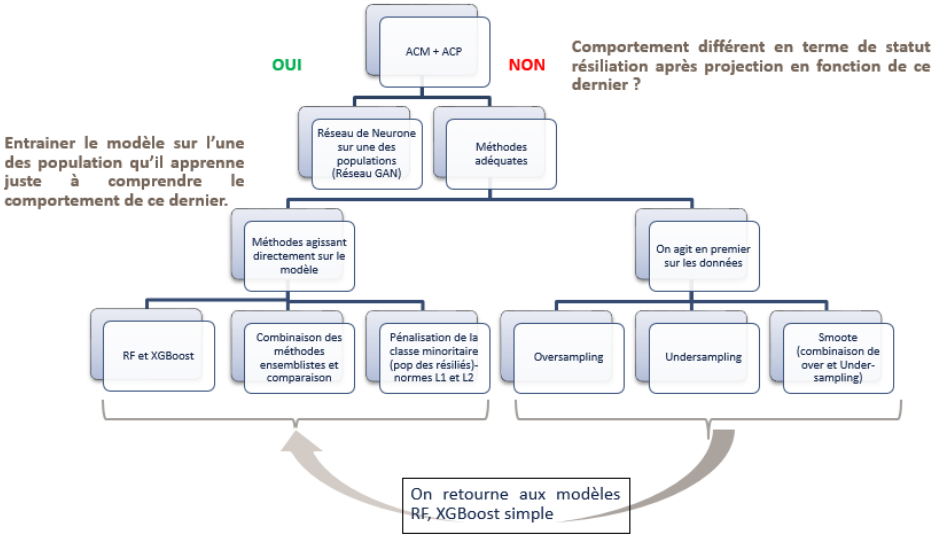


FIGURE 4.2 – Les différentes étapes de l'étude

4.2 L'analyse factorielle de Données Mixtes - AFDM

Comme nous l'avons évoqué dans le canevas méthodologique, le point de départ de la détection des profils à fort risque de résiliation est l'analyse factorielle de données mixtes. L'objectif est de voir si les contrats résiliés et non résiliés se distinguent clairement les uns des autres. L'AFDM est une combinaison de l'ACP - Analyse en Correspondance Principale et l'ACM - Analyse en Correspondance Multiple qui sont utilisées respectivement sur des données quantitatives et qualitatives. Puisque nous avons les deux types de données, nous optons pour l'AFDM. Toutes les variables sont actives à l'exception de la variable **Resiliation**. Les variables quantitatives sont normalisées.

Sans trop nous attarder sur l'aspect théorique de l'AFDM, nous avons procédé à sa mise en oeuvre. Comme nous le verrons à la fin de cette section, les résultats ne sont pas concluants car l'AFDM ne permet pas de distinguer clairement les deux groupes. Néanmoins, il convient de discuter un peu des résultats obtenus.

Nous n'interpréterons que les 2 premiers axes de l'AFDM. Ils n'expliquent que 5,7% de la variabilité (inertie) totale contenue dans les données.

Pour interpréter (ou caractériser) un axe factoriel dans une AFDM avec les variables de l'analyse, on repère celles qui sont les plus liées à cet axe. L'intensité de liaison entre l'axe factoriel et une variable se mesure par le rapport de corrélation.

Pour interpréter un axe factoriel dans une AFDM avec les modalités de l'analyse, on repère celles (pour les variables actives) ayant contribué le plus à la construction de cet axe.

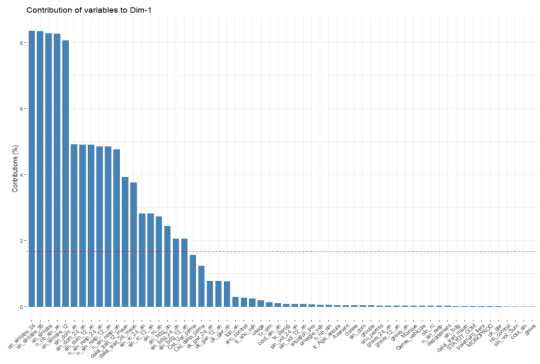


FIGURE 4.3 – La contribution de variables au premier axe factoriel

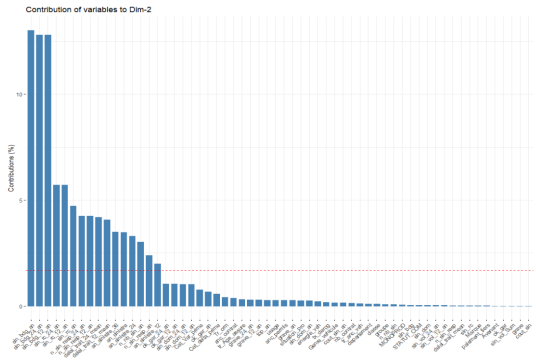


FIGURE 4.4 – La contribution de variables au deuxième axe factoriel

On y voit que les **antécédents de sinistres**, **le sinistre dommage** et les **délais de traitement** sont les plus liés à l'axe 1. Les sinistres **Bris de glaces** et **responsabilité civile** sont plutôt plus liés à l'axe 2. On remarque là que ce sont les variables qui décrivent la sinistralité du contrat et sa gestion sont celles qui décrivent les deux premiers axes comme nous pouvons le voir sur la figure A.12.

Enfin, nous avons projeté les individus dans les deux premières dimensions tout en les distinguant selon que les contrats soient résiliés ou non. L'objectif est de voir si les deux classes se comportent différemment dans les données. Nous avons réalisé cela en nous basant uniquement sur les données quantitatives donc sur une ACM.

FIGURE 4.5 – Projection des contrats sur les deux premiers axes factoriels

La figure 4.5 montre que les deux classes ne sont pas distinguables. Et par conséquent, nous ne pouvons pas entraîner des réseaux GAN - Generative Adversarial Networks - sur les données non résiliés mais plutôt, comme annoncé dans le canevas méthodologique, nous allons employer des méthodes ensemblistes du type Random Forest.

4.3 Généralités sur les méthodes d'apprentissage automatique

L'objectif de ce chapitre est de détecter les profils à forte propension de résilier en assurance auto à l'aide des méthodes d'apprentissage automatique. Mais avant de passer aux différentes méthodes que nous mettrons en oeuvre, nous avons jugé qu'il est bien de commencer par comprendre de manière générale, ce que c'est qu'un modèle d'apprentissage automatique et les différentes formes qu'on a.

4.3.1 L'apprentissage automatique

L'apprentissage automatique ou communément appelée en anglais *machine learning* est une branche de l'intelligence artificielle qui se base sur les mathématiques, statistiques afin de donner aux ordinateurs la capacité d'apprendre à partir des données disponibles. En d'autres termes, ce sont des méthodes qui apprennent par elles-mêmes sans qu'on ne les programmes préalablement explicitement pour les tâches qu'elles exécutent. L'un des premiers algorithmes connu du domaine est le Perceptron introduit par F. Rosenblatt dans le célèbre papier *« The perceptron : a probabilistic model for information storage and organization in the brain »* (1958).

C'est un domaine très récent et qui est en pleine progression. En effet, avec l'avènement des big data (les données des sites internet, les gros volumes de données), les outils analytiques classiques du statisticien trouvent ses limites. Les méthodes d'apprentissage automatiques, en revanche, arrivent à déceler les similitudes, les ressemblances et l'ensemble des informations dans un amas de données. Rappelons-le, ce sont des méthodes qui apprennent, donc généralement, plus il y a de données, mieux elles apprennent et donc mieux elles améliorent leurs performances.

De manière générale, leur principe est simple et se fait en 2 phases. La première dite la phase d'entrainement ou d'apprentissage consiste à donner en entrée du modèle des données pour qu'il apprenne à les comprendre et à détecter les ressemblances, les groupes. La seconde est celle de prédiction. En effet, maintenant que le modèle a appris, on lui donne de nouvelles données et sur la base de ses connaissances des données, le modèle prédit un comportement d'intérêt ciblé par l'utilisateur.

4.3.2 Les grands types d'apprentissage

Il existe 2 grands types d'apprentissage d'après la configuration des données en entrée lors de la phase d'apprentissage. Ce sont *l'apprentissage supervisé* et *l'apprentissage non supervisé*.

Apprentissage supervisé

On parle d'apprentissage supervisé lorsque les données d'entrainement sont étiquetées. Il est souvent le cas dans les problèmes de classification où les données sont présentées sous forme d'une collection de couples (X_i, Y_i) où X_i est l'ensemble des variables explicatives et Y_i la variable d'intérêt. Ainsi, préalablement, les données sont telles que les classes liées à X_i sont connues. Le modèle n'a donc qu'à apprendre et à la fin, lors de la deuxième phase, connaissant uniquement les variables explicatives, tenter de prédire la classe correspondante.

Intrinsèquement dans les algorithmes, c'est plutôt la probabilité d'appartenir aux différentes classes qui est calculée.

Apprentissage non supervisé

On parle d'apprentissage non supervisé lorsque les données d'apprentissage ne sont pas couplées au préalable. En effet, c'est à l'algorithme de découvrir seul les groupes , les différentes classes et la structure des données. La construction des groupes homogènes à partir d'un jeu de données ou le *data clustering* est un exemple classique d'apprentissage non supervisé.

4.3.3 Le biais et la variance d'un prédicteur

Le biais et la variance sont deux notions importantes dans la construction d'un prédicteur. En apprentissage supervisé, le meilleur prédicteur est celui qui trouve un juste milieu entre ces deux notions.

Le bais mesure la différence entre les vraies valeurs et les valeurs prédites par le modèle sur les données d'apprentissage. C'est donc une mesure de l'erreur. Un bais élevé peut être le résultat d'un algorithme qui n'arrive pas à comprendre les relations entre les données : on parle alors de sous-apprentissage.

La variance d'un prédicteur mesure sa capacité à s'adapter à de nouvelles données. La variance sert donc à mesurer la sensibilité des modèles aux fluctuations dans les données d'apprentissage. Une variance élevée signifie que le modèle "apprend" bien à connaître les relations entre les données d'entrainement mais n'arrive pas à révéler ces relations sur de nouvelles données : c'est le surapprentissage.

Dans notre problème, il s'agit d'un problème d'apprentissage supervisé car pour chacune des observations, nous connaissons préalablement si le contrat est résilié ou non. Les deux algorithmes d'apprentissage supervisé que nous utiliserons sont la forêt aléatoire - Random Forest en anglais - et l'XGBoost. Mais avant d'appliquer ces algorithmes sur nos

données, il est important de comprendre leurs fonctionnements et leurs constructions. Une forêt étant littéralement une combinaison de plusieurs arbres, nous allons d'abord voir la construction d'un arbre de décision : arbre CART.

4.4 L'arbre de décision CART

4.4.1 Notations et position du problème

Nous nous plaçons dans le cadre de la classification supervisée. Notre problème est de prédire la classe -résilié ou non résilié - des contrats en portefeuille. Considérons à cet effet :

— $Y \in \mathcal{Y}$ où $\mathcal{Y} = \{0, 1\}$; 0 pour les contrats non résiliés et 1 pour les contrats résiliés ;
— $X \in \mathcal{X}$ qui désigne la collection des p variables explicatives ; lesquelles pouvant être un mélange des variables qualitatives et quantitatives.

Notre problème de classification supervisée se formule comme suit :

— nous nous intéressons au couple aléatoire (X, Y) ;
— Nous avons les données d'échantillon $\mathcal{E}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ qui sont les réalisations de n variables aléatoires de même loi que (X, Y) ; les X_i sont appelées les données d'entrée et les Y_i, celles de sortie,
— nous cherchons à calculer les probabilités *a posteriori* pour que Y appartiennent à chacune des classes conditionnellement à une réalisation de $X = x$ fixé c'est à dire :

$$\forall c \in \{0, 1\},\ \mathbf{P}(Y = c | X = x).$$

Bien évidemment, ces probabilités ne sont pas toujours égales à 1 pour une classe et 0 pour les autres à cause du bruit contenu dans les données d'entrée. Mais la fixation d'un seuil permet de palier à ce problème et enfin, d'affecter la classe prédite selon les informations d'entrée x.

4.4.2 Les arbres de décision : l'algorithme CART

Les prérequis de la compréhension de l'algorithme des forêts aléatoires sont les principes et les modes de construction des arbres de décision. Un arbre de décision a pour objectif d'expliquer une variable à partir des informations disponibles. Dans cette sous section, nous allons nous attarder sur l'arbre CART qui signifie - en anglais - *Classification And Regression Trees*, introduit par le statisticien américain Leo Breiman et al. en 1984. Nous nous plaçons dans le cadre de la classification car l'arbre CART peut aussi être utilisé en régression.

Le principe de CART

Le principe de CART est de partitionner de manière récursive les individus de l'espace d'entrée $\mathcal{X}$ en deux groupes les plus homogènes possibles du point de vue de la variable à prédire Y. Le partitionnement est basé sur une hiérarchie de la capacité prédictive des variables explicatives. Cela permet essentiellement d'avoir une visualisation des résultats dans un arbre et enfin formuler des règles orientées métier pour expliquer notre variable d'intérêt.

Pour définir les règles de décision, plusieurs itérations sont nécessaires. A chaque itération, on divise les individus en 2 groupes homogènes pour expliquer Y.

- Première division : elle est définie par la variable explicative qui fournit la meilleure séparation des individus ; ce qui définit alors les sous-populations et est représenté par un « noeud» de l'arbre ;
- La deuxième division : Elle se fait au sein de chacune des deux sous-populations obtenues à la première division. De même, elle est obtenue par la meilleure variable explicative qui sépare bien chaque sous-population ;
- .... ;
- .... ;
- La division terminale : On continue de segmenter les populations précédentes jusqu'à ce qu'aucune division ne soit possible. On obtient alors au bout de chaque branche de l'arbre les « feuilles ».

Sur la figure 4.6, que nous avons réalisée sur nos données, il est bien observable l'arborescence de l'arbre, les noeuds et les feuilles.

Construction de l'arbre CART

La construction de l'arbre CART se fait en deux phases : la construction de l'arbre maximal et la phase d'élagage ; la première permet d'avoir un ensemble de modèles parmi lesquels nous chercherons à sélectionner le meilleur ; et la seconde permet de construire une suite de sous-arbres optimaux à partir de l'arbre maximal.

La construction de l'arbre maximal

A partir de la racine de l'arbre CART, on obtient les deux noeuds fils en partitionnant l'espace en deux parties à l'aide de la variable X^j selon la découpe suivante :

- Si X^j est une variable quantitative alors : $\{X^j \leq d\} \cup \{X^j \geq d\}$

- Si X^j est une variable qualitative, alors : $\{X^j \in d\} \cup \{X^j \in \bar{d}\}$

où $j \in \{1, ..., p\}$ et $d \in \mathbf{R}$ pour les variables quantitatives et d et $\bar{d}$ sont des ensembles non vides formant une partition des modalités de la variable qualitative X^j. Pour l'explication de la suite de la construction, nous allons nous baser sur les variables qualitatives. La réplication pour les variables quantitatives est assez triviale.

La partition $\{X^j \in d\} \cup \{X^j \in \bar{d}\}$ signifie que toutes les modalités de la $j^{\text{ième}}$ variable explicative qui sont dans l'ensemble d définissent le noeud gauche et les autres le noeud droit de la racine.

La construction de cette première ramification consiste à trouver le meilleur couple (j, d) qui minimise la fonction d'impureté du noeud : l'indice de GINI Φ définie pour un noeud t :

$$\Phi(t) = \hat{p}^t{}_0(1 - \hat{p}^t{}_0) + \hat{p}^t{}_1(1 - \hat{p}^t{}_1)$$

où $\hat{p}^t{}_1$ et $\hat{p}^t{}_0$ représentent respectivement les proportions d'observations de la classe 1 et 0 contenues dans le noeud t. Cela conduit à maximiser pour chaque noeud t la fonction

$$\Phi(t) - \left(\frac{\#t_L}{\#t} \Phi(t_L) + \frac{\#t_R}{\#t} \Phi(t_R) \right)$$

avec t_R le noeud de droite et t_L, celui de gauche.

L'objectif est de chercher à diminuer la fonction d'impureté de Gini afin d'avoir les noeuds homogènes et purs. Un noeud homogène est un noeud qui comporte les observations de la même classe.

Après avoir effectué cette première division, l'algorithme effectue la même tâche au niveau de chaque noeud et ainsi de suite jusqu'à obtenir un noeud terminal indivisible appelé "feuille". De manière générale, l'utilisateur peut fixer une condition d'arrêt des ramifications. Le plus souvent, il peut par exemple fixer le nombre d'observations minimal dans un noeud. On obtient alors à la fin un arbre totalement développé appelé arbre maximal. La classe majoritaire de chaque feuille définit une règle de décision. La combinaison de ces règles définit un prédicteur.

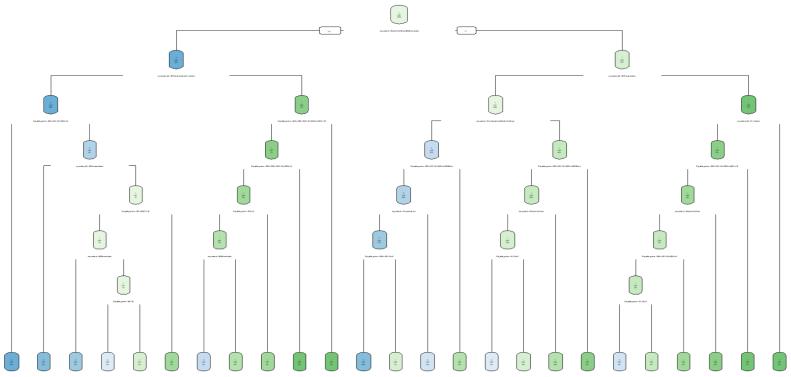


FIGURE 4.6 – Exemple de l'arbre maximale CART construit sur 4 variables

La phase d'élagage

La phase d'élagage de l'arbre maximal consiste à choisir le meilleur sous-arbre élagué parmi l'ensemble des sous-arbres ayant la même racine que l'arbre maximal. C'est une sorte de sélection de modèles dans le sens où tous les sous-arbres ayant la même racine que l'arbre maximal sont des modèles admissibles. De ce fait, la phase d'élagage permet de sélectionner le sous-arbre qui minimise le taux de mauvais classement.

Cette étape fait un compromis entre deux indicateurs d'un prédicteur du type arbre de décision. En effet, l'arbre maximal possède une très grande variance mais a un biais très faible. *A contrario*, un arbre constitué uniquement de la racine a une variance très faible mais n'arrivera pas à bien prédire les classes : un biais très élevé. L'élagage permet de trouver un juste milieu entre biais de prédiction et variance. Le sous arbre optimal de l'arbre maximal 4.6 est ci-dessous sur la figure 4.8. Nous pouvons observer qu'il a moins de feuilles par rapport à l'arbre maximal.

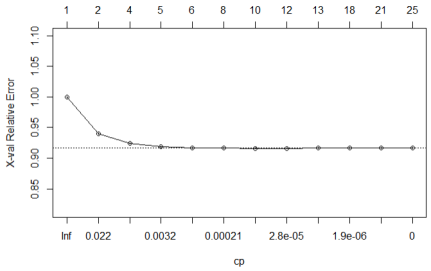


FIGURE 4.7 – Taux d'erreur en fonction de la profondeur d'arbre

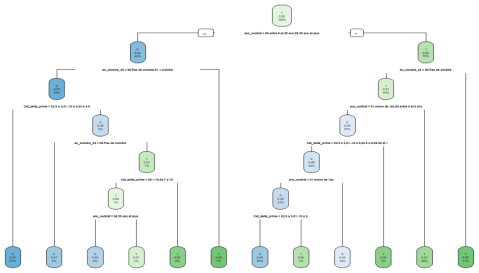


FIGURE 4.8 – L'arbre de la figure 4.6 élagué

Sous R, la construction et l'élagage de l'arbre maximal se font grâce au package rpart. La fixation de la bonne valeur du paramètre *cp* permet d'obtenir le bon arbre qui donne le meilleur compromis entre la variance et le biais et d'éviter le surapprentissage. Ce paramètre définit la complexité de l'arbre et son choix consiste à observer la courbe d'évolution de l'erreur de classement comme sur la figure 4.7.

Les limites de l'arbre CART

L'arbre CART a de nombreuses limites. En effet, les prédicteurs basés sur CART sont très versatiles et très sensibles aux données d'entrainement. Les prédicteurs basés sur CART changent facilement d'opinions. Une faible fluctuation introduite dans les données générera un résultat différent ; en d'autres termes, c'est un modèle qui a une variance élevée. Dans le jargon des data scientistes, l'arbre CART fait partie de la famille des prédicteurs faibles ou mieux les *weak predictors* en anglais.

Pour palier à la versatilité de CART, Leo Breiman va introduire le concept des méthodes ensemblistes qui vont se baser sur les *weak predictors* pour construire des méthodes

plus robustes et qui donnent de meilleurs résultats. Ces méthodes, comme nous le verrons, permettent de réduire la variance du prédicteur. Random Forest et XGBoost font partie de cette famille.

4.5 Le principe des méthodes ensemblistes

Le principe des méthodes ensemblistes est très simple et très intuitif. Il est basé sur la combinaison de plusieurs *weak predictors*. Cela permet en effet d'obtenir des résultats plus robustes. En effet, si l'on combine la prédiction de plusieurs prédicteurs, l'on n'est au moins sûr de faire mieux que le plus mauvais de tous. L'intuition derrière les méthodes ensemblistes s'illustre bien avec cet exemple : *« si un médecin (et un seul) vous annonce que vous avez probablement une maladie grave qui nécessite une intervention chirurgicale, que feriez-vous ? [..] avant de prendre la moindre décision, vous "demandez" à d'autres personnes/ experts leur avis avant de vous faire le vôtre. Parce que aussi experte que soit une personne dans un domaine, sa voix seule ne suffit pas, en général, pour prendre la bonne décision»* [4].

C'est exactement la même chose qui se fait au niveau des méthodes ensemblistes. De manière générale, il y a deux phases dans la mise en place de ces modèles :

— Construction de plusieurs *weak predictors* ;
— la phase d'agrégation de ces prédicteurs.

La phase d'agrégation dans le cadre de la classification ressemblerait aux résultats d'une élection démocratique. Elle consiste à attribuer la classe du vote majoritaire de la population des *weak-predictors*. Par exemple, si nous construisons 1000 arbres CART et que 501 prédisent la classe 0 et 499 , la classe 1, alors notre méthode ensembliste basée sur les 1000 arbres CART va prédire la classe 0 pour l'observation concernée.

Toutefois, il est important de faire attention entre la comparaison des méthodes ensemblistes et les *weak-predictors*. En effet, l'on perd une partie d'explicabilité du modèle mais l'on gagne en performance du modèle.

4.6 La forêt aléatoire : le Random Forest

4.6.1 Historique et motivations

Contrairement à la régression logistique qui suppose une loi *a priori*, les forêts aléatoires - *Random Forest en anglais* - , introduites par Leo Breiman dans les années 2000, sont une méthode non paramétrique et sont très utilisées en statistique moderne pour résoudre les problèmes de classification. Elles font partie de la famille des méthodes en-

semblistes et ont de très bonnes performances. En effet, les évaluations menées par des journaux spécialisés comme *Journal of Machine Learning Research* ou encore *Pattern Recognition Letters* sur les algorithmes de classification révèlent que les forêts aléatoires figurent dans les 2 et 3 meilleurs [14].

Dans le papier « *Do we Need Hundreds of Classifiers to Solve Real World Classification Problems ?* », les auteurs Fernández-Delgado et al. [10] ont évalué 179 classifieurs en 2014 provenant de 17 familles notamment l'analyse discriminante, l'analyse bayésienne, les réseaux de neurones ou encore les k-voisins les plus proches. Ils en sont arrivés à la conclusion selon laquelle les forêts aléatoires sont les meilleurs classifieurs (3 des 5 meilleurs algorithmes de classification sont basés sur les forêts aléatoires). C'est la première raison qui nous a poussé à utiliser les forêts aléatoires pour classer nos contrats selon leur risque de résiliation.

De plus, avec les données déséquilibrées, un seul prédicteur (la régression logistique ou l'arbre CART) a une variance élevée. Dans le souci de minimiser cette variance, nous avons choisi d'utiliser les méthodes de la famille ensembliste dont les forêts aléatoires.

4.6.2 L'algorithme du Random Forest

Le Random Forest repose sur l'assemblage de plusieurs estimateurs à base de l'arbre de décision. Au lieu de trouver un moyen pour résoudre les problèmes d'instabilité de CART, Breiman a eu l'idée de combiner plusieurs arbres CART pour former une "forêt " - *forest* - d'arbres. Sa construction s'appuie sur deux principes fondamentaux : le *bagging* et le *sampling*.

$$Random\ Forest = Bagging + Sampling$$

Bagging

La dénomination *Bagging* vient de la contraction des mots anglais ***B**oostrap* et ***Agg**regat**ing***. Introduite par Leo Breiman en 1996, il constitue une méthode à part entière de la famille des méthodes ensemblistes. Son principe est le suivant :

— A partir de l'échantillon d'entrainement initial $\mathcal{E}_n$, on tire de manière indépendante, b échantillons boostrap $\left(\mathcal{E}_n^{\Theta_1}, \ldots, \mathcal{E}_n^{\Theta_b}\right)$. L'échantillonnage par boostrap $\mathcal{E}_n^{\Theta_l}$ est par exemple obtenu en faisant un tirage aléatoire, successif et avec remise de n observations dans l'échantillon initial $\mathcal{E}_n$. De ce fait, chaque observation a la même probabilité $p = 1/n$ d'être tirée et Θ_l représente ce tirage aléatoire ;
— On entraîne sur chacun des b échantillons, un arbre de décision du type CART. L'on obtient alors b *weak predictors* $\left(\hat{h}(.,\mathcal{E}_n^{\Theta_1}), \ldots, \hat{h}(.,\mathcal{E}_n^{\Theta_b})\right)$;

— La dernière étape du principe de bagging est l'agrégation des b prédicteurs construits à l'étape précédente. Cette étape permet d'avoir un prédicteur plus performant $\hat{h}_{BAG}(.)$. L'on peut démontrer facilement que la variance du modèle agrégé est plus faible que celle d'un arbre CART simple, pris seul. La démonstration est en fin de cette sous-section.

Nous pouvons résumer l'étape de bagging de la construction du Random Forest sur la figure ci-dessous.

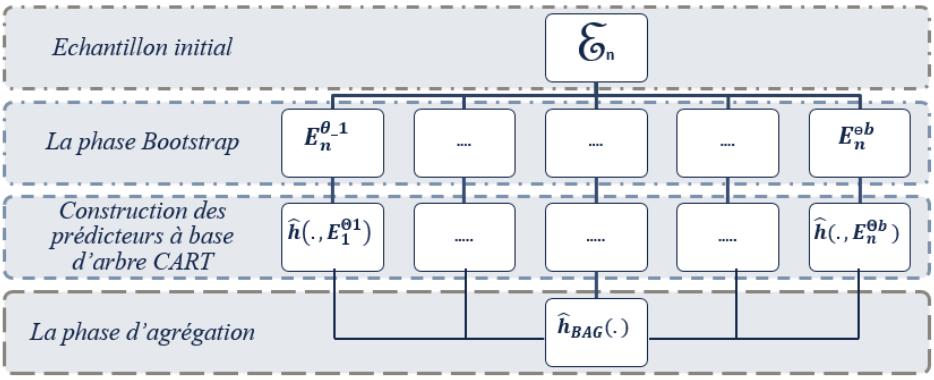


FIGURE 4.9 – Illustration de l'algorithme Bagging

Sampling

Même si l'algorithme de Bagging donne des résultats satisfaisants, il lui est souvent reproché une forte ressemblance entre les prédicteurs générés sur les échantillons boostrap. En effet, le fait que les prédicteurs soient construits sur le même nombre de variables, même si les observations sont tirées aléatoirement, introduit une corrélation entre les prédicteurs dans le sens que certains parmi eux peuvent avoir les mêmes performances et règles de décision.

Dans le but de réduire la corrélation entre les arbres, la phase de la construction des prédicteurs se fait dans l'algorithme de Random Forest différemment. En effet, les divisions au niveau des noeuds d'un arbre CART sont déterminées par un jeu de m variables tirées aléatoirement parmi l'ensemble des p variables explicatives : c'est le sampling. Les m variables sont tirées sans remise et uniformément ; ainsi chaque variable à une probabilité $1/p$ d'être tirée. Deux dilemmes sont à concilier lors de cette phase :

— m ne doit pas être trop faible car dans le cas échéant, les arbres ne porteraient pas assez d'informations ;
— m ne doit pas être trop grand car dans le cas échéant, les arbres de la forêt seraient suffisamment corrélés.

Il faut donc trouver un juste milieu qui concilie ces deux extrêmes pour avoir un bon prédicteur agrégé. Notons que le choix des variables introduit un second aléa en ajout à celui du tirage boostrap de la méthode bagging. C'est cela qui confère le surnom "aléatoire" - *random en anglais* à la forêt d'arbres. En général, la racine carrée du nombre de variables est considérée comme raisonnable.

Remarque 1 : Le nombre m de variables à tirer est fixé dans la mise en oeuvre du Random Forest et est identique à tous les arbres mais il convient de mentionner que les m variables qui interviennent dans les divisions de deux noeuds sont différentes généralement. C'est un paramètre très important dans la mise en oeuvre de l'algorithme.

Remarque 2 : Les arbres CART construits dans le cas du Random Forest sont des arbres maximaux et non élagués. Un cas particulier d'une forêt d'arbres construite avec $m = p$ revient tout simplement à faire du bagging d'arbres CART non élagués.

Avant de mettre en oeuvre le modèle, nous voulons démontrer que le Random Forest permet de réduire la variance. Pour cela, considérons b échantillons et, par conséquent, b arbres CART. Pour ne pas alourdir les calculs, nous allons noter ici T_i le $i^{\text{ème}}$ arbre. Supposons que pour tout $i \neq j$, $Corr(T_i, T_j) = \rho > 0$ et la variance de l'arbre T_i est égale à $Var(T_i) = Var(T_j) = \sigma^2$. Alors, la variance de la forêt aléatoire qui peut s'écrire sous la forme $F = \frac{1}{b}\sum_{i=1}^{b} T_i$ est :

$$\begin{aligned}
Var(F) &= Var(\frac{1}{b}\sum_{i=1}^{b} T_i) \\
&= \frac{1}{b^2}\sum_i \sum_j Corr(T_i, T_j) \\
&= \frac{1}{b^2}\sum_i \left(\sum_{i\neq j} Corr(T_i, T_j) + Var(T_i)\right) \\
&= \frac{1}{b^2}\sum_i \left((b-1)\rho\sigma^2 + \sigma^2\right) \\
&= \frac{b(b-1)\rho\sigma^2 + b\sigma^2}{b^2} \\
&= \rho\sigma^2 + \frac{(1-\rho)}{b}\sigma^2 \\
&= \sigma^2\left[\frac{1}{b} + \rho(1-\frac{1}{b})\right]
\end{aligned}$$

D'après l'expression de la variance, on peut voir que moins les arbres sont corrélés, plus faible est la valeur de ρ est donc plus le Random Forest permet de réduire la variance.

4.6.3 Mise en place du modèle et le calibrage des paramètres

Les paramètres

Bien que le Random Forest soit un modèle très simple à utiliser avec très peu de paramètres à contrôler, il n'en est pas moins vrai que son calibrage permet non seulement d'éviter les problèmes de surapprentissage - *overfitting en anglais* mais aussi d'obtenir un meilleur modèle prédictif. Les deux paramètres les plus importants sont le nombre de variables m et le nombre b d'arbres de la forêt. Ces paramètres sont respectivement désignés sous R par *mtry* et *ntree* dans la fonction *randomforest* du package du même nom. Le choix de la profondeur de la forêt est fait en observant la courbe d'évolution de l'erreur de généralisation en fonction de *ntree*.

L'erreur de généralisation du Random Forest

L'algorithme du Random Forest est très riche. Il ne fait pas que construire le prédicteur mais calcule en parallèle, en son sein une estimation de son erreur de généralisation. Cette erreur est dénommée dans les sorties de l'algorithme l'erreur Out-Of-Bag (OOB) où "Out-Of-Bag" signifie "en dehors du boostrap". Il est calculé lors de l'étape bagging (d'où le mot "Bag" dans sa dénomination). Son calcul se fait suivant le procédé suivant dans l'algorithme du Random Forest :

- Dans l'échantillon d'apprentissage $\mathcal{E}_n$, il fixe une observation (X_i, Y_i) ;
- Il construit les prédicteurs - les arbres CART - sur tous les échantillons bootstrap ne contenant pas l'observation (X_i, Y_i). En d'autres termes, on construit les prédicteurs sur les échantillons pour lesquels l'observation (X_i, Y_i) est "Out-Of-Bag" ;
- L'algorithme agrège les prédicteurs précédents pour construire une règle de décision finale ;
- Ensuite, il réalise une prédiction $\hat{Y}_i$ de Y_i ; il refait les opérations précédentes pour toutes les observations de l'échantillon d'apprentissage $\mathcal{E}_n$ pour obtenir les prédictions de tous les Y_i ;
- Enfin, il calcule l'erreur de prédiction. En classification, il s'agit de la proportion des observations mal prédites - classées $(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{Y}_i \neq Y_i})$. C'est cette quantité qui est l'erreur Out-Of-Bag.

Remarquons que l'erreur OOB est calculée sur des données qui n'ont pas encore été rencontrées par les prédicteurs. Cette construction a l'avantage de ne nécessiter aucun découpage des données comme dans le cadre de la validation croisée ou calcul de l'erreur par échantillon test. C'est l'un des avantages du Random Forest.

Nous utiliserons cette erreur pour définir la profondeur de notre forêt. En effet, lorsque

cette erreur n'évolue plus, au lieu de "peupler" la forêt avec un grand nombre d'arbres avec les risques de surapprentissage, il vaut mieux prendre le nombre d'arbres à partir duquel cette erreur ne diminue plus significativement.

Choix des paramètres du Random Forest

Dans la littérature, le choix de m pour le sampling est fait de sorte à ce que $m = \sqrt{(p)}$ avec p le nombre total de variables explicatives. Dans notre cas, nous allons prendre la partie entière de $\sqrt{(p)}$.

Pour obtenir la valeur idéale de *ntree* , nous avons réalisé dans un premier temps un modèle de Random Forest avec sa valeur par défaut ($ntree = 500$). L'objectif est de regarder à partir de quelle valeur l'erreur OOB n'évolue plus de manière significative. D'après la figure 4.10, , on remarque qu'à partir de 50 arbres, l'erreur OOB devient stable. Par conséquent, nous fixons la valeur de $ntree = 50$.

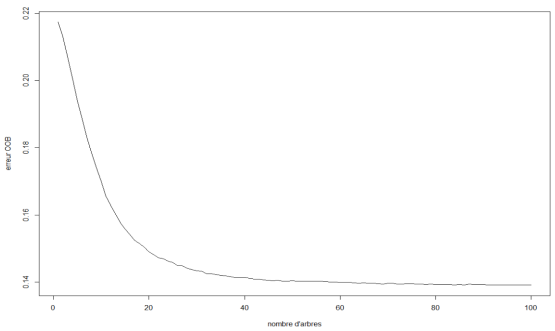


FIGURE 4.10 – L'erreur Out Of Bag par nombre d'arbres

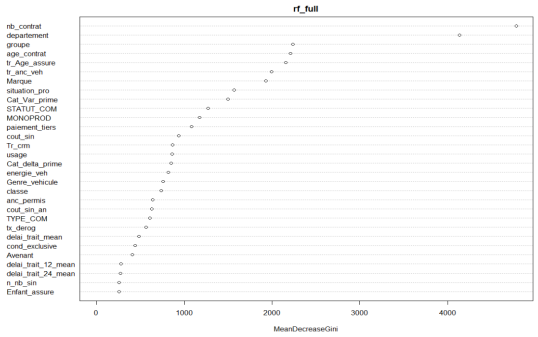


FIGURE 4.11 – Importance des variables - Random Forest

Le Random Forest nous donne la possibilité de regarder les variables qui ont le plus contribué à la segmentation de l'espace des variables explicatives. C'est ce qui est observable sur la figure 4.11. Nous remarquons que le département, le nombre de contrats, le groupe du véhicule, l'ancienneté du contrat, l'âge de l'assuré ou encore la marque du véhicule et la situation socioprofessionnelle de l'assuré sont les variables qui ressortent comme étant les plus importants. Le fait que ces variables ressortent s'inscrit dans l'une des limites du Random Forest qui privilégie les variables ayant un nombre élevé de modalités.

... Zoom sur l'importance des variables

Les méthodes ensemblistes étant le résultat d'une agrégation de règles faibles, l'interprétabilité et l'explication de ces modèles est altéré. L'importance des variables est l'une des méthodes qui permet d'expliquer l'apport des variables. Elle est systématiquement implémentée dans ces algorithmes. Loin de développer la théorie de calcul de l'importance des variables, nous nous permettons de donner l'intuition derrière cette mesure.

Nous pouvons définir de manière intuitive l'importance d'une variable comme le poids de cette dernière dans la capacité prédictive du modèle. Autrement, une variable a beaucoup d'importance lorsque absente dans la prédiction dans la collection des variables explicatives de la variables cible, augmente considérablement l'erreur de prédiction (OOB) et ce, de manière proportionnelle à son importance. Nous sommes conscients que cette définition est critiquable. C'est Leo Breiman [8] qui va introduire les premières méthodes de calcul en 2001.

4.6.4 Les résultats et prédiction

Le modèle à l'entrainement

Nous entraînons notre modèle sur les données d'entrainement de l'échantillon 2. Nous avons gardé 80% de l'échantillon d'entrainement pour entraîner le modèle et 20% pour sa validation. Les résultats obtenus sont sur la figure ci-dessous. Les ordres de grandeurs des données de la matrice de confusion sont anonymisés.

Entrainement du modèle

	Résiliation	Prédiction 0	1	Class error
Observation	0	159709	1979	1,22%
	1	24588	4621	84,18%

Erreur Out-of-Bag
13,92%

validation du modèle

	Résiliation	Prédiction 0	1	Class error
Observation	0	31961	369	1,14%
	1	4769	1081	81,52%

Erreur Out-of-Bag
13,46%

FIGURE 4.12 – Les résultats du modèle - Random Forest

Nous avons un taux d'erreur (taux de mal classé ou encore Erreur Out Of Bag) de 13,92% pour l'entrainement et 13,46% pour la validation. Ce qui nous donne la certitude qu'il n'y a pas eu de surapprentissage. De plus, ces taux d'erreur sont assez faibles pour conclure que nous avons obtenu un bon prédicteur. Mais, en observant la décomposition des résultats par classe, les conclusions semblent différentes. Une analyse approfondie de la matrice de confusion nous montre que notre modèle arrive à bien reconnaître les contrats non résiliés (1,22% d'erreur) mais peine à bien classer les contrats résiliés (seulement 15,82% de paires concordantes).

Pour confirmer ces informations, nous avons calculé les indicateurs d'évaluation de la performance d'un classifieur.

Choix des indicateurs de performance

Puisque nous sommes dans le cadre d'un problème avec des données déséquilibrées, nous nous baserons essentiellement sur 3 mesures de performances pour juger nos prédicteurs. Ce sont notamment le rappel, la précision et l'indicateur AUC. En effet, notre objectif est de connaître la capacité de notre modèle à identifier les contrats résiliés. De ce fait, le rappel semble être le meilleur indicateur. De plus, la précision s'impose comme indicateur car elle permet de voir la valeur prédictive des contrats résiliés. Enfin, l'AUC nous permettra de juger du pouvoir discriminant des prédicteurs. Pour rappel, l'AUC mesure, dans notre cas, la probabilité que le modèle place un contrat comme étant résilié devant un contrat non résilié.

Ce sont ces indicateurs que nous utiliserons tout au long du reste du mémoire.

Le pouvoir prédictif du modèle

Nous avons testé le modèle sur notre échantillon test. Les résultats obtenus confirment les conclusions de la discussion du modèle ci-dessus. La proportion des observations qui sont bien classées par notre modèle est de 85,62%. Cela est à prendre avec précaution. En effet, ce chiffre est gonflé avec le taux de bonne prédiction des contrats non résiliés. Nous avons un rappel de 10%. En d'autres termes, seulement 10% de nos contrats résiliés sont bien identifiés par le modèle. Toutefois, le modèle est bien précis car parmi la population prédite comme étant des contrats résiliés, 77,33% sont effectivement des contrats résiliés (confère tableau 4.1).

Modèle	*Accuracy*	*Rappel*	*Précision*	AUC
Random Forest	85,62%	0,10	0,7733	0,66

TABLE 4.1 – Quelques indicateurs du Random Forest

Pour avoir le pouvoir prédictif du modèle, nous avons construit la courbe ROC et avons regardé aussi la courbe de Rappel - Précision. Comme nous pouvons le constater sur les figures 4.13 et 4.14, la courbe ROC n'est pas très concave avec un AUC de 0.66. Ce qui, d'après notre grille de qualification à la page 59, signifie que notre prédicteur discrimine passablement les contrats. La courbe ROC quant à elle n'est pas non plus assez proche des bords nord et est du cadran Rappel-Précision.

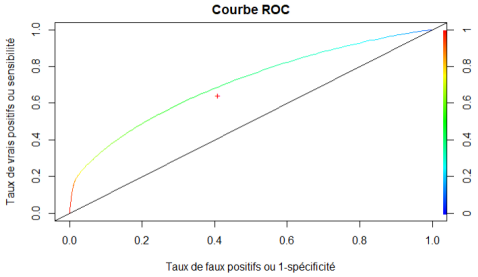


FIGURE 4.13 – La courbe ROC du Random Forest

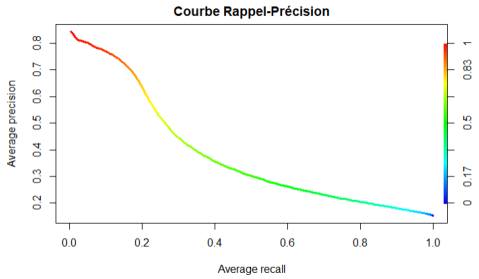


FIGURE 4.14 – Courbe de Rappel- Précision du Random Forest

Nous avons obtenu de meilleurs résultats avec l'algorithme du Random Forest par rapport à la régression logistique. Ces performances sont dues au fait que le Random Forest permet d'avoir une plus faible variance du prédicteur.

Pour améliorer davantage les résultats, nous avons opté pour un modèle ensembliste qui agit à la fois sur la variance et le biais du prédicteur. C'est l'algorithme du XGBoost.

4.7 Extreme Gradient Boosting : XGBoost

XGBOOST (*e**X**treme **G**radient **BOOST**ing*) est la seconde méthode ensembliste que nous mettons en oeuvre dans ce mémoire. Il est très prisé par les data scientistes et les statisticiens car il offre de meilleures performances et dispose de plusieurs paramètres permettant de mieux optimiser son algorithme. Tianqi Chen et Carlos Guestrin rapportaient dans le papier [17] que XGBoost est le plus utilisé dans les compétitions sur Kaggle en 2015. Ils ont donné l'exemple de la compétition KKDCup 2015 où les 10 premiers gagnants ont tous utilisés XGBoost. Il est basé sur l'approche initiale de Gradient Boosting Machine - GBM - développé par Friedman (1999) dans son oeuvre *« Greedy Function Approximation : A Gradient Boosting Machine »* Jérome H. Friedman [7]. Son implémentation a été effectuée par Tianqi Chen et Carlos Guestrin (2016) [17] avant de voir la participation de plusieurs développeurs à travers le monde.

4.7.1 Les motivations du choix du XGBoost pour notre problème

Bien que très complexe, plusieurs raisons nous ont poussés à faire le choix de l'utilisation de l'algorithme XGBoost dans notre travail.

Premièrement, il est prouvé qu'il donne de meilleures performances car agit à la fois sur la variance et le biais du prédicteur. Cela est certainement dû à la conception mathématique très poussée de cet algorithme.

Deuxièmement, nous travaillons avec des données déséquilibrées et l'algorithme du XGBoost dispose des paramètres de régularisation qui permettent de pénaliser les différentes classes de nos données.

Une autre motivation qui justifie notre choix se trouve dans les liens existant entre les différentes variables. En effet, bien que nous avons légèrement diminué le nombre de variables au préalable avec l'étude des corrélations, il n'en est pas moins vrai que nous disposons d'un nombre élevé de variables et des dépendances significatives existent entre certaines parmi elles. Pour rappel, nous avions gardé certaines variables même si la corrélation entre ces dernières s'avère importante. Ainsi, les résultats issus d'un modèle du type régression logistique avec ces variables seraient biaisés. XGBoost permet de s'affranchir de ces contraintes.

Outre les motivations susmentionnés, XGBoost dispose aussi d'autres avantages :

— **La parallélisation des calculs** : L'algorithme du XGBoost permet de faire un traitement parallèle des différentes tâches. Par conséquent, XGBoost est très rapide comparativement à l'algorithme du Random Forest par exemple ;
— **La flexibilité** : Les utilisateurs du XGBoost ont la possibilité de définir leurs objectifs d'optimisation et des critères d'optimisations personnalisés. Par ailleurs, au vu de ses multiples paramètres disponibles, les utilisateurs ont une grande flexibilité pour agir sur l'un ou l'autre en fonction de la structure de leurs données ;
— **La prise en charge des valeurs manquantes** : L'algorithme permet de gérer de manière quasi-autonome les valeurs manquantes. En effet, il essaie différents moyens lorsqu'il rencontre une valeur manquante et s'améliore au fur et à mesure qu'il construit les arbres. Ainsi, il sait quel noeud prendre lorsqu'il rencontre une valeur non renseignée ;
— **Validation croisée** : L'algorithme dans sa conception, permet une validation croisée à chaque itération du processus.

Les lignes précédentes ont relaté toute la richesse de l'algorithme XGBoost. Dans la suite de cette section, nous allons dans un premier temps décliner le principe de fonctionnement de l'algorithme, sa fonction objective et faire un zoom sur certains de ses paramètres qui sont intéressants dans notre étude avant de passer à l'application.

4.7.2 Le principe de l'algorithme XGBoost

XGBoost repose sur l'agrégation d'une famille de règles faibles (arbres de décisions CART de faible profondeur en général) - les *weak predictors*. A la différence du Random Forest qui construit les arbres de manière complètement indépendante, l'algorithme du XGBoost est conçu de manière séquentielle. En effet, après chaque estimation, l'algorithme s'améliore à l'itération suivante de sorte à améliorer l'erreur du prédicteur précédent : c'est le principe du boosting. Une métaphore dans le livre [4] résume bien ce principe : *« le boosting, c'est un peu l'expérience de la vie... Premièrement, on fait des erreurs puis on les corrige. La personnalité est souvent forgée par les erreurs commises, que l'on s'efforce*

de corriger lorsqu'on est confronté à des situations similaires. Toutes nos erreurs vont ainsi avoir des poids plus ou moins forts dans leur contribution à notre personnalité finale. Deuxièmement, notre vie n'est pas gérée par une seule règle universelle qui couvre toutes les situations, mais par de multiples petites règles simples, dont l'assemblage se révèle très puissant ».

Ainsi, à chaque itération, l'algorithme cherche à améliorer sa prédiction des observations mal classées. Pour intégrer cela, chaque observation de la base d'entrainement a un poids qui est égale à $1/n$ à la première itération. Les observations mal prédites à l'itération m ont des poids supérieurs à l'itération $m+1$.

La probabilité de résiliation finale est une moyenne pondérée des probabilités des prédicteurs des itérations.

4.7.3 La fonction objectif

La fonction objectif de l'algorithme XGBoost est très complexe en ce sens où elle est la combinaison d'une fonction de perte et d'un terme de régularisation qui permet une meilleure prédiction surtout dans le cas des données déséquilibrées. Plus formellement, notons M le nombre d'itérations, l la fonction de perte convexe qui mesure la différence entre la valeur observée y et celle prédite $\hat{y}$ et n le nombre d'observations. La fonction objectif que le modèle cherche à minimiser s'écrit :

$$\mathcal{L}() = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{m=1}^{M} \Omega(\delta_m) \quad = \text{Terme d'erreur} + \textbf{Terme de régularisation}$$

avec

$$\Omega(\delta) = \alpha|\delta| + \frac{1}{2}\lambda \|w\|^2,$$

où $|\delta|$ est le nombre de feuilles dans l'arbre CART δ et w le vecteur des valeurs attribuées à chacune de ses feuilles.

Les coefficients α et λ sont interprétés respectivement comme le coefficient de pénalisation du type Lasso (norme L1) et le coefficient de régularisation de Ridge (norme L2). Ainsi, l'interprétation du terme de régularisation devient triviale. Le premier terme permet de réduire la complexité du modèle et le second terme permet de lisser les poids des feuilles et ainsi éviter le surapprentissage.

4.7.4 Les paramètres du modèle

Pour avoir un bon modèle, il est important de bien choisir les paramètres. XGBoost dispose de plusieurs paramètres. D'après la documentation sur les paramètres, les plus importants pour notre problème sont les suivants :

— Les paramètres α et λ qui sont des paramètres de régularisation. Ils permettent de réduire la complexité du modèle et de minimiser les risques de surapprentissage ;
— *nrounds* qui fixe le nombre d'itérations de l'algorithme ;
— *scale_pos_weight* qui permet de tenir compte du déséquilibre entre les classes de la variable à prédire ;
— *max_depth* qui fixe la profondeur des arbres ;
— *eval_metric* qui permet à l'utilisateur de fixer ses métriques d'évaluation du modèle. Il y a plusieurs métriques disponibles parmi lesquelles nous pouvons citer l'erreur de prédiction et l'AUC. Nous utiliserons ces deux indicateurs pour évaluer notre modèle.

La grande difficulté liée à l'utilisation du XGBoost est le choix des bons paramètres afin d'obtenir un modèle performant et éviter le surapprentissage.

Pour le choix des paramètres, nous avons opté pour un entrainement du modèle avec à la fois les données d'entrainement et des données de validations. Cela permet d'observer simultanément l'évolution des métriques d'évaluation et de fixer le nombre d'itérations optimal pour éviter le surapprentissage. Des données d'entrainement issues de l'échantillonnage 2, nous avons pris 20% pour constituer notre échantillon de validation.

Il faut rappeler aussi que l'algorithme du XGBoost ne fonctionne qu'avec des données du type numérique. De ce fait, nous avons converti toutes nos variables qualitatives du type chaîne de caractère en facteurs et les avons ensuite converties en valeurs numériques allant de 1 à autant de modalités possible de la variable considérée.

Les valeurs des paramètres

Nous avons mis en oeuvre 2 modèles XGBoost, le premier étant un modèle basique avec les paramètres par défaut et le second est le résultat d'un processus d'optimisation des paramètres susmentionnés.

Le premier modèle - XGBoost simple : Avec cette lanière d'entraîner le modèle, nous avons obtenu $nrounds = 202$ car, comme nous pouvons le remarquer sur les figures 4.15 et 4.16, le taux de bon classement et l'AUC des données de validations n'augmentent plus à partir de la $202^{\text{ième}}$ itération. En effet, si l'on dépasse cette itération, le taux de bonne prédiction (Accuracy) de la base d'entrainement continue d'augmenter alors que celui de la base validation n'augmente plus : c'est le surapprentissage.

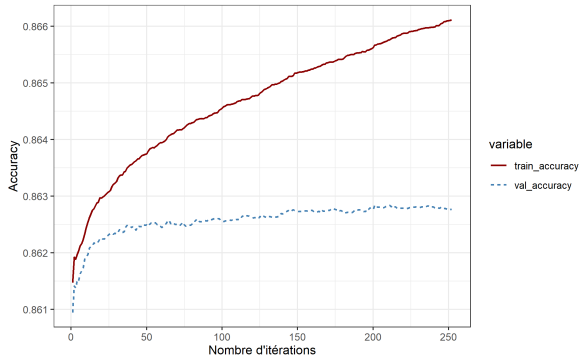


FIGURE 4.15 – Le taux de paires concordantes en fonction du nombre d'itérations

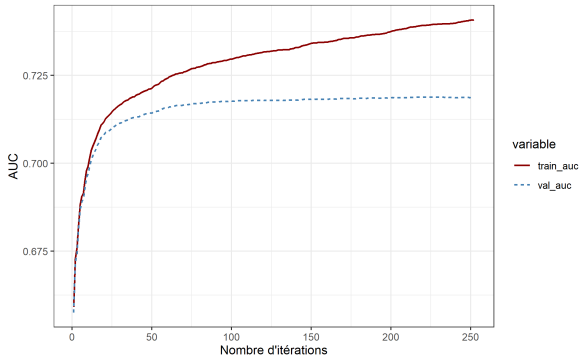


FIGURE 4.16 – AUC en fonction du nombre d'itérations

Le second modèle - XGBoost tuné : Le second modèle mis en oeuvre cherche à optimiser les paramètres de régularisation α et λ ainsi que le paramètre qui permet de tenir compte du déséquilibre entre les classes : *scale_pos_weight*. Pour trouver les bons paramètres, nous avons réalisé la manoeuvre suivante :

- Réalisation de 200 modèles XGBoost sur les mêmes échantillons d'entrainement et de validation ; avec des paramètres choisis de manière aléatoire :
  - les paramètres α et β sont les réalisations d'une loi uniforme entre 0 et 1 ;
  - *max_depth* est un choix aléatoire d'un nombre entre l'intervalle $[\![2\ ;\ 10]\!]$;
  - *scale_pos_weight* est aussi choisi aléatoirement dans l'intervalle $[\![1\ ;\ 7]\!]$. Le choix de la borne supérieure est basée sur la documentation du package xgboost qui conseille d'utiliser la valeur résultante du rapport de la classe négative (contrats non résiliés) et de la classe positive (contrats résiliés). Dans notre cas, cette valeur est légèrement supérieure à 6.
- Pour chaque modèle, nous enregistrons les métriques d'évaluation (AUC et le taux d'erreur) et nous constituons une base de données qui reprend ces métriques et les valeurs des paramètres afférents ;
- Enfin, nous avons trié la base de données selon la meilleure valeur de l'AUC ;
- les valeurs des paramètres correspondants au meilleur AUC sont les paramètres du modèle XGBoost tuné.

Les valeurs des paramètres à l'issue de manoeuvre décrite ci-dessus sont consignées dans le tableau ci-dessous. L'obtention du nombre d'itérations optimal se fait avec la même technique que celle utilisée sur la méthode 1.

λ	α	scale_pos_weight	nrounds	max_deppth
0,300	0,1269	1	121	6

TABLE 4.2 – Les paramètres du modèle 2 - XGbbost tuné

4.7.5 Les résultats des modèles

Les résultats obtenus avec XGBoost semblent meilleurs comparativement au Random Forest. En effet, nous obtenons un AUC de l'ordre de 0.71 avec une précision de 0.70. Les deux modèles arrivent à détecter environ 18% des contrats résiliés.

Modèles	*Accuracy*	*Rappel*	*Précision*	*AUC*
XGBoost Simple	86,22%	0.18209	0,69941	0.7169621
XGBoost tuné	86,23%	0,18195	0,70139	0,7164821

TABLE 4.3 – Quelques indicateurs du Random Forest

Le modèle tuné a une meilleure précision par rapport au modèle avec les paramètres de base. Toutefois, les différences ne sont pas très significatives. Peut-être devrions-nous augmenter le nombre de modèles pour obtenir des paramètres qui donnent de meilleurs résultats ?

D'après les lignes précédentes, nous avons mis en oeuvre deux méthodes d'apprentissage supervisé (Random Forest et XGBoost). Il en découle que XGBoost donne de meilleures performances. Dans la section suivante, nous allons tenter d'améliorer les résultats de ces deux méthodes avec des techniques de ré-échantillonnage. Nous allons procéder de manière méthodique : dans un premier temps, nous essayerons d'améliorer les résultats du Random Forest ; si les résultats obtenus ne dépassent pas ceux du XGBoost, alors nous pourrons par transitivité dire que le meilleur modèle est celui basé sur l'algorithme du XGBoost.

4.8 Ré-échantillonnage et amélioration des résultats

Comme nous l'avons rappelé à chaque commentaire des résultats obtenus, nous avons des données déséquilibrées et cela influence les performances de nos modèles. Pour éventuellement améliorer les performances des différents modèles, nous avons décidé d'agir sur les données à l'aide des techniques de ré-échantillonnage. Cette façon de faire est qualifiée d'approche externe dans le sens où cela n'agit pas directement sur les classifieurs. Par conséquent, il est beaucoup plus souple et flexible. Il existe couramment 3 méthodes de ré-échantillonnage : l'undersampling, l'oversampling et le smote qui est une combinaison des deux premières.

4.8.1 Les techniques de ré-échantillonnage

La technique d'oversampling consiste à créer de nouvelles observations de la classe minoritaire (ici les contrats résiliés) pour rééquilibrer les observations des deux classes. L'inconvénient de cette méthode est qu'il y a un risque de "sur-information" dans la base

d'apprentissage. Par ailleurs, cela pourrait allonger davantage le temps de calcul des modèles. Nous avons donc écarté cette première méthode.

La technique d'undersampling consiste à supprimer des observations de la classe majoritaire (ici les contrats non résiliés). En se faisant, on diminue le nombre d'observations dans la base d'apprentissage par rapport à celle initiale. L'inconvénient de cette méthode est qu'elle génère une perte d'informations.

La dernière technique que nous allons présenter est celle dit SMOTE (Synthetic Minority Oversampling Technique) qui est une combinaison de l'undersampling et de l'oversampling. Contrairement à l'oversampling qui crée de nouvelles observations de la classe minoritaire avec un tirage aléatoire des observations existantes, le SMOTE crée de **manière artificielle** de nouvelles observations de la classe minoritaire. Leur construction est basée sur l'algorithme des *k- plus proches voisins*.

Nous avons testé les 3 types de ré-échantillonnage et c'est l'undersampling 50% qui donne les meilleures performances.

4.8.2 Undersampling 50% et le Random Forest

Nous avons un nombre d'observations très élevé dans la base d'entrainement. De ce fait, créer de nouvelles observations (oversampling) pourrait peut-être ne rien ajouter globalement aux informations disponibles, sinon un ajout de bruits qui pourrait fausser les prédictions. La technique SMOTE aussi a donné de moins bons résultats car, la création d'observations artificielles pourrait introduire des informations non conformes aux comportements réels des contrats. Ce sont probablement les raisons qui font que l'undersampling est la méthode qui convient le plus à nos données.

Le modèle à l'entrainement

Après l'entrainement du modèle, nous avons obtenu de meilleurs résultats. Il y a une diminution du taux d'erreur qui passe de 13,92% à 7,66%. L'erreur de classe des contrats résiliés a aussi considérablement diminué, passant de 84,18% à 10%. A l'inverse, l'erreur de la classe des contrats non résiliés a augmenté ; ce qui est normal.

Le pouvoir prédictif du modèle

Pour analyser le pouvoir prédictif du modèle, nous avons testé ce dernier sur notre échantillon test. Les résultats obtenus sont meilleurs que ceux obtenus sur les données déséquilibrées.

	Accuracy	*Rappel*	*Précision*	*AUC*
RF-Undersampling	70,24%	0,56	0,27	0,705

TABLE 4.4 – Indicateurs du RF Undersampling 50%

les résultats du tableau 4.4 montre une diminution du taux de bon classement qui passe de 85,62% à 70,24% par rapport au modèle sur l'échantillon brut. Néanmoins, nous gagnons en rappel qui est très important dans notre analyse. En effet, notre modèle arrive à détecter les résiliations (56%). Néanmoins, il est moins précis car parmi l'ensemble des contrats prédits comme étant résiliés, seuls 27% sont des contrats réellement résiliés. Nous pouvons constater cela sur la courbe ROC et Rappel-Précision ci-dessous.

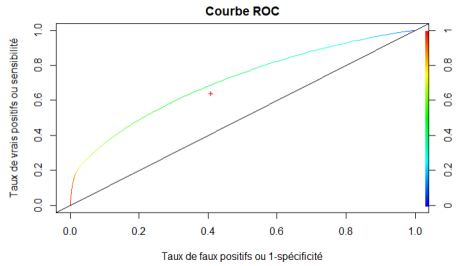


FIGURE 4.17 – La courbe ROC du Random Forest - Undersampling

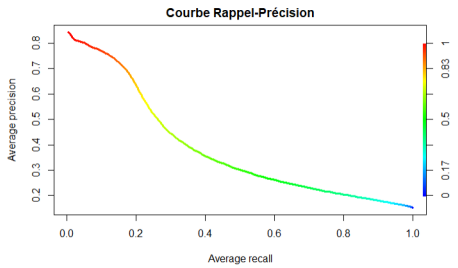


FIGURE 4.18 – La courbe de Rappel Précision - Random Forest - Undersampling

La valeur de l'AUC est de 0,705, ce qui fait que nous pouvons classer ce modèle comme étant un classifieur acceptable.Toutefois, ce modèle a un pouvoir discriminant inférieur à celui du XGBoost simple,par exemple (AUC=0.7169).

4.9 Comparaison des modèles et choix du meilleur modèle

Plusieurs modèles sont mis en oeuvre depuis le chapitre 2. Ces différents modèles ont des performances différentes et il est nécessaire de les comparer et donc , faire un choix du meilleur modèle selon notre problématique. Le tableau 4.9 résume les indicateurs d'évaluation des différents prédicteurs.

Modèles	*Accuracy*	*Rappel*	*Précision*	*AUC*
Random Forest	85,62%	0,10	0,77	0,66
RF- Undersampling	70,24%	0,56	0,27	0,705
XGBoost Simple	86,22%	0,183	0.699	0,7169
XGBoost tuné	86,23%	0,182	0,701	0.7172
Modèle 1 - LR	84,46%	0,0016	0,252	0.596
Modèle 2 - LR	84,62 %	0,0004	0,337	0.615

TABLE 4.5 – Comparaison des modèles

Pour comparer les prédicteurs, nous allons regarder le rappel, la précision et l'AUC. Notre objectif est de détecter à temps les contrats qui sont susceptibles d'être résiliés. Une fois ces contrats détectés, la direction compétente peut mieux les cibler et mener des actions pour les retenir selon le profil des clients. La comparaison de ces 3 indicateurs nous emmène à considérer 2 modèles (RF-Undersampling et XGBoost tuné) selon les cas :

— Le modèle RF-Undersampling a un $AUC = 0.705$ et semble classer des contrats comme étant résiliés alors qu'ils ne le sont pas car il a une faible précision (0.27) et un rappel élevé (0.56). En d'autres termes, parmi les contrats qui sont prédits comme étant résiliés, seuls 27% le sont réellement. De ce fait, si le coût de rétention d'un client en portefeuille est plus faible que de laisser le client partir (résilier son contrat), il vaut mieux utiliser ce modèle pour prédire la propension à résilier des différents clients. Nous rappelons que le coût de rétention d'un client peut être mesuré par exemple avec le montant de ses gestes commerciaux, de ses avantages ou encore des dérogations tarifaires dont il bénéficie ;

— Le modèle XGBoost tuné a le pouvoir discriminant le plus élevé ($AUC = 0.7172$). Toutefois, il a un rappel faible mais avec une grande précision (0.701). Ainsi, si le coût de rétention est très élevé, il vaut mieux opter pour un modèle très précis.

Nous postulons que le coût de rétention des clients est élevé et par conséquent, le meilleur modèle pour prédire les résiliations doit être un modèle qui soit le plus précis possible. Ainsi, dès qu'un contrat est identifié comme étant susceptible d'être résilié, qu'il y ait une forte probabilité qu'il le soit réellement. Par conséquent, le modèle qu'il convient de choisir ici est l'XGBoost avec une optimisation des hyperparamètres : "XGBoost tuné". Par ailleurs, ce postulat est soutenable. En effet, d'après les résultats de la régression logistique, nous avons constaté que le profil des clients qui résilient le plus ont un CRM élevé et une forte sinistralité. Les contrats avec ces caractéristiques sont des contrats qui coûtent chers à l'assureur.

4.10 Conclusion

Durant ce chapitre, nous avons mis en oeuvre différents modèles pour prédire la probabilité de résiliation des contrats. Bien que l'algorithme du Random Forest donne des résultats satisfaisants, XGBoost semble être le meilleur prédicteur qui discrimine le mieux les contrats.

Bien qu'il soit possible de retracer la règle de décision pour un arbre, la phase d'agrégation des deux méthodes mises en oeuvre font qu'on perd une part d'interprétabilité des résultats. Toutefois, de nouvelles techniques comme la *SHAP Value* permettent de voir au-delà de l'importance des variables explicatives. Nous insisterons sur l'explicabilité des résultats dans le chapitre suivant.

Outre la notion d'interprétabilité des résultats, nous allons dans la suite de ce mémoire mettre en oeuvre une méthode innovante et originale pour améliorer les résultats du XGBoost et faire une discussion sur le seuil de séparation des contrats.

Détection des contrats résiliés avec XGBoost à l'aide d'un scoring de sinistralité avec BiRank

D'après les chapitres précédents, la sinistralité joue un rôle important dans l'explication de la résiliation. De plus, le modèle XGBoost donne le meilleur prédicteur des contrats résiliés. Fort de ces conclusions, nous avons décidé d'exploiter au mieux les données de sinistralité afin d'améliorer les indicateurs de performances de la méthode XGBoost. Pour ce faire, nous allons construire de nouvelles variables à partir de ces données puis les introduire dans l'algorithme de XGBoost.

La construction de nouvelles variables, communément appelée *feature engineering* en anglais, de ce mémoire est basée sur la méthodologie utilisée dans l'article *Social network analytics for supervised fraud detection in insurance* de M. Óskarsdóttir et al. [11]. Les auteurs ont utilisé l'algorithme BiRank de He et al. [18], qui est lui-même une version personnalisée de l'algorithme PageRank de Google, pour détecter les fraudes dans les sinistres qui surviennent en assurance non-vie. Ils ont dans un premier temps construit un réseau biparti entre les sinistres et les différentes parties qui interviennent dans la gestion du sinistre telles que les experts, les garages (par exemple le garage dans lequel le véhicule a été réparé si réparation a lieu). Ensuite, ils ont construit un scoring avec l'algorithme BiRank en se basant sur le réseau biparti précédent et en extraire des variables. Ils ont utilisé ces variables pour améliorer la détection des fraudes qui est aussi un sujet avec des données très déséquilibrées.

En s'inspirant de ce travail, notre analyse dans ce chapitre se déroulera en trois étapes. Dans un premier temps, nous allons construire un graphe biparti entre les contrats sinistrés et les différents types de sinistres qu'ils ont subi. Ensuite, nous utiliserons ce réseau biparti pour construire un scoring de sinistralité de tous les contrats. Enfin, nous utiliserons ce scoring comme une nouvelle variable que nous introduirons dans notre modèle de XGBoost.

5.1 Réseau de graphe biparti entre les contrats et les types de sinistre

5.1.1 Formalisation des données

Les données qui interviennent dans cette analyse de réseau biparti sont des données de sinistralité. Nous avons scindé notre base de données de travail en deux groupes : les contrats sinistrés et les contrats non sinistrés. Ensuite, pour chaque contrat sinistré, nous récupérons les garanties sinistrées et leurs charges. Ce sont ces garanties que nous avons dénommées les types de sinistre.

Nous avons au total 5 types de sinistres pour chaque contrat sinistré : Responsabilité Civile (RC), Dommages (DOM), Vol (VOL), Bris de Glaces (BDG) et les autres garanties (AUG). Les données se présentent comme sur le tableau suivant :

RC	*DOM*	*VOL*	*BDG*	*AUG*
0	4580	0	0	0
...	...	...	...	...
455500	2500	0	0	0
0	2021	0	258	0

TABLE 5.1 – Formalisation des données pour le réseau Bipartite

Chaque ligne d'observation correspond à un contrat et les colonnes représentent les charges des garanties sinistrées.

Dans ce chapitre, la construction du scoring avec l'algorithme BiRank se fera uniquement sur les contrats sinistrés. Une analyse de ce scoring permettra d'allouer un score aux contrats non sinistrés.

5.1.2 Définition et notations

Les graphes sont un langage universel qui permet de représenter des relations entre des entités. Ces entités peuvent être de même nature ou de nature différente. Nous pouvons donner l'exemple de la relation entre les courtiers et l'ensemble des assureurs.

On désigne par graphe biparti, un graphe qui permet de modéliser la relation entre deux entités de nature différente. Dans un tel graphe, il ne peut y avoir de lien qu'entre deux entités de types différents. Dans notre étude, ces deux entités sont les contrats sinistrés et les types de sinistres.

Sur la figure 5.1, nous avons un exemple de graphe biparti construit avec notre structure de données avec le package igraph. Les cercles représentent les sommets, les traits reliant ces cercles représentent les arêtes. De manière générale, on peut avoir un poids sur

chaque arête. Ce poids mesure l'intensité de la liaison entre les 2 entités. On parle alors de graphe pondéré ou valué.

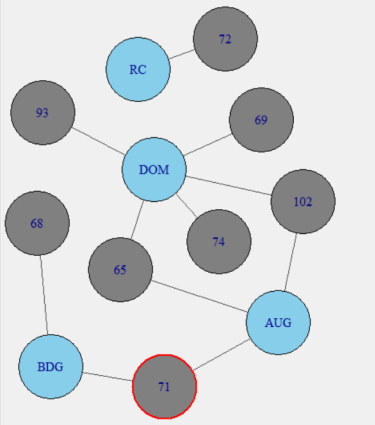


FIGURE 5.1 – Un exemple de réseau de graphe biparti entre les contrats et les types de sinistres

Les cercles bleus représentent les types de sinistres et les cercles gris les contrats. Le contrat 71 est lié aux sinistres du types BDG et AUG. Les sinistres aussi sont liés aux contrats. Par exemple, le sinistre DOM est lié aux contrats 65, 69, 74, 93 et 102. Comme dans la définition précédente, nous pouvons observer que le graphe est bien biparti dans le sens où il n'y a que des connexions entre un sinistre et contrat.

Notations

Les différentes notations mathématiques qui interviennent dans la formalisation du graphe et de l'algorithme BiRank sont les suivantes :

Formalisation du graphe : Désignons par $G = (C \cup S, A)$ un graphe biparti de noeuds $C \cup S$ et d'arête A. Les noeuds C désignent les contrats et les noeuds S, les types de sinistres. Chaque arête A lie un sommet du type C à un sommet du type S. Les sommets S peuvent être AUG, BDG, DOM, RC ou VOL et les sommets en C sont les identifiants (les nombres).

Soient n_C le nombre de sommets distincts en C avec c_i un sommet individuel en C, $i \in \{1, ..., n_C\}$; et $n_S = 5$ le nombre de sommets en S et les types de sinistres sont notés s_j, $j \in \{1, ..., n_S\}$.

La matrice des poids : Les arêtes sont représentées par la matrice des poids $\mathbf{W} = (w_{i,j})$, avec $i \in \{1, ..., n_C\}$, $j \in \{1, ..., n_S\}$. Par conséquent, $\mathbf{W}$ a n_C lignes et n_S colonnes représentants respectivement les sommets en C et en S. Le graphe mis en oeuvre dans ce chapitre est pondéré. En effet, si le noeud c_i est connecté ou lié au noeud s_j, alors $w_{i,j}$ est égale à la charge du type de sinistre. Dans le cas contraire, $w_{i,j} = 0$. Le graphe étant non

orienté, alors $w_{i,j} = w_{j,i}; \forall i, j \in [\![1\ ;\ n_C]\!] \times [\![1\ ;\ n_S]\!]$. La matrice $\mathbf{W}$ se présente comme dans le tableau 5.1.

Définition du degré pondéré des sommets : Pour un noeud c_i, on note d_i son degré. Pour rappel, le degré d'un noeud dans un graphe non pondéré est le nombre de sommets qui lui sont connectés directement. Dans notre cas, le degré est tout simplement la somme des poids des arêtes issus du sommet c_i, soit $d_i = \sum_{j=1}^{n_S} w_{i,j}$. Autrement, d_i est la charge de tous les sinistres du contrat c_i. Ces degrés sont résumés dans une matrice diagonale à n_C éléments $\mathbf{D_C}$ avec $(\mathbf{D_C})_{i,i} = d_i$.

De la même manière, on définit le degré pondéré du type de sinistre s_i par $d_j = \sum_{i=1}^{n_C} w_{i,j}$ et la matrice diagonale à n_S éléments D_S avec $(\mathbf{D_S})_{j,j} = d_j$. De manière concrète, d_j représente la charge globale des sinistres du type s_j de tous les contrats sinistrés.

En résumé, le graphe biparti construit est non orienté et pondéré. Le poids de l'arête est la charge de sinistre.

Le graphe construit a une densité de $6.882318e-06$. La densité d'un graphe mesure la proportion d'arêtes dans le graphe par rapport au nombre d'arêtes possibles. On dira alors que le graphe est dense lorsque cette valeur est élevée. Dans le cas contraire, le graphe est dit creux. Bien que la densité de notre graphe soit faible, il est difficile de conclure qu'il s'agit d'un graphe creux.

5.2 Construction d'un scoring de sinistralité avec l'algorithme BiRank

Après cette première partie sur la construction du graphe biparti valué, nous allons construire un scoring de sinistralité avec le graphe précédent. Le scoring est une technique qui consiste à affecter une valeur appelée score à chaque contrat dans notre cas. C'est une technique très utilisée notamment dans le marketing ou encore par les géants Facebook, Google et Twitter. Par exemple, l'algorithme de recommandation des chansons sur YouTube est basé sur des techniques de scoring.

5.2.1 Intuition et objectif

L'intuition derrière la construction d'un scoring de sinistralité est l'hypothèse selon laquelle tous les sinistres n'ont pas le même impact quant à la propension à résilier. Notre graphe étant valué avec les charges de sinistres, il est intuitif de se dire que les sinistres ayant une charge élevée n'auraient pas le même impact que les sinistres générant un coût faible. De plus, la deuxième information importante dans ce graphe est le type de sinistre.

Au delà des coûts, il est intuitif de se dire qu'un sinistre du type BDG n'aurait pas le même impact qu'un sinistre du type DOM. Les variables en nombre de sinistres de ces différents types peuvent donc être des sources de bruits.

Ainsi, le scoring de la sinistralité va permettre de synthétiser l'information sur la sinistralité globale du contrat et plus encore, va constituer une variable supplémentaire dans la collection des variables explicatives.

5.2.2 L'algorithme BiRank

Développé par He et al. [18], l'algorithme BiRank est spécialement conçu pour construire un scoring avec des graphes bipartis. Le score est construit pour chacun des sommets donc ici, à la fois un score pour chaque contrat et un score pour chaque type de sinistre. Pour des raisons de simplicité, nous notons c_i et s_j les scores des sommets c_i et s_j. Ainsi, à la sortie, l'algorithme renvoie deux vecteurs c et s de taille respective n_C et n_S.

.... Le principe de calcul des scores

Le calcul des scores des sommets se fait de manière itérative et simultanée. L'idée de base est la suivante : si un sommet c_i a un score élevé, alors, il doit être connecté à un sommet s_j de score élevé. De là découle la règle de calcul suivante : le score d'un sommet c_i est la somme pondérée des scores des sommets s_j avec lesquels il est connecté. Partant de là, la formule des scores est la suivante :

$$c_i = \sum_{j=1}^{n_S} w_{i,j} s_j \quad et \quad s_j = \sum_{i=1}^{n_C} w_{i,j} c_i$$

Pour assurer la convergence des scores, He et al. [18] ont proposé la normalisation symétrique. En se faisant, ils ont réussi à lisser le poids d'une arête par le degré de ses deux sommets connectés simultanément. Les scores normalisés sont :

$$c_i = \sum_{j=1}^{n_S} \frac{w_{i,j}}{\sqrt{d_i}\sqrt{d_j}} s_j = Ns \quad et \quad s_j = \sum_{i=1}^{n_C} \frac{w_{i,j}}{\sqrt{d_i}\sqrt{d_j}} c_i = N^{\mathrm{t}} c$$

avec $N = D_C^{-\frac{1}{2}} \mathbf{W} D_S^{-\frac{1}{2}}$ et N^{t} la transposée de la matrice N.

Dans sa conception, l'algorithme permet d'intégrer des informations *a priori* sur les vecteurs c et s notés c_0 et s_0. Pour tenir compte de ces informations *a priori*, les formules itératives s'écrivent en allouant un poids à ces informations. Ces poids sont mesurés avec les paramètres α et β et les formules précédentes s'écrivent :

$$c_i = \alpha Ns + (1-\alpha)c_0 \quad et \quad s_j = \beta N^{\mathrm{t}} c + (1-\beta)s_0$$

Dans notre cas, nous avons pris $\alpha = \beta = 1$.

.... L'algorithme BiRank

L'algorithme prend en entrée le graphe biparti $G = (C \cup S, A)$ et sa matrice de poids $\mathbf{W}$. La sortie est une fonction $f : C \cup S \longmapsto \mathbf{R}$ qui alloue un nombre réel à tous les sommets de G. De manière algorithmique, le processus d'itération et la conception de l'algorithme sont les suivants :

Entrées : Le graphe biparti $G = (C \cup S, A)$ avec sa matrice des poids $\mathbf{W}$, les hyperparamètres α et β
Sorties : Les vecteurs de score c et s
Calcul de la matrice de normalisation : $N = D_C^{-\frac{1}{2}} \mathbf{W} D_S^{-\frac{1}{2}}$;
Les vecteurs c et s sont initialisés aléatoirement;
tant que *Condition d'arrêt non vérifiée* **faire**
  $c \leftarrow Ns$;
  $s \leftarrow N^{\mathrm{t}}c$
retourner c *et* s

Algorithme 1 : Algorithme de BiRank avec $\alpha = \beta = 1$

5.2.3 Quelques résultats du scoring

Nous avons appliqué l'algorithme BiRank sur le graphe biparti issu des contrats sinistrés. Les scores obtenus sont de l'ordre de 10^{-4}. Pour avoir des valeurs présentables, nous avons multiplié le score par 10^4. En effet, en se faisant, nous ne modifions en rien les informations apportées par le scoring. Un premier aperçu des résultats du tableau 5.2 montre le score qui varie de 0,00001 à 617,435 avec une valeur médiane de 4,939. La moyenne des scores est de 5,337.

Min	Médiane	Moyenne	Max
0,000	4,939	5,337	617,435

TABLE 5.2 – Statistique descriptive

Pour tester la pertinence des résultats, nous avons calculé le score moyen en fonction de la variable cible et les résultats sont consignés dans le tableau 5.3. Dans la population des contrats sinistrés, les contrats résiliés ont un score plus élevé que ceux non résiliés en moyenne. Le test d'égalité de moyenne de Student donne une p-valeur $< 2,2e-16$. De ce fait, les deux moyennes sont donc significativement différente.

Types de contrats	Score moyen
Non résiliés (Résiliation=0)	4,982
Résiliés (Résiliation=1)	7,352

TABLE 5.3 – Score moyen en fonction de la résiliation

Ces résultats sont cohérents car nous savons *a priori* qu'un contrat sinistré a plus de chance d'être résilié. De ce fait, nous nous attendons à avoir un score moyen de sinistralité plus élevé pour les contrats résiliés. Le graphe étant valué, il convient de mentionner que la valeur du score est une fonction croissante en w_i : les charges des sinistres.

5.3 Une nouvelle variable explicative

Notre base de données initiale contient à la fois des contrats sinistrés et des contrats n'ayant eu aucun sinistre. Nous avons calculé dans la section précédente un score de sinistralité uniquement sur les contrats sinistrés. Ainsi, pour avoir une variable qui s'étend sur toute la base de données, il nous faut construire un scoring de "non" sinistralité sur les contrats non sinistrés.

... Quelle score affecté aux contrats non sinistrés ?

Pour distinguer les contrats sinistrés des contrats non sinistrés, nous avons décidé d'affecter une valeur ne se situant pas dans l'intervalle des valeurs prises par le score construit précédemment. D'après le tableau 5.2, le score est inférieur à $617,435$ et non négatif. Par conséquent, nous avons décidé d'allouer le score de 10000 aux contrats non sinistrés.

... Comment justifier cette valeur de 10 000 ?

Comme mentionné en début de ce chapitre, nous nous baserons sur le modèle XGBoost. Or, ce modèle se base sur l'arbre CART. Ainsi, lorsqu'il verra une concentration de la valeur 10 000 dans les données, il comprendra que cela définit une modalité unique. Il pourra alors les traiter comme telle dans la division des noeuds.

En définitive, la nouvelle variable que nous prendrons pour la suite de ce chapitre est la variable *score* qui prend des valeurs décrites dans la section 5.2.3 pour les sinistrés et 10 000 pour les autres.

5.4 Amélioration des résultats de prédiction des résiliations avec XGBoost

Nous avons mis en oeuvre le modèle XGBoost avec en plus des variables explicatives initiales, le score construit dans la section précédente. Pour voir l'effet de la variable score, nous avons comparé deux modèles. Le premier modèle est un modèle dans lequel nous avons considéré à la fois les variables de sinistralité et le score de sinistralité ; le second considère uniquement la variable score avec une suppression des variables de sinistralité

qui ont servi à la construction du score. Nous les nommons respectivement *XGBoost full* et *XGBoost Score*.

L'optimisation des hyperparamètres s'est faite avec le même procédé que dans la section 4.7.4. Les paramètres obtenus sont dans le tableau 5.4.

Modèles	λ	α	scale_pos_weight	nrounds	max_depth
XGBoost full	0,88	0,93	2	464	5
XGBoost Score	0,709	0,78	2	562	5

TABLE 5.4 – Les paramètres du modèle 2 - XGBoost tuné

Les résultats consignés dans le tableau 5.4 montrent une nette amélioration de la prédiction des résiliations. Les deux modèles donnent un taux de détection de $28,7\%$ environ avec une précision de 74%. Les classifieurs obtenus à l'issu de cette étape peuvent être qualifiés de classifieurs acceptables avec un AUC de $0,76$; ce qui représente une amélioration de $6,40\%$ par rapport au modèle sans introduction du score (XGBoost tuné). Toutefois, c'est le modèle *XGBoost full* qui donne la meilleure prédiction d'après les critères de précision et d'AUC.

Modèles	*Accuracy*	*Rappel*	*Précision*	*AUC*
XGBoost full	87,62 %	0,28660	0,74333	0.7631912
XGBoost Score	87,62%	0,28819	0,74119	0,7622784

TABLE 5.5 – Amélioration des résultats du XGBoost

5.5 Interprétabilité du modèle : SHAP Value

Il est souvent reproché aux méthodes de machine learning d'être des boites noires. Cette dénomination vient du fait que l'interprétabilité des résultats est souvent compliquée. Ils reçoivent en entrée des données et produisent à la sortie d'excellents résultats en apprenant des données et de leurs structures. Contrairement aux modèles GLM qui offrent une facilité d'interprétabilité, il est souvent difficile en machine learning de décorreler par exemple la probabilité de résiliation de chaque modalité d'une variable donnée.

Plusieurs méthodes ont été ainsi mises en place par les chercheurs pour palier à la remarque précédente. Parmi ces méthodes, on distingue la SHAP Value qui signifie SHapley Additive exPlanations qui a été introduit par Lunderberg et al. en 2017 [15] dont le fondement théorique est en annexe A.3 (Encadré 2).

SHAP Value permet de mesurer l'impact de chaque variable sur la variable cible. Nous nous permettons de faire une digression métaphorique pour illustrer l'idée du SHAP Va-

lue. L'idée de la SHAP Value est de considérer chaque variable de la base de données comme un joueur d'une équipe de football par exemple (prenons l'exemple de l'équipe de France à la coupe du monde - Russie 2018) ; la base de données étant l'équipe. Chaque joueur contribue au résultat final obtenu par l'équipe. La somme de ces contributions donne la valeur obtenue par la variable cible (score du match de la finale par exemple). Ainsi, la victoire 4-2 obtenue contre la Croatie est la somme des efforts de tous les joueurs (Mbappe, Griezmann, Pogba etc...).

Cette digression illustre bien l'idée de la SHAP value. Par ailleurs, pour chaque variable explicative, SHAP calcule l'impact de chacune des variables explicatives sur la variable cible. Nous pouvons par exemple observer si une variable participe plus à prédire les contrats résiliés ou les contrats non résiliés.

Nous avons réalisé la SHAP Value sur notre modèle *XGBoost full* à l'aide du package SHAPforxgboost.

Les résultats de la figure 5.2 montrent l'importance des variables ; une importance qui est mesurée avec la valeur moyenne des SHAP Values sur les variables. On remarque que la variable *Prime_totale* est la plus influente avec le *nb_contrat* et désigne respectivement la somme des primes des contrats que le client a chez GAN Assurances et le nombre total de contrats qu'il détient.

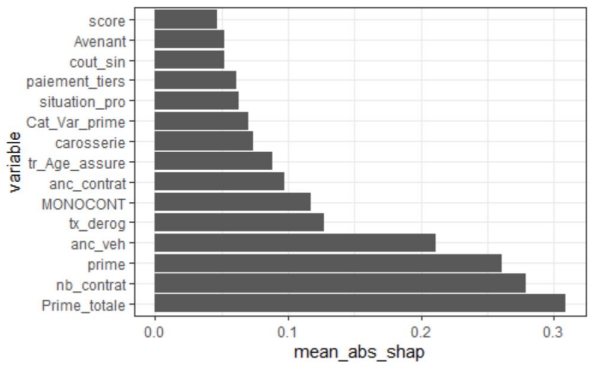


FIGURE 5.2 – Importance des variables à l'aide de SHAP Value- XGBoost full

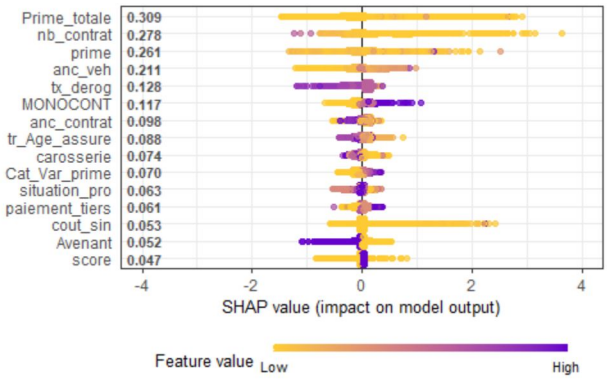


FIGURE 5.3 – SHAP Value XGBoost full

Interprétation de la SHAP Value

La figure 5.3 représente la SHAP Value des 15 premières variables les plus influentes. Chaque point sur une ligne représente une observation de la variable considérée.

Par exemple, la variable *Cat_Var_Prime* qui mesure la variation de la prime en euros d'une version du contrat sur une autre semble très discriminante. En effet, les contrats ayant une variation de prime élevée (en bleu) ont une SHAP Value positive donc ont une

probabilité de résilier élevée. Les variations faibles tendent à générer une probabilité de résiliation faible. Ces résultats confirment l'intuition que nous pouvons avoir et que nous avons d'ailleurs confirmé dans la section de statistique descriptive.

La variable *MONOCONT* qui indique si le client à un seul contrat ou non chez GAN (1 ou 0) est aussi très discriminante. Les clients ayant un seul contrat chez GAN (en bleu) ont une SHAP Value élevée et donc une probabilité de résiliation plus élevée. Les multi-équipés chez GAN résilient moins.

Globalement, plus l'ancienneté du contrat est élevée, moins on résilie son contrat. Toutefois, l'impact de cette variable est moindre.

Enfin, la variable *score*, construite avec BiRank apparaît dans les variables les plus influentes. Cela justifie d'ailleurs l'amélioration significative des résultats obtenus par nos modèles.

Pour regarder un peu plus en détail les résultats, nous avons séparé les données en deux sous groupes : les contrats résiliés uniquement et les contrats sinistrés et avons calculé les contributions des variables explicatives. Les résultats obtenus sont représentés sur les figures 5.4 et 5.5.

En considérant uniquement les contrats résiliés, la variable *score* sort des top 15 des variables les plus influentes au profit du type d'énergie utilisé par le véhicule. Pour les contrats sinistrés, sans surprise, le coût total de sinistre (*cout_sin*) et le score deviennent plus influents comparativement à leurs contributions dans la prédiction avec le modèle basé sur toutes les données.

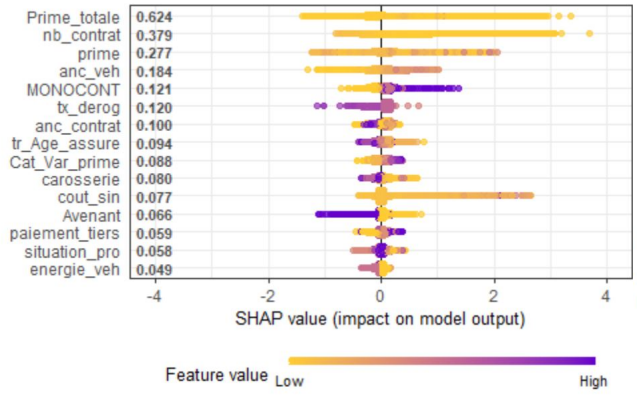


FIGURE 5.4 – SHAP Value pour XGBoost full sur les contrats résiliés uniquement

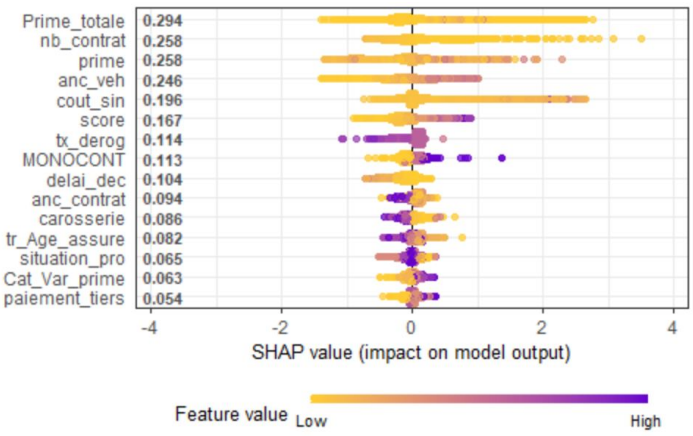


FIGURE 5.5 – SHAP Value XGBoost full sur les contrats sinistrés

5.6 Influence du seuil de séparation des classes sur les performances

Avec les données déséquilibrées, le seuil de probabilité de séparation des classes de défaut ($seuil = 0,5$) n'est pas toujours le meilleur. L'objectif étant de prédire au mieux les contrats résiliés, une discussion sur le seuil s'impose.

En se basant sur le modèle XGBoost full, nous avons fait varier le seuil de $0,2$ à $0,5$ et avons prédit les résiliations. Nous avons ensuite récupéré les indicateurs d'évaluation (précision, rappel).

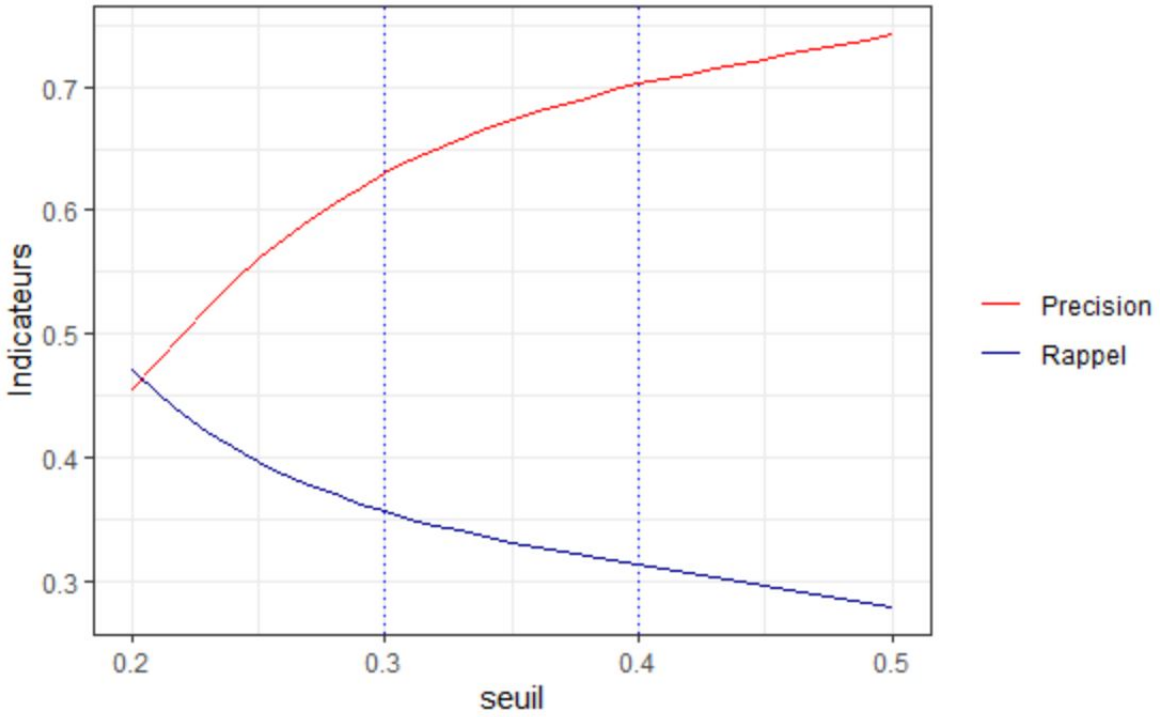


FIGURE 5.6 – Variation de la précision et du rappel en fonction du seuil

L'observation de la figure 5.6 montre qu'une diminution du seuil entraîne une augmentation du rappel et une diminution de la précision. En effet, dès lors que le seuil diminue, un grand nombre de contrats sont prédits comme étant des contrats résiliés. Les contrats se trouvant vers la limite du seuil initial mais qui sont des contrats non résiliés, se retrouvent alors prédits comme étant résiliés par le modèle. Cela introduit une impureté supplémentaire dans le prédicteur et par conséquent, une diminution de la précision.

Or comme nous l'avons mentionné en fin du chapitre précédent, nous privilégions la précision à la capacité du modèle à classer un contrat comme étant résilié. De ce fait, nous ne somme pas prêt à faire trop de compromis en précision. Partant de cet objectif, nous avons décidé de tolérer une imprécision maximale de 30% ; ce qui équivaut à une précision

minimale de 70%. C'est ainsi que nous obtenons un seuil de probabilité 0,4. D'après notre simulation, ce seuil offre un rappel de 0,31 et une précision de 0,7019.

5.7 Opérationnalisation du travail de détection

La détection des contrats à fort risque de résiliation doit être un exercice récurrent que les assureurs doivent faire. Ainsi, nous proposons ici une piste d'automatisation du processus de détection que l'assureur peut exécuter mensuellement par exemple. Notre canevas d'opérationnalisation peut-être effectué sur R à l'aide du module Rshiny.

- Récupération des données brutes sur les contrats en faisant attention aux données personnelles. En effet, R est un logiciel open source et il n'est pas conseillé de laisser traîner des données d'identification des clients sur cela ;
- Mettre en place un interface graphique sur Rshiny qui permet d'importer ces données et qui donne à la sortie une base de données contenant les contrats susceptibles d'être résiliés. Le back-end ou le dorsal de ce formulaire Rshiny peut contenir les étapes suivantes :
  - Une fonction qui fait le traitement des données (regroupement de modalités, création de nouvelles variables) ;
  - une fonction qui construit le graphe biparti tout en formalisant au préalable les données ;
  - une fonction qui construit le score de sinistralité avec BiRank ;
  - une fonction qui permet de faire la prédiction des résiliations et qui stocke dans un dataframe les contrats dont la probabilité dépasse un seuil fixé par l'utilisateur. Dans notre cas, le seuil est de 0,4.
- les contrats récupérés nécessitent un second traitement. En effet, l'on peut utiliser les critères de CRM > 1, antécédent de sinistralité > 2, ancienneté du contrat et l'évolution de la prime de la régression logistique pour subdiviser la base de données. Les contrats qui ont des critères différents de ceux de la régression logistique peuvent être mis en ce que nous avons dénommée "surveillance de résiliation". Ce sont principalement ces contrats qui peuvent subir des gestes de la direction commerciale pour être retenus en portefeuille.

Le pilotage de ce processus peut être assuré par un actuaire qui surveille le produit assurance auto. Des études de détection de seuil de probabilité de séparation peuvent être réalisées. C'est d'ailleurs ce pourquoi l'actuaire devra avoir la main sur le paramètre "seuil" de l'application Rshiny.

Conclusion Générale

Dans ce mémoire, nous avons pour objectif de prédire les résiliations. Nous avons fait une première étude descriptive pour détecter les variables potentielles qui discriminent le mieux la propension à résilier. Nous avons mis l'accent sur l'année 2020 qui est particulièrement marquée par la crise sanitaire liée au covid. Pour avoir des résultats robustes qui reflètent la réalité, nous avons finalement supprimé les données de 2020 pour nous concentrer sur la modélisation de la probabilité de résiliation sur les données allant de 2017 à 2019.

Une première modélisation faite grâce à la régression logistique nous a permis de déceler les profils des contrats les plus résiliés. Quatre variables permettent de définir une première règle :

— Le CRM ;
— L'antécédent de sinistre sur les 24 derniers mois ;
— L'ancienneté du contrat ;
— La variation de prime en % .

Les contrats qu'il faut surveiller sont notamment les contrats ayant un CRM supérieur à 1 avec une ancienneté comprise entre 1 et 3 ans avec au moins 2 sinistres lors des 24 derniers mois pour une revalorisation tarifaire de plus de 3%. Ces contrats sont ceux qui ont le plus de chance d'être résiliés lors de la version en cours du contrat.

Néanmoins, il faut rappeler que la régression logistique ne nous a pas permis de bien prédire les résiliations si bien qu'il explique mieux le phénomène par rapport aux méthodes de machine learning qui ne suppose aucune loi *a priori*. La raison principale de ces résultats est la structure déséquilibrée des données.

Les méthodes ensemblistes pour minimiser la variance du prédicteur

Dans un contexte de données déséquilibrées, les méthodes ensemblistes semblent donner de meilleurs résultats d'après la littérature. C'est ainsi que nous avons mis en oeuvre deux méthodes ensemblistes : le Random Forest et le XGBoost.

Le Random Forest a donné de meilleurs résultats par rapport à la régression logistique. Des techniques de ré-échantillonnage ont été appliquées pour améliorer les performances de ce modèle. Toutefois, c'est finalement le modèle du XGBoost qui prédit mieux les résiliations. La raison principale est qu'en plus d'agir sur la variance comme le fait Random Forest, XGBoost agit sur le biais de prédiction dans son processus de boosting.

Une méthodologie innovante pour améliorer les performances du XGBoost

Convaincu que XGBoost permet d'obtenir le meilleur prédicteur, une méthodologie innovante a été mise en oeuvre pour améliorer ses performances. A partir d'un graphe biparti valué de sinistralité, nous avons construit un score de sinistralité avec l'algorithme BiRank. Ce score constituera une nouvelle variable qui permettra d'améliorer considérablement nos résultats.

Les limites des modèles utilisés

Si les méthodes d'agrégation d'arbres sont très performants et très stables, ils sont souvent critiqués pour leur interprétabilité. En effet, en agrégeant les arbres CART, on perd la traçabilité de la règle de décision finale. Heureusement, de nouvelles techniques ont été développées pour expliquer les méthodes dites de "boites noires" à l'instar de la SHAP Value.

La SHAP Value nous a permis d'interpréter les résultats et d'observer l'influence de l'ensemble des variables sur la probabilité de résiliation. Les résultats observés sont cohérents avec les *a priori* que nous avons avec la statistique descriptive.

Ainsi, à la fin de ce mémoire, nous avons proposé une opérationnalisation de cette étude qui permettrait à chaque assureur de mieux suivre son portefeuille et de mieux fidéliser les clients. En résumé, nous proposons l'algorithme du XGBoost pour détecter les contrats susceptibles d'être résiliés. Les contrats détectés doivent alors passer un second filtre : les critères de résiliation obtenus avec la régression logistique. Enfin, les contrats ne vérifiant pas ces critères doivent être suivis de prêt pour d'éventuels gestes commerciaux ou mieux d'actions visant à les retenir en portefeuille.

Les pistes d'amélioration de ce travail

Cette étude a plusieurs pistes d'améliorations :

— Le premier point d'amélioration concerne la construction du scoring de sinistralité. En effet, nous avons uniquement construit ce scoring sur les sinistres du contrat en cours. Ce scoring pourrait être généralisé sur l'ensemble de la sinistralité du contrat en prenant en compte les antécédents de sinistres ;

— Une autre piste d'amélioration serait de chercher à comprendre les déterminants de la résiliation par motif. Cela permettrait à la direction compétente de savoir quelle action prendre selon les différents motifs. Toutefois, des exercices de fiabilisation des données devront être réalisés car comme nous l'avions souligné, il apparaît que le gestionnaire n'enregistre pas correctement les vrais motifs.

Pour une meilleure rétention des clients, cette étude est un outil d'aide à la décision pour la direction technique, marketing et aussi pour les agents GAN. Nous avions clairement identifié les critères de résiliations. Toutefois, dans son application telle que nous l'avons décrite dans la section 5.7, il faudra faire attention entre les bons et les mauvais risques [1]. Ainsi, selon les contrats détectés, les agents peuvent, dès lors qu'ils jugent que c'est un "bon" risque, appliquer les outils de fidélisation habituels. Ils pourront par exemple faire des rabais tarifaires à certains contrats qui ont une revalorisation supérieure à la normale (la politique tarifaire fait une revalorisation de 2,75% annuelle actuellement). D'ailleurs, comme c'est déjà le cas à GAN, les agents disposent chaque année des enveloppes qui leur permettent de choisir un certain nombre de contrats sur lesquels ils n'appliqueront pas de revalorisation. Aussi, cette étude leur permettra de choisir les bons contrats.

1. le risque ici concerne les conséquences financières qu'un évènement du type accident de circulation, dommages etc peut engendrer pour le contrat ciblé.

Bibliographie

[1] A. KASSAMBARA, « *Machine Learning Essentials, Practical Guide in R* » , Edition 1.

[2] C. BOUQUET & P. MENARD, « *Assurance Automobile - Optimisation des ressources à l'échéance* » , Mémoire de l'institut des actuaires, CEA, 2011.

[3] D. W. HOSMER & S. LEMESHOW, « *Applied Logistic regression* », Second Edition, Wiley, 2000.

[4] E. BIERNAT & M. LUTZ, « *Data science : Fondamentaux et études de cas - Machine learning avec Python et R* », Edition Eyrolles.

[5] INSEE, « *Évolution de la population Bilan démographique 2020 - Tableaux rétrospectifs : Chiffres détaillés* », paru le 29/03/2021 sur le site `https://www.insee.fr/fr/statistiques/5007690?sommaire=5007726`.

[6] J. GRAU, I. GROSSE & J. KEILWAGEN, « *PRROC : computing and visualizing Precision-Recall and Receiver Operating Characteristic curves in R* » , Bioinformatics ,March 2015. Click here to get the pdf

[7] J. H. FRIEDMAN, « *Greedy function approximation : A Gradient Boosting Machine* », The Annals of Statistics, Vol. 29, No. 5 , pp. 1189-1232, oct. 2001.

[8] L. BREIMAN, « *Random Forests* », Machine Learning, 45, 5–32, 2001.

[9] M. CHEVALIER, C. LAUNAY & B. MAINGUY, « *La Bancassurance — Analyse de la situation de la Bancassurance dans le monde* » Focus, octobre 2005

[10] M. FERNANDEZ-DELGADO, E. CERNADAS & S. BARRO, « *Do we Need Hundreds of Classifiers to Solve Real World Classification Problems ?* », Journal of Machine Learning Research, October 2014. Click here to get the pdf

[11] M. ÓSKARSDÓTTIR, W. AHMED, K. ANTONIO, B. BAESENS, R. DENDIEVEL, T. DONAS & T. REYNKENS, « *Social network analytics for supervised fraud detection in insurance* », Cornell University, arXiv :2009.08313 [cs.SI], 15 september 2020.

[12] P. OTTOU, « *Méthodes d'apprentissage automatique appliquées au provisonnement ligne à ligne en assurance non-vie* », Mémoire d'actuariat - Institut des Actuaires, Université Paris Dauphine, 2017.

[13] R. RAKOTOMALALA, « *Pratique de la Régression Logistique : Régression Logistique Binaire et Polytomique* » Version 2.0, Université de Lyon, 2017.

[14] R. GENUER & J. POGGI, « *Arbres CART et Forêts aléatoires, Importance et sélection de variables* », hal-01387654v2, 2017.

[15] S. M. LUNDBERG, & S. LEE, « *A unified approach to interpreting model predictions* », University of Washington, 2017.

[16] S. M. LUNDBERG, G. G. ERION, & S. LEE, « *Consistent individualized feature attribution for tree ensembles* », University of Washington, 7 March 2019.

[17] T. CHEN & C. GUESTRIN, « *XGBoost : A Scalable Tree Boosting System* », Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 2016, p. 785–794.

[18] X. HE, M. GAO, M. KAN & D. WANG, « *BiRank : Towards Ranking on Bipartite Graphs* »,IEEE Transactions on knowledge and data engineering, arXiv :1708.04396v1 [cs.IR], 15 August 2017.

Annexe

A.1 Statistiques descriptives

A.1.1 Âge moyen par département et par année

L'âge moyen par département semble expliquer les disparités géographiques observées entre les départements en termes de propension à résilier. De manière globale, plus l'âge moyen du département est faible, plus le taux de résiliation de ce département est élevé.

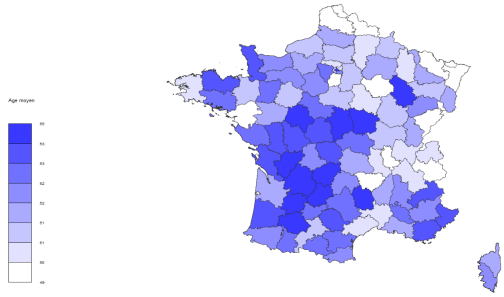


FIGURE A.1 – Âge moyen par département en 2017

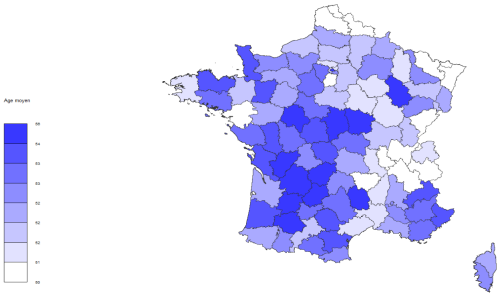


FIGURE A.2 – Âge moyen par département en 2018

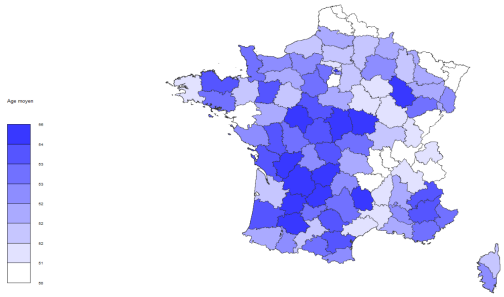


FIGURE A.3 – Âge moyen par département en 2019

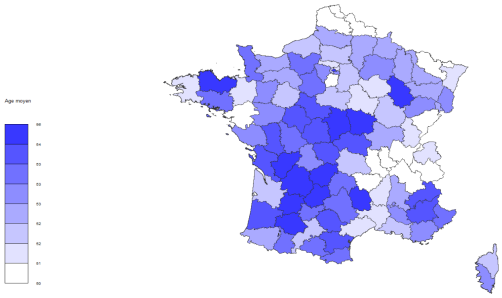


FIGURE A.4 – Âge moyen par département en 2020

A.1.2 Âge moyen par zone d'habitation

Il y a une légère différence entre les différentes zones d'habitation. Les zones rurales ont un âge légèrement plus élevé en moyenne par rapport aux zones urbaines. Par conséquent, le taux de résiliation dans les zones urbaines est plus élevé que dans les zones rurales.

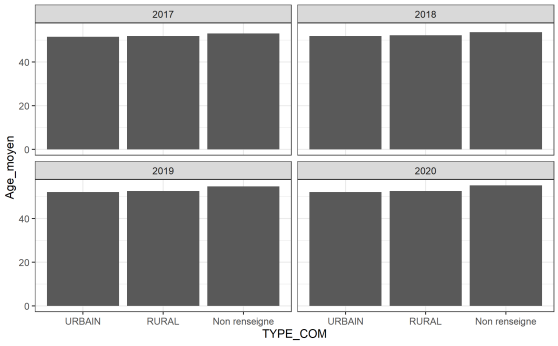


FIGURE A.5 – Âge moyen par zone d'habitation

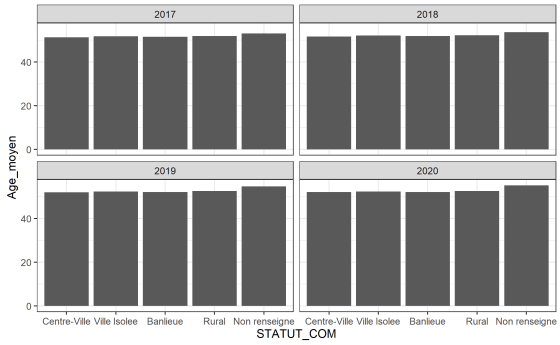


FIGURE A.6 – Âge moyen par zone d'habitation - plus fine

A.1.3 Distribution de l'âge des assurés par zone d'habitation

Les histogrammes représentant l'âge des assurés par zone d'habitation permet de se rendre compte de la légère différence entre les ruraux et les assurés vivant dans les villes en termes d'âge.

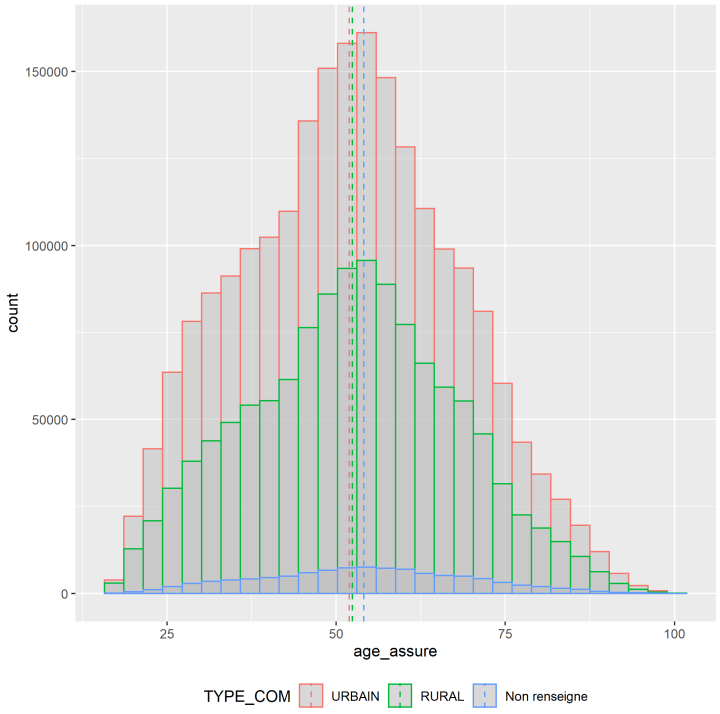


FIGURE A.7 – Distribution de l'âge des assurés par zone d'habitation

A.1.4 Taux de résiliation par âge de l'assuré

Les résultats de la figure A.8 confirment les conclusions précédentes. En effet, nous observons globalement une diminution du taux de résiliation avec l'âge ; les jeunes de $18-24$ ans étant les plus risqués et les assurés âgés entre 75 et 80 ans, les moins risqués. Les assurés les plus prépondérants dans le portefeuille GAN sont ceux ayant un âge entre 50 et 55 ans.

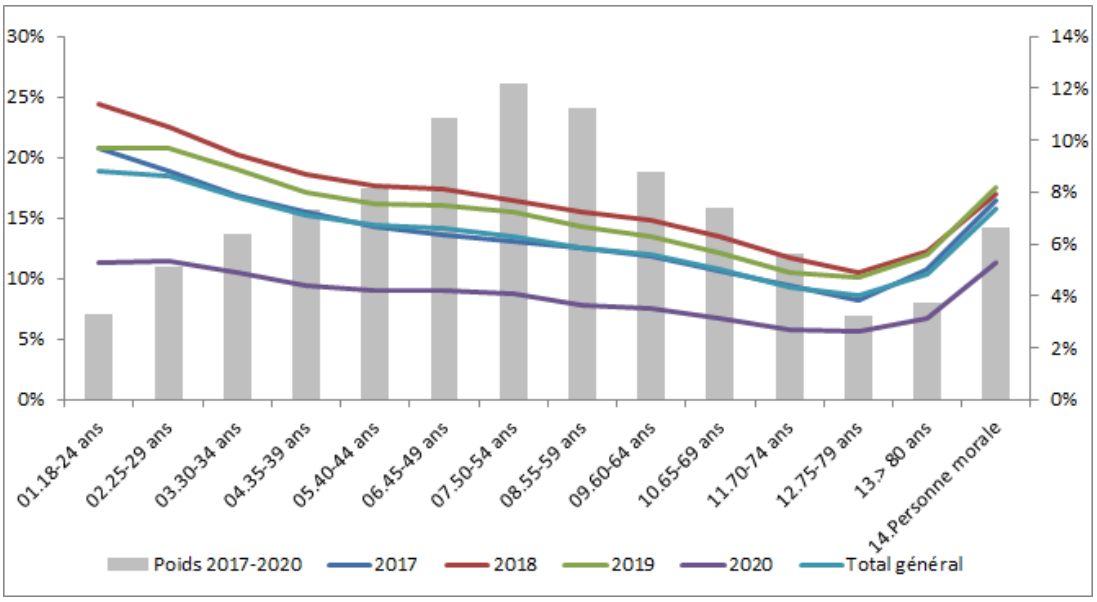


FIGURE A.8 – Taux de résiliation par âge de l'assuré

A.1.5 Matrice de corrélation des 104 variables

Les corrélations entre les différentes variables sont difficiles à apprécier graphiquement. Toutefois, plus la couleur du carré défini par deux variables tend vers le bleu, plus ces deux variables sont corrélées. L'intensité des liaisons entre les variables est calculée par :

— le ρ de Pearson pour la corrélation entre 2 variables quantitatives ;
— le V de Cramer pour la mesure de l'intensité de liaison entre 2 variables qualitatives ;
— le carré du rapport de corrélation η^2 pour mesurer la corrélation entre une variable quantitative et une variable qualitative.

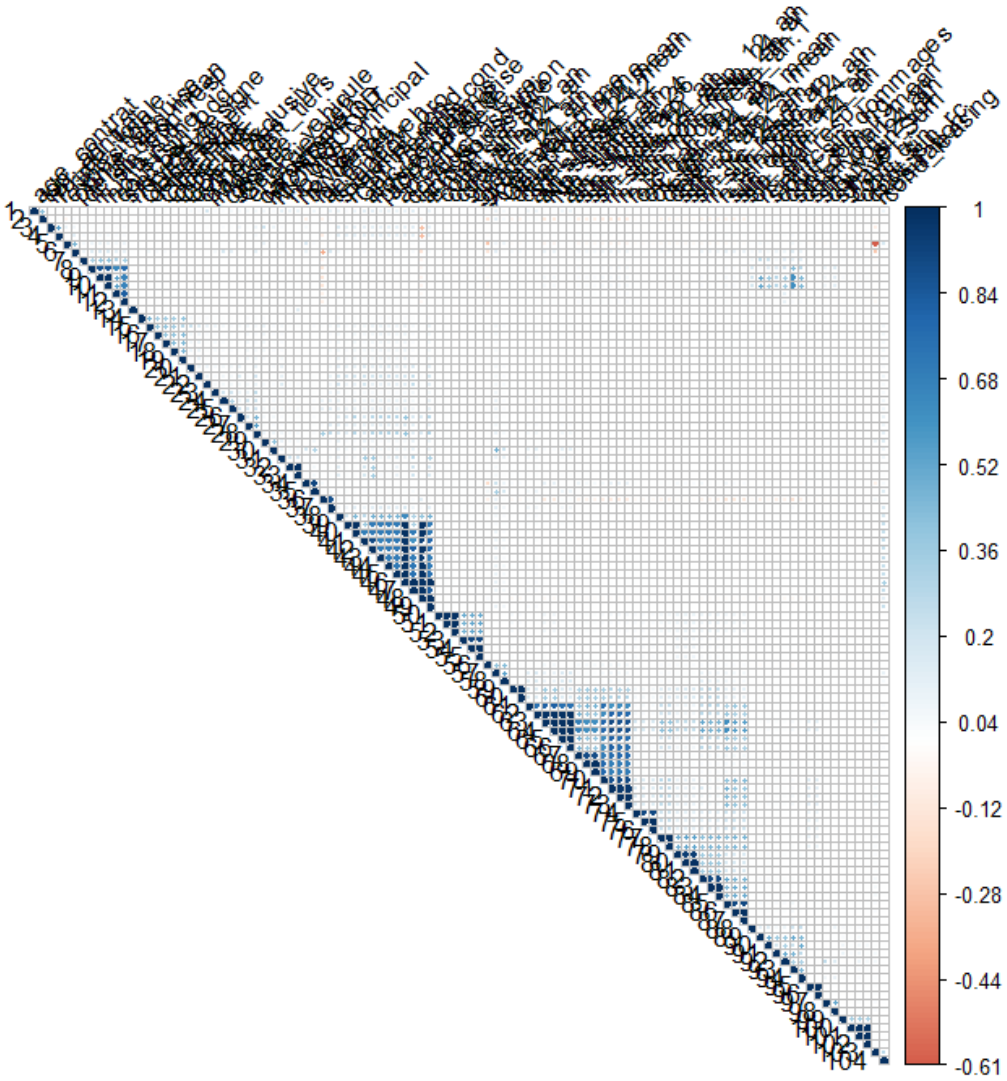


FIGURE A.9 – Matrice de corrélation des 104 variables

A.2 Les motifs et les variables

A.2.1 Regroupement des motifs de résiliation

Motifs	Les sous-motifs
Vente de véhicule	Vente du bien, Vente véhicule art L121-11 du CA
Initiative de l'assuré	Résiliation à l'échéance par MR, Initiative de l'assuré Résiliation à Loi Chatel
Concurrence	April, Areas CMA, AVIVA,AXA Azur, Banque Populaire, Beneficiaire ACS/Apria Caisse d'épargne, CIC, Direct assurances, GMF Groupe Zephir, La Suisse, LCL Matmut, MFA, MRA, Mutuelle des motards Mutuelle du Poitou, Mutuelle Saint Christophe Prevadies, Reprise autre assureur Crédit mutuel, Résil.Concurrence Autre cie Pacifica, GROUPAMA AGF, MAAF, MACIF MDM, UAP,SAMDA, SMACL, Société Générale Swiss Life, Winterthur, Zurich
Remplacement	Ajout nouveau risque Remplacement Remplacement même type contrat Contrat transféré
Hamon	Hamon date classique
Changement de situation	Cessation activité - Retraite, Changement de situation Changement type de contrat, Cessation exploitation
Destruction de véhicule	Résiliation véhicule détruit, résiliation véhicule hors usage Disparition du risque art L121-9 du CA
Compagnie	Nullité, Nullité fausse déclaration Redressement erreur, Remise ne vigueur Résiliation Absence Protection Juridique Résiliation administrative, Résiliation contentieux
Autres motifs	Résiliation décès art L121-10 du CA Migration,Reprise non agricole

TABLE A.1 – Le regroupement des motifs de résiliation

A.2.2 Dictionnaire des variables

Variable	libellé
Résiliation	La variable indicatrice de la résiliation du contrat
prime	La prime du contrat
Prime_totale	La prime de tous les contrats en cours qu'a un client chez GAN
MONOCONT	Indique si le client a un seul contrat ou non chez GAN
MONOPROD	Indique si le client a souscrit aux contrats d'un seul produit ou non
nb_contrat	Le nombre total de contrats que détient le client chez GAN
tr_Age_assure	Âge de l'assuré regroupé par tranche
tr_anc_veh	Ancienneté du véhicule regroupée par tranche
CRM	Coefficient de Réduction et de Majoration
tx_derog	Taux de dérogation (somme des avantages et gestes commerciaux)
Avenant	Présence d'avenant pour un contrat considéré
ok_gar	Le nombre de sinistres dont la charge correspond uniquement aux frais d'experts
ok_gar_24	Le nombre d'antécédent de sinistres des 24 derniers mois dont la charge totale = frais d'experts
cout_sin	Charge de sinistres du contrat
cout_g	Charge des sinistres graves du contrat
cout_sin_rc	Charge de sinistres de la garantie RC
cout_sin_dommages	Charge de sinistres de la garantie dommages
cout_sin_bdg	Charge de sinistres de la garantie bris de glaces
cout_sin_vol	Charge de sinistres de la garantie vol
energie_veh	Énergie utilisée par le véhicule assuré
departement	Le département dans lequel se situe le garage
an_sinistre	Antécédent de sinistres sur toute la période
an_sinistre_12	Antécédent de sinistres sur les 12 derniers mois
an_sinistre_24	Antécédent de sinistres sur les 24 derniers mois
an_sinistre_36	Antécédent de sinistres sur les 36 derniers mois
an_sinistre_48	Antécédent de sinistres sur les 48 derniers mois
Marque	Marque du véhicule
Genre_vehicule	Le genre du véhicule
groupe	Le groupe du véhicule
cond_exclusive	Indique si le véhicule est conduit uniquement par l'assuré
cond_leasing	Indique si le véhicule assuré est un véhicule de location de longue durée
paiement_tiers	Indique la fréquence de paiement de la prime
Cat_Var_prime	Évolution de la prime en euros
Cat_delta_prime	Évolution de la prime en %
situation_pro	La situation socio-professionnelle de l'assuré
delai_trait_mean	Le délai moyen de traitement d'un sinistre pour les contrats sinistrés
delai_trait_12_mean	Le délai moyen de traitement d'un sinistre pour les contrats sinistrés dans les 12 derniers mois
delai_trait_24_mean	Le délai moyen de traitement d'un sinistre pour les contrats sinistrés dans les 24 derniers mois

TABLE A.2 – Dictionnaire des variables importantes

A.3 Encadrés méthodologiques

Encadré 1 : GLM

Le GLM - Generalized Linear Models est une généralisation de la régression linéaire simple. Sa mise en place nécessite des hypothèses.

La famille exponentielle

Pour mettre en place un GLM, il est indispensable d'avoir des distributions appartenant à la famille exponentielle.

Forme générale

Les lois de la famille exponentielle sont des lois de probabilité à deux paramètres θ et ϕ dont la densité par rapport à la mesure dominante (mesure de comptage $\mathbf{N}$ ou mesure de Lebesgue sur R) est de la forme :

$$f\left(y \mid \theta, \phi\right) = \exp\left(\frac{y\theta - b\left(\theta\right)}{a\left(\phi\right)} + c\left(y, \phi\right)\right), \tag{A.1}$$

où :

- $a\left(.\right)$ est une fonction non nulle définie sur $\mathbf{R}$,
- $b\left(.\right)$ est une fonction définie sur $\mathbf{R}$ et 2 fois dérivable,
- $c\left(.\right)$ est une fonction définie sur $\mathbf{R} * \mathbf{R}$ et 2 fois dérivable,
- θ est appelé paramètre naturel ou d'intérêt et ϕ est appelé paramètre de dispersion ou de nuisance.

Quelques exemples

La loi normale : La loi gaussienne de moyenne μ et de variance σ^2, $\mathcal{N}(\mu, \sigma^2)$ appartient à la famille exponentielle avec $\theta{=}\mu$,$\phi{=}\sigma^2$,$a\left(\phi\right){=}\phi$, $b\left(\theta\right){=}\theta^2/2$ et $c\left(y, \theta\right){=}-\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log\left(2\pi\sigma^2\right)\right)$; $y{\in}\ \mathbf{R}$.

La loi de Poisson : La loi de Poisson de moyenne λ, $\mathcal{P}\left(\lambda\right)$ appartient à la famille exponentielle avec $\theta{=}\log\left(\lambda\right)$,$\phi{=}1$,$a\left(\phi\right){=}1$, $b\left(\theta\right){=}\exp\left(\theta\right){=}\lambda$ et $c\left(y, \theta\right){=}-\log\left(y!\right)$; $y \in \mathbf{N}$. Nous avons aussi les lois binomiale négative, Gamma qui font partie de la famille exponentielle.

Encadré 1 : GLM - Suite

Hypothèses et principe du GLM

Le principe des GLM est simple. Au lieu de modéliser directement la variable dépendante, c'est plutôt une fonction de l'espérance de cette variable (fonction de lien) qui est modélisée.
Considérons Y_i, $i = 1, 2, ..., n$ la variable à expliquer et X_i=$(1, X_{i1} =, X_{i2}, ..., X_{ip})$ le vecteur des variables explicatives. Y_i s'apprête à un modèle GLM si :

- la loi conditionnelle de Y_i sachant $X_{i1} = x_{i1}, X_{i2} = x_{i2}, ..., X_{ip} = x_{ip}$ est telle qu'il existe une fonction de lien h strictement monotone de R vers R et des coefficients $(\beta_0, \beta_1, ..., \beta_p)$ tels que :

$$h\left(\mathbb{E}\left(Y_i\right)\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_i j \tag{A.2}$$

- La loi de probabilité de Y_i doit appartenir à la famille exponentielle.

Dans le cadre de la régression logistique, cette fonction de lien est la fonction *logit* définie par $h(x) = \log(\frac{x}{1-x})$.

Estimation des paramètres du modèle

L'estimation des paramètres d'un modèle GLM se fait avec la méthode de maximum de vraisemblance. Ainsi, il suffit de trouver les zéros de la dérivée première de la vraisemblance [a]tout en vérifiant les conditions de second degré. La vraisemblance s'écrit :

$$L\left(\theta_1, \theta_2, ..., \theta_n; \phi_1, \phi_2, ..., \phi_n; y_1, y_2, ..., y_n\right) = \prod_{i=1}^{n} f\left(y_i; \theta_i; \phi_i\right)), \tag{A.3}$$

En pratique, il est compliqué d'obtenir analytiquement des solutions des équations résultant des conditions de premier ordre. La méthode de Newton-Raphson permet d'obtenir des estimateurs des coefficients solutions.

a. La vraisemblance est le produit de fonction de densité.

Encadré 2 : Calcul de la SHAP Value

La SHAP Value, développé en 2017 par Lunderberg et al. [15] est une mesure de l'effet marginale d'une variable sur le modèle qui a été inspirée de la valeur SHapley introduite en 1951 en théorie des jeux. En effet, elle permet de mesurer l'impact de chaque variable sur la prédiction de la variable cible. Après avoir démontré que les mesures de la contribution des variables, utilisées jusque là comme par exemple l'importance des variables, sont non robustes, les auteurs ont proposé la SHAP Value qui est une mesure robuste.

Dans sa formulation, l'effet global des variables d'un modèle sur une variable cible serait la somme des effets ϕ_j de chacune de ces variables explicatives X^j qui s'écrit :

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(M! - |S|! - 1)!}{M!} \left[f_x(S \cup \{j\}) - f_x(S) \right] \text{ [16]}$$

avec :

- N : l'ensemble de toutes les variables explicatives ;
- M : le nombre totale de variables explicatives ;
- S : un ensemble de variables explicatives ;
- f_x : la fonction de prédiction à l'instant x c'est-à-dire $f_x(S) = \mathbf{E}[f_x|x_S]$.

La SHAP value a les propriétés suivantes :

- **Additivité :** La somme de des effets de toutes les variables donne la prédiction du modèle à une constante près ϕ_0 ;
- **Cohérence :** Si l'effet d'une variable sur le modèle devient plus important suite à un changement de modèle, l'attribution assignée à cette variable ne doit pas baisser ;
- **Variables nulles sans effet :** Si une variable de l'exemple considéré est à zéro, alors la variable ne doit pas avoir d'impact pour cet exemple.

Suite à cela, la prédiction $\hat{y}$ de l'observation y du modèle peut s'écrire :

$$\hat{y}_i = \phi_0 + \sum_{j=1}^{N} \phi_j y_i$$

A.4 Les sorties .R de la régression logistique

Le modèle 1 est celui qui a à la fois la plus faible déviance (158003) et le plus faible AIC (1585085). Toutefois, il faut remarquer une valeur très élevée de la déviance qui mesure la différence entre la valeur observée et la valeur prédite du modèle. C'est l'une des conséquences des données déséquilibrées avec la régression logistique.

Pour les deux modèles, toutes les variables sont significatives avec le test de Wald.

A.4.1 Sortie R Régression logistique - Modèle 1 et 2

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6244  -0.6175  -0.5342  -0.4202   3.4821

Coefficients:
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                    -1.605480   0.006692 -239.920  < 2e-16 ***
anc_contrat02.entre 1 et 3 ans  0.153310   0.006284   24.397  < 2e-16 ***
anc_contrat03.entre 3 et 5 ans  0.141348   0.007029   20.110  < 2e-16 ***
anc_contrat04.entre 5 et 9 ans  0.048159   0.006700    7.188 6.56e-13 ***
anc_contrat05.entre 9 et 20 ans -0.141103   0.006975  -20.231  < 2e-16 ***
anc_contrat06.20 ans et plus    -0.233382   0.012746  -18.310  < 2e-16 ***
tr_Age_assure01.<= 34 ans        0.046882   0.007095    6.608 3.90e-11 ***
tr_Age_assure03.50-59 ans       -0.063480   0.005823  -10.901  < 2e-16 ***
tr_Age_assure04.60-69 ans       -0.159638   0.006879  -23.207  < 2e-16 ***
tr_Age_assure05.70-79 ans       -0.382051   0.009527  -40.100  < 2e-16 ***
tr_Age_assure06.80 ans et plus  -0.231071   0.013265  -17.420  < 2e-16 ***
tr_Age_assure07.Personne morale  0.264804   0.009198   28.789  < 2e-16 ***
tr_anc_veh00.<=1 ans            -0.732313   0.011967  -61.195  < 2e-16 ***
tr_anc_veh01.1 ans              -0.329796   0.011553  -28.547  < 2e-16 ***
tr_anc_veh02.2 ans              -0.104958   0.010706   -9.804  < 2e-16 ***
tr_anc_veh03.3 ans              -0.134420   0.010644  -12.628  < 2e-16 ***
tr_anc_veh04.4-6 ans            -0.188756   0.006616  -28.530  < 2e-16 ***
tr_anc_veh07.7-8 ans            -0.163000   0.007284  -22.377  < 2e-16 ***
tr_anc_veh09.15-19 ans           0.209185   0.006167   33.922  < 2e-16 ***
tr_anc_veh10.>=20 ans            0.220525   0.007128   30.939  < 2e-16 ***
paiement_tiersTrimestriel        0.104582   0.016055    6.514 7.33e-11 ***
paiement_tiersSemestriel        -0.062443   0.006097  -10.241  < 2e-16 ***
paiement_tiersAnnuel            -0.173541   0.005431  -31.955  < 2e-16 ***
Tr_crm01.entre 0.5 et 0.72       0.205206   0.006959   29.489  < 2e-16 ***
Tr_crm02.entre 0.73 et 1         0.366748   0.008002   45.830  < 2e-16 ***
Tr_crm03.entre 1 et 2            0.796323   0.035902   22.180  < 2e-16 ***
Cat_delta_prime00.<-10           -0.023535   0.009707   -2.424   0.0153 *
Cat_delta_prime01.-10 a 0        -0.219494   0.010591  -20.725  < 2e-16 ***
Cat_delta_prime03.4 a 6          0.268265   0.008595   31.212  < 2e-16 ***
Cat_delta_prime04.7 a 19         0.248333   0.009241   26.874  < 2e-16 ***
Cat_delta_prime05.20 et +        0.085276   0.013417    6.356 2.07e-10 ***
energie_vehelectrique            0.308220   0.058827    5.239 1.61e-07 ***
energie_vehessence              -0.052782   0.004946  -10.672  < 2e-16 ***
energie_vehGPL - Non renseigne  -0.054102   0.012871   -4.204 2.63e-05 ***
Genre_vehicule21 - 22            0.092602   0.008673   10.678  < 2e-16 ***
Genre_vehicule23                 0.304500   0.029302   10.392  < 2e-16 ***
Genre_vehicule30 à 32           -0.179646   0.007546  -23.806  < 2e-16 ***
Genre_vehicule33 et +           -0.455145   0.019029  -23.918  < 2e-16 ***
tx_derog                        -4.035114   0.038620 -104.482  < 2e-16 ***
an_sinistre_2401.1 sinistre      0.287046   0.006689   42.913  < 2e-16 ***
an_sinistre_2402.2 sinistres et + 0.767668   0.014621   52.506  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1632413  on 1908962  degrees of freedom
Residual deviance: 1585003  on 1908922  degrees of freedom
AIC: 1585085
```

FIGURE A.10 – Sortie R Régression logistique - Modèle 1

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.8106  -0.6141  -0.5338  -0.4387   2.8849

Coefficients:
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                    -1.643023   0.006556 -250.630  < 2e-16 ***
anc_contrat02.entre 1 et 3 ans  0.193044   0.006211   31.082  < 2e-16 ***
anc_contrat03.entre 3 et 5 ans  0.190737   0.006971   27.363  < 2e-16 ***
anc_contrat04.entre 5 et 9 ans  0.106199   0.006662   15.940  < 2e-16 ***
anc_contrat05.entre 9 et 20 ans -0.082303   0.006873  -11.975  < 2e-16 ***
anc_contrat06.20 ans et plus    -0.173610   0.012634  -13.742  < 2e-16 ***
tr_Age_assure01.<= 34 ans        0.060405   0.007034    8.588  < 2e-16 ***
tr_Age_assure03.50-59 ans       -0.062716   0.005748  -10.911  < 2e-16 ***
tr_Age_assure04.60-69 ans       -0.167102   0.006779  -24.650  < 2e-16 ***
tr_Age_assure05.70-79 ans       -0.379936   0.009237  -41.132  < 2e-16 ***
tr_Age_assure06.80 ans et plus  -0.225742   0.012812  -17.619  < 2e-16 ***
tr_Age_assure07.Personne morale  0.278515   0.009186   30.320  < 2e-16 ***
tr_anc_veh00.<=1 ans            -0.746319   0.012325  -60.554  < 2e-16 ***
tr_anc_veh01.1 ans              -0.371116   0.011565  -32.091  < 2e-16 ***
tr_anc_veh02.2 ans              -0.118904   0.010592  -11.226  < 2e-16 ***
tr_anc_veh03.3 ans              -0.141544   0.010506  -13.473  < 2e-16 ***
tr_anc_veh04.4-6 ans            -0.195818   0.006600  -29.667  < 2e-16 ***
tr_anc_veh07.7-8 ans            -0.169358   0.007216  -23.471  < 2e-16 ***
tr_anc_veh09.15-19 ans           0.229610   0.006020   38.143  < 2e-16 ***
tr_anc_veh10.>=20 ans            0.251639   0.006923   36.350  < 2e-16 ***
paiement_tiersTrimestriel        0.136008   0.015871    8.570  < 2e-16 ***
paiement_tiersSemestriel        -0.088970   0.006099  -14.588  < 2e-16 ***
paiement_tiersAnnuel            -0.183622   0.005370  -34.192  < 2e-16 ***
Tr_crm01.entre 0.5 et 0.72       0.176862   0.006895   25.649  < 2e-16 ***
Tr_crm02.entre 0.73 et 1         0.344462   0.007922   43.481  < 2e-16 ***
Tr_crm03.entre 1 et 2            0.762461   0.036061   21.144  < 2e-16 ***
Cat_delta_prime00.<-10           0.056353   0.009801    5.750 8.95e-09 ***
Cat_delta_prime01.-10 a 0       -0.143406   0.010226  -14.023  < 2e-16 ***
Cat_delta_prime03.4 a 6          0.139905   0.007768   18.010  < 2e-16 ***
Cat_delta_prime04.7 a 19         0.189272   0.008781   21.554  < 2e-16 ***
Cat_delta_prime05.20 et +        0.128711   0.013754    9.358  < 2e-16 ***
energie_vehelectrique            0.322292   0.054518    5.912 3.39e-09 ***
energie_vehessence              -0.057569   0.004862  -11.841  < 2e-16 ***
energie_vehGPL - Non renseigne  -0.061069   0.012753   -4.788 1.68e-06 ***
Genre_vehicule21 - 22            0.091884   0.008878   10.349  < 2e-16 ***
Genre_vehicule23                 0.324872   0.030209   10.754  < 2e-16 ***
Genre_vehicule30 à 32           -0.158924   0.007606  -20.894  < 2e-16 ***
Genre_vehicule33 et +           -0.264213   0.014059  -18.793  < 2e-16 ***
tx_derog                        -1.908893   0.026222  -72.796  < 2e-16 ***
an_sinistre_2401.1 sinistre      0.236537   0.006195   38.180  < 2e-16 ***
an_sinistre_2402.2 sinistres et +  0.664761   0.012479   53.270  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1674447  on 1953821  degrees of freedom
Residual deviance: 1631598  on 1953781  degrees of freedom
AIC: 1631680
```

FIGURE A.11 – Sortie R Régression logistique - Modèle 2

A.4.2 AFDM

Les variables qui définissent les deux premiers axes de l'analyse factorielle avec les données mixtes sont principalement les données de sinistralité. Les variables qui décrivent les sinistres qui ont touché la garantie "Bris de Glaces" sont celles qui ont le plus contribué au deuxième axe.

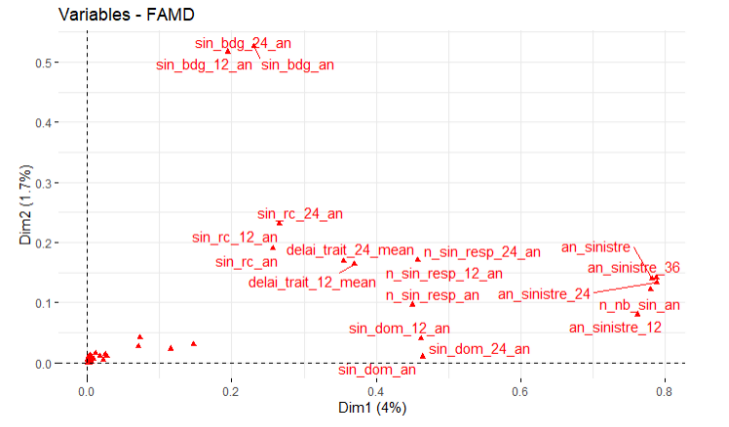


FIGURE A.12 – Les variables dans le plan des deux premières axes

A.5 Quelques résultats du XGBoost

XGBoost est le modèle qui détecte le mieux les résiliations. Il a l'avantage d'agir non seulement sur le biais du prédicteur, mais aussi sur sa variance. En effet, l'amélioration des résultats à chaque itération du boosting permet de minimiser davantage le biais et l'agrégation des prédicteurs issus des itérations permet de réduire la variance.

En introduisant la variable score, nous obtenons une courbe ROC plus concave et une courbe Rappel-Précision plus proche du cadran Nord et Est.

A.5.1 La courbe ROC

La courbe ROC représente le rappel en fonction du taux de faux positifs en faisant varier le seuil de 0 à 1. La croix rouge sur la courbe désigne les coordonnées du XGBoost Score dans ce plan d'indicateurs.

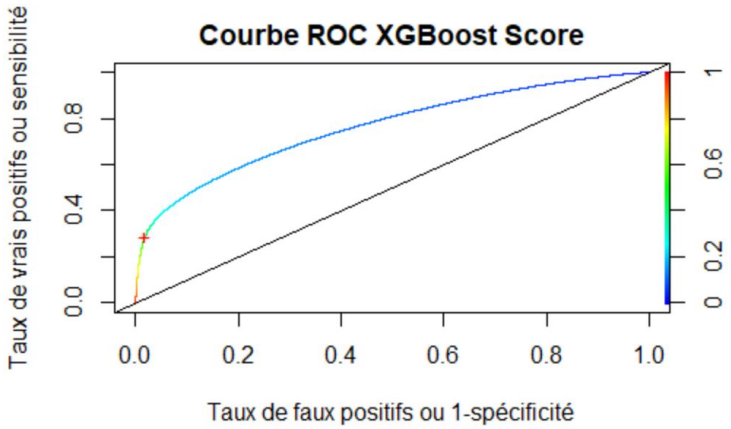


FIGURE A.13 – Les variables dans le plan des deux premières axes

A.5.2 La courbe Rappel-Précision

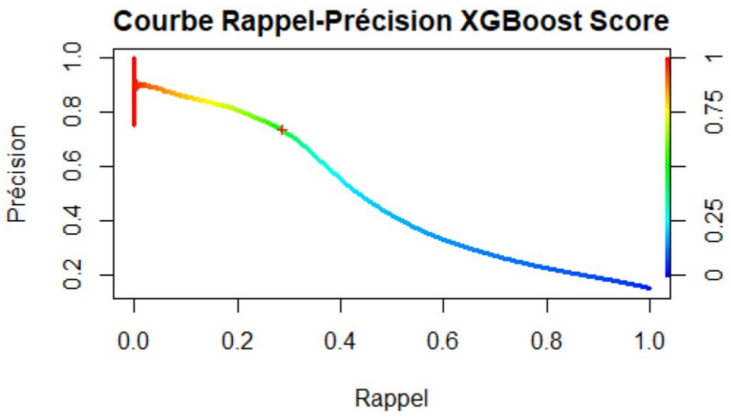


FIGURE A.14 – Les variables dans le plan des deux premières axes

Table des figures

Liste des tableaux