EURIA – EURo Institut d'Actuariat

Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 21 Septembre 2018

Par : Fatima-Zohra Zouggagh
Titre : Tarification automobile à l'aide de modèles de *machine learning* et apport des données télématiques.

Confidentialité : Oui (Durée: 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membre présent du jury de l'Institut des Actuaires :
Florence Picard
Sonia Guélou
Romain Nobis
Signature :

Entreprise :
Galea & Associés
Signature :

Membres présents du jury de l'EURIA :
Franck Vermet

Directeur de mémoire en entreprise :
Florence Chiu
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

Secrétariat :

Bibliothèque :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
F +33 (0)2 98 01 66 57
euria@univ-brest.fr

Résumé

Dans un contexte concurrentiel comme celui de l'assurance automobile, la tarification représente un défi de taille pour l'assureur. Celui-ci doit élaborer des modèles tarifaires qui doivent refléter au mieux le risque auquel il sera exposé, tout en segmentant aussi finement que possible son portefeuille.

Par ailleurs, dans un univers en pleine progression devant la multiplication des données, les actuaires se doivent d'adapter leurs outils pour la prédiction des comportements de leurs assurés.

Dans ce contexte, l'objectif de ce mémoire s'inscrit autour de deux points majeurs. Le premier consiste en une étude comparative des performances des modèles économétriques classiques, à savoir les modèles linéaires généralisés (GLM), avec des modèles plus innovants : CART, *Random Forest*, *Le Gradient Boosting Machine (GBM)* et l'*eXtreme Gradient Boosting (XGBoost)* ; et ce pour la prédiction du **nombre de sinistres** et du **coût du sinistre** d'un portefeuille de responsabilité civile en automobile. La seconde problématique de ce mémoire est d'étudier l'impact de l'ajout de données télématiques, récupérées auprès d'une entreprise spécialisée dans le comportement des conducteurs. L'enjeu est de comparer la performance des prédictions du **nombre de sinistres** et du **coût du sinistre** en ajoutant ces nouvelles variables avec celle obtenue sans ce complément d'information. L'opportunité d'utiliser des données externes sera également mise en relief.

Mots clefs: Tarification, assurance automobile, GLM, *machine learning*, télématiques, CART, GBM, XGBoost, segmentation, *Data Science*, données externes.

Abstract

In a competitive environment such as car insurance, pricing represents a major challenge for the insurer, insofar as he must develop pricing models that best reflect his risk exposure, while segmenting as finely as possible his portfolio.

Moreover, in a rapidly improving world given the data proliferation, actuaries must adapt their tools to predict their policyholders' behavior.

In this context, the purpose of this thesis involves two major points. The first point consists of a comparative study of the performance of classical econometric models, namely Generalized Linear Models (GLM), with more innovative models such as CART, *Random Forest*, *Gradient Boosting Machine (GBM)* and *eXtreme Gradient Boosting (XGBoost)*; this applied to predicting **claims number** and **claims cost** of a car liability insurance portfolio. The second point is around assessing the impact of adding telematics data recovered from a company specialized in driver behavior. The challenge is to compare the performance of the predictions of **claims number** and **claims cost** with and without adding this additional information. The opportunity of using external data will also be highlighted.

Keywords: Pricing, car insurance, GLM, *machine learning*, telematics, CART, GBM, XGBoost, segmentation, *Data Science*, external data.

Synthèse

Dans un secteur aussi concurrentiel et saturé comme le marché de l'assurance automobile, l'assureur est poussé à établir une tarification précise de son portefeuille afin de contenir les risques inhérents à son activité. Cette tarification doit s'appuyer sur une segmentation fine du portefeuille pour lutter contre l'antisélection. Ainsi, un assureur qui optimise sa tarification sera en mesure d'attirer de nouveaux assurés en leur proposant le juste prix tout en maintenant son équilibre technique. Une sous-tarification ou une sur-tarification conduirait forcément à une sélection adverse qui consiste à faire fuir les bons risques pour ne garder que les contrats déficitaires.

Depuis plusieurs années, les modèles linéaires généralisés (GLM) constituent un outil capital pour la tarification et sont ainsi largement déployés au sein des compagnies d'assurance. Aujourd'hui, le monde des données connaît une profonde révolution avec l'ouverture conséquente des données. Un grand nombre d'algorithmes issus de la théorie de l'apprentissage statistique a été mis en place pour traiter et analyser ces données, ils permettent ainsi de répondre à des problématiques de prédiction ou de classification. Ces travaux ont donné naissance à la théorie de l'apprentissage statistique. Cependant, l'engouement vis-à-vis de ces nouvelles méthodes novatrices s'accompagne de conclusions mitigées quant à leur application opérationnelle et leur interprétabilité. Elles sont d'ailleurs souvent qualifiées de méthodes *«boîte noire»*. A ce stade, l'implémentation des méthodes d'apprentissage statistique au sein des équipes de tarification est restreinte.

Par ailleurs, afin d'établir un modèle tarifaire, il est essentiel de retenir les meilleures variables pouvant expliquer la sinistralité. Alors que les tarifs reposaient sur un faible nombre de variables communiquées par l'assuré, il est facilement possible aujourd'hui d'accéder à un grand panel de variables externes pouvant enrichir les modèles existants et renforcer leurs capacités prédictives.

Objectifs du mémoire :

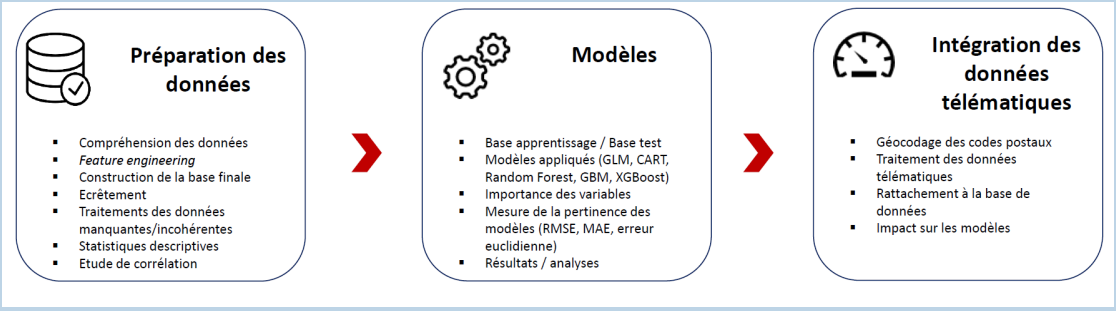
Ce mémoire se propose d'étudier deux problématiques :

- Mise en oeuvre sur un portefeuille d'assurance RC automobile d'une approche GLM et d'un panel d'algorithmes de *maching learning* pour évaluer la pertinence de ceux-ci et comparer leur efficacité dans la modélisation du **nombre de sinistres** et le **coût des sinistres** (hors sinistres graves).

- Étude de l'apport des données télématiques sur les prédictions du **nombre de sinistres** et du **coût des sinistres**. Ces données ont été récupérées auprès d'une entreprise spécialisée dans la collecte des données comportementales.

Démarche suivie :

L'étude est alors articulée selon les étapes suivantes :



Modélisation du nombre de sinistres et du coût des sinistres sans utilisation des données externes

L'étude repose sur les données du portefeuille d'un assureur automobile français relatives à la garantie responsabilité civile. Avant d'entamer les modélisations, une première étape a consisté à préparer la base de travail. En effet, nous avons adapté la structure des bases "contrats" et "sinistres" afin de les fusionner pour obtenir une base finale sur laquelle seront établis nos modèles. La base "contrats" contient l'ensemble des informations relatives au conducteur et à son véhicule. La base "sinistres" dispose de l'ensemble des informations liées à la sinistralité. Plusieurs études préliminaires ont été réalisées afin de vérifier la qualité de la base de données et de repérer les éventuelles corrélations entre les variables tarifaires qui pourraient biaiser le modèle GLM.

Un modèle basé sur le GLM a été réalisé pour estimer séparément le **nombre de sinistres** et le **coût des sinistres**. Ce modèle servira de référence dans le sens où les performances des modèles de *machine learning* lui seront comparées à travers notamment des mesures d'écarts des erreurs comme la RMSE, la MAE et l'erreur euclidienne.

Les approches de *machine learning* qui ont été testées au cours de cette étude sont :

— Les arbres de décision (CART) ;

— Les forêts aléatoires ou *Random Forest* ;

— *Le Gradient Boosting Machine (GBM)* ;

— L'*eXtreme Gradient Boosting (XGBooost).*

Dans un premier temps, ces modèles ont été exécutés sans effectuer d'optimisation des paramètres (hyperparamétrage). Ensuite, les modèles ont été calibrés sur la base d'apprentissage en testant plusieurs combinaisons de valeurs d'hyperparamètres. Les performances des modèles ainsi construits sont comparées sur la base de test avec celles des GLM établis grâce à la **RMSE**. Les résultats sont donnés au tableau suivant :

	Nombre de sinistres	Coût du sinistre
GLM	**0,2210493**	1449,768
CART	0,2217749	**1447,239**
Random Forest	0,2216543	1452,612
GBM	0,2220635	1551,913
XGBoost	0,2216571	1448,542

A travers les modèles construits, nous pouvons constater que les méthodes d'apprentissage statistique testées dans cette étude offrent des performances comparables à celles du GLM. En effet, que ce soit pour le **nombre de sinistres** ou le **coût des sinistres**, nous avons constaté des RMSE proches.

Concernant la modélisation du **nombre de sinistres**, nous constatons que le modèle dont l'erreur est la plus faible (RMSE= 0.2210493) est bien le **GLM**, il est suivi de **Random Forest** avec une valeur de RMSE égale à 0.2216543.

Quant à la modélisation du **coût des sinistres**, le modèle **CART** est meilleur avec une valeur de RMSE égale à 1447.239, il est suivi du **XGBoost** dont la RMSE vaut 1448.542.

Ensuite, les variables qui contribuent le plus à chacun des modèles ont été mises en exergue.

Importance des variables pour les modèles du nombre de sinistres

Le tableau suivant permet de résumer l'importance des variables [1] qui apparaissent pertinentes pour la modélisation du **nombre de sinistres** pour chaque modèle. Ces variables sont mises en évidence avec la couleur **bleu**. Par ailleurs, les variables qui apparaissent en premier dans chaque modèle sont présentées en **rouge**.

1. Le détail de chaque variable est expliqué dans la partie 3.1

GLM	CART	Random Forest	GBM	XGBoost
CRM	**CRM**	CRM	CRM	**CRM**
Zone	Zone	Zone	Zone	Zone
Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B
Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule
Genre	Genre	Genre	Genre	Genre
Formule	Formule	Formule	Formule	Formule
Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage
Classe_SRA	Classe_SRA	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA
Novice	Franchise auto	Franchise auto	Franchise auto	Franchise auto

Pour conclure, il serait préférable de garder le modèle GLM pour la modélisation du **nombre de sinistres**. Ce modèle offre de bonnes performances et est notamment facilement interprétable.

Importance des variables pour les modèles du coût des sinistres

Comme pour le **nombre de sinistres**, nous résumons l'importance des variables pour chaque modèle dans le tableau suivant. Les variables importantes qui contribuent à la construction des modèles sont mises en évidence avec la couleur **bleu**. Les variables qui apparaissent en premier dans les modèles sont présentées en **rouge**.

GLM	CART	Random Forest	GBM	XGBoost
CRM	CRM	CRM	CRM	CRM
Zone	Zone	Zone	Zone	Zone
Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B
Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule
Genre	Genre	Genre	Genre	Genre
Formule	**Formule**	Formule	Formule	**Formule**
Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage
Classe_SRA	Classe_SRA	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA
Novice	Franchise auto	Franchise auto	Franchise auto	Franchise auto

Une amélioration potentielle de ces prédictions serait d'intégrer des données externes à notre base de données, car elles capteront des informations supplémentaires non communiquées à la souscription du contrat. C'est dans ce contexte que nous intégrerons à notre portefeuille des données issues de la télématique dans le but de mesurer leurs apports dans les modélisations du **nombre de sinistres** et du **coût des sinistres**.

Modélisation du nombre de sinistres et du coût des sinistres après intégration des données télématiques

Comme mentionné précédemment, avec l'avènement du *Big Data* et l'ouverture des données *open source* ainsi que le développement des outils connectés, les assureurs peuvent enrichir leurs tarifs par l'apport de données externes. Ces données leur permettent une analyse plus fine du risque.

Dans ce mémoire, et dans le cadre d'un partenariat avec une entreprise nommée **Ellis Car** spécialisée dans l'étude du comportement des conducteurs, le cabinet de conseil Galea & Associés a réussi à récupérer les données liées à la télématique afin d'étudier leurs impacts sur les prédictions de la sinistralité.

Un traitement préalable à l'utilisation de ces données a été effectué afin de remédier à la problématique des données manquantes. Une comparaison des performances des modélisations sur la base des RMSE est donnée sur le tableau suivant :

	Sans données télématiques		**Avec données télématiques**	
	Nombre de sinistres	**Coût du sinistre**	**Nombre de sinistres**	**Coût du sinistre**
GLM	**0,2210493**	1449,768	**0,2195956**	1449,171
CART	0,2217749	**1447,239**	0,2202025	1447,146
Random Forest	0,2216543	1452,612	0,2199053	1450,462
GBM	0,2220635	1551,913	0,2204248	1539,457
XGBoost	0,2216571	1448,542	0,2199385	**1444,946**

Pour le **nombre de sinistres**, le modèle GLM reste le meilleur avec une amélioration de la RMSE après ajout des données externes (qui passe de 0.2210493 à 0.2195956). Il est suivi du *Random Forest* qui est également amélioré, sa RMSE passe de 0.2216543 à 0.2199053.

Quant à la prédiction du **coût des sinistres**, ce n'est plus le modèle CART qui ressort comme meilleur mais le modèle **XGBoost** avec un gain en prédiction nettement amélioré (RMSE est égale à 1444.946). Il est ensuite suivi de CART dont la RMSE vaut 1447.146. Le GLM ne vient qu'en $3^{ème}$ position avec une RMSE égale à 1449.171.

Importance des variables pour les modèles du nombre de sinistres après intégration des données télématiques

Nous représentons l'importance des variables (**en bleu**) de chaque modèle après intégration des données télématiques. Nous mettons en évidence (**en rouge**) les variables dont la pertinence est la plus importante dans la prédiction.

Il apparaît que pour la modélisation du **nombre de sinistres**, aucune variable télématique ne se distingue par rapport aux variables caractéristiques du contrat de l'assuré et de son véhicule. En effet, bien qu'elles contribuent à l'amélioration des prédictions des modèles, elles ont un degré de pertinence moins élevé que les variables du portefeuille.

GLM	CART	Random Forest	GBM	XGBoost
CRM	**CRM**	CRM	CRM	CRM
Ancienneté du véhicule	Ancienneté du permis B	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule
Ancienneté du permis B	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Genre	Genre	Genre	Genre	Genre
Zone	Zone	Zone	Zone	Zone
Novice	Novice	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Classe_SRA	Classe_SRA	Formule	Formule	Formule
Formule	Formule	Groupe_SRA	Groupe_SRA	Ancienneté du permis B
Groupe_SRA	Groupe_SRA	Ancienneté du permis B	Ancienneté du permis B	Franchise dommage
Age du conducteur	Franchise dommage	Franchise dommage	Franchise dommage	roadDensity
Franchise dommage	roadDensity	roadDensity	roadDensity	Franchise dommage
trunkDistance	trunkDistance	trunkDistance	trunkDistance	trunkDistance
totalDistance	totalDistance	totalDistance	totalDistance	totalDistance
pavedDistance	pavedDistance	pavedDistance	pavedDistance	pavedDistance
crossing	crossing	crossing	crossing	crossing
meanSpeed	meanSpeed	meanSpeed	meanSpeed	meanSpeed
motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance
cobblestoneDistance	cobblestoneDistance	cobblestoneDistance	cobblestoneDistance	cobblestoneDistance
accident	accident	accident	accident	accident

Importance des variables pour les modèles du coût des sinistres après intégration des données télématiques

Nous représentons l'importance des variables (**en bleu**) sur le tableau après intégration des données télématiques. Nous mettons en évidence (**en rouge**) les variables dont la pertinence est la plus importante dans la prédiction.

GLM	CART	Random Forest	GBM	XGBoost
CRM	CRM	CRM	CRM	**Formule**
Ancienneté du véhicule	Ancienneté du permis B	Ancienneté du véhicule	Ancienneté du véhicule	CRM
Ancienneté du permis B	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Genre	Genre	Genre	Genre	Genre
Zone	Zone	Zone	Zone	Zone
Novice	Novice	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Classe_SRA	Classe_SRA	Formule	Formule	Ancienneté du véhicule
Formule	**Formule**	Groupe_SRA	Groupe_SRA	Ancienneté du permis B
Groupe_SRA	Ancienneté du véhicule	Ancienneté du permis B	Ancienneté du permis B	Groupe_SRA
Age du conducteur	Franchise dommage	Franchise dommage	Franchise dommage	populationDensity
Franchise dommage	roadDensity	roadDensity	roadDensity	Franchise dommage
trunkDistance	trunkDistance	roundabout	trunkDistance	traffic_signals
totalDistance	totalDistance	totalDistance	totalDistance	totalDistance
pavedDistance	pavedDistance	traffic_signals	stop	pavedDistance
crossing	crossing	crossing	crossing	crossing
meanSpeed	meanSpeed	meanSpeed	meanSpeed	meanSpeed
motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance
cobblestoneDistance	cobblestoneDistance	trunkdistance	cobblestoneDistance	cobblestoneDistance
accident	accident	accident	roundabout	accident

L'algorithme XGBoost est le modèle faisant intervenir le plus les variables télématiques. Pour les modèles GLM, *Random Forest* et GBM, elles interviennent avec une intensité faible (mises en évidence avec la couleur **orange**). Par ailleurs, la variable représentant la distance totale "totalDistance" apparaît parmi les variables les plus importantes dans CART.

Conclusions générales

Il est vrai qu'aujourd'hui, la majorité des assureurs basent leurs tarifs sur des analyses GLM, les modèles de *machine learning* étant globalement peu déployés. Pourtant, ces méthodes s'avèrent souvent pertinentes, voire parfois plus performantes que les approches classiques. Il serait relativement intéressant de tester les deux familles d'approches lors des revues des tarifs et de déterminer au cas par cas celle qui est la plus pertinente.

Concernant l'apport des données télématiques, ce mémoire montre que l'ajout de données externes potentiellement accessibles à tous les acteurs du marché peut permettre d'améliorer significativement la pertinence d'un tarif.

Limites / axes d'amélioration

Il serait intéressant d'intégrer des données télématiques supplémentaires comme un indicateur de sinistralité par commune ou encore, d'effectuer une classification non-supervisée pour définir un zonier sur le risque relatif à chaque commune et intégrer uniquement cette variable au portefeuille.

Par ailleurs, cette étude a été appliquée à un portefeuille RC automobile, il serait alors judicieux d'envisager d'autres garanties.

Enfin, la distinction entre les sinistres corporels et matériels n'a pas été relevée dans ce mémoire, une éventuelle piste d'amélioration pourrait être la distinction entre les deux.

Executive summary

In a sector as competitive and saturated as the car insurance market, the insurer is required to establish a precise pricing of its portfolio in order to contain the inherent risks to its activity. This pricing must be based on a fine segmentation of the portfolio to fight against adverse selection. Thus, an insurer that optimizes its pricing will be able to attract new policyholders by offering a fair price while maintaining its technical balance. Under-pricing or over-pricing would inevitably lead to adverse selection, which consists of driving away the good risks to keep only the bad ones in the portfolio.

For many years, Generalized Linear Models (GLM) have been a key tool in pricing and are widely deployed within insurance companies. Today, the world of data is experiencing a profound revolution with the explosion of massive databases. A large number of algorithms has been set up, which objective was either related to the prediction of values or to the classification of individuals. This work gave birth to the theory of statistical learning. However, the enthusiasm for these new innovative methods is accompanied by mixed conclusions about their operational application and interpretability and they are often referred to as *« black box»* methods. Moreover, as the performance of the GLM is generally good, the implementation of statistical learning methods within the pricing teams is narrowed.

Furthermore, in order to establish a pricing model, it is essential to retain the best explanatory variables of the claims. While the tariffs were based on a small number of variables communicated by the insured party, it is now easily possible to access a large panel of external variables able to enrich the existing models and strengthen their predictive scope.

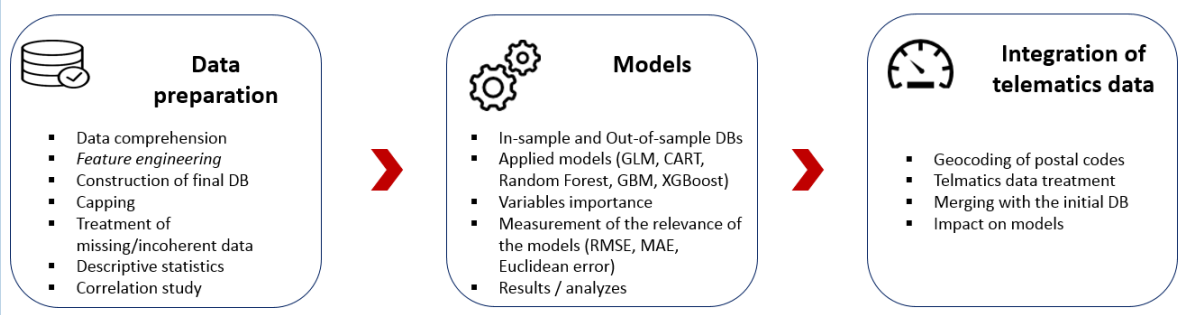
Objectives of the thesis :

This thesis proposes to examine two main issues :

- The implementation on a car insurance portfolio of a GLM approach and a panel of *Data Science* algorithms to evaluate the relevance of these and compare their efficiency in the modeling of **claims number** and **claims cost**.
- The study of telematics data contribution on the predictions of the **claims number** and **claims cost**. This data was retrieved from a company specializing in driver behavior.

Adopted approach :

The study is then built around the following steps :



Modeling the claims number and claims cost without the use of external data

The study is based on data from the portfolio of a french automobile liability insurance. Before starting the modeling, a first step consisted in the preparation of an operable database. This was done by adapting the structure of the databases "contracts" and "claims" to bring them together to obtain a final database on which our models will be constructed. The "contracts" database contains all information relating to the driver's contract and his vehicle. The "claims" database has all the information related to the claims history. Several preliminary studies were conducted to assess the quality of the database and identify possible correlations between tariff variables that could bias the GLM.

First of all, a GLM was performed to model separately the **claims number** and **claims cost**. It will be our reference model, and the performances of statistical learning models will be compared to it, in particular through error measures such as RMSE, MAE and the Euclidean error.

The *Data Science* models tested in this study are :

— CART decision trees.

— *Random Forest* ;

— *Gradient Boosting Machine (GBM)* ;

— *eXtreme Gradient Boosting (XGBooost).*

As a first step, these models will be applied without calibrating their parameters. Then, they will be optimized on the in-sample database by testing several combinations of multiple

values of hyper-parameters. The performances of the models thus constructed are compared on the out-of-sample database with those of the GLM previously established in terms of RMSE. The results are given in the following table :

	Claims number	Claims cost
GLM	**0,2210493**	1449,768
CART	0,2217749	**1447,239**
Random Forest	0,2216543	1452,612
GBM	0,2220635	1551,913
XGBoost	0,2216571	1448,542

Through the built models, we can see that the tested statistical learning methods offer performances comparable to those of the GLM. Indeed, whether for the **claims number** or the **claims cost**, we notice that the RMSE are close.

Regarding the modeling of **claims number**, we found that the model with the smallest error (RMSE = 0.2210493) is indeed the **GLM**, it is followed by **Random Forest** with an RMSE value equal to 0.2216543.

As for the modeling of **claims cost**, the best model is **CART** with an RMSE equal to 1447.239, followed by **XGBoost** which RMSE equals 1448.542.

Next, we highlighted the variables that contribute most to each model.

Variables importance for the claims number models

The following table summarizes the importance of the variables that emerge in each model used for **claims number**, they are highlighted in **blue**. In addition, the variables that appear first in each model are shown in **red**.

GLM	CART	Random Forest	GBM	XGBoost
CRM	**CRM**	CRM	CRM	**CRM**
Zone	Zone	Zone	Zone	Zone
Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B
Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule
Genre	Genre	Genre	Genre	Genre
Formule	Formule	Formule	Formule	Formule
Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage
Classe_SRA	Classe_SRA	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA
Novice	Franchise auto	Franchise auto	Franchise auto	Franchise auto

To conclude, it would be better to keep the GLM model for modeling the **claims number**. This model offers good performance and is particularly easy to interpret.

Variables importance for the claims cost models

As with the **claims number**, we summarize the importance of the variables for each model in the following table. The most important variables that contribute to the construction of the models are highlighted with **blue**. Variables that appear first in models are shown in **red**.

GLM	CART	Random Forest	GBM	XGBoost
CRM	CRM	CRM	CRM	CRM
Zone	Zone	Zone	Zone	Zone
Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B
Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule
Genre	Genre	Genre	Genre	Genre
Formule	**Formule**	Formule	Formule	**Formule**
Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage
Classe_SRA	Classe_SRA	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA
Novice	Franchise auto	Franchise auto	Franchise auto	Franchise auto

A potential improvement of these predictions would be to take into account external data. Indeed, these variables capture complementary information that wasn't communicated at the time of the subscription of the policy. It is in this context that we will integrate data from telematics into our portfolio, in order to measure their contribution to modeling the **claims number** and **claims cost**.

Modeling the claims number and claims cost after the integration of telematics data

As mentioned earlier, with the advent of *Big Data* and the increasingly available data sources and the development of connected tools, insurers can enrich their tariffs with external data. This data enables insurers to finer their risk analysis.

For this thesis, and as part of a partnership with a company named **Ellis Car** specialized in driver behavior, the consulting firm Galea & Associés managed to recover data related to telematics to study their impact on the predictions of claims.

Pretreatment of this data has been done to address the issue of missing data.

A comparison of models performance based on RMSE is given in the following table :

	Without telematic data		**With telematic data**	
	Claims number	**Claims cost**	**Claims number**	**Claims cost**
GLM	**0,2210493**	1449,768	**0,2195956**	1449,171
CART	0,2217749	**1447,239**	0,2202025	1447,146
Random Forest	0,2216543	1452,612	0,2199053	1450,462
GBM	0,2220635	1551,913	0,2204248	1539,457
XGBoost	0,2216571	1448,542	0,2199385	**1444,946**

For **claims number**, the GLM remains better with an improvement of the RMSE after adding the external data (it goes from 0.2210493 to 0.2195956). It is followed by the *Random Forest* which has also improved, its RMSE goes from 0.2216543 to 0.2199053.

As for the prediction of **claims cost**, it is no longer the CART model the better one but the **XGBoost** model with a much improved prediction gain (the RMSE is equal to 1444.946). It is then followed by CART which RMSE equals 1447.146. The GLM only comes in $3rd$ position with an RMSE equal to 1449.171.

Variables importance for claims number models with the integration of telematics data

We represent the importance of the variables for each model after integration of telematics data (in **blue**). We highlight (in **red**) the variables which relevance is most important in the prediction.

It appears that for the modeling of claims number, no behavioral variable stands out from the characteristic variables of the policy of the insured and his vehicle. Although they contribute to improve model prediction, they have a lower degree of importance than portfolio variables.

GLM	CART	Random Forest	GBM	XGBoost
CRM	**CRM**	CRM	CRM	CRM
Ancienneté du véhicule	Ancienneté du permis B	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule
Ancienneté du permis B	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Genre	Genre	Genre	Genre	Genre
Zone	Zone	Zone	Zone	Zone
Novice	Novice	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Classe_SRA	Classe_SRA	Formule	Formule	Formule
Formule	Formule	Groupe_SRA	Groupe_SRA	Ancienneté du permis B
Groupe_SRA	Groupe_SRA	Ancienneté du permis B	Ancienneté du permis B	Franchise dommage
Age du conducteur	Franchise dommage	Franchise dommage	Franchise dommage	roadDensity
Franchise dommage	roadDensity	roadDensity	roadDensity	Franchise dommage
trunkDistance	trunkDistance	trunkDistance	trunkDistance	trunkDistance
totalDistance	totalDistance	totalDistance	totalDistance	totalDistance
pavedDistance	pavedDistance	pavedDistance	pavedDistance	pavedDistance
crossing	crossing	crossing	crossing	crossing
meanSpeed	meanSpeed	meanSpeed	meanSpeed	meanSpeed
motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance
cobblestoneDistance	cobblestoneDistance	cobblestoneDistance	cobblestoneDistance	cobblestoneDistance
accident	accident	accident	accident	accident

Variables importance in claims cost models with telematics data

We represent the importance of the variables (**in blue**) on the table after the integration of telematics data. We highlight (**in red**) the variables which relevance is most important in the prediction.

GLM	CART	Random Forest	GBM	XGBoost
CRM	CRM	CRM	CRM	**Formule**
Ancienneté du véhicule	Ancienneté du permis B	Ancienneté du véhicule	Ancienneté du véhicule	CRM
Ancienneté du permis B	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Genre	Genre	Genre	Genre	Genre
Zone	Zone	Zone	Zone	Zone
Novice	Novice	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Classe_SRA	Classe_SRA	Formule	Formule	Ancienneté du véhicule
Formule	**Formule**	Groupe_SRA	Groupe_SRA	Ancienneté du permis B
Groupe_SRA	Ancienneté du véhicule	Ancienneté du permis B	Ancienneté du permis B	Groupe_SRA
Age du conducteur	Franchise dommage	Franchise dommage	Franchise dommage	populationDensity
Franchise dommage	roadDensity	roadDensity	roadDensity	Franchise dommage
trunkDistance	trunkDistance	roundabout	trunkDistance	traffic_signals
totalDistance	totalDistance	totalDistance	totalDistance	totalDistance
pavedDistance	pavedDistance	traffic_signals	stop	pavedDistance
crossing	crossing	crossing	crossing	crossing
meanSpeed	meanSpeed	meanSpeed	meanSpeed	meanSpeed
motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance
cobblestoneDistance	cobblestoneDistance	trunkdistance	cobblestoneDistance	cobblestoneDistance
accident	accident	accident	roundabout	accident

We can see that the XGBoost algorithm is the one involving telematics variables the most. For the GLM, *Random Forest* and GBM models, they operate at a low intensity (highlighted in **orange**). In addition, the variable that represents the total distance "totalDistance" appears among the most important variables in CART.

General conclusions

It is true that today, the majority of insurers only base their tariff on GLM as machine learning models aren't that much deployed. However, these methods are often relevant and sometimes even more performant than conventional approches. It would undoubtedly be interesting to test the two approaches during tariff reviews and determine, on a case-by-case basis, which is the most relevant.

Regarding the contribution of telematics data, this thesis shows that it would be conceivable for the insurer to integrate the behavioral data, since it brings a real gain in terms of predictive capabilities.

Limits / axis of improvement

It would be interesting to integrate additional telematics data such as an indicator of claims by township. Also, it is possible to do an unsupervised classification in order to determine a *zonier* on the relevant risk of each township and add afterwards this variable to the portfolio.

Moreover, this thesis is applied to a car liability insurance portfolio. It would be interesting to consider other insurance products.

Finally, there hasn't been a distinction between injury claims and material claims, so we could improve this point by taking them separatly in consideration.

Remerciements

Je tiens tout d'abord à adresser mes vifs remerciements à Norbert Gautron pour m'avoir accordé sa confiance et m'avoir permis d'effectuer mon mémoire au sein du cabinet Galea & Associés. Je le remercie également pour ses précieux conseils tout au long de ce stage.

Je voudrais également remercier ma tutrice de stage Florence Chiu et Léonard Fontaine pour leur patience, leur encadrement régulier et avisé tout au long de ces travaux et aussi pour leurs nombreuses relectures.

Je souhaite remercier Nathalie Ramos pour son suivi, ses encouragements et ses conseils.

Je remercie Romain Boyer Chammard pour ses idées et ses directives pertinentes.

Je tiens à remercier Franck Vermet, directeur des études à l'EURIA, pour la qualité de ses cours et pour son implication dans mon mémoire, ses précieux conseils m'ont été d'une grande aide pendant ces travaux.

Je remercie tous les consultants du cabinet pour leur accueil chaleureux, et tout particulièrement mes collègues d'îlot : Mélissande Sanchez, Tanguy Aucoin et Florian Cabocel pour leur bonne humeur, leur dynamisme, leur professionnalisme et pour avoir rendu l'environnement de travail chaleureux et agréable tous les jours pendant mon stage.

J'adresse mes plus sincères remerciements à mes parents et mes soeurs qui ont toujours cru en moi, surtout pendant les périodes de doutes. Ce travail est l'expression de ma profonde gratitude.

Enfin, je remercie Ali et Inas qui m'ont soutenue et supportée tout au long de ces travaux et qui m'ont toujours poussée à donner le meilleur de moi-même durant toutes ces années.

Table des matières

Introduction

Le marché de l'assurance automobile est aujourd'hui particulièrement tendu et concurrentiel. En effet, il y a de plus en plus de nouveaux entrants alors que le marché est déjà saturé, notamment par la présence des bancassureurs. Par ailleurs, les coûts des sinistres sont toujours plus élevés. La Loi Hamon de 2015 a constitué un tournant pour le marché de l'assurance automobile. En effet, cette loi stipule que l'assuré a désormais le droit de résilier son contrat d'assurance automobile à tout moment un an après la souscription. Cette mesure donne ainsi plus de flexibilité aux assurés pour changer d'assurance et cela conduit à accentuer la concurrence entre les organismes assureurs, ce qui peut se traduire par une baisse des primes d'assurance en faveur des assurés. Aussi, l'essor des comparateurs en ligne permet aux assurés d'individualiser leurs risques à la recherche du meilleur prix.

Dans ce contexte, l'enjeu pour les différents acteurs est donc double mais contradictoire. Il s'agit de conserver leurs parts de marché tout en maintenant l'équilibre technique. Le processus de tarification apparaît comme l'outil clé pour atteindre ces deux objectifs. Un assureur capable d'optimiser sa tarification pourra attirer des adhérents en leur proposant le juste prix sans pour autant dégrader son ratio combiné. A l'inverse, si la tarification est inadaptée dans le sens où l'assureur sur-tarifie ses contrats, la conséquence serait une perte de son portefeuille. S'il sous-tarifie, il fera face à un déficit technique. Au global, il devient nécessaire à l'assureur d'adopter une tarification précise de son portefeuille qui doit s'appuyer sur une segmentation à la fois poussée et pertinente.

Depuis les années 80, l'état de l'art en matière de tarification automobile consiste à utiliser des modèles paramétriques, à savoir les modèles linéaires généralisés (GLM). Ces approches ont permis d'améliorer sensiblement l'appréhension de la prime pure d'un assuré et sont aujourd'hui déployées chez quasiment tous les acteurs du marché.

Ces dernières années, d'autres approches ont été introduites pour aborder ces questions. Il s'agit des méthodes *data science*. Bien que la littérature scientifique et actuarielle se soit largement penchée sur ces modèles, les approches de *data science* sont aujourd'hui peu déployées opérationnellement dans les équipes de tarification, et ce pour des raisons multiples : la bonne performance des outils GLM existants, le coût d'entrée des nouveaux algorithmes qui nécessitent d'adapter les systèmes d'information des compagnies d'assurance, ou encore l'aspect « boîte noire » des approches de *machine learning*. On peut néanmoins regretter cet état de fait, les approches de *data science* ayant prouvé dans de nombreux cas leur très bonnes capacités

prédictives, en particulier face au flux massif des données disponibles via l'ouverture des données *open source*, des données télématiques et des technologies de collecte de données.

Le premier objectif de ce mémoire est de mettre en oeuvre sur un portefeuille automobile, une approche GLM et un panel d'algorithmes de *data science* pour évaluer la pertinence de ceux-ci et comparer leur efficacité.

La deuxième problématique abordée dans ce mémoire est l'étude des meilleures variables à retenir dans l'établissement d'un tarif. Traditionnellement, les assureurs disposaient de peu de données. Les tarifs reposaient sur un faible nombre de variables fournies par l'assuré comme le coefficient bonus/malus, le type du véhicule, l'ancienneté du permis, la zone géographique etc. L'avènement du *big data* amène à changer cette vision. Les données externes de plus en plus nombreuses sont disponibles et peuvent venir enrichir les modèles existants. Ce point a fait l'objet de plusieurs études en assurance habitation mais a été peu testé en assurance automobile à ce jour. Nous avons alors cherché à tester l'opportunité d'ajouter à un tarif automobile existant des données externes pour établir dans quelle mesure elles permettent d'améliorer son caractère prédictif.

Dans un premier temps, nous rappelons quelques éléments de contexte relatifs à l'assurance automobile, puis nous présenterons l'utilité des nouvelles techniques télématiques en assurance et les aspects réglementaires qui y sont liées.

La seconde section décrit les différents outils de tarification que nous avons retenus pour ce mémoire. Nous présenterons alors les modèles linéaires généralisés (GLM) qui constitueront notre modèle de référence, ainsi que les méthodes de *data science* : les arbres de décision CART, les forêts aléatoires ou *Random Forest*, le *Gradient Boosting Machine* et finalement, l'*eXtreme Gradient Boosting (XGBoost)*. Ces différents outils sont utilisés pour modéliser le **nombre de sinistres** et le **coût des sinistres** d'un portefeuille RC automobile.

La troisième partie présente ce portefeuille, les données disponibles, les traitements effectués et les corrélations entre les variables.

Dans la quatrième partie, nous appliquerons à ces données des outils décrits dans la troisième section et nous comparerons l'efficacité de ces différentes approches.

Enfin, dans une dernière section, nous analyserons l'apport des données externes. Nous nous baserons sur des données fournies par l'entreprise *Ellis Car*, spécialiste en télématique automobile. La base de données d'*Ellis Car* permet, dans chaque zone géographique, d'établir des indicateurs tels que la fréquence de freinages brusques, des accélérations etc. Ces données, potentiellement disponibles pour tous les assureurs, viennent enrichir notre base de données initiale. Nous testons dans quelle mesure elles permettent d'améliorer les modèles établis à la section précédente.

Chapitre 1

Généralités et contexte

1.1 Description du marché de l'assurance automobile en France

1.1.1 Présentation du marché

Le marché de l'Assurance français est considérable et devient le leader européen devant l'Allemagne et l'Italie avec un chiffre d'affaires de 208 milliards d'euros en 2017. L'assurance des biens et responsabilité progresse de 2,3% avec des revenus de 54 Md€ en 2017 contre 53 Md€ en 2016 et une croissance de 2,8% pour l'assurance automobile.

Affaires directes en Md €	2016	2017	Evolution
Cotisations en Assurances de biens et de responsabilité	**53,2**	**54,5**	**2,3%**
- dont Particuliers	33,5	34,5	3,0%
- dont Professionnels	19,7	20	1,2%

FIGURE 1.1 – Cotisations en Assurance de biens et de responsabilité

Affaires directes par branches (Md€)	2016	2017
Automobile	**20,7**	**21,3**
Dommages aux biens	**17,8**	**18,1**
- Particuliers	10,2	10,5
- Professionnels	7,6	7,6
- Agricoles		
Transport	**0,9**	**0,8**
Responsabilité civile générale	**3,6**	**3,6**
Construction	**2,1**	**2,1**
Catastrophes naturelles	**1,6**	**1,6**
Divers (crédits, protection juridique,assistance)	**6,5**	**6,9**
TOTAL	**53,2**	**54,5**

FIGURE 1.2 – Cotisations en Assurance de biens et de responsabilité (par branches)

L'assurance automobile est obligatoire depuis 1958. Cette obligation ne concerne que la garantie « responsabilité civile », qui permet l'indemnisation des dommages causés aux tiers par la faute du conducteur du véhicule ou d'un de ses passagers. Elle possède une place centrale

dans le secteur de l'assurance en France avec un chiffre d'affaires de 21,3 milliards d'euros, donnant lieu à une intense concurrence entre les différentes compagnies. Ce secteur ne cesse de croître, notamment avec l'augmentation croissante du parc automobile. Au 1^{er} Janvier 2017, le parc automobile français était constitué de 39 millions de véhicules, révèle le dernier décompte du CCFA [1]. Cela correspond à 1,2% de véhicules immatriculés supplémentaires par rapport à l'année 2016.

En termes d'accidentologie, le nombre d'accidents ainsi que leur gravité diminuent grâce à l'augmentation de la sécurité routière, à la réglementation accrue mais aussi à l'automatisation croissante des voitures. Le nombre de morts sur les routes a subi une légère diminution entre 2016 et 2017 avec 29 décès en moins. [2]

Bilan de l'année 2017	Accidents corporels	Tués à 30 jours	Blessés	dont blessés hospitalisés
Année 2017	58 613	3 448	73 384	27 732
Année 2016	57 522	3 477	72 645	27 187
Différence 2017 / 2016	1 091	-29	739	545
Evolution 2017 / 2016	1.9%	-0.8%	1.0%	2.0 %

TABLE 1.1 – La mortalité routière entre 2016 et 2017

Cependant, la baisse de la sinistralité [3] n'est pas encore suivie par une baisse des coûts (figure 1.3). Cela est dû à la hausse des frais de main d'oeuvre, des prix des pièces détachées et surtout à la revalorisation des compensations dont bénéficient les victimes d'accidents corporels, augmentant ainsi la charge pour les assureurs.

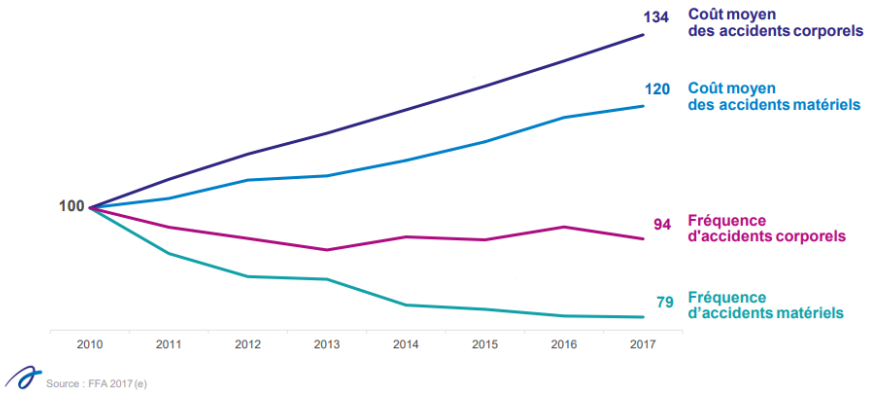


FIGURE 1.3 – Évolution des coûts et fréquences de sinistres sur une base de 100 en 2010

1. Comité des Constructeurs Français d'Automobiles
2. Données ONISR : Observatoire national interministériel de la sécurité routière
3. Source FFA

A plus long terme, les indemnités versées par les assureurs devraient réduire, notamment grâce aux avancées technologiques qui sont en évolution permanente. Ainsi, un engagement a été pris par plusieurs constructeurs automobiles afin qu'il n'y ait plus aucun mort ou blessé dans leurs véhicules d'ici cinq ans (Volvo promet le « zéro décès » à l'horizon 2020). De même, la perspective d'un véhicule 100% autonome se concrétise petit à petit, Google, Uber, et Tesla étant les constructeurs les plus avancés. Les programmes de ce type se multiplient également au sein des constructeurs traditionnels ou des équipementiers automobiles. Cependant, cela implique de réfléchir aux nouvelles problématiques soulevées par ces futures technologies. Ainsi, l'émergence des véhicules intelligents va contraindre les pouvoirs publics à repenser les règles de circulation ainsi que les régimes de responsabilité. Par exemple, en cas d'accident, un changement des lois en vigueur est envisagé afin de transférer la responsabilité du conducteur au constructeur.

Globalement, la sinistralité de la branche automobile s'améliore avec un ratio combiné en baisse de 3,4 points (figure 1.4). Le ratio reste cependant structurellement supérieur à 100 %. L'assurance automobile apparaît comme un marché difficile.

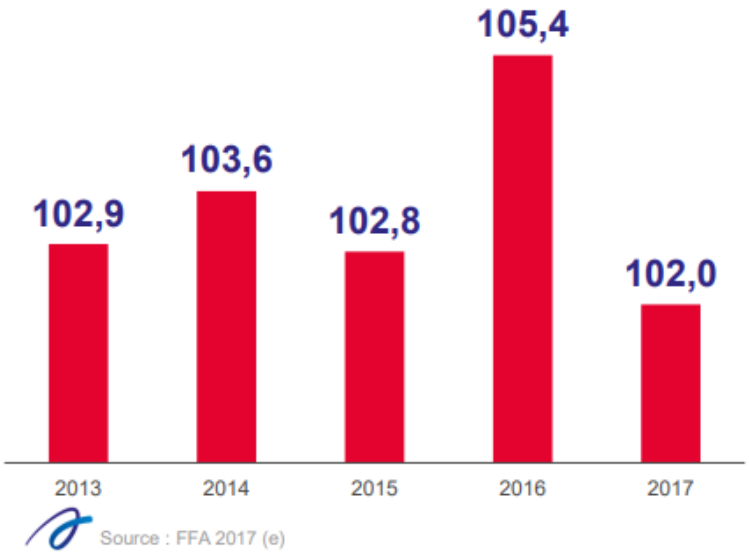


FIGURE 1.4 – Évolution du ratio combiné en assurance automobile

De plus, la compétitivité du secteur ne fait qu'évoluer avec le développement de l'assurance directe, l'émergence des voitures connectées et la concurrence entre les sociétés d'assurance. Pour cela, l'assureur est tenu de trouver la meilleure approche pour segmenter le plus justement possible son offre et de capter au maximum les « bons » risques en affinant ses modèles tarifaires, notamment en incluant de nouvelles variables par exemple.

1.1.2 La garantie responsabilité civile en assurance automobile

1.1.2.1 Définition de la responsabilité civile

La responsabilité civile relève du domaine du droit qui vise à réparer le non respect d'une obligation ou d'un devoir envers autrui. Traditionnellement, on distingue la responsabilité civile délictuelle de la responsabilité civile contractuelle.

La responsabilité civile délictuelle :

Elle est engagée lorsqu'il s'agit d'un fait juridique, c'est-à-dire un événement susceptible de produire des effets de droit (un accident par exemple). Elle se fonde sur l'obligation de réparer un dommage selon les articles 1240 et 1241 du Code Civil en cas :

- de faute : *"tout fait quelconque de l'homme, qui cause à autrui un dommage, oblige celui par la faute duquel il est arrivé, à le réparer."*[4]
- de négligence, d'omission, ou d'imprudence :*"Chacun est responsable du dommage qu'il a causé non seulement par son fait, mais encore par sa négligence ou par son imprudence."*[5]

Exemple : Une personne "A" fait des travaux sur son toit, elle pose son marteau dont elle n'a plus besoin momentanément, sauf que la zone où elle l'a posé est en pente. Le marteau tombe et assomme une personne "B". "B" est alors en droit de demander une indemnisation à "A" en raison d'un préjudice subi.

La responsabilité civile contractuelle

Elle est engagée lorsqu'il s'agit d'un acte juridique, c'est-à-dire une manifestation de volonté destinée à produire des effets de droit (un contrat par exemple). Elle est entraînée en cas de mauvaise exécution ou d'inexécution totale ou partielle des obligations nées d'un contrat. Selon l'article 1147 du Code Civil : *""Le débiteur est condamné, s'il y a lieu, au paiement de dommages et intérêts, soit à raison de l'inexécution de l'obligation, soit à raison du retard dans l'exécution, toutes les fois qu'il ne justifie pas que l'inexécution provient d'une cause étrangère qui ne peut lui être imputée, encore qu'il n'y ait aucune mauvaise foi de sa part."*

Exemple : Une personne "A" souhaite faire repeindre son appartement, pour cela elle engage un peintre "B". L'obligation de "B" est de refaire les peintures selon ce qui a été convenu dans le contrat (couleur de la peinture par exemple). L'obligation de "A" est de payer "B". Dans cet exemple, la responsabilité de "A" ou du peintre "B" peut être engagée si l'une des parties ne remplit pas ses obligations.

1.1.2.2 La responsabilité civile en assurance automobile

En assurance automobile, la garantie responsabilité civile est la garantie minimale proposée à un conducteur qui cherche à s'assurer. Elle est obligatoire depuis 1958, et a pour but de réparer

4. Article 1240 du Code Civil
5. Article 1241 du Code Civil

les dommages matériels (dégâts causés à un véhicule) ou corporels (blessures de la victime) que le véhicule du conducteur pourrait causer à des tiers. Par ailleurs, cette garantie indemnise les passagers du conducteur responsable, quel que soit le lien qu'ils ont avec lui.

Les voitures concernées :

Les propriétaires d'un véhicule terrestre à moteur sont dans l'obligation de l'assurer pour pouvoir le faire circuler. Ce véhicule peut être :

- Une voiture (particulière, utilitaire ou sans-permis) ;
- Un 2 ou 3 roues (moto ou scooter) ou un quad, même non-homologué (une mini-moto par exemple) ;
- Une tondeuse auto-portée avec un siège permettant au conducteur de manœuvrer l'engin.

Les sanctions en cas de défaut d'assurance :

Les risques encourus dans le cas où le véhicule n'est pas assuré sont :

- Une amende de 3 750 euros ;
- Une suspension de permis de conduire (jusqu'à 3 ans) ;
- L'annulation du permis de conduire et l'interdiction de le repasser pendant 3 ans (au plus) ;
- L'interdiction de conduire certains véhicules, même s'ils ne nécessitent pas le permis de conduire ;
- Une peine de travail d'intérêt général.

1.2 Les télématiques dans la personnalisation du tarif

1.2.1 Le contexte des télématiques en assurance

La télématique de véhicule est la technologie d'enregistrement, d'envoi, de réception et de stockage des informations via des appareils de télécommunication. Cette technologie a de nombreuses applications, notamment dans l'assurance automobile à travers l'assurance dite *Blackbox* ou par les offres de *Pay As/How You Drive (PHYD)* ou *Usage Based Insurance (UBI)*. L'intérêt des compagnies pour cette technologie est d'autant plus grand qu'elle permet de mieux évaluer les risques, et de pouvoir effectuer un travail de prévention beaucoup plus poussé auprès de leurs clients.

D'après l'étude menée par *Market Insights Reports*, l'utilisation de la télématique dans le secteur de l'assurance devrait connaître une croissance fulgurante dans les années à venir. Dans son rapport intitulé *« Global Insurance Telematics Market »*, le cabinet de recherche signale que le marché de l'assurance télématique évalué en 2017 à 1,05 Md USD devrait atteindre 5,83 Md USD d'ici 2025, soit un taux de croissance annuel de 21%.

L'*Usage Based Insurance (UBI)* est une innovation récente développée par les assureurs automobiles afin d'adapter la prime en fonction des comportements de conduite des conducteurs.

Le kilométrage ainsi que ces comportements sont scrutés en utilisant un odomètre ou des boîtiers de télécommunication (télématiques) qui sont installés par le conducteur dans un port spécifique du véhicule, comme des terminaux qui se branchent sur l'allume-cigare et font aussi office de chargeurs, pour encourager les assurés à les utiliser. D'autres boîtiers sont déjà installés dans le véhicule par les constructeurs. L'idée principale des télématiques dans l'assurance automobile est de contrôler en temps réel le comportement du conducteur. Ces boîtiers permettent d'évaluer la qualité de conduite en enregistrant des paramètres de vitesse, d'accélération, de freinage ou encore de virage. La compagnie d'assurance analyse les données grâce à des programmes d'apprentissage statistique et calcule une prime personnalisée en fonction des informations collectées. Dans son rapport d'étude de 2013 sur l'*UBI*, le groupe de conseil en stratégie Ptolemus a estimé que d'ici 2020, l'assurance télématique représenterait plus de 35 millions de polices d'assurance. [6].

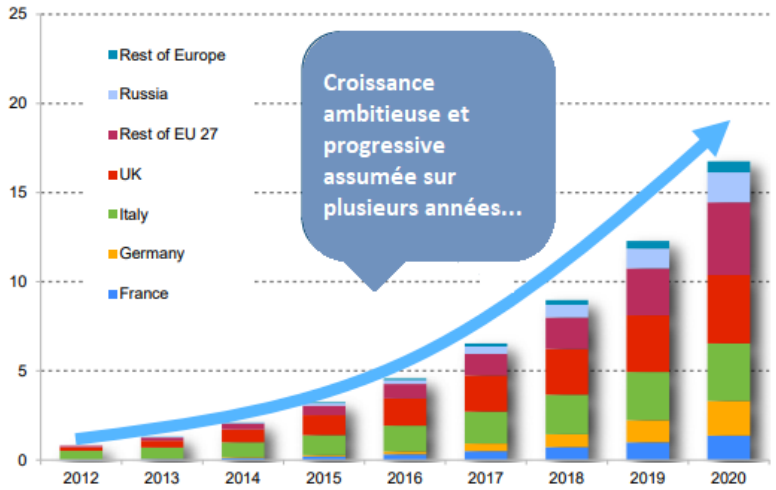


FIGURE 1.5 – Croissance attendue des offres de *Usage Based Insurance*

Les premiers programmes de l'*Usage Based Insurance* ont commencé à se développer aux Etats-Unis il y a une décennie, quand Progressive et General Motors Assurance Company (GMAC) ont lancé leurs offres proposant un rabais en fonction du kilométrage combiné à une technologie GPS et des systèmes cellulaires qui collectent les kilomètres parcourus. Ces offres étaient (et sont toujours) souvent associées à des prestations supplémentaires comme une assistance routière et la récupération de véhicules volés. Les évolutions récentes de la technologie ont augmenté l'efficacité des télématiques, permettant aux assureurs de mesurer le nombre de kilomètres parcourus par les assurés ainsi que leur style de conduite . Il en a résulté l'émergence de plusieurs offres variées de l'*Usage Based Insurance*, comme le *Pay As You Drive (PAYD), Pay How You Drive (PHYD), Pay As You Go, et Distance Based Insurance.*

En France, *Amaguiz* de Groupama et *IDMacif* de Macif ont été les premiers à se lancer dans le secteur de l'automobile connectée permettant au conducteur de prendre en compte

6. Ptolemus USAGE-BASED INSURANCE Global Study 2013

uniquement le kilométrage parcouru. Depuis, Direct Assurance a lancé son application mobile *You Drive* basée sur le *PAYD*, de même que Axa avec son application mobile *AxaDrive*. Depuis Avril 2018, pour une durée de 15 mois, la Matmut a lancé un test de télématique embarquée *« Matmut Connect Auto »* auprès de 2000 volontaires parmi ses assurés automobile. Il s'agit d'une expérimentation qui s'appuie sur la technologie et l'expertise du groupe Michelin, l'offre DDI *(Driving Data to Intelligence)* qui permet d'analyser les comportements de conduite. La Matmut est un acteur engagé de la sécurité routière et compte utiliser ce dispositif afin de *« rendre la route plus sûre.» « C'est pourquoi l'objectif principal est de fournir au fur et à mesure aux assurés participant des conseils personnalisés, par exemple sur les thèmes de l'anticipation et de l'adaptation aux conditions... et ce afin de les aider à améliorer leur conduite.»*

Au niveau européen, les italiens sont parmi ceux qui payent le plus cher leur police d'assurance auto du fait de la présence de forts niveaux d'aléa moral et de fraudes. Grâce à la télématique, les assureurs italiens peuvent proposer une assurance moins chère et mieux segmentée.

En France, l'aléa moral est moins marqué et les primes sont moins élevées. Toutefois, on peut retrouver le contexte du marché italien sur la tranche des jeunes conducteurs : du fait de leur faible expérience en risque automobile, ils sont amenés à payer des primes plus chères et l'aléa moral est plus fort. L'assurance *YouDrive* de Direct Assurance propose une prime qui peut varier de +10 à -50% dès le premier mois en fonction de la qualité de conduite. Elle cible les conducteurs qui ont moins de 7 ans de permis, qui représentent une clientèle très hétérogène dont l'âge peut aller de 20 à 35 ans, voire plus.

Un autre frein qui bloque l'évolution du *Pay As You Drive (PAYD)* en France est la réglementation. En 2005, la MAAF a lancé une offre *PAYD* qui s'est vue stoppée par une décision de la *CNIL (Commission Nationale Informatique et Liberté)*. Cette décision est dûe au fait que le contrat proposé prévoyait la géolocalisation permanente des conducteurs ainsi que la détention illégale de données comme les dépassements de vitesses maximales autorisées, que seule l'autorité publique peut avoir en possession.

Si la croissance des offres télématiques sur le marché français des particuliers n'est pas aussi significative que celle observée en Italie, les perspectives sont tout autre sur le marché des flottes automobiles. En effet, dans la plupart des cas, le véhicule n'appartenant pas au conducteur et la police d'assurance n'étant pas souscrite par lui, il est donc moins enclin à prendre soin de son véhicule. Par conséquent, le risque d'aléa moral est plus élevé et une offre télématique pourrait notablement réduire ce risque en responsabilisant le conducteur. Aujourd'hui, il existe une panoplie d'offres télématiques sur le marché des flottes (*SoFleet, Fleetmatics, Quartix, Ocean, etc.)* mais rares sont celles qui sont couplées à une offre d'assurance.

1.2.2 Avantages et respect de la vie privée

Les avantages :

Les programmes *UBI* offrent plusieurs avantages aux assureurs, aux assurés et à la société. On en cite quelques uns sur le tableau 1.2.

Société	Assureurs	Assurés
• Baisser la fréquence des accidents et la sévérité	• Corriger les erreurs dans la classification des risques	• Réduire les primes
• Localiser les véhicules volés	• Renforcer l'exactitude des prix	• Adopter un comportement prudent à la suite d'un accident
• Etablir la faute pour une meilleure équité dans le règlement de sinistres	• Attirer des risques favorables	• Profiter des services à valeur ajoutée comme la surveillance des conducteurs adolescents ou des services d'urgence
• Réduire le temps de réponse lors de la survenance d'un accident	• Réduire les coûts de sinistres	

TABLE 1.2 – Les avantages à adopter les télématiques

Le fait de lier étroitement les primes d'assurance au véhicule individuel ou à la flotte automobile permet aux assureurs de proposer une tarification plus adaptée. Ceci permet une meilleure accessibilité financière aux conducteurs à faible risque dont la plupart sont aussi des conducteurs à faible revenu. Par ailleurs, cela donne la possibilité aux assurés d'avoir un contrôle sur le prix de leurs primes en les incitant à réduire les kilomètres parcourus et à adopter un comportement plus vigilant sur la route. En effet, une conduite plus sûre et peu de kilomètres parcourus contribuent à réduire le nombre des accidents, la congestion routière et l'émission de CO2 des véhicules, ce qui est bénéfique à la société.

L'usage des télématiques aide les assureurs à estimer plus rigoureusement les coûts des sinistres et à baisser la fraude en leur permettant d'analyser les données relatives à la conduite (les freinages brusques, la vitesse et le temps) lors de la survenance d'un accident. Ces données supplémentaires peuvent également être utiles aux assureurs pour affiner ou distinguer les produits *UBI*. Enfin, cela peut aussi permettre aux flottes de déterminer les trajets les plus efficaces, dans le but de rationaliser des coûts considérables relatifs au personnel, à l'essence et à l'entretien des véhicules.

Vie privée :

Un des reproches récurrents fait à la télématique porte sur la confidentialité des données. En effet, une partie des assurés est réticente à l'idée de partager les informations les concernant avec les assureurs, que ce soit leur comportement au volant (comme la vitesse ou le freinage) ou les données relatives à leur position. Cependant, les services basés sur la localisation attirent de plus en plus les jeunes générations technophiles. Les smartphones, les appareils GPS et les appareils de télépéage sont répandus dans plusieurs pays.
Pendant que des conducteurs ignorent comment leurs données sont collectées et utilisées, d'autres sont au courant et n'y voient pas d'inconvénients en échange de services qui leur sont profitables. Comme l'a expliqué en 2016, Arthur Dutel : *« En France comme en Europe, 28% des automobilistes sont prêts à livrer leurs données, mais attendent 25% de réduction de leur prime en retour, ce qui n'est pas techniquement possible».* Dans d'autres cas, des automobilistes acceptent les services basés sur la localisation à condition que les données collectées ne soient pas utilisées contre eux.

Pour les flottes automobiles, des considérations de confidentialité importantes sont à prendre en compte. Par exemple, les données de localisation d'une flotte peuvent contenir des informations concurrentielles sensibles. Toutefois, les problèmes de confidentialité sont différents lorsque le conducteur est un employé qui est tenu de respecter les politiques de l'entreprise comme condition de son recrutement.

Les informations obtenues par les objets connectés doivent être manipulées avec la plus grande précaution. Leur utilisation soulève de nombreux problèmes. En effet, les français sont de plus en plus préoccupés par la protection de leurs données personnelles, en particulier avec l'essor de la communication par voie numérique. Le secteur de l'assurance n'y échappe pas car le recueil d'informations constitue le coeur même de l'exercice du métier. Il est à souligner que les objets connectés représentent des indicateurs de risque et un outil de ciblage marketing. L'assurance connectée devient un vecteur de performances dans la mesure où l'appropriation par les assurés des objets connectés modifiera la relation avec l'assureur. Pour cela, il est nécessaire de connaître les limites de la collecte, la légitimité de l'accès et le traitement des données. Dans le paragraphe suivant, nous allons revoir quelques aspects de la loi actuelle.

1.2.3 La protection des données, enjeu de taille pour l'assurance

Le déploiement des véhicules connectés constitue une source potentielle de progrès pour la mobilité et la sécurité routière comme nous l'avons cité dans les paragraphes précédents. Toutefois, ces systèmes doivent présenter les meilleures garanties possibles en termes d'intégrité et de sécurité numérique afin de respecter la confidentialité des données à caractère personnel des usagers. En Europe, la protection de ces données est largement affirmée par le droit.

Tout d'abord, il est essentiel de rappeler ce que l'on entend par « donnée à caractère personnel ».

On appelle une donnée à caractère personnel (DCP) [7] toute donnée susceptible de permettre l'identification d'une personne, directement ou indirectement. Elle entraîne sous cette qualification les données qui lui sont associées, imposant leur protection.
La *Commission Nationale de l'Information et des Libertés, CNIL*, est une autorité chargée à veiller à la protection de ces données. Elle dispose d'un pouvoir de contrôle sur place et de sanction administrative et analyse les conséquences des nouveautés technologiques sur la vie privée. Soucieux de contrôler les échanges et le stockage de données, le milieu assurantiel est à la pointe dans ce domaine. La *CNIL* a établi un « Pacte de Conformité- Assurance » [8] présenté en Novembre 2014 à l'ensemble de la profession. Ce pack dote le secteur de l'assurance d'une méthode de travail et apporte un nouveau mode de régulation à l'autorité en charge de veiller à la protection des données personnelles. Il est composé d'une part des « autorisations uniques » pour lesquelles les organismes sont tenus de prendre un engagement de conformité en ce qui concerne la collecte des données, et d'autre part des « normes simplifiées » qui permettront aux organismes de faire une déclaration simplifiée de conformité auprès de la CNIL. Les informations liées à la vie privée des automobilistes telles que la géolocalisation du véhicule qui permet la connaissance des lieux fréquentés par le conducteur ou encore ses habitudes de consommation sont souvent utilisées par les constructeurs dans le but d'affiner et de cibler leur offre. Cette utilisation sans que le client en ait conscience, peut représenter une intrusion dans sa vie privée.

Ainsi, afin de protéger les données personnelles des automobilistes, la présidente de la CNIL a présenté en Octobre 2017 un « Pack de Conformité : Véhicules connectés et données personnelles » [9]. Ce pack complète le règlement européen de Mai 2018.

Cette boîte à outils rappelle :

— Les types de données personnelles appliquées au véhicule connecté ;
— La définition du traitement des données et les nouvelles dispositions s'y appliquant ;
— Le principe de détermination du responsable de traitement ;
— Les droits des usagers des véhicules connectés basés sur ce principe : les données provenant du véhicule et de son habitacle appartiennent à l'utilisateur.

Le 25 Mai 2018, le règlement général sur la protection des données (RGPD) a été mis en application et a comme enjeu de réguler l'usage commercial des données personnelles. Ce règlement européen rend le pouvoir au client et renforce les droits des individus à contrôler leurs propres données. Il impose de nouvelles règles de sécurisation et d'utilisation des données :

7. RGPD, art.4 al1 : Est une DCP *« toute information se rapportant à une personne physique identifiée ou identifiable ; est réputée être une « personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale ».*
8. https ://www.cnil.fr/sites/default/files/typo/document/PACK_ASSURANCE_complet.pdf
9. https ://www.cnil.fr/sites/default/files/atoms/files/pack_vehicules_connectes_web.pdf

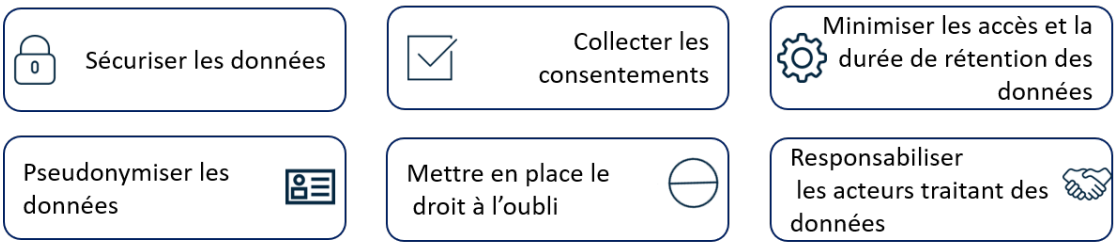


FIGURE 1.6 – Les règles de sécurisation et d'utilisation des données par le RGPD

Par ailleurs, le RGPD fait particulièrement référence aux données de localisation [10] comme donnée identifiante. Cette donnée permet d'étudier le risque routier, elle prend alors une dimension explicitement identifiante, ce qui n'était pas directement le cas dans la législation. D'ailleurs, dans le « Pack de Conformité : Véhicules connectés et données personnelles », parmi les données considérées comme données à caractère personnel : les données de géolocalisation, les données techniques liées à l'état du véhicule et des pièces, les données biométriques du conducteur et les données liées à l'utilisation du véhicule par le conducteur ou les occupants. On peut citer comme exemple de données à caractère personnel relatives au conducteur :

- Les données qui permettent d'identifier son mode de conduite comme l'action sur le frein, sur le clignotant, l'activation et l'utilisation d'une aide ou pas ;
- Les données qui permettent de déterminer ses déplacements : trajets habituels, lieux fréquentés etc ;
- Les données comme la position, la direction et la vitesse.

Considérant le nombre d'éléments importants qui sont susceptibles de transmettre des données du conducteur et par conséquent l'identifier, la réglementation et les autorités de régulation (CNIL) insistent sur les techniques de confidentialité. Pour cela, l'anonymisation ou à défaut, lorsque certaines données sont nécessaires pour atteindre les finalités fixées, la pseudonymisation, sont fortement encouragées par la réglementation.

L'anonymisation :

L'anonymisation consiste à modifier le contenu ou la structure des données de façon irréversible de sorte à ce qu'il soit impossible de ré-identifier les personnes même après traitement, les données perdent ainsi la qualification de DCP. Cependant, cette méthode reste compliquée à appliquer car plus le volume de données croît, plus il devient possible de ré-identifier une personne par ses comportements. Or la CNIL précise que « pour qu'une solution d'anonymisation soit efficace, elle doit empêcher toutes les parties d'isoler un individu dans un ensemble de données, de relier entre eux deux enregistrements dans un ensemble de données (ou dans deux ensembles de données séparés) et de déduire des informations de cet ensemble de données ».

10. RGPD, art.4

Par conséquent, certaines entreprises optent pour une solution qui représente un compromis : la pseudonymisation.

La pseudonymisation :

Introduit par le RGPD, le concept de pseudonymisation ne rend pas les données complètement anonymes ni complètement identifiables non plus. En effet, elle consiste à séparer les données de leurs propriétaires respectifs afin que tout lien avec une identité ne soit possible sans une information supplémentaire. Il est à noter que la pseudonymisation des données a un point faible qui réside dans le fait de générer une clé d'identification qui permet d'établir un lien entre les différentes informations des personnes. Ces clés d'identification doivent être stockées avec un contrôle d'accès performant. En effet, si une clé d'identification est mal protégée, cela engendrerait des risques que les conducteurs soient ré-identifiés par des tiers non légitimes.

Enfin, les entreprises peuvent opter pour les techniques d'anonymisation ou de pseudonymisation selon leurs besoins mais aussi selon la nature des données collectées.

1.3 La tarification non vie dans la littérature actuarielle

1.3.1 La nécessité de segmenter

L'assurance repose fondamentalement sur l'idée que la mutualisation des risques entre des assurés est possible. Cette mutualisation n'a de sens qu'au sein d'une population de risques homogènes. C'est pour cela que les assureurs sont contraints de segmenter leurs portefeuilles. Tout d'abord, il est essentiel de rappeler la définition de la segmentation. La segmentation consiste à analyser et contrôler l'adaptation des primes aux sinistres suivant des classes de risques homogènes, de façon à en tirer des conséquences du point de vue technique. La segmentation permettra de prendre des mesures techniques à chacun des niveaux de segmentation, tant en tarification, qu'en souscription.

La segmentation des contrats d'assurance est essentielle du point de vue technique. Elle permet d'attribuer une prime plus élevée par exemple pour les assurés qui ont eu des sinistres supérieurs en montant ou en nombre, ce qui permet ainsi de responsabiliser chaque assuré. D'autre part, dans une vision économique, la segmentation permet à l'assureur d'acquérir un avantage concurrentiel lorsque son tarif est basé sur le risque. Cela lui permet de plus d'éviter le phénomène d'anti-sélection. Dans un marché aussi concurrentiel que le marché de l'assurance automobile, une insuffisance ou une absence de segmentation exposerait justement les assureurs à ce phénomène. Les bons risques auraient tendance à fuir vers les concurrents qui distingueraient ces profils dans leur tarification et leur proposeraient un tarif plus avantageux. Les mauvais risques quant à eux, resteraient dans le portefeuille car ils estimeraient que le tarif qui leur serait appliqué ailleurs serait moins avantageux.

Afin de mieux comprendre ce concept de segmentation de tarifs, nous allons l'illustrer sur un exemple simple de deux assureurs A et B. L'assureur A décide de faire payer plus cher les conducteurs à risque qui sont déterminés à l'aide d'informations collectées à la souscription, comme les informations sur le conducteur ou son véhicule ; en contrepartie, il diminue la prime pour les conducteurs à faible risque. L'assureur B quant à lui, ne fait aucune segmentation. Il réalise alors des profits avec les conducteurs à faible probabilité de sinistre, et des pertes avec les autres. Les conducteurs prudents ayant une faible probabilité d'avoir un sinistre (les bons risques) iront s'assurer auprès de A car il propose des primes plus faibles. Par contre, comme l'assureur B propose un tarif unique, il n'attirera vers lui que les conducteurs ayant un comportement à risque (les mauvais risques) et donc, une forte probabilité de sinistre. Ces mauvais conducteurs trouveront que le tarif proposé par B est bien plus intéressant que celui proposé par A. L'assureur B ne conservera donc que les mauvais conducteurs pour lesquels son tarif est inssufisant.

Pour conclure, afin de lutter contre l'anti-sélection, les assureurs ont intérêt à segmenter leur tarif pour diminuer l'hétérogénéité des risques au sein de chaque classe. L'assuré devra alors payer une prime plus en adéquation avec son niveau de risque qui correspondra au coût du risque moyen de la classe à laquelle il appartient. En revanche, il est à noter qu'il n'existe pas de bonne ou mauvaise segmentation. Il convient de rester prudent et de ne pas segmenter à l'extrême notamment pour les raisons suivantes :

— Les classes construites seront sans doute plus homogènes mais ne contiendront pas assez d'effectif, ce qui pénalise les opérations de mutualisation, le tarif sera par ailleurs peu robuste ;

— D'un point de vue commercial, une segmentation très fine peut rendre le tarif moins lisible par le réseau commercial, ce qui peut rallonger considérablement la durée de l'entretien de vente causant ainsi un handicap commercial.

Au final, le degré de segmentation que retient l'assureur dépend essentiellement :

— Des pratiques de marché, dans le sens où, si un critère est largement utilisé, l'assureur doit autant que possible l'inclure lui aussi dans sa segmentation afin d'éviter l'anti-sélection ;

— De la volonté à conserver un degré de mutualisation des risques importants sur certaines catégories d'assurés. En effet, un assureur peut choisir selon son implication sociale de ne pas segmenter selon un critère donné.

1.3.2 État de l'art

Au sein des compagnies d'assurance non-vie, les méthodes de tarification s'appuient généralement sur des modèles économétriques. Il s'agit de caractériser la relation de dépendance entre la sinistralité d'un portefeuille et différents facteurs de risque liés à l'assuré ou à son véhicule. L'approche de la prime pure chargée est la plus utilisée par les assureurs, et sera détaillée par la suite dans le paragraphe 2.1.3. Une étape essentielle avant d'entamer la construction des modèles consiste à analyser les données afin de valider la bonne adéquation de leurs propriétés avec les exigences de l'assureur.

Classiquement, on applique le modèle collectif qui suppose que la prime pure suit une distribution composée fréquence-sévérité. Cela se traduit par une modélisation séparée de la fréquence du sinistre par assuré et du coût moyen de ces sinistres. Historiquement, les modèles qui répondent à cette problématique sont **les modèles linéaires généralisés (GLM)**. Introduits par John NELDER et Robert WEDDERBURN (1972), leur objectif était d'élargir le cadre limité de la régression linéaire multiple. Une grande variété de distributions est alors proposée par ces modèles, qui va au-delà de la loi Normale des modèles linéaires. Leur facilité d'interprétation ainsi que leur capacité à mettre en évidence l'impact des variables explicatives sur la variable réponse en font des outils de choix pour les actuaires dans le domaine de la tarification automobile. Ils sont par ailleurs plus robustes que d'autres techniques de modélisation prédictive et moins susceptibles de faire du surapprentissage sur des bases de données peu volumineuses.

Les GLM se distinguent des modèles linéaires classiques à travers leur prise en compte partielle des effets non linéaires grâce au choix d'une fonction de lien qui transforme la relation de dépendance représentée par une structure linéaire entre la variable à modéliser et les variables explicatives. Généralement, c'est la fonction logarithme qui est choisie pour les modélisations en assurance non-vie. Un des inconvénients des GLM est la non prise en compte des effets croisés entre des prédicteurs qui évoluent simultanément [GUILLOT, 2015], c'est au statisticien de les vérifier à priori manuellement dans la formule de régression dans son modèle. Si par exemple dans un modèle, l'actuaire dispose de 5 variables explicatives prenant chacune 8 modalités, il sera amené alors à tester $8^5 = 32768$ interactions possibles en évaluant leur significativité. La gestion des interactions nécessite donc le recours à de nouvelles extensions des modèles classiques. Afin de traiter cette difficulté, il va falloir s'affranchir du caractère additif des modèles économétriques. Pour cela, l'utilisation des arbres de décision constitue une amélioration à cette problématique. Cette approche sera détaillée par la suite dans ce mémoire.

Aujourd'hui, le monde des données connait une profonde révolution avec l'explosion de données massives. Des équipes de chercheurs en informatique ont mis au point un grand nombre d'algorithmes dont l'objectif était soit la prédiction de valeurs, soit la classification d'individus. Ces travaux ont donné naissance à la théorie de l'apprentissage statistique (*machine learning*). L'intérêt pour les données massives pousse les actuaires à s'intéresser à d'autres approches issues de la théorie statistique de l'apprentissage. [PAGLIA, 2010] propose ainsi dans la situation classique de la tarification d'un contrat d'assurance automobile, une comparaison entre les approches classiques par GLM et une méthode basée sur la théorie de l'apprentissage statistique. La théorie de l'apprentissage statistique ne formule qu'une seule hypothèse : les données à

prédire sont générées de façon identique et indépendante par un processus à partir du vecteur des variables explicatives, contrairement à l'approche classique qui nécessite de formuler des hypothèses sur la distribution des données. Le but est alors de construire un algorithme qui va apprendre à prédire la valeur de la variable cible en fonction des variables explicatives. Il résulte de cet apprentissage une fonction qui fait intervenir les variables explicatives et un paramètre de complexité. Ce paramètre représente par exemple dans la théorie des arbres de décision le nombre de noeuds. Grâce à cette fonction qui devient de plus en plus complexe à mesure que l'algorithme apprend, des singularités de la structure de données comme les interactions entre variables peuvent être modélisées.

Il convient aussi de noter que, lors de la modélisation de la fréquence, il peut y avoir une difficulté liée aux classes fortement déséquilibrées. En effet, comme c'est le cas dans ce mémoire, les contrats non sinistrés représentent la majorité du portefeuille. La qualité de la prédiction de la fréquence de sinistres par contrat est mesurée par la capacité du modèle utilisé à détecter correctement une minorité d'observations, c'est-à-dire les polices sinistrées. Les modèles traditionnels ne sont toujours pas assez robustes en présence de ces classes minoritaires car la vraisemblance est une quantité moyennée, et sa maximisation ne permet pas de prendre en compte correctement l'information qui ressort de ces observations. Cet enjeu de détection de classes déséquilibrées constitue l'un des défis majeurs du *machine learning* .

Enfin, les modèles d'apprentissage statistique améliorent généralement la prédiction de la prime pure d'un contrat d'assurance non-vie grâce à deux qualités majeures. D'abord, ils sont capables de modéliser les structures de dépendance présentes dans les données alors que celles-ci doivent être explicitement précisées dans les modèles GLM. De plus, le modèle produit est optimisé, non pas pour donner le meilleur ajustement sur la base de données, mais pour réduire l'erreur de la valeur prédite sur une autre base (indépendante de la base sur laquelle le modèle a été construit) ce qui ne peut que renforcer la robustesse des résultats prédits.

Chapitre 2

Présentation des modèles de tarification

L'objet de ce chapitre est de poser les bases théoriques essentielles à la bonne compréhension des outils utilisés dans le cadre de la modélisation du **nombre de sinistres** et du **coût du sinistre**.

2.1 Modélisation du risque en assurance automobile

2.1.1 Le modèle collectif

La tarification en assurance non-vie consiste à estimer une prime pure payée par l'assuré à l'assureur en échange d'un transfert du risque. Comme la prime dépend du risque, il s'agit ainsi de regrouper les différents contrats en sous-portefeuilles, ou classes de risques homogènes. On parle de tarification *à priori* si ces classes sont constituées à partir des informations sur l'assuré ou sur le bien assuré disponible a priori. Dans l'approche collective, les contrats ne sont pas distingués et les charges de sinistres individuelles ne sont pas forcément connues. La charge de sinistres S est déterminée à l'aide du nombre total de sinistres et des montants de chacun de ces sinistres. Autrement dit, la prime pure est le résultat du produit de l'espérance du nombre de sinistres et de leur coût moyen.

Dans le modèle collectif, on exprime la sinistralité de la façon suivante à l'aide de :

- S : variable aléatoire du montant total des sinistres d'un risque au cours d'une période (en général 1 an) ;
- N : variable aléatoire du nombre de sinistres ;
- X_i : variable aléatoire représentant le montant du sinistre i.

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

La prime pure se définit comme l'espérance de S, soit :

$$\mathbb{E}[S] = \mathbb{E}[\sum_{i=1}^{N} X_i] = \mathbb{E}[N]\mathbb{E}[X]$$

La prime pure permet ainsi une modélisation distincte de la fréquence et du coût. Ces deux quantités ne dépendent pas des mêmes facteurs de risque, il paraît donc intuitif de les modéliser séparément.

2.1.2 Démarche de la tarification classique avec GLM

La tarification en non-vie vise à modéliser la fréquence de survenance des sinistres ainsi que leurs coûts en fonction des différentes caractéristiques du contrat (profil de l'assuré, type du véhicule...). Il convient alors de suivre les étapes suivantes :

- **Étude de la qualité du portefeuille :** Il s'agit de vérifier que la base utilisée dans l'étude est de bonne qualité. En effet, un contrôle de premier niveau permet de détecter les anomalies présentes dans la base de données ;
- **Retraitement de la base de données :** Retraiter les anomalies détectées dans l'étape d'analyse de la qualité du portefeuille ;
- **Analyses univariées des variables tarifaires :** Étude de la fréquence de survenance et du coût moyen par segment afin de vérifier l'intérêt tarifaire de chacune des variables ;
- **Sélection des variables utiles à la modélisation :**
  - ☐ Étude des corrélations entre les variables tarifaires : Des variables très corrélées entre elles pourraient biaiser le modèle. Il convient alors de supprimer certaines variables.
  - ☐ Méthode de sélection en régression : Utiliser une méthode de sélection afin de mesurer l'impact des variables sur la modélisation car l'analyse des corrélation n'est pas toujours suffisante.
- **Modélisation de la prime pure :** Afin de modéliser la prime pure, le coût et la fréquence de survenance sont modélisés séparément à l'aide de modèles statistiques adaptés par exemple le GLM. L'association des résultats obtenus permet alors de modéliser la prime pure associée à chaque assuré en appliquant la formule suivante :

  $$Prime\ Pure = \frac{Nombre\ de\ sinistres}{Risque\ année} * \frac{Coût\ des\ sinistres}{Nombre\ de\ sinistres}$$

  où « Risque année » correspond à l'exposition du contrat.

  Ceci est équivalent à :

  $$Prime\ Pure = Fréquence * Coût\ moyen$$

2.1.3 Le passage de la prime pure à la prime commerciale

La prime commerciale est calculée en appliquant un chargement correspondant aux différents coûts et frais de l'assureur : taxes, coût de la réassurance, frais (ces frais comportent autant les frais de gestion des sinistres que la rémunération des agents généraux ou des courtiers).

De manière générale, chaque compagnie applique ses propres règles pour le calcul du chargement commercial.

2.1.4 Les modèles linéaires généralisés (GLM)

2.1.4.1 Définition

Le Modèle Linéaire Généralisé est la principale méthode de tarification adoptée depuis plus d'une vingtaine d'années. Son usage est aujourd'hui quasi-systématique en assurance non-vie. Ces modèles constituent une extension des modèles linéaires. Ils permettent de modéliser une relation non-linéaire entre la variable à expliquer et les variables explicatives. La figure 2.1 constitue un récapitulatif non exhaustif. Les régressions utilisées dans le cadre de ce mémoire sont mises en évidence avec la couleur rouge.

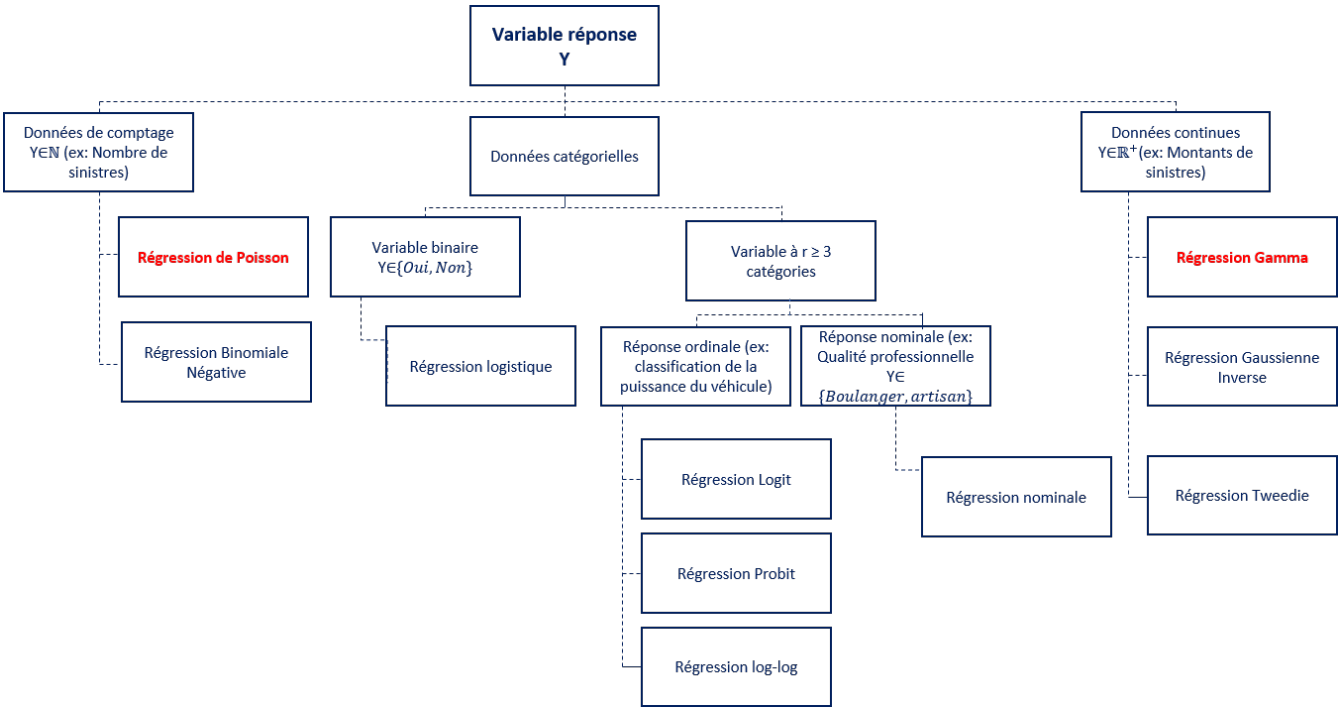


FIGURE 2.1 – Schéma classique de choix de régression pour un GLM

A travers une fonction lien g, les modèles linéaires généralisés (GLM) permettent de modéliser la relation entre la variable réponse Y et les variables explicatives X :

$$g(\mathbb{E}[Y_i|X_i]) = X_i^t\beta$$

Les GLM sont caractérisés par trois composantes :

- La **composante aléatoire**, qui se définit par la distribution de probabilité de la variable réponse Y ;
- La **composante déterministe** $X_i^t\beta$, ou prédicteur linéaire, qui se définit par une fonction linéaire des variables explicatives ;
- La **fonction de lien**, qui exprime une relation fonctionnelle entre l'espérance mathématique de Y notée η et les variables explicatives X.

La composante aléatoire :

La distribution de probabilité de la variable réponse Y est généralement supposée appartenir à la famille exponentielle, c'est-à-dire que sa densité s'exprime de la façon suivante :

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

avec :

$$\begin{aligned}
\theta_i &\in \mathbb{R} = paramètre\ canonique\ ou\ de\ la\ moyenne \\
\phi &\in \mathbb{R} = paramètre\ de\ dispersion \\
a\ &fonction\ définie\ sur\ \mathbb{R}\ non\ nulle \\
b\ &fonction\ définie\ sur\ \mathbb{R}\ deux\ fois\ dérivable \\
c\ &fonction\ définie\ sur\ \mathbb{R}^2
\end{aligned}$$

Les lois binomiales, de Bernoulli, de Poisson, Normale, Gamma ou Gaussienne inverse sont alors éligibles comme on peut le voir sur le tableau 2.1.

TABLE 2.1 – Caractéristiques de distribution usuelle de la famille exponentielle

Distribution de Y_i	θ_i	ϕ	$a_i(\phi)$	$b(\theta_i)$
Normale (μ_i, σ^2)	μ_i	σ^2	ϕ	$\frac{\theta^2}{2}$
Poisson (μ_i)	$log(\mu_i)$	1	ϕ	$exp(\theta_i)$
Binomiale $\frac{1}{m_i}(m_i; \mu_i)$	$log(\frac{\mu_i}{1-\mu_i})$	$\frac{1}{\mu_i}$	ϕ	$log(1 + exp(\theta_i))$
Gamma (μ_i, α)	$\frac{-1}{\mu_i}$	α^{-1}	ϕ	$-log(-\theta)$
Inverse Gaussienne (μ_i, σ^2)	$\frac{-1}{2\mu_i^2}$	σ^2	ϕ	$-(-2\theta)^{\frac{1}{2}}$

- **La composante déterministe :**

  La composante déterministe, aussi connue sous le nom de *composante systémique* relie le paramètre η aux variables explicatives X. Cela se traduit par :

$$g(\eta) = \beta^t X = \beta_1 X_1 + ... + \beta_p X_p$$

- **La fonction de lien :**

  La relation entre la composante aléatoire et le prédicteur linéaire est exprimée par la troisième composante appelée fonction de lien g qui doit être différentiable et monotone. Notons $\mu_i = \mathbb{E}(Y_i)$, on a alors :

$$g(\mu_i) = \eta_i \quad \text{ou} \quad \mu_i = g^{-1}(\eta_i) = g^{-1}(X_i^t \beta)$$

  Le choix de la fonction de lien dépend significativement de la nature du problème et des données étudiées. Le tableau 2.2 récapitule quelques lois de probabilité appartenant à la famille exponentielle ainsi que leur fonction de lien canonique.
  En assurance non-vie et plus particulièrement en tarification, il est fréquent d'utiliser une fonction de lien log et ce pour les raisons suivantes :

  1. Elle garantit la cohérence des supports de la variable de réponse et de l'exponentielle du prédicteur linéaire ;
  2. Elle permet d'obtenir une structure tarifaire « multiplicative » qui apparait comme la forme la plus adaptée aux données d'assurance non-vie.

TABLE 2.2 – Fonctions de liens canoniques associées aux principales lois de probabilité de la famille exponentielle

Loi	Nom du lien	Fonction de lien
Normale	Lien identité	$g(\mu) = \mu$
Poisson	Lien log	$g(\mu) = ln(\mu)$
Gamma	Lien inverse	$g(\mu) = \frac{1}{\mu}$
Binomiale	Lien logit	$g(\mu) = log(\frac{\mu}{1-\mu})$

2.1.4.2 Estimation des paramètres

L'objectif des GLM est d'estimer les coefficients de régression $\beta_0, \beta_1, ..., \beta_p$. Pour cela, on utilise la méthode de maximum de vraisemblance détaillée ci-dessous.

Considérons la variable réponse notée Y_i indépendante et issue d'une famille exponentielle. La vraisemblance s'écrit :

$$L(y_1, ..., y_n; \theta, \phi) = exp\left(\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right)$$

Notons $L = L(y_1, .., y_n; \theta, \phi)$, nous obtenons l'expression de la log-vraisemblance suivante :

$$log(L) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$$

Il faut donc maximiser cette dernière expression, en calculant tout d'abord la dérivée en fonction des paramètres β_j :

$$\frac{\partial \log(L)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j}(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi))$$

Les équations de vraisemblance sont alors :

$$\frac{\partial \log(L)}{\partial \beta_j} = 0 \quad \forall j = 1, ..., p$$

La résolution de ces équations requiert une méthode itérative telle que **la méthode de Newton-Raphson**.

2.1.4.3 Significativité des variables

La significativité des coefficients associés aux variables explicatives peut être testée à l'aide du test de Wald.

Soit le test suivant :

$$\begin{cases} H_0 : \forall j, \beta_j = 0 \\ H_1 : \exists j, \beta_j \neq 0 \end{cases}$$

La statistique de Wald s'écrit alors :

$$W = \frac{\hat{\beta}_j}{\hat{\sigma}\left(\hat{\beta}_j\right)}$$

Sous H_0, la statistique du test suit approximativement une loi Normale $N(0, 1)$.

Le test de Wald peut aussi être défini ainsi :

$$Z = \left(\frac{\hat{\beta}_j}{\hat{\sigma}\left(\hat{\beta}_j\right)} \right)^2$$

Sous H_0, cette statistique suit asymptotiquement une loi de khi-deux à un degré de liberté.

2.1.4.4 Sélection des variables

Afin de sélectionner les variables qui influent le plus sur la variable réponse. il existe plusieurs types d'algorithmes de sélection de variables dans la littérature, dont les plus couramment utilisés sont les méthodes pas à pas suivantes :

- **La méthode ascendante (Forward) :** Le premier modèle est celui qui ne comporte que le terme constant, ensuite les variables exogènes sont introduites pas à pas, c'est-à-dire celles dont l'introduction provoque la plus grande augmentation du R^2.
- **La méthode déscendante (Backward) :** Le premier modèle étudié est celui comportant toutes les variables, puis la variable la moins significative est enlevée, c'est-à-dire celle dont la suppression provoque la plus faible diminution du R^2. Pour que le processus n'élimine pas toutes les variables, on peut introduire le même seuil que dans la méthode forward.
- **La méthode progressive (Stepwise) :** Cette méthode consiste en un mélange des méthodes forward et backward.

Parmi les critères les plus utilisés pour la sélection, on évoque les critères **AIC** et **BIC** :

- **L'Akaike Information Criterion (AIC) :** Il est défini par la formule suivante :

$$AIC = -2log(L) + 2k$$

où $log(L)$ constitue la log-vraisemblance maximisée et k représente le nombre de paramètres. L'AIC prend donc en compte à la fois la qualité de l'ajustement à travers la fonction de vraisemblance (qui dépend des coefficents du GLM) et de la complexité du modèle (via le nombre de variables retenues).

- **Le Bayesian Information Criterion (BIC) :** Le critère AIC a tendance à choisir les modèles avec de nombreuses variables explicatives. Afin de pallier ce problème, le nombre de paramètres dans la formule du BIC est multiplié par le logarithme du nombre d'observations $log(n)$ et permet ainsi d'appliquer une pénalité plus sévère afin de privilégier l'utilisation de modèles avec moins de variables explicatives. Ce critère est défini par la formule suivante :

$$BIC = -2log(L) + klog(n)$$

Ces deux critères doivent être minimisés. Un modèle sera considéré comme meilleur s'il présente un critère plus petit par rapport aux autres modèles testés.

2.1.4.5 Validation et comparaison des modèles

La déviance

Le modèle estimé est comparé au modèle dit saturé, c'est-à-dire le modèle possédant autant de paramètres que d'observations et estimant donc exactement les données. La déviance est définie à partir de la log-vraisemblance de ces deux modèles :

$$D = 2(log(L_{sat}) - log(L))$$

La statistique de Pearson

Un test de χ^2 est également utilisé pour comparer les valeurs observées Y_i à leurs valeurs prédites par le modèle. La statistique du test s'écrit :

$$\chi^2_{pearson} = \sum_{i=1}^{N} \frac{(Y_i - \hat{\mu_i})^2}{V\hat{A}R(\hat{\mu_i})}$$

Le modèle est jugé de mauvaise qualité si ces critères sont supérieurs au quantile d'une loi du χ^2 à $n - p - 1$ degrés de libertés.

2.2 La *Data Science* dans la tarification

2.2.1 Emergence des données et conséquences

La quantité de données accessibles aux actuaires ne fait que croître. Cette tendance est dûe à la digitalisation croissante de la société. Ces données sont utilisées en vue de détecter des règles, des associations, des tendances inconnues ou cachées. Pour cela, des méthodes scientifiques et des outils de calcul ont été développés et regroupés dans le *data mining*, que l'on peut traduire par *fouille de données*, et qui a récemment évolué vers la *data science* avec l'apport de nouvelles méthodes théoriques et l'arrivée de nouvelles problématiques et de nouvelles données. Le *data mining* est l'art d'extraire des informations, voire des connaissances, à partir des données. Il intervient dès que, partant de données brutes, on tente d'aller du connu vers l'inconnu et de se livrer à des prédictions ou des analyses de tendances fouillées.

De la souscription à la gestion des sinistres, de la tarification au suivi des risques, la donnée est au cœur de l'activité d'assurance. Les volumes colossaux de données sont traités en parallèle par des machines performantes. Le développement et l'ampleur des nouvelles technologies comme les objets connectés (voitures, maisons, bracelets), les réseaux sociaux et les offres de la nouvelle ère (*Pay As you Drive, Pay As You Live*, etc.) transforment petit à petit le monde assurantiel. Il s'agit potentiellement d'un des plus grands défis de la prochaine décennie.

L'assurance ne peut plus exister sans l'utilisation de la *data science* car elle permet à l'assureur de mieux connaître ses assurés ou ses futurs assurés. Ce principe même constitue d'une part, la clé de développement de son portefeuille à travers notamment une tarification ciblée ou une prédiction sur les besoins et comportements de ses assurés et d'autre part, cela lui permet une meilleure maîtrise de ses risques à travers la détection de fraude, d'indemnisation ou de prévention par exemple.

Les assureurs disposent d'un avantage grâce à leur accès privilégié aux données. Le rôle de l'assureur est de coupler son savoir-faire avec ces nouvelles techniques de traitement de données afin de détecter les nouveaux comportements face aux risques et de proposer les offres de demain.

Enfin, comme nous l'avons précisé dans le paragraphe 1.3.1, il est essentiel de segmenter les tarifs pour des raisons de concurrence de marché. Les assureurs cherchent ainsi à déterminer la donnée qui leur permettra d'avoir la connaissance la plus fine de leur sinistralité. De ce fait, la collecte de données tant internes qu'externes ainsi que les capacités de traitement et d'analyse de celles-ci représentent des enjeux forts pour les assureurs.

2.2.2 Démarche d'un projet *Data Science*

Le *machine learning* est la science de la conception et de l'application d'algorithmes capables d'apprendre des informations du passé. Il utilise des algorithmes complexes qui sont itérés sur de larges bases de données et analyse les tendances (*patterns*) dans les bases. L'algorithme

permet aux machines de répondre plus facilement aux différentes situations sur lesquelles elles n'ont pas été explicitement programmées.

Il faut du temps pour l'homme pour lire, collecter, catégoriser et traiter les informations. Le *machine learning* apprend aux machines à identifier et jauger l'importance de ces informations à la place des humains. En particulier, dans les cas où les données doivent être obligatoirement analysées et exploitées dans un intervalle de temps précis, avoir l'appui des machines permet à l'homme d'être plus efficace et d'agir avec confiance.

Les méthodes de *machine learning* convertissent les informations complexes en un format simple aidant ainsi les responsables à prendre des décisions. L'utilisateur apprend le système de *machine learning* en ajoutant en permanence des données et de l'expérience. Ainsi, au coeur du*machine learning*, il y a un cycle de 3 parties : Apprentissage/Test/Prédiction ou classification.

De par la connaissance des sujets abordés et la maîtrise des outils informatiques, l'intervention d'un expert est indispensable à la bonne réalisation d'un projet *Data Science*.

Les étapes indispensables à la bonne mise en place d'un projet *Data Science* sont décrites ci-après :

1. **Compréhension du périmètre de l'étude** : Cette phase consiste à voir quelles sont les connaissances sur le sujet et bien cerner les besoins du commanditaire du projet ;
2. **Extraction de la base de données** : Cette étape consiste à extraire la base sur laquelle le *Data Scientist* va pouvoir mettre en place les modèles ;
3. **Nettoyage de la base** : Cette phase vise à traiter les erreurs, les fautes de saisies, les valeurs aberrantes etc ;
4. **Exploration des données** : Cette phase consiste à étudier les distributions des variables, de leurs interactions etc ;
5. **Transformation des données** : Cette étape consiste à réduire la dimension (dans le cadre d'une *ACP* par exemple), découpage en classes etc ;
6. **Choix des modèles** : Selon le but final de l'étude (interprétabilité, prédictibilité...), les modèles mis en œuvre ne seront pas forcément les mêmes ;
7. **Tests sur les différents modèles** : Cette étape vise vérifier la qualité d'ajustement des modèles et leur qualité de prévision ;
8. **Choix du modèle final** : Cette étape consiste à choisir le meilleur modèle en se basant sur les tests qui viennent d'être réalisés.

2.3 Présentation de la théorie des modèles de *Data Science*

L'apprentissage statistique permet d'identifier les sous-ensembles homogènes. Par conséquent, quand un nouvel individu est soumis au système, il peut être affecté à l'une de ces classes.

2.3.1 Typologie des modèles

Régression / Classification

— On parle de régression lorsque la variable à prédire est continue, comme le coût du sinistre ;

— On parle de classification lorsqu'il s'agit de prédire la classe d'une variable discrète, comme des cas de fraudes, de résiliations, etc.

Apprenstissage supervisé / non supervisé

Il existe deux problèmes en *Data Science* :

— **L'apprentissage supervisé :** Ce problème présuppose que l'on dispose d'un ensemble d'exemples où les variables explicatives sont distinguées des variables à expliquer. L'algorithme supervisé a pour objectif de construire, à partir de la base d'apprentissage, un modèle qui permet la prédiction des variables réponses d'un nouvel échantillon (appelé base de test).

— **L'apprentissage non-supervisé :** Dans ce cas, la variable réponse n'est pas connue. L'algorithme doit trouver la structure présente dans la base de données. Ainsi, le modèle regroupe les individus en sous-groupes homogènes. Dans ce cas, l'apprentissage ne peut plus se faire à partir d'une indication qui peut être préalablement fournie par un expert, mais uniquement à partir des fluctuations observables dans les données. On cite par exemple le *clustering* qui fait partie de ces algorithmes.

L'ensemble des modèles élaborés dans ce mémoire sont des modèles d'apprentissage supervisé.

Le but de cette section est d'expliquer quelques notions fondamentales et communes aux algorithmes qui sont utilisés dans ce mémoire :

- **L'algorithme CART ;**
- **Les forêts aléatoire ou Random Forest ;**
- **Le Gradient Boosting Machine ;**
- **L'extreme Gradient Boosting ou XGBoost.**

2.3.2 Quelques notions de *Data Science*

Avant de commencer la présentation des modèles théoriques, il est important de rappeler quelques notions de la *Data Science*.

Le dilemme biais-variance :

Les modèles d'estimation commettent généralement deux erreurs quand ils prédisent une variable :

- Le biais : correspond à l'erreur entre la variable observée et celle prédite par le modèle ;
- La variance : est liée à la généralisation du modèle sur d'autres données.

Le compromis *biais-variance* réside dans le fait qu'un modèle simple (variance faible) risque le sous-apprentissage (biais élevé y compris sur les données d'entraînement). Un modèle complexe (variance élevée) risque le sur-apprentissage (biais faible sur les données d'entraînement mais élevé sur de nouvelles données). L'intérêt est de trouver un modèle « intermédiaire », on peut le voir par exemple sur l'illustration 2.2, au point d'inflexion de la courbe orange où le biais de prédiction est le plus faible et la généralisation [1] est meilleure.

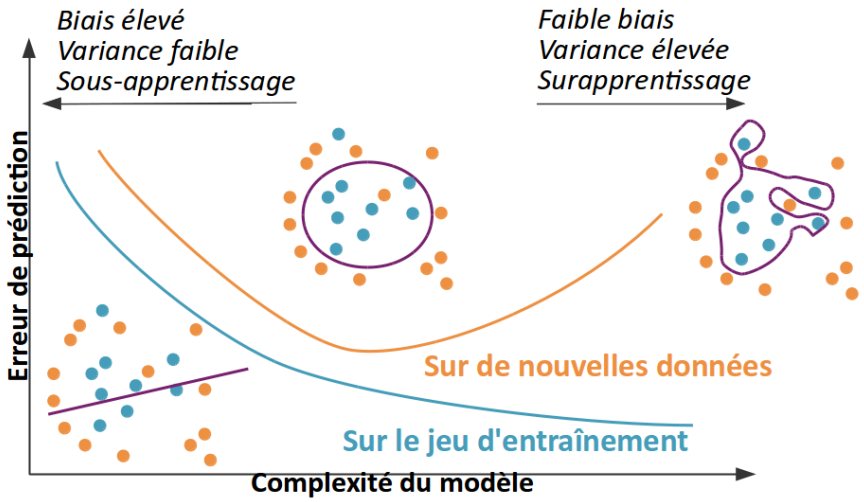


FIGURE 2.2 – Le compromis biais-variance

On suppose qu'il existe une fonction f reliant les variables explicatives X à la variable à prédire Y :

$$Y = f(X) + \epsilon$$

où ϵ est un vecteur aléatoire d'espérance nulle et de variance σ^2.

1. On entend par la généralisation, la capacité du modèle à faire des prédictions sur de nouvelles bases de données et non pas seulement sur le jeu de données sur lequel il a été construit.

Avec les données d'apprentissage, nous créons la fonction de prédiction $\hat{f}$ de telle sorte que si on donne une matrice d'entrée X au modèle, on obtient une estimation de la variable réponse par $\hat{f}(X)$. L'erreur de prédiction du modèle est l'écart entre la réalisation $f(X)$ et la prédiction $\hat{f}(X)$.
L'erreur quadratique se traduit par l'équation suivante :

$$\begin{aligned}\mathbb{E}[(f(X)-\hat{f}(X))^2] &= \mathbb{E}[f(X)^2 - 2f(X)\hat{f}(X) + \hat{f}(X)^2] \\ &= \mathbb{E}[f(X)^2] + \mathbb{E}[\hat{f}(X)^2] - 2f(X)\mathbb{E}[\hat{f}(X)] \\ &= Var(f(x)) + \mathbb{E}[f(X)]^2 + Var(\hat{f}(X)) + \mathbb{E}[\hat{f}(X)]^2 - 2f(X)\mathbb{E}[\hat{f}(X)] \\ &= \sigma^2 + Var(\hat{f}(X)) + \mathbb{E}[f(X) - \hat{f}(X)]^2 \\ &= \sigma^2 + Var(\hat{f}(X)) + Biais^2(\hat{f}(X))\end{aligned}$$

avec :

- $Var(\hat{f}(X)) = \mathbb{E}[\hat{f}(X)^2] - \mathbb{E}[\hat{f}(X)]^2$
- $Biais(\hat{f}(X)) : \mathbb{E}[f(X) - \hat{f}(X)]$

Nous pouvons alors constater que l'erreur quadratique moyenne est bien en fonction du biais et de la variance.

- **Découpage de la base de données :**

Classiquement, la base de données est divisée en un échantillon d'apprentissage sur lequel le modèle est construit, et un échantillon de test (indépendant de la base d'apprentissage) qui sert à mesurer la qualité et la pertinence du modèle. Dans le cadre de notre étude, nous avons obtenu ces deux échantillons par un tirage aléatoire simple et sans remise en prenant :

- 80% de la base initiale pour constituer la base d'apprentissage ;
- 20% de la base initiale pour constituer la base de test.

- **La validation croisée :**

Le principe de la validation croisée aussi appelée *k fold Cross Validation* est de découper la base de données en k échantillons ayant la même taille. L'un de ces sous-ensembles est considéré comme la base de test, les $k-1$ sous-ensembles restants sont utilisés comme base d'apprentissage. Cette opération est répétée en prenant chacun des k échantillons comme base de test. Finalement, les modèles ainsi construits ont chacun un score de performance. Ensuite un score de performance moyen est alors calculé en prenant leur moyenne.

L'illustration 2.3 résume l'approche décrite ci-dessus.

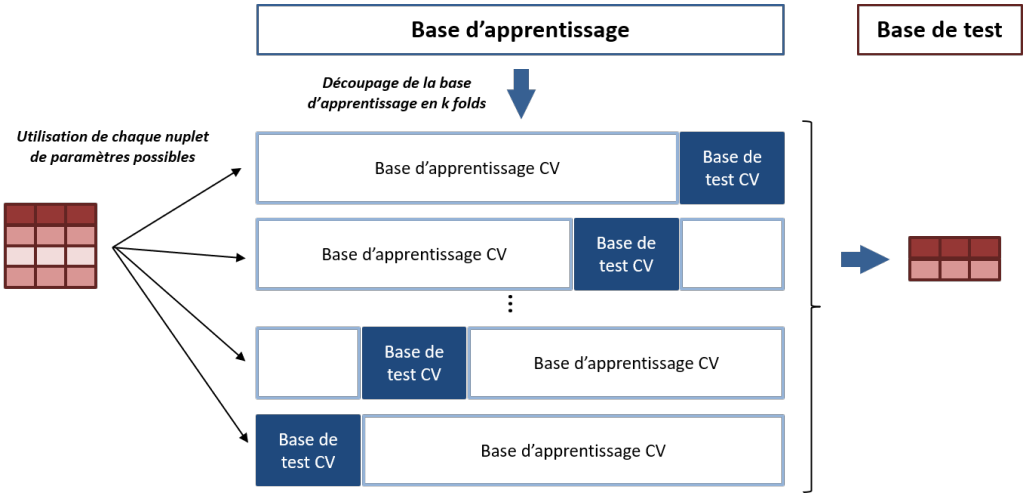


FIGURE 2.3 – Le principe de la validation croisée

Dans le but de déterminer les hyperparamètres des modèles, deux méthodes de validation croisée peuvent être utilisées :

- La méthode **Grid Search** :

  Elle constitue la méthode classique de recherche exhaustive des paramètres optimaux. L'idée est de choisir un nombre fini de valeurs à tester pour chaque paramètre. L'algorithme teste alors chaque combinaison possible.

  Exemple : Dans le cas de la forêt aléatoire, nous pouvons chercher le meilleur paramètre relatif aux nombres d'arbres **ntrees** $\in \{100, 200, 300\}$. Cette méthode sera retenue pour la suite de l'étude.

- La méthode **Random Search** :

  Contrairement à la première méthode, le *Random Search* remplace la recherche exhaustive des valeurs du paramètre par énumération de toutes les combinaisons possibles, par une recherche alétoire dans un intervalle de valeurs.

  Exemple : Dans le cas de la forêt alétoire, nous cherchons le nombre d'arbres **ntrees** $\in [100, 300]$. Le **Random Search** est alors plus long à mettre en place.

Indicateur de performances dans le cas de la régression :

Rappelons que l'erreur d'estimation d'un modèle correspond à l'écart entre la valeur observée et la valeur prédite. On compare les valeurs obtenues par un indicateur (ou plusieurs) pour déterminer le meilleur modèle pour un jeu de données d'entraînement, ou le cas échéant pour un contexte donné si l'on est en recherche d'une généralisation ou de l'élaboration d'une méthode.

L'idée est donc d'obtenir un modèle avec une erreur d'estimation minimale.

En notant :

- y_i : La valeur à prédire (nombre de sinistres i ou coût du sinistre i) ;
- $\hat{y}_i$: La valeur prédite (nombre de sinistres i ou coût du sinistre i).

Les indicateurs de performance des modèles prédictifs classiquement utilisés en régression sont :

- **Mean Squared Error (MSE) :**

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

.

On utilise souvent la **« RMSE » (Root Mean Squared Error)** qui est la racine carrée de la **MSE**. Ces indicateurs présentent l'avantage de pénaliser plus fortement les fortes erreurs (à travers le carré) que d'autres mesures de performance.

- **Mean Absolute Error (MAE) :** Elle est donnée par l'expression suivante :

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

- **L'erreur euclidienne :**
  Cette métrique correspond à la racine carrée de la somme du carré des distances séparant les valeurs observées des valeurs prédites. Elle est donnée par la formule suivante :

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

2.3.3 Les arbres de décision CART

Les arbres de décisions, aussi appelés *Arbres de Régression et de Classification (CART)*, sont des méthodes d'apprentissage statistique utilisées pour la régression ou la classification. Ils sont efficaces et très populaires et ont été développés initialement par Leo Breiman *et al.* en 1984. Ils constituent la base de plusieurs modèles de prédiction en *Data Science* tels que les forêts aléatoires ou le Gradient Boosting Machine.

Le but principal d'un arbre est d'expliquer une variable à partir de variables continues ou discrètes. Ceci peut être représenté par une matrice X avec m observations et n variables, associée à un vecteur Y à expliquer. Y peut prendre des valeurs :

- Numériques : Dans ce cas, on parle d'arbre de régression ;
- Qualitatives : Dans ce cas on parle d'arbre de classification.

Les arbres de décision consistent à séparer les observations suivant une hiérarchie d'arbre de telle façon à minimiser une fonction de coût telle que la **MSE** pour les arbres de régression et **le coefficient de GINI** pour les arbres de classification.

Dans ce mémoire, nous nous limiterons au cas des arbres de régression puisque les variables Y à prédire sont des variables numériques : **nombre de sinistre** et **coût des sinistres**.

2.3.3.1 Principe de construction de l'arbre

L'arbre de décision débute par un noeud initial (appelé aussi *racine*), puis se découpe en deux branches menant à deux noeuds et ainsi de suite jusqu'à ce que l'arbre atteigne une condition d'arrêt.

On obtient alors un emboîtement de rectangles qui définissent une partition de la population. Les noeuds terminaux (appelés *feuilles*), se situent en bas de l'arbre. Ils regroupent des ensembles homogènes d'observations, ces feuilles partagent des combinaisons de modalités de variables explicatives ayant un effet commun sur la variable réponse permettant à la variable d'intérêt de prendre des valeurs aussi homogènes que possible.

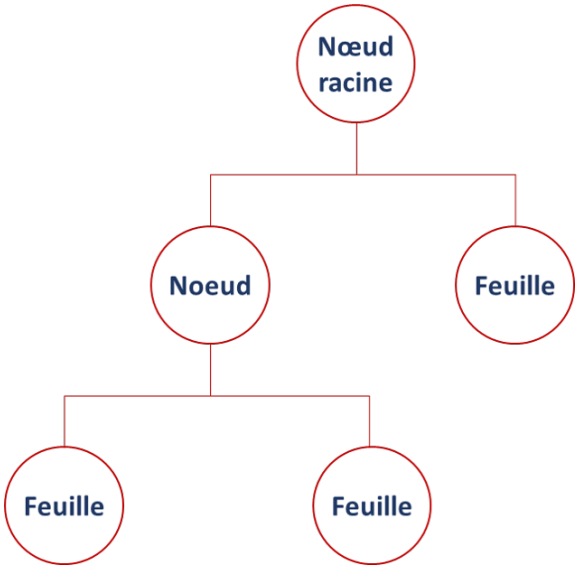


FIGURE 2.4 – Illustration simplifiée d'un arbre CART

Notons :

- Y : La variable réponse ;
- r : Le nombre de covariables ;
- X_j : Les covariables, avec $1 \leq j \leq r$;
- π_0 : La quantité que l'on veut prédire.

Souvent, la quantité que l'on veut prédire est :

$$\pi_0 = \mathbb{E}[Y|X = x]$$

Il est possible de choisir une autre quantité comme un quantile. Il faut alors choisir le bon critère de mesure de l'homogénéité des noeuds.

Lorsque la quantité est l'espérance, la fonction de perte que l'on utilise est l'erreur de généralisation des moindres carrés ou *mean squared error* (MSE) :

$$\mathbb{E}[(\pi(x) - Y)^2]$$

Comme l'espérance est la solution de la minimisation de l'erreur quadratique, la quantité d'intérêt choisie est solution de l'équation suivante :

$$\pi_0(x) = arg\min_{\pi(x)} \mathbb{E}[\phi(Y, \pi(x))|X = x]$$

avec $\phi(Y, \pi(x)) = (Y - \pi(x))^2$

En pratique, les calculs d'espérance sont faits de façon empirique. Ainsi, on cherche :

$$\pi_n(x) = arg\min_{\pi(x)} \mathbb{E}_n[\phi(Y, \pi(x))|X = x]$$

On note :

- $V_n(.)$: La variance empirique ;
- $E_n(.)$: L'espérance empirique.

$$\mathbb{E}_n(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{et} \quad V_n(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{E}_n[Y])^2$$

Comme $\mathbb{E}[(Y - \mathbb{E}[Y])^2] = Var[Y]$, ceci revient à opter pour la variance empirique comme critère de sélection.

Afin de construire l'arbre, il faut segmenter chaque noeud en deux noeuds fils, en veillant à minimiser la variance des deux nouveaux noeuds. Au niveau de chaque noeud construit, le nouvel estimateur de $\mathbb{E}[Y]$ devient l'espérance empirique de l'ensemble des observations du noeud en question.

On note :

- t_{gauche} : le noeud fils de gauche ;
- t_{droite} : le noeud fils de droite ;
- Q_{gauche} : la proportion d'individus dans le noeud de gauche ;
- Q_{droite} : la proportion d'individus dans le noeud de droite.

Dans le but de trouver la meilleure variable explicative ainsi que son meilleur seuil, on résout le problème suivant :

$$\min_{X_i \leq j} \; Q_{gauche} \; * \; V(t_{gauche}) + Q_{droite} \; * \; V(t_{droite})$$

Afin de pouvoir séparer l'ensemble en deux sous-ensembles plus homogènes, on teste chaque seuil et chaque variable explicative et on choisit le couple qui minimise la variance des deux nouveaux noeuds. Ce processus est réitéré sur chaque nouveau noeud avec le nouvel ensemble associé.

En résumé, l'algorithme de création de l'arbre maximal est alors :

1. On se place au niveau de la racine de l'arbre (population initiale).
2. La somme pondérée de la variance des noeuds fils est minimisée en faisant un test sur chaque covariable et chaque seuil.
3. Dans le cas où la segmentation n'est plus possible, on s'arrête. Sinon on recommence l'étape 2 sur chaque noeud fils.

2.3.3.2 Elagage de l'arbre

Un arbre peut continuer à pousser indéfiniment en segmentant de plus en plus les données en partitions jusqu'à ce que l'algorithme n'ait plus de critères de séparation. Cependant, exiger que toutes les observations soient parfaitement rangées amène nécessairement au surapprentissage. Ce risque survient lorsqu'il y a un sur-ajustement vis-à-vis de la base d'apprentissage. Par conséquent, le modèle ne peut pas être facilement généralisé à d'autres données indépendantes.

Ainsi, l'élagage de l'arbre permet de supprimer les feuilles qui n'apporteraient rien à l'analyse. Cette opération reste néanmoins compliquée car les capacités prédictives de l'arbre ne doivent pas être impactées notamment en supprimant un nombre trop important de noeuds. Des opérations d'**élagage** sont pratiquées afin de pallier ce problème.

L'élagage à priori consiste à arrêter de rajouter des noeuds lorsque la profondeur de l'arbre excède un seuil particulier caractérisant la complexité maximale de l'arbre de décision ou le fait que le nombre d'observations par feuille n'est pas suffisant pour représenter les différentes classes. L'élagage à postériori quant à lui, part de l'arbre de complexité maximale, puis supprime les noeuds les uns après les autres jusqu'à revenir à une unique classe couvrant tout le portefeuille.

Il s'agit de minimiser la quantité $c(T) + \alpha|T|$ avec :

- $c(T)$: la somme des variances de chacune des classes ;
- α : paramètre de complexité de l'arbre (coût en terme d'erreur de l'addition d'un noeud dans le modèle). Le cas particulier $\alpha = 0$ correspond à l'arbre saturé. En calibrant α, il est possible de trouver un sous-arbre robuste.

2.3.3.3 Les limites des arbres de décision

Le Surapprentissage :

On dit qu'il y a surapprentissage lorsqu'un modèle est trop dépendant de l'échantillon de données sur lequel il est calibré. Il s'imprègne alors de toutes les caractéristiques des données

qu'il a parcouru, en particulier des points exceptionnels voire aberrants. Il retient alors ces traits exceptionnels et les considère comme des comportements normaux, ce qui génère un biais.

On retrouve ce problème dans les arbres de décisions lorsqu'il y a un très grand nombre de feuilles/noeuds. Afin de contourner ce risque, l'élagage de l'arbre permet de ne pas prendre en compte des individus atypiques non pertinents à l'étude. Par ailleurs, les modèles agrégés permettent de sélectionner le critère qui est redondant sur l'ensemble des arbres construits et mesurent sa contribution dans la construction de la variable d'intérêt dans chaque arbre.

Sensibilité aux optimums locaux :

Ce point constitue l'une des plus importantes faiblesses des arbres de régression. L'algorithme étudie les variables explicatives successivement. Les nœuds sont construits de façon enchaînée et un critère choisi pour figurer à un emplacement de l'arbre n'est plus réétudié par la suite. Ce qui suggère que modifier en amont la construction d'un critère fort remet en question la construction de l'intégralité de l'arbre. L'agrégation des modèles permet ici aussi de remonter les variables les plus déterminantes dans l'explication de la variable cible.

2.3.4 Les forêts aléatoires ou *Random Forest*

Le but de l'algorithme des forêts aléatoires est de conserver la plupart des atouts des arbres de décision en éliminant leurs inconvénients, en particulier leur sensibilité au surapprentissage, leur instabilité et la complexité des opérations d'élagage. C'est un algorithme de classification ou de régression non-paramétrique qui s'avère à la fois très flexible et très robuste. Le principe est le suivant : au lieu d'avoir un estimateur complexe censé tout faire, on en construit plusieurs de moindre qualité individuelle. Ces différents estimateurs sont ensuite réunis pour fournir une vision globale.

L'algorithme des forêts aléatoires peut être vu comme :

$$Random forest = tree\ bagging + feature\ sampling$$

en notant que le *tree bagging* constitue un algorithme qui consiste à assembler des arbres de décision qui ont été construits sur la base d'un tirage aléatoire parmi les observations. Les forêts aléatoires ajoutent au *tree bagging* un échantillonnage sur les variables du problème (*feature sampling*).

En utilisant les forêts aléatoires, on retrouve aisément la polyvalence des arbres de décisions. En effet, elles peuvent être utilisées en :

- Classification : Le résultat final est obtenu en faisant un vote sur chaque arbre ;
- Régression : Le résultat est obtenu en effectuant une moyenne sur les résultats de tous les arbres.

Construction de la forêt aléatoire :

- **Tree bagging :**

Le *bagging* consiste à sélectionner, à chaque itération, un échantillon avec remise de la base d'apprentissage, puis de construire un arbre classique (sans élagage) sur cet échantillon. L'opération d'agrégation des différents arbres bâtis constitue la méthode du *bagging*. La taille de l'échantillon constitué à partir du *bagging* est généralement la même que celle de la base initiale, mais pour des raisons de temps de calcul, il est possible d'envisager un sous-échantillonnage.

Grâce à un classement ou à une régression, le *bagging* trouve le modèle « moyen » approchant le mieux les données d'intérêt.

La figure 2.5 résume le principe du *bagging*.

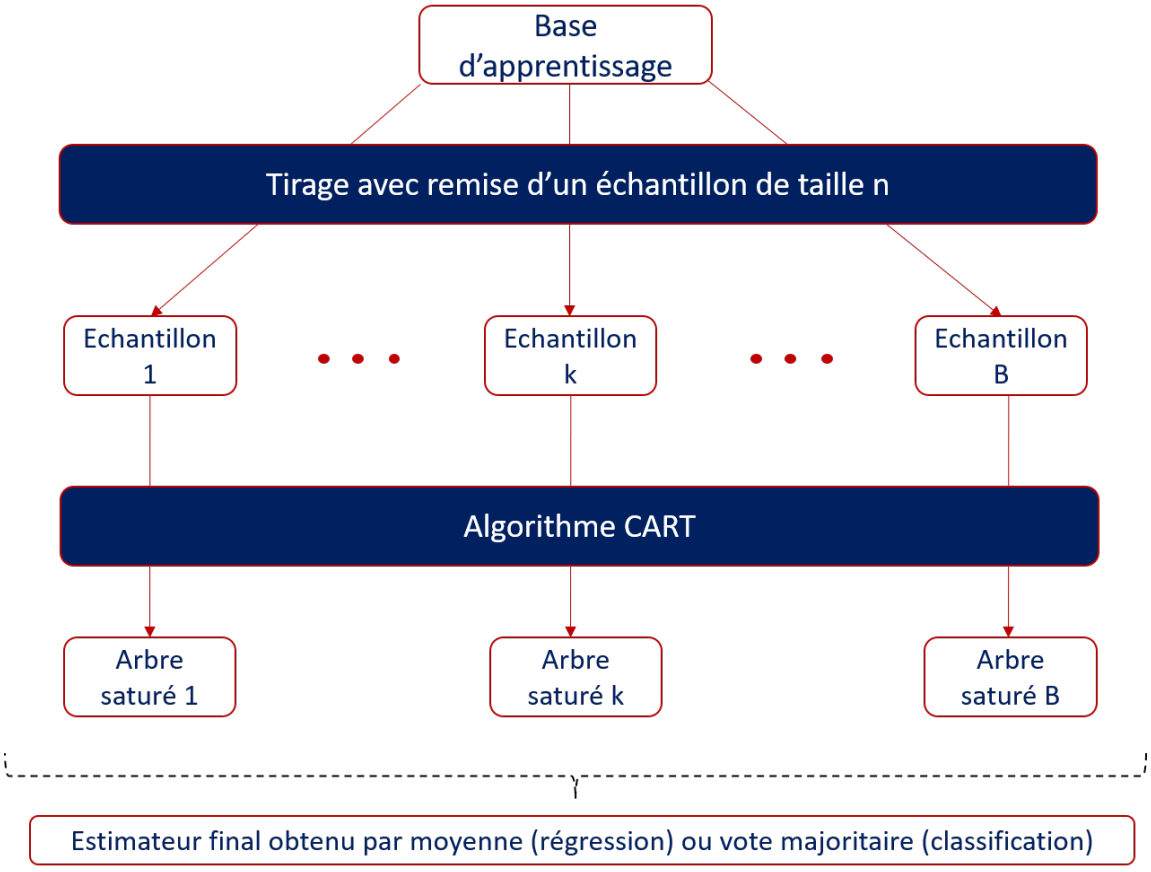


FIGURE 2.5 – Algorithme du *Bagging*

- **Feature sampling :**

La méthode des forêts aléatoires est similaire au *bagging* dans la manière d'agréger les modèles. Cependant, en plus du tirage aléatoire sur les lignes, le *random forest*, ainsi que d'autres méthodes ensemblistes qui se basent sur des arbres de décision, introduisent un tirage aléatoire sur les variables à utiliser. En effet, à chaque palier, lors de la génération d'un nouveau noeud, afin de chercher la division binaire optimale, l'algorithme parcourt un sous-ensemble aléatoire des variables candidates.

Ceci permet d'éviter le risque de corrélation des arbres construits par un simple *bagging*. Ce risque est d'autant plus important lorsque certaines variables ont un pouvoir explicatif élevé et apparaissent dans plusieurs arbres construits. La taille de ce sous-espace est, par défaut, $\sqrt{n}$ pour un problème de classification à n variables et $\frac{n}{3}$ pour un problème de régression. Ajouter cet aléa dans la construction des arbres permet de les rendre plus indépendants et donc de réduire la variance de l'estimation.

Une moyenne de B variables indépendantes et identiquement distribuées, chacune de variance σ^2, a une variance de $\frac{\sigma^2}{B}$. En excluant l'hypothèse d'indépendance des variables, ce qui est souvent le cas dans la réalité, et en notant ρ le coefficient de corrélation des variables, la variance de l'ensemble est égale à :

$$V_{forêt\ aléatoire} = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Si les échantillons sont indépendants, alors $\rho = 0$ et la variance converge vers 0 quand B grandit. La réduction du premier terme ne sera possible que grâce au *feature sampling* qui aura pour effet de baisser la corrélation entre les arbres et donc de réduire la variance de l'ensemble.

La figure 2.6 illustre l'algorithme des forêts aléatoires.

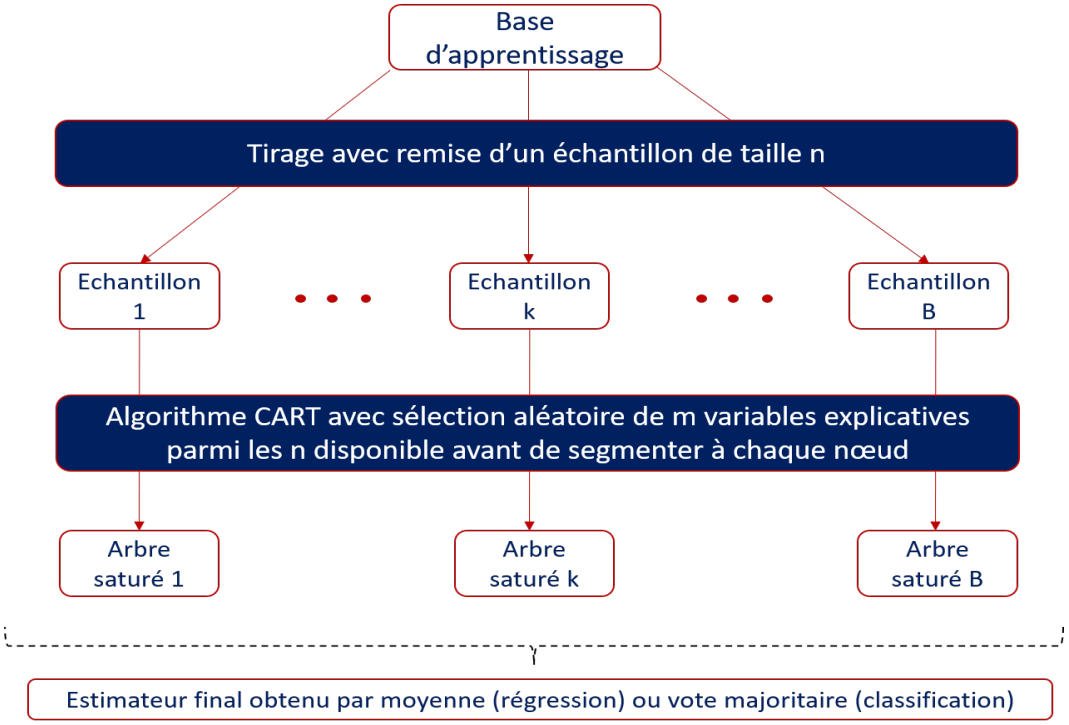


FIGURE 2.6 – Algorithme des *forêts aléatoires*

Nous pouvons résumer le principe de la construction des forêts aléatoires dans l'algorithme suivant :

Algorithme 1 : Génération d'une forêt alétoire

Entrées : Deux paramètres d'apprentissage :

- Nombre d'itérations : B
- Taille du sous-espace : r

pour $i \leftarrow 1$ *à* B **faire**
  Sélectionner un échantillon avec remise de la base initiale;
  Initialiser l'arbre R_k s'appuyant sur cet échantillon avec sa racine (moyenne globale des Y_i)
  pour chaque *Noeud terminal* D *de* R_k **faire**
    Sélectionner un sous-espace aléatoire de taille r des variables explicatives;
    Déterminer la division optimale parmi l'ensemble des partitions binaires des r variables envisagées
  fin
fin

Sorties : Retourner le modèle agrégé :$R = \frac{1}{B}\sum_{i=1}^{B} R_i$

- **Choix des variables :**

  Les forêts aléatoires permettent de sélectionner des variables et de mesurer leur importance.

  - **Erreur Out-of-bag** :

    Soit une observation (x_i, y_i) issue de la base d'apprentissage. Soit F_i^* la sous forêt construite avec un échantillon ne contenant pas cette observation. L'erreur Out-of-bag est définie comme l'erreur de prédiction moyenne des arbres composant la sous forêt F_i^* sur l'observation (x_i, y_i). En d'autres termes, pour chaque individu, le vote ou la prédiction moyenne ne sont pris en compte que sur les arbres qui ont été construits sur un échantillon n'incluant pas cet individu.

  - **Importance des variables** :

    L'erreur Out-of-bag permet de donner un ordre d'importance aux variables explicatives X dans la prédiction de Y. L'objectif étant de donner l'importance de l'information marginale qu'apporte une variable dans la construction d'une forêt aléatoire. La mesure d'importance d'un régresseur correspond à l'augmentation marginale de l'erreur out-of-bag dûe à la permutation des observations de cette variable. Si le fait de réarranger les valeurs d'une certaine variable n'impacte pas la précision de la prédiction, cela signifie qu'elle est peu importante.

2.3.5 Les méthodes de boosting

2.3.5.1 Boosting

Le *boosting* adopte le même principe que le *bagging*. C'est une méthode d'agrégation développée par Freund et Schapire (1996) qui repose sur des stratégies adaptatives (*adaboost* pour *adaptative boosting*). Il cherche à optimiser l'affectation des poids en fonction des prévisions. Il crée ainsi des classifieurs faibles h_i de façon à obtenir le classifieur H tel que :

$$H = signe\left(\sum_t \alpha_t h_t\right)$$

Alors que le Random Forest construit plusieurs arbres en parallèle, le *boosting* construit des arbres en série, c'est-à-dire que chaque arbre généré (sauf le premier) a accès à son prédécesseur, ou plus précisément à l'**erreur** de son prédécesseur.

Le nouvel arbre construit aura pour but de se concentrer sur les lacunes de son prédécesseur désormais dévoilées, en donnant plus de poids aux données mal prédites.

2.3.5.2 Le Gradient Boosting Machine

Le but du *Gradient Boosting Machine* est de trouver une combinaison linéaire d'arbres optimaux. Il va construire une nouvelle base corrélée avec le gradient de la fonction de perte.

Il convient toutefois de rappeler le principe de l'algorithme d'optimisation de la descente du Gradient. Cet algorithme vise à déterminer un minimum local en partant d'un point aléatoire puis de se déplacer dans la direction de la plus forte pente. Finalement, l'algorithme finira par converger et donnera ainsi le minimum.

Partant du même principe que l'algorithme *Adaboost*, *le Gradient Boosting Machine* construira une série de modèles de façon à ce que chaque modèle soit ajouté à la combinaison pour améliorer la prédiction. Dans ce cas, le pas sera orienté dans la direction du gradient de la fonction perte, lui-même approché par un arbre de régression ou de classification.

2.3.5.3 Une variante : *eXtreme Gradient Boosting (XGBoost)*

L' *eXtreme Gradient Boosting* constitue une généralisation du *Gradient Boosting Machine*. Il est à l'heure très populaire car permet d'obtenir de bonnes performances dans différentes situations (classification ou régression).

En effet, *XGBoost* apporte des avantages intéressants par rapport au *Gradient Boosting Machine*. D'abord, son implémentation est parallèle, ce qui minimise le temps d'entraînement comparé à celui du *Gradient Boosting Machine*. Ensuite, le *XGBoost* permet l'utilisation d'autres algorithmes sous-jacents, comme les modèles linéaires contrairement au *Gradient Boosting Machine* qui n'implémente que des arbres de régression. Aussi, de manière similaire au *Gradient Boosting Machine*, *XGBoost* construit des arbres en séries en vue de minimiser le biais, tout en contrôlant la variance.

Résumé de la démarche mise en place au cours du mémoire

Modélisations envisagées

Comme nous l'avons mentionné auparavant, les techniques actuarielles usuelles de tarification en non-vie visent à modéliser la fréquence des sinistres ainsi que leurs coûts en fonction des différentes caractéristiques du contrat. La prime pure se déduit en calculant le produit de la fréquence par le coût moyen.

Dans le cadre de ce mémoire, le portefeuille dont nous disposons est un portefeuille automobile en assurance responsabilité civile. L'enjeu de ce mémoire est de comprendre et d'expliquer la sinistralité en assurance automobile pour notre garantie. Pour cela, nous modéliserons le **nombre de sinistres** ainsi que le **coût du sinistre** à travers une approche économétrique classique à savoir le **GLM**. Les résultats seront ensuite comparés à quatre méthodes issues de l'apprentissage statistique : **CART, *Random Forest, Gradient Boosting Machine (GBM)* et XGBoost**. Enfin, nous intégrerons à notre portefeuille des données comportementales issues de la télématique afin d'étudier leur impact sur la sinistralité.

L'importance des variables contribuant à chaque modèle sera mise en exergue.

Démarche globale

Le schéma ci-dessous permet de présenter de manière très succincte la démarche mise en place au cours de ce mémoire :

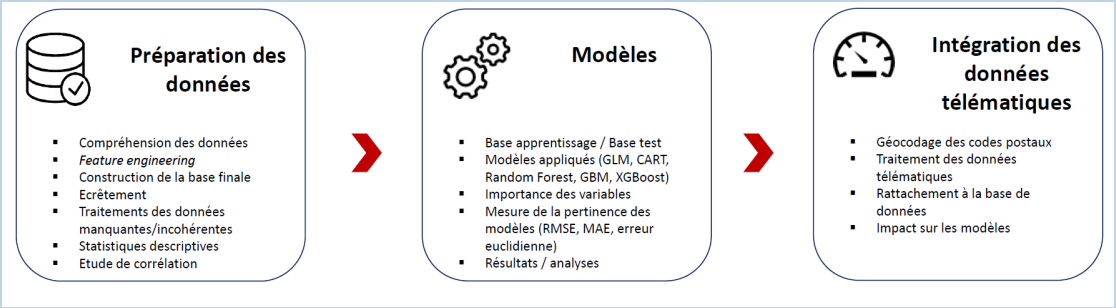
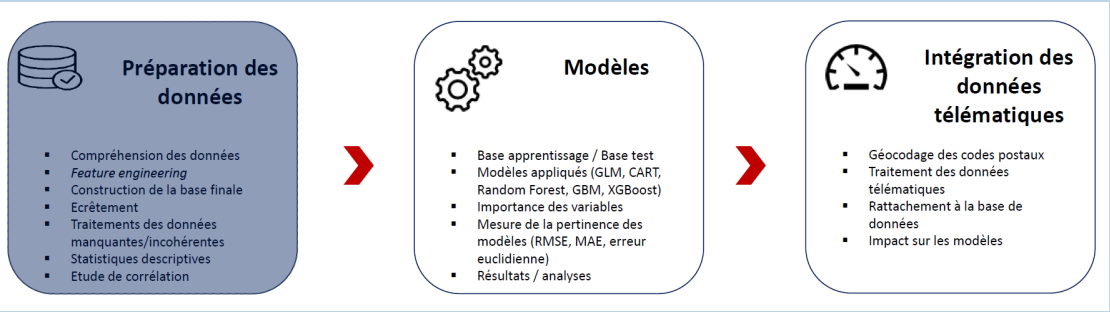


FIGURE 2.7 – Présentation de la démarche globale suivie

Comme dans tout projet *Data Science*, deux phases principales sont à considérer. La première consiste à préparer la base de données et la deuxième est liée à l'utilisation optimale des modèles. Nous intégrerons dans la dernière partie, des données externes liées à la télématique afin d'évaluer à nouveau les prédictions de nos modèles et mesurer le pouvoir explicatif de ces nouvelles variables comportementales sur la sinistralité.

Chapitre 3

Etude et préparation des données



Le traitement et l'analyse de la base de données constituent une étape essentielle avant d'établir les modèles tarifaires. En effet, nous devons construire en amont une base de données adaptée qui permettra d'identifier les variables cibles et d'avoir un premier avis sur les variables discriminantes avant d'entamer la modélisation. Cette analyse reposera sur une étude de corrélation.

L'objet de ce chapitre est de présenter la méthode de construction de la base d'étude, toutes garanties confondues, ainsi que les différents retraitements et tests de cohérence effectués.

Pour rappel, nous nous focaliserons dans la suite de l'étude uniquement sur la garantie **RC automobile**.

3.1 Présentation des données

Le jeu de données qui nous a été remis pour l'étude correspond au portefeuille automobile d'un assureur de taille moyenne allant de 2008 à 2017. Ce jeu de données est constitué de deux bases sous format *SAS*, une base **"contrats"** et une base **"sinistres"** :

— La base **"contrats"** : Elle contient l'ensemble des informations relatives au contrat, au conducteur et à son véhicule, elle comporte 1 895 013 lignes ;

— La base **"sinistres"** : Elle dispose de l'ensemble des informations liées à la sinistralité, elle comporte 1 232 822 lignes.

3.1.1 La base "contrats"

La base **"contrats"** est construite de façon à ce qu'elle représente autant de lignes que de garanties souscrites.

Nous pouvons classer les données de la base **"contrats"** selon deux catégories :

- Les données liées au contrat ;
- Les données liées au conducteur et au sociétaire.

✗ **Les données liées au contrat :**

— La date de début de situation : elle représente la date où le contrat a été souscrit ;

— La date de fin de situation : elle représente la date où la version du contrat prend fin ;

— L'identifiant global : il est construit en concaténant «le numéro du sociétaire», « le numéro du contrat » et «le numéro de la version du contrat au moment du sinistre » ;

— La formule : elle correspond au niveau de garanties souscrit par l'assuré ;

— Parc : cette variable prend deux modalités, «P» s'il s'agit d'un parc automobile et «I» sinon ;

— La cotisation émise pour chaque garantie ;

— Le niveau de franchise ;

— La remise commerciale accordée ;

— Le code secteur : Il s'agit du numéro de secteur commercial de l'agence ayant souscrit le contrat.

✗ **Les données liées au conducteur et au sociétaire :**

— La date d'obtention du permis ;

— La date de naissance ;

— Le secteur professionnel ;

— La qualité professionnelle ;

- Le nom et le prénom ;
- Novice : Il s'agit d'une variable qui reflète le caractère novice du conducteur, c'est-à-dire si celui-ci a moins de 3 ans d'ancienneté de permis. Cette variable prend la modalité « NOVICE » et « PAS NOVICE » ;
- Les garanties souscrites ainsi que les cotisations associées ;
- Le code postal ;
- La zone géographique ;
- Les informations liées au véhicule :
  - La classe SRA : Il s'agit d'une nomenclature nationale qui est en fonction de la valeur du véhicule, elle représente la valeur à neuf du véhicule et permet l'estimation du coût de sinistres matériels en cas de perte totale,
  - Le groupe SRA : Il s'agit d'une nomenclature nationale qui reflète la dangerosité intrinsèque du véhicule, elle contribue à la détermination des tarifs des garanties responsabilité civile,
  - La carrosserie,
  - Le genre du véhicule : Camionnettes, véhicule spécialisé, etc.,
  - L'usage du véhicule : Cette variable prend deux modalités selon qu'il s'agisse d'un usage professionnel ou privé,
  - La puissance,
  - Le coefficient bonus-malus,
  - La date de mise en circulation.

3.1.2 La base "sinistres"

Dans la base **"sinistres"**, pour chaque garantie sinistrée, il y a autant de lignes que de mouvements effectués relatifs au sinistres comme : une intervention d'un expert, une réévaluation de la provision, une clôture etc. Ainsi, une ligne sélectionnée de la base **"sinistres"** constitue une image du sinistre à une date renseignée. Nous verrons dans le prochain paragraphe comment regrouper les données de la base **"sinistres"** afin qu'une ligne corresponde à un sinistre.

Comme pour la base **"contrats"**, les données de la base **"sinistres"** peuvent être catégorisées selon deux groupes :

- Les données liées au conducteur sinistré ;
- Les données liées au sinistre.

✗ **Les données liées au conducteur sinistré :**

- Le nom et le prénom ;
- La date de naissance ;
- La date d'obtention du permis.

✗ **Les données liées au sinistre :**

- L'identifiant global : similaire à l'indentifiant global mentionné dans la description de la base **"contrats"** ;
- Le numéro du sinistre ;
- La date de survenance du sinistre ;
- La date d'ouverture du sinistre : il s'agit de la $1^{ère}$ date de saisie du sinistre ;
- Le code clôture : il s'agit d'un code qui détermine si le sinistre est ouvert, clos ou annulé ;
- La date de mouvement : un mouvement peut correspondre à un paiement, une réouverture etc ;
- Le code de la garantie sinistrée ;
- Les différents coûts : détaillés en paiements, recours, provisions pour paiements, prévisions de recours et le coût total.
- Honoraires : il s'agit des frais d'expertise si un expert a été mandaté.

3.2 Construction de la base de données finale

Nous avons vu dans les paragraphes précédents les caractéristiques des deux tables dont nous disposons. Ces tables sont agrégées selon des règles différentes avec un identifiant unique qui permet de les relier : l'identifiant global. Le processus de constitution de la base finale est présenté sur le schéma 3.1.

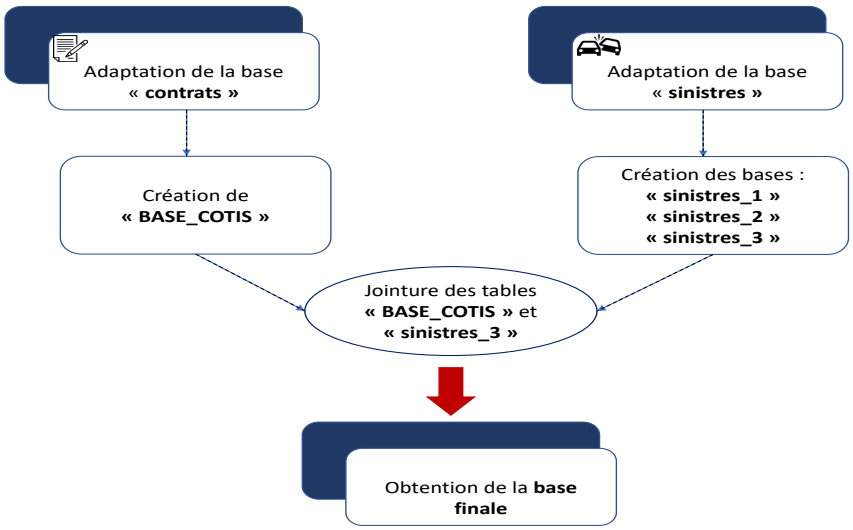


FIGURE 3.1 – Schéma simplifié de la construction de la base finale

Les bases **"sinistres_1", "sinistres_2", "sinistres _3"** et **"BASE_COTIS"** sont présentées par la suite.

3.2.1 Adaptation de la base "contrats"

Objectif :

La base **"contrats"** n'est pas agrégée par garantie mais par contrats. Nous souhaitons que la table ne contienne qu'une ligne par garantie et ce, pour chaque année de souscription en veillant à conserver toutes les autres informations associées à chaque garantie.

Avant d'entamer la modification de la base **"contrats"**, nous avons d'abord créé des variables qui nous seront utiles pour la suite de l'étude à partir de variables déjà présentes dans la base de données, ce procédé est connu sous le nom de *feature engineering*.

- Risque année :
  Il correspond à l'exposition du contrat, c'est-à-dire le prorata de présence de l'assuré dans le portefeuille. En effet, le risque année permet de tenir compte du fait qu'un sinistre déclaré par exemple pour une police exposée sur un mois est plus mauvais qu'un sinistre déclaré sur une police annuelle. Ce risque année est calculée de la façon suivante :

$$Risque\ Année = min\left(\frac{Date\ de\ fin\ de\ situation - Date\ de\ début\ de\ situation}{365}; 1\right)$$

- Age du conducteur :
  Cette variable est définie pour chaque année, par la relation suivante :

$$Age\ du\ conducteur = \frac{Date\ de\ début\ de\ situation - Date\ de\ naissance\ du\ conducteur}{365}$$

- Ancienneté du permis B :
  Elle est calculée pour chaque année, de la façon suivante :

$$Ancienneté\ du\ permis\ B = \frac{Date\ de\ début\ de\ situation - Date\ du\ permis\ B}{365}$$

- Ancienneté du véhicule :
  Elle est obtenue pour chaque année, par la relation suivante :

$$Ancienneté\ du\ véhicule = \frac{\substack{Date\ de\ début \\ de\ situation} - \substack{Date\ de\ mise \\ en\ circulation\ du\ véhicule}}{365}$$

Nous allons présenter désormais les traitements effectués sous *SAS* par des photographies « avant/après » afin de faciliter la compréhension de la configuration de la base **"contrats"**. Il est à noter que des cadrages ont été effectués avec le service comptable pour valider les coûts. Le nombre de sinistres a également été cadré avec le service de gestion et d'autres sources internes.

Nous prenons l'exemple d'un contrat « ID_GLOBAL=C0000523001201 » ayant souscrit à différentes garanties pour différentes années de souscription. La structure de la base **"contrats"** est représentée dans le tableau 3.1.

TABLE 3.1 – Structure de la base **"contrats"**

ID_GLOBAL	DATE_DEBUT_SITUATION	DATE_FIN_SITUATION	GARANTIE_100	GARANTIE_165	...	GARANTIE_N
C0000523001201	01/01/2008	31/12/2008	91.39	0	.	17.28
C0000523001201	01/01/2009	31/12/2009	88.36	0	.	23.91
C0000523001201	01/01/2010	31/12/2010	86.59	0	.	23.43
C0000523001201	01/01/2011	31/12/2011	86.59	0	.	23.42
C0000523001201	01/01/2012	31/12/2012	88.67	0	.	24.01

Nous agrégeons la base **"contrats"** par garantie en mettant en colonne la variable « COTISATION » qui correspond à la cotisation de chacune des garanties présentes dans le contrat.

La table " **BASE_COTIS** " dont l'extrait est présenté dans le tableau 3.2 est alors créée.

Table 3.2 – Structure de la **BASE_COTIS**

GARANTIE	ID_GLOBAL	DATE_DEBUT_SITUATION	DATE_FIN_SITUATION	COTISATION	ANNEE_PROD	RA	Autres variables
100	C0000523001201	01/01/2008	31/12/2008	91.39	2008	1	xxx
100	C0000523001201	01/01/2009	31/12/2009	88.36	2009	1	xxx
100	C0000523001201	01/01/2010	31/12/2010	86.59	2010	1	xxx
100	C0000523001201	01/01/2011	31/12/2011	86.59	2011	1	xxx
100	C0000523001201	01/01/2012	31/12/2012	88.67	2012	1	xxx
200	C0000523001201	01/01/2008	31/12/2008	14.4	2008	1	xxx
200	C0000523001201	01/01/2009	31/12/2009	16.86	2009	1	xxx
200	C0000523001201	01/01/2010	31/12/2010	16.52	2010	1	xxx
200	C0000523001201	01/01/2011	31/12/2011	16.52	2011	1	xxx
200	C0000523001201	01/01/2012	31/12/2012	16.93	2012	1	xxx
…	…	…	…	…	…	…	…

Nous pouvons remarquer qu'une ligne représente désormais une garantie et une année.

3.2.2 Adaptation de la base "sinistres"

Objectif :

La base **"sinistres"** est organisée selon les mouvements qui ont lieu. Les données correspondent aux montants cumulés des dépenses et recours et aux stocks de provisions. Nous avons donc, pour un même sinistre, plusieurs dates de mouvement. Le but est de modifier la base **"sinistres"** de sorte à ce qu'on obtienne qu'une ligne par sinistre.

La base **"sinistres"** pour un contrat dont l'identifiant est par exemple « ID_GLOBAL= C0971386000201 » est présentée dans le tableau 3.3.

TABLE 3.3 – Structure de la base **sinistres**

ID_GLOBAL	NUMERO_SINISTRE	CODE_GARANTIE	DATE_SURVENANCE	ANNEE_SURVENANCE	COUT_SINISTRE	DATE_MONTANT	ANNEE_MONTANT	Autres variables
C0971386000201	203000309	100	06/01/2008	2008	1500	31/01/2008	2008	xxx
C0971386000201	203000309	100	06/01/2008	2008	1500	13/05/2008	2008	xxx
C0971386000201	203000309	100	06/01/2008	2008	606	20/05/2008	2008	xxx
C0971386000201	203017006	170	20/07/2008	2008	0	31/07/2008	2008	xxx
C0971386000201	203017006	170	20/07/2008	2008	0	14/08/2008	2008	xxx
C0971386000201	203017006	170	20/07/2008	2008	0	20/08/2008	2008	xxx
C0971386000201	205013002	200	08/06/2010	2010	331	22/06/2010	2010	xxx
C0971386000201	203017006	350	20/07/2008	2008	100	31/07/2008	2008	xxx
C0971386000201	203017006	350	20/07/2008	2008	100	13/08/2008	2008	xxx
C0971386000201	203017006	350	20/07/2008	2008	189	14/08/2008	2008	xxx
C0971386000201	203017006	350	20/07/2008	2008	89	20/08/2008	2008	xxx
C0971386000201	203017006	530	20/07/2008	2008	0	20/08/2008	2008	xxx

Nous ne gardons donc que la date de mouvement la plus récente, une fois positionnés à la date d'étude.

Par ailleurs, les sinistres sans suite dont le coût est compris entre -1€ et 1€ sont supprimés. En effet, la présence de ces sinistres peut être dûe aux raisons suivantes :

- Une erreur existe sur la déclaration du sinistre de la part de l'assuré. Ce sinistre représente un sinistre « sans suite » car il a dû être enregistré lors de la déclaration, mais par la suite il a été refusé dans le traitement.
- La compagnie d'assurance note un état de sinistre « en cours » lorsqu'il s'agit d'identifier la nature de celui-ci en attendant les justificatifs de l'assuré ou les éléments de réponse de la compagnie adverse.
- Dans le cas où l'assuré est jugé non responsable du sinistre, un recours est alors fait. Une fois la nature du sinistre identifiée et le dossier complet, ce type de sinistre dont le coût est compris entre -1€ et 1€ est en état « clos ».

Le tableau 3.4 représente la base **"sinistres_1"** qui illustre les modifications citées ci-dessus.

TABLE 3.4 – Structure de la base **"sinistres_1"**

ID_GLOBAL	NUMERO_SINISTRE	CODE_GARANTIE	DATE_SURVENANCE	ANNEE_SURVENANCE	COUT_SINISTRE	DATE_MONTANT	ANNEE_MONTANT	Autres variables
C0971386000201	203000309	100	06/01/2008	2008	606	20/05/2008	2008	xxx
C0971386000201	203017006	350	20/07/2008	2008	89	20/08/2008	2008	xxx
C0971386000201	205013002	200	08/06/2010	2010	331	22/06/2010	2010	xxx

Certaines garanties ont leurs cotisations comprises dans la cotisation d'une autre garantie, nous affectons donc les sinistres à cette cotisation.

Exemple :

La garantie RC automobile qui est codée en 100 comprend également les garanties : 110 (Rente viagère), 150 (Recours assuré). Nous créons alors une variable « Garantie » qui prendra comme modalité "100" si le « CODE_GARANTIE » vaut 100, 110 et 150.

Nous calculons ensuite la somme des « COUT_SIN » de chacune des garanties. (voir tableau 3.5 qui représente la base ainsi modifiée que l'on nommera base **"sinistres_2"**.)

Table 3.5 – Structure de la base **"sinistres_2"**

ID_GLOBAL	NUMERO_SINISTRE	GARANTIE	DATE_SURVENANCE	ANNEE_SURVENANCE	CT_SINISTRE	GAR_100	…	GAR_N	Autres variables
C0971386000201	203000309	100	06/01/2008	2008	606	606	.	xxx	xxx
C0971386000201	203017006	350	20/07/2008	2008	89	0	.	xxx	xxx
C0971386000201	205013002	200	08/06/2010	2010	331	0	.	xxx	xxx

Nous faisons ensuite une jointure par les variables « ID_GLOBAL» et « GARANTIE » entre la **"BASE_COTIS"** présentée dans le tableau 3.2 et la base **sinistres_2"**, en respectant le fait que la date de survenance du sinistre doit être comprise entre la date de début et la date de fin de situation du contrat.

Par ailleurs, nous comptons pour chaque contrat précédent le nombre de sinistres ainsi que les coûts associés.

Avec ces traitements, nous obtenons une base que l'on nommera **"sinistres_3"** (voir tableau 3.6).

Table 3.6 – Structure de la base **"sinistres_3"**

ID_GLOBAL	DATE_DEBUT_SITUATION	DATE_FIN_SITUATION	GARANTIE	ANNEE_PROD	NB_SINISTRES	COUT_SIN	COUT_100	…	COUT_N
C0971386000201	01/01/2008	31/12/2008	100	2008	1	606	606	.	0
C0971386000201	01/01/2008	31/12/2008	350	2008	1	89	0	.	0
C0971386000201	01/01/2010	19/07/2010	200	2010	1	331	0	.	0

Dans le but d'obtenir une base complète avec toutes les informations sur le risque, nous relions à chaque garantie les sinistres appropriés. Pour cela, nous faisons un regroupement de la **"BASE_COTIS"** construite dans le tableau 3.2 et la base **"sinistres_3"** donnée par le

tableau 3.6 par « GARANTIE », « DATE_DEBUT_SITUATION », « DATE_FIN_SITUATION » et « ID_GLOBAL ».

Par ailleurs, nous rajoutons à notre nouvelle base toutes les caractéristiques du contrat et du véhicule (classe_SRA, code postal...).

Enfin, dans le but de modéliser le coût du sinistre, il a fallu isoler les sinistres graves des sinistres attritionnels afin de ne pas biaiser les modélisations. En accord avec l'entreprise, un seuil de 10 000€ a été fixé, et une variable « COUT_ECR » a été créée en prenant le minimum entre le coût du sinistre et 10 000€. Le pourcentage de la surcrête sera par la suite intégré dans le calcul de la prime finale.

Le tableau 3.7 représente la structure de la base d'étude ainsi constituée. Elle contient 17 152 977 observations.

TABLE 3.7 – Structure de **la base d'étude**

GARANTIE	ID_GLOBAL	DATE_DEBUT_SITUATION	DATE_FIN_SITUATION	COTISATION	ANNEE_PROD	RA	NB_SINISTRES	COUT_SIN	COUT_ECR	COUT_100	...	COUT_N	Autres variables
100	C0971386000201	01/01/2008	31/12/2008	155.48	2008	1	1	606	606	606	.	xxx	xxx
200	C0971386000201	01/01/2008	31/12/2008	20.45	2008	1	0	0	0	0	.	xxx	xxx
220	C0971386000201	01/01/2008	31/12/2008	24.52	2008	1	0	0	0	0	.	xxx	xxx
230	C0971386000201	01/01/2008	31/12/2008	12.26	2008	1	0	0	0	0	.	xxx	xxx
350	C0971386000201	01/01/2008	31/12/2008	170.97	2008	1	1	89	89	0	.	xxx	xxx
.	C0971386000201	01/01/2008	31/12/2008	7.55	2008	1	0	0	0	0	.	xxx	xxx
.	C0971386000201	01/01/2008	31/12/2008	5.75	2008	1	0	0	0	0	.	xxx	xxx
.	C0971386000201	01/01/2008	31/12/2008	20.6	2008	1	0	0	0	0	.	xxx	xxx
xxx	C0971386000201	01/01/2008	31/12/2008	22.02	2008	1	0	0	0	0	.	xxx	xxx
xxx	C0971386000201	01/01/2008	31/12/2008	9.32	2008	1	0	0	0	0	.	xxx	xxx
xxx	C0971386000201	01/01/2008	31/12/2008	1.1	2008	1	0	0	0	0	.	xxx	xxx

3.2.3 Définition de la base finale retenue pour l'étude

Maintenant que le base d'étude a été construite, il s'agit de ne retenir que la garantie **RC automobile** que l'on retrouve dans la colonne « GARANTIE= 100 ».

Par ailleurs, le périmètre que nous considérons est celui des véhicules à 4 roues (voitures particulières, véhicules automoteurs spécialisés et camionnettes) hors parc automobile pour les années allant de 2011 à 2017, car le volume de données est plus conséquent sur ces années.

3.2.4 Tests de cohérence

Afin de s'assurer de la cohérence interne entre les observations dans la base agrégée, nous effectuons différents tests dans le but de détecter les valeurs aberrantes, les valeurs manquantes et les valeurs incohérentes.

— Age du conducteur : Des tests de cohérence entre la date de permis et l'âge du conducteur ont été réalisés. Nous ne gardons que les conducteurs dont l'âge est supérieur ou égal à 18 ans et inférieur ou égal 110 ans.

— Ancienneté du permis B : Nous retenons pour cette variable les valeurs comprises entre 0 et 92 ans. Les valeurs manquantes ainsi que les valeurs contenues en dehors de cet intervalle représentent 0.01% de la base finale. Elles sont considérées comme étant aberrantes. Nous les retraitons en les remplaçant par la médiane.

— Ancienneté du véhicule : Nous retenons pour cette variable les valeurs comprises entre 0 et 40 ans. Les valeurs manquantes ainsi que les valeurs contenues en dehors de cet intervalle sont considérées comme étant aberrantes. Elles représentent 0.1% de la base finale. Nous les retraitons en les remplaçant par la médiane.

— Classe_SRA : Selon la convention SRA, les valeurs "X", "Y" et "Z" sont considérées comme étant aberrantes. Ces modalités, ainsi que les valeurs manquantes de la variable Classe_SRA, représentent 0.4%. Elle sont alors remplacées par la modalité la plus fréquente dans cette variable.

— Puissance : Les valeurs manquantes dans cette variable représentent 0.03%. Nous les remplaçons par la médiane.

— Le coefficient bonus-malus (CRM) : Des tests de cohérence entre l'ancienneté de permis B et le coefficient bonus-malus ont été réalisés. Le coefficient de 0.5 est en théorie acquis entre la $13^{ième}$ et la $14^{ième}$ année d'anciennecté de permis B. Cependant, il existe deux exceptions :

  - Lorsqu'un sociétaire entre en portefeuille en fin d'année (par exemple en Novembre), il bénéficie du bonus s'il n'a pas de sinistre dès le mois de Janvier de l'année qui suit.
  - Auparavant, lorsqu'un sociétaire avait passé la conduite accompagnée, il bénéficiait d'un bonus dès l'entrée en portefeuille, lui permettant ainsi de gagner un an.

  Par conséquent, ne seront considérées comme incohérentes que les anciennetés de permis inférieures à 12 ans avec un coefficient bonus-malus inférieur ou égal à 0.5.

— RA : Nous ne considérons dans la suite de l'étude que les risques années > 0.

Enfin, après les traitements décrits ci-dessus, nous obtenons une base finale **RC automobile** pour les exercices allant de 2008 à 2017 et comptant ainsi 317 236 lignes.

3.3 Statistiques descriptives

3.3.1 Analyse univariée

Il est utile d'étudier l'analyse univariée afin de permettre une appréciation des données à notre disposition. Cette analyse permet d'indiquer de surcroît les segments les plus à risque notamment en considérant la fréquence et du coût moyen correspondants.

Pour cela, nous présentons dans ce paragraphe quelques analyses univariées. D'autres sont aussi présentes en annexe B.

3.3.2 Classe SRA

A priori, la classe SRA semble pertinente pour la fréquence, un peu moins pour le coût moyen.

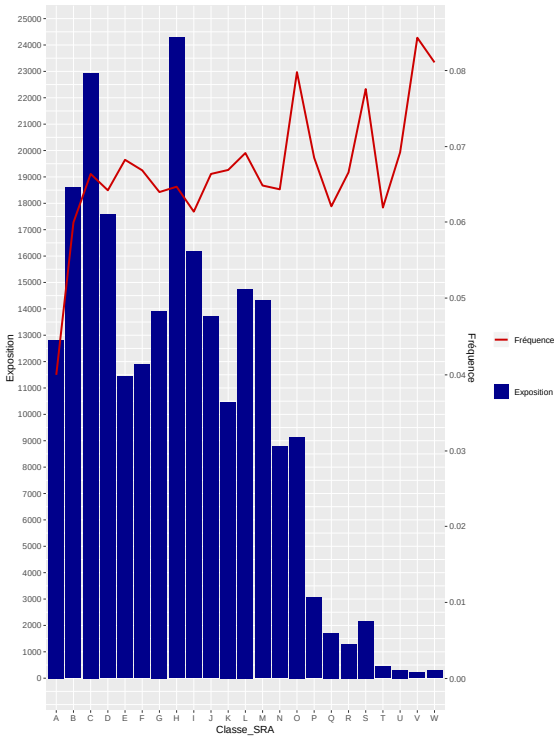


FIGURE 3.2 – Fréquence par Classe_SRA

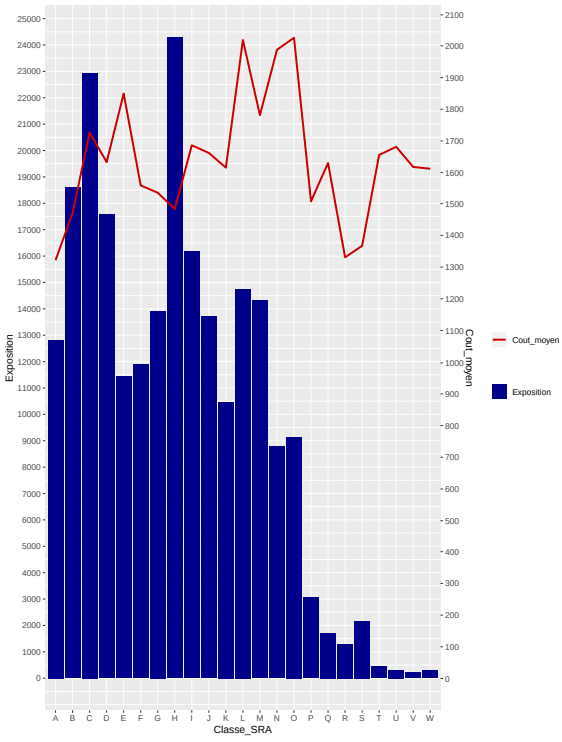


FIGURE 3.3 – Coût moyen par Classe_SRA

3.3.3 Le coefficient bonus-malus (CRM)

La fréquence de sinistralité augmente en fonction du coefficient bonus-malus de l'assuré. Par ailleurs, un assuré ayant un coefficient bonus malus de 0.5 aura un coût moyen plus faible qu'un assuré ayant un coefficient entre 0.5 et 0.9.

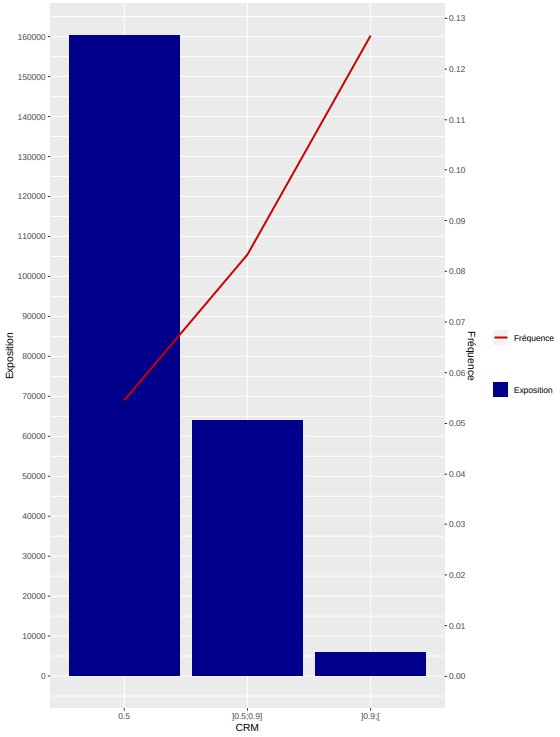


FIGURE 3.4 – Fréquence par CRM

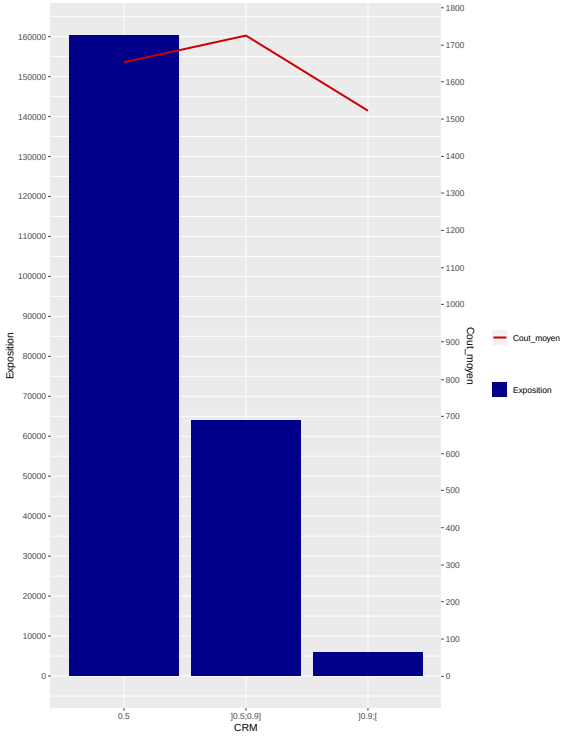


FIGURE 3.5 – Coût moyen par CRM

3.3.4 Zone

Le graphique suivant montre que les assurés de la zone 2 ont un coût moyen plus élevé (près de 1500€) que ceux de la zone 8 qui ont un coût moyen de 1700 €. A priori, la zone semble être pertinente pour la fréquence mais moins pour le coût.

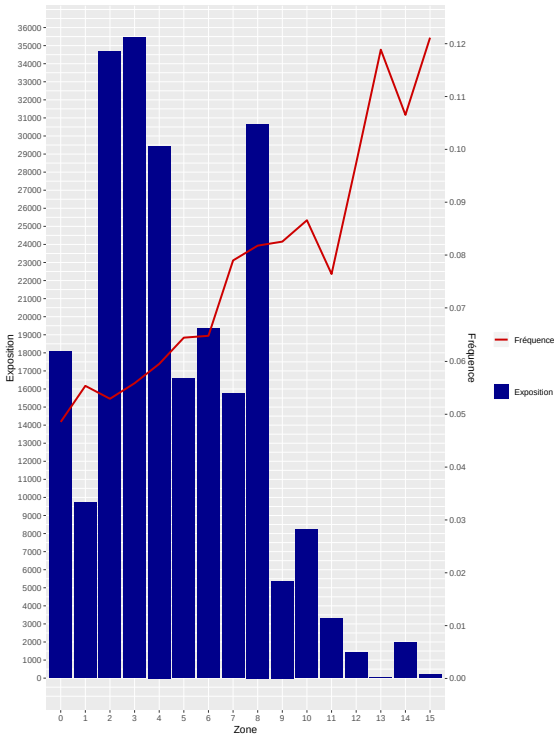


FIGURE 3.6 – Fréquence par Zone

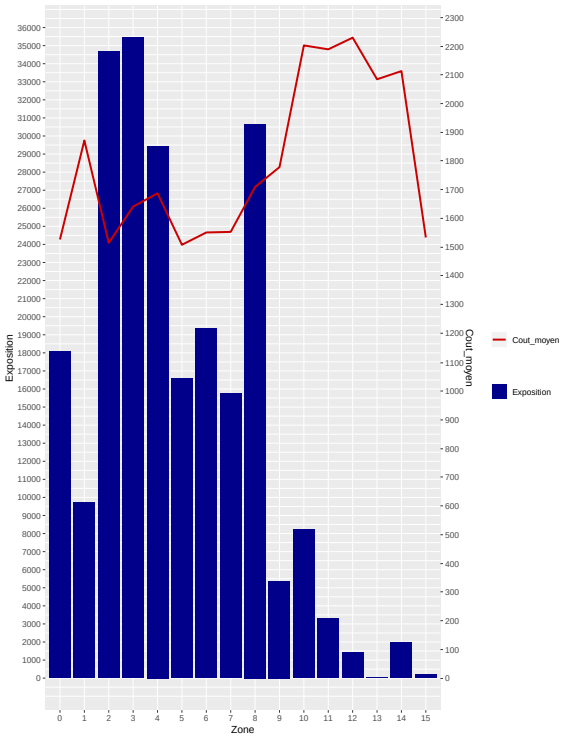


FIGURE 3.7 – Coût moyen par Zone

3.3.5 Analyse des corrélations

Dans cette partie, il s'agit de détecter les dépendances entre les variables explicatives en étudiant les interactions entre celles-ci. Afin de ne pas biaiser le modèle GLM, il est nécessaire de retirer certaines variables dans le cas d'une dépendance trop importante.

Pour cela, deux types de corrélations sont utilisées :

— Le ρ de Pearson ;

— Le V de Cramer.

3.3.5.1 Définitions

Le rho (ρ) de Pearson

La corrélation de Pearson a pour but de mesurer la dépendance entre deux variables quantitatives.

Considérons deux variables quantitatives X et Y prenant respectivement les valeurs $(x_i)_{i=1...n}$ et $(y_i)_{i=1...n}$. Le ρ de Pearson est alors donné par la formule suivante :

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

On calcule l'estimateur sans biais de ρ_{XY} qui est :

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

avec σ_X et σ_Y les corrélations empiriques respectives de X et Y et :

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

et $\hat{\sigma}_{XY}$ la covariance empirique de X et Y :

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

- **Le V de Cramer**

Le V de Cramer se base sur le test d'indépendance du χ^2. Ce test permet de mesurer l'existence d'une relation entre deux variables qualitatives. Les variables quantitatives sont assimilées à des variables qualitatives contenant un grand nombre de modalités. Dans notre base de données, toutes les variables quantitatives sont discrètes, nous calculons alors la corrélation entre une variable quantitative et une variable qualitative à l'aide du V de Cramer, en considérant les variables quantitatives comme étant des variables qualitatives avec autant de classes que nécessaire.

Le test d'indépendance du χ^2

L'hypothèse que vérifie le test du χ^2 en considérant deux variables qualitatives X et Y est : «H_0 : Les variables X et Y sont indépendantes ». Ainsi, il faut s'assurer que la statistique du test suit une distribution du χ^2 avec probabilité α (en pratique= 5%).
La statistique du test du $\chi 2$ est basée sur le tableau de contingence des deux variables considérées, elle est définie par la formule suivante :

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})}{\frac{n_{i.}n_{.j}}{n}}$$

où :

- n_{ij} : Nombre d'observations présentant la $i^{ième}$ valeur pour la première variable et la $j^{ième}$ valeur pour la seconde variable ;
- $n_{i.} = \sum_j n_{ij}$;
- $n_{.j} = \sum_i n_{ij}$;

- $n = \sum_{ij} n_{ij}$ est le nombre total des observations.

Cependant, la valeur de χ^2 ne permet pas de quantifier la dépendance entre deux variables, cette mesure varie entre 0 et $+\infty$, par conséquent, une mesure telle que le V de Cramer est davantage pertinente car elle permet de normaliser cette valeur par le χ^2 maximal théorique qui correspond à un tableau de contingence comportant une seule valeur non nulle par ligne et par colonne. Le V de cramer est alors compris entre 0 et 1 :

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n * min(r-1, c-1)}}$$

avec :

- r : Nombre de lignes du tableau de contingence ;
- c : Nombre de colonnes du tableau de contingence.

3.3.5.2 Mesures des corrélations entre les variables

Après une analyse univariée des variables, il serait intéressant de visualiser la structure sous-jacente des données. Afin d'éviter une perte de précision, il est conseillé d'étudier ces relations de colinéarité et de les traiter.

Pour cela, nous calculons une matrice de corrélation des variables explicatives ; cependant, celle-ci n'est calculable que pour les variables numériques à l'aide du ρ de Pearson.

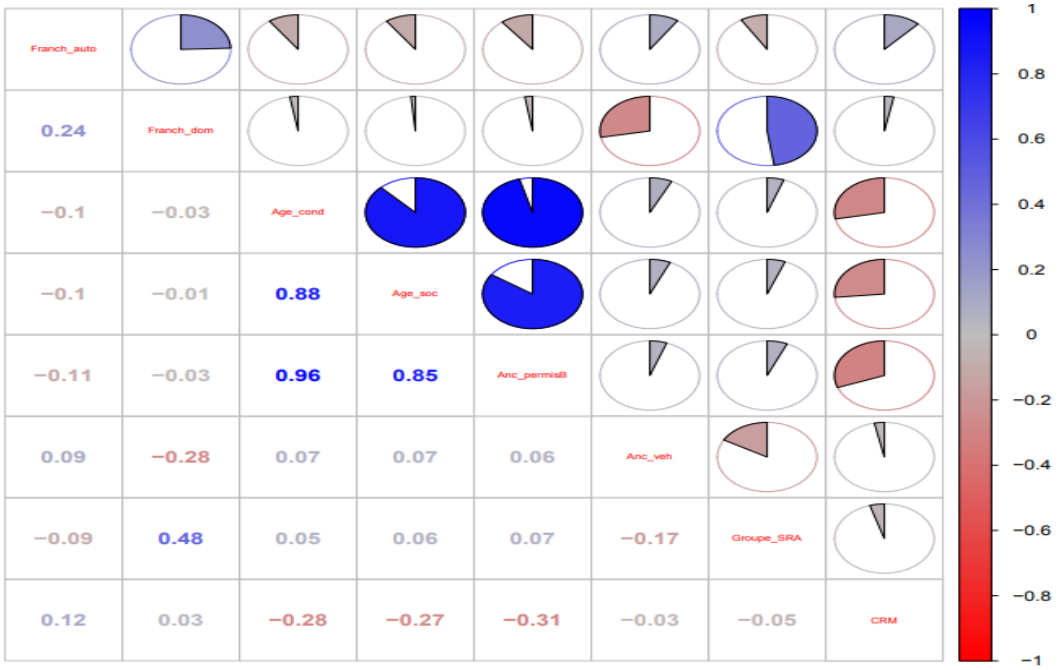


FIGURE 3.8 – Corrélation des variables numériques

Certaines corrélations paraissent intuitives (figure 3.8) :

— L'âge du conducteur et l'âge du sociétaire sont fortement corrélés, ceci s'explique par le fait que généralement, le conducteur est lui même sociétaire. Nous ne considérons dans la suite de notre étude que l'âge du conducteur ;

— L'âge du conducteur et l'ancienneté du permis présentent une corrélation quasi-totale, ce qui semble évident ;

— Nous observons également une corrélation négative entre le CRM et l'âge du conducteur. Plus une personne est jeune, plus son CRM est proche de 100% car lors d'une première assurance, le CRM est de 100%, puis il diminue avec le nombre d'années sans sinistre responsable ou augmente en cas de sinistre.

Un complément à cette première analyse serait de rajouter les variables qualitatives. A l'aide du **V de Cramer**, nous calculons la mesure d'intensité de l'association entre deux variables dès lors que l'une des deux est qualitative. Les variables quantitatives à notre disposition sont toutes discrètes et ont donc été transformées en variables qualitatives.

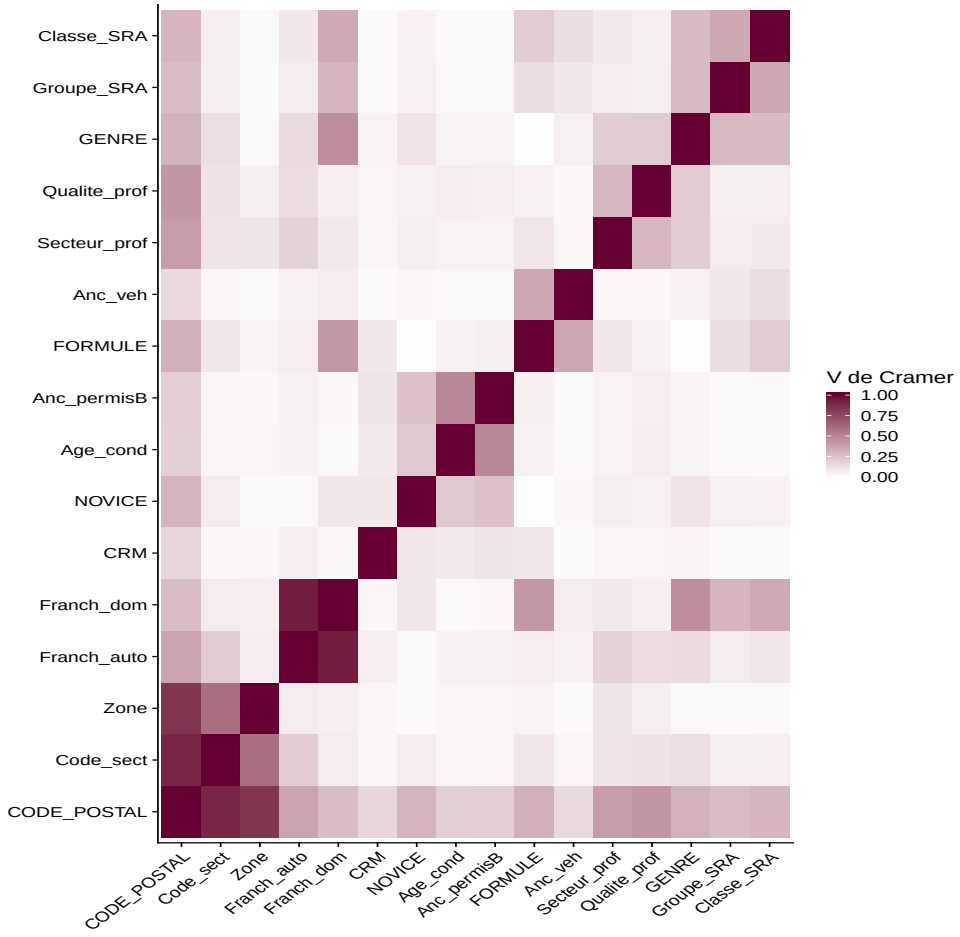


FIGURE 3.9 – Corrélation des variables à l'aide du V de Cramer

Cette analyse met en évidence quelques liens intuitifs entre certaines variables :

— Corrélation entre les variables concernant le véhicule (Groupe SRA, Classe SRA) ;

— La formule et l'ancienneté véhicule ;

Par ailleurs, certaines variables sont corrélées "naturellement" comme le code postal et la zone. Par la suite, le code postal n'est pas retenu dans l'étude car l'information est déjà contenue dans la variable zone.

Il sera néanmoins utile dans la deuxième étude qui s'articule autour de la tarification avec intégration de données externes. En effet, cette variable servira de clé pour faire une jointure entre notre base de données et celles des données télématiques récupérées.

Les conclusions liées aux différentes variables corrélées évoquées ci-dessus sont étudiées par la suite dans la modélisation par GLM.

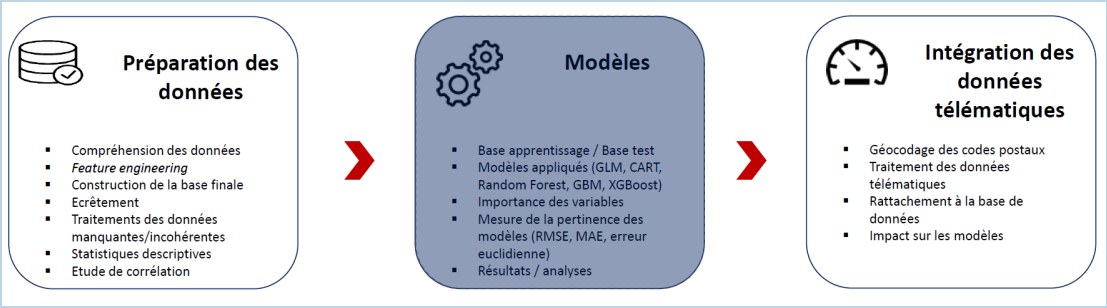
Les variables liées à la catégorie socio-professionnelle de l'assuré n'ont pas été retenues. Il s'agit de : code secteur, qualité professionnelle et secteur professionnelle. Cette décision est motivée par le fait de vouloir éviter le risque de fraude à la souscription. En effet, certains secteurs professionnels et certaines qualités professionnelles peuvent s'avérer proches, par conséquent, l'agent pourra aussi choisir de classer l'assuré dans la modalité la moins chère.

En résumé, les variables retenues dans la suite de la modélisation sont :

— Ancienneté du permis B ;
— Formule ;
— Genre ;
— Novice ;
— Franchise automobile ;
— Franchise dommage ;
— Age du conducteur ;
— Ancienneté du véhicule ;
— Groupe SRA ;
— Classe SRA ;
— CRM ;
— Zone ;
— Nombre de sinistres ;
— Coût du sinistre ;
— Risque année ;

Chapitre 4

Application des modèles sans données externes



Dans ce chapitre, nous souhaitons modéliser le **nombre de sinistres** ainsi que le **coût du sinistre** séparément pour la garantie RC automobile en usage professionnel à l'aide des modèles que nous avons présentés dans le chapitre 2. Le premier modèle que nous mettons en place est **le GLM**, il nous servira de modèle de base afin de le comparer aux modèles d'apprentissage statistique pour refléter leurs apports et leurs limites. Ceci est d'autant plus important que la concurrence, très rude entre les acteurs du marché de l'assurance automobile, incite à une sophistication toujours plus grande de la tarification et de l'analyse des risques. Ensuite, les algorithmes basés sur les arbres seront appliqués. Il s'agit de :

- CART ;
- Random Forest ;
- Gradient Boosting Machine (GBM) ;
- eXtreme Gradient Boosting (XGBoost).

Comme il s'agit d'un problème de régression, la comparaison de nos modèles se fera sur la base des métriques d'indicateurs d'erreurs : la racine de l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE) et l'erreur euclidienne. Un modèle sera dit plus pertinent qu'un autre si sa métrique est plus faible sur la base test.

Dans tout ce chapitre, la base d'étude a été divisée en deux : **une base d'apprentissage contenant 80% des données**, et **une base de test contenant 20% des données**. La modélisation du **coût du sinistre** concerne uniquement les sinistres dont le coût est strictement positif. Pour le **nombre de sinistres**, la modélisation se fait sur toute la base de données.

Les modèles GLM sont élaborés sur le logiciel *SAS*, pour la partie des méthodes d'apprentissage statistiques, les travaux sont effectués sous *R*.

4.1 Les modèles linéaires généralisés (GLM)

Depuis plusieurs années, les modèles linéaires généralisés (GLM) sont devenus incontournables dans plusieurs compagnies d'assurance. Elles sont de plus en plus nombreuses à s'équiper de logiciels coûteux afin de permettre une maniabilité plus aisée et d'alléger les temps de calculs.

Comme précisé au début de ce chapitre, nous présentons dans un premier temps la modélisation du **nombre de sinistres** puis celle du **coût du sinistre**. Pour cela, nous utilisons la fonction *GENMOD* de *SAS*.

4.1.1 Méthodologie

La démarche que nous suivons pour la modélisation est la suivante :

1. Choix de la loi pour la modélisation ;
2. Sélection des variables ;
3. Traitement des variables ;
4. GLM avec les variables retenues ;
5. Importance des variables ;
6. Pertinence du modèle : AIC, calcul de la RMSE, la MAE et l'erreur euclidienne.

4.1.2 Modélisation du nombre de sinistres avec GLM

Choix de la loi pour la modélisation du nombre de sinistres

Pour modéliser le **nombre de sinistres**, on utilise traditionnellement un modèle de Poisson. En effet, cette loi a l'avantage d'être bien adaptée pour modéliser un processus de comptage.

Pour la modélisation du **nombre de sinistres** déclarés sur un exercice (et ceci est valable aussi pour la modélisation du **coût du sinistre**), **la fonction lien « log »** est la mieux adaptée et ce pour les raisons suivantes :

- Elle permet d'assurer une structure tarifaire « multiplicative » qui est la forme la plus adaptée aux données d'assurance non-vie ;
- Le modèle tarifaire obtenu simplifie l'interprétation de l'influence des critères tarifaires sur la fréquence des sinistres.

Par ailleurs, nous avons comparé le AIC d'un modèle avec la loi Poisson et celui d'une loi Binomiale Négative :

Table 4.1 – Choix entre la loi Poisson et Binomiale Négative

	AIC	Deviance
Poisson	95 149.5251	72 118.8684
Binomiale Négative	95 151.5251	72 118.8684

Il n'y a pas une grande différence entre les AIC. Notre choix pour la loi Poisson se confirme. Ensuite, la variable **log(RA)** [1] a été créée. Cette variable sert d'offset [2] à la procédure *GENMOD* de *SAS*.

Sélection des variables

Maintenant que la loi et la fonction lien sont choisies, et avant d'entamer la modélisation, il est primordial d'analyser les variables explicatives. En effet, un modèle de régression est meilleur lorsque les variables explicatives sont significatives. Autrement dit, elles expliquent le phénomène étudié (ici le **nombre de sinistres**).

A l'aide de la procédure *GLMSELECT* sous *SAS*, nous omettrons les variables qui ne sont pas significatives (*p-value > 0.05*) en adoptant la méthode *sélection stepwise* expliquée dans la partie 2.1.4.4.

1. RA représente l'exposition, voir (3.2.1)
2. Lorsqu'une variable explicative X a une relation linéaire avec la variable réponse, cette variable X est déclarée en offset dans le modèle.

Les résultats sont présentés dans le tableau 4.2.

TABLE 4.2 – Résultat de la séletion des variables avec *GLMSELECT*

Stepwise selection summary						
Etape	**Effet saisi**	**AIC**	**AICC**	**BIC**	**Valeur F**	**P-value**
0	Intercept	-513 409.05	-513 409.05	-767 198.06	0.00	1.0000
1	CRM	-513 947.71	-513 947.71	-767 736.73	541.23	<.0001
2	Zone	-514 348.15	-514 348.15	-768 137.21	28.72	<.0001
3	Anc_veh	-514 490.80	-514 490.80	-768 279.84	144.68	<.0001
4	GENRE	-514 617.52	-514 617.52	-768 406.55	65.37	<.0001
5	FORMULE	-514 671.98	-514 671.98	-768 461.00	20.15	<.0001
6	Classe_SRA	-514 718.26	-514 718.24	-768 507.27	4.10	<.0001
7	Age_cond	-514 751.18	-514 751.16	-768 540.17	34.91	<.0001
8	Anc_permisB	-514 761.67	-514 761.65	-768 550.65	12.48	0.0004
9	Groupe_SRA	-514 770.62	-514 770.60	-768 559.61	10.95	0.0009
10	NOVICE	-514 776.18	-514 776.16	-768 565.16	7.55	0.0060
11	Franch_auto	-514 777.20*	-514 777.18*	-768 566.18*	3.02	**0.0820**

Comme la variable Franchise automobile a une *p-value >0.05*, nous décidons ne pas l'inclure dans le modèle.

Traitement des variables

✗ Age du conducteur et ancienneté du permis B :

La démarche retenue agit de telle sorte que si plusieurs variables sont fortement corrélées, leur influence n'est prise en compte qu'une seule fois grâce à un regroupement ou une élimination de variables.

Par exemple, l'âge du conducteur étant fortement corrélé à l'ancienneté de permis B, le modèle ne prendra pas en compte ces deux caractéristiques séparément (ou indépendamment), mais soit l'une, soit l'autre, soit le regroupement des deux en fonction de la pertinence dans le modèle.

Nous avons alors testé un modèle incluant la variable ancienneté de permis B sans la variable âge du conducteur, ensuite un modèle avec la variable âge du conducteur sans l'ancienneté du permis B et finalement le croisement des deux variables. Il en

ressort que le modèle ayant le plus petit AIC est celui contenant l'ancienneté du permis B sans l'âge du conducteur. **La variable « âge du conducteur » sera alors éliminée du modèle.**

✗ **Prise en compte du caractère « Novice » :**

Le caractère novice est regroupé avec le coefficient bonus-malus (CRM) et l'ancienneté du permis B. En d'autres termes, la méthode GLM ne fournira qu'un seul coefficient tarifaire pour chaque croisement des trois variables. Le tarif pour un cas type sera donc différent selon l'ancienneté de permis B, le coefficient bonus-malus et le caractère novice.

Il sera différencié de la façon suivante :

— Les novices en fonction de leur ancienneté de permis :
  - 1 an d'ancienneté de permis au plus,
  - de plus de 1 an à 2 ans inclus d'ancienneté de permis,
  - plus de 2 ans strictement d'ancienneté de permis.

— Les non novices en fonction de leur ancienneté de permis B et de leur coefficient bonus-malus.

✗ **Regroupement des modalités de la variable formule**

D'après l'analyse univariée de la variable formule et selon des stratégies commerciales, nous avons décidé d'effectuer les regroupements suivants :

Création de « Formule2 » telle que :

— Si la formule est dans les modalités 'EKO' ou 'ESS', alors la Formule2 prend la classe 'ESS' ;

— Si la formule est dans les modalités 'TRK' ou 'S', alors la Formule2 prend la classe 'TRK'.

✗ **Formule et ancienneté du véhicule**

Les variables formule et ancienneté du véhicule sont corrélées (35%). Nous suivons alors la même démarche que pour l'âge du conducteur et l'ancienneté de permis B. Le modèle avec croisement des deux variables donne le AIC le plus faible. Nous gardons ce croisement pour la suite de la modélisation du **nombre de sinistres**.

GLM avec les variables retenues

Grâce à la procédure *GENMOD* de *SAS*, nous mettons en application le modèle GLM pour **le nombre de sinistres**.

La variable « Groupe_SRA » est sortie avec une *p-value* > 0.05, nous l'avons donc ommise du modèle et nous avons relancé la procédure sans cette variable. Le critère AIC diminue, et toutes les variables restantes sont significatives. (Voir tableau 4.3)

TABLE 4.3 – Résultat de la procédure *GENMOD* pour la modélisation du **nombre de sinistres**

	Statistique LR pour Analyse de Type 1			
Source	**Ecart**	**DDL**	**Khi-2**	**Pr > Khi-2**
Intercept	73 643.6206			
PERM_CRM	72 795.4211	55	848.20	<.0001
VHL_FORM	72 425.6118	81	369.81	<.0001
GENRE	72 342.6655	2	82.95	<.0001
Zone	72 001.2763	15	341.39	<.0001
Classe_SRA	71 902.8046	22	98.47	<.0001

Importance des variables

L'importance des variables ayant contribué à la construction du modèle GLM pour le **nombre de sinistres** est déduite à l'aide du gain en AIC dans la figure 4.1.

Nous pouvons constater que le croisement des variables ancienneté de permis B, le CRM et le caractère novice (« PERM_CRM ») est prépondérant dans la modélisation du **nombre de sinistres**.

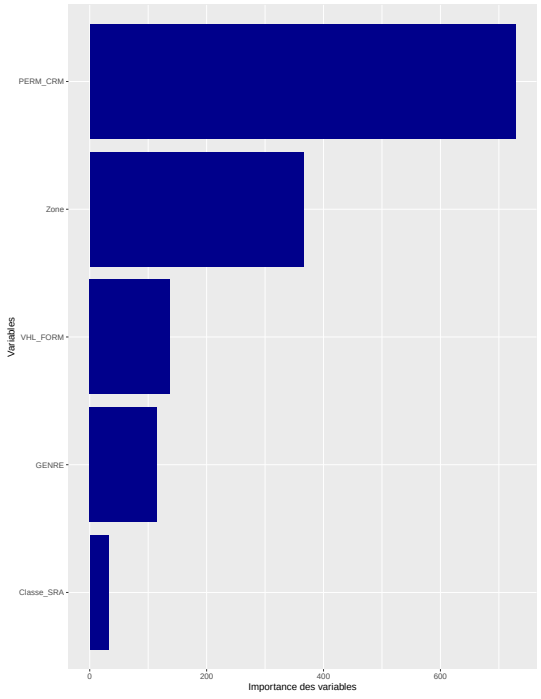


FIGURE 4.1 – Importance des variables dans la modélisation du **nombre de sinistres** par GLM

- **Mesure de la pertinence du modèle**

  Nous calculons les métriques de mesure de l'erreur de prédiction suivantes : RMSE, MAE et l'erreur euclidienne. (Tableau 4.4)

TABLE 4.4 – Erreurs obtenues pour la prédiction du **nombre de sinistres** avec GLM

	RMSE	MAE	Erreur euclidienne
Test	0.2210493	0.08853187	55.67941

4.1.3 Modélisation du coût du sinistre avec GLM

La méthode de mise en place du modèle pour le **coût du sinistre** est la même que celle du **nombre de sinistres**. Dans un souci de clarté, nous présenterons uniquement les résultats principaux.

- **Choix de la loi pour la modélisation du coût du sinistre**

  Traditionnellement, on utilise la loi Gamma pour modéliser le **coût du sinistre**, elle permet de modéliser des données continues positives à queue épaisse.
  Nous retenons comme pour le **nombre de sinistres**, la fonction lien « log ».

- **Sélection des variables**

  Une sélection des variables à l'aide de la procédure *GLMSELECT* a été effectuée.
  Les variables dont la *p-value* <0.05 ont été retenues. Il s'agit des variables CRM et zone.

- **Traitement des variables**

  ✗ **Regroupement de la variable CRM**
  Une variable CRM2 est créée telle que :
  - Si le CRM est égal à 0.5, CRM2 prend la classe "0.5" ;
  - Si le CRM est compris entre 0.5 et 0.6, CRM2 prend la classe "]0.5 ;0.6]" ;
  - Si le CRM est compris entre 0.6 et 0.9, CRM2 prend la classe "]0.6 ;0.9]" ;
  - Si le CRM est supérieur à 0.9, CRM2 prend la classe ">0.9".

  ✗ **Regroupement de la variable Zone**
  Les modalités de la variable zone ont été regroupées créant ainsi la variable zone2 :
  - Si la zone est dans 0, 2, 3, 5, 6, 7, 15 alors zone2 prend la classe "0-2-3-5-6-7-15" ;
  - Si la zone est dans 1, 4, 8, 9, 13 alors zone2 prend la classe "1-4-8-9-13" ;
  - Si la zone est dans 10, 11, 12, 14 alors zone2 prend la classe "10-11-12-14".

- **GLM avec les variables retenues**

Toujours à l'aide de la procédure *GENMOD* sous *SAS*, nous mettons en application le GLM pour **le coût du sinistre**. Le résultat est dans le tableau 4.5.

TABLE 4.5 – Résultat de la procédure *GENMOD* pour la modélisation du **coût du sinistre**

	Statistique LR pour Analyse de Type 1			
Source	**Ecart**	**DDL**	**Khi-2**	**Pr > Khi-2**
Intercept	-179 356.39			
zone2	-179 342.25	2	14.14	<.0001
CRM2	-179 321.97	3	20.28	<.0001

Importance des variables

Notre modèle pour **le coût du sinistre** ne comportant que deux variables, le gain en AIC ne sera pas représenté étant donné qu'elles sont toutes les deux importantes dans la construction du modèle. Notons que la variable CRM est présente dans les deux modèles.

Mesure de la pertinence du modèle

La RMSE, la MAE et l'erreur euclidienne sont les suivantes pour ce modèle :

TABLE 4.6 – Erreurs obtenues pour la prédiction du **coût du sinistre** avec GLM

	RMSE	**MAE**	**Erreur euclidienne**
Test	1449.768	969.6993	76054.11

Conclusion :

Les mesures d'erreurs calculées dans cette section pour le **nombre de sinistres** et le **coût du sinistre** seront comparées aux modèles d'apprentissage statistique. L'idée est de voir si ces algorithmes sont mieux adaptés à la structure tarifaire de notre portefeuille RC automobile en usage professionnel, autrement dit, si ces approches minimisent les erreurs.

4.2 Les modèles d'apprentissage statistique

Comme nous l'avons précisé dans la partie théorique des modèles, les modèles GLM sont très répandus dans la tarification, que ce soit en assurance automobile ou en assurance santé. Ces modèles ont la particularité de détecter les effets non linéaires et prennent en compte le caractère non gaussien dans la distribution des résidus.

Cependant, bien que performants par rapport aux modèles de régression classique, les contraintes imposées par ces modèles telles que les interactions entre les variables explicatives ou encore les contraintes liées à la structure du risque font que ces modèles peuvent conduire à des résultats non pertinents.

L'idée dans cette partie est d'appliquer les méthodes d'apprentissage statistique précédemment introduites dans la partie théorique, afin de prédire le **nombre de sinistres** et le **coût du sinistre** séparément. Il sera question de comparer les résultats de ces modèles avec ceux de notre modèle de base qui est le GLM.

4.2.1 Les arbres de décision : CART

4.2.1.1 Méthodologie

Afin de s'affranchir des limites que présentent les modèles GLM, nous construisons un modèle CART à la fois pour modéliser le **nombre de sinistres** et pour le **coût du sinistre**. Nous rappelons par ailleurs que le principe des arbres de régression est de déterminer des sous-groupes dans la population du portefeuille de l'assureur selon leurs variables explicatives. Chacun des groupes constitués est étiqueté par la valeur moyenne des sorties des observations qu'il contient.

L'algorithme CART est construit à l'aide de la fonction *rpart* du logiciel R représenté dans le tableau 4.7.

TABLE 4.7 – Application sur R de la fonction *rparts*

rpart (Formule, data=,method=, control=)	
Formule	Elle est de type : variable à prédire = prédicteur1+…+prédicteurN
data=	Base d'apprentissage
method=	*anova* pour un arbre de régression
control=	Paramètres optionnels pour contrôler l'arbre : rpart.control(cp=,xval=,...) où cp représente le paramètre de complexité et xval la validation croisée

L'approche de modélisation que nous suivons dans la construction de l'arbre est la même pour la modélisation du **nombre de sinistres** et celle du **coût du sinistre**. Elle est résumée dans les étapes suivantes :

1. **Construction de l'arbre maximal ;**
2. **Elagage de l'arbre maximal :**
   Avec le modèle saturé (arbre maximal), le nombre de feuilles est extrêmement élevé, ce qui peut conduire au phénomène de **surapprentissage** dans les arbres dû à une arborescence profonde et complexe, d'où l'intérêt d'élaguer ceux-ci. La complexité [3] de l'arbre peut être pénalisée grâce au paramètre de complexité **cp** qui s'apparente au α présenté au paragraphe 2.3.3.2. Ce critère est optimisé par la validation croisée effectuée à l'aide du paramètre **xval**.
3. **Visualisation de l'arbre optimal ainsi construit ;**
4. **Importance des variables** :
   L'arbre obtenu permet de mettre en évidence les variables les plus discriminantes ayant contribué directement à la construction et à la prévision. Toutefois, certaines variables n'apparaissent pas dans l'arbre optimal du fait qu'elles aient pu être, pour plusieurs noeuds, concurrentes des variables discriminantes.
5. **Test de pertinence :**
   Nous calculons comme pour le GLM, la RMSE, la MAE et l'erreur euclidienne sur la base test afin de juger de la pertinence du modèle.

4.2.1.2 Application : modélisation du nombre de sinistes

Construction de l'arbre maximal

Comme précisé dans la méthodologie ci-dessus, la construction d'un arbre maximal comporte un nombre très important de feuilles. Cet arbre (figure 4.2) est illisible et

3. On dit qu'un arbre est complexe lorsqu'il présente un grand nombre de noeuds.

présente peu d'intérêt dans le cadre de l'étude effectuée car l'arbre ne peut pas être appliqué sur une autre base que la base d'apprentissage. Il n'est donc pas robuste à l'égard de la base test.

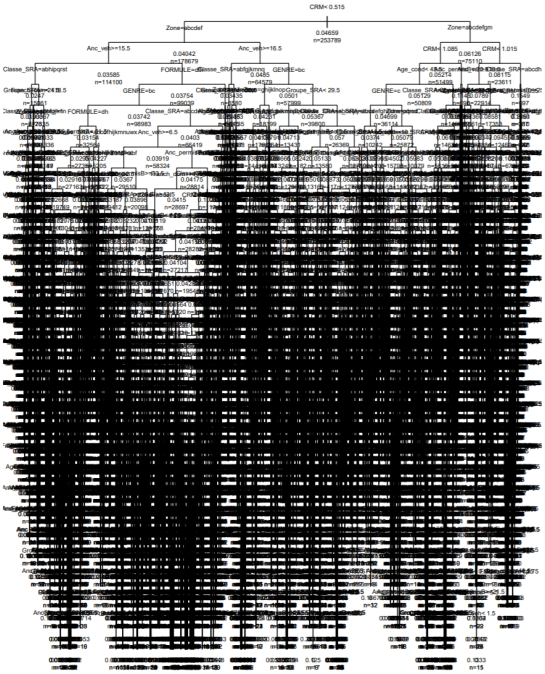


FIGURE 4.2 – Présentation de l'arbre saturé pour le **nombre de sinistres**

Elagage de l'arbre

Après avoir construit l'arbre saturé avec un **cp=0**, nous effectuons désormais une validation croisée en utilisant la fonction *rpart* de R. Nous gardons le paramètre **xval** fixé à la valeur par défaut= 10. La base d'apprentissage est donc découpée en 10 parties et le modèle apprend à chaque fois sur 9 d'entre elles, la dixième sert à la validation.

La démarche d'obtention du critère de complexité est une optimisation "à posteriori". La valeur de **cp** est déduite en adoptant la règle de Breiman qui consiste à choisir un **cp** inférieur à :

$$min\ (erreur\ relative) + std\ (erreur\ relative) \tag{4.1}$$

La valeur qui satisfait cette règle est **cp=0.0002202201**.

Visualisation de l'arbre final

L'arbre ainsi élagué est donné dans la figure 4.3

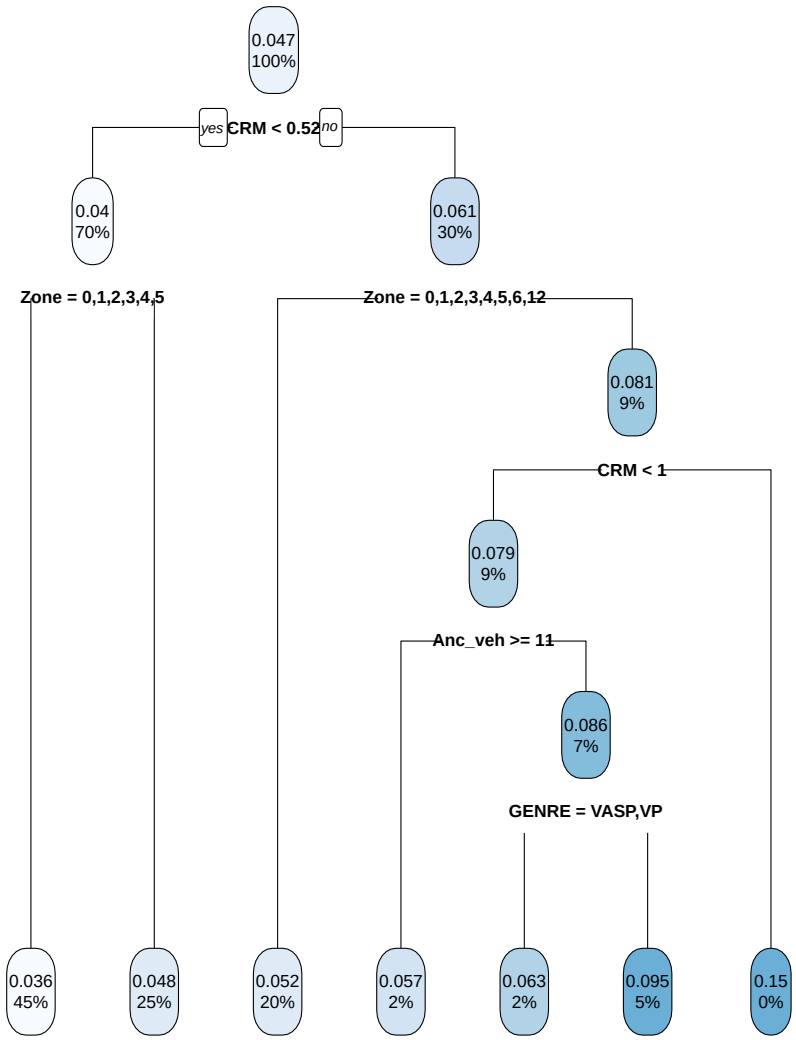


FIGURE 4.3 – Arbre retenu pour la modélisation du **nombre de sinistres**

Nous pouvons voir par exemple, que les personnes vivant dans les zones 0,1,2,3,4,5 et ayant un CRM < 0.52 sont ceux qui ont un nombre de sinistres le plus faible.

Par ailleurs, afin de permettre une meilleure visualisation de l'arbre, nous explicitons la lecture de celui-ci avec le schéma 4.4.

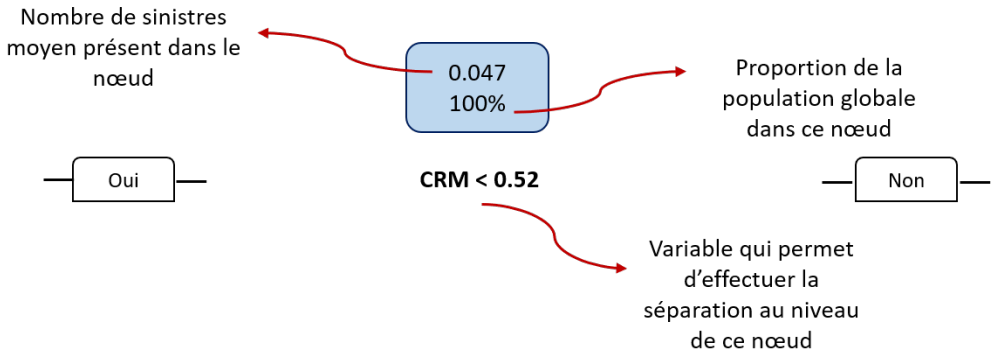


FIGURE 4.4 – Lecture de l'arbre

Importance des variables

Le graphique 4.5 met en exergue une certaine disparité entre les variables en termes d'importance dans le modèle. Nous pouvons constater que certains facteurs se démarquent nettement du reste des régresseurs dans la contribution à la construction de l'arbre final : CRM, zone et ancienneté du permis B. Il est à noter aussi que l'arbre construit ne comporte que 7 feuilles, donc 7 prédictions pour le **nombre de sinistres**.

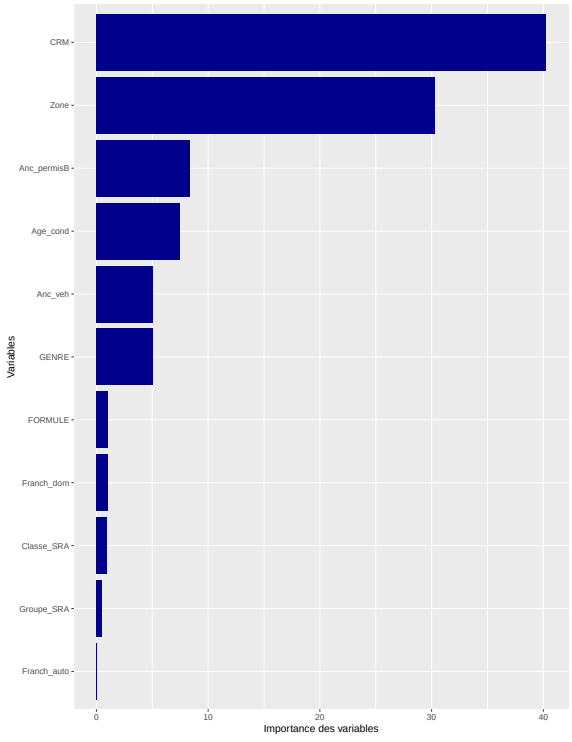


FIGURE 4.5 – Importance des variables dans la modélisation du **nombre de sinistres** par CART

- **Mesure de la pertinence du modèle**

  Afin de déterminer la pertinence du modèle, nous calculons les différentes métriques sur la base test (tableau 4.8) présentées dans le paragraphe théorique 2.3.2. Ces indicateurs nous permettront de comparer nos modèles.

TABLE 4.8 – Erreurs obtenues pour la prédiction du **nombre de sinistres** avec CART

	RMSE	MAE	Erreur euclidienne
Test	0.2217749	0.08937047	55.86219

Nous pouvons constater qu'il y a une légère différence entre les valeurs obtenues pour les erreurs en comparaison avec le GLM. Le modèle GLM pour le **nombre de sinistres** est légèrement meilleur avec une RMSE égale à 0.2210493.

4.2.1.3 Application : modélisation du coût du sinistre

Dans le but de mettre en place un modèle de régression d'arbre CART pour modéliser le **coût du sinistre**, le principe est similaire à celui employé pour le **nombre de sinistres**. En construisant l'arbre, nous obtenons la valeur moyenne des coûts pour chaque assuré tout en mettant en avant les variables explicatives qui ont le plus impacté l'estimation.

- **Construction de l'arbre maximal**

  L'arbre maximal est de nouveau construit. Nous maintenons les mêmes conclusions que pour le **nombre de sinistres.** Cet arbre est illisible et nous ne pouvons pas en tirer d'informations sur la prédiction. Nous ne le présenterons pas ici.

- **Elagage de l'arbre**

  Nous procédons à l'élagage de l'arbre maximal en utilisant la validation croisée avec une valeur de **xval** fixée à 10.
  A l'aide de la règle de Breiman, nous obtenons un **cp** optimal égal à **0.0015306.**

- **Visualisation de l'arbre final**

  Avec le **cp** obtenu dans l'étape de l'élagage de l'arbre, nous avons un arbre optimal qui ne comporte que **3 feuilles.**
  Il est toutefois à noter que retenir un **cp** plus faible permettrait d'avoir plus de complexité et donc plus de valeurs prédites avec une erreur relative aussi petite. Ainsi, à erreur minimale équivalente, nous retenons un **cp** plus faible car cela permet d'avoir plus de variables utiles au modèle tarifaire tout en maintenant le même niveau d'optimalité de l'arbre. Par ailleurs, cela permet de diviser la population en des sous-groupes homogènes et distincts évitant ainsi les risques d'anti-sélection.

Le paramètre de complexité finalement retenu est donc **cp=0,001460914**.
L'arbre final pour la modélisation du **coût du sinistre** est donné dans la figure 4.6.

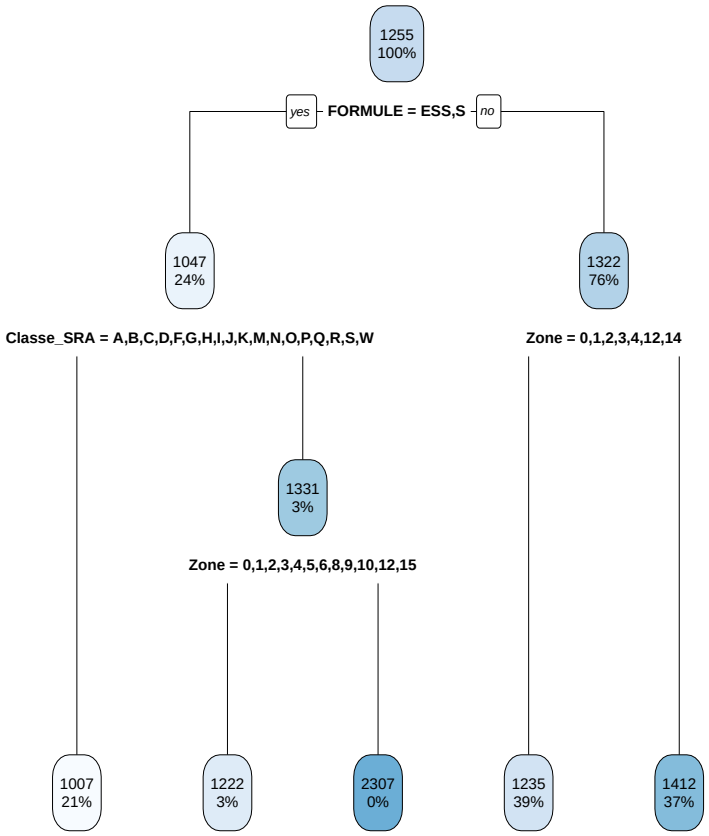


FIGURE 4.6 – Arbre retenu pour la modélisation du **coût du sinistre**

Par exemple, nous pouvons voir que les assurés ayant choisi la formule « ESS » ou « S » dont le véhicule appartient à la Classe_SRA « A » ont en moyenne un coût de sinistre plus faible, égal à 1007 €.

Importance des variables

Nous pouvons constater que les variables formule, zone et classe_SRA sont les plus importantes pour la modélisation du **coût du sinistre** avec CART. Ainsi, la zone est importante pour le modèle du **nombre de sinistres** comme pour le **coût du sinistre**.
La figure 4.7 illutre l'importance des variables ayant contribué à la construction de l'arbre.

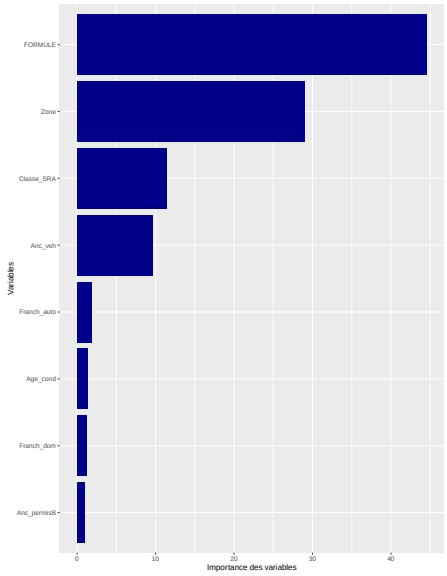


FIGURE 4.7 – Importance des variables dans la modélisation du **coût du sinistre** par CART

Mesure de la pertinence du modèle

Afin de déterminer la pertinence du modèle, nous calculons les différentes métriques présentées dans le tableau 4.9.

TABLE 4.9 – Erreurs obtenues pour la prédiction du **coût du sinistre** avec CART

	RMSE	MAE	Erreur euclidienne
Test	1447.239	952.3669	75921.45

La RMSE, la MAE et l'erreur euclidienne sur la base test sont plus faibles pour le CART comparé à celles du GLM. Nous concluons alors que le modèle CART est meilleur que le GLM pour la modélisation du **coût du sinistre**.

4.2.1.4 Limites du modèle CART

L'arbre CART présente deux limites majeures :

- La structure de l'arbre n'est pas robuste. En effet, si on change une variable, et plus particulièrement celles qui se trouvent au niveau plus haut de l'arbre, cela modifierait complètement la structure de l'arbre.

- Il peut être difficile d'appliquer cette méthode dans la tarification étant donné que le nombre de valeurs prédites est très réduit. Il serait donc plus intéressant d'appliquer ces méthodes dans une problématique de classification plutôt que de régression.

A présent, nous allons appliquer les modèles qui se basent sur les arbres de décision (méthodes de *bagging* et de *boosting*) pour modéliser le **nombre de sinistres** et le **coût du sinistre** afin de voir si la pertinence des modèles évolue par rapport aux GLM et CART.

4.2.2 Utilisation du package *h2o*

h2o est une bibliothèque open source d'algorithmes de maching learning et d'analyse de données que les entreprises peuvent utiliser pour construire des modèles sur des données massives. L'utilisation de cette plateforme intègre de manière native la parallélisation de plusieurs algorithmes standards de *machine learning* tels que le *Gradient Boosting Machine* ou encore le *Random Forest*. Il peut être implémenté sous le logiciel *R* ainsi que d'autres comme *Python* ou *Java*.

L'un des avantages de *h2o* réside dans le fait qu'il permet d'utiliser des *clusters* d'ordinateurs afin d'augmenter la capacité et la puissance de calculs sur de larges bases de données.

Nous utilisons alors ce package sous *R* dans la modélisation du **nombre de sinistres** et du **coût du sinistre** pour les algorithmes de *Random Forest* et *Gradient Boosting Machine*.

4.2.3 Les forêts aléatoires *(Random Forest)*

Rappelons que la méthode de *Random forest* s'appuie sur la construction de plusieurs arbres de régression sur la base d'échantillons de bootsrap, pour ensuite en faire une moyenne. La particularité des forêts aléatoires réside dans le fait de s'intéresser à un nombre réduit de variables pour chaque arbre, choisi aléatoirement au lieu de les prendre toutes en compte.

4.2.3.1 Méthodologie

A l'aide du package *h2o*, nous mettons en application l'algorithme du *Random Forest* comme décrit sur le tableau 4.10.

TABLE 4.10 – Fonction *Random Forest* sous h2o

h2o.randomForest (y=y.dep,x=x.dep, training_frame=,nfolds=, ntrees=,max_depth=,mtries=)	
y=y.dep	La variable à prédire : "Nb_sin" ou "COUT_ECR"
x=x.dep	Les variables explicatives
training_frame=	La base d'apprentissage sous format *h2o*
nfolds=	Nombre de partitions dans la validation croisée
ntrees=	Nombre d'arbres retenus
max_depth=	Profondeur maximale de chaque arbre
mtries=	Nombre de variables tirées aléatoirement à chaque nœud de chaque arbre

Dans le but d'éviter le surapprentissage et de chercher la meilleure forêt avec les paramètres optimaux, une étape d'hyperparamétrage est essentielle.

Le package *h2o* propose deux méthodes pour rechercher les meilleurs hyperparamètres : **Le Grid Search et le Random Search** (cf, paragraphe 2.3.2). La méthode adoptée dans ce mémoire pour le calibrage des hyperparamètres est celle du **Grid Search**.

Dans la modélisation du **nombre de sinistres** ainsi que le **coût du sinistre**, nous adopterons la méthodologie suivante :

1. Lancer un premier modèle sans hyperparamétrage et calculer les métriques RMSE, MAE et erreur euclidienne sur ce modèle afin de les comparer au modèle optimisé :
   - **mtries**= $p/3$, avec p le nombre de variables explicatives,
   - **nfolds**= 5, nombre de partitions dans la validation croisée,
   - **ntrees**= 100 arbres,
   - **max_depth**=30.
2. Procéder à l'hyperparamétrage à l'aide de la fonction *h2o.grid* :
   - **ntrees** : Nous testons un nombre d'arbres entre 10 et 400 avec un pas de 10, autrement dit, nous testons les valeurs $\{10, 20, 30, ..., 400\}$,
   - **mtries** : Nous testons les valeurs entre 5 et 18 avec un pas de 2, autrement dit, nous testons les valeurs dans $\{5, 7, 9, 11, 13, 15, 17\}$,
   - **max_depth** : Nous fixons la profondeur de l'arbre à 10.
3. Procéder à l'élaboration de la forêt aléatoire une fois que les hyperparamètres ont été déterminés. Cependant, afin de ne pas tomber dans le phénomène du surapprentissage, nous traçons un graphique qui reflète l'évolution de la RMSE en fonction du nombre d'arbres (en allant jusqu'à 500 arbres) et en précisant le paramètre **stopping_rounds** à 500, c'est-à-dire que si la RMSE n'évolue pas après 500 itérations, l'algorithme arrête

d'apprendre. Cette étape nous permettra de prendre une décision sur le paramètre **ntrees** qu'a renvoyé le **grid search**.

4. Déterminer l'importance des variables : Il est souvent reproché aux forêts aléatoires leur côté « boîte noire » du fait que le modèle produit n'ait pas la forme d'un arbre ou d'une combinaison de variables, dans lequel on pourrait voir directement l'importance d'une variable. Toutefois, il est possible de visualiser l'importance de chaque variable à l'aide de la fonction **h2o.varimp**. En effet, pour chaque arbre, nous mesurons l'erreur quadratique moyenne des individus n'appartenant pas à l'échantillon sur lequel l'arbre a été construit. Parmi ces individus *"out of bag"*, nous remplaçons la valeur de chaque variable par sa valeur pour un autre individu, ensuite l'algorithme mesure le taux d'erreur sur ces individus ainsi désorganisés. Une variable sera considérée importante dans la construction de l'arbre si la différence entre le taux d'erreur avant et après perturbation est importante. L'importance de la variable dans la forêt correspond à la moyenne sur l'ensemble des arbres de la différence des taux d'erreurs avant et après perturbation, elle est ensuite divisée par l'écart type de cette différence s'il n'est pas nul.
5. Mesurer la pertinence des prédictions.

4.2.3.2 Application : modélisation du nombre de sinistres

Modèle sans hyperparamétrage

Un premier modèle est lancé comme défini dans la méthodologie ci-dessus, avec un nombre de variables **mtries** égal au paramètre par défaut 4 (p étant égal à 12 variables, le paramètre par défaut $p/3$ vaut 4). Les métriques de pertinence du modèle arbitraire sont représentées dans le tableau 4.11.

TABLE 4.11 – Résultats du modèle arbitraire *Random Forest* pour le **nombre de sinistres**

	RMSE	MAE	Erreur euclidienne
Test	0.2331003	0.09364877	58.71491

Il s'agira alors d'améliorer la performance du modèle non optimisé ci-dessus en effectuant un hyperparamétrage.

Hyperparamétrage

Sur 118 modèles construits, les paramètres optimaux renvoyés par le **grid search** sont :

- **mtries**= 5 ;
- **ntrees**= 400.

Le graphique 4.8 affiche une stabilisation de l'erreur RMSE au-delà de 200 arbres, ce qui justifie donc que le choix de 400 arbres conduit à un modèle robuste. Par ailleurs, la RMSE obtenue avec un modèle à 200 arbres est supérieure à celle d'un modèle à 400 arbres. Nous maintenons alors le paramètre **ntrees** à 400.

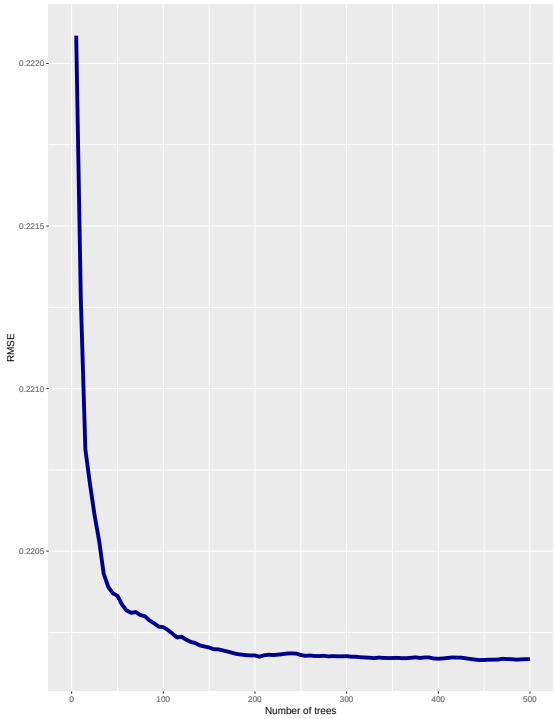


FIGURE 4.8 – Evolution de l'erreur en fonction du nombre d'arbres pour la modélisation du **nombre de sinistres**

Nous retenons dans la suite une forêt à 400 arbres pour la modélisation du **nombre de sinistres**.

Importance des variables

La classe SRA ressort importante au côté du CRM, de la zone et de l'ancienneté du véhicule.

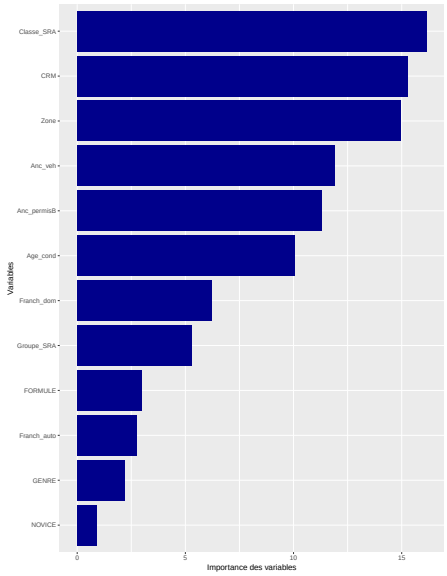


FIGURE 4.9 – Importance des variables dans la modélisation du **nombre de sinistres** par *Random Forest*

Mesure de la pertinence du modèle

Comme pour les modèles précédents, nous calculons les métriques RMSE, MAE et l'erreur euclidienne pour la forêt retenue. Les résultats sont présentés dans le tableau 4.12.

TABLE 4.12 – Erreurs obtenues pour la modélisation du **nombre de sinistres** avec *Random Forest*

	RMSE	MAE	Erreur euclidienne
Test	0.2216543	0.08918285	55.83182

Nous pouvons remarquer que la RMSE s'améliore par rapport à la RMSE du modèle arbitraire. Par ailleurs, en comparant aux résultats de CART, *Random Forest* semble légèrement meilleur.

4.2.3.3 Application : modélisation du coût du sinistre

La démarche mise en place pour la modélisation du **coût du sinistre** est en tout point identique à celle employée pour la modélisation du **nombre de sinistres**.

- **Modèle sans hyperparamétrage**

Nous commençons tout d'abord par un modèle arbitraire avec les mêmes paramètres que pour la modélisation du **nombre de sinistres**. Le résultat des erreurs est donné dans le tableau 4.13.

TABLE 4.13 – Résultats du modèle arbitraire *Random Forest* pour le **coût du sinistre**

	RMSE	MAE	Erreur euclidienne
Test	1482.045	990.2006	77747.36

Comme pour le **nombre de sinistres**, il s'agira d'améliorer les performances du modèle non optimisé à l'aide d'un calibrage des paramètres.

- **Hyperparamétrage**

Sur 160 modèles consrtuits, les paramètres optimaux renvoyés par le **grid search** sont :

- **mtries**=5 ;
- **ntrees**=400.

Par ailleurs, en traçant la RMSE en fonction du nombre d'arbres, nous obtenons la figure 4.10.

Ce graphique justifie encore une fois le choix du nombre d'arbres retenus par le **grid search** (400).

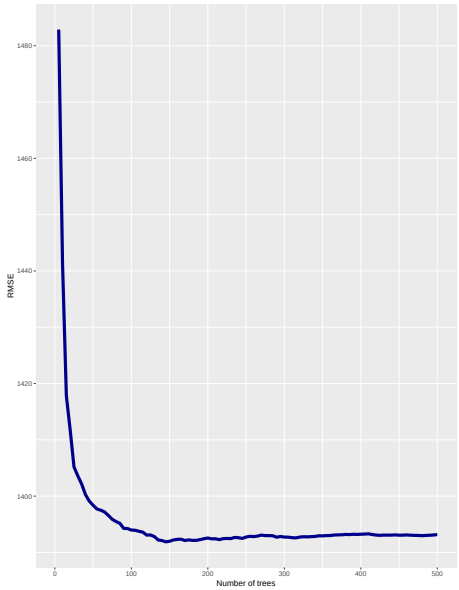
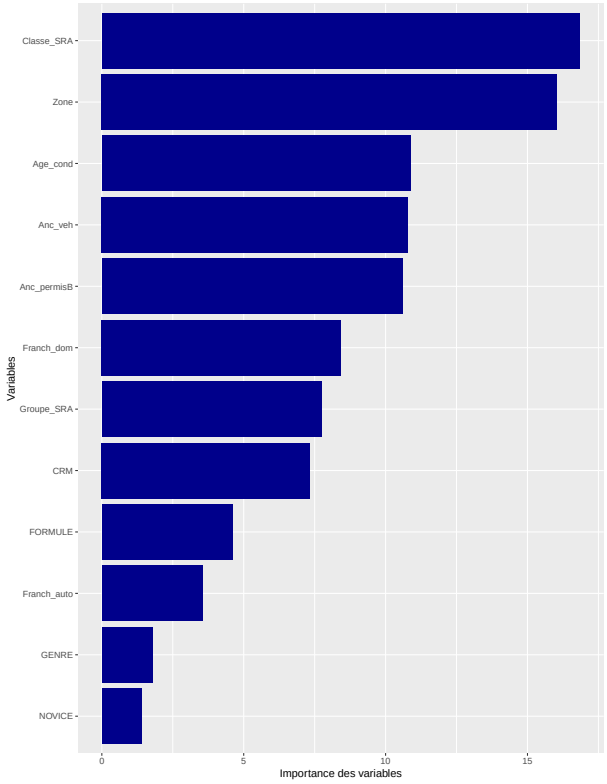


FIGURE 4.10 – Evolution de l'erreur en fonction du nombre d'arbres pour la modélisation du **coût du sinistre**

Importance des variables



Les variables les plus importantes sont précisées dans la figure 4.11.

La classe SRA est la plus importante, elle est suivie de la zone, de l'âge du conducteur et de l'ancienneté véhicule.

Contrairement à la modélisation du coût avec CART, la formule du véhicule n'est pas aussi importante.

FIGURE 4.11 – Importance des variables dans la modélisation du **coût du sinistre** par *Random Forest*

Mesure de la pertinence du modèle

Les erreurs calculées pour ce modèle sont représentées dans le tableau 4.14.

TABLE 4.14 – Erreurs obtenues pour la modélisation du **coût du sinistre** avec *Random Forest*

	RMSE	MAE	Erreur euclidienne
Test	1452.612	954.5662	76203.3

Le modèle avec hyperparamétrage donne de meilleurs résultats que le modèle sans optimisation des paramètres.
Par ailleurs, la RMSE est légèrement moins bonne que celle du modèle CART pour le coût.

4.2.4 Le *Gradient Boosting Machine* : GBM

Après l'implémentation des forêts aléatoires, nous mettons en oeuvre une deuxième méthode d'aggrégation : le boosting d'arbres de décision.

4.2.4.1 Méthodologie

Toujours à l'aide du package *h2o*, nous utilisons la fonction **h2o.gbm** pour modéliser **le nombre de sinistres** ainsi que **le coût du sinistre**. Cette fonction est explicitée sur le tableau 4.15.

TABLE 4.15 – Fonction *Gradient Boosting Machine* sous h2o

h2o.gbm (y=y.dep,x=x.dep, training_frame=,nfolds=, ntrees=,max_depth=,mtries=, learn_rate=)	
y=y.dep	La variable à prédire : "Nb_sin" ou "COUT_ECR"
x=x.dep	Les variables explicatives
training_frame=	La base d'apprentissage sous format h2o
nfolds=	Nombre de partitions dans la validation croisée
ntrees=	Nombre d'arbres retenus
max_depth=	Profondeur maximale de chaque arbre
mtries=	Nombre de variables tirées aléatoirement à chaque nœud de chaque arbre
learn_rate=	Aussi appelé *schrinkage*, il s'agit du taux d'apprentissage

Pour le GBM, nous mettrons en application les étapes suivantes :

1. Lancer un premier modèle sans hyperparamétrage et calculer les métriques RMSE, MAE et erreur euclidienne sur ce modèle dans le but de le comparer au modèle optimisé :
   - **mtries**= $p/3$, avec p le nombre de variables explicatives,
   - **nfolds**= 5, nombre de partitions dans la validation croisée,
   - **ntrees**= 100 arbres,
   - **max_depth**=30,
   - **learn_rate**= Le taux d'apprentissage correspond au pourcentage de correction qu'apporte un arbre supplémentaire à la prédiction. Ce taux est compris entre 0 et 1. *Exemple* : Supposons que l'arbre actuel prédise 200€ et que le deuxième arbre prédise 270 €, la correction serait de +70. Avec un learn rate de 1, la prédiction corrigée devient : 200+1*70=270 €. Empiriquement, il a été remarqué que de faibles valeurs du learn rate conduisaient à de meilleurs résultats. Pour le modèle sans hyperparamétrage, ce paramètre sera fixé à **0.01**.

2. Procéder à l'hyperparamétrage à l'aide de la fonction *h2o.grid* :
   - **ntrees** : Nous testons un nombre d'arbres entre 10 et 300 avec un pas de 10, autrement dit, nous testons les valeurs $\{10, 20, 30, ..., 300\}$,
   - **max_depth** : Nous fixons la profondeur de l'arbre à 10.
   - **learn_rate** : Nous testons les valeurs entre 0.01 et 0.05, autrement dit, nous testons les valeurs $\{0.01, 0.02, 0.03, 0.04, 0.05\}$.
3. Tracer comme pour les forêts aléatoires et une fois que les hyperparamètres ont été déterminés, l'évolution de la RMSE en fonction du nombre d'arbres dans le but de faire un choix sur le paramètre **ntrees** final à retenir ;
4. Déterminer l'importance des variables ayant contribué à l'élaboration du modèle.
5. Mesurer la pertinence du modèle retenu.

4.2.4.2 Application : modélisation du nombre de sinistres

Modèle sans hyperparamétrage

Un premier modèle est lancé comme défini dans la méthodologie ci-dessus, avec **mtries** égal au paramètre par défaut 4. Les métriques de pertinence du modèle arbitraire sont représentées dans le tableau 4.16.

TABLE 4.16 – Résultats du modèle arbitraire *Gradient Boosting Machine* pour le **nombre de sinistres**

	RMSE	MAE	Erreur euclidienne
Test	0.22404	0.08975003	56.43274

Hyperparamétrage

Sur les 65 modèles construits, les paramètres optimaux sont :

— **ntrees**= 140 ;

— **learn_rate**=0.05 ;

— **max_depth**= fixé à 10.

En traçant l'évolution de la RMSE en fonction de **ntrees** (figure 4.12), nous pouvons voir que l'erreur se stabilise un peu avant 100 arbres. Nous avons décidé de retenir un modèle avec **ntrees = 100 arbres** car la RMSE est la même entre un modèle à 100 arbres et un autre à 140.

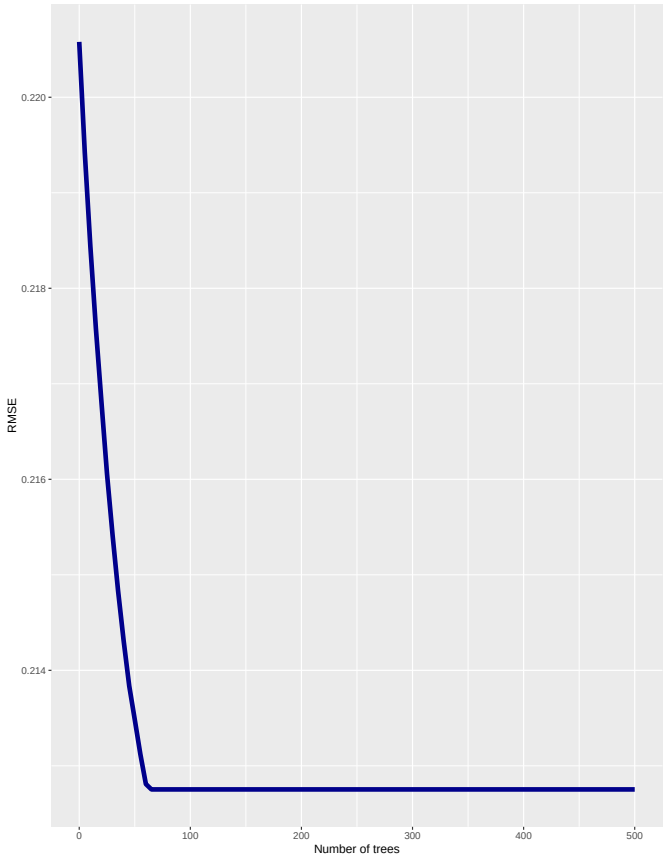


FIGURE 4.12 – Evolution de l'erreur en fonction du nombre d'arbres pour la modélisation du **nombre de sinistres**

Importance des variables

L'importance des variables pour la modélisation du **nombre de sinistres** est représentée dans la figure 4.13.

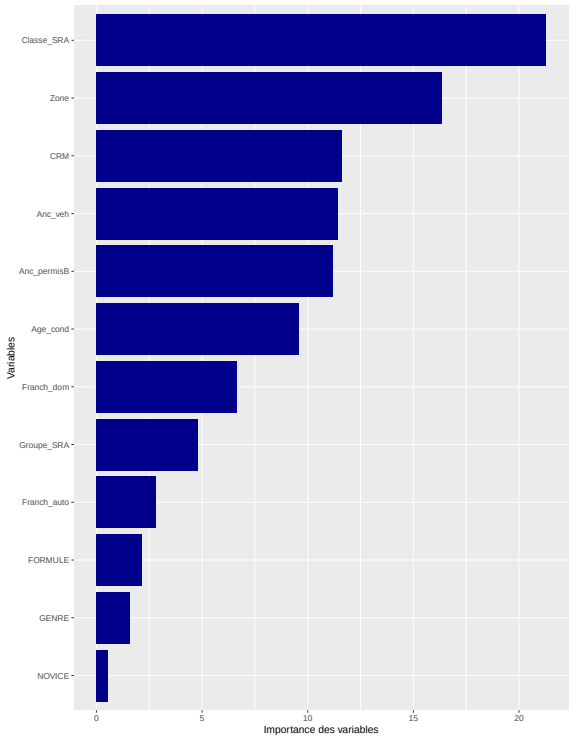


FIGURE 4.13 – Importance des variables dans la modélisation du **nombre de sinistres** par *Gradient Boosting Machine*

Il en ressort que comme pour le modèle *Random Forest*, la classe SRA, la zone, le CRM et l'ancienneté du véhicule sont les plus importantes pour prédire le **nombre de sinistres**.

Mesure de la pertinence du modèle

Les erreurs calculées pour ce modèle sont représentées au tableau 4.17. Nous pouvons remarquer que, bien que la RMSE de la base test s'améliore par rapport au modèle arbitraire construit auparavant, le *Gradient Boosting Machine* reste le modèle le moins pertinent pour la modélisation du **nombre de sinistres** en comparaison avec le *Random Forest* ou le *CART*.

TABLE 4.17 – Erreurs obtenues pour la modélisation du **nombre de sinistre** avec *Gradient Boosting Machine*

	RMSE	MAE	Erreur euclidienne
Test	0.2220635	0.08910118	55.93488

4.2.4.3 Application : modélisation du coût du sinistre

La démarche mise en place pour la modélisation du **coût du sinistre** est en tout point identique à celle employée pour modéliser le **nombre de sinistres**.

Modèle sans hyperparamétrage

Comme pour la modélisation du **nombre de sinistres**, nous commençons tout d'abord par un modèle arbitraire avec les mêmes paramètres que pour la modélisation du **nombre de sinistres**. Le résultat des erreurs est donné dans le tableau 4.18.

TABLE 4.18 – Résultats du modèle arbitraire *Gradient Boosting Machine* pour le **coût du sinistre**

	RMSE	MAE	Erreur euclidienne
Train	1131.183	749.5634	118725.8
Test	1459.52	967.5886	76565.73

Hyperparamétrage

Sur les 250 modèles construits, les paramètres optimaux sont :

- **ntrees**= 490 ;
- **learn_rate**= 0.05 ;
- **max_depth**= fixé à 10.

En traçant l'évolution de la RMSE en fonction de **ntrees** (figure 4.14), nous pouvons voir que l'erreur commence à se stabiliser à partir de **ntrees= 450**.

Nous avons décidé de retenir un modèle avec **ntrees = 490 arbres**, valeur ressortie avec le **grid search**. Nous pouvons s'attendre à ce que le modèle soit sujet au surapprentissage.

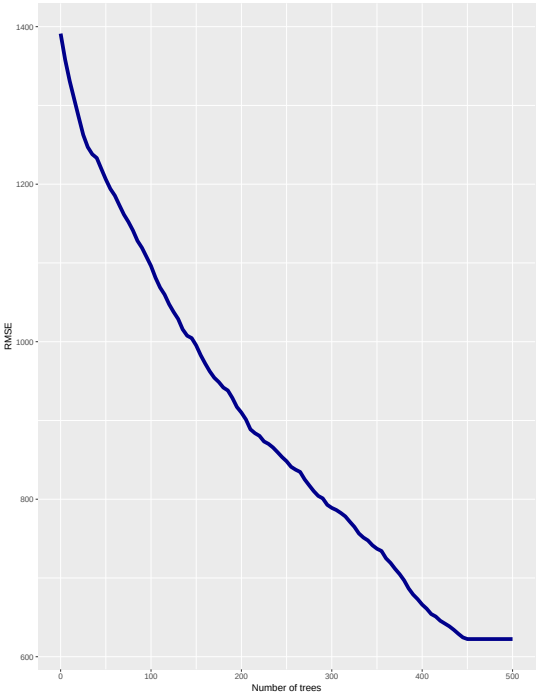


FIGURE 4.14 – Evolution de l'erreur en fonction du nombre d'arbres pour la modélisation du **coût du sinistre**

Importance des variables

L'importance des variables pour la modélisation du **coût du sinistre** est représentée dans la figure 4.15.

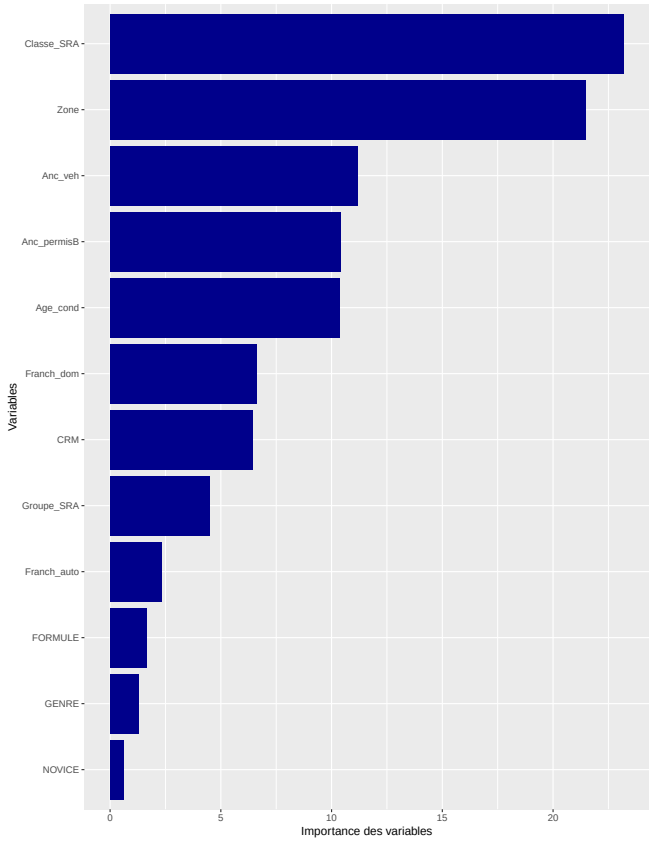


FIGURE 4.15 – Importance des variables dans la modélisation du **coût du sinistre** par *Gradient Boosting Machine*

La classe SRA est la variable la plus discriminante, suivie de zone, ancienneté véhicule et ancienneté du permis B.

Mesure de la pertinence du modèle

La RMSE, la MAE ainsi que l'erreur euclidienne augmentent sur la base test par rapport au modèle sans hyperparamétrage. Nous constatons que bien que la RMSE sur la base d'apprentissage baisse (elle passe de 1131.183 à 622.4372) en effectuant l'hyperparamétrage, le modèle **surapprend**. En effet la RMSE passe d'une valeur de 622.4372 à 1551.913. Par conséquent, ce modèle ne se généralise pas à la base test même si les paramètres sont optimisés.

Nous pouvons donc conclure que le *Gradient Boosting Machine* n'est pas adapté à la modélisation du **coût du sinistre**.

TABLE 4.19 – Erreurs obtenues pour la modélisation du **coût du sinistre** avec *Gradient Boosting Machine*

	RMSE	MAE	Erreur euclidienne
Train	622.4372	384.4409	65329.21
Test	1551.913	1045.911	81412.59

4.2.5 *eXtreme Gradient Boosting* : XGBoost

L'*eXtreme Gradient Boosting* généralise le GBM à d'autres fonctions que des arbres de régression, il a aussi la particularité d'utiliser une formalisation du modèle plus régularisée pour limiter le surapprentissage. Ceci lui permet généralement de donner de meilleurs résultats que le GBM.

Avant de mettre en oeuvre cet algorithme, il est essentiel de rendre binaire ($\{0, 1\}$) les variables catégorielles. En effet, cette étape est indispensable pour l'application du XGBoost. Un traitement a été effectué dans ce sens.

4.2.5.1 Méthodologie

La fonction XGBoost est présentée dans le tableau 4.20.

TABLE 4.20 – Fonction *XGBoost* sous R

xgboost(data=,label=,params=,nrounds=)	
data=	Base d'apprentissage sans la variable à prédire
label=	Base d'apprentissage avec la variable à prédire
params=	Liste des paramètres à optimiser : • **eta** : Le taux d'apprentissage compris entre 0 et 1 • **max_depth** : Profondeur de l'arbre • **gamma** : La régularité du modèle, plus le paramètre sera grand et plus le modèle sera lissé • **colsample_bytree** : Pourcentage de variables prises en compte dans la construction de chaque arbre • **min_child_weight** : Ce paramètre correspond au nombre minimum d'observations présentes dans chaque nœud pour poursuivre le développement de l'arbre.
nrounds=	Le nombre d'itérations, ici c'est le nombre d'arbres à implémenter

Nous suivrons la méthodologie suivante pour la modélisation du **nombre de sinistres** et du **coût du sinistre** :

1. Lancer un premier modèle par défaut avec :
   - **eta**=0.3,
   - **max_depth**=6,
   - **gamma**= 0,
   - **colsample_bytree**= 1,
   - **min_child_weight**=1,
   - **nrounds**= 100,
   - **subsample**= 1, ce paramètre correspond au pourcentage d'observations prises en compte dans chaque arbre.
2. Pocéder à l'hyperparamétrage afin de calibrer les paramètres ;
3. Déterminer l'importance des variables ;
4. Mesurer la pertinence des prédictions.

4.2.5.2 Application : modélisation du nombre de sinistres

Modèle sans hyperparamétrage

Les résultats du modèle par défaut sont donnés dans le tableau 4.21.

TABLE 4.21 – Résultats du modèle arbitraire XGBoost pour le **nombre de sinistres**

	RMSE	MAE	Erreur euclidienne
Test	0.2226936	0.08925384	56.09361

Hyperparamétrage

De nouveau le paramètre **nrounds** a été fixé à 100. A l'aide de la fonction *train* du package *caret*, nous testons les valeurs suivantes pour les paramètres cités dans la méthodologie :

- **eta** : 0.1, 0.05, 0.06, 0.07 ;
- **max_depth** : 2, 4, 6, 8, 10 ;
- **gamma** : 0, 1 ;
- **colsample_bytree** : 0.8,
- **min_child_weight** : 0,
- **subsample** : 0.75.

Par ailleurs, une validation croisée est réalisée sur la base d'un découpage en 5 partitions, les graphiques (4.16) suivants présentent les RMSE suivant les différentes valeurs des paramètres testés.

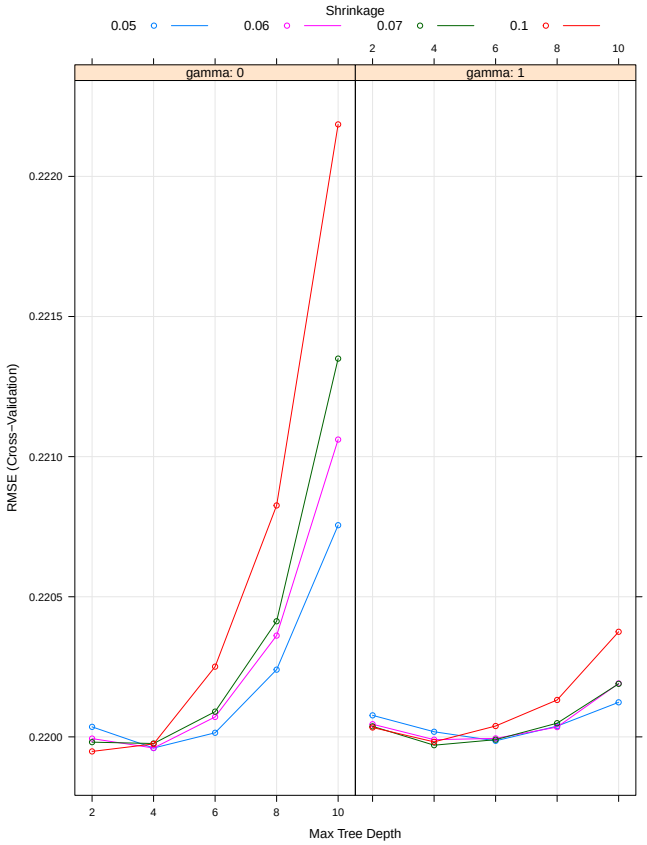


FIGURE 4.16 – RMSE en fonction des différents paramètres XGBoost pour le **nombre de sinistres**

Ainsi, les paramètres qui minimisent la RMSE sont :

- **eta** : 0.1 ;
- **max_depth** : 2 ;
- **gamma** : 0 ;
- **colsample_bytree** : 0.8,
- **min_child_weight** : 0,
- **subsample** : 0.75.

Comme le nombre d'itérations est assez sensible et qu'une légère modification de celui-ci pourrait modifier les résultats, nous effectuons une seconde validation croisée en prenant comme métrique la RMSE afin de fixer le paramètre **nrounds**. La figure 4.17 permet de visualiser cette évolution.

Nous pouvons constater que la RMSE commence à se stabiliser à peu près après 40 itérations. Afin d'éviter le surapprentissage, nous fixons la valeur de **nrounds** à 40.

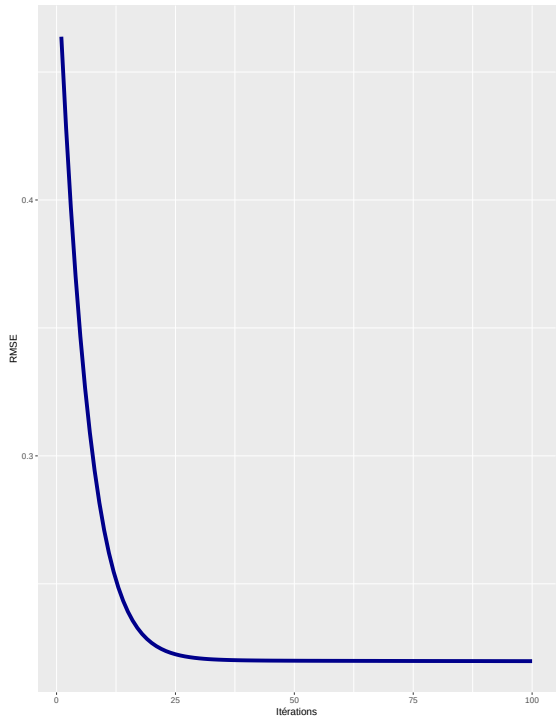


FIGURE 4.17 – Evolution de la RMSE sur la base d'apprentissage en fonction du nombre d'itérations

Importance des variables

XGBoost permet aussi d'analyser l'importance des variables. Il calcule le gain au niveau de chaque noeud, celui-ci représente la contribution de la variable sélectionnée. Ainsi, en parcourant tous les splits de tous les arbres, ces gains sont sommés et agrégés par variable.

La figure 4.18 précise l'importance de chacune des variables, avec également les modalités les plus importantes.

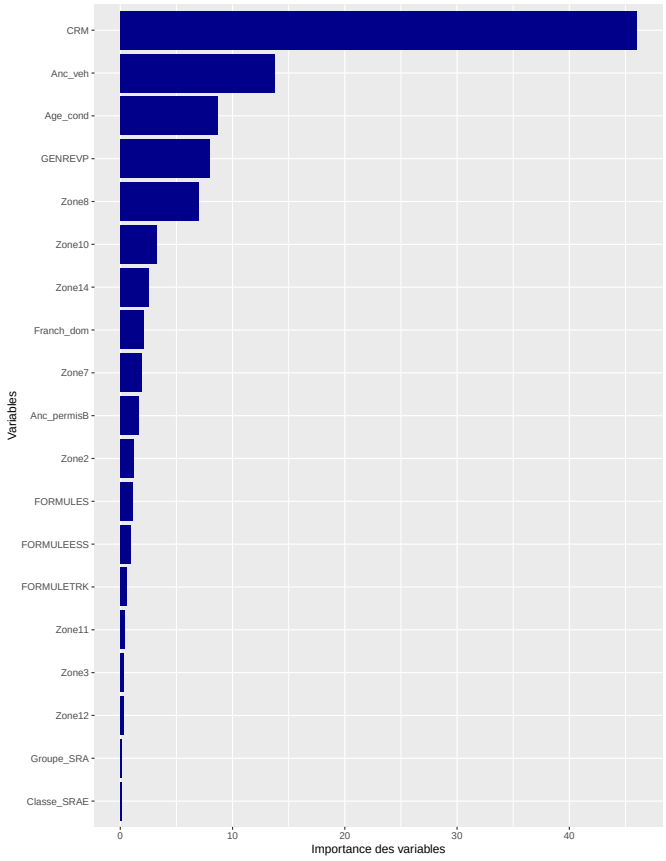


FIGURE 4.18 – Importance des variables dans la modélisation du **nombre de sinistres** par XGBoost

Les 3 variables les plus importantes dans le modèle du **nombre de sinistres** avec XGBoost sont : le CRM, l'ancienneté véhicule et l'âge du conducteur.

Mesure de la pertinence du modèle

Afin de mesurer la pertinence du modèle, nous calculons les trois métriques habituelles : RMSE, MAE et erreur euclidienne présentées dans le tableau 4.22.

TABLE 4.22 – Erreurs obtenues pour la modélisation du **nombre de sinistres** avec XGBoost

	RMSE	MAE	Erreur euclidienne
Test	0.2216571	0.095454	55.83252

Sur la base test, les trois métriques s'améliorent dans le modèle optimisé en comparaison avec le modèle arbitraire, sauf pour la MAE qui connait une légère hausse sur la base test.

4.2.5.3 Application : modélisation du coût du sinistre

L'approche pour la modélisation du **coût du sinistre** est en tout point similaire à celle du **nombre de sinistres**.

- **Modèle sans hyperparamétrage**

  Dans un premier temps, nous mettons en oeuvre un modèle sans optimisation des paramètres, les résultats sont donnés dans le tableau 4.24.

TABLE 4.23 – Rédultats du modèle arbitraire XGBoost pour **le coût du sinistre**

	RMSE	MAE	Erreur euclidienne
Test	1461.98	934.51	76694.75

- **Hyperparamétrage**

  Les mêmes combinaisons de paramètres pour le **nombre de sinistres** ont été testées, à savoir :

  — **eta** : 0.1, 0.05, 0.06, 0.07 ;
  — **max_depth** : 2, 4, 6, 8, 10 ;
  — **gamma** : 0, 1 ;
  — **colsample_bytree** : 0.8,
  — **min_child_weight** : 0,
  — **subsample** : 0.75.

  Par ailleurs, la validation croisée a été réalisée sur la base d'apprentissage en 5 partitions (5 folds).

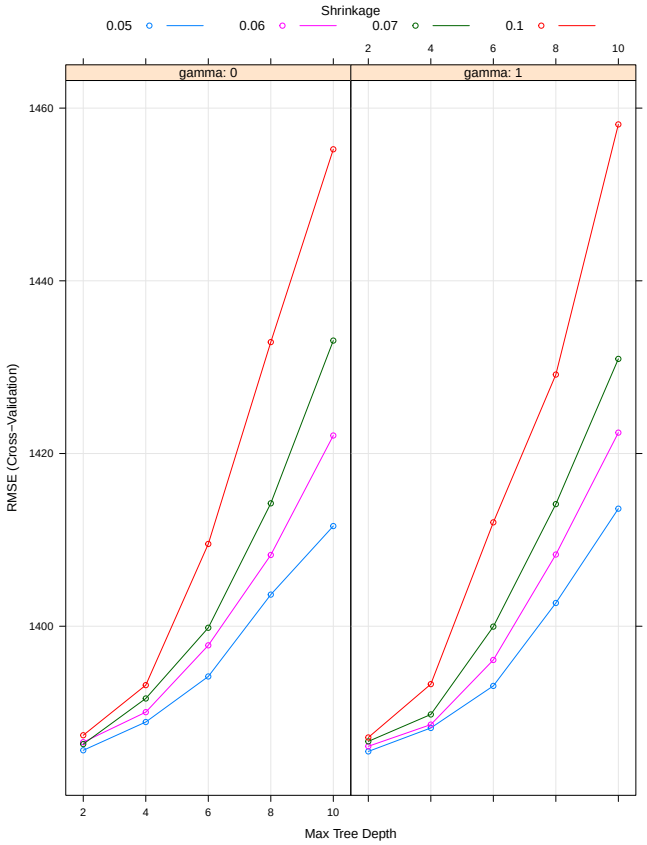


FIGURE 4.19 – RMSE en fonction des différents paramètres XGBoost pour le **coût du sinistre**

Nous avons retenu les paramètres suivants selon le critère de la RMSE :

- **eta** : 0.05 ;
- **max_depth** : 2 ;
- **gamma** : 1 ;
- **colsample_bytree** : 0.8,
- **min_child_weight** : 0,
- **subsample** : 0.75.

Le nombre d'itérations est fixé à l'aide d'une deuxième validation croisée. Le graphique 4.20 permet de visualiser l'évolution de la RMSE en fonction du nombre d'itérations :

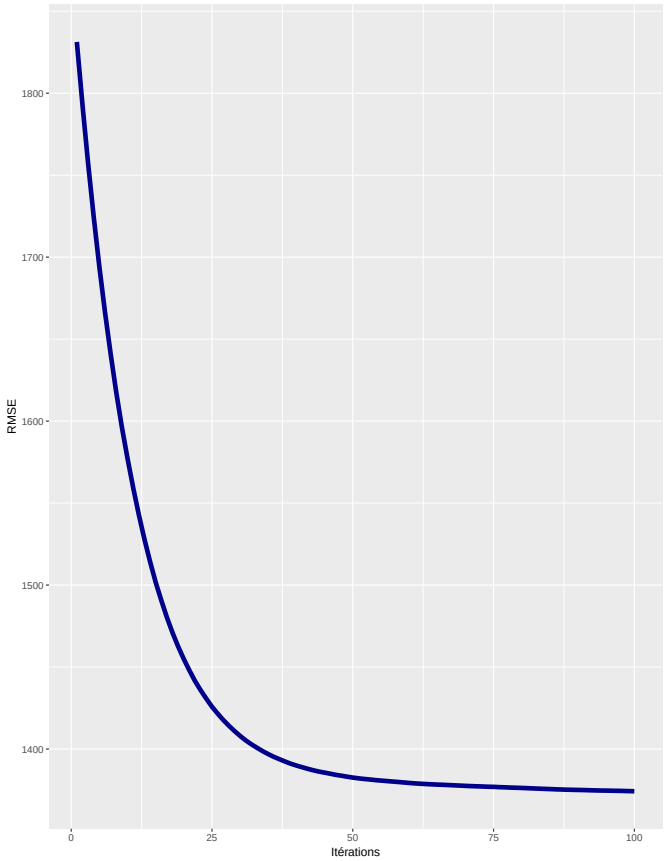


FIGURE 4.20 – Evolution de la RMSE en fonction du nombre d'itérations

La valeur de la RMSE diminue fortement au début, puis elle a tendance à ralentir. Nous avons fixé **nrounds** à 100.

Les paramètres du modèle ainsi fixés, nous représentons désormais l'importance des variables.

Importance des variables

La figure 4.21 permet de résumer l'importance des variables pour ce modèle avec une précision sur la modalité prise par chaque variable :

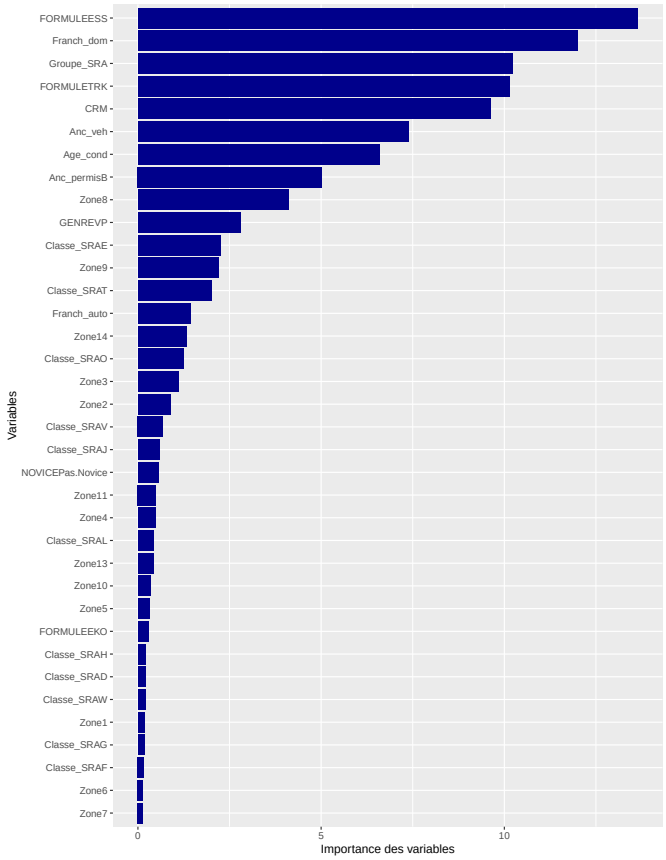


FIGURE 4.21 – Importance des variables dans la modélisation du **coût du sinistre** par XGBoost

La formule "ESS", la franchise dommage, le groupe SRA, le CRM, la formule tous risques "TRK" ainsi que l'ancienneté du véhicule sont celles qui contribuent le plus au modèle du coût.

Mesure de la pertinence du modèle

Avec les paramètres retenus, nous calculons comme pour les modèles précédents, la RMSE, la MAE ainsi que l'erreur euclidienne. Elles sont présentées dans le tableau 4.24.

TABLE 4.24 – Erreurs obtenues pour la modélisation du **coût du sinistre** avec XGBoost

	RMSE	MAE	Erreur euclidienne
Test	1448.542	950.323	75989.81

Nous pouvons constater que la RMSE et l'erreur euclidienne baissent par rapport au modèle arbitraire. La MAE est en légère hausse.

4.3 Synthèse et comparaison des modèles sans données externes

4.3.1 Comparaison des résultats des modèles

A présent, nous allons présenter dans ce paragraphe une synthèse de tous les modèles construits dans la partie précédente.

Le tableau 4.25 présente une comparaison des **RMSE** calculées auparavant sur la base test. Le choix de la RMSE est motivé par le fait que par sa définition, cet indicateur présente l'avantage de pénaliser plus fortement les fortes erreurs à travers la forme quadratique de la RMSE.

TABLE 4.25 – Synthèse des erreurs des modèles sur la base test

	Nombre de sinistres	Coût du sinistre
GLM	**0,2210493**	**1449,768**
CART	0,2217749	**1447,239**
Random Forest	**0,2216543**	1452,612
GBM	0,2220635	1551,913
XGBoost	0,2216571	**1448,542**

Concernant la modélisation du **nombre de sinistres**, nous constatons que le modèle dont l'erreur est la plus faible (RMSE= 0.2210493) est bien le **GLM**, il est suivi de **Random Forest** avec une valeur de RMSE égale à 0.2216543.

Quant à la modélisation du **coût du sinistre**, c'est le modèle **CART** qui l'emporte avec une valeur de RMSE égale à 1447.239, il est suivi de **XGBoost** dont la RMSE vaut 1448.542.

4.3.2 Importance des variables et comparaison graphique des modèles

4.3.2.1 Nombre de sinistres

Pour la modélisation du **nombre de sinistres**, d'après les tableaux présentés dans la section précédente, il ressort que le modèle **GLM** est meilleur, suivi de *Random Forest*. Le tableau 4.26 permet de résumer l'importance des variables qui ressortent pour la modélisation du **nombre de sinistres** pour chaque modèle, elles sont mises en évidence avec la couleur **bleu**. Par ailleurs, les variables qui apparaissent en premier dans chaque modèle sont présentées en **rouge**.

TABLE 4.26 – Tableau récapitulant les variables les plus importantes dans la modélisation du **nombre de sinistres**

GLM	CART	Random Forest	GBM	XGBoost
CRM	**CRM**	CRM	CRM	**CRM**
Zone	Zone	Zone	Zone	Zone
Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B
Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule
Genre	Genre	Genre	Genre	Genre
Formule	Formule	Formule	Formule	Formule
Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage
Classe_SRA	Classe_SRA	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA
Novice	Franchise auto	Franchise auto	Franchise auto	Franchise auto

Nous représentons dans la figure 4.22 une comparaison graphique du modèle GLM et *Random Forest* avec les observations réelles du **nombre de sinistres** en fonction de la variable « Classe_SRA ». Ce choix est motivé par le fait qu'elle est la plus importante dans le modèle *Random Forest*, mais ne ressort pas dans le GLM. Le but est de voir s'il est possible de déceler des différences de prédictions entre le GLM et *Random Forest* vu qu'en termes de RMSE, ils sont à peu près équivalents.

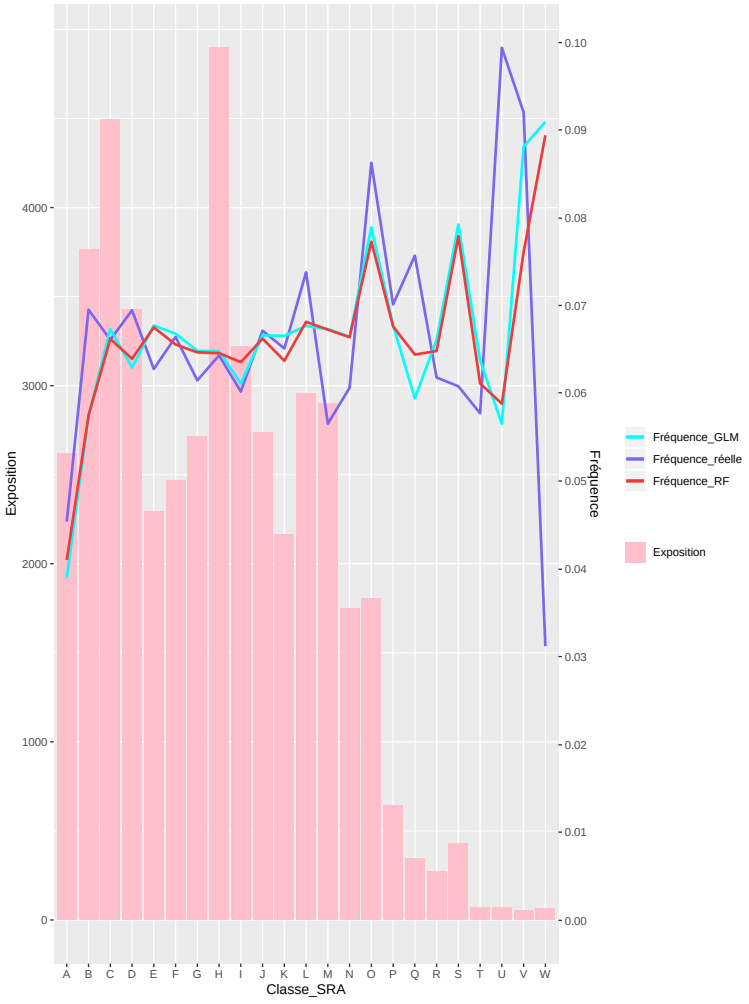


FIGURE 4.22 – Comparaison des observations rééelles et prédites par GLM et *Random Forest* pour la **fréquence des sinistres** en fonction de la Classe_SRA

Globalement, les prédictions du **nombre de sinistres** par GLM et *Random Forest* ne sont pas très éloignées. De la classe A à C, les deux modèles sont aussi pertinents l'un que l'autre. Pour les classes J,K et L, nous remarquons que le GLM (en bleu) prédit un peu plus que la courbe réelle (en violet) contrairement au *Random Forest* (en rouge). Sur les classes Q et R, *Random*

Forest se rapproche plus de la courbe réelle que le GLM.

En conclusion, il est préférable de maintenir le modèle GLM pour la modélisation du **nombre de sinistres**, car comme on le verra plus tard (tableau 4.28), ce dernier est facile à mettre en oeuvre tant en termes de temps d'execution qu'en termes d'interprétabilité.

4.3.2.2 Coût du sinistre

Pour le **coût du sinistre**, nous constatons que le GLM n'est plus en première place. En effet, l'arbre de décision **CART** présente une RMSE de 1447.239, il est suivi du modèle XGBoost avec une RMSE de 1448.542 ensuite du GLM avec une RMSE de 1449.768.

Comme pour le **nombre de sinistres**, nous résumons l'importance des variables pour chaque modèle (tableau 4.27). Les variables importantes qui contribuent à la construction des modèles sont mises en évidence avec la couleur **bleu**. Les variables qui apparaissent en premier dans les modèles sont présentées en **rouge**.

TABLE 4.27 – Tableau récapitulant les variables les plus importantes dans la modélisation du **coût du sinistre**

GLM	CART	Random Forest	GBM	XGBoost
CRM	CRM	CRM	CRM	CRM
Zone	Zone	Zone	Zone	Zone
Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B	Ancienneté du permis B
Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule
Genre	Genre	Genre	Genre	Genre
Formule	**Formule**	Formule	Formule	**Formule**
Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage	Franchise dommage
Classe_SRA	Classe_SRA	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA	Groupe_SRA
Novice	Franchise auto	Franchise auto	Franchise auto	Franchise auto

Les modèles CART, XGBoost ainsi que le GLM sont présentés graphiquement pour la variable « Formule » dans la figure 4.23. Les modèles CART et XGBoost suivent la tendance

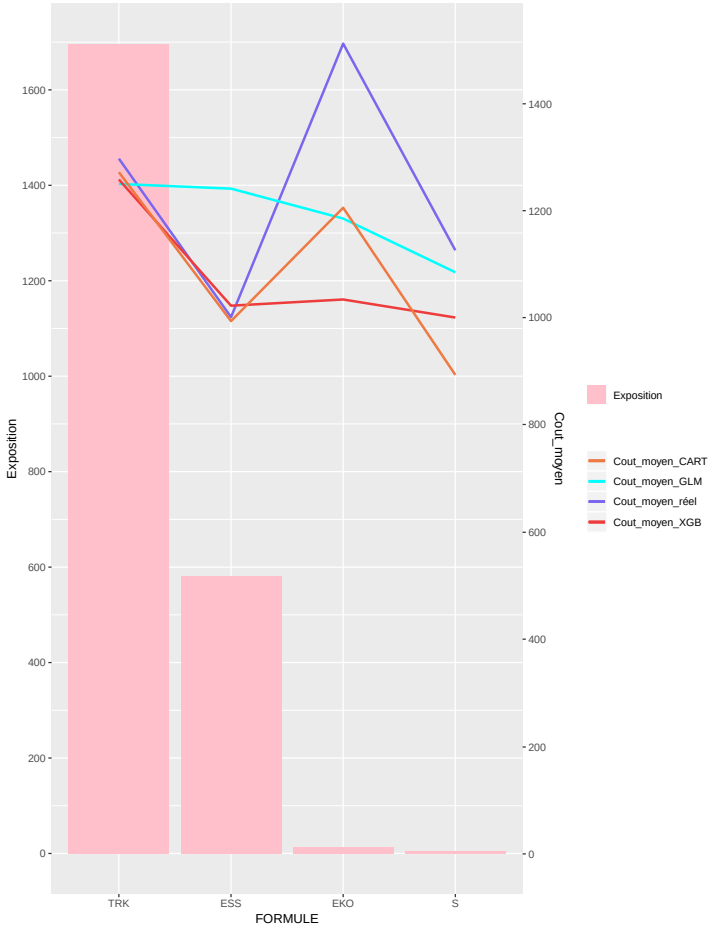


FIGURE 4.23 – Comparaison des observations rééelles et prédites par GLM et *Random Forest* pour le **coût moyen** en fonction de la Formule

du coût moyen réel entre les formules "TRK" et "ESS". En revanche, le GLM prédit à peu près le même coût moyen pour les deux formules alors qu'en évidence, la formule "TRK" a un coût moyen réel plus élevé que "ESS".

Si l'assureur souhaite quand même maintenir le modèle GLM pour prédire le **coût du sinistre**, il serait intéressant d'appliquer un coefficient correcteur à la prime finale afin d'ajuster ce problème.

Entre les formules "ESS" et "EKO", CART est celui qui se rapproche le plus de la courbe des coûts moyens réels.

4.3.3 Comparaison des modèles : Facilité d'explication, facilité de paramétrage, interprétabilité et pouvoir prédictif

Nous avons pu élaborer dans ce chapitre les prédictions du **nombre de sinistres** et du **coût du sinistre** ainsi qu'une comparaison de leurs RMSE sur la base de test. Le modèle optimal n'est pas forcément celui qui s'ajuste le mieux à la base de données, mais celui qui réduit l'erreur de la valeur prédite sur une autre base, ce qui est donc de nature à renforcer la robustesse des résultats prédits. Cependant, d'autres critères entrent en compte lorsqu'il s'agit de choisir le meilleur modèle. En effet, bien que les méthodes de *machine learning* conduisent souvent à des estimations plus précises, leurs résultats restent difficilement interprétables contrairement aux GLM. C'est d'ailleurs pour cette raison qu'on les qualifie de modèle *boîte noire*.

Nous comparons dans le tableau 4.28 les avantages et inconvénients des modèles implémentés.

Lecture du tableau :

Plus un modèle a de signes "+" pour un critère étudié, mieux c'est. Le maximum va jusqu'à 5 signes "+".

Table 4.28 – Comparaison des différents modèles testés

Modèle	Vitesse d'apprentissage	Facilité d'explication de l'algorithme	Facilité de paramétrage	Pouvoir prédictif	Interprétabilité des résultats
GLM	+++++	+++++	+++++	+++	+++++
CART	+++++	+++++	++++	+++	++++
Random Forest	++	+	++	++++	+
GBM	++	+	++	+	+
XGBoost	+	+	+	+++++	+

✗ Vitesse d'apprentissage et facilité de paramétrage

Etant donné que les modèles de *Random Forest*, *Gradient Boosting Machine* et XGBoost sont des modèles agrégés, la vitesse de leur apprentissage est particulièrement ralentie par rapport aux autres modèles (GLM et CART).

Néanmoins, la vitesse d'apprentissage n'est pas le seul critère à prendre en compte lorsqu'il s'agit de comparer le temps d'implémentation de chaque modèle. En effet, un travail de calibrage des hyperparamètres a également été analysé dans les modèles. Voici quelques enseignements tirés à ce sujet au cours de ce mémoire :

— Le modèle GLM est le plus rapide en termes d'exécution. Néanmoins, utiliser ce modèle nécessite une étude statistique approfondie du portefeuille, ce qui peut demander un certain temps ;

- Le modèle CART est moins contraignant, que ce soit pour le calibrage des paramètres ou en termes de temps d'execution. En effet, des premiers résultats peuvent être obtenus en passant une durée assez restreinte sur ces étapes ;
- La calibration des paramètres de la forêt aléatoire est également assez rapide, le fait de, par exemple, fixer un nombre élevé d'arbres dès le départ permet d'améliorer la performance du modèle rapidement, cependant le temps d'exécution de l'algorithme est particulièrement lent ;
- Le *Gradient Boosting Machine* est moins contraignant à calibrer par rapport à *XGBoost*, cependant, il s'agit du seul modèle qui ne s'améliorait pas après avoir été optimisé, il est également sujet au surapprentissage et ce, même après l'étape d'hyperparamétrage ;
- L'*eXtreme Gradient Boosting* est sans doute le modèle qui demande le plus de temps à calibrer du fait de la multitude des paramètres dont il dispose. Il est nécessaire de tester plusieurs combinaisons de paramètres afin de déceler celle qui permet d'obtenir le meilleur modèle. Comme pour le *Random Forest*, le temps d'exécution de XGBoost est assez conséquent du fait qu'il soit issu d'un modèle agrégé.
- Certains algorithmes nécessitent le traitement des données en amont (dummisation des variables qualitatives pour l'*XGBoost* par exemple). Par ailleurs, les systèmes d'informations aujourd'hui sont rigides et il est difficile d'intégrer de nouvelles variables même en *feature engineering*.

✗ **Facilité d'explication de l'algorithme**

Avant de mettre en place un modèle, il est nécessaire pour l'utilisateur de maîtriser l'algorithme derrière le modèle en question. La théorie des modèles exposée au chapitre 2 permet de conclure que les modèles linéaires généralisés (GLM) sont les plus faciles en termes de « facilité d'explication de l'algorithme ». Ils sont suivis de l'algorithme CART. En effet, la lecture d'un arbre est beaucoup plus simple et appréciable, on peut y voit clairement les variables explicatives qui contribuent directement à la constitution de l'arbre. Aussi, la valeur de la variable à expliquer est directement affichée sur l'arbre.

✗ **Pouvoir prédictif**

Dans le cadre de ce mémoire, les différents modèles pour le **nombre de sinistres** et le **coût du sinistre** ont été comparés à l'aide des métriques RMSE, MAE et erreur euclidienne sur la base test. Dans le tableau récapitulatif 4.25, nous nous sommes focalisés sur la comparaison des RMSE. Il en est ressorti que pour la modélisation du nombre de sinistres, le modèle GLM est le plus performant avec une RMSE faible par rapport aux autres modèles, il est suivi de *Random Forest*. Concernant le **coût du sinistre**, c'est le modèle CART qui est performant, suivi du modèle agrégé XGBoost.

✗ **Interprétabilité des résultats**

La lecture des arbres de décisions CART ne nécessite pas de compétences particulières en statistiques, en effet la segmentation effectuée peut être facilement visualisée sur l'arbre. Cette propriété est évidemment perdue par l'agrégation d'arbres.

Conclusion :

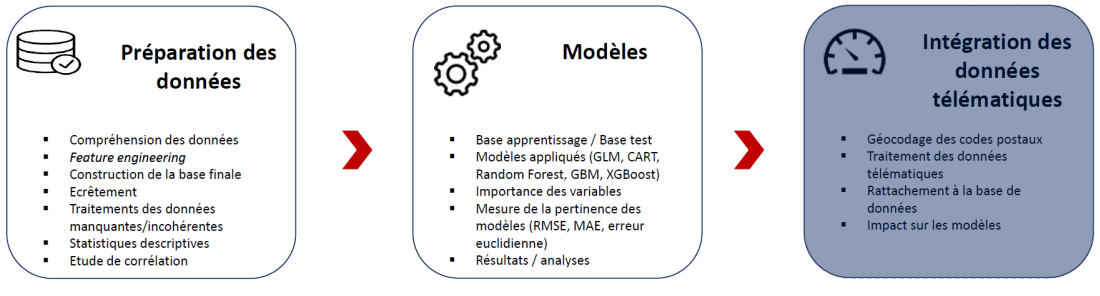
Dans un secteur fortement concurrentiel comme celui de l'assurance automobile, les assureurs se doivent d'améliorer de façon permanente leur segmentation tarifaire dans le but d'adapter au mieux leur prime au risque assuré. A travers les modèles construits, nous pouvons constater que les méthodes d'apprentissage statistique présentées dans ce chapitre offrent des performances comparables à celles du GLM. En effet, que ce soit pour le **nombre de sinistres** ou le **coût des sinistres**, nous avons remarqué que les RMSE sont proches.

Par ailleurs, il existe des différences quant à l'importance des variables. Par exemple, pour la modélisation du **coût du sinistre**, l'arbre CART (ressorti comme étant le meilleur), fait intervenir la variable « Formule » qui n'a pas été employée dans la modélisation par les GLM.

Une amélioration potentielle de ces prédictions serait d'intégrer des données externes à notre base de données. C'est dans ce contexte que s'inscrit le chapitre suivant où nous intégrerons à notre portefeuille des données issues de la télématique dans le but de mesurer leur apport dans les modélisations du **nombre de sinistres** et du **coût des sinistres**.

Chapitre 5

Intégration des données issues de la télématique



L'idée dans cette section est d'intégrer les données issues de la télématique dans notre portefeuille de responsabilité civile en usage professionnel. Le but est de mesurer l'apport de ces données dans les modèles précédemment présentés.

Nous commencerons tout d'abord par évoquer le contexte de l'acquisition des données télématiques issues de l'entreprise **Ellis Car**. Ensuite, en intégrant ces données à notre portefeuille, nous refaisons les mêmes modèles que dans le chapitre 4 et nous calculerons les métriques de mesure d'écarts entre les observations et les prédictions à la fois pour le **nombre de sinistres** et pour le **coût du sinistre**.

5.1 Contexte de l'étude et acquisition des données télématiques

Avec l'avènement du *Big Data* et l'utilisation de nouvelles sources de données telles que les capteurs ou la télématique, les sources de données externes, le digital ou encore les réseaux sociaux, les organisations en profitent pour appliquer les techniques de machine learning à de nouveaux aspects des opérations d'assurance.

Ces données donnent la possibilité d'acquérir des variables complémentaires facilitant ainsi une qualification plus fine du risque comparée à celle obtenue à partir de données transmises par le réseau et/ou l'assuré. D'autant plus que tout apport d'information pouvant renforcer la caractérisation du profil de l'assuré est susceptible d'éviter l'antisélection.

Il existe plusieurs manières permettant d'acquérir les données selon leur accessibilité. Nous pouvons par exemple citer :

- Les données publiques dites *Open Data* qui sont mises à disposition par le gouvernement [1]. Cependant, même ces données publiques peuvent parfois être payantes ;
- L'INSEE propose également un grand nombre d'informations, agrégées par code INSEE (grossièrement équivalent au code postal) ;
- *OpenStreetMap* [2] constitue la plus grande source d'informations géographiques disponible gratuitement. Il s'agit d'un projet communautaire fondé en 2004 à UCL qui vise à créer une base de données ouverte de l'ensemble de la planète, en se basant sur les contributions de volontaires bénévoles.

Pour ce mémoire, et dans le cadre d'un partenariat avec l'entreprise **Ellis Car**, le cabinet de conseil Galea & Associes a réussi à récupérer des données liées à la télématique afin de nous en servir pour cette partie de l'étude.

5.1.1 Présentation brève de l'entreprise Ellis Car

Ellis Car est une entreprise qui utilise le *Big Data* pour analyser les comportements des conducteurs et réduire le risque d'accidents. Ainsi, à l'aide de l'ensemble des données collectées (flux RSS des sociétés d'autoroute et capteurs sur le véhicule), l'application **Ellis Car** est capable de détecter les comportements qui sont mathématiquement déviants et les zones à risques. Les données collectées intéressent non seulement les automobilistes qui peuvent revoir leurs comportements et ajuster leur conduite afin de réduire leurs coûts, mais aussi les constructeurs et les assureurs. En effet, en connaissant ces informations, les assureurs seront en mesure de connaître l'assuré et de réduire le risque du portefeuille.

1. Plusieurs données sont regroupées selon des sujets variés sur le site : www.data.gouv.fr
2. http ://openstreetmap.fr/

5.1.2 Acquisition des données et description des nouvelles variables

Dans le but de vérifier l'impact de l'ajout de données télématiques à notre portefeuille de responsabilité civile en usage professionnel, et étant donné que les données collectées par l'application **Ellis Car** sont géographiques, il a fallu passer par une étape de géocodage des codes postaux disponibles dans notre base de données. Ces codes postaux, rappelons-le, représentent le domicile de l'assuré.

Le géocodage est un procédé qui est connu des sciences cartographiques et qui est largement documenté, et de nombreux services externes le rendent accessible [3]. Le géocodage transforme une adresse suivant un format pré-défini (dans notre cas, un code postal) en des coordonnées géographiques (x, y) qui représentent la latitude et la longitude. Ce retraitement permet alors de rattacher chaque sinistre à un point précis de la France.

La figure 5.1 résume les étapes d'acquisition de la base enrichie des données télématiques.

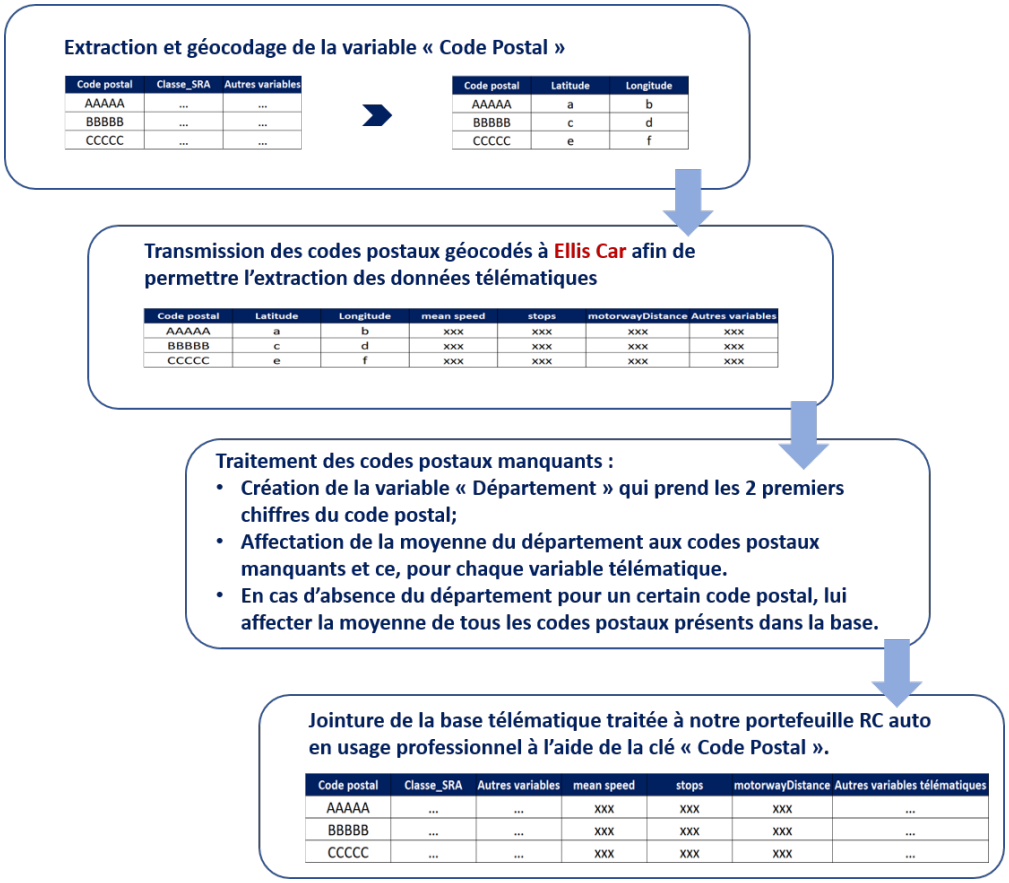


FIGURE 5.1 – Acquisition des données télématiques

3. Dans ce mémoire, nous avons principalement utilisé les sites suivants : https ://www.laposte.fr/particulier/outils/trouver-un-code-postal **et** https ://www.coordonnees-gps.fr

Les informations concernant chaque variable sont extraites par coordonnées GPS par **Ellis Car sur un rayon de 2 kilomètres pendant les 3 dernières années**. Il s'agit de :

- totalDistance : Distance totale de routes en mètres ;
- motorwayDistance : Distance d'autoroutes en mètres ;
- trunkDistance : Distance de routes nationales en mètres ;
- pavedDistance : Distance de routes recouvertes de pavés, de béton ou de bitume en mètres ;
- cobblestoneDistance : Distances recouvertes de pavés ronds à base de pierre de la taille d'un galet, exprimée en mètres ;
- traffic_signals : Nombre de feus rouges ;
- stop : Nombre de panneaux "stop" ;
- crossing : Nombre de passages piétons ;
- traffic_calming : Nombre de ralentisseurs, ou "dos-d'âne" ;
- roundabout : Nombre de ronds-points ;
- roadDensity : Densité des routes ;
- populationDensity : Densité de la population vivant dans le rayon considéré ;
- meanSpeed : Vitesse moyenne sur les 3 dernières années ;
- countOfSpeedExcess : Nombre d'excès de vitesse sur les 3 dernières années ;
- countOfAccelerationExcess : Nombre d'accélérations effectuées par le conducteur ;
- countOfbrakingExcess : Nombre de feinages brusques du conducteur ;
- accident : Indicateur de dangerosité d'accidents, cette variable est sans unités. Elle permet de cartographier les accidents. Par exemple, si dans un certain endroit A l'indicateur vaut 100, et dans un autre endroit B, l'indicateur vaut 110, on dira que l'endroit B est une zone plus dangereuse que A.

5.2 Modélisation du nombre de sinistres et du coût des sinistres en intégrant les données télématiques

Objectif :

Comme nous l'avons évoqué auparavant, tout apport d'information permettant de renforcer le profil d'un assuré serait utile pour éviter le risque d'antisélection. Alors que les variables explicatives de la première partie (chapitre 4) étaient limitées aux caractéristiques de l'assuré ou de son véhicule, nous avons à présent rajouté des variables comportementales issues de la télématique afin de mesurer l'impact de celles-ci sur nos prédictions.

Pour cela, nous présenterons pour le **nombre de sinistres** et pour le **coût du sinistre**, les modèles : GLM, *CART*, *Random Forest*, *GBM* et *XGBoost* et les erreurs associées (RMSE, MAE et erreur euclidienne) pour chaque modèle. Nous tâcherons de nous focaliser sur les résultats renvoyés par les modèles étant donné que la démarche reste la même que celle suivie pour le chapitre 4.

5.2.1 Les modèles linéaires généralisés (GLM)

5.2.1.1 Modélisation du nombre de sinistres après intégration des données télématiques

- **Modèle GLM avec les variables retenues**

  Tout d'abord, nous avons remarqué que les données issues de la télématique sont fortement corrélées. Ceci s'explique par le fait qu'elles sont rattachées à l'aide de la même variable : le code postal. Néanmoins, il a été choisi de conserver l'ensemble de ces variables externes à ce stade ; l'étape de sélection des variables permettra d'en éliminer quelques unes.

  Ainsi, nous rajoutons ces variables à celles gardées auparavant dans la modélisation du **nombre de sinistres** (partie 4.1.2).

  Parmi les variables télématiques considérées, la procédure de sélection de variables n'en choisit que deux significatives ($p - value < 0.05$), il s'agit de : la distance totale des routes en mètres (totalDistance) et des distances recouvertes de pavés ronds (cobblestoneDistance). (Voir tableau 5.1). Par ailleurs, la loi Poisson est maintenue avec le lien «log».

TABLE 5.1 – Résultat de la procédure *GENMOD* pour la modélisation du **nombre de sinistres** avec données télématiques

	Statistique LR pour Analyse de Type 1			
Source	**Ecart**	**DDL**	**Khi-2**	**Pr > Khi-2**
Intercept	73 846.3634			
PERM_CRM	72 918.7782	55	55	<.0001
VHL_FORM	72 500.6511	81	81	<.0001
GENRE	72 433.0350	2	2	<.0001
Zone	72 089.2648	15	15	<.0001
Classe_SRA	72 018.2248	22	22	<.0001
totalDistance	72 014.1692	1	1	0.0440
cobblestoneDistance	72 007.2048	1	1	0.0083

Importance des variables

Le gain en AIC des variables ayant contribué au modèle est représenté dans la figure 5.2.

Nous pouvons remarquer que les deux variables télématiques rajoutées n'apportent que très peu au modèle.

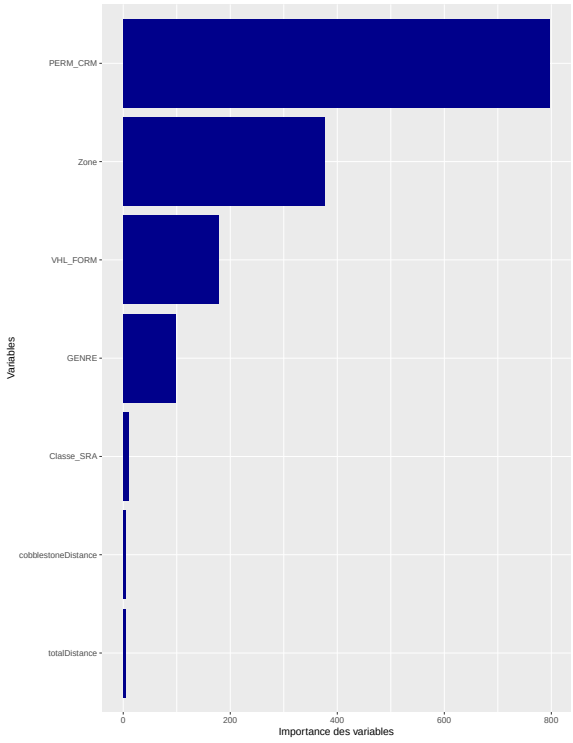


FIGURE 5.2 – Importance des variables dans la modélisation du **nombre de sinistres** par GLM après intégration des données télématiques

- **Pertinence du modèle**

  Les erreurs du modèle après intégration des variables télématiques deviennent alors :

TABLE 5.2 – Erreurs obtenues pour la prédiction du **nombre de sinistres** avec GLM après intégration des données télématiques

	RMSE	MAE	Erreur euclidienne
Test	0.2195956	0.08825353	55.31326

La RMSE du modèle sur la base test s'améliore en passant de 0.2210493 à 0.2195956. La MAE et l'erreur euclidienne baissent également.

5.2.1.2 Modélisation du coût du sinistre après intégration des données télématiques

- **Modèle GLM avec les variables retenues**

  Les variables télématiques sélectionnées pour la modélisation du **coût du sinistre** en plus de celles retenues dans la partie 4.1.3 sont : La distance de routes recouvertes de pavés, béton ou bitume en mètres (pavedDistane) et les distances recouvertes de pavés ronds à base de pierre de la taille d'un galet, en mètres (cobblestoneDistance). Les résultats du modèle sont présentés dans le tableau 5.3.

TABLE 5.3 – Résultat de la procédure *GENMOD* pour la modélisation du **coût du sinistre** après intégration des données télématiques

	Statistique LR pour Analyse de Type 1			
Source	**Ecart**	**DDL**	**Khi-2**	**Pr > Khi-2**
Intercept	-179355.72			
zone2	-179346.42	2	9.30	<.0001
CRM2	-179310.39	3	36.03	<.0001
pavedDistance	-179301.97	1	8.41	0.0037
cobblestoneDistance	-179285.75	1	16.23	<.0001

Importance des variables

Le gain en AIC des variables contribuant au modèle est représenté dans la figure 5.3.

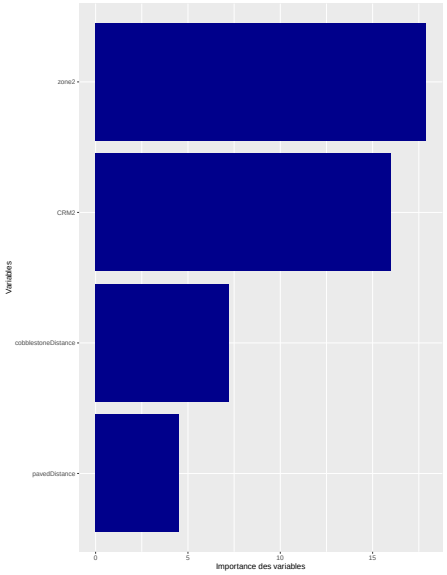


FIGURE 5.3 – Importance des variables dans la modélisation du **coût du sinistre** par GLM après intégration des données télématiques

Pertinence du modèle

La pertinence du modèle est mesurée par la RMSE, la MAE et l'erreur euclidienne. (Tableau 5.4).

TABLE 5.4 – Erreurs obtenues pour la prédiction du **coût du sinistre** avec GLM après intégration des données télématiques

	RMSE	MAE	Erreur euclidienne
Test	1449,171	968.3022	76161.09

La RMSE sur la base test s'améliore très légèrement en passant d'une valeur de 1449.768 à 1449.171. Par ailleurs, la MAE passe de 969.6993 à 968.3022. L'erreur euclidienne quant à elle présente une légère hausse et passe de 76054.11 à 76161.09.

Objectif :

Dans la suite et afin de mesurer l'impact des données télématiques sur les modèles d'apprentissage statistique, l'ensemble des variables télématiques qui nous ont été transmises ont été intégrées aux modèles avec les données liées à l'assuré ainsi que les caractéristiques de son véhicule.

5.2.2 Modélisation avec CART

L'approche suivie pour la modélisation du **nombre de sinistres** et du **coût du sinistre** est la même que celle présentée dans la partie 4.2.1.

5.2.2.1 Modélisation du nombre de sinistres

Nous commençons d'abord par l'élagage de l'arbre maximal.

Selon la règle de Breiman, la valeur du critère de complexité est : **cp=0.000323438**. L'arbre construit est représenté dans la figure 5.4.

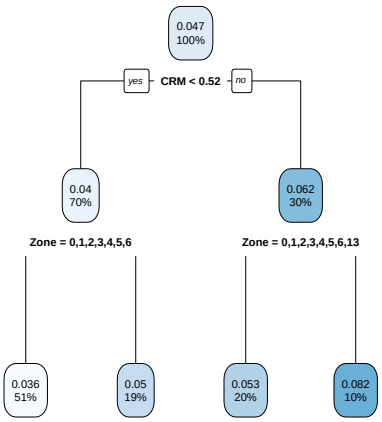


FIGURE 5.4 – Arbre retenu pour la modélisation du **nombre de sinistres** après intégration des données télématiques

Importance des variables

Les variables qui contribuent à la construction de l'arbre sont représentées dans la figure 5.5.

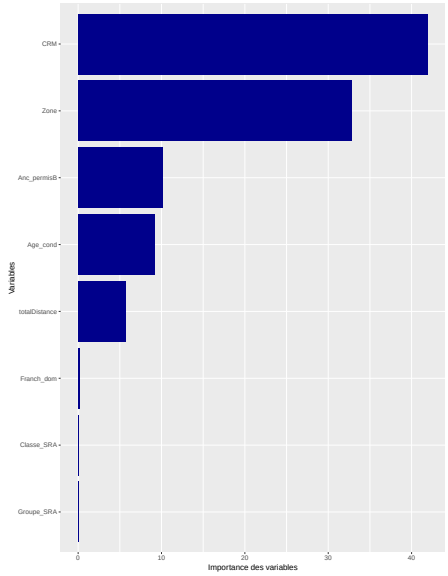


FIGURE 5.5 – Importance des variables dans la modélisation du **nombre de sinistres** par CART après intégration des données externes

Les variables CRM et Zone tout comme dans la partie 4.2.1.2 restent prédominantes dans la construction de l'arbre pour le modèle du **nombre de sinistres**. La variable totalDistance apparait en $5^{ième}$ position. Elle aurait pu être visible également sur l'arbre de la figure 5.4 en choisissant un autre paramètre de complexité **cp** qui permettrait d'avoir plus de feuilles (autrement dit, plus de valeurs prédites). Nous avons cependant décidé de garder le **cp** qui respecte la règle de Breiman décrite par l'équation 4.1 de la partie 4.2.1.2.

Pertinence du modèle

La RMSE, la MAE ainsi que l'erreur euclidienne sont présentées dans le tableau 5.5.

TABLE 5.5 – Erreurs obtenues pour la prédiction du **nombre de sinistres** avec CART après intégration des données télématiques

	RMSE	MAE	Erreur euclidienne
Test	0.2202025	0.08906019	55.46612

Nous pouvons remarquer que la RMSE passe à une valeur légèrement plus faible (de 0.2217749 à 0.2202025). Il en est de même pour la MAE et l'erreur euclidienne.

5.2.3 Modélisation avec Random Forest

La méthodologie suivie pour la modélisation du **nombre de sinistres** ainsi que celle du **coût du sinistre** est en tout point similaire à celle établie dans la partie 4.2.3.1. Nous commencerons tout d'abord par présenter un modèle arbitraire que nous optimiserons par la suite. Ensuite, l'importance de chaque prédicteur impliqué dans la modélisation sera présentée.

5.2.3.1 Modélisation du nombre de sinistres

Modèle sans hyperparamétrage

Rappelons que notre modèle sans hyperpamètrage, tout comme dans la première partie, est caractérisé par :

- **mtries**= $p/3$, avec p le nombre de variables explicatives. Notons qu'en rajoutant les variables d'**Ellis Car**, le nombre de variables explicatives s'élève à 29 ;
- **nfolds**=5 ;
- **ntrees**= 100 arbres ;
- **max_depth**=30.

Les erreurs obtenues pour ce modèle sont représentées dans le tableau 5.6.

TABLE 5.6 – Résultats du modèle arbitraire *Random Forest* pour le **nombre de sinistres** avec les variables télématiques

	RMSE	MAE	Erreur euclidienne
Test	0.2343106	0.09169112	59.01976

Hyperparamétrage

A présent, nous procédons à l'hyperparamétrage des paramètres :

- **mtries** : Comme le nombre de variables a augmenté, nous élargissons notre plage de valeurs en faisant varier ce paramètre entre 5 et 20 avec un pas de 2. Les valeurs testées sont $\{5, 7, 9, 11, 13, 15, 17, 19\}$.
- **ntrees** : Nous faisons varier ce paramètre, comme dans le chapitre 4 entre 10 et 400 avec un pas de 10 ;
- **max_depth** : La profondeur de l'arbre est fixée à 10.

Sur 84 modèles testés, les paramètres optimaux renvoyés par le **grid search** sont :

- **mtries**= 5 ;
- **ntrees**= 340.

Nous pouvons remarquer sur le graphique 5.6, que l'erreur quadratique moyenne (RMSE) se stabilise à partir de la valeur ntrees=300, nous décidons de retenir la valeur renvoyée par le **grid search** qui est 340, car en comparant avec une forêt à 300 arbres, la RMSE augmente.

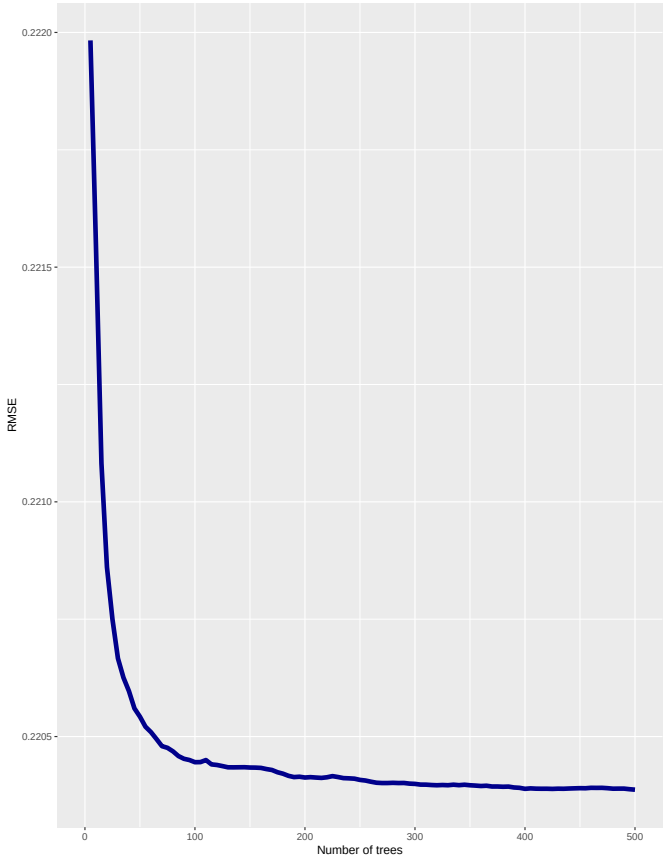


FIGURE 5.6 – Evolution de l'erreur en fonction du nombre d'arbres pour la modélisation du **nombre de sinistres** en rajoutant les données télématiques

Importance des variables

Les variables les plus importantes sont précisées dans la figure 5.7.

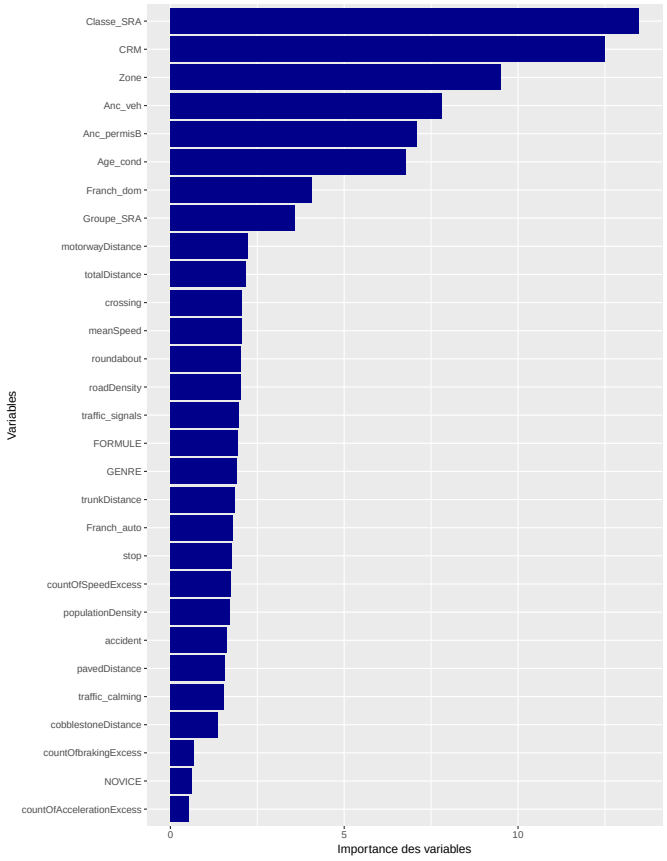


FIGURE 5.7 – Importance des variables dans la modélisation du **nombre de sinistres** par *Random Forest* avec les variables télématiques

Force est de constater que les variables du portefeuille ressorties auparavant dans la partie 4.2.3.2 sont toujours aussi importantes. Il s'agit de la classe SRA, le CRM, la zone, l'ancienneté du véhicule, l'ancienneté du permis B, l'âge du conducteur, la franchise dommage et le groupe SRA.

Nous pouvons aussi remarquer que certaines variables issues de la télématique sont importantes à peu près au même degré, comme la distance d'autoroutes (motorwayDistance), la distance totale de routes (totalDistance), le nombre de passages piétons (crossing), la vitesse moyenne (meanSpeed) ou encore le nombre de ronds-points (roundabout), mais elles n'interviennent qu'à partir de la $9^{ième}$ position.

Mesure de la pertinence du modèle

Les erreurs calculées pour ce modèle sont représentées dans le tableau 5.7.

TABLE 5.7 – Erreurs obtenues pour la modélisation du **nombre de sinistres** avec *Random Forest* avec les variables télématiques

	RMSE	MAE	Erreur euclidienne
Test	0.2199053	0.08890695	55.39126

En comparant la RMSE sur la base test du modèle *Random Forest* pour le **nombre de sinistres** obtenue dans la partie 4.2.3.2, nous constatons que celle-ci s'améliore de près de 1%. En effet, la RMSE sans intégration des variables externes vaut 0.2216543, en ajoutant les nouvelles informations sur le comportement, elle baisse pour atteindre une valeur de 0.2199053, soit un gain relatif de 1%.

5.2.3.2 Modélisation du coût du sinistre

Modèle sans hyperparamétrage

Les erreurs obtenues pour le modèle sans optimisation des paramètres sont données dans le tableau 5.8.

TABLE 5.8 – Résultats du modèle arbitraire *Random Forest* pour le **coût du sinistre** avec les variables télématiques

	RMSE	MAE	Erreur euclidienne
Test	1495.325	998.3633	78444.02

Hyperparamétrage

Sur 293 modèles, les paramètres optimisés retenus pour la modélisation du **coût du sinistre** sont :

- **mtries**= 5 ;
- **ntrees**= 290.

Par ailleurs, le graphique 5.8 représentant l'évolution de la RMSE en fonction du nombre d'arbres justifie le choix de ntrees=290.

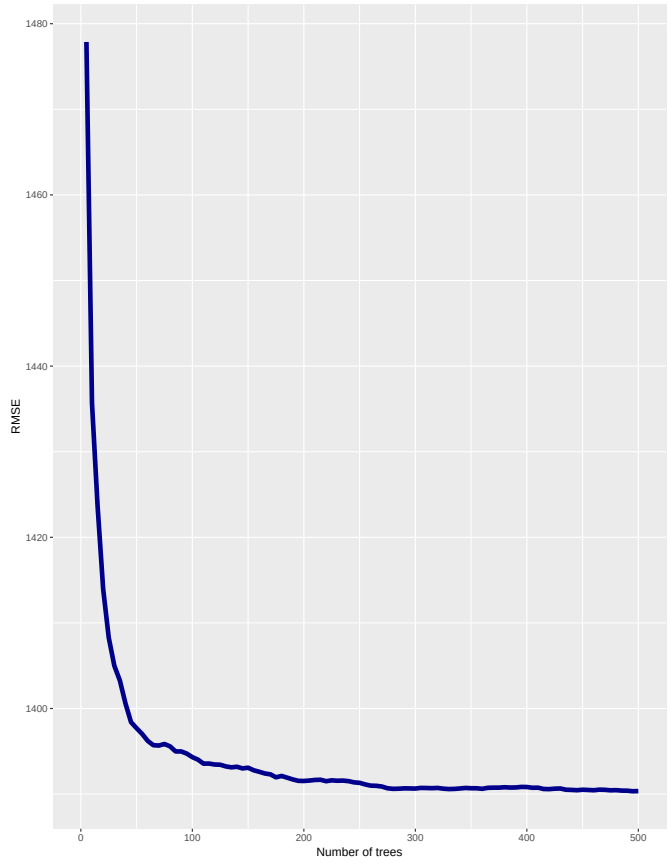


FIGURE 5.8 – Evolution de l'erreur en fonction du nombre d'arbres pour la modélisation du **coût du sinistre** en rajoutant les données télématiques

Importance des variables

Comme pour le **nombre de sinistres**, nous présentons dans la figure 5.9 la contribution des variables dans la construction du modèle pour le **coût du sinistre**.

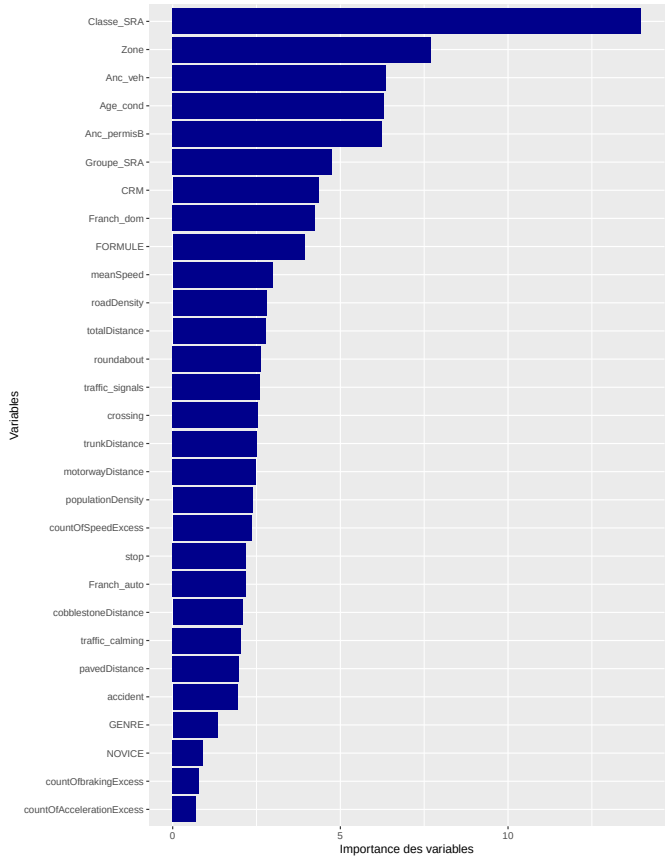


FIGURE 5.9 – Importance des variables dans la modélisation du **coût du sinistre** par *Random Forest* avec les variables télématiques

De par les variables liées à l'assuré et à son véhicule (classe SRA, CRM, zone, ancienneté du véhicule...), nous pouvons remarquer que certaines variables télématiques contribuent à la modélisation du **coût du sinistre** comme la vitesse moyenne, la densité des routes, le nombre de ronds-points ou le nombre de feux rouges à une intensité à peu près égale.

Mesure de la pertinence du modèle

La performance du modèle est obtenue en calculant les métriques RMSE, MAE et l'erreur euclidienne sur la base test (tableau 5.9). En comparant avec les erreurs renvoyées lors de la modélisation sans les données télématiques, nous constatons qu'il y a une légère amélioration. En effet, on passe d'une RMSE sur la base test de 1452.612 à 1450.462. La MAE et l'erreur euclidienne baissent également.

TABLE 5.9 – Erreurs obtenues pour la modélisation du **coût du sinistre** avec *Random Forest* avec les variables télématiques

	RMSE	MAE	Erreur euclidienne
Test	1450.462	954.0693	76137.2

5.2.4 Modélisation avec *Gradient Boosting Machine* (GBM)

Les mêmes étapes que dans 4.2.4 ont été suivies pour la modélisation du **nombre de sinistres** et du **coût du sinistre**.

5.2.4.1 Modélisation du nombre de sinistres

- **Modèle sans hyperparamétrage**

  En maintenant les mêmes paramètres que dans la partie 4.2.4.2, les erreurs calculées pour le modèle sans hyperparamétrage sont représentées dans le tableau 5.10.

TABLE 5.10 – Résultats du modèle arbitraire *Gradient Boosting Machine* pour le **nombre de sinistres** après intégration des données télématiques

	RMSE	MAE	Erreur euclidienne
Test	0.2221867	0.08863876	55.96591

- **Hyperparamétrage**

  Les paramètres optimaux renvoyés par le **grid search** sont :

  — **ntrees**= 250 ;
  — **learn_rate**=0.04 ;
  — **max_depth**= fixé à 10.

  Par ailleurs, en traçant l'évolution de la RMSE en fonction de **ntrees** (figure 5.10), nous pouvons voir que l'erreur se stabilise un peu avant 100 arbres. Nous avons décidé de retenir un modèle avec **ntrees = 100 arbres** car la RMSE est la même entre un modèle à 100 arbres et un autre à 250.

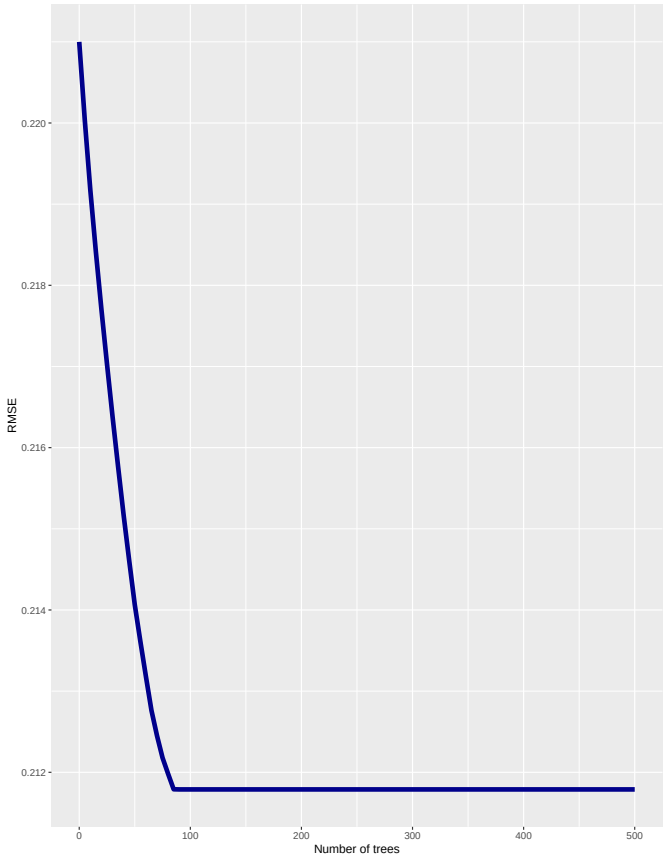


FIGURE 5.10 – Evolution de l'erreur en fonction du nombre d'arbres pour la modélisation du **nombre de sinistres** après intégration des données télématiques

Importance des variables

La figure 5.11 présente l'importance des variables pour ce modèle.

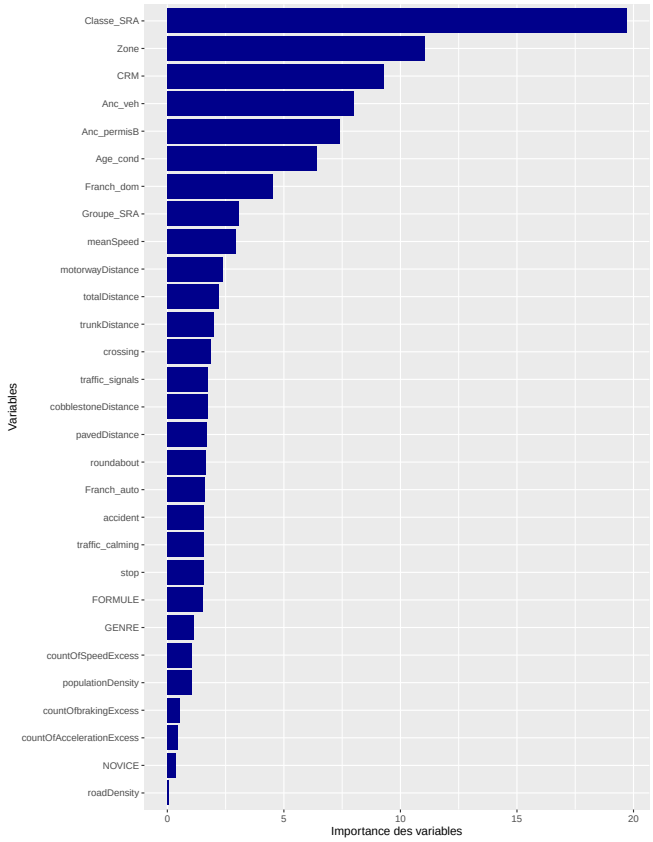


FIGURE 5.11 – Importance des variables dans la modélisation du **nombre de sinistres** par *Gradient Boosting Machine* après intégration des données télématiques

Les 8 premières variables importantes sont similaires à celles apparues dans la modélisation du **nombre de sinistres** sans données télématiques. Une variable externe apparaît à la $9^{ième}$ position, il s'agit de la vitesse moyenne (meanSpeed), elle est suivie de la distance d'autoroutes en mètres (motorwayDistance).

Mesure de la pertinence du modèle

La pertinence du modèle est représentée dans le tableau 5.11.

TABLE 5.11 – Erreurs obtenues pour la modélisation du **nombre de sinistres** avec *Gradient Boosting Machine* après intégration des données télématiques

	RMSE	MAE	Erreur euclidienne
Test	0.2204248	0.08864572	55.52212

Bien que la RMSE, la MAE et l'erreur euclidienne baissent en intégrant les données

externes, le modèle GBM reste le moins pertinent en comparaison avec GLM, CART ou *Random Forest.*

5.2.4.2 Modélisation du coût du sinistre

Modèle sans hyperparamétrage

Avec les paramètres par défaut, nous obtenons les résultats du modèle sans hyperparamétrage dans le tableau 5.12 :

TABLE 5.12 – Résultats du modèle arbitraire *Gradient Boosting Machine* pour le **coût du sinistre** après intégration des données télématiques

	RMSE	MAE	Erreur euclidienne
Train	1084.567	710.1878	113833.1
Test	1465.271	962.3759	76867.38

Hyperparamétrage

Sur les 233 modèles construits, les pramètres optimaux sont :

- **ntrees**= 500 ;
- **learn_rate**= 0.05 ;
- **max_depth**= fixé à 10.

Importance des variables

L'importance des variables pour la modélisation du **coût du sinistre** après intégration des données télématiques est représentée dans la figure 5.12.

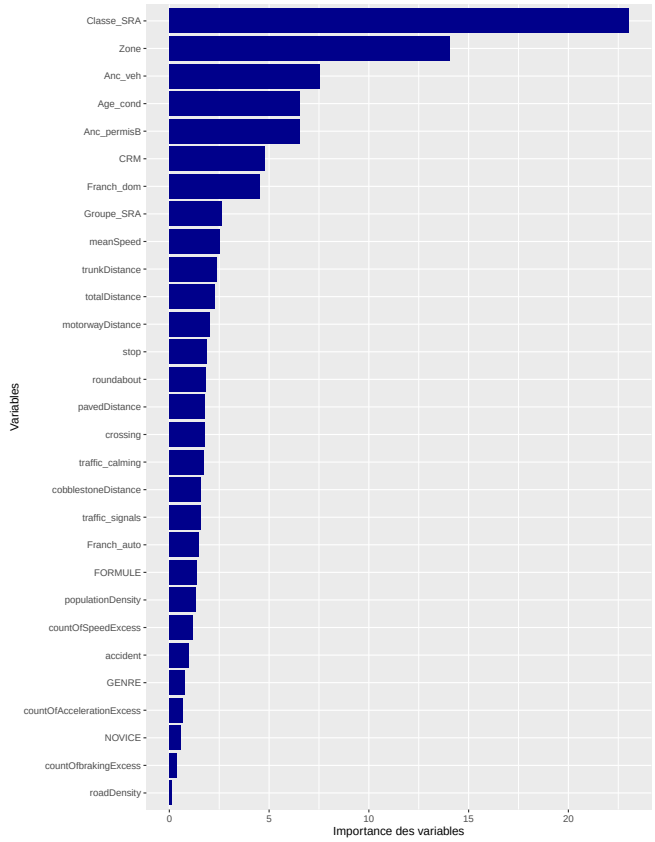


FIGURE 5.12 – Importance des variables dans la modélisation du **coût du sinistre** par *Gradient Boosting Machine* après intégration des données externes

Là encore, nous pouvons remarquer l'apparition de la vitesse moyenne (meanSpeed) à la $9^{ième}$ position, suivie cette fois-ci de la distance de routes nationales (trunkDistance).

Mesure de la pertinence du modèle

Sur le tableau 5.13, nous constatons la même conclusion que dans la partie 4.2.4.3. Le modèle est sujet au surapprentissage et ne se généralise pas à la base test. En effet, la RMSE après hyperparamétrage passe de 574,5311 à 1539,457. Il s'agit du modèle le moins adapté pour la modélisation du **coût du sinistre**.

TABLE 5.13 – Erreurs obtenues pour la modélisation du **coût du sinistre** avec *Gradient Boosting Machine* après intégration des données télématiques

	RMSE	MAE	Erreur euclidienne
Train	574.5311	354.9087	60301.14
Test	1539.457	1024.769	80759.18

5.2.5 Modélisation avec XGBoost

La méthodologie suivie pour la modélisation du **nombre de sinistres** ainsi que celle du **coût du sinistre** est en tout point similaire à celle établie dans la partie 4.2.5.1.

5.2.5.1 Modélisation du nombre de sinistres

Modèle sans hyperparamétrage

Nous commencerons tout d'abord par un modèle par défaut (c.f le point 1 de la méthodologie décrite dans 4.2.5.1).

Les résultats sont donnés dans le tableau 5.14.

TABLE 5.14 – Résultats du modèle arbitraire *XGBoost* pour le **nombre de sinistres** avec les variables télématiques

	RMSE	MAE	Erreur euclidienne
Test	0.2201646	0.09077469	55.45656

Hyperparamétrage

Nous procédons à l'optimisation de notre modèle, pour cela nous faisons varier les paramètres suivants :

- **eta** : 0.1,0.02,0.03,0.04,0.05,0.06,0.07,0.001,0.0001 ;
- **max_depth** : 2, 4, 6, 8, 10 ;
- **gamma** : 0, 1 ;
- **colsample_bytree** : 0.8,
- **min_child_weight** : 0,
- **subsample** : 0.75.

Par ailleurs, une validation croisée est réalisée sur la base d'un découpage en 5 partitions, le graphique (5.13) suivant présente les RMSE suivant les différentes valeurs des paramètres testées.

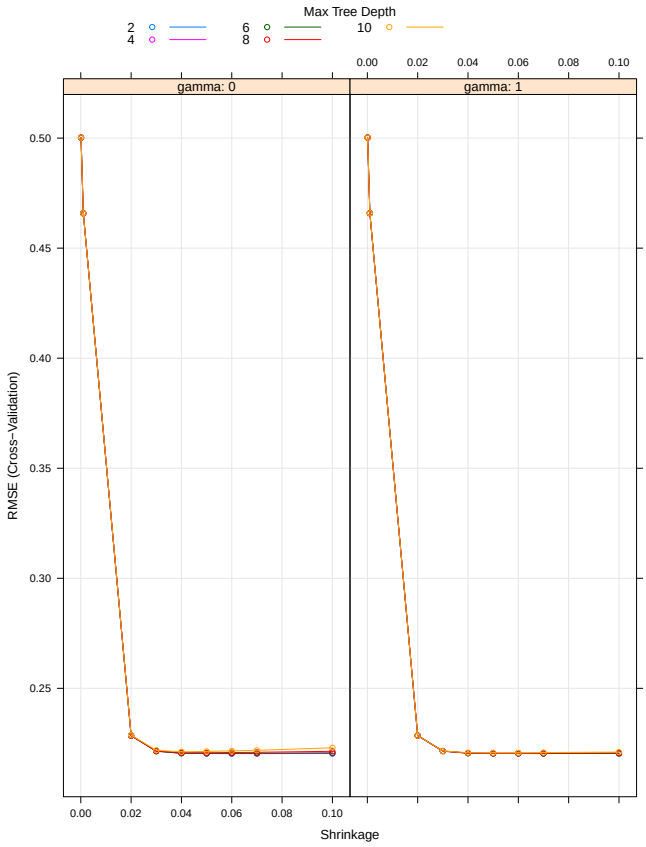


FIGURE 5.13 – RMSE en fonction des différents paramètres XGBoost pour le **nombre de sinistres**

Ainsi, les paramètres qui minimisent la RMSE sont :

- **eta** : 0.06 ;
- **max_depth** : 2 ;
- **gamma** : 0 ;
- **colsample_bytree** : 0.8 ;
- **min_child_weight** : 0 ;
- **subsample** : 0.75.

La figure 5.14 montre que la stabilisation de la RMSE est observée à partir de 60 d'arbres. A l'aide d'une validation croisée, nous obtenons la valeur du paramètre **nrounds** qui vaut 100 itérations.

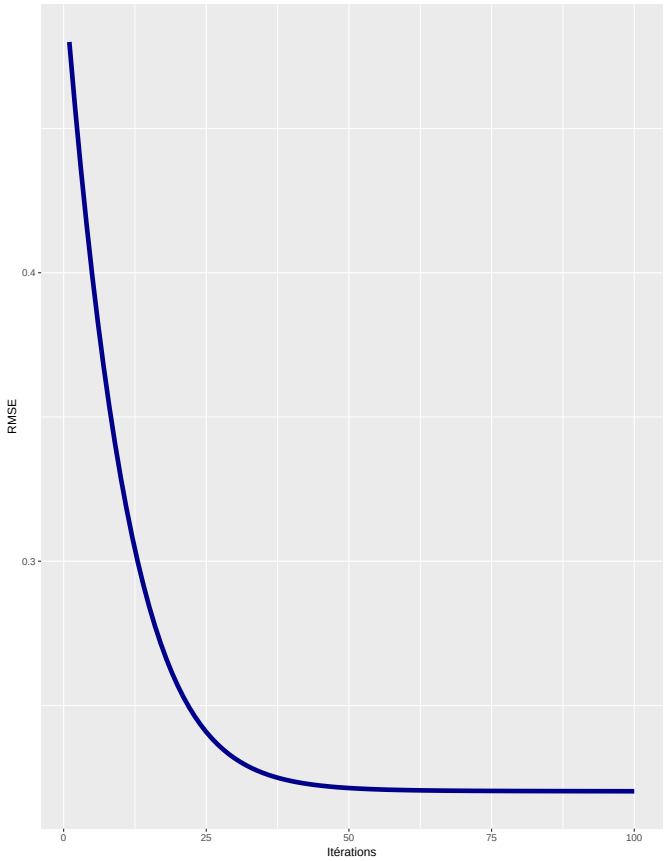


FIGURE 5.14 – Evolution de la RMSE sur la base d'apprentissage avec données télématiques en fonction du nombre d'itérations pour la modélisation du **nombre de sinistres**

Importance des variables

La figure 5.15 permet de résumer l'importance des variables pour ce modèle avec une précision sur la modalité prise par chaque variable :

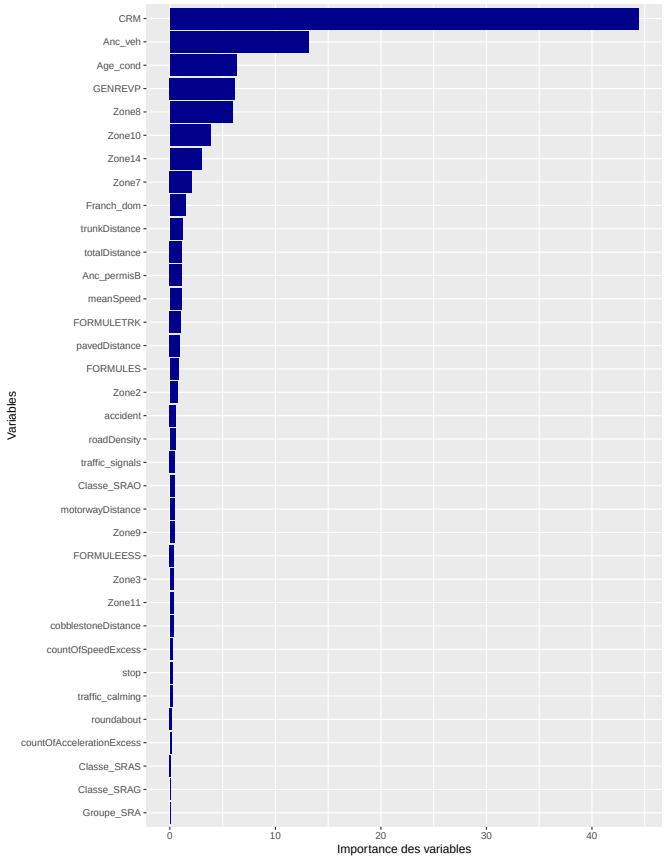


FIGURE 5.15 – Importance des variables dans la modélisation du **nombre de sinistres** par XGBoost en intégrant les variables télématiques

Ici encore, nous pouvons constater que la majorité des variables contribuant à la modélisation du **nombre de sinistres** après intégration des données externes sont les variables liées à l'assuré et à son véhicule (CRM, ancienneté du véhicule, âge du conducteur, genre du véhicule ou encore la zone). La distance de routes nationales (trunkDistance) apparait à la $10^{ième}$ position mais avec une faible intensité.

Mesure de la pertinence du modèle

Avec les paramètres retenus, nous calculons comme pour les modèles précédents, la RMSE, la MAE ainsi que l'erreur euclidienne. Elles sont présentées dans le tableau 5.15.

TABLE 5.15 – Erreurs obtenues pour la modélisation du **nombre de sinistres** avec *XGBoost* avec les variables télématiques

	RMSE	MAE	Erreur euclidienne
Test	0.2199385	0.08975844	55.39961

En comparant les erreurs obtenues à celles du modèle sans données externes, les erreurs baissent avec le modèle comportant les données télématiques. En particulier, nous pouvons constater que la RMSE test passe de 0.2216571 à 0.2199385, soit un gain relatif de 1%.

5.2.5.2 Modélisation du coût du sinistre

Modèle sans hyperparamétrage

Toujours en s'appuyant sur la même méthodologie que pour la modélisation du nombre de sinistres, les résultats du modèle non optimisé sont présentés dans le tableau 5.16 :

TABLE 5.16 – Résultats du modèle arbitraire *Random Forest* pour le **nombre de sinistres** avec les variables télématiques

	RMSE	MAE	Erreur euclidienne
Test	1460.842	941.1952	76635.08

Hyperparamétrage

L'optimisation des paramètres se fait de la même manière que pour la modélisation du **nombre de sinistres** avec données externes. (cf. 5.2.5.1)

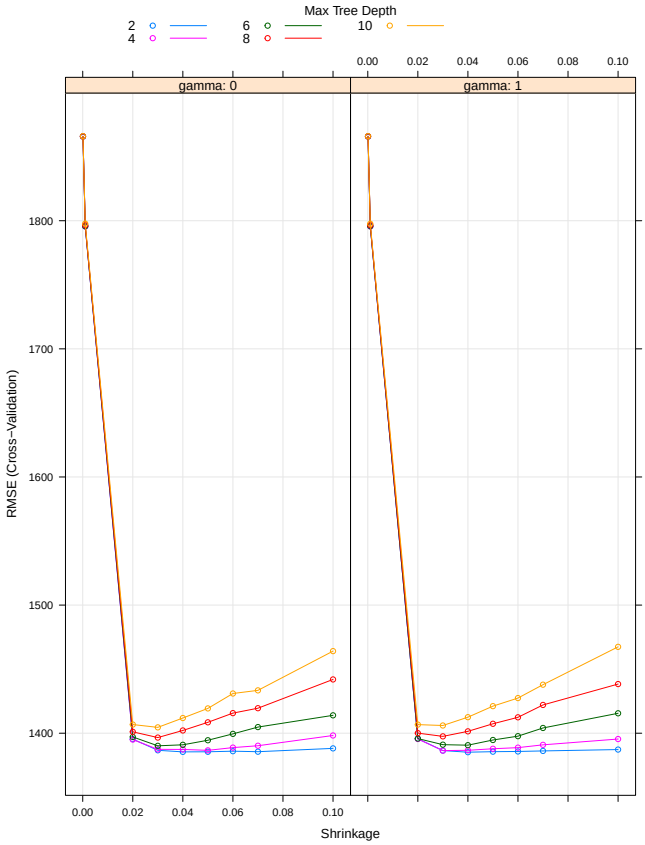


FIGURE 5.16 – RMSE en fonction des différents paramètres XGBoost pour le **coût du sinistre** avec les variables télématiques

Les paramètres retenus sont :

- **eta** : 0.04 ;
- **max_depth** : 2 ;
- **gamma** : 1 ;
- **colsample_bytree** : 0.8,
- **min_child_weight** : 0,
- **subsample** : 0.75.

Par ailleurs, la validation croisée permet de retenir une valeur de nombre d'itérations **nrounds**= 109, résultat appuyé par la figure 5.17.

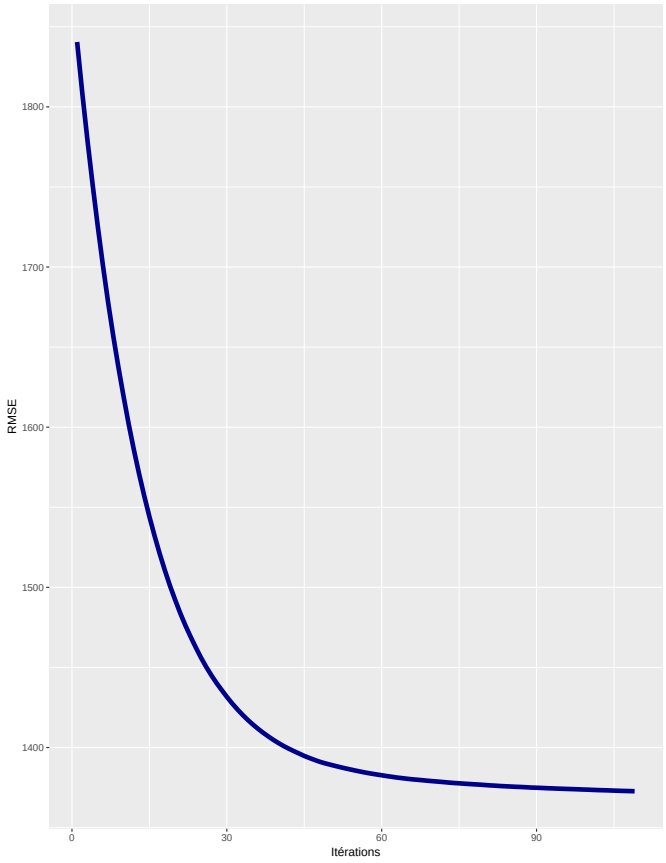


FIGURE 5.17 – Evolution de la RMSE sur la base d'apprentissage avec données télématiques en fonction du nombre d'itérations pour la modélisation du **coût du sinistre**

Importance des variables

La figure 5.18 résume l'importance des variables pour ce modèle.

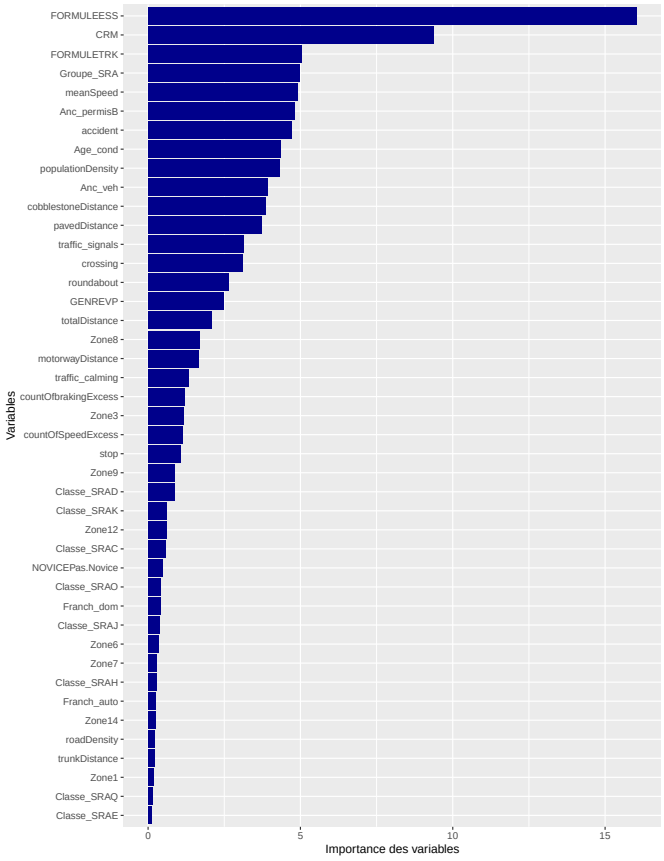


FIGURE 5.18 – Importance des variables dans la modélisation du **coût du sinistre** par XGBoost en intégrant les variables télématiques

Nous pouvons constater que pour le coût du sinistre, contrairement au nombre de sinistres, parmi les 10 variables les plus importantes figure la vitesse moyenne (meanSpeed), l'indicateur de dangerosité des accidents (accident) et la densité de population (populationDensity).

Mesure de la pertinence du modèle

Avec les paramètres retenus, nous calculons comme pour les modèles précédents, la RMSE, la MAE ainsi que l'erreur euclidienne. Elles sont présentées dans le tableau 5.17.

TABLE 5.17 – Erreurs obtenues pour la modélisation du **coût du sinistre** avec *XGBoost* avec les variables télématiques

	RMSE	MAE	Erreur euclidienne
Test	1444,946	942,8719	75801,16

La performance de ce modèle s'améliore par rapport au modèle sans compléments d'informations sur le comportement du conducteur. En effet, on passe d'une RMSE de 1448.542 à 1444.946. Soit un gain relatif de 0.25%. Par ailleurs, la MAE et l'erreur euclidienne baissent également dans ce modèle.

5.2.6 Synthèse et comparaison des modèles après intégration des données télématiques

Tout comme dans la partie 4.3, nous présentons dans le tableau 5.18 un récapitulatif de la RMSE obtenue sur la base test pour les modèles GLM, CART, *Random Forest*, GBM et XGBoost, après intégration des données externes.

TABLE 5.18 – Synthèse des erreurs des modèles sur la base test avant/après intégration des données télématiques

	Sans données télématiques		**Avec données télématiques**	
	Nombre de sinistres	**Coût du sinistre**	**Nombre de sinistres**	**Coût du sinistre**
GLM	**0,2210493**	1449,768	**0,2195956**	1449,171
CART	0,2217749	**1447,239**	0,2202025	1447,146
Random Forest	0,2216543	1452,612	0,2199053	1450,462
GBM	0,2220635	1551,913	0,2204248	1539,457
XGBoost	0,2216571	1448,542	0,2199385	**1444,946**

Pour le **nombre de sinistres**, le modèle GLM reste meilleur avec une amélioration de la RMSE après rajout des données externes (elle passe de 0.2210493 à 0.2195956). Il est suivi du *Random Forest* qui est également amélioré, sa RMSE passe de 0.2216543 à 0.2199053.

Quant à la prédiction du **coût du sinistre**, ce n'est plus le modèle CART qui ressort comme meilleur mais le modèle **XGBoost** avec un gain en prédiction nettement amélioré (la RMSE est égale à 1444.946). Il est ensuite suivi de CART dont la RMSE vaut 1447.146. Le GLM ne vient qu'en $3^{ième}$ position avec une RMSE égale à 1449.171.

5.2.7 Importance des variables dans les modèles après intégration des données externes

5.2.7.1 Nombre de sinistres

Nous représentons l'importance des variables (**en bleu**) de chaque modèle sur le tableau 5.19 après intégration des données télématiques. Nous mettons en évidence (**en rouge**) les variables dont la pertinence est la plus importante dans la prédiction. Les variables télématiques

liées au comportement n'apparaissent pas aussi importantes comme les variables liées à l'assuré et à son véhicule pour la prédiction du **nombre de sinistres**. Néanmoins, nous les mettons dans le tableau (**en orange**).

TABLE 5.19 – Tableau récapitulant les variables les plus importantes dans la modélisation du **nombre de sinistres** après intégration des données télématiques

GLM	CART	Random Forest	GBM	XGBoost
CRM	**CRM**	CRM	CRM	**CRM**
Ancienneté du véhicule	Ancienneté du permis B	Ancienneté du véhicule	Ancienneté du véhicule	Ancienneté du véhicule
Ancienneté du permis B	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Genre	Genre	Genre	Genre	Genre
Zone	Zone	Zone	Zone	Zone
Novice	Novice	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Classe_SRA	Classe_SRA	Formule	Formule	Formule
Formule	Formule	Groupe_SRA	Groupe_SRA	Ancienneté du permis B
Groupe_SRA	Groupe_SRA	Ancienneté du permis B	Ancienneté du permis B	Franchise dommage
Age du conducteur	Franchise dommage	Franchise dommage	Franchise dommage	roadDensity
Franchise dommage	roadDensity	roadDensity	roadDensity	Franchise dommage
trunkDistance	trunkDistance	trunkDistance	trunkDistance	trunkDistance
totalDistance	totalDistance	totalDistance	totalDistance	totalDistance
pavedDistance	pavedDistance	pavedDistance	pavedDistance	pavedDistance
crossing	crossing	crossing	crossing	crossing
meanSpeed	meanSpeed	meanSpeed	meanSpeed	meanSpeed
motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance
cobblestoneDistance	cobblestoneDistance	cobblestoneDistance	cobblestoneDistance	cobblestoneDistance
accident	accident	accident	accident	accident

Il apparaît que pour la modélisation du **nombre de sinistres**, aucune variable liée au comportement ne se distingue par rapport aux variables caractéristiques du contrat de l'assuré ou de son véhicule. En effet, bien qu'elles contribuent à l'amélioration des prédictions des modèles, elles ont un degré d'importance moins élevé que les variables du portefeuille.

5.2.7.2 Le coût du sinistre

Nous représentons l'importance des variables (**en bleu**) sur le tableau 5.20 après intégration des données télématiques. Nous mettons en évidence (**en rouge**) les variables dont la pertinence est la plus importante dans la prédiction.

TABLE 5.20 – Tableau récapitulant les variables les plus importantes dans la modélisation du **coût du sinistre** après intégration des données télématiques

GLM	CART	Random Forest	GBM	XGBoost
CRM	CRM	CRM	CRM	**Formule**
Ancienneté du véhicule	Ancienneté du permis B	Ancienneté du véhicule	Ancienneté du véhicule	CRM
Ancienneté du permis B	Age du conducteur	Age du conducteur	Age du conducteur	Age du conducteur
Genre	Genre	Genre	Genre	Genre
Zone	Zone	Zone	Zone	Zone
Novice	Novice	**Classe_SRA**	**Classe_SRA**	Classe_SRA
Classe_SRA	Classe_SRA	Formule	Formule	Ancienneté du véhicule
Formule	**Formule**	Groupe_SRA	Groupe_SRA	Ancienneté du permis B
Groupe_SRA	Ancienneté du véhicule	Ancienneté du permis B	Ancienneté du permis B	Groupe_SRA
Age du conducteur	Franchise dommage	Franchise dommage	Franchise dommage	populationDensity
Franchise dommage	roadDensity	roadDensity	roadDensity	Franchise dommage
trunkDistance	trunkDistance	roundabout	trunkDistance	traffic_signals
totalDistance	totalDistance	totalDistance	totalDistance	totalDistance
pavedDistance	pavedDistance	traffic_signals	stop	pavedDistance
crossing	crossing	crossing	crossing	crossing
meanSpeed	meanSpeed	meanSpeed	meanSpeed	meanSpeed
motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance	motorwayDistance
cobblestoneDistance	cobblestoneDistance	trunkdistance	cobblestoneDistance	cobblestoneDistance
accident	accident	accident	roundabout	accident

Nous pouvons constater que l'algorithme XGBoost est celui qui fait intervenir le plus les variables télématiques. Pour les modèles GLM, *Random Forest* et GBM, elles interviennent avec une intensité faible (mises en évidence avec la couleur **orange**). Par ailleurs, la variable qui représente la distance totale "totalDistance" apparaît parmi les trois variables les plus importantes dans CART.

Conclusion :

Dans ce chapitre, nous avons pu étudier l'impact de l'utilisation des données externes sur les prédictions du **nombre de sinistres** et du **coût du sinistre**.

En ce qui concerne la modélisation du **nombre de sinistres**, nous avons remarqué une légère amélioration des prédictions sur l'ensemble des modèles construits. En particulier, la RMSE du modèle GLM passe d'une valeur de 0.2210493 à une valeur de 0.2195956, faisant de celui-ci le meilleur modèle pour la prédiction du **nombre de sinistres**.

Pour le **coût des sinistres**, il s'est avéré que l'algorithme XGBoost a été le plus performant après l'ajout des données télématiques. En effet, sa RMSE vaut 1444.946, améliorant ainsi le modèle CART (le meilleur modèle sans données externes) dont la RMSE est de 1447.239. Par ailleurs, certaines variables comportementales influencent de manière notable le modèle XGBoost comme la vitesse moyenne ou la densité de population.

Nous pouvons alors souligner que l'apport des données externes est important car elles apportent un gain réel en termes de capacité prédictive.

Ainsi, comme la tarification constitue un enjeu majeur pour l'assureur, il serait envisageable d'intégrer ces données sur le comportement à son portefeuille, même si cela pourrait s'accompagner d'un travail de traitement sur ces données au préalable.

Par ailleurs, il serait intéressant d'intégrer des données relatives à la sinistralité comme « un indicateur du nombre de sinistres par densité de population » ou encore des données propres au trajet « domicile-travail » de l'assuré. Aussi, il aurait été pertinent d'extraire les données télématiques non pas par « code postal » comme cela a été fait par l'entreprise **Ellis Car**, mais en agrégeant par le trajet « domicile-travail » quotidiennement effectué par les assurés.

Par ailleurs, il est possible d'effectuer une classification non-supervisée (du type k-means par exemple) des communes à partir des données externes, l'idée est de définir des groupes de communes classés du moins risqué au plus risqué, autrement dit, il s'agirait de construire un zonier sur les communes. Ensuite, cette variable créée serait intégrée au portefeuille de l'assureur au lieu d'ajouter l'ensemble des variables télématiques.

Conclusion

La tarification des produits d'assurance représente un enjeu central pour les assureurs, particulièrement dans un marché automobile extrêmement concurrentiel et qui représente la première source de chiffre d'affaires en assurance de biens et de responsabilité. L'assureur est donc poussé à élaborer une tarification qui soit fine et précise sans pour autant supprimer le principe fondamental de la mutualisation, afin de mieux cerner les risques inhérents à son activité. Pour acquérir de nouveaux assurés au sein de son portefeuille tout en maintenant sa rentabilité, l'assureur doit maîtriser la segmentation de ses risques.

Dans cette optique, ce mémoire se propose de tester différentes méthodes de tarification, à travers la modélisation de deux quantités : **le nombre de sinistre** et le **coût des sinistres** d'un portefeuille RC automobile. Les outils de tarification employés par les compagnies d'assurance reposent principalement sur des méthodes économétriques traditionnelles : les modèles linéaires généralisés (GLM). Aujourd'hui, un certain nombre d'algorithmes innovateurs ont vu le jour, notamment du fait d'une ouverture des données de plus en plus conséquente. Ils ont pour objectif soit la prédiction de valeurs, soit la classification d'individus. Ces algorithmes ont donné naissance à la théorie de l'apprentissage statistique (*machine learning*).

Nous avons alors réalisé une étude comparative des performances réalisées par les modèles traditionnels (GLM) qui nous serviront de modèle de base, et de quelques modèles de *machine learning* : les arbres de décision CART, les forêts aléatoires, le *Gradient Boosting Machine* et l'*eXtreme Gradient Boosting Machine*. L'étude menée montre que les approches *data science* sont pertinentes et offrent des performances comparables à celles des GLM. Selon les cas, l'une ou l'autre approche apparaît plus appropriée. Pour les prédictions du **nombre de sinitres**, c'est le GLM qui ressort meilleur. Pour le **coût des sinistres**, c'est CART qui a des performances légèrement meilleures que le GLM.

Aujourd'hui, la majorité des assureurs ne basent leurs tarifs que sur des analyses GLM. Pourtant, les méthodes de *data science* s'avèrent souvent pertinentes, voire parfois supérieures aux approches classiques. Il serait sans doute intéressant de tester les deux familles d'approches lors des revues des tarifs et de déterminer au cas par cas celle qui est la plus pertinente. Ce mémoire illustre cette idée sur deux exemples.

Dans la seconde partie de ce mémoire, la base de données utilisée auparavant a été enrichie avec des variables issues de la télématique. Ces données ont été récupérées auprès d'une

entreprise spécialisée dans les comportements des conducteurs et qui oeuvre à réduire le risque d'accidents. En effet, tout apport d'information pouvant renforcer la caractérisation du profil de l'assuré est susceptible d'éviter l'antisélection. L'objectif est d'analyser les performances des modèles établis précédemment quand nous leur rajoutons ces données externes. Nous avons alors constaté qu'au global, la performance de tous les modèles s'améliorent. En particulier, le modèle *XGBoost* apporte une amélioration nettement meilleure que le modèle CART pour le **coût des sinistres** avec une contribution notable des données télématiques. Concernant le **nombre de sinistres**, le modèle GLM reste meilleur avec une légère amélioration par rapport au modèle sans données externes.

Enfin, nous avons pu relever dans ce mémoire que les données télématiques constituent une nouvelle opportunité pour les assureurs, d'autant plus qu'elles sont accessibles à tout le monde. Elles permettent d'effectuer des tarifs sur les petits portefeuilles malgré le peu d'informations dont ils disposent, et d'enrichir davantage les tarifs des gros portefeuilles.

Il serait alors intéressant d'intégrer des variables relatives à la sinistralité par commune ou encore, effectuer une classification non-supervisée pour définir un zonier sur le risque relatif à chaque commune et intégrer uniquement cette variable au portefeuille.

Annexe A

Annexe A

A.1 L'apport de l'assurance automobile connectée pour les français

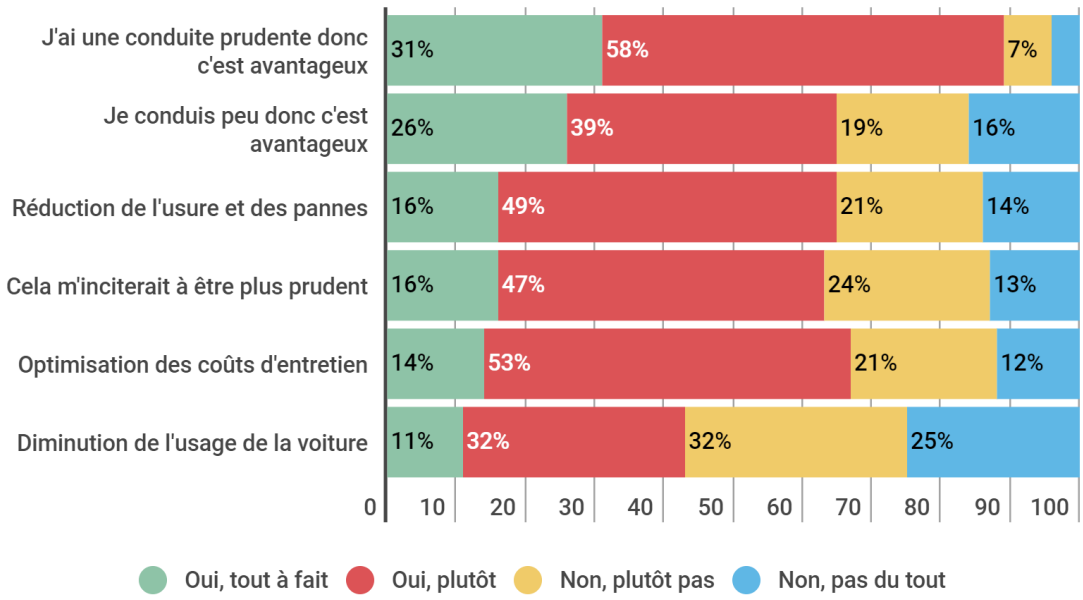


FIGURE A.1 – L'apport de l'assurance automobile connectée pour les français

A.2 La présence de la télématique en Europe

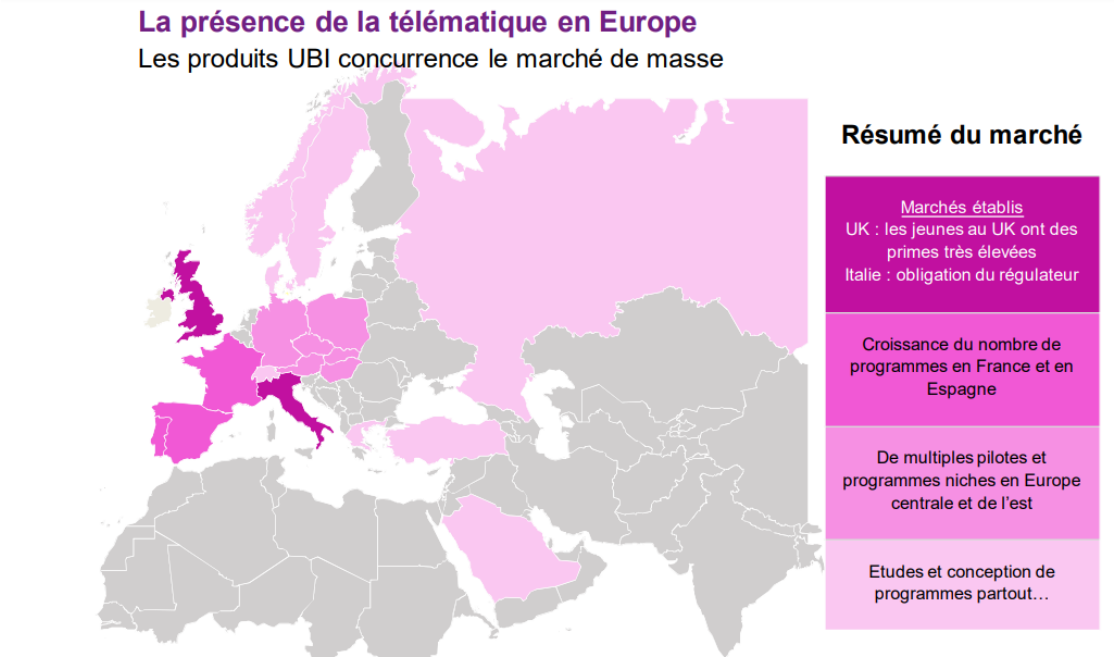


FIGURE A.2 – La présence de la télématique en Europe, *source :Willis Towers Watson*

A.3 Comment la télématique change l'assurance auto ?

FIGURE A.3 – Comment la télématique change l'assurance auto ?

A.4 L'avenir de la télématique

FIGURE A.4 – L'avenir de la télématique, *source :Willis Towers Watson*

Annexe B

Quelques statistiques univariées

B.1 Ancienneté du véhicule

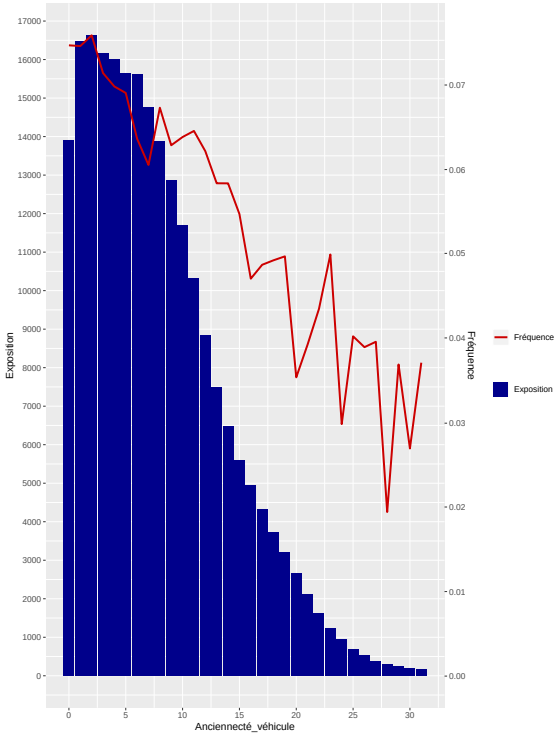


FIGURE B.1 – Fréquence par Ancienneté de véhicule

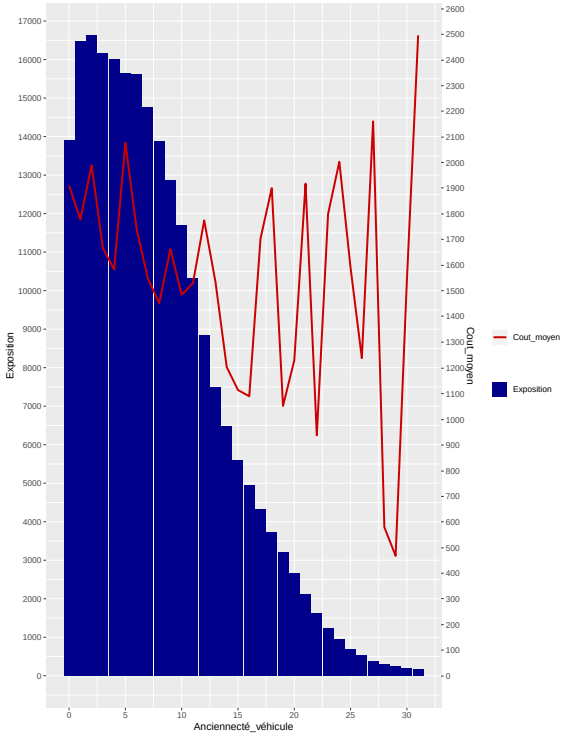


FIGURE B.2 – Coût moyen par Ancienneté de véhicule

B.2 Genre du véhicule

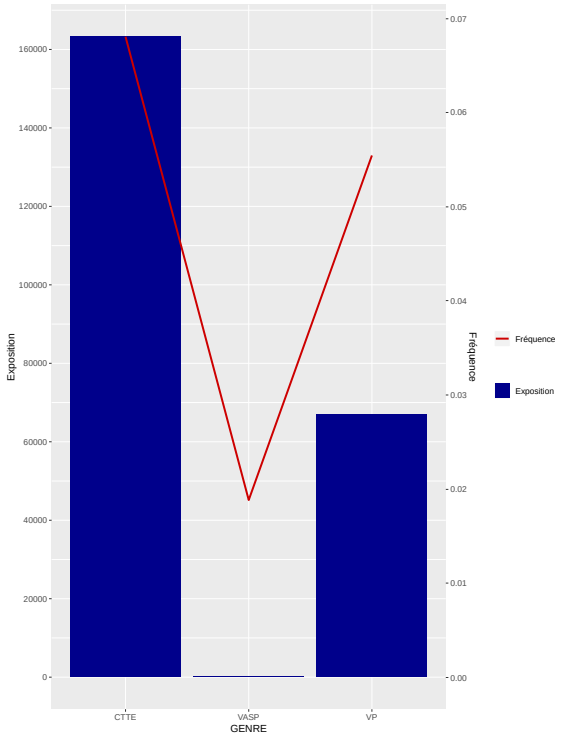


FIGURE B.3 – Fréquence par Genre

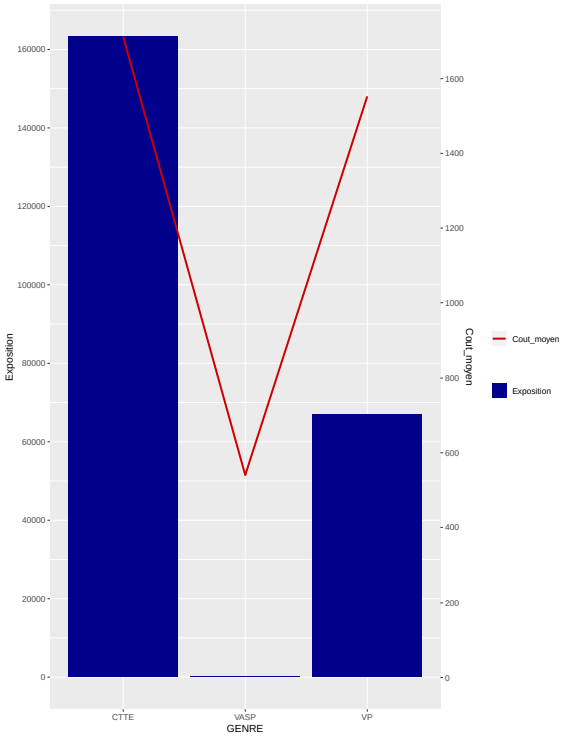


FIGURE B.4 – Coût moyen par Genre

B.3 Ancienneté de permis B

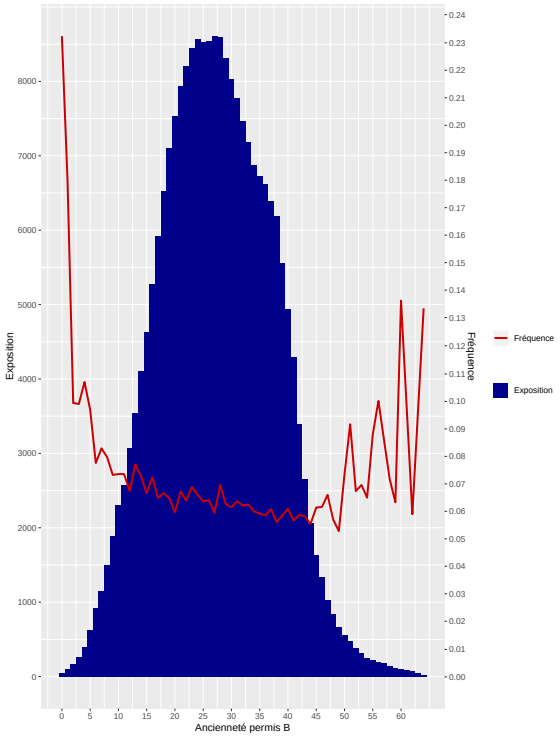


FIGURE B.5 – Fréquence par Ancienneté de permis B

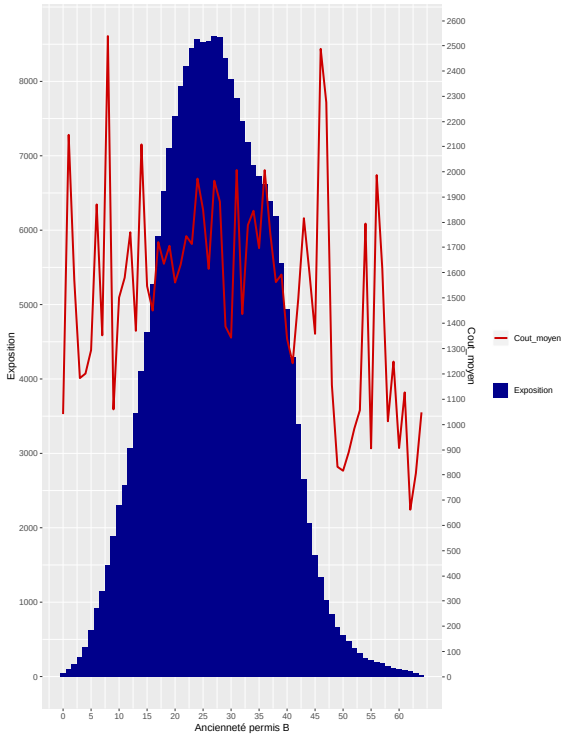


FIGURE B.6 – Coût moyen par Ancienneté de permis B

B.4 Formule

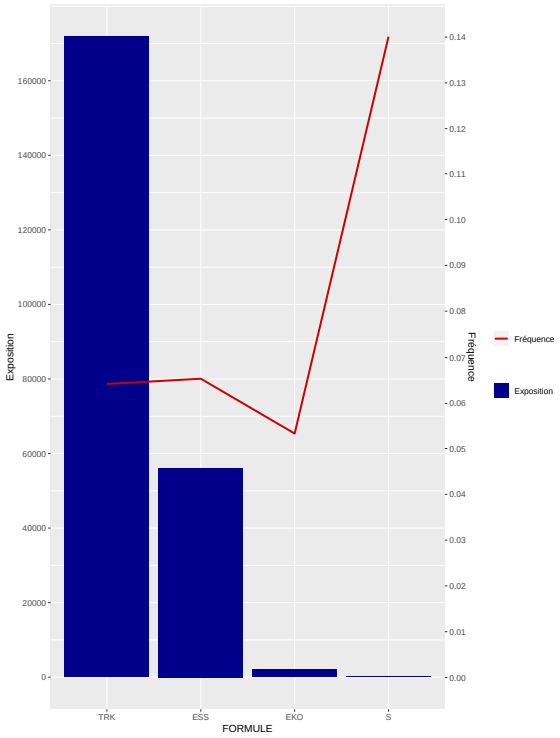


FIGURE B.7 – Fréquence par Formule

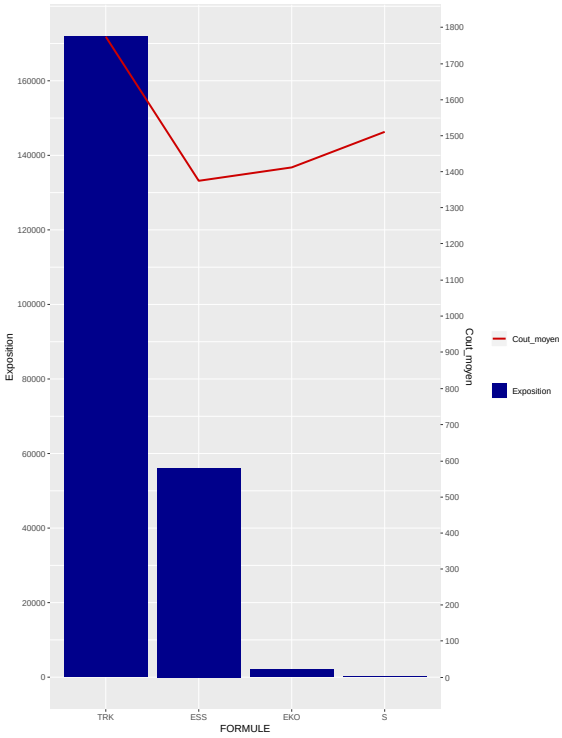


FIGURE B.8 – Coût moyen par Formule

Bibliographie

[ALLAIRE, 2016] ALLAIRE, G. (2016). La protection des données, enjeu de taille pour l'assurance. *La Revue du Courtage - mars 2016*, (921):44–51.

[BARBASTE, 2017] BARBASTE, M. (2017). Une méthode de provisionnement individuel par apprentissage automatique. *Mémoire Institut des Actuaires*.

[BELLINA, 2014] BELLINA, R. (2014). Méthodes d'apprentissage appliquées à la tarification non-vie. *Mémoire Institut des Actuaires*.

[BERAUD-SUDREAU, 2017] BERAUD-SUDREAU, G. (2017). Construction d'un zonier en mrh à l'aide d'outils de data-science. *Mémoire Institut des Actuaires*.

[BOLLACHE, 2015] BOLLACHE, V. (2015). Tarification automobile : des évolutions apportées dans le cadre de la loi hamon. *Mémoire Institut des Actuaires*.

[BOUCHÉ, 2014] BOUCHÉ, T. (2014). Modèle de propension des assurés par rapport aux risques de sinistres corporels graves en assurance automobile. *Mémoire Institut des Actuaires*.

[BOUTTIER, 2015] BOUTTIER, F. (2015). Modélisation de la prime pure de la garantie rc automobile. *Mémoire Institut des Actuaires*.

[EZZAKRAOUI, 2015] EZZAKRAOUI, N. (2015). Tarification d'un contrat santé collective et estimation de l'effet d'une surcomplémentaire. *Mémoire Institut des Actuaires*.

[FAUCHET, 2016] FAUCHET, G. (2016). Profil bancaire et risque en assurance automobile. *Mémoire Institut des Actuaires*.

[GRARI, 2015] GRARI, V. (2015). Impact des données exogènes sur la tarification en santé. *Mémoire Institut des Actuaires*.

[GUILBOT *et al.*, 2018] GUILBOT, M., VASLIN, L. et ARREGLE, E. (2018). Véhicule connecté, communicant, automatisé et protection des données à caractère personnel des usagers. *Congrès ATEC ITS-France. Les rencontres de la mobilité intelligente*.

[GUILLOT, 2015] GUILLOT, A. (2015). Apprentissage statistique en tarification non-vie : quel avantage opérationnel ? *Mémoire Institut des Actuaires*.

[Jacques et Rain, 2012] JACQUES, L. et RAIN, E. (2012). Du modèle glm à une approche darwinienne : Nouvelle génération de concepts et d'indicateurs pour l'optimisation de renouvellement auto. *Mémoire Institut des Actuaires*.

[KARATEKIN, 2014] KARATEKIN, O. (2014). Tarification et mesure de l'antisélection en assurance santé collective. *Mémoire Institut des Actuaires*.

[KOLASA et PSAUMEE, 2011] KOLASA, S. et PSAUMEE, J. (2011). Tarification en iard et nouvelles contraintes de rentabilité : Etude d'un produit flotte automobile. *Mémoire Institut des Actuaires*.

[LAGADEC, 2009] LAGADEC, F. (2009). Tarification d'un contrat de complémentaire santé par un modèle linéaire généralisé. *Mémoire Institut des Actuaires*.

[LANTZ, 2015] LANTZ, B. (2015). *Machine Learning with R*. Packt publishing.

[LE BOUCHER, 2016] LE BOUCHER, B. (2016). Tarification i.a.r.d et open data. *Mémoire Institut des Actuaires*.

[LE CAMUS, 2014] LE CAMUS, M. (2014). Méthodes de provisionnement en rc corporelle automobile. *Mémoire Institut des Actuaires*.

[LUO, 2015] LUO, Y. (2015). Amélioration de la modélisation de sinistres graves à l'aide d'une approche d'apprentissage. *Mémoire Institut des Actuaires*.

[LUTZ et BIERNAT, 2015] LUTZ, M. et BIERNAT, E. (2015). *Data Science : fondamentaux et études de cas : Machine Learning avec Python et R*. Eyrolles.

[NANA NJOYA, 2016] NANA NJOYA, E. S. (2016). Prédiction des comportements de rachat en épargne individuelle : une approche machine learning. *Mémoire Institut des Actuaires*.

[NGUYEN, 2008] NGUYEN, T. T.-V. (2008). Flottes automobiles :un nouveau modèle de tarification. impact de la conservation sur la distribution du ratio sinistres à primes. *Mémoire Institut des Actuaires*.

[NIAMKE, 2014] NIAMKE, D. G. (2014). La tarification des branches rc automobiles à développement long. *Mémoire Institut des Actuaires*.

[NICOLLE, 2017] NICOLLE, C. (2017). Tarification au trajet à l'aide de l'open data. *Mémoire Institut des Actuaires*.

[OSSENI, 2014] OSSENI, Z. (2014). Optimisation de la prise en compte de la sinistralité dans la tarification automoteur agricole. *Mémoire Institut des Actuaires*.

[PAGLIA, 2010] PAGLIA, A. (2010). Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. *Mémoire Institut des Actuaires*.

[PAGLIA *et al.*, 2011] PAGLIA, A., Phelippe-Guinvarc'h et V, M. (2011). Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. *Bulletin français d'Actuariat*, 11(22):49–81.

[PARIENTE, 2017] PARIENTE, J. (2017). Modélisation du risque géographique en assurance habitation. *Mémoire Institut des Actuaires*.

[Périclès,] Périclès. Machine learning : Du glm à l'arbre de cart en passant par le random forest. *Périclès Actuarial*.

[SADOUN, 2016] SADOUN, K. (2016). Apport des télématiques dans la segmentation tarifaire en assurance automobile. *Mémoire Institut des Actuaires*.

[TUFFERY, 2015] TUFFERY, S. (2015). *Modélisation prédictive et apprentissage statistique avec R*. Éditions Technip.

[VALEAU, 2016] VALEAU, A. (2016). Développement de la modélisation en complémentaire santé individuelle dans le contexte de l'accord national interprofessionnel du 11 janvier 2013. *Mémoire Institut des Actuaires*.

[VEGNI, 2011] VEGNI, M. (2011). Modélisation du coût des sinistres extrêmes en assurance automobile. *Mémoire Institut des Actuaires*.

[WEI, 2018] WEI, J. (2018). Homogénéité des risques : application à la sinistralité en assurance non-vie. *Mémoire Institut des Actuaires*.

[WICKHAM et GROLEMUND, 2016] WICKHAM, H. et GROLEMUND, G. (2016). *R for Data Science*. Éditions O'Reilly.

Table des figures

Liste des tableaux