Master Actuariat de Dauphine

Mémoire présenté devant l'Université Paris Dauphine pour l'obtention du diplôme du Master Actuariat et l'admission à l'Institut des Actuaires

le 26 Janvier 2018

Par : Maxence de LIGNAUD de LUSSAC

Titre : Comparaison de modèles prédictifs pour l'évaluation des coûts matériels automobiles

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

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

Membre présent du jury de l'Institut des Actuaires :

Signature :

Entreprise :

Nom : Prim'Act

Signature :

Directeur de mémoire en entreprise :

Nom : Frédéric PLANCHET & Inès Burr

Signature :

Membres présents du jury du Master Actuariat de Dauphine :

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

Signature du responsable entreprise :

Secrétariat :

Bibliothèque :

Signature du candidat :

Prim' Act

DAUPHINE UNIVERSITÉ PARIS

Mémoire d'actuariat

Maxence de Lignaud de Lussac

Comparaison de modèles prédictifs pour l'évaluation des coûts matériels automobiles

---

Tuteur de mémoire : Frédéric Planchet

Résumé

Dans le contexte actuel de compétitivité de l'assurance automobile, les sociétés d'assurance recherchent activement les moyens permettant d'améliorer leurs ratios combinés. En effet, les produits d'assurance automobile constituent une part importante de l'activité non vie, et l'assuré verra un gage de qualité chez l'assureur qui propose des tarifs concurrentiels sur ce type de produit. Ainsi, être performant sur les produits d'assurance automobile est un enjeu important pour l'assureur, dont l'exigence se retrouve jusque chez les prestataires de services.

Ce mémoire s'inscrit dans le cadre d'un travail dont l'objectif est d'évaluer la performance d'un réseau d'experts face à ses concurrents. Il s'articule également autour des problématiques d'explication du coût de sinistres matériels à l'aide de covariables. C'est pourquoi plusieurs modèles sont mis en concurrence afin de comparer leurs pouvoirs prédictifs respectifs.

Pour répondre à la première problématique, un modèle de régression type GLM (*Generalized Linear Model*) est construit. A chaque variable est estimé un coefficient permettant d'évaluer l'influence de celle-ci sur le coût du sinistre.
Ensuite, dans l'optique d'améliorer le pouvoir prédictif du GLM, celui-ci est inclus dans une méthode d'agrégation de modèle, le RGLM (*Random Generalized Linear Model*), en lui associant les caractéristiques des forêts aléatoires. Il a été possible, après plusieurs modifications de l'algorithme, d'améliorer substantiellement le pouvoir prédictif du GLM. Toutefois, cela a nécessité une complexification du modèle rendant l'interprétation plus délicate.
Enfin, une méthode d'apprentissage non paramétrique est utilisée : le *gradient boosting model* (GBM). Parmi les trois, il possède la meilleure qualité de prédiction, mais au détriment de la traçabilité des calculs et de l'explication du résultat.

Cette étude a mis en avant la performance des algorithmes de machine learning, que ce soit en les alliant à des techniques de régression plus classique (avec le GLM) ou en les utilisant tel quel (avec le GBM). Néanmoins, cette performance a un prix : l'interprétabilité des résultats, qui dans un contexte pratique peut être rédhibitoire.

Mots clefs: assurance automobile, assurance non-vie, réseau d'experts, GLM, RGLM, boosting, machine learning, prédictions, performance.

Abstract

In today's context of competitivity in automobile insurance, insurance companies are actively seeking means to lower their combined ratio. Indeed, automobile insurance's product constitute a large part of the non-life insurance market, meaning that insured persons will see a guarantee of quality with the Insurer that offer a competitive price on those products. Hence, performing on those products is a crucial stake for Insurers, which explain their requirement toward their service providers.

This memory thesis fits into a work environment which objective is to evaluate the gap of performance between two networks of experts. It is structured around the explanation of material disasters' cost using suitable covariables. Therefore, we let several models compete to compare their predictive power.

To answer the first problematic, a generalized linear model (GLM) is built. With a GLM, the influence of each variable on the disaster's cost is estimated through a coefficient.
Then, to improve the predictive power of the GLM, we add some characteristics of random forest to it, hence forming a random generalized linear model (RGLM), a model aggregation. After many changes of the algorithm, the predictive power of the GLM have been increased substantially. However, a complexification of the model was needed, making the interpretation of the results a little bit more delicate.
Finally, a non-parametric model is used: the gradient boosting model (GBM). Among the others, this model has the best quality of prediction, but to the detriment of the computation's traceability and the explanation of the result.

This study exposes the performance of machine learning's algorithms, whether coupled with more traditional regression techniques (GLM) or used by itself (GBM). Nonetheless, this performance has a price: the interpretability of results, which could be crippling in a practical context.

Key words: automobile insurance, non-life insurance, network of experts, GLM, RGLM, boosting, machine learning, predictions, performance.

Synthèse

Depuis quelques années, les bancassureurs s'attaquent au marché de l'assurance non-vie, et notamment celui de l'assurance automobile, accroissant ainsi la concurrence déjà forte sur ce domaine. Avec un ratio combiné d'en moyenne 105 % depuis plus de 10 ans dans cette branche de l'assurance, les assureurs redoublent d'efforts afin d'être capable de proposer des tarifs compétitifs. Cette exigence est portée jusqu'aux prestataires de services, dont la compétitivité est mise à l'épreuve.

Dans ce contexte, ce mémoire a pour objectif de modéliser l'écart de performance entre deux réseaux d'experts sur un portefeuille de sinistres matériels. Ajouté à cela, on s'intéresse à la construction et l'optimisation de modèles prédictifs, en étudiant l'impact d'une introduction progressive de techniques propres au machine learning.

1. Cadre de travail

Les données utilisées dans le cadre de l'étude proviennent de 3 assureurs, permettant d'avoir un volume important d'observations : 876 935 après retraitement. Cependant, les variables spécifiques à chaque assureur ont dû être écartées dans un souci d'homogénéisation de la base de données.

L'objectif de la modélisation est d'expliquer le montant d'expertise à l'aide d'une dizaine de variables explicatives, qui fournissent des informations concernant le véhicule, le type d'accident, le garage affilié, le lieu, la date d'occurrence du sinistre et le réseau d'experts ayant pris en charge le sinistre. C'est au travers de l'étude de l'influence de cette dernière qu'une estimation de l'écart de performance pourra être fournie.

La mise en place de la modélisation requiert la création de trois échantillons : apprentissage, validation et test. Le premier est utilisé pour la construction du modèle, le second pour l'optimisation des paramètres et le dernier pour l'évaluation de la qualité de prédiction. Ces échantillons sont communs à chaque modèle, afin d'avoir un périmètre commun propice à une comparaison objective. D'autre part, la qualité de prédiction est estimée en grande partie par l'erreur moyenne quadratique (MSE : *Mean squared error*), appuyée par une mesure de corrélation et d'un calcul de la norme L1.

2. Modèle linéaire généralisé (GLM)

On commence la modélisation par l'ajustement d'un GLM. Ce choix se repose sur la structure du GLM, qui permet d'apporter une réponse à l'estimation de l'écart de performance entre les deux réseaux d'experts. En effet, le GLM estime un coefficient par variable qui mesure l'impact de celle-ci sur le coût moyen. En utilisant une fonction de lien logarithmique, la relation devient assez intuitive :

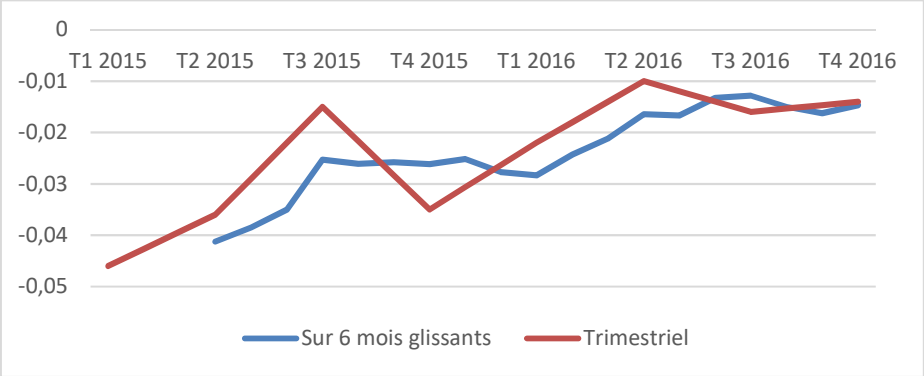
$$E[Y|X] \approx (1+\beta_1)^{X_1} * \ldots * \left(1+\beta_p\right)^{X_p}.$$

Pour la loi de Y | X, la loi Gamma est choisie pour sa flexibilité et son adéquation aux variables aléatoires asymétriques et positives.

Enfin, une étude des valeurs extrêmes a été menée, complétée par une analyse statistique de segments du portefeuille, mettant en évidence la nécessité d'un seuil à droite. Un raisonnement similaire est appliqué pour les sinistres à faibles coûts, et un écrêtage des sinistres de plus de 30 000€ et de moins de 334€ est alors effectué. Cet écrêtage est conservé pour les autres modèles.

Le GLM est alors construit à l'aide d'une sélection de variable selon le critère AIC, et l'écart de performance est estimé. Il est par exemple possible de produire une étude trimestrielle :

Fig. 1 : Evolution de l'écart de performance sur 6 mois glissants



La validation de ces résultats se fait au travers de l'étude des résidus de Pearson, qui dans notre cas sont de bonnes qualités, confirmant ainsi la robustesse du modèle.

3. RGLM

Afin d'améliorer le pouvoir prédictif du GLM, on s'intéresse au RGLM, méthode d'agrégation de modèle utilisant les caractéristiques des forêts aléatoires.

La première étape consiste à bootstrapper plusieurs fois l'échantillon d'apprentissage, à l'aide d'un tirage avec remise. C'est un moyen pour le modèle d'acquérir de l'expérience en créant une multitude de situations similaires aux observations. Ensuite, une sélection aléatoire de q variables est effectuée sur chaque échantillon bootstrappé, afin de créer des combinaisons de variables qui sont pas réalisables autrement. Enfin, un GLM identique à celui présenté auparavant est ajusté sur chaque échantillon bootstrappé, et les prédictions finales du RGLM sont déterminées comme la moyenne des prédictions de chaque GLM.

On constate donc l'apparition de deux nouveaux paramètres : le nombre d'échantillons bootstrappé et le nombre q de variables sélectionnées aléatoirement. Le premier est fixé à 100 pour des raisons de complexité, tandis que le deuxième est optimisé à l'aide de l'échantillon de validation. Cette optimisation conduit à conserver toutes les variables, et la qualité de prédiction du RGLM est équivalente à celle du GLM, ce qui suggère de retravailler l'algorithme pour mieux tirer parti de ses fonctionnalités.

Une deuxième application du RGLM est donc effectuée, appelée « RGLM amélioré », en introduisant les termes d'interactions de second ordre entre les variables. Pour deux variables explicatives x et y, cela revient à ajouter une nouvelle variable $z = x * y$. Le nombre de variables dans le modèle augmente donc significativement (passage de 9 à 45), afin d'apporter plus d'intérêt au paramètre de sélection aléatoire des variables. La méthode de sélection de variables est alors remplacée par une régression pénalisée de type LASSO, qui est plus efficace en termes de temps de calcul. D'autres points mineurs ont été modifiés, comme l'apparition forcée de facteurs discriminants et l'optimisation de l'agrégation en favorisant les modèles de meilleure qualité.

L'optimisation du nombre de variables sélectionnées aléatoirement donne encore une fois un net avantage à la conservation de toutes les variables, comme le tableau suivant en atteste.

Tab. 1. Erreurs du RGLM « amélioré » sur l'échantillon de validation

q	3	4	5	6	7	8	**9**
COR	0,59	0,61	0,64	0,66	0,65	0,66	**0,66**
MSE	3 479 941	3 334 662	3 155 130	3 053 032	2 959 632	2 931 436	**2 907 608**
L1	786	772	758	747	735	729	**722**

Néanmoins, le pouvoir prédictif du modèle s'est grandement amélioré en passant d'une erreur moyenne quadratique de 3 639 483 à 2 903 323 sur l'échantillon test.

Une troisième et dernière application du RGLM est faite, en modifiant cette fois-ci uniquement la méthode d'agrégation de modèle. Celle-ci ne retient désormais que le meilleur modèle. Cette méthode semble être celle qui propose la meilleure qualité de prédiction, tout en valorisant la sélection aléatoire de variable :

Tab. 2. Erreurs du RGLM « meilleur modèle » sur l'échantillon test

q	3	4	5	6	7	**8**	9
COR	0,62	0,62	0,64	0,65	0,66	**0,66**	0,66
MSE	3 135 499	3 141 007	2 993 409	2 955 974	2 909 052	**2 900 468**	2 902 814
L1	759	758	726	728	721	**716**	717

Il est à noter cependant que chaque version du RGLM est incapable de retranscrire une évaluation précise de l'écart de performance, soit à cause de la phase d'agrégation qui empêche le calcul d'un coefficient, soit à cause des termes d'interactions qui brouillent l'estimation de l'action de la variable.

4. Gradient boosting model (GBM)

Le dernier modèle à être ajusté sur les données est le Gradient boosting model (GBM). Le principe est d'appliquer itérativement des arbres de régression afin de corriger au fur et à mesure les prédictions, dans le sens d'une descente de gradient. Le GBM a donc une tendance naturelle à surapprendre, d'où l'importance de la décomposition de l'échantillon principal.

Contrairement aux deux autres modèles, le GBM possède une multitude de paramètres à optimiser, parmi lesquels on note entre autres : la profondeur maximale des arbres de régression, le coefficient d'apprentissage, une sélection aléatoire de colonnes, une sélection aléatoire de lignes et le nombre d'itérations de l'algorithme. De plus, la plupart de ces paramètres sont interdépendants, ce qui signifie que pour obtenir une optimisation de bonne qualité, il n'est pas suffisant d'optimiser un à un les paramètres.

De ce fait, on définit des bornes d'existences pour chaque paramètre et on recherche dans cette grille la combinaison menant au meilleur modèle. Il semble alors nécessaire d'introduire un facteur aléatoire dans la recherche car le nombre de combinaisons est trop important pour les parcourir itérativement. Après 12 heures de recherche aléatoire, 235 modèles différents ont pu être testés. On peut ainsi évaluer le gain de performance de l'algorithme par rapport à une application « simple » utilisant des paramètres par défaut :

Tab. 3. Tableau récapitulatif de la qualité de prédiction des modèles testés

	COR	MSE	L1
GLM	0,55	3 639 942	770
RGLM simple	0,55	3 639 483	770
RGLM amélioré	0,66	2 903 323	716
RGLM méthode "meilleur modèle"	0,66	2 900 468	716
GBM simple	0,72	2 477 139	661
GBM optimisé	**0,73**	**2 411 535**	**651**

Le GBM présente ainsi une qualité de prédiction supérieure à celle des deux autres modèles. Il permet également d'estimer l'influence relative des variables explicatives, qui est une fonction du nombre de fois où la variable a servi dans la scission d'un nœud, et de l'amélioration qu'une telle scission apporte à la qualité de l'arbre. Néanmoins, l'estimation étant relative, elle semble moins pertinente qu'une estimation brut telle que le GLM propose.

5. Conclusion

Parmi les 3 modèles testés, il semble que le GLM soit le plus à même de répondre à la problématique d'évaluation de l'écart de performance entre les deux réseaux d'experts. Sa structure simple permet une interprétation immédiate de l'influence de la variable et de son impact sur le coût d'un sinistre matériel.

En ce qui concerne la qualité de prédiction, le GBM détient les meilleures performances, suivi du RGLM, puis du GLM. D'autre part, sa structure est propice à l'analyse de grands volumes de données en un temps de calcul et une gestion de mémoire efficaces. Cependant, l'interprétation des résultats est assez délicate, surtout lorsqu'il est demandé d'expliquer la valeur d'une prédiction.

Le RGLM peut être considéré comme un entre deux, en conservant une interprétabilité correcte, tout en proposant une qualité de prédiction supérieure au GLM.

Enfin, il a pu être remarqué dans cette étude une incapacité des modèles à retranscrire correctement le coût du sinistre. Suite à une observation plus poussée de ce phénomène, il a été conclu que le nombre de variables explicatives disponibles n'est pas suffisant pour correctement distinguer certains sinistres, ce qui affecte substantiellement les performances des modèles prédictifs. Toutefois, cela n'impacte pas l'étude de l'écart de performance, qui se limitait de toute façon aux variables observées.

Globalement, le RGLM et le GBM peuvent constituer des alternatives intéressantes au GLM, particulièrement dans une problématique Big Data mais également dans un cadre statistique plus classique.

Synthesis

For the past few years, bank insurers tend to focus on the non-life insurance market, especially automobile insurance, thus increasing the already strong competition on this area. With a combined ratio that has been around 105% for the past decade in automobile insurance, insurers intensify efforts to propose competitive rates. This requirement reaches service providers as well, whose competitivity is tested more carefully.

In this context, the aim of this memory is to model the performance gap between two experts' networks on a portfolio of material disasters. Added to this, we are interested in building and optimizing predictive models, while studying the impact of a gradual introduction of machine learning's techniques.

1. Work environment

The data used in this memory come from 3 insurers, meaning that the volume of observations is quite important: 876 935 after reprocessing. However, some variables are specific to the insurer it originated from. Therefore, they had to be removed to homogenize the database.

The goal of modeling is to explain the expert estimate with the help of a bunch of explanatory variables, which provide information on the vehicle, the type of the accident, the garage affiliated, the area, the date and the experts' network of the disaster. It is through the study of the influence of the latter variable that it will be possible to provide an estimation of the performance gap.

The implementation of modeling requires the creation of three samples: learning, validation and test. The first one is used to build the model, the second one to optimize parameters and the last one to assess the quality of prediction. Those samples are common to each model, to establish a perimeter conductive to an objective comparison. Furthermore, the quality of prediction is estimated by the mean squared error (MSE), supported by the correlation measurements and the computation of the L1 norm.

2. Generalized linear model (GLM)

We start the modeling part by adjusting a GLM. This choice is base on the structure of this method, which is able to bring an answer to the estimation of the performance gap between the networks of experts. Indeed, the GLM estimate a coefficient for each variable that measures its impact on the disaster's cost. With the use of a logarithmic link function, the relation becomes rather intuitive:

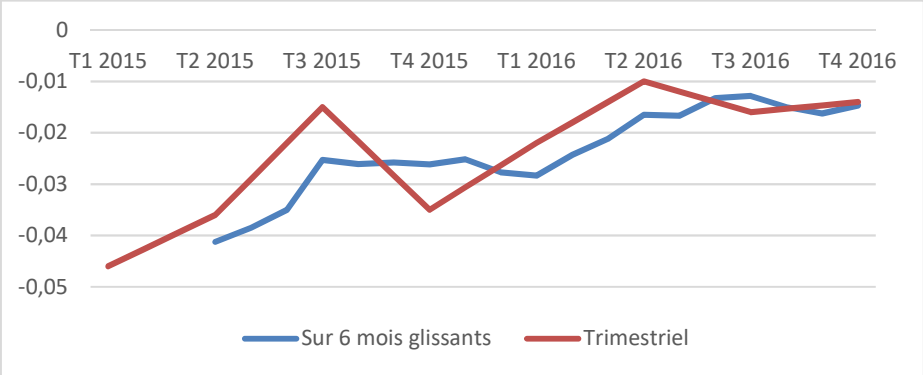
$$E[Y|X] \approx (1+\beta_1)^{X_1} * \ldots * \left(1+\beta_p\right)^{X_p}.$$

For the distribution law of Y | X, the Gamma law is chosen for its flexibility and its adequacy to asymmetric and positive random variables.

Finally, a study of extreme values is carried out, completed by a statistic analysis of the portfolio's segments, which lead to the necessity of establishing a threshold. A similar reasoning is applied to the smaller disasters, and a clipping has been set to remove disasters above 30 000€ and below 334€. This clipping is kept for the other models.

To finish the construction of the GLM, a variable selection through the AIC criterion is applied, thus allowing to estimate the performance gap. For instance, it is possible to produce a quarterly survey:

Fig. 1 : Evolution of the performance gap over a sliding 6-months period



To validate the results, we checked the Pearson residuals, which seem to be of good quality, thus confirming the model's robustness.

3. RGLM

So as to increase the predictive power of GLM, we introduce random generalized linear models (RGLM), which is method of model aggregation based on characteristics of random forests.

The first step is to bootstrap many times the learning sample, through a draw with replacement. This way the model will get experience by creating multiple situations that resemble observations. Then, a random selection of q variables is performed on each bootstrapped sample, which generates combinations of variables that weren't feasible otherwise. Finally, we adjust a GLM, similar to the one presented before, on each bootstrapped sample. The final predictions are then computed as the mean of the predictions of each GLM.

This method introduces two new parameters: the number of bootstrap samples and the number q of variables randomly selected. The first one is set at 100 to limit the complexity of the algorithm, whereas the second one is optimized with the validation sample. This optimization leads to keeping all variables, and the RGLM's quality of prediction is

equivalent to the GLM's, suggesting that we should do some work on the RGLM's algorithm to take advantage of its functionalities.

Therefore, a second application of the RGLM is made, called "Enhanced RGLM", that brings the interaction terms of second order between variables. Hence, for two explanatory variables x and y, a third variable $z = x * y$ is added to the relation. The number of variables in the model increase significantly (from 9 to 45) to bring more interest to the parameter of variables' random selection. We replace also the method of variable selection by a LASSO penalized regression, which is more effective in terms of computation. Other minor points have been changed, like forcing the appearance of a discriminating factor and the method of aggregation, that now favors models with higher quality.

The optimization of the number of variables randomly selected gives once again a clear lead to keeping all variables, as shown in the next table.

Tab. 1. « Enhanced RGLM » errors on the validation set

q	3	4	5	6	7	8	**9**
COR	0,59	0,61	0,64	0,66	0,65	0,66	**0,66**
MSE	3 479 941	3 334 662	3 155 130	3 053 032	2 959 632	2 931 436	**2 907 608**
L1	786	772	758	747	735	729	**722**

Nevertheless, the predictive power of the model has increased a lot, with the reduction of the mean squared error from 3 639 483 to 2 903 323 on the test set.

A third and last application of the RGLM is done, this time modifying only the aggregation method. From now on, it only keeps the best model. This method seems to be the one that has the best quality of prediction, while valuing the variables' random selection step:

Tab. 2. « Best model RGLM » errors on the test set

q	3	4	5	6	7	**8**	9
COR	0,62	0,62	0,64	0,65	0,66	**0,66**	0,66
MSE	3 135 499	3 141 007	2 993 409	2 955 974	2 909 052	**2 900 468**	2 902 814
L1	759	758	726	728	721	**716**	717

That being said, it is important to note that none of the RGLM versions is capable of evaluating precisely the performance gap between the networks of experts.

4. Gradient boosting model (GBM)

The last model that is adjusted to the data is the Gradient boosting model (GBM). The principle is to apply iteratively regression trees to correct predictions progressively, following a gradient-based descent minimization. Thus, GBM has a natural tendency to overfit, that is why splitting the original sample is important.

By contrast with the two other models, GBM have a multitude of parameters to optimize, which include among others: maximal depth of regression trees, learning rate, random

selection of columns, random selection of lines and the number of iterations. Moreover, most of these parameters are interdependent, which means the optimization of parameters can't be made one by one.

Thus, we determine maximum limits for each parameter and search in a grid the combination leading to the best model. It seems necessary to introduce a random factor in this search as the number of combinations is to high to test them all iteratively. After 12 hours of random grid searching, 235 models were tested. It is now possible to assess the performance gain of the optimized GBM compared to a "simple" application that is using default parameters:

Tab. 3. Summary of models' quality of prediction on the test set

	COR	MSE	L1
GLM	0,55	3 639 942	770
Simple RGLM	0,55	3 639 483	770
Enhanced RGLM	0,66	2 903 323	716
RGLM "best model" method	0,66	2 900 468	716
Simple GBM	0,72	2 477 139	661
Optimized GBM	**0,73**	**2 411 535**	**651**

The optimized GBM seems to be the best model in terms of quality of prediction. Besides, it allows to estimate the relative influence of explanatory variables, through the number of times each variable was used to split a node and the improvement it brought to the regression tree. However, the estimation being relative, it seems this measure serves little purpose compared to a gross estimation as in the GLM.

5. Conclusion

Among the 3 models tested, it seems that the GLM is the one that answer the best the problematic of evaluating the performance gap between two networks of experts. Indeed, its simple structure allows a clear interpretation of the associated variable's influence, and its impact on the cost of a material disaster.

In terms of quality of prediction, the GBM has the best performance, followed by the RGLM and then by the GLM. It is important to note that the structure of the algorithm suits well the analysis of great volume of data, with a reasonable computing time and an efficient memory management. Nevertheless, the interpretation of the results might be a bit delicate, especially when it is requested to explain a particular value of a prediction.

The RGLM is an in-between, keeping a correct interpretability while proposing a quality of prediction superior to the GLM.

Lastly, it has been observed throughout this study that the models have difficulty predicting correctly the disaster's cost. By taking a deeper look, it has been revealed that

the explanatory variables available were not sufficient to correctly distinguish some disasters, thus affecting substantially the performance of our predictive models. Nonetheless, it should not impact the study of the performance gap, whose work is limited to the observed variables anyway.

Overall, the RGLM and the GBM constitute interesting alternative to the GLM, particularly in a Big Data problematic but also in a more traditional statistic work environment.

Remerciements

Je tiens à remercier dans un premier temps Frédéric PLANCHET, directeur associé chez Prim'Act et membre agrégé de l'institut des actuaires, pour la confiance et l'intérêt qu'il a su porter à mon travail. Ses conseils m'ont été d'une aide précieuse, et ce fut un réel plaisir de travailler avec lui.

Je remercie également le cabinet Prim'Act pour son accueil, mais aussi pour l'excellent cadre de travail qui m'a permis d'être dans les meilleurs conditions pour mener à bien mon mémoire.

Un grand merci à Inès BURR, ma tutrice chez Prim'Act, pour sa présence, son suivi et son soutien. Une attention particulière également à Joëlle SIAKA et Paul LACOUR, consultants chez Prim'Act, pour leurs conseils et leurs convivialités.

J'adresse également mes remerciements au corps professoral du département MIDO de l'université Paris Dauphine, pour sa qualité d'enseignement qui a su rendre ses trois dernières années très enrichissantes. Je n'oublie pas non plus toutes les personnes qui ont su m'accompagner durant ma scolarité.

Enfin, je remercie infiniment ma famille, pour son soutien infaillible durant toutes ces années.

Table des matières

RÉSUMÉ 3
ABSTRACT 4
SYNTHÈSE SYNTHÈSE DU MÉMOIRE EN 4 PAGES MAX 5
SYNTHESIS 5
REMERCIEMENTS 10
TABLE DES MATIÈRES 16
INTRODUCTION GÉNÉRALE 20
CHAPITRE 1 : CADRE DE TRAVAIL ET DONNÉES 22
I. GÉNÉRALITÉS 22
I.a. Objectif de l'étude 22
I.b. Modélisation 22
I.b.1. Principe de la modélisation 23
I.b.2. Comparaison du pouvoir prédictif des modèles – cas du sur-apprentissage 24
II. PRÉPARATION DES DONNÉES 26
II.a. Présentation des données 26
II.a.1. Contexte des variables 28
II.a.2. Création de variable 29
II.b. Retraitement des variables 29
II.b.1. Le code postal 31
II.b.2. Le kilométrage 32
II.b.3. L'issue 33
II.b.4. La variable cible 33
II.b.5. L'âge du véhicule 33
II.b.6. Les taux horaires 34
II.b.7. Le montant d'expertise 36
II.b.8. Le département 36
II.b.9. La marque 36
II.b.10. Bilan du nombre d'observations supprimées 37
II.c. Biais entre la variable cible et l'assureur 38
II.d. Composition du portefeuille de sinistres par réseau 38
II.e. Liaisons des variables explicatives avec la variable à expliquer 40

II.f. Liaisons entre les variables explicatives 41
II.f.1. V de cramer 42
II.f.2. La variable assureur 43
CHAPITRE 2 : MODÉLISATION **46**
I. MODÈLE LINEAIRE GENERALISÉ **46**
I.a. Théorie du GLM 46
I.b. Paramétrisation du GLM 47
I.b.1. Prérequis 47
I.b.2. Optimisation des variables 55
I.c. Application et résultats du GLM 56
I.c.1. Périmètre d'application du modèle 56
I.c.2. Estimation de la performance du réseau 56
I.c.3. Validation du modèle 59
I.d. Pouvoir prédictif du GLM 60
I.d.1. Périmètre 60
I.d.2. Résultats de la mesure d'erreur 60
I.e. Conclusion 61
II. RGLM **62**
II.a. Principe du RGLM 62
II.a.1. Contexte 62
II.a.2. Algorithme 63
II.a.3. Estimation de l'écart de performance des réseaux 65
II.b. Application du RGLM 66
II.b.1. Introduction de nouvelles mesures d'erreurs 66
II.b.2. Optimisation des paramètres 67
II.b.3. Résultats et comparaison avec le GLM 68
II.c. Proposition d'amélioration de l'algorithme et son application 68
II.c.1. Ajout des termes d'interactions 68
II.c.2. Facteurs discriminants 69
II.c.3. Méthode de sélection des variables 69
II.c.4. Agrégation des prédictions 71

II.c.5. Optimisation du nombre q de variables sélectionnées aléatoirement 72
II.c.6. Résultats du RGLM amélioré 73
II.d. Méthode alternative d'agrégation : le meilleur modèle 74
II.e. Conclusion 76
III. STOCHASTIC GRADIENT BOOSTING 77
III.a. Méthode du boosting 77
III.a.1. Idée générale du boosting 77
III.a.2. Le Stochastic Gradient Boosting 78
III.a.3. Implémentation de l'algorithme 79
III.a.4. Influence des variables explicatives 80
III.b. Paramétrage du modèle 81
III.b.1. Retraitement utilisé 81
III.b.2. Première application du boosting 81
III.b.3. Optimisation des paramètres 84
III.c. Application et résultats 92
III.c.1. Résultats du modèle optimisé 92
III.c.2. Importance relative des variables 93
III.d. Conclusion 95
IV. ANALYSE DES RÉSULTATS 96
IV.a. Evaluation de l'écart de performance des réseaux 96
IV.a.1. GLM 96
IV.a.2. RGLM 96
IV.a.3. GBM 97
IV.a.4. Conclusion 97
IV.b. Prédiction des modèles 98
IV.b.1. Récapitulatif des résultats de l'étude 98
IV.b.2. Analyse des prédictions 99
IV.b.3. Limites de la modélisation 101
IV.c. Cohérence des résultats 104
CONCLUSION GÉNÉRALE 105
ANNEXE 107
A. PRÉPARATION DES DONNÉES 107

B. GLM 108
C. RGLM 109
D. BOOSTING 110
E. ANALYSE DES RÉSULTATS 112
BIBLIOGRAPHIE 114

Introduction générale

Avec un chiffre d'affaires total de 20,8 milliards d'euros en 2016 et une progression stable d'environ 2 % par an, le marché de l'assurance automobile a de bons arguments pour attirer les investisseurs. Cependant, le contexte actuel de compétitivité a de quoi noircir le tableau, avec notamment la présence de près de 100 acteurs sur le marché, d'un ratio combiné d'en moyenne 105 % (FFA, 2016), de la loi Hamon, de règles de solvabilité de plus en plus exigeantes…

Pour faire face à ces défis, les assureurs redoublent de mesures pour redresser leurs résultats techniques. La diversification des offres en est une. Elle inclue une dimension service à l'assurance automobile, qu'on retrouve à travers la garantie bris de glace par exemple.
D'autre part, il est possible d'observer une hausse des tarifs, principalement due à une augmentation des coûts des sinistres matériels et corporels, ainsi qu'à la fin de la baisse des fréquences sinistres.
Enfin, les assureurs jouent sur la concurrence des prestataires de service, avec par exemple des taux horaires avantageux auprès de certains garages.

Un des objectifs de ce mémoire est d'intervenir sur ce dernier point, en fournissant une méthode permettant de modéliser l'influence d'un prestataire de service dans le coût du sinistre. Plus particulièrement, on s'intéresse à l'influence d'un réseau d'experts dans le cadre d'un portefeuille de sinistres matériels.

Aux défis précédemment cités, il ne faut pas oublier d'ajouter les nouveaux challenges de l'assurance automobile : le Big Data. En 2020, selon une étude de Deloitte, la part du chiffre d'affaire allouée aux véhicules connectés atteindrait 12 %, soit 2,2 milliards d'euros (Legoff, 2016). En ce qui concerne l'actuariat, cette perspective constitue un enjeu technique important. Les compétences nécessaires pour traiter ce genre de problème, à savoir les statistiques en grande dimension, sont assez différentes de ce qui est actuellement utilisé. En plus de la partie mathématique et statistique, l'informatique joue un rôle important en pratique pour arriver à une méthode performante.

Une partie de ce mémoire est donc consacrée à l'exploration de nouveaux modèles prédictifs, en associant le Machine Learning avec les méthodes de régression actuellement utilisées. L'application de ces modèles sera faite dans un cadre classique de prédiction du coût de sinistres matériels, afin de tester la performance de ces algorithmes inhabituels.

L'étude est divisée en deux chapitres.

Le premier consiste à présenter le cadre de travail et la structure des données. On y rappelle l'objectif de l'étude, puis les bases de la modélisation afin de poser un cadre propice à une bonne évaluation des différentes méthodes utilisées. Ensuite, les données sont analysées et retraitées afin de préparer au mieux l'étape de modélisation.

Le deuxième chapitre, réservé à la modélisation, est composé de 4 parties. Trois parties sont consacrées à la présentation, la paramétrisation, l'optimisation et l'application des modèles, à savoir : le modèle linéaire généralisé (GLM : *generalized linear model*), l'agrégation de modèle RGLM (*Random generalized linear model*), et un modèle de boosting reposant sur une descente de gradient, le *Stochastic Gradient Boosting Model* (SGBM ou GBM). Enfin, la dernière partie est dédiée à l'analyse des résultats de ces modèles.

Ainsi, nous aurons un bon aperçu des avantages et des inconvénients de chaque modèle vis-à-vis des données utilisées, ce qui permettra d'en apprendre plus sur leur fonctionnement dans le cadre particulier d'un portefeuille de sinistres matériels.

CHAPITRE 1 : CADRE DE TRAVAIL ET DONNEES

Ce chapitre a pour but d'introduire au mieux la modélisation et les objectifs poursuivis. Dans un premier temps, il est présenté les principes de la modélisation ainsi que les objectifs à atteindre lors de cette étude. La deuxième partie, quant à elle, est destinée à la présentation des données disponibles à l'étude. Il y est précisé la nature et la signification de chaque variable, suivi de la partie retraitement des données, afin d'en retirer les valeurs aberrantes. Enfin, les différentes variables ont été mises en relation afin d'étudier les liaisons qui peuvent exister entre elles. L'objectif de cette manœuvre est d'extraire un maximum d'informations en amont, pour être dans les meilleures conditions de modélisation possibles.

I. GENERALITES

Cette partie détaille le cadre général retenu et notamment les méthodes qui permettent de comparer les modèles entre eux. Cette partie est volontairement détaillée afin que l'étude proposée par la suite soit la plus claire et la plus cohérente possible.

I.a. Objectif de l'étude

L'objectif est double. Dans un premier temps, on s'intéresse à la reconstitution du coût moyen en optimisant le pouvoir prédictif du modèle. Dans un deuxième temps, il s'agit de mesurer l'impact d'une variable en particulier, le réseau d'expert, sur le coût du sinistre lorsque le modèle le permet.

De ce fait, nous ne disposons pas d'un portefeuille d'assurés mais d'un portefeuille de sinistres. Le montant des sinistres n'est relatif qu'aux dégâts matériels subit par le véhicule, évalué par l'expert lors de son intervention. On parle de manière équivalente de « coût du sinistre » et de « montant d'expertise ».

Pour plus de simplicité et préserver la confidentialité de l'étude, les réseaux d'experts sont différenciés en leur attribuant une lettre. Le premier, dont il est demandé d'étudier la performance, est référencé par « réseau A », et les concurrents par « réseau B ».

I.b. Modélisation

Cette section présente de manière générale ce qu'est la modélisation, comment elle est appréhendée, et quels sont les outils pratiques utilisés afin de comparer les modèles. Cette partie s'inspire du mémoire réalisé par R. Bellina (Bellina, 2014).

I.b.1. Nécessité de modéliser

Afin d'évaluer l'écart de performance entre les réseaux, une méthode rapide consisterait à regarder le coût moyen des sinistres pour chaque réseau. Le réseau ayant le montant d'expertise le plus bas serait alors le plus performant des deux. Cette méthode est valable à condition de s'assurer que les deux réseaux traitent des sinistres de même nature et en même proportion. En pratique, ce n'est pas le cas. Par exemple, admettons qu'un réseau soit présent dans une zone géographique où le niveau de vie est élevé. Il est probable que ce réseau traite des sinistres qui concernent des véhicules plus récents et donc pour un type d'accident équivalent le coût sera plus important. Ainsi, pour être objectif quant à la comparaison de la performance de chaque réseau, il faut pouvoir prendre en compte les dissimilarités relatives à la nature des sinistres traités par chaque réseau.

De ce fait, il semble important de modéliser le coût moyen en fonction de plusieurs variables préalablement choisies. Plutôt que de simplement regarder le coût moyen par réseau, le précédent exemple rend compte de l'importance de l'âge du véhicule et de la localisation du sinistre dans la modélisation du coût moyen d'un sinistre.

I.b.2. Principe de la modélisation

L'idée est simple, mais la réalisation en est complexe : trouver une relation reliant la variable réponse $Y \in \mathbb{R}^n$, le montant de l'expertise, et les caractéristiques du sinistre $X = (X_1, \dots, X_p) \in \mathbb{R}^{n \times p}$. Autrement dit, on cherche une fonction f telle que $Y = f(X)$, l'objectif étant de pouvoir prédire la valeur de Y lorsque X est connu.

Dans la réalité, il est probable que la relation soit plutôt de la forme $Y = h(X, Z)$ avec h une fonction peu ou pas régulière et Z des variables inobservées. C'est pour cela qu'il n'existe pas de modélisation parfaite. Le simple fait de chercher une relation entre Y et X induira un bruit tant par les variables non prises en compte que par la recherche de la fonction f, qui ne fera qu'approximer la « vraie » fonction.

En pratique, il y a en général assez de variables dans les bases de données pour expliquer la variable réponse, et la difficulté réside plutôt dans la sélection des variables les plus pertinentes, car il n'est pas possible de toutes les garder.

La recherche de la fonction f dépend de la méthode utilisée, mais répond à la même problématique d'optimisation : minimiser l'erreur de prédiction quantifiée par une fonction de perte L.

$$f = \underset{\varphi \in \mathcal{F}}{\operatorname{argmin}} L\big(Y, \varphi(X)\big).$$

Plusieurs fonctions de pertes sont envisageables, la plus répandue étant l'erreur quadratique moyenne : $L(u, v) = E[(u - v)^2]$.

En effet, cette fonction de perte est pratique car elle réduit fortement le nombre de possibilités pour la fonction f. Si Y est de carré intégrable, il y a existence et unicité de la projection orthogonale de Y dans l'espace L^2 qui se trouve être l'espérance de Y conditionnellement à X.

En termes mathématiques, cela donne :

$$\forall Z\ intégrable\ \sigma(X) - mesurable,$$
$$E[(Y-Z)^2] \geq E[(Y-E[Y|X])^2]$$

Ce qui veut dire que $f(X) = E[Y|X]$ minimise l'erreur quadratique moyenne. Ainsi, l'espérance conditionnelle de Y sachant X est considérée comme la meilleure prédiction de Y lorsque X est connu, au sens de l'erreur quadratique moyenne. Ce résultat est important, et certains types de modélisation tels que l'approche GLM l'utilisent comme fondement de leurs théories.

Néanmoins, il est à noter que certains modèles ne fournissent pas d'expression précise de la fonction f, en se contentant de fournir une classification permettant de prédire la variable d'intérêt à travers la segmentation effectuée.

I.b.3. Comparaison du pouvoir prédictif des modèles – cas du sur-apprentissage

En pratique, on dispose de réalisations pour les variables Y et X, que l'on note $(y_i, x_i)_{i\in[\![1,n]\!]}$. Lorsqu'on discrétise le problème ci-dessus, on obtient :

$$\underset{f\ \in\ \mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2$$

Pour certains modèles, notamment les modèles basés sur l'apprentissage statistique, ce problème tel qu'il est posé est trop permissif. Par exemple, la fonction $f(x) = \sum_{i=1}^{n} y_i \mathbb{1}_{\{x=x_i\}}$ est bien une solution du problème, mais ne convient pas pour des raisons évidentes : une fonction de ce type n'a aucun pouvoir prédictif, puisqu'elle attribue un montant d'expertise nul à toutes nouvelles observations.

Il est donc nécessaire d'ajouter une contrainte à ce critère pour éviter les solutions de ce genre.

De manière générale, il est possible d'éviter le sur-apprentissage en utilisant la validation croisée. Plusieurs approches existent ; celle qui est retenue consiste à découper l'ensemble de la base de données en trois échantillons distincts. Cette approche est notamment présentée dans l'article de A. Paglia (Paglia, 2011).

Le découpage se présente de la sorte :

- L'échantillon d'apprentissage : il représente environ 60% de la base de données et est utilisé pour la construction du modèle.

- L'échantillon de validation : il représente environ 20% de la base de données. L'optimisation des paramètres du modèle est réalisée sur cet échantillon.

- L'échantillon test : il représente 20% de la base de données, et est utilisé pour l'évaluation des prédictions.

Chacun de ces échantillons remplit un rôle important lors de la modélisation.

L'échantillon de validation joue le rôle de pivot entre l'échantillon test et l'échantillon d'apprentissage. Il permet d'optimiser les paramètres sans toucher à l'échantillon test. Son objectif premier est d'éviter le sur-apprentissage, en introduisant une seconde mesure de la qualité du modèle. On s'assure ainsi de la non-dépendance du modèle à l'échantillon d'apprentissage.

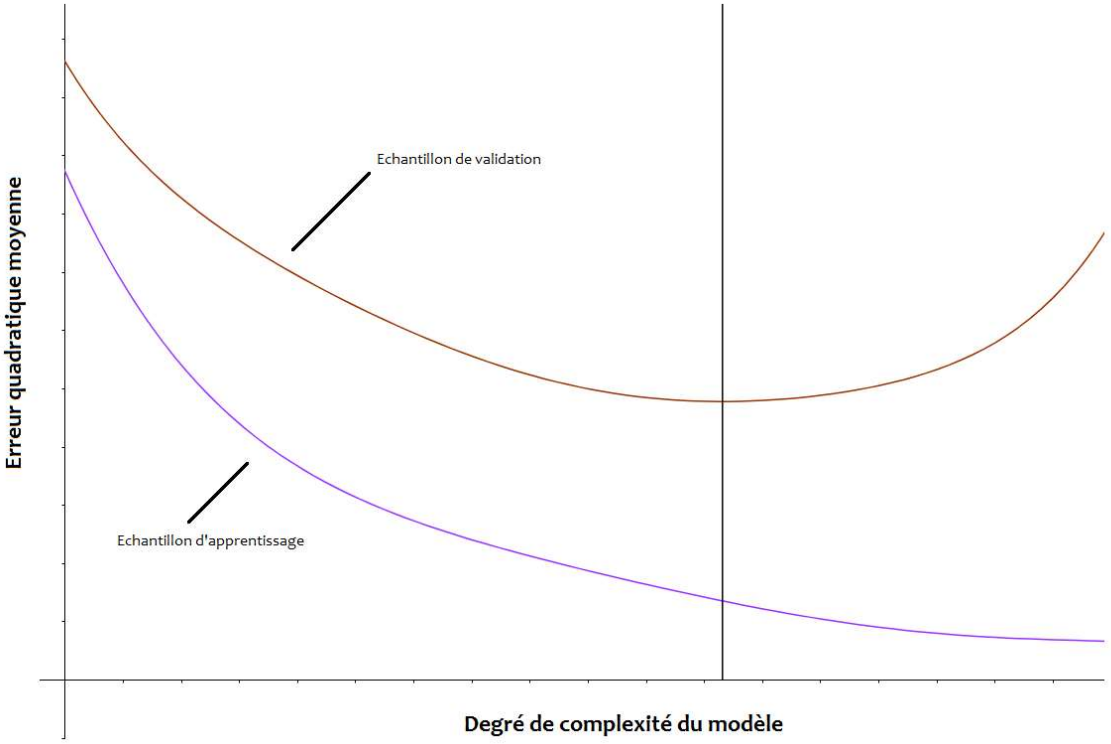
Une fois les paramètres optimisés, le modèle se sert de l'échantillon d'apprentissage pour effectuer son étape de calibration. Celle-ci dépend du modèle, et peut se traduire par l'évaluation de coefficients nécessaires à la construction d'une fonction réponse, ou encore à une segmentation du portefeuille.

Enfin, l'échantillon test est un échantillon qui ne sera utilisé que pour l'évaluation des prédictions. Il n'intervient donc pas dans la construction du modèle. Cet échantillon permet de créer artificiellement une situation nouvelle, similaire à celles observées sur l'échantillon d'apprentissage et de validation, afin d'évaluer la qualité du modèle dans un cadre différent de celui qui a servi à sa construction.

Il est courant de retrouver l'échantillon de validation et l'échantillon test regroupés en un seul et unique échantillon. En effet, deux échantillons suffisent pour la calibration et l'optimisation d'un modèle. Néanmoins, un troisième échantillon est nécessaire si l'on souhaite procéder à une comparaison de plusieurs modèles. Cela permet d'avoir un périmètre commun et surtout nouveau pour l'évaluation des performances de chaque modèle.

En ce qui concerne l'optimisation des paramètres, l'importance de l'échantillon de validation est illustrée dans la figure ci-dessous :

Fig. 1 : Ajustement des paramètres du modèle



Cet exemple est tiré d'une application rapide du Boosting, qui est étudié plus loin dans ce mémoire. Lors de l'ajustement de certains paramètres, ici le nombre d'arbres utilisés, il arrive que le modèle colle de plus en plus aux données et rend ainsi l'erreur sur l'échantillon d'apprentissage peu pertinente pour évaluer la qualité du modèle. L'idée est donc d'ajuster les paramètres en optimisant également l'erreur sur l'échantillon de validation.

Enfin, on dira qu'un modèle est meilleur qu'un autre si son pouvoir prédictif est supérieur, ce qui se traduit par le modèle ayant la plus petite erreur quadratique moyenne sur l'échantillon test.

II. PREPARATION DES DONNEES

II.a. Présentation des données

Le jeu de données disponible est constitué des données de 7 assureurs, chacun utilisant le réseau A. Ces données ont été rassemblées entre le quatrième trimestre de 2014 (T4 2014) et le premier trimestre de 2017 (T1 2017).

Les données du T4 2014 et du T1 2017 sont écartées car certains assureurs n'ont pas communiqué leurs données sur ces trimestres. Il a été jugé plus prudent de ne pas les

prendre en compte lors de l'étude afin d'éviter de biaiser l'estimation. La base de données allant du T1 2015 au T4 2016 comporte ainsi 1 603 690 lignes.

L'agrégation de plusieurs bases provenant d'assureurs différents nous permet de disposer d'un volume d'observations important, ce qui a l'avantage de limiter les problèmes de segmentations trop fines du portefeuille. Cependant, il a été nécessaire de supprimer certaines variables, dans un souci d'harmonisation de la base agrégée.

Suite à la suppression des variables spécifiques à chaque assureur, voici les variables restantes de la base de données :

Tab. 1. Liste des variables communes à chaque assureur

typeSinistre	11 modalités spécifiant la nature du sinistre (cf. Annexe)
kilometrage	Kilométrage du véhicule
ageVehicule	L'âge du véhicule au moment du sinistre
issue	REP (réparation) ou VEI (véhicule économiquement irréparable)
tauxHoraire1	Coût horaire du garage pour une opération courante
tauxHoraire2	Coût horaire du garage pour une opération complexe
tauxHoraire3	Coût horaire du garage pour une opération de haute technicité
tauxHoraireP	Coût horaire du garage pour refaire la peinture
assureur	Assureur prenant en charge le sinistre
mois	Mois de survenance du sinistre
trimestre	Trimestre de survenance du sinistre
marque	Marque du véhicule
cdpostexp	Code postal du lieu de survenance du sinistre
département	Département du lieu de survenance du sinistre
coutREP	Montant d'expertise en cas de réparation
coutVEI	Montant d'expertise pour un véhicule économiquement irréparable
cible	Sinistre expertisé par le réseau d'experts A (Oui/Non)

Lors d'un sinistre, si le montant des réparations est supérieur à la valeur du véhicule, la procédure VEI (véhicule économique irréparable) s'enclenche. L'assureur est alors tenu de proposer le rachat du véhicule. L'assuré peut accepter, auquel cas il est indemnisé de la VRADE (valeur de remplacement à dire d'expert), ou refuser et les travaux de réparation devront être suivis et seront indemnisés dans la limite de la VRADE. L'objectif de cette procédure est de limiter le trafic de cartes grises et renforcer la sécurité routière en empêchant des véhicules dangereux de circuler.

Il est à noter que les variables disponibles sont assez peu nombreuses par rapport à une base de données d'assurance automobile traditionnelle.

Par exemple, on constate l'absence :

- D'informations concernant le conducteur : les données concernant l'assuré sont totalement absentes, que ce soit son âge, son niveau de bonus-malus, le type d'assurance souscrite, son lieu de domicile, sa situation familiale, son expérience en tant que conducteur, sa situation professionnelle...
- Des caractéristiques du véhicule : le type du véhicule (voiture, camion, moto, 4x4 entre autres), le carburant utilisé, la gamme de prix (luxueuse ou non) …
- De détails plus spécifiques concernant l'accident : la gravité du sinistre, l'emplacement du choc, la météo au moment du sinistre …

Il est vrai que certaines variables comme le type du sinistre et la marque du véhicule renseignent une partie de ces informations. Néanmoins, il est probable que l'information utile soit le coût du véhicule neuf et, au sein d'une même marque, les gammes de prix peuvent être très variables.

Le type du sinistre renseigne une caractéristique générale du sinistre mais le manque d'informations détaillées peut affecter la précision des prédictions.

II.a.1. Contexte des données

Il est important de noter que la demande de données a été faite dans le cadre d'une problématique particulière.
D'une part, il a fallu tenir compte de la difficulté de certains assureurs à effectuer une extraction trimestrielle des données demandées, ce qui a eu pour conséquence de réduire le nombre de variables finales.
D'autre part, l'origine de la demande de données concerne l'évaluation de l'écart de performance entre les deux réseaux d'experts. Ainsi, certaines variables qui auraient pu être demandées ont été écartées de l'étude, car elles n'ont pas été jugées pertinentes pour répondre à cette problématique. On note par exemple la pièce de moteur remplacée et l'expertise à distance. En effet, l'expert a une forte influence sur ces deux points, et intégrer ces deux variables reviendrait à les retirer de la mesure de l'écart de performance.

Ce contexte aura donc un impact sur le bon déroulement du deuxième objectif de ce mémoire, à savoir l'ajustement de modèles prédictifs. En effet, cette problématique est propre à l'écriture de ce mémoire, et a été ajoutée compte tenu de l'intérêt que celle-ci constitue dans les enjeux actuels de l'assurance IARD.

II.a.2. Création de variables

II.a.2.1. La variable à expliquer : montant de l'expertise

Afin de mener à bien cette étude, il est nécessaire de créer la variable à expliquer faisant référence au montant de l'expertise, que l'on nomme montExp. Sa valeur est égale à celle de coutREP si le montant d'expertise est celui de la réparation du véhicule, ou égale à celle de coutVEI si la procédure VEI a été acceptée par l'assuré. Il est ainsi inutile de garder ces deux variables, car le montant d'expertise et l'issue suffisent à couvrir leurs informations.

II.a.2.2. La densité

Il semblerait qu'une variable renseignant si le sinistre a eu lieu en zone rurale ou en ville soit utile dans la modélisation. Cependant, il semble difficile d'accéder à cette information en l'état actuel des choses. Le code postal est mal renseigné, ce qui rend difficile son exploitation, et l'échelle du département ne paraît pas adaptée à l'information recherchée.

Néanmoins, Il est possible de créer une variable renseignant la densité de population du département d'occurrence du sinistre. Dans ce sens, on relie la variable département à la densité de population associée. Pour cela, on utilise la population par département au 1er Janvier 2016 fournie par l'INSEE (INSEE, 2017).

L'avantage de la variable ainsi construite est qu'elle permet aux différents modèles de relier les différents départements entre eux à travers la mesure de leurs densités. On verra par la suite que le format numérique de cette variable possède un atout non négligeable en comparaison au département.

Il se peut que la variable densité se révèle moins utile que le département, ou autres variables à disposition. Cela ne pose pas de problèmes particuliers, puisque chaque modèle utilisé possède une phase de sélection de variables. Ainsi, si la densité n'est pas suffisamment pertinente, elle sera automatiquement écartée. Il n'y a donc pas de réel risque a ajouté de l'information de cette façon.

II.b. Retraitement des variables

La phase de retraitement des données est assez particulière, dans le sens où il n'existe pas de méthode universelle ou de théorie assurant une efficacité minimum. Cette partie relève donc majoritairement de l'expérience de l'expert et de sa compréhension de la structure des données.

Il va notamment être nécessaire d'arbitrer les valeurs manquantes et les valeurs mal renseignées.

Lorsqu'une modalité n'est pas remplie, il est possible d'ignorer la ligne ou de lui attribuer une valeur par défaut. Pour la seconde option, on peut choisir de lui attribuer la moyenne, une valeur aléatoire de la loi de probabilité empirique de la variable, ou toutes autres statistiques qui se révèleraient pertinentes.
Lorsqu'une modalité est mal remplie, il est nécessaire d'imaginer un retraitement automatique, en s'assurant qu'il n'y ait pas d'anomalies générées par ce biais.

Dans notre cas, il a été choisi de supprimer les lignes lorsqu'une information importante vient à manquer, et de retraiter prudemment les informations mal renseignées. Par prudemment, il est entendu un remplacement lorsque l'interprétation de la valeur à modifier est évidente, et une suppression de la ligne associée lorsque ce n'est pas le cas. Lorsque le nombre de lignes à supprimer par cette méthode est trop important, il a été décidé d'attribuer une valeur par défaut, comme c'est le cas pour le taux horaire.

Cette méthode a été choisie afin d'assurer une bonne maîtrise de l'information véhiculée par les variables. Etant donné la taille de la base de données, 1 603 690 lignes, la méthode est jugée raisonnable sachant qu'elle ne supprime que 9 % des observations.

Il faut être conscient du biais instauré par un retraitement. Modifier une valeur revient à supposer la manière dont un évènement s'est déroulé et supprimer une ligne revient à supposer que l'évènement n'a pas eu lieu. Dans les deux cas, le retraitement altère la réalité et influe sur l'estimation. Néanmoins, il est nécessaire d'effectuer cette étape sans quoi la modélisation ne serait pas possible. « Entre deux maux, il faut choisir le moindre »[1], et ici la méthode utilisée a été jugée comme telle.

Pour le retraitement, on commence par regarder les statistiques de base de chaque variable pour trouver un angle d'approche :

[1] Aristote, Ethique à Nicomaque, Livre II, 9.4.

Fig. 2 : Vue d'ensemble des données

```
  cdpostexp              marque        kilometrage                    cible         typeSinistre      ageVehicule
       :  25106   RENAULT   :235910   Min.   :      0   CABESTANY        :     1   ACT    :1081993   Min.   :-13.155
92600  :  11757   PEUGEOT   :209968   1st Qu.:  42118   N                :611169   AUT    : 138684   1st Qu.:  3.357
44400  :   9879   CITROEN   :151095   Median :  92348   O                :992518   ACP    : 137363   Median :  6.418
13730  :   9866   VOLKSWAGEN: 91496   Mean   : 106888   ROISSY EN FRANCE :     1   APC    : 132215   Mean   :  7.389
30900  :   8805   FORD      : 60313   3rd Qu.: 154196   ST VICTORET      :     1   VLP    :  31531   3rd Qu.: 10.330
21000  :   7416   AUDI      : 58803   Max.   :9999999                              GRE    :  22063   Max.   :116.988
(Other):1530861   (Other)   :796105   NA's   :29465                                (Other):  59841   NA's   :3453
issue          tauxHoraire1        tauxHoraire2      tauxHoraire3      tauxHoraireP       TRIMESTRE            ASSUREUR
   :      1   Min.   :     0.00   Min.   :   0.00   Min.   :   0.00   Min.   :   0.00   T4 2015:224732   Assureur_1:545773
   :  10361   1st Qu.:    42.20   1st Qu.:  41.00   1st Qu.:   0.00   1st Qu.:  43.00   T1 2015:211623   Assureur_2:181636
REP:1322889   Median :    46.00   Median :  45.89   Median :   0.00   Median :  46.70   T1 2016:207929   Assureur_3:233845
RNP:  26374   Mean   :    48.15   Mean   :  43.66   Mean   :  28.05   Mean   :  47.04   T4 2016:207658   Assureur_4: 34989
VEI: 244065   3rd Qu.:    56.00   3rd Qu.:  56.13   3rd Qu.:  51.00   3rd Qu.:  57.65   T2 2016:193854   Assureur_5: 60783
              Max.   :192699.00   Max.   :1938.35   Max.   :9999.00   Max.   :3000.00   T2 2015:188900   Assureur_6:369890
                                                                                        (Other):368994   Assureur_7:176774
     Mois          département         montExp
Min.   :201501   13      :  57426   Min.   : -1048.0
1st Qu.:201507   59      :  48593   1st Qu.:   661.4
Median :201512   33      :  43169   Median :  1149.1
Mean   :201556   77      :  42110   Mean   :  1817.6
3rd Qu.:201606   6       :  41715   3rd Qu.:  2083.3
Max.   :201612   (Other):1365394   Max.   :340000.0
                 NA's    :   5283   NA's   :29050
```

Etant donné la taille de la base de données, il n'est pas étonnant de trouver autant d'anomalies. Le nombre de valeurs non renseignées parmi toutes les variables étant relativement faible, elles sont supprimées :

Fig. 3 : Suppression des lignes ayant des valeurs manquantes

```
Valeurs non renseignées de kilometrage :  29465 lignes à supprimer
Valeurs non renseignées de ageVehicule :  3408 lignes à supprimer
Valeurs non renseignées de département :  5077 lignes à supprimer
Valeurs non renseignées de montExp :  27431 lignes à supprimer
```

Cette mesure concerne donc les variables kilométrage, ageVéhicule, département et montExp. Au total, 65 381 observations comportant une variable non renseignée seront supprimées.
Désormais, le retraitement des données est effectué variable par variable.

II.b.1. Le code postal

Actuellement, le code postal du lieu du sinistre possède un grand nombre de modalités, 12 234 pour être exact. Il est intéressant de se poser la question concernant la pertinence de cette variable. En effet, on relève deux problèmes associés à cette variable :

- Le code postal n'est pas unique à la commune. Il en existe 6329 différents d'après la base officielle des codes postaux (LA POSTE, 2017) communiquée par le gouvernement français. Les communes n'ayant pas de bureau distributeur de courriers sont rattachées à une autre qui en possède un. Cela pose problème lorsqu'un petit village est rattaché à une ville moyenne. En effet, l'information pertinente semble être la distinction campagne / ville du lieu du sinistre. Cette information n'est donc pas systématiquement véhiculée par le code postal.

- Le code postal est mal renseigné dans la base de données. Celui-ci possède 12 234 modalités différentes. Or, il n'en existe que 6329 uniques. Après vérification, près d'un tiers des observations de la base de données ont un code postal non conforme.

Le simple fait du mauvais renseignement de la variable empêche sa bonne prise en compte, et son intérêt limité n'encourage pas un retraitement extensif de la variable. C'est pourquoi il a été décidé de l'écarter de l'étude.

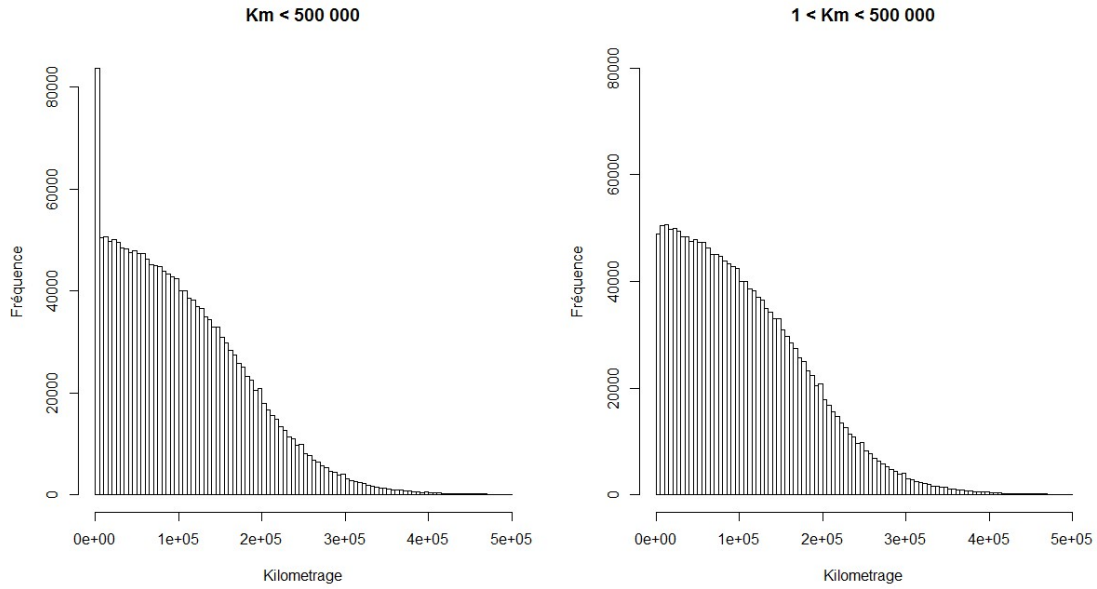
II.b.2. Le kilométrage

La variable kilométrage correspond au kilométrage du véhicule au moment du sinistre.

La valeur maximale de cette variable, 9 999 999 km, suggère de retraiter les lignes dont le kilométrage est trop élevé. On va alors poser un seuil arbitraire assez large, de 500 000 km, afin de retirer les valeurs mal renseignées. 2 319 observations sont concernées et supprimées en conséquence.

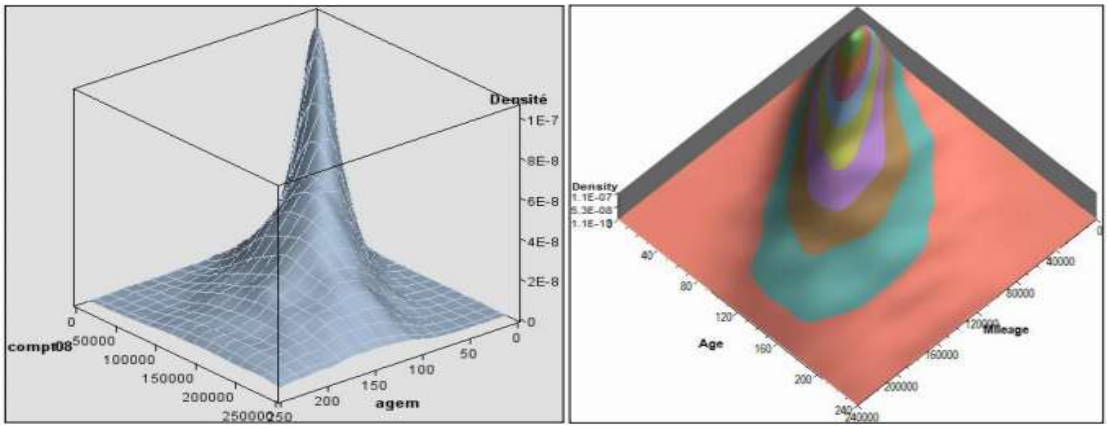
En traçant l'histogramme du kilométrage, on se rend compte d'une autre anomalie :

Fig. 4 : Histogrammes du kilométrage



On constate une concentration importante de sinistres pour des véhicules ayant un kilométrage nul. En supprimant ces observations (26 482 lignes), la nouvelle distribution du kilométrage semble plus harmonieuse. Pour justifier ce retraitement, on s'appuie de la figure suivante :

Fig. 5 : Densités bivariées (vue de face et du dessus) du parc automobile français en 2008



Ce graphique est tiré de la thèse de KOLLI Z. (KOLLI, 2012). On observe ainsi que la distribution des véhicules par kilométrage (figure 5) est sensiblement similaire à la distribution harmonisée des véhicules sinistrés par kilométrage (figure 4). En d'autres termes, plus le nombre de véhicules est important dans une tranche de kilométrage, plus la sinistralité est importante, ce qui semble cohérent.

II.b.3. L'issue

La variable issue indique si le véhicule est en réparation ou économiquement irréparable. De ce fait, on s'attend à avoir uniquement deux facteurs pour cette variable, REP et VEI. Il convient ainsi de supprimer les issues qui divergent de ces deux possibilités, soit 5 726 observations.

II.b.4. La variable cible

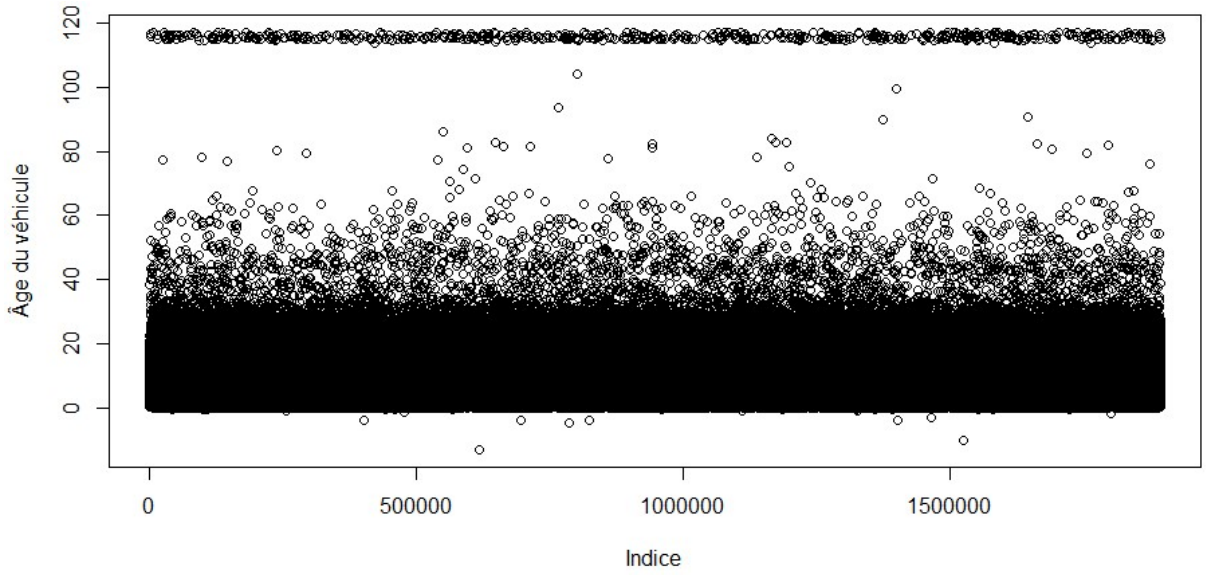
La variable cible indique si le sinistre a été expertisé ou non par le réseau d'experts. On supprime ainsi les observations dont la modalité est autre que « Oui » ou « Non ». Elles sont au nombre de trois.

II.b.5. L'âge du véhicule

L'âge du véhicule renseigne l'ancienneté du véhicule à la date du sinistre. Lorsque l'on regarde la vue d'ensemble des données, on s'aperçoit que certains véhicules ont une ancienneté négative. Ils sont au nombre de 31 et les lignes associées sont supprimées.

Une autre anomalie est mise en évidence lorsqu'on se penche sur l'observation de l'âge du véhicule ligne par ligne :

Fig. 6 : Âge du véhicule en fonction de l'indice de la ligne



Il y a une concentration anormale autour de la valeur « 120 ans » pour la variable âge du véhicule. Un seuil de 80 ans est alors fixé pour pallier à cette anomalie, qui conduit à supprimer 450 lignes supplémentaires.

II.b.6. Les taux horaires

Les variables taux horaires correspondent aux tarifs du garage, et sont divisées en 3 selon le niveau de complexité de la réparation du véhicule. Le taux horaire 1 correspond au taux de facturation pour les opérations simples, telles que la dépose et les contrôles de bases. Le taux horaire 2 correspond aux opérations plus complexes qui demandent une compétence technique spécifique ainsi que du matériel. On y trouve notamment les changements de pare-brise, et les redressages de carrosserie. Enfin, le taux horaire 3 est le taux de facturation utilisé pour les opérations de haute technicité, telles que le redressage des éléments de structure du véhicule et le contrôle de la géométrie des trains roulants. Le taux horaire P correspond quant à lui au taux de facturation lors d'une opération de peinture sur le véhicule.

Suite à l'analyse des statistiques de base de ces variables, plusieurs points retiennent l'attention :

- Présence de taux horaire très élevé, de l'ordre de 9 000 € ;
- 5,5 % des observations n'ont aucun taux horaire de renseigné, 3,3 % pour les sinistres en réparation ;
- 50 % des observations ont les 4 taux horaires de renseignés ;
- 30 % des observations ont les 4 taux horaires strictement identiques ;
- 2,5 % des observations ont des taux horaires anormalement faibles (<15 €) ;
- 12,8 % des sinistres VEI ont au moins un taux horaire de renseigné.

Le dernier point fait écho à un problème de structure dans la corrélation de nos variables : les taux horaires ne sont pas censés être renseignés lorsque le sinistre est VEI, puisqu'aucune réparation n'a lieu. Cela signifie que ces 4 variables sont par nature complètement corrélées avec la variable issue. D'autre part, les points 3 et 4 mettent en évidence un problème de renseignement de la variable : la majorité des sinistres étant des sinistres à faible coût, il est étonnant de retrouver 50 % des sinistres avec un taux horaire 3 utilisé et il est d'autant plus étonnant de retrouver les 4 taux horaires égaux pour un bon nombre de lignes.

Compte tenu de ces informations, il a été décidé de créer une variable qualitative « taux horaire » unique pour remplacer l'issue et les 4 taux horaires. Cette idée est motivée par la constatation que les taux horaires ne sont pas renseignés lorsque l'opération est effectuée sur le véhicule, ce qui rend nul leur distinction. Du moins, cette dernière ne semble pas suffisamment systématique pour qu'elle soit prise en compte dans un modèle.

On divise ainsi le retraitement en deux :

Pour les sinistres VEI :

On assigne la valeur par défaut « NUL ».

Pour les sinistres en réparation :

Les taux horaires supérieurs à 300 € étant peu nombreux, les 47 observations correspondantes sont simplement supprimées. Ensuite, on construit la variable taux horaire, ayant pour valeur la moyenne des valeurs non nulles parmi les 4 taux horaires. On remarque ensuite que 30 % des observations ont un taux horaire faible ou nul. Ce type d'anomalies requiert une méthode différente de retraitement que la simple suppression des observations. Ainsi, lorsque le taux horaire est inférieur à 15 €, on affecte une valeur par défaut égale à la moyenne des taux horaires supérieurs à 15 €.
Enfin, pour que le type de la variable soit cohérent pour tous les sinistres, la variable est découpée en trois modalités en fonction des quantiles : « Faible », « Moyen » et « Fort ».

Ainsi, le taux horaire est pourvu de 4 classes : « NUL » pour les sinistres VEI, « Faible », « Moyen » et « Fort » pour les sinistres en réparation. Les variables tauxHoraire1, 2, 3, P et issue sont supprimées.

Ce retraitement a l'avantage de supprimer les variables fortement corrélées pour les remplacer par une seule variable, et la perte d'informations induite par ce procédé est considérée comme raisonnable. Il règle également tous les points ci-dessus qui posaient problème jusqu'alors.

II.b.7. Le montant d'expertise

La variable montExp est la variable renseignant le coût effectif des réparations, ou dans le cas VEI, la valeur du véhicule avant sinistre. Il n'est donc pas concevable d'avoir des valeurs négatives ou nulles pour cette variable. C'est pourquoi les observations qui en ont sont supprimées (6 090 lignes).

II.b.8. Le département

Le département renseigné dans la base de données correspond au département de survenance du sinistre. Parmi les différentes modalités de la variable département, un certain nombre sont vides ou mal renseignées, en faisant référence à des départements qui n'existent pas. Dans ce deuxième cas, on utilise le code postal dont on extrait les deux premiers chiffres (les trois premiers si c'est un DOM-TOM) afin de retrouver le département du lieu de sinistre. De manière générale, les départements sont uniformisés selon une unique nomenclature, afin d'éviter d'avoir plusieurs facteurs pour le même lieu de sinistre.

29 049 lignes sont supprimées pour les départements vides et 103 lignes dont le département est mal renseigné et non retraitable avec le code postal sont également supprimées.

II.b.9. La marque

La variable marque renseigne la marque du véhicule sinistré. Dans la base de données, elle comporte 2698 niveaux ce qui suggère que son renseignement n'est pas uniforme. En effet, parmi ces niveaux on remarque par exemple la présence de « CITROEN », « CITROEN C3 », « CITOREN », « CITR », « CITROEN AX », « CITREN » et bien d'autres. L'objectif est de regrouper toutes les modalités qui font référence à la même marque en une seule.

La première méthode imaginée pour ce retraitement est la suivante : dans un premier temps, toutes les marques sont mises en majuscule et les espaces inutiles sont supprimés, afin d'uniformiser les données. Puis, on couple la variable avec une liste de marques préalablement constituée, en associant les premiers caractères entre eux. On obtient ainsi une variable marque ne possédant plus que 156 facteurs. On note que 13 marques représentent 85 % des observations de la base de données. Le problème de cette méthode est que certaines lignes sont exclues, alors qu'elles auraient pu être déduites, comme par exemple « PGT » signifiant « PEUGEOT ». Il semble néanmoins que la majorité des observations éludées par cette méthode est liée à des marques qui ne

sont pas retraitables. 8 895 observations sont écartées en utilisant cette méthode de retraitement, correspondant à un peu moins de 0,5% de la base de données.

Une méthode plus élaborée consiste à utiliser la distance de Levenshtein (cf le sujet d'étude sur la distance de Levenshtein (François, 2014)). Elle propose une approche pour quantifier la distance entre deux chaînes de caractères A et B en comptant une distance de 1 pour chaque opération élémentaire.
Les opérations élémentaires sont au nombre de 3 :

- Le remplacement d'un caractère ;
- La suppression d'un caractère ;
- L'ajout d'un caractère.

Le but est de minimiser le nombre d'opérations élémentaires nécessaires pour passer de la chaîne A à la chaîne B. Cette méthode a le mérite d'attribuer une valeur à tous les niveaux de « marque », et permet de fixer une valeur par défaut pour les marques mal renseignées. Cependant, cette méthode induit un bon nombre d'erreurs. Par exemple, l'algorithme attribue à la valeur « ASTA » la marque « AMERICAN AUSTIN-BANTAM », dont il est difficile de justifier la pertinence.

La première méthode renvoie une valeur uniquement lorsqu'on est sûr de pouvoir identifier la marque, mais délaisse certaines valeurs qui auraient pu être identifiées. La deuxième méthode amplifie la recherche afin d'attribuer le plus de valeurs possibles, au détriment de la précision. Par extension, cette méthode crée alors des valeurs fausses.

La première méthode est plus prudente et le nombre de lignes écartées par celle-ci est raisonnable. Elle est donc retenue.

II.b.10. Bilan du nombre d'observations supprimées

Le nombre lignes effectivement supprimées n'est pas égale à la somme des observations supprimées suite au retraitement de chaque variable, car certaines d'entre elles comportent plusieurs variables mal renseignées. Au final, 139 545 observations, soit environ 9 % de la base de données, ont été écartées suite au retraitement des variables. Il reste ainsi 1 464 145 lignes dans la base de données.

Pour récapituler, voici le tableau des variables retenues ainsi que leurs types :

Tab. 2. Liste des variables retenues

Variable à expliquer	Variables explicatives	Type des variables
montExp	ageVehicule kilométrage densité	Variables « quantitatives »
	Cible tauxHoraire typeSinistre trimestre mois assureur marque département	Variables « qualitatives »

Ce qui fait 11 variables explicatives pour une variable à expliquer. Elles ne seront pas toutes utilisées pour la modélisation, certaines étant redondantes.

II.c. Biais entre la variable cible et l'assureur

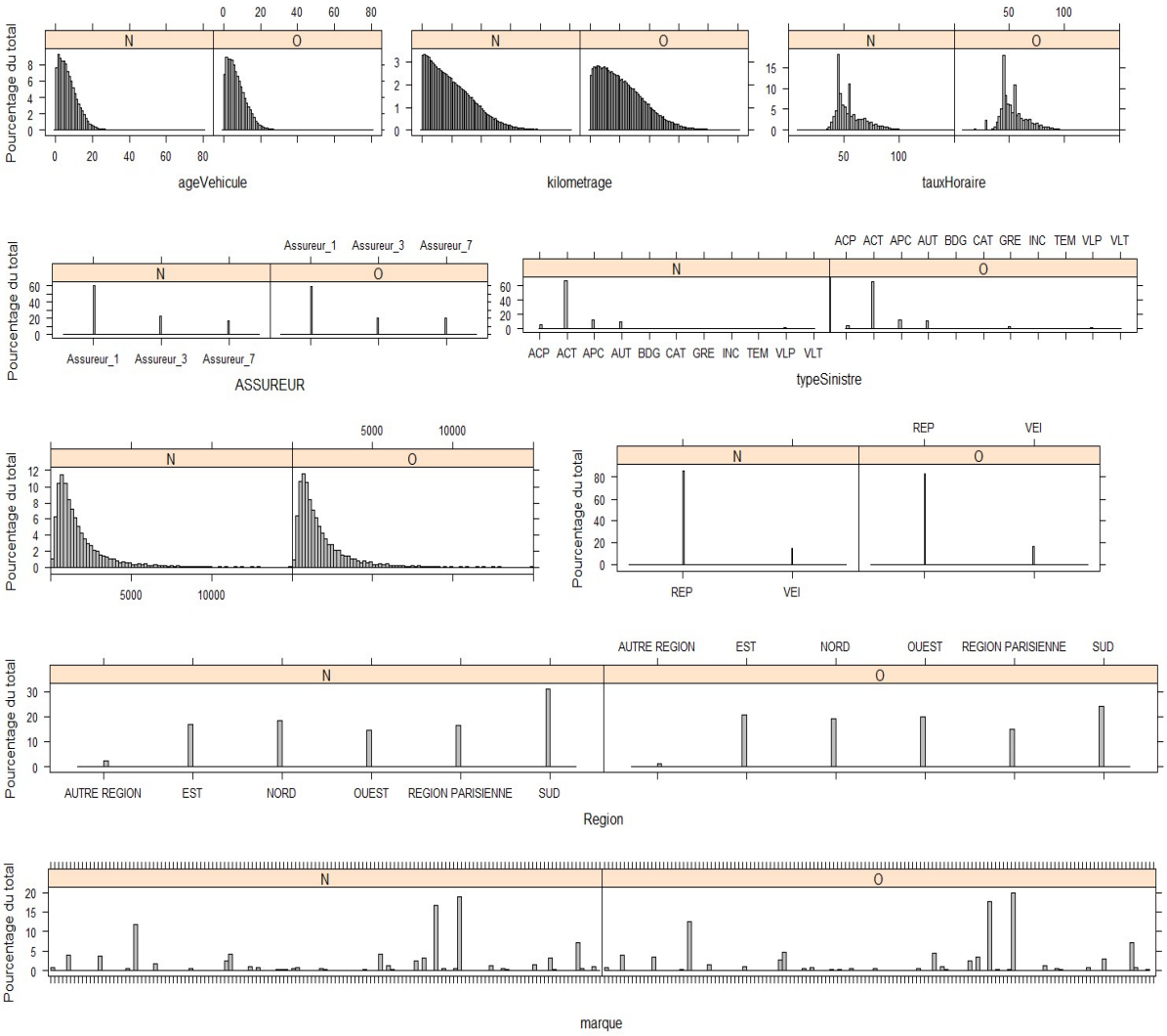
Il est observé que certains assureurs n'utilisent que le réseau A. Or, il semble nécessaire qu'un sinistre puisse être expertisé également par la concurrence pour sa bonne prise en compte.
On pourrait argumenter le fait que toutes observations aideraient à mieux expliquer le comportement de la variable cible. Cependant, ces observations proviennent d'autres assureurs, ce qui leurs attribuent une structure particulière qui risque de perturber l'évaluation de l'écart de performance. Il est donc jugé plus prudent de ne pas les conserver ces données. Le fait que la base de données soit suffisamment grande permet également de motiver ce choix.

Ainsi, on supprime de la base les données relatives aux 4 assureurs concernés, et on ne garde que celles de l'assureur 1, 3 et 7. La base est ainsi réduite à 876 935 observations, ce qui reste suffisant pour l'étude.

II.d. Composition du portefeuille de sinistres par réseau

Il est intéressant de regarder la répartition et les statistiques de bases des variables en fonction du réseau d'experts. L'objectif est double : s'assurer qu'il ne reste pas de valeurs aberrantes nécessitant un retraitement, et avoir une vue d'ensemble des données et des potentielles dissimilarités entre les deux réseaux.

Fig. 7 : Répartition des variables par réseau



Cette figure montre bien qu'a priori, il n'y a pas d'effets de structure dans la répartition des variables par réseau. Cela semble montrer qu'il n'y a pas de biais lors de l'attribution des sinistres par assureur, c'est-à-dire que chacun des réseaux est amené à traiter des sinistres en moyenne similaires.

La différence de montant d'expertise moyen est assez importante pour qu'elle soit soulignée. Le réseau A possède un montant d'expertise de 1 944 € en moyenne, comparé au réseau B qui est en moyenne à 1 987 €.

Il serait hâtif de conclure que le réseau A est plus performant que le réseau B, même après avoir pris en considération que les sinistres traités sont en apparence similaires au regard des variables observées. Il n'est pas exclu que de petites dissimilarités, une fois

cumulées, aient un impact important. C'est pourquoi il est important de vérifier ce résultat par la modélisation, afin d'avoir tous les éléments permettant de justifier un possible écart de performance entre les deux réseaux.

II.e. Liaisons des variables explicatives avec la variable à expliquer

Il est intéressant de regarder le lien qui peut exister entre le montant de l'expertise et le reste des variables explicatives. Il sera ainsi possible de déterminer les variables les plus influentes. Deux outils sont utilisés pour évaluer la robustesse de la liaison entre deux variables : le V de Cramer et le test du χ^2.

Etant donné que le montant de l'expertise est une variable quantitative, il aurait été possible d'utiliser le coefficient de corrélation lorsque la variable explicative est également quantitative. Cependant, la majorité des variables étant qualitatives, on discrétise les variables continues afin de pouvoir appliquer le V de Cramer et n'utiliser ainsi qu'une seule mesure. Le découpage utilisé est basé sur les 3 quantiles, ce qui permet d'avoir 3 classes avec chacune le même effectif.

Le test du χ^2 permet de tester l'indépendance de la variable explicative à la variable montExp. L'hypothèse nulle est alors H_0 : "$Les\ deux\ variables\ sont\ indépendantes$". La statistique du χ^2 est alors calculée, ainsi que sa p-valeur associée. On rappelle que la p-valeur correspond à la probabilité d'obtenir une valeur identique ou plus extrême de cette statistique en supposant que l'hypothèse nulle soit vraie. Ainsi, une faible p-valeur conduit à rejeter l'hypothèse nulle. Il est important de noter qu'une forte p-valeur ne valide pas l'hypothèse nulle, mais ne la rejette pas non plus.

Le V de Cramer quant à lui mesure la force du lien entre deux variables. Il est calculé à partir de la statistique du χ^2 : $V = \sqrt{\frac{\chi^2}{n*[\min(x,y)-1]}}$ avec :

- X^2 la statistique du χ^2
- n le nombre d'observations
- x le nombre de modalités de la première variable
- y le nombre de modalités de la seconde variable

On estime que le pouvoir discriminant de la variable explicative est très bon lorsque le V de Cramer est supérieur à 0,20, bon entre 0,10 et 0,20, correct entre 0,05 et 0,10 et faible en deçà.

Tab. 3. V de Cramer et test du χ^2

variable	V de cramer	p value Khi2
tauxHoraire	0,226	<2.2e-16
typeSinistre	0,103	<2.2e-16
département	0,068	<2.2e-16
ageVehicule	0,066	<2.2e-16
densite	0,052	<2.2e-16
ASSUREUR	0,043	<2.2e-16
marque	0,043	<2.2e-16
kilometrage	0,031	<2.2e-16
TRIMESTRE	0,011	<2.2e-16
cible	0,005	0,001

Premièrement, on constate que les p-valeurs du test du χ^2 sont faibles, ce qui encourage à retenir ces variables pour la modélisation du montant d'expertise. En effet, une faible p-valeur met en évidence une dépendance significative entre la variable à expliquer (le montant d'expertise) et la variable explicative.

D'autre part, le V de Cramer met en avant le pouvoir discriminant de la variable tauxHoraire. Cela s'explique en grande partie par le fait que la nouvelle variable taux horaire a récupéré l'information de l'ancienne variable issue, qui renseignait si le sinistre était VEI ou en réparation. Le coût moyen d'un sinistre VEI étant bien plus important que le coût moyen d'une réparation, il est normal de retrouver la variable tauxHoraire comme la variable la plus discriminante. A l'inverse, les variables assureur, marque, cible, kilométrage et trimestre semblent avoir un pouvoir discriminant plutôt faible.

En particulier, la variable cible possède le V de Cramer le plus faible, ce qui suggère une influence modérée sur le coût final du sinistre. Néanmoins, la p-valeur du test du χ^2 indique une dépendance significative entre la variable cible et la variable d'intérêt. Ces deux résultats ne sont pas en contradiction. Nous savions que l'influence de la variable cible serait de second ordre par rapport aux autres variables, car le contraire n'aurait pas nécessité une étude de performance. Un des objectifs de cette étude est justement d'évaluer l'influence de cette variable une fois que toutes les autres aient été prises en compte. Il est encourageant d'avoir une p-valeur de cet ordre pour la variable cible, et cela justifie l'étude qui a été mise en place pour déterminer son influence sur le coût du sinistre.

II.f. Liaisons entre les variables explicatives

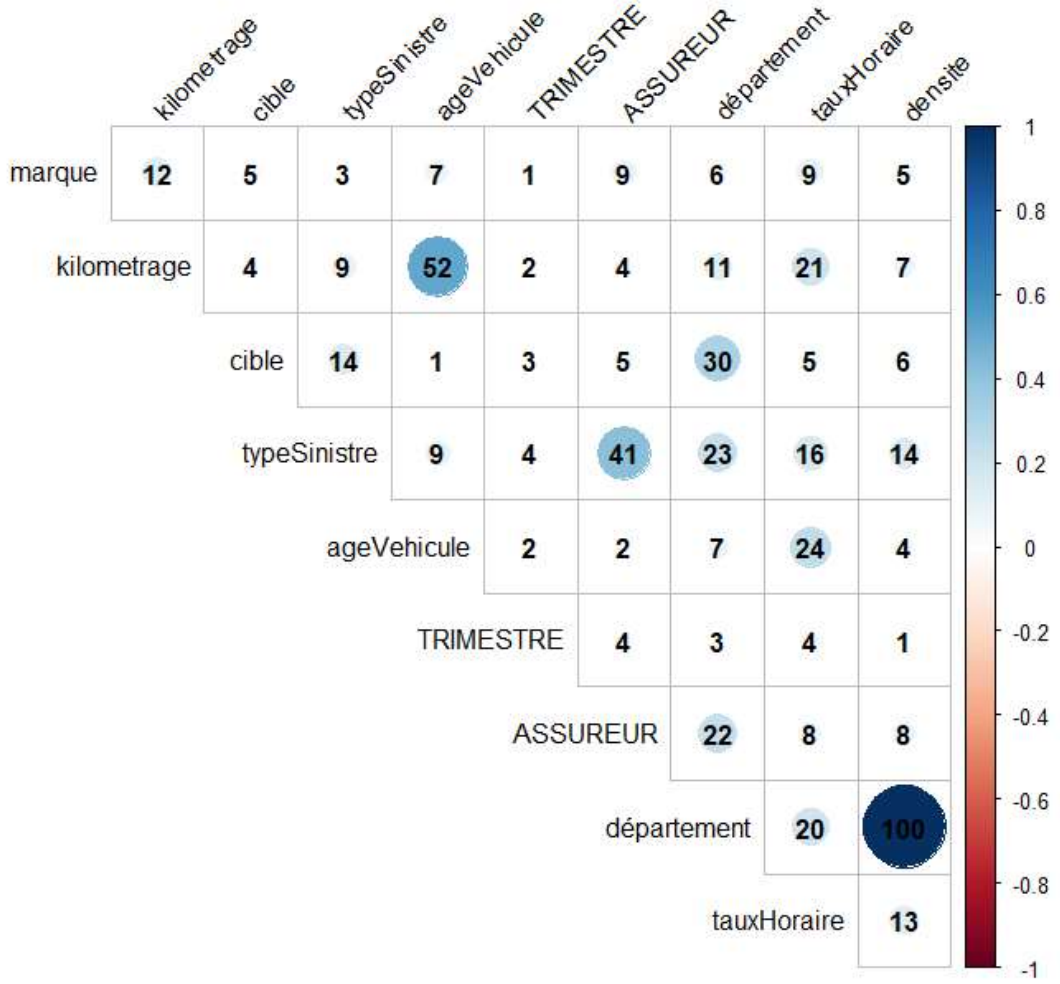
Il est intéressant de regarder les liens entre les différentes variables explicatives, sachant que certains modèles ont besoin pour leurs bons fonctionnements que ceux-ci ne soient pas trop forts. A titre d'exemple, lorsque la corrélation est maximale entre deux variables,

l'estimation des coefficients d'un GLM est en général mauvaise et le résultat de l'estimation est difficilement interprétable.

II.f.1. V de Cramer

De la même manière que précédemment, on calcule le V de Cramer mais ici par couple de variables explicatives.

Tab. 4. V de Cramer des variables explicatives



On observe ainsi une colinéarité totale entre le département et la densité, expliquée par le fait que ces deux variables sont liées par construction. Il sera donc nécessaire de choisir entre l'une et l'autre au moment de la modélisation.

Parmi les variables les plus liées, on note le couple ageVehicule – kilométrage. En effet, plus le véhicule est vieux, plus il a eu l'occasion de rouler et d'augmenter son kilométrage. Vient ensuite le couple type du sinistre – assureur, qui est un peu plus atypique et qui sera détaillé un peu plus loin.

D'autre part, on note une assez forte dépendance entre la variable cible et le département. Cela est sans doute lié au fait que l'activité du réseau d'experts est particulièrement concentrée autour du lieu de son siège.

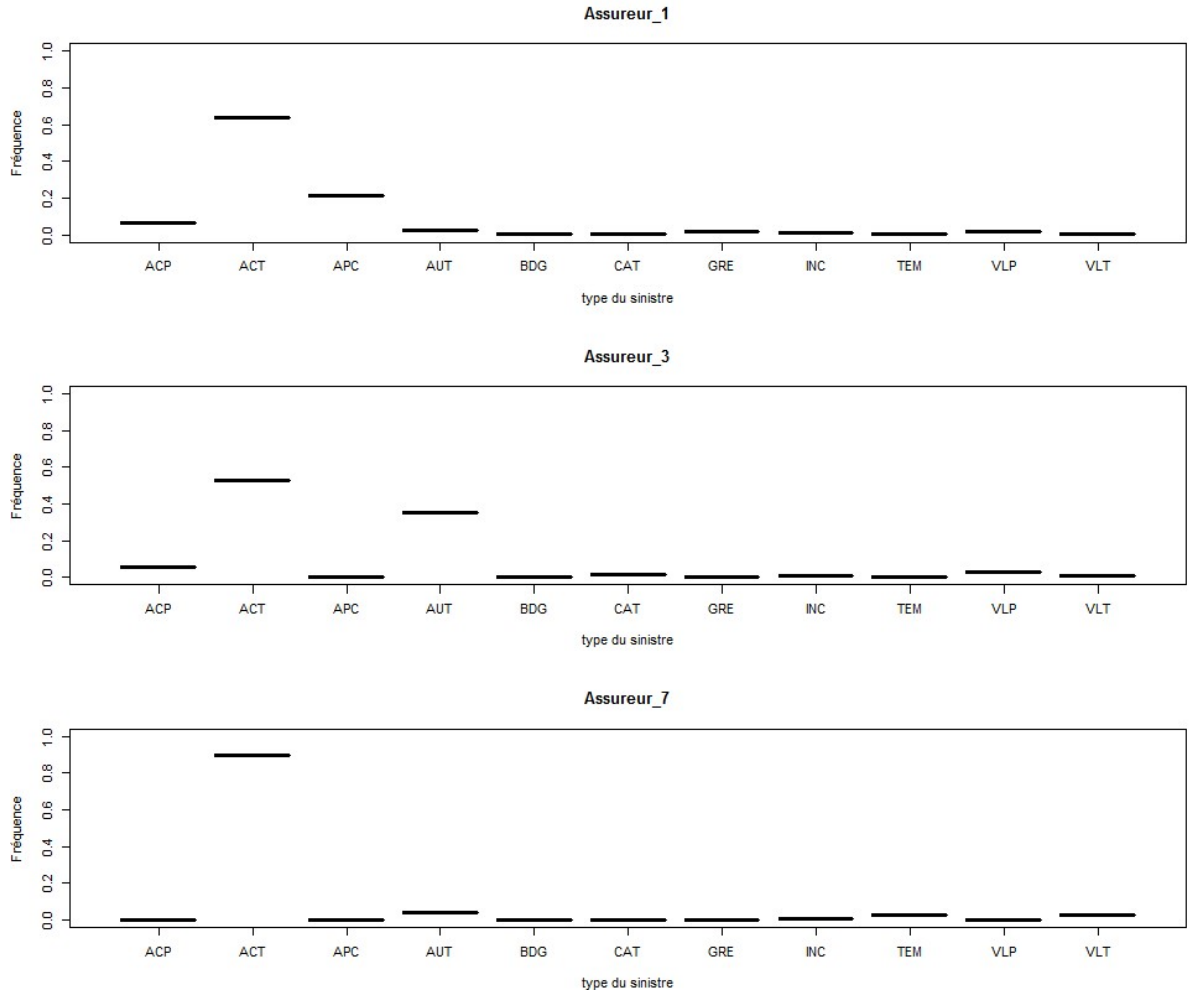
Lorsque le V de Cramer est important entre deux variables, il est probable que l'une d'entre elles tienne son pouvoir discriminant de l'autre variable. Il est alors intéressant de se demander s'il est pertinent de garder les deux variables. Pour le moment, les deux variables sont conservées. Hormis le cas des variables département et densité, nous considérons que les V de Cramer observés n'ont pas de valeurs assez fortes pour exclure d'emblée une variable.

II.f.2. La variable assureur

Le précédent tableau révèle que deux variables ont un lien particulier avec la variable assureur : le taux horaire et le type du sinistre. Le premier s'explique par les taux horaires négociés avec certains garages, propres à chaque assureur. Le deuxième pourrait s'expliquer par les activités propres à chaque assureur, mais il peut tout de même être intéressant de creuser un peu plus.

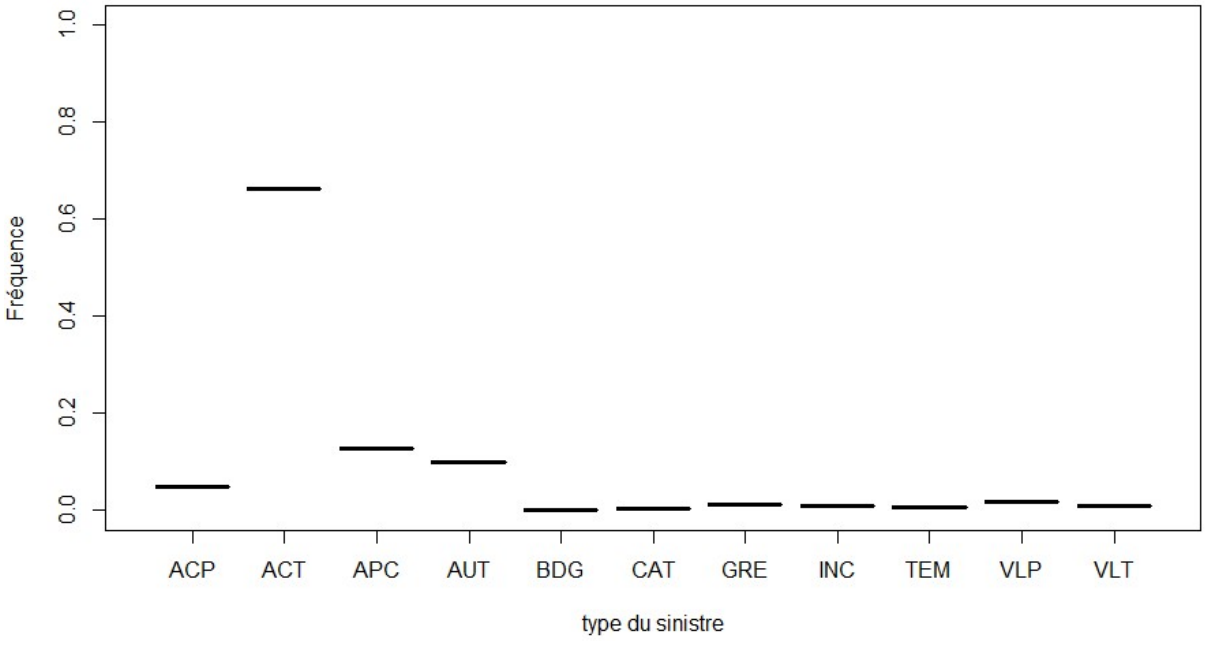
Statistiquement, on considère qu'un lien est fort entre deux variables lorsque la distribution de la première variable en ayant pris connaissance de la seconde change substantiellement. Cela permet d'observer d'un autre point de vue le lien entre le type du sinistre et l'assureur :

Fig. 8 : Distribution du type du sinistre par assureur



Il faut comparer ces 3 distributions à la distribution « globale » de la variable typeSinistre :

Fig. 9 : Distribution de la variable typeSinistre



On peut voir que le renseignement de la variable assureur altère significativement la distribution de la variable type du sinistre, principalement au niveau des modalités ACT (accident avec tiers), APC (accident perte de contrôle) et AUT (autre accident). Inversement, étant donné une valeur pour le type du sinistre, il est possible de la rattacher à un assureur au regard des distributions qui sont relativement distinguables.

Enfin, il faut garder à l'esprit qu'il y a une forte asymétrie dans le volume de données par assureur. En effet, sur 876 935 observations, l'assureur 1 en détient 60 %, l'assureur 3 : 22 % et l'assureur 7 : 18 %. Il a pu être observé dans cette partie que les données par assureur sont assez disparates. Cela peut s'expliquer par le fait que chaque assureur a sa propre politique économique, une implantation géographique différente ainsi qu'une population d'assurés spécifiques qui les amènent à traiter des sinistres de nature différente.

CHAPITRE 2 : MODELISATION

I. MODELE LINEAIRE GENERALISE

Etant donné la problématique de reconstitution du coût moyen par un modèle prédictif, il est relativement classique de modéliser le problème par un modèle linéaire généralisé (GLM : *Generalized Linear Model*). En effet, cette approche a l'avantage d'avoir une implémentation accessible et de présenter des résultats facilement interprétables. Cette méthode est d'autant plus intéressante qu'elle permet d'estimer l'influence de chaque variable dans le coût final, et en particulier celui de la variable cible permettant de répondre au deuxième objectif : l'évaluation de la performance du réseau d'experts.

Ainsi, l'objectif de cette partie est de présenter un modèle simple, efficace, qui servira de comparaison pour les modèles plus complexes. Cette partie s'inspire des travaux relatifs à l'assurance non-vie de M. Denuit et A. Charpentier (Charpentier – Denuit, 2005), dans lequel ils présentent notamment les modèles type GLM.

I.a. Théorie du GLM

Comme son nom l'indique, le GLM est une version généralisée du modèle de régression linéaire classique. Ce modèle a fait son apparition avec John Nelder et Robert Wedderburn en 1972, afin de répondre à la complexification des problèmes statistiques par une modélisation plus flexible que ne l'est la régression linéaire.

L'objectif est d'être capable d'établir une relation entre la variable réponse Y, ici le montant d'expertise, et les variables explicatives $X = (X_1, \dots, X_p)^t$, que sont l'issue, la marque, l'âge du véhicule ... L'idée est de modéliser l'espérance de Y conditionnellement à X, car cette valeur correspond à la meilleure approximation de Y lorsque X est connu. Pour cela, on choisit une loi de distribution pour la variable Y | X parmi celles de la famille exponentielle, caractérisée par une densité de la forme :

$$f(y|\theta) = a(\theta)b(y)\exp\big(\eta(\theta).T(y)\big),$$

où y est la variable aléatoire et θ un paramètre. La famille exponentielle est souvent retrouvée dans sa forme simplifiée, qu'on appelle alors la famille exponentielle naturelle, où η et T sont chacun l'identité :

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(u,\phi)\right) ; y \in S,$$

S est le support de la variable réponse, θ le paramètre naturel et ϕ le paramètre de dispersion. La famille exponentielle englobe la majorité des lois usuelles utilisées dans la modélisation statistique.

Une fois la loi de Y choisie dans la famille exponentielle, nous supposons l'existence d'une relation de la forme :

$$g(E[Y|X]) = \beta X,$$

où β est le paramètre à estimer et g une fonction de lien qu'il faut choisir. Chacune des lois de distribution de la famille exponentielle possède une fonction de lien canonique, vérifiant $g(E[Y_i]) = \theta_i$.

On retrouve la régression linéaire classique en choisissant la loi avec la fonction de lien canonique associée qui est l'identité. Ainsi, on comprend bien l'intérêt d'un GLM par rapport à la régression linéaire, qui permet entre autres de sortir du cadre restrictif de la loi normale qui n'est pas toujours adaptée aux données. Dans notre cas, le montant d'expertise serait mal représenté par une loi normale, notamment, car la loi normale est symétrique et autorise les valeurs négatives.

Une fois le cadre posé, il suffit d'estimer les coefficients β par maximum de vraisemblance, qui en pratique se fait itérativement avec la méthode de Newton-Raphson.

I.b. Paramétrisation du GLM

I.b.1. Prérequis

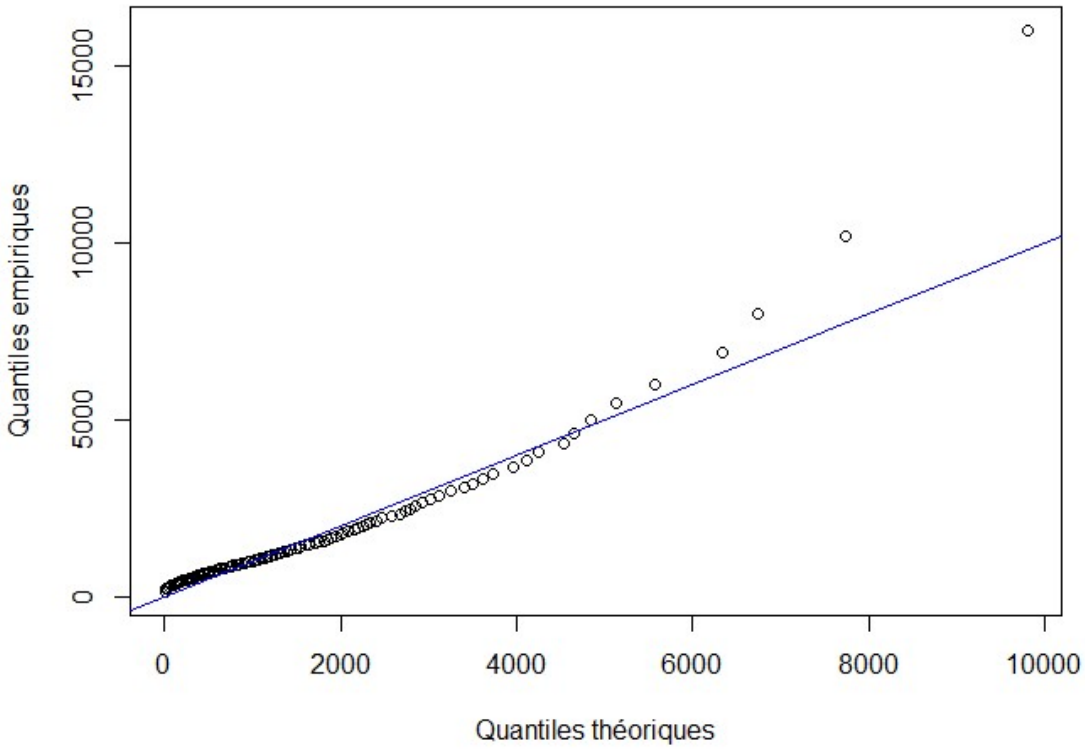
I.b.1.1. Sinistres graves

Lors d'une tarification IARD, une étape courante avant la modélisation est d'écarter les sinistres graves et d'en faire l'étude à part, afin qu'ils ne perturbent pas le modèle en induisant une forte variance. Il est important de garder à l'esprit que cette tarification automobile est particulière dans le sens où le coût du sinistre modélisé correspond au coût de réparation du véhicule. Ainsi les sinistres traités sont majorés par le prix du véhicule neuf (exemple : le sinistre le plus grave observé coûte 250 000 €). On est donc loin de la tarification automobile habituelle, qui propose une garantie responsabilité civile et dont le coût du sinistre peut atteindre plusieurs millions d'euros.

Ceci étant dit, il est intéressant de vérifier si la théorie des valeurs extrêmes s'applique à ce cas. Pour cela, on s'appuie sur le mémoire de R. Boyer Chammard (Boyer Chammard, 2007) qui traite de la modélisation de la sévérité des sinistres. Dans un premier temps, on

trace un diagramme quantile – quantile de la loi exponentielle et de paramètre ajusté à $\frac{1}{\frac{1}{n}\sum_{i=1}^{n} y_i}$. Cela permet d'avoir une première idée concernant l'épaisseur de la queue de distribution du montant d'expertise. En effet, si la courbe est convexe, cela voudra dire que la queue de distribution est moins épaisse que l'exponentielle. Si c'est une droite, alors les coûts suivent une loi exponentielle. Si la courbe est concave, la queue de distribution est plus lourde.

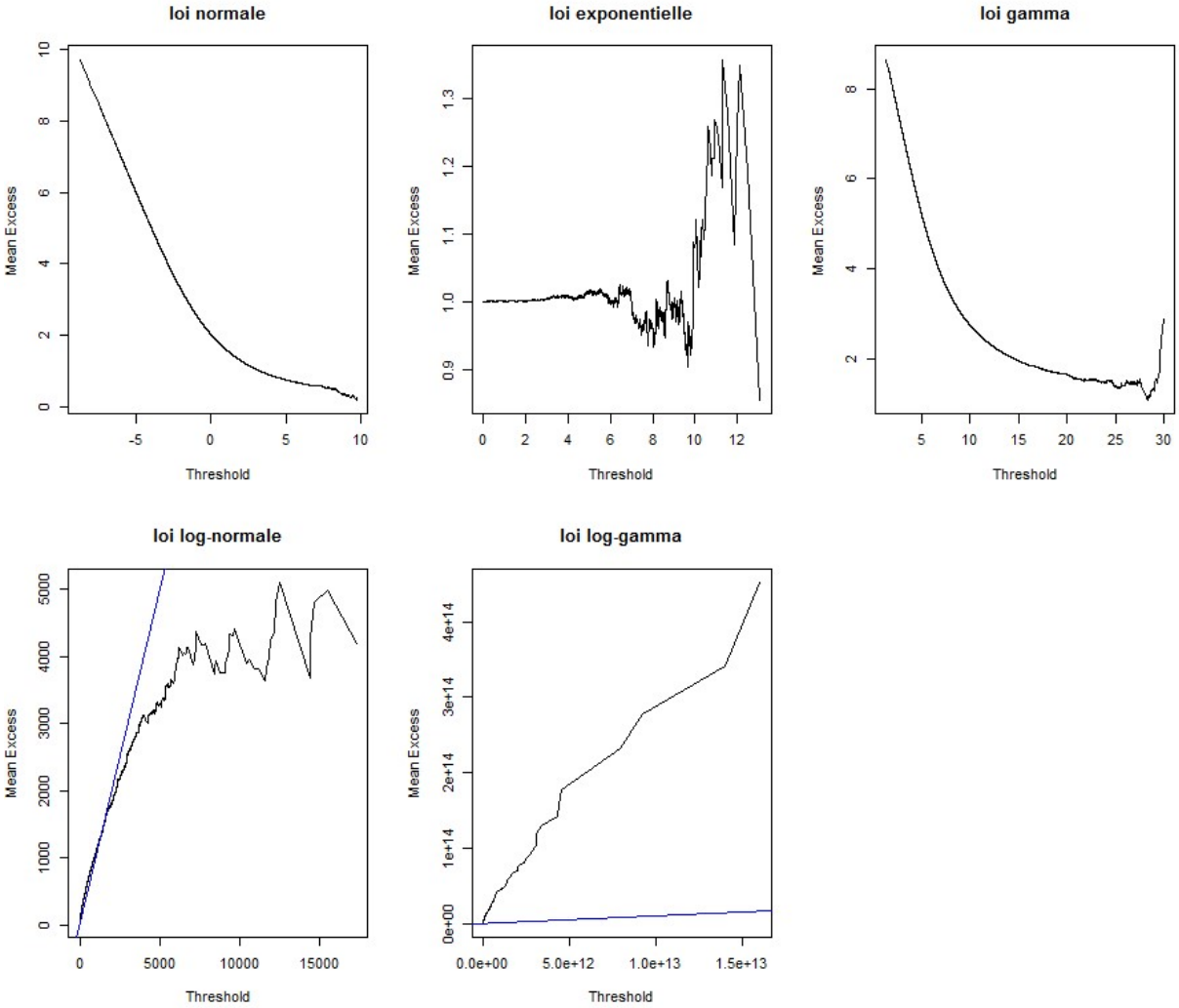
Fig. 10 : QQ-plot du montant d'expertise avec une exponentielle



Le diagramme est dans un premier temps linéaire, et se termine de manière convexe.

Il est intéressant de compléter cette analyse en traçant l'espérance résiduelle $e(s) = E(C - s|C > s)$, où C représente le montant d'expertise. En d'autres termes, cette valeur représente la moyenne des coûts des sinistres excédents le seuil s. L'allure de son évolution donnera une bonne indication sur l'épaisseur de la loi de distribution. Nous avons ainsi tracé l'espérance résiduelle estimée (*mean excess plot*) pour quelques lois usuelles :

Fig. 11 : Mean excess plot des lois usuelles

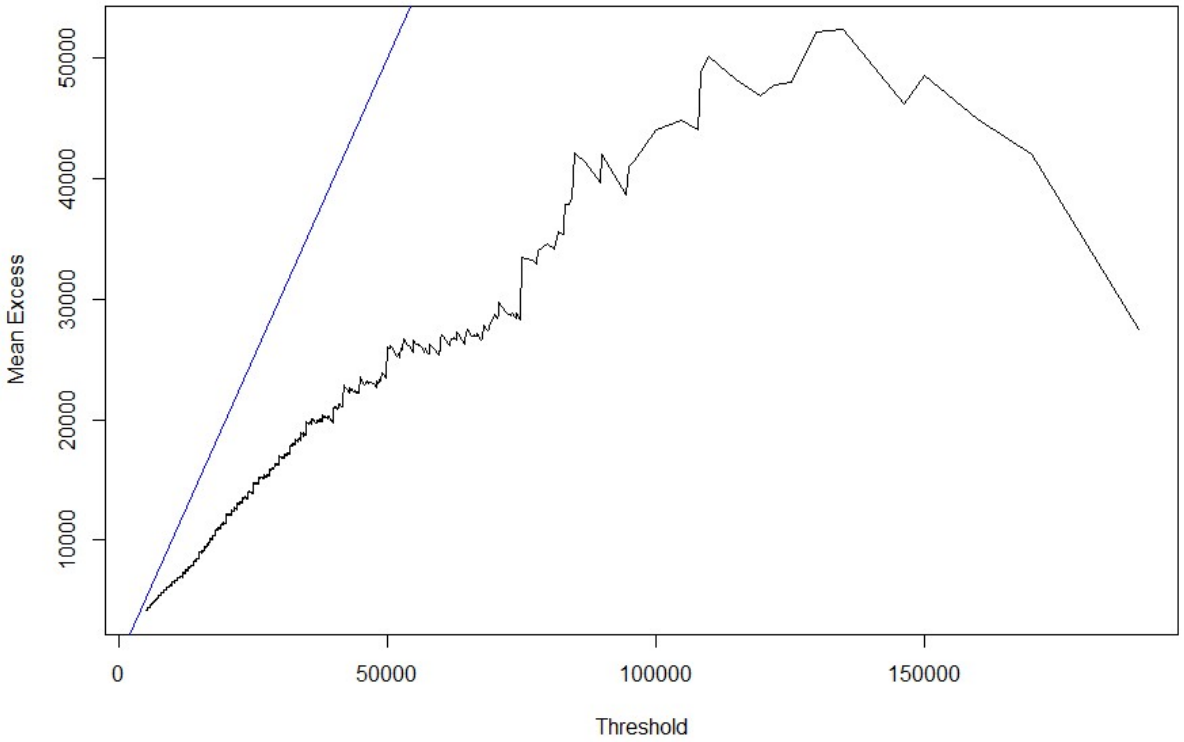


Il est à noter que les fortes variations observées à l'extrémité droite de chaque graphique sont dues aux fluctuations de la moyenne, qui est calculée sur un nombre d'observations de plus en plus faible. Ainsi, l'espérance résiduelle de :

- la loi normale est décroissante et tend vers 0 ;
- la loi exponentielle est constante autour de la valeur 1/λ (où λ est le paramètre de la loi) ;
- la loi gamma est décroissante qui tend vers son paramètre d'intensité ;
- la loi log-normale est croissante et concave ;
- la loi log-gamma est croissante et linéaire.

Comparons maintenant ces graphiques au tracé de l'espérance résiduelle estimée du coût du sinistre.

Fig. 12 : Mean excess plot du montant d'expertise



La croissance est relativement linéaire, jusqu'au point d'inflexion, se situant autour du seuil de 130 000 €, où l'on observe une décroissance importante. Il est à noter que passé ce seuil, il n'y a plus que 10 observations dans la base de données, ce qui induit une forte variance dans l'espérance résiduelle. L'allure de ce graphique présente le plus de similitude avec le graphique de la loi log-normale, qui a une queue de distribution assez épaisse.

Ainsi, le QQ-plot et le Mean Excess Plot montrent que le portefeuille de sinistres est constitué de beaucoup de sinistres « moyennement graves », et qu'ils ont un fort impact sur le poids total du portefeuille. En effet, 5 % des sinistres représentent plus d'un quart du coût total. Cela montre, malgré la faible fréquence de sinistres exceptionnels, que les sinistres graves ont un impact fort sur le coût moyen.

Etant donné toutes ces informations, il semble intéressant de distinguer les sinistres graves dans cette étude. Ce procédé apporte de nombreux atouts statistiquement parlant, comme une stabilité accrue des coefficients estimés. En effet, les sinistres graves ont tendances à fortement perturber les segments de population de faible taille. Prenons en exemple un segment particulier du portefeuille de sinistres : le véhicule est récent, d'une marque peu utilisée, d'un faible kilométrage ; le sinistre a eu lieu en région parisienne (zone à forte densité de population) ; l'accident est arrivé avec perte du contrôle du véhicule (APC), et en résulte que le véhicule est économiquement irréparable. 18 sinistres ont exactement ces propriétés.

Pour que le modèle fonctionne correctement, les coûts de ces sinistres doivent être relativement homogènes, car le modèle fournira une réponse similaire pour ces sinistres. Or, les coûts observés varient fortement :

Fig. 13 : Segment de 18 sinistres homogènes de risque, classés par montant d'expertise

250 000 €	200 040 €	50 000 €	37 000 €	29 725 €	25 000 €	23 030 €	18 000 €	15 600 €
15 500 €	14 000 €	14 000 €	13 500 €	10 304 €	9 000 €	8 700 €	7 750 €	7 000 €

La moyenne des coûts des sinistres de ce segment est de 41 563 €, ce qui n'est pas en adéquation avec les coûts de la majorité des sinistres. Si on retire les 2 sinistres les plus graves, afin qu'ils soient traités à part, on obtient deux segments :

Fig. 14 : Nouveaux segments avec un seuil arbitraire

250 000 €
200 040 €

Segment 1

50 000 €	37 000 €	29 725 €	25 000 €	23 030 €	18 000 €	15 600 €	15 500 €
14 000 €	14 000 €	13 500 €	10 304 €	9 000 €	8 700 €	7 750 €	7 000 €

Segment 2

La moyenne des coûts dans le premier segment est de 225 020 €, et dans le deuxième de 18 631 €. Cet exemple purement illustratif montre bien l'intérêt d'écrêter les données afin d'homogénéiser les segments. En pratique, il n'est pas possible de fixer un seuil pour chaque segment, il est donc nécessaire de fixer un seuil arbitrairement qui sera appliqué sur toute la base.

I.b.1.2. Choix du seuil à droite

Il est nécessaire de choisir un seuil adapté aux données. Si le seuil est trop faible, une partie des sinistres moyens se mélangera aux sinistres graves, sous estimant ainsi leurs impacts. S'il est trop haut, cela pourrait entraver une potentielle étude des sinistres graves, ou encore rendre inopérant l'écrêtage.

Il est possible de déterminer le seuil de plusieurs manières suivant l'objectif que l'on souhaite remplir.

Si le but est de modéliser à part les sinistres graves, comme c'est souvent le cas pour les lois à queue de distribution lourde, il est nécessaire de garder assez d'observations pour que la modélisation soit pertinente. Dans ce cas, on utilise le *Mean Excess Plot* pour déterminer le seuil, en repérant le moment où la fonction devient linéaire.

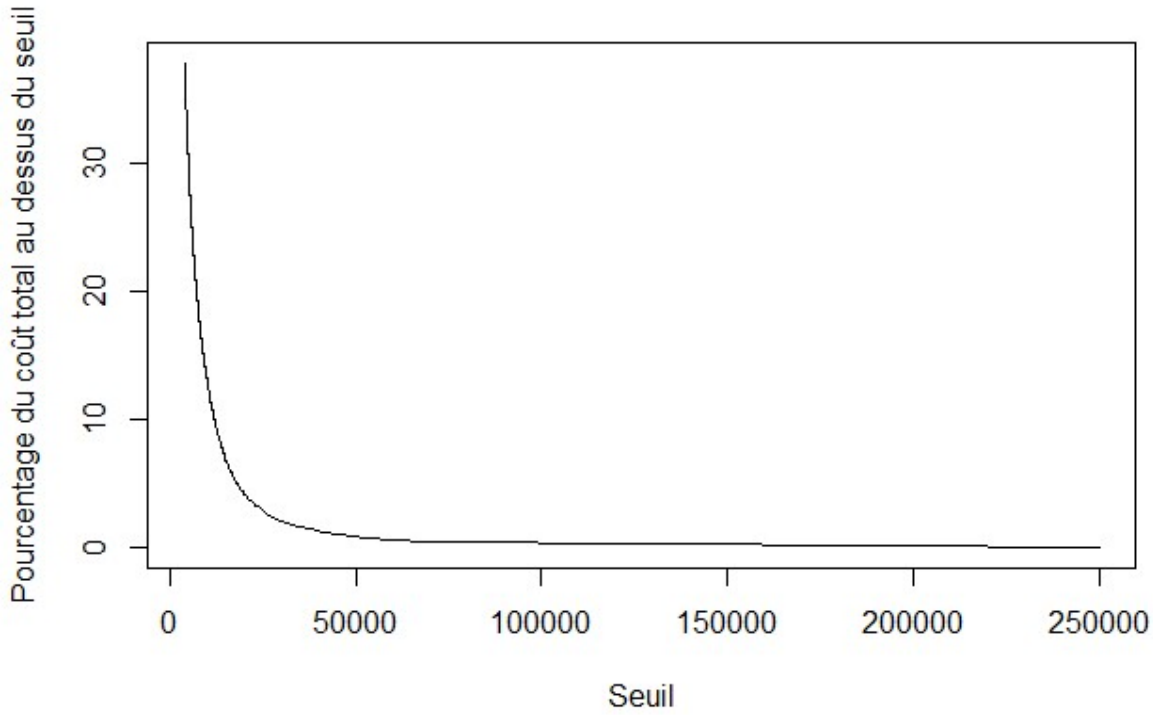
L'autre méthode, utilisée pour les queues de distributions moins lourdes, consiste à fixer un seuil élevé : on considère alors que l'apparition d'un sinistre exceptionnel dans un segment constitue « un coup de malchance », et qu'il n'y a pas de raison de perturber sa distribution plus qu'une autre.

Pour cette méthode, il est nécessaire de garder en tête que la moyenne des coûts sera légèrement sous-évaluée, et de ce fait on pourra éventuellement la redresser par la suite en ajoutant une constante calculée à partir de l'historique des sinistres exceptionnels écartés. Pour les prédictions, on attribue aux grands risques un montant unique mutualisé sur l'ensemble des segments. L'avantage est que les prédictions sont bien plus précises en utilisant ce principe.

La deuxième méthode semble celle qui convient le mieux au cadre de l'étude. En effet, le portefeuille de sinistres présente une proportion importante de sinistres moyens et peu de sinistres graves. Afin d'avoir assez de sinistres pour modéliser les sinistres graves, il serait nécessaire de fixer un seuil bas qui engloberait une bonne partie des sinistres moyens, ce qui n'est pas satisfaisant. D'autre part, l'objectif initial étant l'écart de performance entre deux réseaux, il semble logique d'écarter les sinistres exceptionnels du fait de leur volatilité.

Pour fixer le seuil à partir duquel les sinistres ne sont plus considérés, il est possible de s'inspirer de l'évolution de la somme des coûts des sinistres au-dessus du seuil, un peu à la manière du Mean Excess Plot. L'idée est de regarder à partir de quel seuil cette valeur n'évolue plus.

Fig. 15 : Coût au-dessus du seuil



La courbe semble stagner un peu avant 50 000 €, il est intéressant d'étayer cet élément par un tableau renseignant les valeurs précises :

Tab. 5. Coût et nombre de sinistres au-dessus du seuil - Tableau

Seuil	Pourcentage du coût au-dessus du seuil	Nombre de sinistres au-dessus du seuil
21000	3,88 %	0,23 %
21500	3,78 %	0,22 %
22000	3,67 %	0,21 %
22174	3,55 %	0,20 %
22793	3,44 %	0,19 %
23083	3,32 %	0,18 %
23833	3,20 %	0,17 %
24167	3,08 %	0,16 %
25000	2,96 %	0,15 %
25400	2,83 %	0,14 %
26000	2,70 %	0,13 %
27000	2,56 %	0,12 %
28000	2,42 %	0,11 %
29040	2,28 %	0,10 %
30000	2,13 %	0,09 %
31800	1,97 %	0,08 %
33109	1,81 %	0,07 %
35000	1,64 %	0,06 %
37981	1,45 %	0,05 %
41095	1,25 %	0,04 %
45545	1,03 %	0,03 %
52680	0,78 %	0,02 %
69000	0,48 %	0,01 %

Pour choisir le seuil, il semble pertinent de faire la balance entre :

- La valeur du seuil, pour éviter d'avoir des segments trop hétérogènes ;
- Le coût total des sinistres écartés ;
- Le nombre de sinistres écartés.

Dans le cas ici présent, un seuil de 30 000 € semble être un bon choix. En effet, seulement 0,09 % des sinistres, soit moins de 800 observations, sont écartés par ce procédé, ce qui correspond à 2,13 % du coût total.

Suite à l'application de ce seuil, la base comporte désormais 876 146 observations, contre 876 935 auparavant.

I.b.1.3. Choix du seuil à gauche

Il est important de noter que la franchise est incluse dans le montant de l'expertise. Cela veut dire que certains sinistres sont reportés dans la base de données mais ne seront pas payés par l'assureur, car pour ceux-ci la franchise est plus grande que le coût. Néanmoins, nous ne disposons pas d'informations ligne à ligne concernant la franchise appliquée. L'information disponible est que seul 33 % des réparations de véhicules légers ont une franchise, et celle-ci coûte en moyenne 334 € TTC.

Compte tenu de cette information, il a été décidé d'appliquer un seuil à gauche égal à 334 €. L'avantage de ce seuil est qu'il permet d'éliminer les anomalies liées aux sinistres faibles, que ce soit pour une franchise non prise en compte ou un montant d'expertise mal renseigné. De plus, ces sinistres sont assez nombreux alors qu'ils représentent un intérêt limité par rapport au coût total du portefeuille : on observe que les 4,5 % des sinistres les plus faibles représentent 0,6 % du coût total.

Dans la même idée que pour l'établissement du seuil à droite, à partir du moment où le pourcentage du coût total écarté reste raisonnable, il est préférable d'écarter ces sinistres de l'étude afin de ne pas perturber inutilement la distribution des segments du portefeuille.

Suite à l'application de ce seuil, la base comporte désormais 835 881 observations, contre 876 146 auparavant.

I.b.1.4. Choix de la loi et de la fonction de lien

Pour la loi de Y | X, la loi gamma a été choisie du fait de son caractère flexible. De plus, elle convient bien aux distributions asymétriques et positives.

Pour la fonction de lien, la fonction « log » est utilisée pour avoir des tarifs multiplicatifs. En effet, avec

$$log(E[Y|X]) = \beta X$$
$$E[Y|X] = \exp(\beta_1 X_1 + \cdots + \beta_p X_p)$$
$$E[Y|X] \approx (1+\beta_1)^{X_1} * \ldots * \left(1+\beta_p\right)^{X_p}$$

Car $e^x = 1 + x + o(x)$ quand x est petit.
L'avantage de cette fonction de lien est qu'elle rend l'interprétation du tarif intuitive. Par exemple, si X_i est une variable à deux modalités, à valeurs dans $\{0,1\}$, et que $\beta_i = -0{,}02$, cela signifie que la présence de X_i diminue le prix d'environ 2 %. Ainsi, la lecture des coefficients renseigne directement l'influence de la variable sur le coût final. Pour évaluer

la performance du réseau d'expert, il suffira de regarder le coefficient associé à la variable cible.

Ces paramètres ont été choisis pour leur simplicité d'implémentation. Il est possible de faire une étude extensive pour optimiser le choix de la loi, tout comme pour le choix du seuil et du lien log.

I.b.2. Optimisation des variables

I.b.2.1. Réduction du nombre de facteurs

Lors de la modélisation, du fait de son caractère pratique, il est courant d'être confronté à certaines contraintes techniques comme la complexité du calcul. Ici, la base de données comporte près de 800 000 lignes. Ajouté à cela, le GLM doit calculer un coefficient par la méthode du maximum de vraisemblance pour chaque modalité de chaque variable. Afin de pouvoir obtenir un résultat en pratique, il est donc nécessaire de faire des choix, parfois assez drastiques, concernant le regroupement des modalités. Les variables concernées sont le département, la marque et le mois.

Les départements sont regroupés en 6 classes : « Nord », « Est », « Ouest, « Sud », « Région Parisienne » et « Autre », et la variable est renommée « Région ».

Quant à la marque, seule les 12 plus importantes représentant 85 % des observations sont conservées tandis que les autres sont regroupées sous une même modalité, nommée « Autre ».

Pour les variables trimestre et mois, elles sont supprimées. En effet, elles semblent peu pertinentes dans le cadre d'une approche prospective.

I.b.2.2. Sélection des variables

Il existe de nombreuses façons de sélectionner les variables, le but étant de ne garder que les plus pertinentes. La méthode retenue repose sur l'AIC du modèle, qui est un critère pénalisant la log-vraisemblance du modèle par le nombre de paramètres retenus.

L'avantage du critère AIC est qu'il n'accepte l'intégration d'un nouveau paramètre qu'à condition que celui-ci améliore significativement le modèle. Cela permet d'obtenir un modèle avec un nombre raisonnable de variables, ce qui est un point important lors de l'implémentation pratique du modèle.

La « sélection ascendante et descendante » sur le critère AIC fonctionne comme suit :

- Initialisation : on part du modèle avec toutes les variables explicatives (ou de manière équivalente, avec aucune variable)
- Etape i : on ajoute (ou retire) la variable qui conduit à la meilleure amélioration de l'AIC du modèle.

L'algorithme s'arrête lorsque l'AIC n'augmente plus par l'ajout ou par la suppression d'une variable. On observe que cette méthode conduit à garder toutes les variables explicatives (la marque, le kilométrage, la variable cible, le type du sinistre, l'âge du véhicule, l'issue, l'assureur, le taux horaire, la densité et la région).

I.c. Application et résultats du GLM

Il est rappelé que la distribution retenue pour la loi conditionnelle de la variable à expliquer sachant les variables explicatives est la loi gamma, avec une fonction de lien logarithmique et un seuil à droite de 30 000 € et un seuil à gauche de 334€.

I.c.1. Périmètre d'application du modèle

Pour rappel, cette étude repose sur deux objectifs : reconstruire le coût moyen à travers un modèle prédictif et estimer l'écart de performance entre les deux réseaux d'experts.

Pour l'estimation de l'écart de performance, l'ensemble de la base est utilisé afin de n'éluder aucune information, le but étant d'avoir la meilleure estimation compte tenue de l'historique des données. Les 3 échantillons d'apprentissage, de validation et de test sont utilisés dans le cadre de la production des prédictions.

I.c.2. Estimation de la performance du réseau

Les résultats du GLM avec une loi gamma et fonction de lien log sont présentés en annexe.

On constate ainsi que, toutes choses égales par ailleurs, le réseau d'experts fait baisser le coût total du sinistre d'environ 2,4 %, correspondant à la valeur du coefficient estimé. Cette mesure se situe dans un intervalle de confiance à 95 % de plus ou moins 0,4 %.

Les coefficients sont cohérents avec le V de Cramer qui a été calculé entre la variable à expliquer et les variables explicatives auparavant. On remarque que la variable taux horaire, notamment sa modalité « tauxHoraireNUL » signifiant que le sinistre est VEI, est bien la variable ayant le plus d'impact sur le coût final.

D'autre part, on remarque que le test de significativité des variables est bon dans son ensemble. Seule une modalité de la densité est jugée non significative. Cela conforte le choix et la sélection des variables pour le modèle.

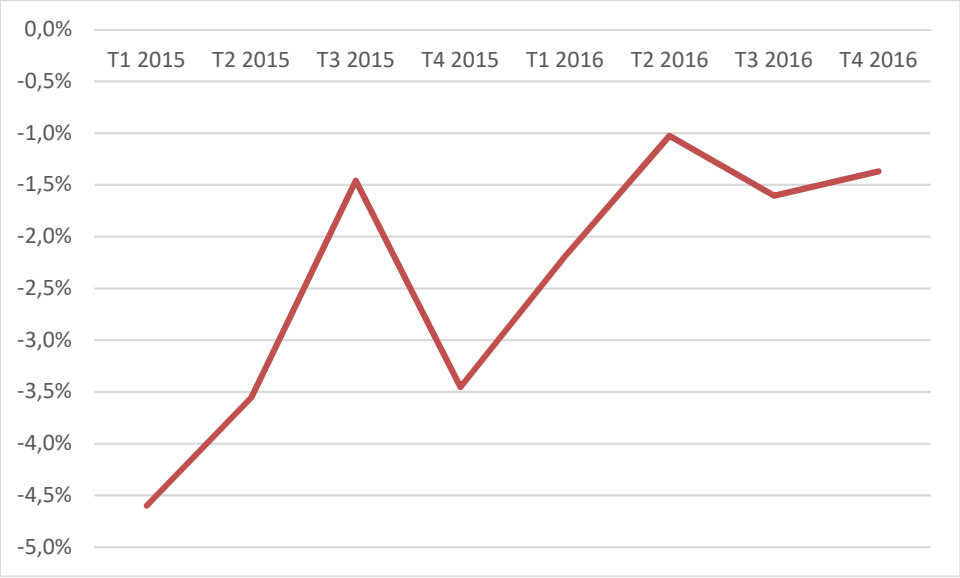
L'estimation du coefficient associé au réseau d'experts peut également se faire par trimestre, ce qui permet de proposer une évolution temporelle de l'écart de performance :

Tab. 6. Coefficient trimestriel de la variable cible

Trimestre	Coefficient	Ecart-type	Effectif
T1 2015	-4,6%	0,50%	115867
T2 2015	-3,6%	0,54%	101329
T3 2015	-1,5%	0,59%	90799
T4 2015	-3,5%	0,51%	120122
T1 2016	-2,2%	0,52%	109334
T2 2016	-1,0%	0,55%	100277
T3 2016	-1,6%	0,57%	92326
T4 2016	-1,4%	0,54%	105827

Et le graphique associé :

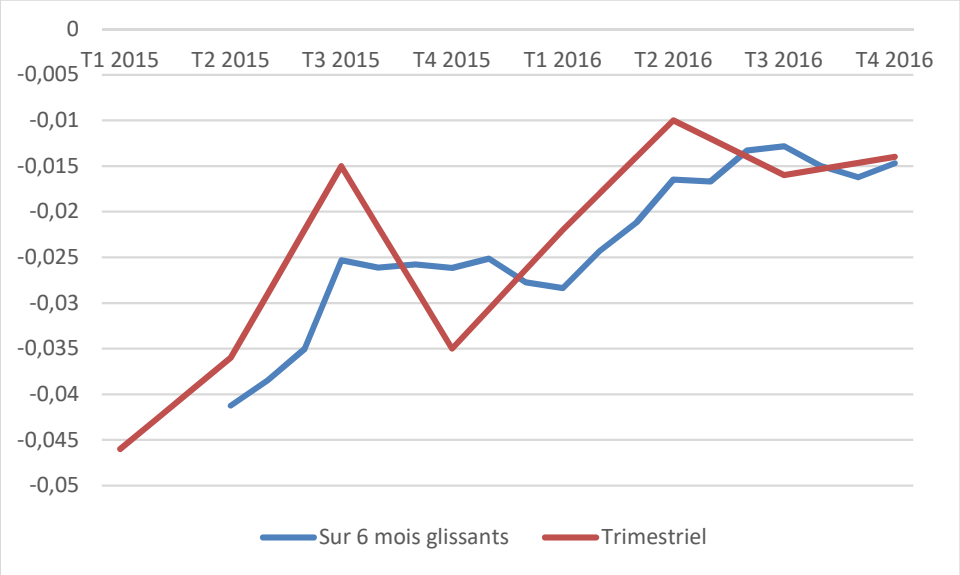
Fig. 16 : Evolution de l'écart de performance au fil des trimestres



Ces résultats indique une tendance à un resserrement de l'écart entre les deux réseaux, néanmoins toujours à l'avantage du réseau d'expert A. Ceci étant dit, étant donné l'écart type, l'intervalle de confiance à 95 % se situe à plus ou moins 1 % de l'estimation, ce qui signifie que l'écart peut basculer au prochain trimestre.

Le coefficient trimestriel est très volatil : il varie de -4,6 % à -1,0 %, et la plus grande variation enregistrée se situe entre le T2 2015 et le T3 2015 où l'indicateur perd 2 points. Afin de proposer une estimation plus propice à une prise de décision, il est possible de mettre en place un coefficient sur 6 mois glissants.

Fig. 17 : Evolution de l'écart de performance sur 6 mois glissants



Il est à préciser que le coefficient sur 6 mois glissants ne commence qu'au T2 2015, étant donné que celui-ci nécessite un historique de 6 mois d'observations.

La vision sur 6 mois glissants est plus stable que la vision trimestrielle, et également plus « saccadée » étant donné que l'estimation se fait tous les mois. On note que l'effet de tendance est correctement restitué, tout autant que les variations qu'on peut observer entre le T2 2015 et T3 2015, et entre le T1 2016 et le T2 2016.

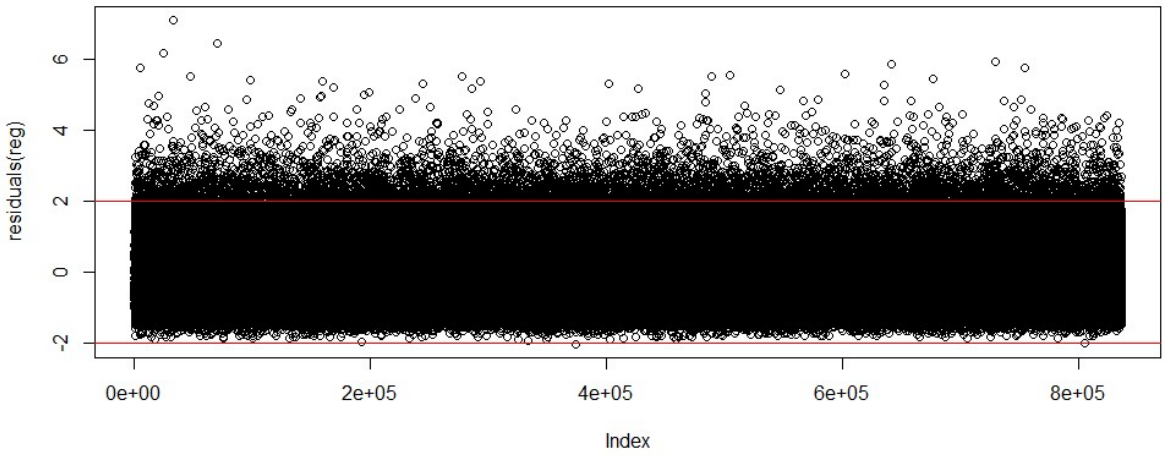
Le désavantage de la vision sur 6 mois glissants est qu'il est plus difficile d'attribuer une date à une variation. L'estimation étant faite sur les 6 derniers mois de manière glissante, une variation de l'estimateur peut autant être dû à l'inclusion du mois le plus récent qu'à l'exclusion du mois le plus ancien.

Néanmoins, au détriment d'une vision instantanée, il est plus aisé de ressortir un avis général quant à l'évolution de l'estimateur, et cela facilite ainsi l'interprétation et la prise de décision, notamment décider de la mise en place ou non d'un plan d'action afin d'inverser la tendance. Si nécessaire, il est toujours possible d'utiliser les deux perspectives.

I.c.3. Validation du modèle

Enfin, une bonne façon d'asserter la validité d'un GLM est d'analyser les résidus.

Fig. 18 : Résidus du glm



En tenant compte de l'écrêtement à droite qui est relativement faible, les résidus sont en apparence plutôt satisfaisants. L'élément important est de vérifier si la majorité des résidus se trouvent dans la bande [-2,2]. En effet, les résidus représentés ici sont les résidus de Pearson, et en théorie ils devraient suivre une loi normale centrée réduite. Il est donc nécessaire de s'assurer qu'ils ne s'écartent pas trop des valeurs d'une telle loi de probabilité. Pour cela, on trace la densité pour mieux s'en apercevoir :

Fig. 19 : Densité des résidus du GLM

On se rend compte que la majorité des résidus se trouvent dans la bande [-2,2], seul 0,73 % qui se trouvant en dehors de celle-ci. Bien que légèrement décalée sur la gauche, l'allure de la densité empirique des résidus ressemble bien à celle d'une normale centrée réduite.

A cela est ajouté un test de normalité de Jarque-Bera, dont la p-valeur est inférieure à $2{,}2 * 10^{-16}$, ce qui encore une fois va dans le sens d'un modèle statistique de bonne qualité.

I.d. Pouvoir prédictif du GLM

I.d.1. Périmètre

On rappelle que pour avoir une mesure comparable du pouvoir prédictif de chaque modèle, il est nécessaire d'utiliser un périmètre commun. Pour ce faire, le modèle est calibré sur l'échantillon test, optimisé sur l'échantillon de validation, et le pouvoir prédictif est évalué sur l'échantillon test. Le critère utilisé pour cette évaluation est l'erreur quadratique moyenne, dont la pertinence a été précisée dans le premier chapitre.

Cette décomposition est peu habituelle lorsqu'il s'agit d'un GLM, mais cela permet de mettre le GLM sur le même pied d'égalité que les autres modèles. En effet, si le pouvoir prédictif du GLM était évalué sur l'ensemble de la base, tandis que les autres modèles ne sont évalués que sur l'échantillon test, le GLM profiterait alors d'un biais d'informations.

I.d.2. Résultats de la mesure d'erreur

Les coefficients du GLM sont estimés à partir de l'échantillon d'apprentissage, puis un vecteur de prédiction est généré en injectant l'échantillon test dans le modèle. Dès lors, les prédictions et les valeurs réelles du montant d'expertise sont comparées à travers l'erreur quadratique moyenne, dont on rappelle la forme :

$$MSE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \widehat{y_i})^2$$

Avec y_i la valeur observée du montant d'expertise du sinistre i, et $\widehat{y_i}$ la valeur prédite par le modèle GLM du montant d'expertise du sinistre i.

Ainsi, on trouve la valeur suivante : $\boldsymbol{MSE = 3\,639\,942}$. Cette valeur n'a un sens que si elle est comparée à celle d'un autre modèle, afin de déterminer le meilleur des deux.

Il est intéressant d'introduire une autre mesure pour avoir une vision plus concrète de l'efficacité du modèle. Pour cela, nous utilisons la mesure de corrélation entre le vecteur des prédictions et celui des observations. Cette mesure est jugée moins précise que la première, mais elle a l'avantage de pouvoir être interprétée indépendamment des autres.

On trouve ici $\boldsymbol{COR = 0,55}$.
Petit rappel sur le coefficient de corrélation : il peut prendre une valeur entre -1 et 1, et lorsque la valeur est dans les extrêmes, cela indique une relation affine entre les deux variables, croissante si la corrélation est positive, décroissante si la corrélation est négative. Globalement, c'est une mesure qui renseigne sur le degré de dépendance linéaire entre les deux variables.

I.e. Conclusion

Un modèle linéaire généralisé avec une loi gamma et une fonction de lien log a été ajusté sur les données. Les avantages de ce modèle sont multiples : il est simple à mettre en place et il permet d'interpréter facilement l'influence d'une variable. L'analyse des résidus et la significativité des variables vont dans le sens d'un modèle robuste statistiquement, ce qui conforte la validité et la significativité des résultats.

Ce type de modèle a montré la nécessité d'un seuil à droite comme à gauche afin d'exclure les éléments isolés qui perturbent de manière importante la distribution d'un segment du portefeuille de sinistres. Afin de garder le même périmètre d'étude, le seuil à droite de 30 000 € et le seuil à gauche de 334 € sont conservés pour la suite.

Concernant la mesure de l'écart de performance, la volatilité des résultats nous a conduit à produire un indicateur sur 6 mois glissants afin de faciliter l'interprétation des résultats et la prise de décisions vis-à-vis de ceux-ci. Au global, le réseau d'experts A semble être plus performant, avec un écart dans le coût final d'environ 2,4 % avec le réseau d'experts B. Néanmoins, l'écart semble se resserrer au fil des trimestres.

Enfin, le pouvoir prédictif du modèle calculé à travers l'erreur quadratique moyenne et le coefficient de corrélation semble être d'une qualité correcte. Ces mesures seront comparées aux autres modèles afin d'asserter plus concrètement de la valeur prédictive du GLM.

La modélisation par une approche GLM n'est pas sans défaut. Son aspect paramétrique nous demande de faire plusieurs hypothèses, comme sur la loi de Y | X, qui peuvent se révéler imprécises. De plus, la construction des variables relève de l'expertise du statisticien, tout comme leur sélection en amont. Certaines méthodes permettent de contourner ces difficultés, et il est intéressant de comparer les performances de celles-ci avec le GLM.

II. RGLM

II.a. Principe du RGLM

II.a.1. Contexte

Le Random Generalised Linear Model, aussi appelé « forêt aléatoire de GLM », emprunte les idées des forêts aléatoires et du GLM, afin de les assembler et d'en retirer un meilleur modèle. Le GLM a su prouver sa stabilité et est largement utilisé en actuariat. L'idée ici n'est pas de remplacer le GLM, mais de l'améliorer en ajoutant des concepts retrouvés dans l'apprentissage statistique, plus particulièrement dans les forêts aléatoires. Le RGLM a obtenu d'excellents résultats en biostatistiques (Song L., 2013), surpassant notamment les forêts aléatoires, les GLM, les arbres de régression, la classification par les K plus proches voisins et les machines à vecteur support sur une douzaine de problématiques différentes. Ces résultats sont encourageants et il est intéressant de voir les performances du RGLM dans le domaine de l'actuariat. Il est donc étudié ici.

La qualité d'un modèle de prédiction est souvent liée à sa précision, parfois au détriment de l'interprétabilité. C'est le cas des forêts aléatoires, qui sont réputées pour la précision de leurs prédictions, mais perdent l'aspect visuel de l'arbre de régression. Elles ont ainsi un effet « boîte noire » dû à la difficulté de relier les prédictions à chaque arbre généré.

A l'inverse, lorsqu'on utilise une sélection de variables unilatérales, le GLM est facilement interprétable. Il associe un coefficient à chaque variable, et la prédiction se fait par la résolution d'une simple équation. Le problème du GLM est qu'il tend à sur-apprendre rapidement, notamment lorsque le nombre de variables est important par rapport au nombre d'observations.

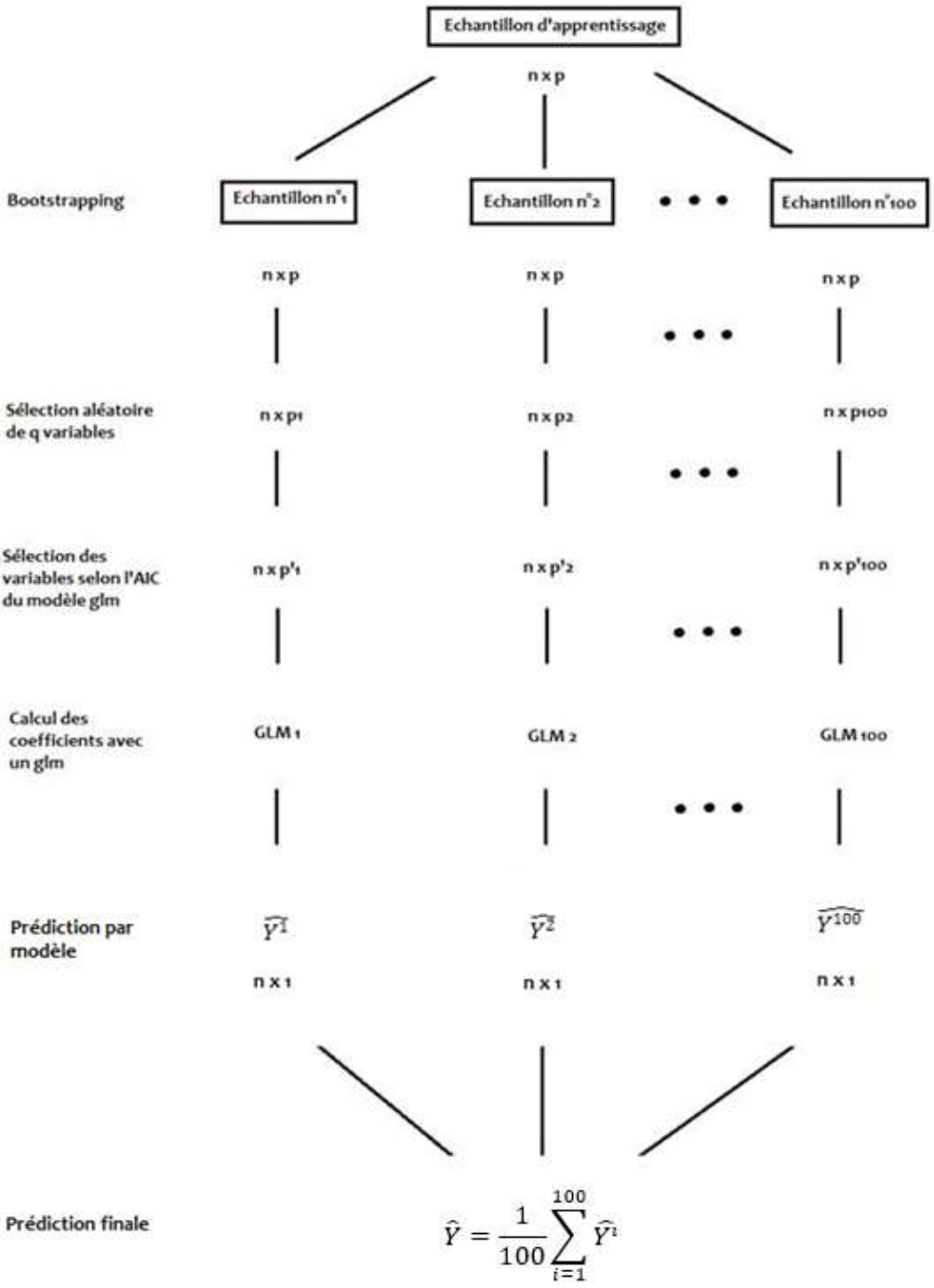
L'idée du RGLM est donc d'allier le GLM et les forêts aléatoires pour obtenir une modélisation qui soit à la fois précise et facile à interpréter.

Pour l'écriture de ce chapitre nous nous appuyons sur le cours d'introduction aux techniques avancées de tarification IARD de F. Planchet et A. Miseray (Miseray – Planchet, 2017).

II.a.2. Algorithme

La logique de modélisation est décrite par le schéma suivant :

Fig. 20 : Schéma descriptif de la méthode RGLM

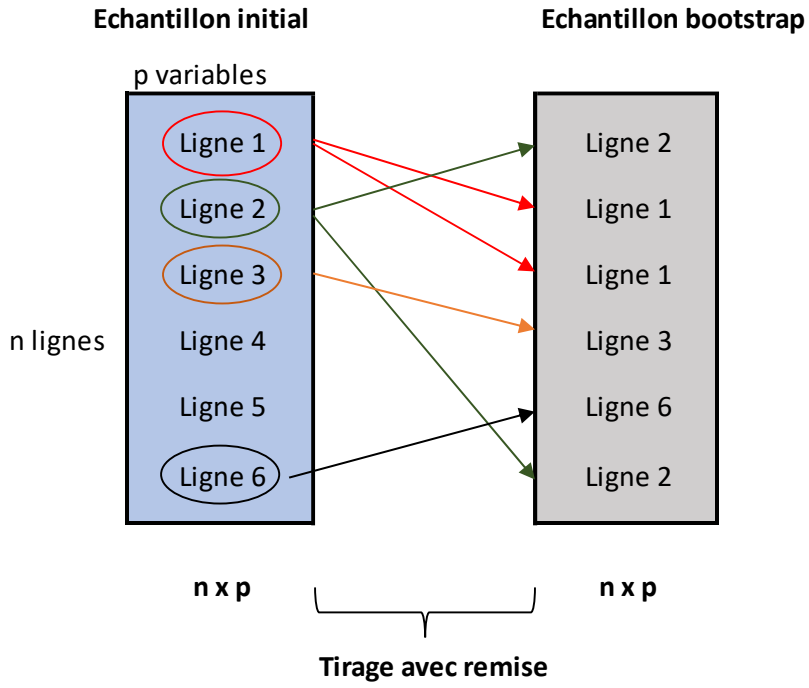


B échantillons bootstrap sont générés :

Dans un premier temps, le RGLM génère plusieurs scenarii similaires aux données observées. L'idée derrière cela est d'augmenter « l'expérience » du modèle en l'habituant au contexte de l'étude.

Pour cela, il faut simuler la loi de distribution que suit l'échantillon initial, afin de pouvoir tirer de nouveaux échantillons. Comme il est difficile d'évaluer la distribution que suit l'échantillon, l'idée du *bootstrap* est d'approcher cette distribution par la distribution empirique, calculée à partir de l'échantillon, en attribuant un poids de 1/n à chaque réalisation. En d'autres termes, pour créer un nouvel échantillon, on effectue n tirages aléatoires avec remise de l'échantillon initial. Avec cela, on obtient un échantillon dont les observations sont bien indépendantes. Voici un exemple pour que ce soit plus clair :

Fig. 21 : Exemple du bootstrapping d'un échantillon de 6 lignes



Ainsi, certaines observations ne se retrouvent pas dans l'échantillon *bootstrap* (les lignes 4 et 5 dans l'exemple) tandis que d'autres apparaissent plusieurs fois (les lignes 1 et 2).

Sélection aléatoire de q variables :

Pour chaque échantillon *bootstrap* généré, on sélectionne q variables aléatoirement parmi les p variables existantes. Chaque échantillon *bootstrap* aura ainsi un ensemble de variables différent. L'idée derrière cela est de pouvoir analyser des compositions de variables auxquelles on n'aurait pas eu accès avec une simple sélection ascendante et descendante.

Sélection des variables selon un critère explicatif :
Une fois les q variables aléatoirement sélectionnées, seule une partie d'entre elles sont définitivement retenues en utilisant une sélection ascendante selon le critère AIC. Il est possible de rajouter un paramètre supplémentaire pour contrôler le nombre de variables retenues par ce critère.

Calcul des coefficients et prédictions par GLM :
Un GLM avec une loi gamma et une fonction de lien log est ajusté sur l'échantillon bootstrappé, reliant le montant d'expertise aux variables explicatives sélectionnées dans la partie précédente. Les coefficients associés à chaque modalité sont ainsi calculés, et permettent d'en déduire les prédictions du GLM.

Prédiction finale :
Les prédictions de tous les GLM ajustés sur les différents échantillons *bootstraps* sont agrégées. La prédiction finale est alors calculée comme la moyenne des prédictions de chaque GLM, permettant à partir de celle-ci de calculer entre autres les erreurs de prédictions.

II.a.3. Estimation de l'écart de performance des réseaux

Il est naturel de se demander si le RGLM, de par sa structure à base de GLM, est un modèle qui permet d'évaluer l'influence des variables explicatives sur la variable à expliquer. On rappelle que pour un GLM classique, on a la relation suivante :

$$E[Y|X] \approx (1+\beta_1)^{X_1} * \dots * \left(1+\beta_p\right)^{X_p} = \prod_{i=1}^{p}(1+\beta_i)^{X_i}$$

Le RGLM n'est en soit qu'une moyenne de GLM. Le point qui complique les choses est l'aléatoire dans la sélection des variables : cela signifie qu'un coefficient n'est pas toujours évalué. On a alors, avec 100 échantillons bootstrappés :

$$E[Y|X] \approx \frac{1}{100}\sum_{j=1}^{100}\left(\prod_{i=1}^{p}\left(1+\beta_{ij}\right)^{X_i}\right)$$

où β_{ij} est le coefficient évalué pour la variable i sur l'échantillon bootstrappé j, potentiellement nul si la variable n'a pas été sélectionnée sur l'échantillon j.

Pour isoler l'action de chaque variable, la question est de savoir s'il est possible de trouver un vecteur γ en fonction de β tel que :

$$(1+\gamma_1)^{X_1} * \ldots * (1+\gamma_p)^{X_p} = \frac{1}{100}\sum_{j=1}^{100}\left(\prod_{i=1}^{p}(1+\beta_{ij})^{X_i}\right)$$

Or, il n'existe pas de solution unique à ce problème, ce qui ne permet pas au RGLM de définir proprement l'influence de la variable cible.

Le RGLM est donc dans l'incapacité de fournir un élément de réponse à l'évaluation de l'écart de performance entre les deux réseaux d'experts.

II.b. Application du RGLM

Dans cette partie est appliqué le RGLM tel qu'il est décrit dans la partie précédente.

II.b.1. Introduction d'une nouvelle mesure d'erreur

Bien qu'ayant appuyé la pertinence de l'erreur moyenne quadratique, il est intéressant de la comparer à d'autres afin de s'assurer que le résultat ne soit pas biaisé par la mesure d'erreur choisie.

De ce fait, on ajoute le coefficient de corrélation présenté précédemment, la norme L_1 (en contraste avec la norme L_2 représentée par l'erreur quadratique moyenne), et l'erreur par segment. Ces deux dernières mesures sont définies de la manière suivante :

$$Erreur\ L_1 = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

$$EPS = \frac{1}{n}\sum_{variable}\sum_{modalité} n_{moda} * (moyenne\ prédite_{moda} - moyenne\ observée_{moda})^2$$

Avec :

- n le nombre de lignes de l'échantillon (ici, celui de l'échantillon de validation) ;
- y_i la valeur observée du montant d'expertise ;
- $\widehat{y_i}$ la valeur prédite du montant d'expertise ;
- n_{moda} le nombre de lignes de l'échantillon tel que la modalité est présente ;
- $moyenne\ prédite_{moda}$ *et* $moyenne\ observée_{moda}$ les moyennes calculées sur l'échantillon telles que la modalité est présente.

L'erreur L1 est une mesure connue, mais bien moins populaire que l'erreur quadratique moyenne, et permet d'avoir un aperçu sur l'écart absolu moyen entre prédictions et observations.

L'erreur par segment est un peu plus particulière. Elle a été conçue spécialement pour ce modèle, afin d'avoir un indicateur permettant de rendre compte de l'écart des prédictions

par segment de modalité. Autrement dit, c'est une mesure permettant de vérifier la capacité du modèle à reconstituer les moyennes par segments de population. Par exemple, elle va comparer l'écart entre la moyenne prédite et observée pour tous les véhicules sinistrés de marque Audi, pondéré par le nombre d'observations associées afin que la mesure d'erreur soit impactée proportionnellement à la taille du sous-échantillon sélectionné.

II.b.2. Optimisation des paramètres

Dans le modèle RGLM, deux paramètres sont laissés au choix de l'utilisateur : le nombre d'échantillon bootstrap utilisé et le nombre q de variables sélectionnées aléatoirement.

Idéalement, il faudrait pouvoir prendre un maximum d'échantillons bootstrap, de l'ordre du millier, pour accroître les performances du modèle. Cependant, les contraintes de temps de calcul nous imposent de réviser ce chiffre et de le fixer à 100. Etant donné que seule une moyenne du résultat de ces modèles est calculée, 100 échantillons semblent suffisants pour avoir une idée précise de l'efficacité du modèle.

Il est également nécessaire d'optimiser le nombre q de variables sélectionnées aléatoirement. Pour cela, on effectue la phase d'apprentissage sur l'échantillon d'apprentissage, et on calcule l'erreur de prédiction sur l'échantillon de validation, qui nous sert à optimiser les paramètres. On dispose de 9 variables explicatives pour une variable à expliquer, le domaine d'existence de q est donc l'intervalle [1 ,9]. On limite ce domaine à l'intervalle [3,9] afin d'avoir un nombre suffisant de variables par modèle, garantissant ainsi une qualité de prédiction minimum.

Ainsi, le modèle est calibré pour 7 valeurs différentes de q, et à chaque fois l'erreur moyenne quadratique est calculée sur l'échantillon de validation, appuyée par la mesure de corrélation, l'erreur L1 et l'erreur par segment, afin de déterminer la valeur de q optimale. Les résultats pour les différentes valeurs de q sont les suivants :

Tab. 7. Erreurs du modèle sur l'échantillon de validation en fonction des valeurs de q

q	COR	MSE	L1	EPS
3	0,49	4429006	952	555683
4	0,51	4256629	917	420588
5	0,52	4143589	892	348378
6	0,54	3905629	940	192241
7	0,54	3786473	808	118412
8	0,55	3680157	790	79601
9	**0,55**	**3601134**	**772**	**51110**

Il semble que conserver toutes les variables, et donc supprimer l'aléatoire lié à la sélection de celles-ci, soit le meilleur choix pour le modèle. Ce résultat est cohérent avec celui présenté dans l'étude biostatistique (Song L., 2013), qui préconise de ne pas sélectionner aléatoirement les variables explicatives lorsque le nombre de celles-ci est inférieur à 10.

II.b.3. Résultats et comparaison avec le GLM

Le RGLM ayant pour base de fonctionnement le GLM, il est intéressant de comparer ses résultats avec celui du GLM précédemment utilisé. Les résultats sont résumés dans le tableau ci-dessous. Ils ont été obtenus en utilisant exactement les mêmes retraitements, en particulier ceux issus de la réduction du nombre de modalités liées à l'utilisation d'un GLM.

Tab. 8. Comparaison des erreurs de prédictions sur l'échantillon test

	COR	MSE	L1	EPS
GLM	0,55	3 639 942	770	49 777
RGLM	0,55	**3 639 483**	770	**49 710**

On constate que d'un point de vue prédictif, les deux modèles implémentés sont quasi équivalents. Il semble que l'élément clé qui a rendu le RGLM inefficace est la sélection aléatoire des variables, qui est jugée inutile lors de l'optimisation de ce paramètre. Cela réduit finalement le RGLM à faire une moyenne de GLM sur des échantillons bootstrappés. Nous allons voir dans la partie suivante s'il n'est pas possible d'adapter le RGLM à la structure des données, et ainsi valoriser la sélection aléatoire de variables.

II.c. Proposition d'amélioration de l'algorithme et son application

II.c.1. Ajout des termes d'interactions

Le RGLM est un modèle qui a fait ses preuves en biostatistique (Song L., 2013), mais le cadre est bien différent du nôtre. En effet, les bases de données de biostatistique possèdent des milliers de variables et quelques dizaines de milliers d'observations, contrairement au cas étudié ici qui ne possède que quelques variables et un très grand nombre d'observations. L'objectif va donc être d'adapter l'algorithme du RGLM à un problème avec peu de variables et beaucoup d'observations.

La première étape est de rajouter artificiellement des variables afin de se rapprocher du cas de la grande dimension. L'idée va être d'introduire des interactions entre chaque variable déjà existante. Pour rappel, lors de l'ajustement d'un GLM on estime les coefficients β tels que : $g(E[Y|X]) = \beta X$. Par exemple, pour deux variables explicatives, cela revient à calculer β_1 et β_2 tels que $g(E[Y|X]) = \beta_1 X_1 + \beta_2 X_2$. Ajouter le terme

d'interaction entre X_1 et X_2 revient à « créer » une troisième variable $Z = X_1 * X_2$, à laquelle on associe un nouveau coefficient β_3 qu'il faudra également estimer. La relation devient ainsi $g(E[Y|X]) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$.

De ce fait, la relation est complexifiée et permet de modéliser des comportements plus riches. Notamment, l'effet d'une variable n'est plus constant, mais va dépendre des valeurs des autres variables. Par exemple, il semble logique que la marque du véhicule ait un impact assez important sur le coût d'un sinistre. Mais il est d'autant plus intéressant de disséquer son influence en fonction du type de sinistre ou du kilométrage : certaines marques sont plus résistantes à tel ou tel type d'accident suivant la sélection de leurs matériaux, et certaines marques s'usent plus vite que d'autres. C'est ce qu'accomplit le terme d'interaction.

II.c.2. Facteurs discriminants

Le point des facteurs discriminants est d'autant plus important que le nombre de variables sélectionnées q est petit. En effet, sur un échantillon *bootstrap*, la question de la pertinence de la modélisation se pose si les variables sélectionnées ne sont pas assez explicatives. Lorsque c'est le cas, la significativité des variables du modèle chute fortement, et l'intercept du modèle devient prédominante. Chaque prédiction n'est alors qu'une légère altération de la moyenne empirique de l'échantillon d'apprentissage.

Afin d'éviter cette situation, plusieurs solutions sont envisageables. Par exemple, il est possible d'établir une liste des variables les plus discriminantes. La sélection aléatoire des variables revient donc à tirer q facteurs aléatoirement, avec q_1 facteurs discriminants et q - q_1 facteurs non discriminants. Ainsi, la probabilité de tirer k variables discriminantes relève de la loi hypergéométrique. On peut alors imaginer un critère à imposer sur q pour qu'on ait une probabilité de sélectionner un facteur discriminant supérieur à un certain seuil.

Le cadre de l'étude est un peu particulier, dans le sens où la variable tauxHoraire, renseignant également si le sinistre est VEI ou non, est beaucoup plus discriminante que toutes les autres. Il a été observé empiriquement que la non inclusion de cette variable au modèle perturbait fortement les prédictions. C'est pourquoi la solution qui a été choisie pour traiter le cas des facteurs discriminants est de forcer l'apparition de la variable tauxHoraire parmi les variables sélectionnées aléatoirement.

II.c.3. Méthode de sélection des variables

Etant donné l'ajout d'un nombre important de variables dans le modèle via les termes d'interactions, il est nécessaire de revoir la méthode de sélection des variables. En effet, les termes d'interaction représentent $\binom{q}{2}$ variables supplémentaires. Par exemple,

lorsque q = 10, le nombre total de variables s'élève à 55. Le problème de la sélection par le critère AIC est qu'à chaque étape, il est nécessaire d'ajuster autant de GLM qu'il y a de variables afin de trouver le retrait qui conduit à la meilleure amélioration de l'AIC. Etant donné que le volume de données est très important, le temps de calcul d'un GLM est non négligeable et augmenter le nombre de variables aussi fortement rend la méthode peu maniable.

Ainsi, on introduit un terme de pénalisation au GLM qui permet de limiter le nombre de paramètres inclus dans le modèle final. On entre ainsi dans le cadre de la régression pénalisée, qui trouve son utilité en actuariat, même si peu utilisée, comme moyen de modéliser des comportements avec une structure particulière comme le suggère le papier de Hugh Miller sur le sujet (Miller, 2015). En pratique, c'est le package H2O (Nykodym, 2015) qui a été utilisé pour implémenter cette pénalisation en vertu de sa flexibilité.
La régression pénalisée se traduit par l'ajout d'une contrainte sur les coefficients estimés :

$$\sum_{j=1}^{p} |\beta_j|^{\delta} \leq C,$$

ou de manière équivalente, l'ajout de la pénalisation suivante lors de la maximisation de la log vraisemblance :

$$\lambda \sum_{j=1}^{p} |\beta_j|^{\delta}.$$

Les valeurs classiques pour δ sont 1 (régression LASSO) ou 2 (régression RIDGE). Lambda quant à lui est le paramètre qui induit l'importance accordée à la pénalisation. En particulier,

- $\lambda = 0$ conduit à la solution obtenue dans le cadre d'un GLM classique sans pénalisation ;
- $\lambda \to +\infty$ conduit à des coefficients β_j tous nuls.

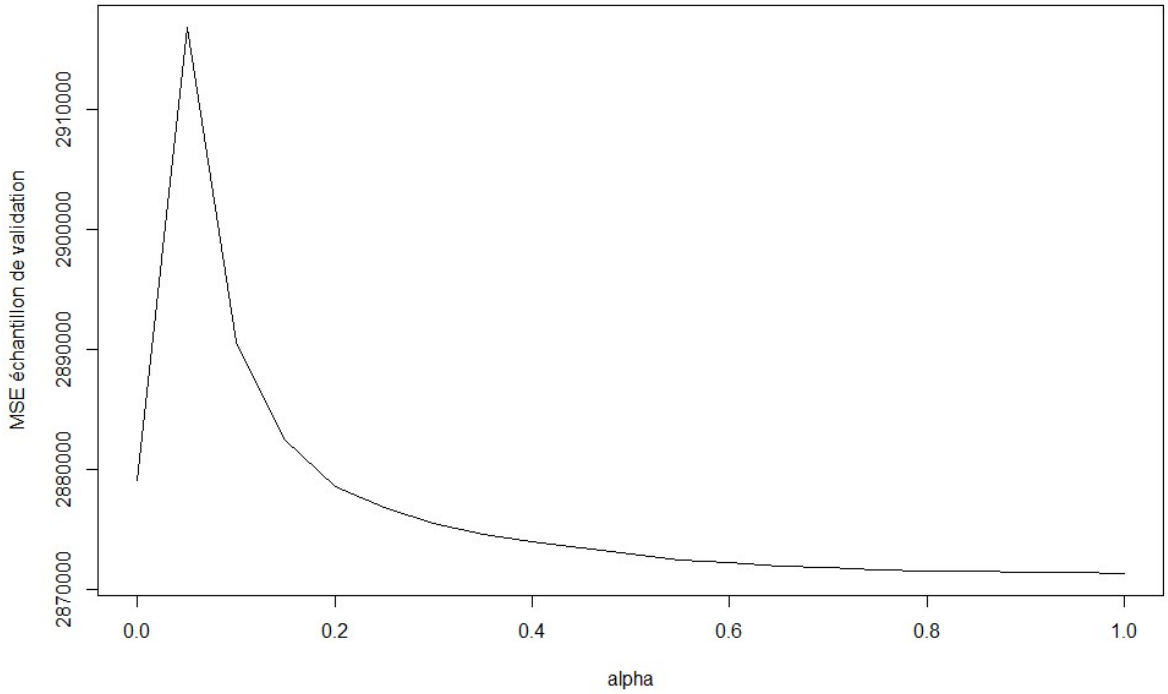
La valeur de λ qui sera retenue est celle qui minimise l'erreur de prédiction en utilisant la validation croisée sur l'échantillon de validation et l'échantillon d'apprentissage.

Pour avoir un modèle encore plus flexible, il est possible d'introduire un paramètre α afin d'utiliser une pénalisation formée du mélange de la régression RIDGE et LASSO :

$$\lambda(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} |\beta_j|^2).$$

On retrouve pour $\alpha = 0$ la pénalisation RIDGE et pour $\alpha = 1$ la pénalisation LASSO. Comme α est un paramètre du modèle, il est nécessaire de tester plusieurs valeurs pour savoir laquelle est optimale. Pour cela, on se réfère à l'erreur sur l'échantillon de validation d'une régression pénalisée avec une loi gamma et fonction de lien log :

Fig. 22 : L'erreur MSE de la régression pénalisée selon alpha



On constate ainsi que la pénalisation optimale est celle de LASSO. Ce résultat est cohérent avec l'objectif souhaité de la pénalisation des coefficients : il est connu que, plutôt que de simplement réduire leurs valeurs, la régression LASSO annule rapidement les coefficients des facteurs les moins discriminants, d'où son utilité lors de la sélection de variables.

Ainsi, la sélection de variable est constituée de l'ajout de la pénalisation LASSO au GLM initial, dont l'impact λ est estimé par validation croisée.

II.c.4. Agrégation des prédictions

Dans l'algorithme initial, une fois les modèles ajustés sur tous les échantillons bootstrapés, la prédiction finale se calcule comme la moyenne des prédictions de ceux-ci. Toujours dans une optique d'amélioration du pouvoir prédictif du modèle, il est intéressant d'optimiser cette partie de l'algorithme en attribuant un poids plus important aux modèles les plus pertinents.

Afin de déterminer la pertinence, on utilise un critère typique des forêts aléatoires : l'out-of-bag error (OOB error). Le principe de cette mesure est de calculer l'erreur de prédiction du modèle sur les observations écartées par la méthode bootstrap. En effet, comme nous avons vu précédemment, l'échantillon bootstrap est obtenu en tirant aléatoirement avec remise les lignes de l'échantillon principal. Cela signifie que certaines observations figurent dans l'échantillon principal mais pas dans l'échantillon bootstrap. Le nombre d'observations écartées est estimé à environ $\frac{1}{e}$ (37 %) de l'échantillon principal. L'OOB constitue ainsi un échantillon de validation naturel permettant d'évaluer la précision de modèle. La raison pour laquelle l'OOB n'a pas été utilisé précédemment est la nécessité de rester sur le même périmètre, afin de comparer les modèles entre eux. Ici, le concept nous est utile pour comparer le pouvoir prédictif de chaque GLM dans leurs échantillons bootstrappés respectifs.

La prédiction finale du RGLM était calculée de la façon suivante :

$$\hat{Y} = \frac{1}{100}\sum_{i=1}^{100} \hat{Y}^i$$

Avec $\hat{Y}^i$ la prédiction du GLM sur l'échantillon bootstrappé i. On cherche donc des poids p_i, indexés sur l'OOB error, tels que :

$$\hat{Y} = \sum_{i=1}^{100} p_i * \hat{Y}^i \quad et \sum_{i=1}^{100} p_i = 1.$$

Ainsi, en notant $OOBe_i$ l'out-of-bag error du glm sur l'échantillon boostrappé i, on pose :

$$p_i = \frac{1/OOBe_i}{\sum_{i=1}^{100}(\frac{1}{OOBe_i})}$$

Le vecteur ainsi créé est bien un vecteur de poids dont la somme vaut 1. De plus, chaque poids est inversement proportionnel à l'erreur sur l'out-of-bag afin de donner plus d'importance aux modèles ayant les erreurs les plus faibles.

II.c.5. Optimisation du nombre q de variables sélectionnées aléatoirement

Pour optimiser q, la même méthode que précédemment est appliquée, à savoir la comparaison de l'erreur quadratique moyenne sur l'échantillon de validation. Il est à noter que l'erreur par segment n'a pas pu être calculée efficacement pour ce modèle, étant donné que cela nécessite la création de sous-tableaux. Le package H2O est

relativement peu permissif sur ce point ce qui entraîne une explosion du temps de calcul. On se contente alors du coefficient de corrélation et de la norme L1 comme mesure d'erreur d'appui.

Tab. 9. Erreurs du modèle sur l'échantillon de validation en fonction des valeurs de q

q	COR	MSE	L1
3	0,59	3 479 941	786
4	0,61	3 334 662	772
5	0,64	3 155 130	758
6	0,66	3 053 032	747
7	0,65	2 959 632	735
8	0,66	2 931 436	729
9	**0,66**	**2 907 608**	**722**

De la même manière, il semble que le modèle optimal ne nécessite pas de sélection aléatoire sur les variables. Ainsi, l'agrégation perd une grande partie de son intérêt, bien que l'ajout des modifications semble avoir amélioré substantiellement le modèle.

Lin Song *et al.* (Song L., 2013) ont proposé une formule permettant de choisir le nombre q de variables sélectionnées aléatoirement. Ils ont estimé que sans terme d'interaction, la sélection aléatoire était inutile sur un tableau de données comportant moins de 10 variables explicatives, ce qui correspond aux résultats obtenus précédemment. Cependant, en appliquant la formule proposée et en incluant les termes d'interactions de second ordre, le nombre optimal de variables sélectionnées aléatoirement devrait être égal à 8. Evidemment, ce résultat dépend de la structure des données, mais aussi des ajustements de la méthode qui ont été faits. C'est pourquoi dans notre cas, 9 reste le meilleur choix.

Sans doute que l'ajout des termes d'interactions de troisième ordre aurait permis de valoriser la sélection aléatoire de variables. Néanmoins l'algorithme utilisé jusqu'ici ne permet pas l'implémentation clé en main de ceux-ci. La mémoire de l'ordinateur sature et le calcul n'aboutit pas. Pour sa bonne exécution, il serait nécessaire d'essayer des techniques plus avancées de parallélisation, étant donné que celles utilisées ici ne suffisent pas.

II.c.6. Résultats du RGLM amélioré

Comme modèle de prédiction, le RGLM amélioré semble être un modèle plus efficace que le GLM et le RGLM simple, comme le montre le tableau de résultats suivant :

Tab. 10. Comparaison des erreurs de prédictions sur l'échantillon test

	COR	MSE	L1
GLM	0,55	3 639 942	770
RGLM simple	0,55	3 639 483	770
RGLM amélioré	**0,66**	**2 903 323**	**716**

On constate ainsi une nette amélioration des résultats du modèle avec les modifications apportées, avec une diminution de l'erreur quadratique moyenne de plus de 20 %.

Néanmoins, la méthode de sélection de variables n'ayant pas été concluante, une méthode alternative d'agrégation est essayée pour s'assurer de l'optimisation de ce résultat.

II.d. Méthode alternative d'agrégation : le meilleur modèle

Il a été observé lors de l'optimisation du nombre de variables sélectionnées aléatoirement que prendre q = 9 est significativement meilleur que prendre une autre valeur.

Initialement, la sélection aléatoire de variables est introduite afin de pouvoir tester des combinaisons de variables écartées par une méthode classique de sélection de variables un peu trop rigide. A cela, on ajoute un apprentissage sur un échantillon boostrappé, pour créer une situation similaire mais légèrement différente de celles de nos données, afin d'éviter une construction déterministe du modèle autour de l'échantillon d'apprentissage.

Dans ce sens, il peut être intéressant de revoir la méthode d'agrégation des modèles. Une autre manière de voir les choses est de simplement prendre le meilleur modèle parmi ceux qui ont été testés, au sens de l'erreur MSE sur l'échantillon de validation, pour chaque valeur de q.

Etant donné qu'on ne retient que le meilleur modèle parmi ceux testés, il n'apparaît plus pertinent de forcer l'apparition de la variable tauxHoraire. Bien qu'il semble difficile de concevoir que le meilleur modèle n'utilise pas cette variable, il n'est pas nécessaire de réduire le champ de possibilités à ce stade.

Ainsi, pour chaque valeur de q entre 3 et 9, 50 modèles ont été construits sur des échantillons bootstrappés. On reporte alors l'erreur de validation des meilleurs modèles :

Tab. 11. Erreurs de validation des meilleurs modèles en fonction de q

q	3	4	5	6	7	8	9
Erreur du meilleur modèle	3 135 498	3 495 148	3 005 071	2 970 974	2 921 903	2 910 419	**2 908 015**
Erreur moyenne	4 564 704	4 229 489	4 238 714	3 794 046	3 450 857	3 089 025	**2 914 018**

Bien que l'erreur moyenne ne soit pas égale à l'erreur de la moyenne des prédictions (par rapport à ce qui a pu être fait auparavant lors de l'agrégation de modèle), on se rend compte grâce à ce tableau que, parmi les nombreux modèles de mauvaise qualité qui ont pu être construits, il existe un modèle dont les performances sont correctes. C'est un résultat intéressant, qui finalement montre qu'on ne perd pas tant de performance en réduisant significativement le nombre de variables. On imagine ainsi qu'avec un grand nombre de variables, cette méthode d'agrégation peut être d'une grande utilité.

On remarque que pour q = 8, les performances sont très proches de l'optimum. A tel point que lorsqu'on compare les erreurs sur l'échantillon test, on s'aperçoit que ce dernier est plus performant :

Tab. 12. Mesures sur l'échantillon test des meilleurs modèles en fonction de q

q	COR	MSE	L1
3	0,62	3 135 499	759
4	0,62	3 141 007	758
5	0,64	2 993 409	726
6	0,65	2 955 974	728
7	0,66	2 909 052	721
8	**0,66**	**2 900 468**	**716**
9	0,66	2 902 814	717

Pour q entre 7 et 9, les résultats sont très similaires, avec une légère supériorité pour q = 8. Il semble qu'à travers cette méthode, il a été possible de montrer que la sélection aléatoire des variables peut permettre dans certains cas d'obtenir un modèle prédictif de meilleure qualité.

Entre autres, l'avantage de cette méthode d'agrégation est que seul un modèle est conservé, et donc les coefficients par variable sont exploitables. Néanmoins, il n'est toujours pas possible d'estimer précisément l'écart de performance des réseaux d'experts à cause des variables d'interactions.

On actualise ainsi le tableau de comparaison des modèles, en distinguant le RGLM avec cette nouvelle méthode d'agrégation :

Tab. 13. Comparaison des erreurs de prédiction sur l'échantillon test

	COR	MSE	L1
GLM	0,55	3 639 942	770
RGLM simple	0,55	3 639 483	770
RGLM amélioré	0,66	2 903 323	716
RGLM méthode "meilleur modèle"	**0,66**	**2 900 468**	**716**

II.e. Conclusion

L'objectif du RGLM était d'introduire des techniques d'apprentissages statistiques dans un modèle type GLM, afin d'améliorer son pouvoir prédictif.

Dans un premier temps, le RGLM a été utilisé tel quel, avec des paramètres par défaut, et les résultats produits de cette manière n'ont pas amélioré ceux du GLM.

Dans une deuxième partie, le RGLM a été amélioré, afin de mieux s'adapter à la structure particulière des données. En effet, le faible nombre de colonnes compromet la phase de sélection aléatoire des variables. Pour pallier à ça, les termes d'interactions entre les variables ont été ajoutés au modèle, pour augmenter artificiellement le nombre de variables du modèle. La méthode de sélection des variables a alors été modifiée, en utilisant une régression pénalisée de type LASSO plutôt qu'un critère sur l'AIC, favorisant une sélection plus rapide qui était nécessaire suite à la complexification du modèle. L'agrégation des prédictions a également été revue, en donnant plus de poids aux modèles de bonne qualité. Dans la même optique, l'apparition de la variable « tauxHoraire » a été forcée dans tous les modèles afin d'assurer une qualité de prédiction minimum.

On constate que de cette façon, les prédictions du modèle sont de bien meilleure qualité que le GLM. Néanmoins, la sélection aléatoire de variables n'apparaît pas à son plein potentiel puisque son optimisation recommande un modèle avec toutes les variables. En revanche, en modifiant la méthode d'agrégation, qui désormais garde uniquement le meilleur modèle testé, ce paramètre regagne de l'intérêt. Le modèle ayant les meilleures prédictions sur l'échantillon test possède une sélection de 8 variables aléatoires, et est plus performant qu'avec l'ancienne méthode d'agrégation.

Finalement, le RGLM, en tant que méthode d'agrégation de modèle, a permis d'améliorer substantiellement le pouvoir prédictif du GLM. La qualité d'interprétation et d'explication reste bonne lorsqu'on utilise la méthode d'agrégation « meilleur modèle ». Néanmoins, il n'est plus possible d'estimer clairement l'influence de la variable réseaux d'experts, étant donné que celle-ci est distribuée à travers toutes les variables d'interactions.

III. STOCHASTIC GRADIENT BOOSTING

Pour une description assez complète du boosting, il est possible de se référer aux travaux de S. Tufféry, issue de son livre « Modélisation prédictive et Apprentissage statistique avec R » (Tufféry, 2015).

III.a. Méthode du boosting

III.a.1. Idée générale du boosting

Suivant l'auteur, il est possible de trouver deux explications différentes quant au procédé du boosting : l'une repose sur la correction des erreurs de prédiction, l'autre sur la minimisation d'une fonction de perte. En réalité, ces deux approches sont équivalentes.

Plus généralement, le boosting est une méthode d'apprentissage automatique qui consiste à appliquer itérativement un algorithme d'apprentissage dit « faible » dans le but de corriger ses erreurs.

Concrètement, on ajuste un premier modèle m_1 sur les données, qui va générer une erreur ε_1. Cette erreur peut être vue comme la différence entre les résultats du modèle et les observations au sens de la fonction de perte quadratique, d'où le rapprochement entre les deux explications précitées.

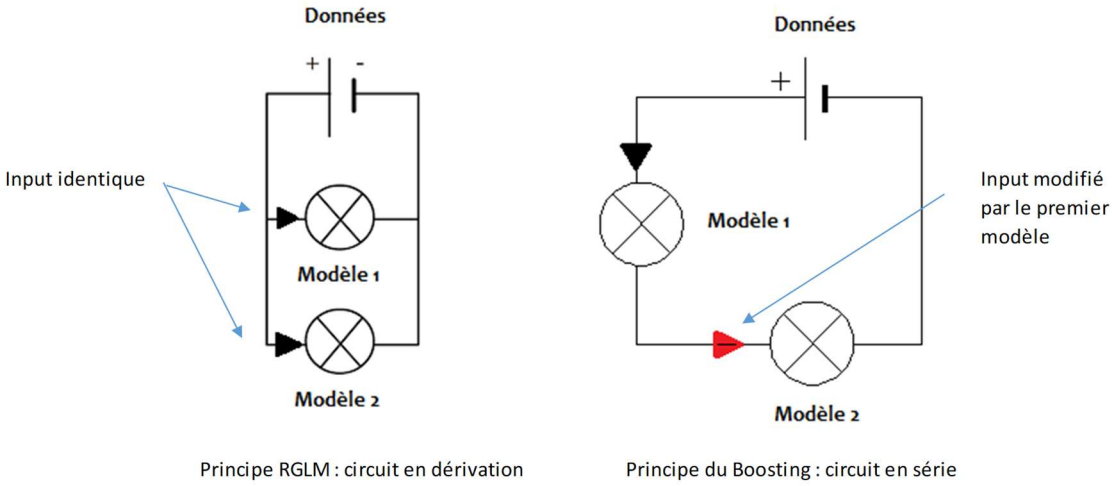
Afin d'améliorer ce modèle, on en ajuste un second m_2 sur ε_1, qui va ainsi générer une erreur ε_2, et ainsi de suite. Ainsi, au bout de T itérations, le modèle devient :

$$m(x) = m_1(x) + m_2(x) + \cdots + m_T(x)$$

La notion de « weak learner » (apprentissage faible) est essentielle ici : étant donné qu'on répète l'algorithme plusieurs fois, il est important que la classification faite ne soit pas trop discriminante afin d'éviter le sur-apprentissage.

Ainsi, le boosting présente une différence de fond importante avec le RGLM : les modèles se succèdent afin d'améliorer les précédents, tandis que le RGLM associe des GLM construits indépendamment les uns des autres. A titre illustratif, on retrouve cette différence entre les circuits en série et les circuits en dérivation :

Fig. 23 : Analogie avec les circuits électriques



Principe RGLM : circuit en dérivation — Principe du Boosting : circuit en série

Il existe de nombreuses versions différentes du boosting, notamment le célèbre algorithme Adaboost présenté par Yoav Freund et Robert E. Schapire (Freund, 1996). Les variations du boosting se situent notamment sur l'algorithme d'apprentissage et la fonction de perte utilisés.

III.a.2. Le Stochastic Gradient Boosting

La méthode de boosting retenue pour cette étude est le Gradient Boosting Model, qui utilise une fonction de perte quadratique associée à la méthode de descente de gradient.

L'objectif du GBM est le même que pour tout algorithme d'apprentissage : minimiser la fonction de perte quadratique

$$f(x) = \underset{\varphi \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \varphi(x_i)\big)^2 = \underset{\varphi \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, \varphi(x_i)\big).$$

En notant $\hat{f}$ l'estimation de f, on présente ici l'algorithme du Stochastic Gradient Boosting Model :

- Initialisation : $\forall i \in [1,n], \hat{f}(x_i) = \frac{1}{n}\sum_{j=1}^{n} y_j$.
- Boucle comportant T itérations :
  - Etape 1 : calcul du gradient. $z_i = \frac{\partial}{\partial \rho} L(y_i, \rho)|_{\rho=\hat{f}(x_i)}$;
  - Etape 2 : Tirage sans remise d'un sous-échantillon de taille p*n de l'ensemble des données, p étant un pourcentage ;
  - Etape 3 : Prédire le gradient $z = (z_1, \ldots, z_N)$ sachant x à partir d'un arbre de régression de profondeur maximum K. On note $g(x) = E[z|x]$;
  - Etape 4 : Actualiser la valeur de $\hat{f}(x_i)$, uniquement pour les observations sélectionnées à l'étape 2 : $\hat{f}(x_i) \leftarrow \hat{f}(x_i) + \lambda * g(x_i)$.

En input du modèle, on retrouve les paramètres suivants :

- Le nombre d'itérations T ;
- La proportion p d'observations sélectionnées aléatoirement ;
- La profondeur maximale des arbres de régression K ;
- Le taux d'apprentissage λ, également appelé le « shrinkage ».

L'algorithme utilisé est donc une adaptation de l'algorithme de descente de gradient auquel certains ajouts ont été effectués :

Tirage sans remise d'un sous-échantillon de taille N :

C'est cet élément qui distingue le Gradient Boosting du Stochastic Gradient Boosting. A l'origine, l'idée d'implémenter un facteur aléatoire dans l'estimation a été proposée par Breiman, en 1996, afin d'améliorer les performances du Bagging. L'idée a ensuite évolué, jusqu'à inclure un sous-échantillonnement aléatoire de l'échantillon d'apprentissage dans l'algorithme du boosting, comme présenté dans l'article de Jérome H. Friedman « Stochastic Gradient Boosting » (Friedman, 1999). Dans cet article, plusieurs cas ont été traités et ont montré la pertinence de ce paramètre.

Arbre de régression :

L'arbre de régression est modifié dans son algorithme afin de favoriser les calculs sur de grands volumes de données. Le résultat renvoyé est néanmoins identique à un arbre de régression classique.

Taux d'apprentissage λ :

Dans une descente de gradient classique, le pas est calculé à chaque itération afin d'optimiser le temps nécessaire à approcher la solution. Ici, le pas est fixe pour toutes les itérations et vaut λ. La raison de ce changement est que le boosting se veut un apprentissage lent, afin d'éviter tout problème lié au sur-apprentissage.

III.a.3. Implémentation de l'algorithme

Il s'est avéré que le volume de données dont nous disposons est trop important pour utiliser le package classique « gbm » sous R. Il a été nécessaire de chercher une alternative avec un algorithme plus optimisé en termes de gestion de la mémoire et permettant une parallélisation poussée des calculs.

Pour cela, nous nous sommes tournés vers le package GBM du module H2O sous R. Le module H2O propose un format qui permet une compression des tableaux de données optimisée pour le calcul, ce qui aide grandement pour la gestion de la mémoire. En plus

de cela, le package GBM de ce module est optimisé pour les calculs sur de grands volumes de données.

En effet, ce package utilise une version améliorée de l'algorithme des arbres de régression, en y ajoutant les composantes du MapReduce. Le MapReduce est un modèle de programmation, élaboré par Google, qui permet de distribuer efficacement les calculs entre différents clusters. Il se révèle être un outil de parallélisation très efficace, constituant ainsi une réponse intéressante aux problématiques de volume du Big Data. Pour plus d'informations sur le MapReduce, il est possible de se référer à l'article de J. Dean et S. Ghemawat (Ghemawat, 2004). Concrètement, l'amélioration notable sur la construction des arbres se situe dans la réduction du nombre de passages dans la base de données. Par exemple, seuls n passages suffisent à la construction d'un arbre de profondeur n.

III.a.4. Influence des variables explicatives

Le gradient boosting propose une fonctionnalité intéressante qu'est l'évaluation de l'importance relative des variables explicatives, dont on peut retrouver le développement dans l'article de J. H. Friedman (Friedman, 2001).

Le raisonnement est le suivant : grâce à la construction du GBM, on dispose désormais d'une fonction $\hat{f}(x)$ capable de prédire un montant d'expertise en fonction des caractéristiques du sinistre. Breiman, Friedman, Olshen et Stone ont proposé une formule pour les arbres de régressions (Breiman, 1983) :

$$\hat{I}_j^2(A) = \sum_{t=1}^{J} i_t^2 1_{v_t=j}$$

Avec :

- J le nombre de nœuds non terminaux de l'arbre A,
- v_t la variable de décision servant à la division du nœud t,
- i_t^2 est le critère d'amélioration des moindres carrés résultant de cette division.

Plus précisément, lorsqu'une région R est subdivisée en deux sous-régions R_1 et R_2, on écrit :

$$i^2(R_1, R_2) = \frac{w_1 w_2}{w_1 + w_2}(\overline{y_1} - \overline{y_2})^2,$$

où $\bar{y}$ la moyenne de la variable réponse et w est le poids de la région correspondante.

Pour généraliser cette idée au boosting, on prend simplement la moyenne de la valeur des $\hat{I}_j^2(A)$ sur tous les arbres constitués :

$$\widetilde{I_J}^2 = \frac{1}{T}\sum_{m=1}^{T} \hat{I}_j^2(A_m).$$

Concrètement, on obtient une mesure $\widetilde{I_J}$ de l'importance de la variable j, qui est fonction du nombre de fois où cette variable est impliquée dans la division d'un nœud, pondérée par l'amélioration que celle-ci apporte à l'arbre.

III.b. Paramétrage du modèle

III.b.1. Retraitements utilisés

Lors des préparatifs de l'ajustement d'un modèle GLM, il a été nécessaire de réduire significativement le nombre de modalités des variables marques et département. En effet, un coefficient était estimé pour chaque modalité de chaque variable.

A la différence d'un modèle de la classe des GLM, le Gradient Boosting est un algorithme de classification. Autrement dit, il segmente la population en plusieurs classes homogènes de risques. Le Gradient Boosting a une réputation de modélisation « off-the-shelf » : c'est un modèle qui nécessite très peu de préparation pour sortir un premier résultat. La raison est qu'il sait gérer les valeurs manquantes et non retraitées, de par sa structure en arbre. Ce n'est pas le cas d'un modèle type GLM qui a besoin pour son bon fonctionnement de vérifier la colinéarité des variables explicatives, de limiter le nombre de modalités ou encore de retraiter les valeurs manquantes.

De ce fait, le retraitement utilisé pour l'application du Gradient Boosting se limite donc à celui présenté dans le chapitre 1. Les réductions de modalités introduites dans la partie sur le GLM ne semblent pas pertinentes pour ce type de modélisation. Cependant, les retraitements précédents, notamment celui des valeurs manquantes, sont conservés. En effet, il est jugé que l'information perdue lors de ces retraitements est de second ordre, et permet dans l'ensemble de mieux maîtriser les domaines d'évolution des variables.

III.b.2. Première application du boosting

L'idée de cette partie est d'appliquer le modèle une première fois avec des paramètres par défaut pour ensuite étudier l'écart de performance avec le modèle optimisé.

Pour initialiser les paramètres, il est nécessaire de se baser sur certains critères issus de l'expérience en observant les contraintes que doivent respecter le modèle. Une première contrainte est la complexité du modèle. Il faut que l'algorithme puisse s'exécuter en un temps raisonnable tout en donnant un résultat de bonne qualité. La complexité est en grande partie contrôlée par le nombre d'itérations T effectuées par l'algorithme, mais elle

est également affectée par un coefficient d'apprentissage trop faible ou par le nombre de nœuds terminaux trop important.

Une autre contrainte importante du modèle est le surapprentissage. De par sa structure, l'algorithme va naturellement coller à l'échantillon d'apprentissage au fur et à mesure des itérations. Il est donc nécessaire de définir un critère d'arrêt, associé à un apprentissage lent afin que chaque itération n'apporte qu'une légère modification du modèle précédent. Par ce procédé il est plus facile de contrôler l'évolution du modèle et ainsi avoir une mesure plus précise du nombre d'itérations optimales à effectuer. Il est possible de ralentir l'apprentissage de l'algorithme en diminuant le coefficient d'apprentissage λ et la profondeur maximale K des arbres de classification.

Le critère choisi pour l'optimisation du nombre d'itération est l'erreur moyenne quadratique sur l'échantillon de validation. Plus spécifiquement, si l'erreur sur l'échantillon de validation n'augmente plus après 50 itérations, alors le modèle s'arrête.

Les paramètres par défaut que nous utilisons sont les suivants :

- Un nombre d'itérations maximal à 10 000 ;
- Un coefficient d'apprentissage fixé à 0,1 ;
- Une profondeur des arbres maximale fixée à 5 ;
- Un tirage sans remise de 50 % de l'échantillon d'apprentissage.

Il est désormais possible d'avoir accès aux premiers résultats du GBM. Le nombre d'arbres optimal au regard du critère d'erreur retenu est de 2822.

Les résultats du GBM simple sont présentés ci-dessous, avec en comparaison les modèles utilisés jusqu'ici :

Tab. 14. Comparaison des erreurs de prédictions sur l'échantillon test

	COR	MSE	L1
GLM	0,55	3 639 942	770
RGLM simple	0,55	3 639 483	770
RGLM amélioré	0,66	2 903 323	716
RGLM méthode "meilleur modèle"	0,66	2 900 468	716
GBM simple	**0,72**	**2 477 139**	**661**

On constate ainsi une très nette amélioration de l'erreur quadratique moyenne sur l'échantillon test avec le GBM simple. Les autres mesures d'erreurs vont également dans ce sens. Ces résultats démontrent la puissance du GBM comme modèle prédictif, surtout en tenant compte du fait que ce modèle n'a été ajusté qu'avec des paramètres choisis par défaut.

Pour aller plus loin, il est intéressant de comparer les résultats en modifiant la méthode d'échantillonnage. Cette comparaison est motivée par le papier de G. Ridgeway, « Generalized Boosted Models » (Ridgeway, 2007), qui annonce un résultat significativement meilleur lorsqu'une validation croisée à 5 échantillons est effectuée. Le principe est le suivant : réunir l'échantillon de validation et d'apprentissage en un seul, diviser cet échantillon en 5 et tour à tour chaque échantillon servira d'échantillon de validation tandis que les 4 autres seront utilisés comme échantillon d'apprentissage. Le critère d'arrêt utilisé est le même que précédemment.

Il est à noter que cette méthode possède un désavantage majeur : le temps de calcul. A titre d'exemple, le calcul standard du GBM avec un échantillon d'apprentissage et de validation prend 1 minute et 31 secondes, tandis que la validation croisée à 5 échantillons prend 37 minutes et 38 secondes pour compléter le calcul. On reste ici dans des ordres de grandeur acceptables, en partie grâce aux bonnes performances de la machine utilisée pour ces calculs (voir annexe). Néanmoins, il est nécessaire de garder ces chiffres en tête pour une application ultérieure de ce modèle : la validation croisée à 5 échantillons ne sera peut-être pas utilisable en pratique.
Comme escompté, la validation croisée améliore la qualité des prédictions sur l'échantillon test :

Tab. 15. Erreur de prédiction sur l'échantillon test du GBM simple avec validation croisée

	COR	MSE	L1
GBM simple CV	0,72	2 460 120	659

L'erreur quadratique moyenne a diminué grâce à cette méthode d'environ 0,7 %. Cela peut sembler dérisoire lorsqu'on voit les écarts entre les différents modèles, mais en termes d'optimisation du modèle cette diminution est non négligeable, d'autant plus que les paramètres n'ont pas été modifiés. Il est à noter que les effets aléatoires des deux modèles ont été fixés afin de pouvoir comparer objectivement le pouvoir prédictif de chacun, uniquement basé sur les différences de méthode.

Il était précisé auparavant que la modification d'un paramètre devait être justifiée par une amélioration de l'erreur de validation, sans utiliser l'échantillon test afin de garder une base de comparaison objective. Ici le cas est un peu particulier, puisque c'est justement la méthode d'échantillonnage qui est remise en question, et l'échantillon d'apprentissage n'est plus commun aux deux approches. On peut noter néanmoins que pour la validation croisée, l'erreur sur les échantillons de validation est en moyenne inférieure à celle du GBM simple (2 462 319 contre 2 468 8115). Il peut être utile pour appuyer son choix de comparer les deux méthodes sur d'autres bases de données, comme l'a fait G. Ridgeway (Ridgeway, 2007), afin de sélectionner la méthode d'échantillonnage la plus performante.

En ce qui concerne cette étude, la validation croisée ne sera utilisée que pour la production du résultat final, afin de gagner du temps sur l'optimisation des paramètres qui est déjà lourde en temps de calcul.

III.b.3. Optimisation des paramètres

III.b.3.1. Recherche aléatoire sur une grille

L'optimisation des paramètres dans un Stochastic Gradient Boosting Model n'est pas chose aisée pour deux raisons :

- La plupart des paramètres sont interdépendants : ajuster la valeur de l'un va modifier la valeur optimale de l'autre.
- Le nombre de paramètres du modèle est très élevé. Nous en avons présenté les principaux mais il en existe d'autres.

La méthode utilisée ici est une recherche par grille. Des bornes sont proposées pour chaque paramètre, créant ainsi de nombreuses combinaisons possibles à l'aide de boucles imbriquées. Malgré la parallélisation des calculs, cette méthode s'avère très gourmande en mémoire et en temps de calculs. Pour pallier au problème du temps de calculs, une recherche aléatoire est utilisée dans la grille préétablie. Ce procédé a été théorisé par J. Bergstra et Y. Bengio dans leur article « Random Search for Hyper-Parameter Optimization » (Bergstra, 2012). L'argument clé est, qu'à temps de calcul équivalent, la recherche aléatoire permet de trouver des modèles performants sur de plus grand espace que ne le fait une recherche par grille classique.

D'autre part, J. Bergstra et Y. Bengio ont également constaté que seuls certains paramètres ont un réel impact sur les performances de l'algorithme. Ils ont également précisé que ces paramètres changent en fonction de la base de données d'étude, ce qui est assez problématique lors d'une recherche simple sur une grille puisque beaucoup de temps est perdu lors de la distinction de cas d'un paramètre non influent. La recherche aléatoire sur une grille permet de pallier à ce genre de problème : pour peu que les bornes soient correctement choisies, un paramètre peu influant n'aura que peu d'impact sur le résultat, comme sur la complexité.

Pour terminer sur la méthode de la recherche aléatoire, il est nécessaire de définir un critère d'arrêt, soit en termes de temps de calculs, soit en termes de modèles calculés. L'objectif est évidemment de tester un maximum de combinaisons possibles afin de s'approcher au mieux de la solution optimale. On verra l'influence de ce paramètre par la suite.

III.b.3.2. Introduction des nouveaux paramètres et recherche de bornes

Certains paramètres n'ont pas été présentés auparavant car ils jouent un rôle secondaire dans l'algorithme et sont en général propres au degré de complexité de celui-ci. Utilisant une recherche aléatoire sur une grille, l'ajout d'un paramètre n'augmente pas la complexité de la recherche, c'est pourquoi un maximum de paramètres est ajouté à la grille, même ceux pouvant sembler peu influants.

Pour déterminer les bornes d'un paramètre, l'idée est de faire varier sa valeur tout en fixant le reste des paramètres. Le paramétrage de base utilisé est celui de la première application du modèle de la partie précédente, et l'outil de comparaison est l'erreur moyenne quadratique sur l'échantillon de validation.

C'est une première approche qui permet d'avoir une idée sur les valeurs intéressantes des paramètres peu connus d'expérience.

- Le pas de discrétisation de l'histogramme de chaque variable.

Lors de la construction de l'arbre de régression, il est nécessaire de trouver la règle de division optimale de chaque nœud. Celle-ci se base sur la construction d'un histogramme par variable, dont le pas de discrétisation dépend du nombre de facteurs de subdivision de la variable. Plus précisément, chaque variable quantitative est transformée en variable qualitative dont le nombre de facteurs est un paramètre du modèle.

Ainsi, pour chaque variable, chaque niveau de l'histogramme est étudié comme potentielle point de rupture. Plus le nombre de facteurs est grand, plus il y aura de possibilité à étudier, donc une complexité accrue, mais plus le point de rupture sera précis. Par défaut, ce paramètre est fixé à 20, et il en existe un pour les variables continues et un pour les variables qualitatives.

D'autres valeurs pour cette variable ont été testées, mais on peut constater que cette variable est peu influente, étant donné qu'une forte variation de celle-ci n'a que peu d'impact sur l'erreur de validation.

Tab. 16. Classement des modèles en fonction du nombre de facteurs

```
Hyper-Parameter Search Summary: ordered by increasing MSE
  nbins     model_ids                mse
1   600  tst4_model_5 2460956.4962919084
2  1200 tst4_model_11  2461255.996887076
3  2300 tst4_model_22 2461436.4820192466
4  1300 tst4_model_12  2461605.436811448
5  2000 tst4_model_19  2462024.611166644

---
   nbins     model_ids                mse
25  1400 tst4_model_13 2466644.2651729384
26   400  tst4_model_3 2466715.3197135897
27   100  tst4_model_0 2467745.7158430503
28   200  tst4_model_1  2467933.303670869
29   300  tst4_model_2  2468867.228306636
30  2100 tst4_model_20  2468917.566456712
```

Il est difficile d'induire clairement une borne à partir de ce tableau, mais étant donné que le paramètre a peu d'influence il n'est pas trop risqué de prendre une borne assez large. Ainsi seront autorisées les valeurs entre 500 et 1 500.

Ce paramètre existe également pour les variables quantitatives. Lorsque le nombre de facteurs en paramètre est supérieur au nombre de facteurs de la variable, alors toutes les modalités de cette variable sont conservées. Si ce n'est pas le cas, certaines modalités sont réunies jusqu'à atteindre le nombre de facteurs maximal. Par défaut, le paramètre est fixé à 1024. Dans notre cas, seules deux variables ont un nombre de modalités significatif qui pourrait être impacté par cette variable : la marque et le département. Pour vérifier cela, une recherche a également été effectuée :

Tab. 17. Classement des modèles en fonction du nombre de facteurs des variables quantitatives

```
Hyper-Parameter Search Summary: ordered by increasing MSE
   nbins_cats     model_ids                mse
1         145 tst7_model_14 2468955.4637203906
2         135 tst7_model_13 2468955.4637203906
3         125 tst7_model_12 2468955.4637203906
4          95  tst7_model_9 2485608.2195481933
5          85  tst7_model_8 2486140.1874559238
6         105 tst7_model_10 2495436.1684483048
7         115 tst7_model_11  2498983.325693518
8          75  tst7_model_7  2506858.143222002
9          45  tst7_model_4 2513011.1970897047
10         55  tst7_model_5  2517306.668929746
11         65  tst7_model_6   2518193.58566017
12         35  tst7_model_3  2531334.554669817
13         15  tst7_model_1  2562345.351921983
14         25  tst7_model_2 2566606.0862876833
15          5  tst7_model_0 2592646.4941334166
```

On constate que garder tous les facteurs produit un modèle de meilleure qualité, avec un écart non négligeable. C'est pourquoi ce paramètre est laissé à sa valeur par défaut, soit 1 024, et n'est pas optimisé.

- La sélection aléatoire de colonnes.

A l'image de la proportion p d'observations sélectionnées aléatoirement, il existe un paramètre permettant de sélectionner aléatoirement les colonnes sans remise. Au début, ce paramètre avait été écarté compte tenu du faible nombre de variables, et donc de colonnes, dans la base de données. De plus, la sélection aléatoire de variables du RGLM n'avait pas été concluante, dont le principe est identique ici.

Ceci étant dit, après avoir testé plusieurs valeurs, on se rend compte qu'activer ce paramètre peut améliorer la qualité du modèle, comme nous le montre la recherche suivante :

Tab. 18. Classement des modèles en fonction de la proportion de colonnes aléatoirement sélectionnées

Hyper-Parameter Search Summary: ordered by increasing MSE

	col_sample_rate	model_ids	mse
1	0.7	tst2_model_6	2461312.801941031
2	0.6	tst2_model_5	2462625.647860262
3	0.9	tst2_model_8	2464138.3774377233
4	1.0	tst1_model_1	2464483.641023643
5	0.8	tst2_model_7	2464900.7149819
6	0.5	tst2_model_4	2465325.5582313132
7	0.4	tst2_model_3	2465448.9397485605
8	0.3	tst2_model_2	2489419.641988703
9	0.2	tst2_model_1	2551253.7582573243
10	0.1	tst2_model_0	2943008.984192401

Le paramètre est donc ajouté au modèle, avec comme bornes [0,4 ; 1]. Le choix des bornes est motivé par l'écart important entre une proportion de 0,4 et 0,3, mais aussi avec la volonté de garder un éventail de possibilités assez large.

Il est à noter que le paramétrage de la sélection aléatoire, de colonnes ou de lignes, est assez complet. Il existe 2 autres paramètres pour modifier la proportion : un en fonction de l'arbre, l'autre en fonction de la couche de l'arbre. Néanmoins ces paramètres n'ont pas été ajoutés.

- Le nombre minimal d'observations par feuille d'arbres.

Il est possible d'imposer un nombre minimal de lignes par feuille, qui est par défaut de 10. De la même manière que pour les autres paramètres, on effectue une recherche pour avoir une idée de la borne à utiliser :

Tab. 19. Classement des modèles en fonction du nombre minimum d'observations par feuille

Hyper-Parameter Search Summary: ordered by increasing MSE

	min_rows	model_ids	mse
1	1.0	tst8_model_0	2463673.1760332575
2	4.0	tst8_model_2	2465679.27901211
3	2.0	tst8_model_1	2466000.97939562
4	8.0	tst8_model_3	2466725.4086665967
5	64.0	tst8_model_6	2467637.66706955
6	16.0	tst8_model_4	2468700.087807552
7	32.0	tst8_model_5	2469946.979450339
8	128.0	tst8_model_7	2472806.888500427
9	256.0	tst8_model_8	2475270.3596672467
10	512.0	tst8_model_9	2490792.8111226005
11	1024.0	tst8_model_10	2513850.0334823374
12	2048.0	tst8_model_11	2592844.397771642
13	4096.0	tst8_model_12	2699128.1014049593
14	8192.0	tst8_model_13	2765946.186050879
15	16384.0	tst8_model_14	2916778.654927906
16	32768.0	tst8_model_15	3333588.378277494
17	65536.0	tst8_model_16	3588652.759667311
18	131072.0	tst8_model_17	4026740.1733577135

On trouve en première position le modèle avec aucune restriction sur le nombre d'observations par feuille. Ce résultat est quelque peu inhabituel puisque ce paramètre est censé être une barrière de dernier recours contre le sur-apprentissage, lorsque le critère d'arrêt de l'arbre ne s'est pas déclenché. On propose ainsi comme valeurs possibles {1,2,5,10,20}.

- La profondeur maximale K des arbres de régression.

La profondeur de l'arbre fait partie des paramètres dont l'impact est important, d'où la nécessité de réduire le plus possible l'écart entre les deux valeurs extrêmes de la borne. Habituellement, une profondeur maximale d'arbre entre 4 et 8 est en général suffisante. Néanmoins il a déjà été observé dans certains cas qu'il peut être intéressant d'augmenter ce paramètre significativement.

D'autre part, il est à noter que ce paramètre influe uniquement sur la profondeur maximale, et pas la profondeur de chaque arbre. Par exemple, il est observé que pour une profondeur maximale de 5, la profondeur moyenne des arbres est de 1,7 avec les paramètres utilisés dans la première application.

Une première recherche de l'optimisation de ce paramètre donne ces résultats :

Tab. 20. Classement des modèles en fonction de la profondeur maximale des arbres

Hyper-Parameter Search Summary: ordered by increasing MSE

	max_depth	model_ids	mse
1	7	tst9_model_3	2453745.5129406666
2	5	tst1_model_1	2464483.641023643
3	9	tst9_model_4	2469888.0528623206
4	11	tst9_model_5	2489161.3907460803
5	13	tst9_model_6	2517797.71237893
6	15	tst9_model_7	2551720.5181803526
7	3	tst9_model_1	2567492.064828534
8	17	tst9_model_8	2578343.529019054
9	19	tst9_model_9	2614029.5981157646
10	21	tst9_model_10	2669223.53028233
11	23	tst9_model_11	2697790.0841237144
12	25	tst9_model_12	2710025.7519092252
13	27	tst9_model_13	2722180.2770810165
14	29	tst9_model_14	2726140.5477014384
15	1	tst9_model_0	3782854.579756631

On se trouve donc dans un cas classique, où on va chercher une profondeur maximale entre 4 et 8.

- La proportion p d'observations aléatoirement sélectionnées.

Il est précisé que le tirage aléatoire est également sans remise.

Tab. 21. Classement des modèles en fonction de la proportion d'observations sélectionnées aléatoirement

Hyper-Parameter Search Summary: ordered by increasing MSE

	sample_rate	model_ids	mse
1	1.0	tst9_model_24	2444825.568001487
2	0.9	tst9_model_23	2448712.557520046
3	0.8	tst9_model_22	2449757.7044083984
4	0.7	tst9_model_21	2453610.684366147
5	0.6	tst9_model_20	2459561.335660451
6	0.5	tst9_model_19	2464483.641023643
7	0.4	tst9_model_18	2473068.0676725414
8	0.3	tst9_model_17	2482353.404918713
9	0.2	tst9_model_16	2494508.535548003
10	0.1	tst9_model_15	2526213.246870193

La même méthode que pour la sélection aléatoire des colonnes est appliquée, qui mène à chercher le paramètre optimal dans les bornes [0,5 ;1].

- Les coefficients d'apprentissages λ et λ_t.

On note une petite différence avec l'algorithme du Stochastic Gradient Boosting : l'introduction d'un facteur dégressif avec le nombre d'itérations λ_t. Plus précisément, la dernière étape qui change la valeur des prédictions $\hat{f}(x_i) \leftarrow \hat{f}(x_i) + \lambda * g(x_i)$ est modifiée par $\hat{f}(x_i) \leftarrow \hat{f}(x_i) + \lambda * \lambda_t * g(x_i)$, avec par exemple $\lambda_t = (0.995)^t$ et $\lambda = 0.1$.

L'idée de cet ajout est la suivante : on constate que le modèle est de meilleure qualité avec un coefficient d'apprentissage faible (0,001 par exemple), mais cela augmente grandement le nombre d'itérations nécessaire pour atteindre la solution. Le facteur dégressif λ_t permet de démarrer les premières itérations avec un coefficient d'apprentissage relativement élevé puis de le décroître rapidement, ce qui permet de travailler un coefficient d'apprentissage faible voir très faible pour la majorité des itérations tout en gardant une convergence assez rapide.

Cette proposition est d'autant plus pertinente lorsqu'on se rappelle du fonctionnement du GBM : le modèle affine les prédictions au fur et à mesure des itérations en modélisant l'écart entre les observations et les prédictions à partir des données. A la première étape, les prédictions sont initialisées à la moyenne, donc l'écart est important. En accélérant l'apprentissage dans les premières itérations, cela permet au modèle d'avoir rapidement une idée de l'ordre de grandeur et des variations, tout en lui permettant de s'affiner correctement en laissant un apprentissage lent par la suite.

Pour illustrer la pertinence du paramètre dégressif, voici les résultats lorsque ce paramètre est activé :

Tab. 22. Classement des modèles en fonction du facteur dégressif d'apprentissage

Hyper-Parameter Search Summary: ordered by increasing MSE

	learn_rate_annealing	model_ids	mse
1	0.995	tst11_model_3	2461723.0795850893
2	0.99	tst9_model_19	2464483.641023643
3	0.999	tst11_model_4	2466729.4459489197
4	1.0	tst11_model_5	2468815.4989898787
5	0.97	tst11_model_1	2507654.9817811525
6	0.95	tst11_model_0	2592743.3734042784

On cherchera ainsi à optimiser le facteur dégressif dans l'intervalle [0,99 ;1], et le coefficient d'apprentissage dans l'intervalle [0,05 ;0,1].

III.b.3.3. Critère d'arrêt de la recherche aléatoire

Voici un tableau récapitulatif des paramètres qui sont utilisés lors de la recherche aléatoire dans une grille :

Tab. 23. Paramètres à optimiser

Paramètre	Notation	Valeurs recherchées
Nombre de subdivisions pour les variables continues	nbins	[500;1500]
Sélection aléatoire de colonnes	ncol_sample_rate	[40%;100%]
Minimum d'observations par feuille	min_rows	{1,2,5,10,20}
Profondeur maximale des arbres	K	[4;8]
Observations sélectionnées aléatoirement	p	[50%;100%]
Coefficient d'apprentissage	λ	[0,05;0,1]
Facteur dégressif	λ_t	[0,99;1]

Il est à noter que compte tenu des domaines d'existence de nos paramètres, il existe $11 * 7 * 5 * 5 * 6 * 51 * 11 = 6479550$ combinaisons possibles. On comprend bien l'intérêt de la recherche aléatoire : parcourir linéairement toutes ces combinaisons n'est pas envisageable.

Il existe deux critères d'arrêt pour la recherche aléatoire : le temps de calcul et le nombre de modèles calculés. Pour le second, il peut être intéressant de calculer par exemple le nombre de modèles nécessaires afin de tomber sur un des 5 % meilleurs modèle avec une probabilité de 99,5 % (voir calcul en annexe). On trouve ainsi un minimum de 149 modèles. Le problème est que le temps de calcul nécessaire à leur construction dépend fortement de la machine utilisée, ce qui peut être handicapant lorsqu'il existe une contrainte telle qu'un délai de livraison.

Ici, il a été préféré de fixer un temps de calcul maximal. Plus le temps est important, plus il y aura de modèles calculés et plus la solution se rapprochera de l'optimum. Dans un cadre pratique, il suffit de fixer un temps de calcul en accord avec les délais de livraison des résultats.

Dans un cadre plus théorique, il semble intéressant de comparer les résultats issus de différents temps de calcul. On présente ainsi les meilleurs modèles parmi tous ceux calculés, au sens de l'erreur moyenne quadratique sur l'échantillon de validation, issue de 3 recherches aléatoires : une de 1 heure, une de 4 heures et une de 12 heures. Le détail des 3 recherches peut être trouvé en annexe.

Tab. 24. Résultats de la recherche aléatoire en fonction du temps de calcul

Temps de calcul	col_sample_rate	lambda	facteur dégressif	K	nbins	min_rows	p	Nombre d'arbres	MSE test
1 heure	0,4	0,062	0,992	8	1400	20	0,8	1375	2 422 460
4 heures	0,4	0,072	0,996	8	1500	1	0,8	2792	2 428 957
12 heures	0,5	0,069	0,994	8	1500	1	1	1853	2 421 259

Il est intéressant de remarquer que le modèle issu de la recherche de 1 heure est plus performant que celui de la recherche de 4 heures : cela met en évidence le facteur aléatoire de la recherche.

Une autre observation intéressante est le nombre de modèles calculés en fonction du temps de calcul. En 1 heure, 44 modèles ont été construits, contre 174 en 4 heures et 235 en 12 heures.
Bien qu'il soit difficile, sans étude complémentaire, d'identifier avec certitude les raisons de la non-linéarité de la relation modèles construit / temps de calcul, il est néanmoins possible d'émettre quelques hypothèses concernant son origine.
En effet, il a été constaté empiriquement que tous les modèles ne sont pas égaux en temps de calcul. La preuve en est que le meilleur modèle trouvé après 4 heures de recherche a besoin de 2792 arbres pour sa construction, contre seulement 1375 et 1853 pour les deux autres. En plus de ça, la vitesse de construction d'un arbre dépend fortement du nombre de colonne et d'observations sélectionnées aléatoirement, mais également de la profondeur maximale autorisée et le nombre d'observations minimum par feuille.

Enfin, parmi les paramètres optimisés, deux résultats sont assez surprenants : la proportion p d'observations sélectionnées aléatoirement et le nombre col_sample_rate de colonnes sélectionnées aléatoirement. En effet, étant donné la structure de nos données, on s'attendait plutôt à trouver p entre 0,5 et 0,7 et à trouver col_sample_rate entre 0,9 et 1. Ce résultat est d'autant plus intéressant quand on le compare avec les résultats du RGLM, qui dispose également d'une sélection aléatoire de colonnes (variables), mais dont le paramètre associé a été considéré comme peu pertinent. La raison de cet écart est probablement dû à la différence de flexibilité entre un GLM et un arbre de régression.

III.c. Application et résultats

III.c.1. Résultats du modèle optimisé

Le GBM optimisé est le modèle retenu après 12 heures de recherche aléatoire parmi les combinaisons de paramètres autorisés. Les caractéristiques de ce modèle sont présentées ici :

Tab. 25. Valeurs des paramètres du modèle optimisé

col_sample_rate	lambda	facteur dégressif	K	nbins	min_rows	p
0,5	0,069	0,994	8	1500	1	1

Pour résumé, le GBM utilisé sélectionne aléatoirement 50 % des colonnes à chaque fois qu'un arbre est construit. Ces arbres ne peuvent dépasser une profondeur de 8 couches, et les feuilles de ces arbres doivent disposer d'au moins une observation, ce qui en

somme revient à ne pas mettre de restriction. Les variables continues sont discrétisées en 1500 facteurs, soit : la densité, le kilométrage, l'ancienneté du véhicule et le montant de l'expertise. Enfin, le coefficient d'apprentissage est de 0,069, et est diminué de 0,006% à chaque itération.

Pour la construction du modèle, la méthode d'échantillonnage est modifiée : on utilise une validation croisée sur 5 échantillons, qui a été constatée plus performante mais également plus lourde en calcul, c'est pourquoi elle n'est utilisée que pour la production finale du résultat. Une fois encore la validation croisée se révèle plus efficace, en abaissant l'erreur MSE du boosting sur l'échantillon test à 2 411 535, contre 2 421 259 avec la méthode d'échantillonnage classique.

En somme, l'optimisation a permis un gain non négligeable de performance : l'erreur sur l'échantillon test de la première application du modèle s'élevait à 2 477 139, et désormais l'erreur du modèle est de 2 411 535, soit une diminution de 2,7 %.

Tab. 26. Tableau comparatif de l'erreur des modèles sur l'échantillon test

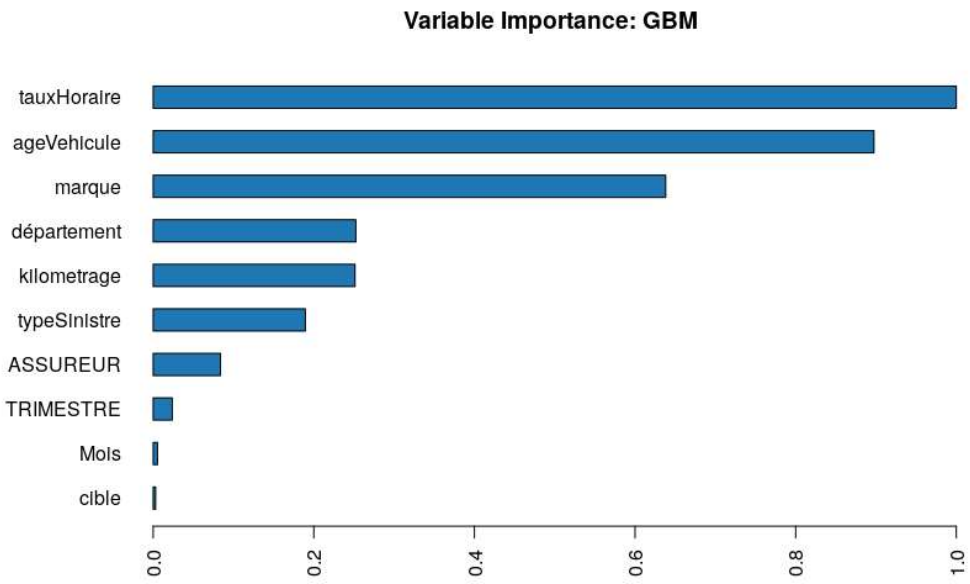
	COR	MSE	L1
GLM	0,55	3 639 942	770
RGLM simple	0,55	3 639 483	770
RGLM amélioré	0,66	2 903 323	716
RGLM méthode "meilleur modèle"	0,66	2 900 468	716
GBM simple	0,72	2 477 139	661
GBM optimisé	**0,73**	**2 411 535**	**651**

Pour une vision plus concrète, la norme L1 est assez parlante : en moyenne, les prédictions issues du GBM optimisé sont de 10 € plus précises que celles du GBM simple. Cette amélioration n'est pas forcément visible au niveau du coût moyen, à cause du phénomène de compensation des écarts, mais atteste bien d'un modèle de meilleure qualité.

III.c.2. Importance relative des variables

On obtient les résultats suivants pour le modèle optimisé, dont le détail est présenté en annexe :

Fig. 24 : Importance relative des variables explicatives pour le GBM optimisé



Il est intéressant de comparer ces résultats avec l'analyse descriptive du V de Cramer des variables explicatives avec les variables à expliquer avant la modélisation. Les caractéristiques principales sont les mêmes, à savoir la prédominance de la variable taux horaire, et en bas du tableau on retrouve le trimestre et la variable cible cible. Néanmoins, on observe que les variables marque et ageVehicule ont joué un rôle important qui n'avait pas été anticipé au début de l'étude.

Il est important cependant de modérer ces résultats. En effet, bien que la hiérarchie des variables semble globalement inchangée, on observe que les valeurs de leurs influences fluctuent fortement en fonction du modèle. Ce n'est pas sans rappeler l'instabilité et l'irrégularité des arbres de régression. En effet, étant donné la partie d'aléatoire présente dans le modèle, il est courant d'observer des changements de comportement importants en comparant des modèles différents mais ayant des performances équivalentes.

Vient alors la question de l'analyse de performance des réseaux d'experts. Est-il possible à travers l'analyse relative des variables explicatives ici présentes, d'en retirer une interprétation de l'influence de la variable dans le coût final du sinistre ? La réponse est probablement non. Une des raisons est que la mesure de l'influence est imprécise, comme cela a été évoqué précédemment, ce qui rend le résultat en lui-même difficile à exploiter et à interpréter. D'autre part, la mesure d'influence utilisée se limite à l'analyse de la construction des arbres. Bien que fortement corrélé, ce n'est pas la même chose qu'analyser directement le poids de la variable dans le coût du sinistre, ce qui ne garantit donc pas que l'analyse soit pertinente.

III.d. Conclusion

A la suite du GLM et du RGLM, un stochastic gradient boostic a été ajusté sur les données afin d'essayer une méthode d'apprentissage automatique, en contraste avec ce qui avait été fait jusqu'ici. Dès la première application, les résultats sont assez surprenants : le modèle atteste d'une qualité de prédiction bien supérieure à celle des deux modèles précédents.

L'optimisation des paramètres n'a fait que confirmer cette observation. A l'aide d'une recherche aléatoire sur une grille de paramètres, il a été possible d'allier performance et temps de calcul. L'optimisation de chaque paramètre a été détaillée afin de s'assurer de la cohérence de la recherche et du résultat.

De plus, le GBM tel qu'il est présenté ici est optimisé pour gérer de grands volumes de données. Il s'exécute rapidement tout en minimisant la mémoire requise, ce qui en fait un modèle de choix pour répondre aux problématiques du Big Data.

Néanmoins, il est à noter que cette performance à un coût : la transparence des résultats. Parmi les centaines d'arbres qui sont construits à la minute, il est difficile de tracer l'évolution de la solution, et de s'assurer du bon déroulement des calculs. De plus, une partie non négligeable de l'algorithme est soumis à l'aléatoire, à travers les paramètres de sélection de colonnes ou de lignes, ce qui parfois peut jouer sur la cohérence de l'interprétation. De la même manière, de nombreuses combinaisons de paramètres peuvent donner lieu à des performances similaires, ce qui complique l'explication du choix de la combinaison utilisée.

En somme, le GBM fournit d'excellents résultats, tant au niveau de l'exécution pratique de l'algorithme que des prédictions, mais il est très difficile de répondre à la question simple : « pourquoi cette valeur de y est prédite pour ces caractéristiques de sinistre ? ».

IV. ANALYSE DES RESULTATS

Deux problématiques dirigent ce mémoire : l'évaluation de l'écart de performance des réseaux d'experts, et l'amélioration du pouvoir prédictif des modèles. C'est donc sur ces deux points que les modèles seront comparés.

IV.a. Evaluation de l'écart de performance des réseaux

IV.a.1. GLM

Le GLM est une méthode tout à fait adéquate à la problématique d'évaluation de l'écart de performance des réseaux. En effet, le GLM de loi gamma et de fonction de lien logarithmique qui a été ajusté propose une interprétation intuitive de l'influence d'une variable. A chaque modalité de chaque variable est estimée un coefficient, qui n'est autre que l'impact sur le coût du sinistre de la présence de cette modalité dans les caractéristiques du sinistre. Ainsi, il est aisé d'en ressortir l'impact de la variable associée aux réseaux d'experts.

IV.a.2. RGLM

Le cas du RGLM dépend de la méthode d'agrégation utilisée. En effet, nous avons montré que lorsque l'agrégation consiste à rassembler les résultats de plusieurs modèles en un seul, alors toute information concernant les coefficients est perdue. Cependant, lorsque la méthode d'agrégation se limite à conserver le meilleur modèle parmi ceux testés, alors on retombe dans l'analyse de coefficients d'une régression pénalisée.

Deux problèmes se posent néanmoins lors de l'estimation de l'influence d'une variable par cette méthode.
Premièrement, nous utilisons les termes d'interactions de second ordre entre les variables, ce qui signifie que l'impact d'une variable n'est plus limité à l'estimation d'un coefficient. Par exemple, si on a deux variables x et y, respectivement la marque et le kilométrage, le modèle prend en compte la relation x*y. Cela signifie que 3 coefficients sont calculés, respectivement β_1, β_2 et β_3, et l'influence de la variable x (la marque) dépend donc de deux coefficients : β_1 et β_3.
Le deuxième problème se situe dans la pénalisation de la régression. La pénalisation altère artificiellement les coefficients, suivant l'importance accordée à celle-ci, afin d'empêcher que ceux-ci ne deviennent trop grands, voir les annuler. C'est le rôle de la pénalisation LASSO que nous utilisons ici, qui est de sélectionner une variable parmi les celles qui sont très corrélées, en annulant le coefficient des autres. L'interprétation est alors gênée : une variable dont le coefficient est nul ne signifie pas qu'elle n'a pas d'impact sur le coût du sinistre, d'autant plus que dans la pénalisation LASSO, la variable qui est conservée est choisie arbitrairement.

A titre illustratif, voici l'estimation des coefficients d'une régression pénalisée LASSO avec termes d'interactions (résultats similaires au RGLM méthode « meilleur modèle »).

Tab. 27. Estimation des coefficients d'une régression pénalisée

Standardized Coefficient Magnitudes: standardized coefficient magnitudes

	names	coefficients	sign
1	ageVehicule_tauxHoraire.Jeune_NUL	0.701333	POS
2	marque_tauxHoraire.BMW_NUL	0.538125	POS
3	typeSinistre.BDG	0.496846	NEG
4	marque_tauxHoraire.AUDI_NUL	0.490758	POS
5	marque_tauxHoraire.MERCEDES_NUL	0.450009	POS

---

	names	coefficients	sign
996	kilometrage.Moyen km	0.000000	POS
997	densite.Faible densite	0.000000	POS
998	densite.Moyenne densite	0.000000	POS
999	ageVehicule.Middle aged	0.000000	POS
1000	cible.N	0.000000	POS
1001	cible.O	0.000000	POS

On note ainsi un très grand nombre de coefficients calculés, 1001 pour être exact, correspondant à tous les couples de modalités possibles. On note l'annulation de 650 coefficients par ce procédé.

IV.a.3. GBM

Il est possible d'estimer l'influence relative des variables explicatives dans un GBM, en prenant en compte le nombre de fois où la variable est intervenue dans la division d'un nœud, et l'amélioration que cette division a apporté au modèle. Le problème de cette estimation est qu'elle ne mesure pas directement l'impact de la variable sur le coût du sinistre, mais seulement l'importance de celle-ci dans la construction du modèle, en comparaison avec les autres variables.

Cette mesure devient intéressante lorsqu'il s'agit de comparer une variable avec une autre qui sert alors d'étalon. Par exemple, si on a une idée précise de l'impact de l'ancienneté du véhicule sur le coût, il peut être intéressant de comparer les coefficients d'influence qui ont été calculé grâce au GBM.
Néanmoins, cette méthode est à utiliser avec précaution. Il a été observé que l'importance relative d'une variable peut fortement fluctuer d'un GBM à l'autre, suivant les paramètres qui ont été utilisés.

IV.a.4. Conclusion

Le GLM semble être le modèle qui répond le mieux à la question de l'évaluation de l'écart de performance entre les deux réseaux d'experts. Ce modèle a l'avantage de fournir une

fonction réponse $\hat{f}$ ayant une structure simple, visualisable, et n'impliquant que les variables fournies par la base de données.

Les fonctions réponses $\hat{f}$ des deux autres modèles sont bien plus complexe, soit dans leurs constructions (GBM) soit dans leurs relations (RGLM). Ces deux modèles disposent certes d'outils permettant d'estimer l'influence des variables, mais semble moins pertinent lorsqu'il s'agit d'estimer leurs impacts sur le coût.

IV.b. Prédiction des modèles

IV.b.1. Récapitulatif des résultats de l'étude

On présente ici le tableau comparatif du pouvoir prédictif des différents modèles construits au cours de ce mémoire :

Tab. 28. Tableau comparatif de l'erreur des modèles sur l'échantillon test

	COR	MSE	L1
GLM	0,55	3 639 942	770
RGLM simple	0,55	3 639 483	770
RGLM amélioré	0,66	2 903 323	716
RGLM méthode "meilleur modèle"	0,66	2 900 468	716
GBM simple	0,72	2 477 139	661
GBM optimisé	**0,73**	**2 411 535**	**651**

Trois mesures ont servi à la comparaison des modèles : la corrélation (COR), l'erreur moyenne quadratique (MSE) et la norme L1.

On constate la présence de trois grands paliers de qualité de prédiction, atteint par le GLM, le RGLM amélioré et le GBM simple. Le RGLM simple étant une version finalement peu modifiée du GLM (la sélection aléatoire de variable se révélant peu pertinente), on peut raisonnablement dire que le changement de méthode est le principal facteur d'amélioration des prédictions.

L'optimisation de chaque algorithme quant à elle permet de tirer le meilleur parti de la méthode utilisée. L'amélioration apportée par celle-ci peut sembler dérisoire lorsqu'on voit à quel point un changement de méthode peut modifier la qualité de prédiction, mais ce n'est pas le cas. En effet, l'avantage de l'optimisation d'un modèle est que celle-ci permet d'améliorer ses performances sans altérer les atouts qu'il possède. Un changement de méthode implique un changement de perspective, et donc des caractéristiques différentes.

En effet, malgré des performances prédictives très différentes, chaque modèle possède son lot d'atouts qui le différencie des autres. Le GLM impose une structure à la relation entre la variable réponse et les variables explicatives, qui demande un certain travail de retraitement et de choix de paramétrisation. Cependant, une fois que cela est fait, l'interprétation et l'explication des résultats deviennent intuitifs et compréhensibles.

Le RGLM quant à lui utilise une base relativement simple, le GLM, qu'il complexifie afin d'en améliorer les performances. Visuellement, le fonctionnement de l'algorithme est assez intuitif, et réussi à tirer parti des avantages du GLM en terme d'interprétabilité. Cependant, ce modèle est plus long à mettre en place et certaines fonctionnalités du GLM peuvent parfois être perdues (comme l'évaluation de l'influence d'une variable).

Enfin, le GBM propose une qualité de prédiction au-dessus des deux autres modèles. Néanmoins, la paramétrisation du modèle est assez lourde, une grande quantité de calculs est effectuée ce qui gêne fortement la traçabilité et la vérification du résultat.

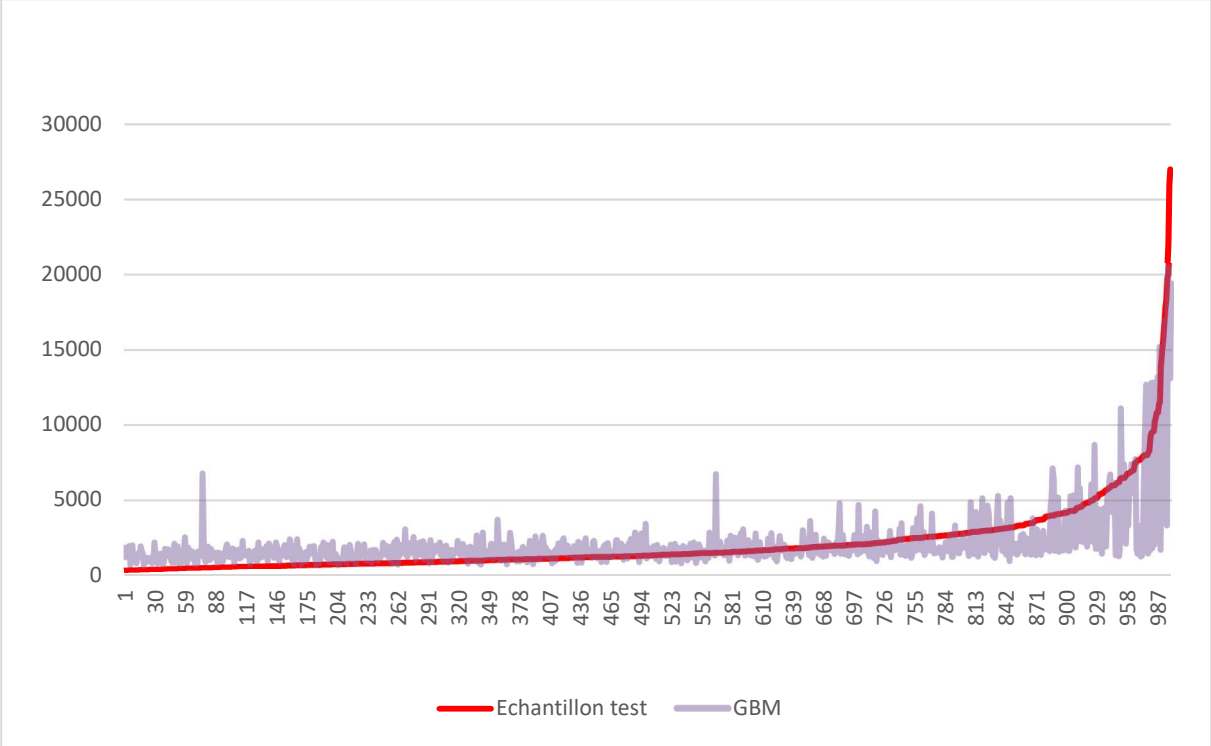
IV.b.2. Analyse des prédictions

Pour la suite de la comparaison, seules les meilleures versions de chaque modèle sont conservées, à savoir : le GLM, le RGLM avec comme méthode d'agrégation le « meilleur modèle », et le GBM optimisé.

L'écriture de cette partie part de la constatation que, même pour le modèle prédictif le plus optimisé, la norme L1 avoisine les 650 €. Pour rappel, la norme L1 correspond à l'écart absolu moyen entre les prédictions et les observations. Une norme L1 aussi élevé attire l'attention, étant donné que les montants de sinistres sont modérés (2012 € en moyenne sur l'échantillon test).

Afin d'identifier la raison pour laquelle la norme L1 est aussi élevée, on effectue un tirage aléatoire sans remise sur l'échantillon test, et on compare les observations aux prédictions du GBM, les autres modèles étant écarté pour le moment pour ne pas encombrer les graphiques.

Fig. 25 : Comparaison des prédictions et des observations dans l'ordre croissant



On constate ainsi une forte variation des prédictions sur les grands sinistres, qu'on peut vérifier en regardant au hasard 10 sinistres graves :

Tab. 29. Comparaison prédictions/observations sur des sinistres graves

Echantillon test	GLM	RGLM	GBM
15000	4034	7663	11495
15600	6775	12873	12658
16500	6211	13183	**9404**
17800	8435	25247	16948
18354,19	1298	1738	**3594**
19766,02	2514	2515	**3325**
20000	4335	11564	20501
22000,01	5121	8740	**14740**
26000	4127	4816	**13068**
27000	8507	15040	19398

On remarque ainsi que les modèles ont beaucoup de mal à identifier, à partir des caractéristiques des sinistres, les scénarii induisant un sinistre grave. Le cas du GLM est le plus flagrant : après vérification, il apparaît que la prédiction maximale enregistrée sur tout l'échantillon test par le GLM est de 10 250 €, ce qui signifie que tous les sinistres graves sont mal prédits. Le RGLM et le RGLM s'en sortent un peu mieux, mais l'adéquation des prédictions est très variable d'un sinistre à l'autre.
De plus, l'imprécision des prédictions ne semblent pas être exclusive aux sinistres graves, bien que les écarts soient bien moins importants.

Tab. 30. Comparaison prédictions/observations sur des petits sinistres

Echantillon test	GLM	RGLM	GBM
335	898	905	939
337	2558	2191	1841
342	1414	1271	1202
342	1273	1232	1322
342	1324	1403	1371
351	2011	1901	1866
353	1829	1968	1959
354	979	970	701
354	1349	1544	1656
355	2251	2217	1990

On remarque ainsi que deux sinistres ayant le même coût observé (354 €) peuvent avoir des prédictions très différentes (701 € et 1656 €).

Enfin, pour appuyer le fait que ce comportement se généralise à tous les sinistres, même les sinistres proches de la moyenne, il est possible de trouver en annexe les prédictions sur un échantillon de 25 sinistres.

IV.b.3. Limites de la modélisation

Il semble que, malgré toutes les optimisations qui ont pu être faites sur les modèles, il n'a pas été possible d'obtenir une qualité de prédiction correcte.

A ce stade, deux hypothèses peuvent expliquer cette observation : soit les modèles ne sont pas adaptés à ce genre de problème, soit les données fournies ne sont pas suffisantes pour prédire précisément le coût d'un sinistre matériel.

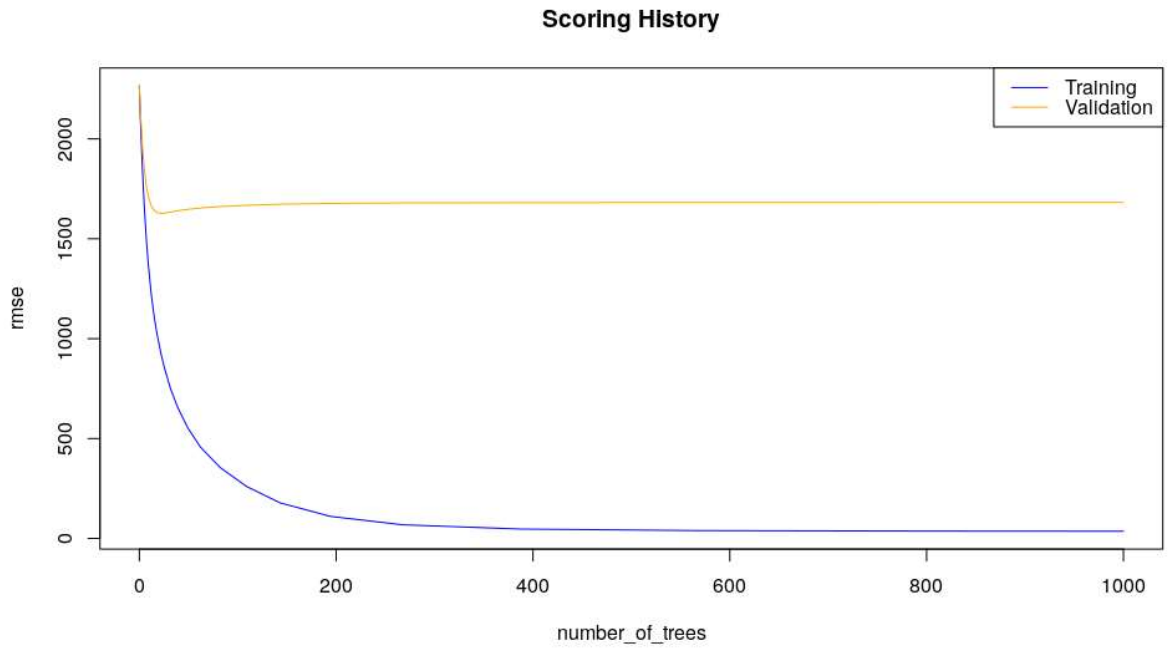
De nombreux éléments semblent être en faveur de la deuxième explication, comme l'observation qu'aucun modèle ne soit capable de prédire deux fois la même valeur pour des sinistres ayant le même coût. Cela semble indiquer qu'il existe une information sous-jacente qui aurait permis de relier ces deux sinistres, mais qui n'est pas présente dans la base de données.

Il est à noter que, là où une segmentation du portefeuille en groupes de risques homogènes améliorerait les performances du GLM et du RGLM, le GBM dispose déjà de sa propre classification, ce qui laisse penser que cette proposition ne résolverait pas tous les écarts de prédictions observés.

D'autre part, on pourrait se demander pourquoi l'étape de validation des modèles, notamment l'étude des résidus du GLM puisqu'elle est explicite, ne relève pas ce manque d'information.
Il faut savoir que lorsque les variables sont introduites dans le modèle, cela lui impose un cadre de travail pour lequel il estime si oui ou non celui-ci est suffisant pour expliquer la variable d'intérêt. Il est tout à fait possible que cette estimation soit positive tout en ignorant l'existence d'une variable discriminante qui aurait changé le cadre de travail. Nous reviendrons sur ce point plus loin.

On peut imaginer plusieurs procédés pour essayer de mettre en évidence l'absence d'une variable discriminante.
Dans le cas du GBM, une idée intéressante est de volontairement faire sur-apprendre le modèle, et vérifier que l'erreur d'apprentissage converge correctement vers 0. Le GBM étant un modèle propice au sur-apprentissage, si le contraire est observé, cela attesterait d'un problème de structure dans les données.

Fig. 26 : Sur-apprentissage d'un GBM



Afin de favoriser le sur-apprentissage, chaque arbre est construit sur 30 couches, avec une moyenne de 25 823 feuilles et le coefficient d'apprentissage a été relevé à 0,1. L'erreur converge bien vers 0, mais si on regarde les prédictions de plus près, on se rend compte que des anomalies persistent. Bien qu'ayant une erreur MSE de 1 280, et une norme L1 de 0,10 € sur l'échantillon d'apprentissage (ce qui confirme bien le sur-apprentissage), une centaine de prédictions ont encore plus de 1 000 € d'écarts avec les observations. Ce genre de résultats est assez étonnant dans un cas avéré de sur-apprentissage. Cela montre que malgré une segmentation très fine des observations,

l'algorithme est incapable de différencier certains sinistres ayant pourtant des coûts très différents.

Cela renvoie à une autre proposition, qui est de s'assurer que la segmentation induite par les variables est suffisante pour bien différencier les sinistres. On a remarqué dans un premier temps que deux sinistres ayant le même montant ont parfois des prédictions très différentes. L'observation opposée est également vraie : il existe des sinistres ayant exactement les mêmes caractéristiques, à savoir la même marque, le même type de sinistre, le même lieu de sinistre, une ancienneté de véhicule et un kilométrage quasi identique, et malgré cela, les montants d'expertises sont différents. Voici quelques exemples de ce qu'on peut trouver dans la base de données :

Tab. 31. Comparaison prédictions/observations sur des petits sinistres

montExp	marque	kilometrage	cible	typeSinistre	ageVehicule	TRIMESTRE	ASSUREUR	Mois	département	tauxHoraire	densite
870,29	RENAULT	15710	O	ACT	0,79945243	T3 2015	Assureur_1	201507	62	TH fort	221,470394
1137,29	RENAULT	15845	O	ACT	0,89801506	T2 2016	Assureur_1	201606	62	TH fort	221,470394
863	RENAULT	26055	O	ACT	1,60985626	T3 2015	Assureur_7	201509	62	TH fort	221,470394
1439	RENAULT	26848	O	ACT	1,83436003	T1 2015	Assureur_7	201503	62	TH fort	221,470394
3273,13	RENAULT	36774	O	ACT	2,23408624	T1 2016	Assureur_3	201602	62	TH fort	221,470394
4578	RENAULT	36776	O	ACT	2,11362081	T1 2016	Assureur_7	201602	62	TH fort	221,470394
413,43	RENAULT	48734	O	ACT	2,9596167	T3 2016	Assureur_1	201609	62	TH fort	221,470394
1753	RENAULT	49480	O	ACT	3,02806297	T1 2015	Assureur_7	201501	62	TH fort	221,470394

Le seul facteur différenciant chaque sinistre d'une paire est soit le moment du sinistre, soit l'assureur. Et pourtant, on peut observer de grande disparité dans le montant de l'expertise.
On peut également observer des sinistres qui sont en tout point identiques, exceptés en ce qui concerne le montant d'expertise. Ce cas est observable lorsqu'un assuré subis plusieurs sinistres durant la même période, mais cela montre le manque de variables explicatives pour différencier ces sinistres puisque seul le montant d'expertise permet de les identifier.

IV.b.4. Conclusion

A travers l'analyse des prédictions et l'analyse statistique de la base de données, il semble que le nombre de variables disponibles soit insuffisant dans l'optique de la construction d'un modèle prédictif de qualité. Le point appuyant le plus cet hypothèse est la présence de sinistres dont les caractéristiques sont identiques mais dont le montant d'expertise est différent.

Cela fait écho à une remarque qui a pu être faite au début de ce mémoire. L'origine de la demande de données était de répondre à la problématique de l'écart de performance entre les deux réseaux. Les variables ont donc été sélectionnées dans ce but précis. Il était donc déjà envisagé que celles-ci ne suffisent pas pour un modèle prédictif.

IV.c. Cohérence des résultats

Les écarts entre les prédictions et les observations ont créé le doute sur la suffisance des variables pour la construction d'un modèle prédictif. A juste titre, on peut se demander si cela remet en cause l'étude de l'écart de performance entre les deux réseaux d'experts.

Les modèles sont construits sur un nombre limité de variable. En notant *X* l'ensemble des variables explicatives et *Y* le coût du sinistre, le modèle fournit une fonction *f*, explicite ou non, tel que $\hat{Y} = f(X)$, $\hat{Y}$ étant les prédictions du modèle.

Cependant, dans la réalité, il est très probable que le coût du sinistre dépende d'autres facteurs, et que la « vraie » relation soit plutôt de la forme $Y = h(X, Z)$, avec Z un ensemble de variables non observées. Dans notre cas, Z pourrait contenir la météo, les informations sur le conducteur, le type de moteur ou encore l'emplacement du choc. Lorsqu'un facteur n'est pas intégré au modèle, les autres variables captent partiellement son information, plus ou moins fortement en fonction de leur corrélation avec celui-ci.

Cela signifie que, pour la méthode GLM, le coefficient estimé d'une variable ne transmet pas uniquement l'influence de celle-ci sur le coût du sinistre, puisque les variables non observées affectent également ce coefficient. Très souvent, les variables non observées ont un impact négligeable sur celui-ci ce qui signifie que la valeur du coefficient est tout de même exploitable.

Si effectivement une variable discriminante est absente du modèle, cela remet en question la valeur du coefficient. Il n'est plus possible de certifier que seule l'influence de la variable observée y est retranscrite.

Néanmoins, il est habituellement supposé que les variables non observées ont un comportement constant au cours du temps, ce qui veut dire que l'analyse des variations des coefficients au cours des trimestres reste pertinente.

Conclusion générale

Dans le cadre d'un portefeuille de sinistres matériels, deux problématiques ont été posées au début de ce mémoire : l'estimation de l'influence d'une variable sur le coût, et la construction d'un modèle prédictif de bonne qualité. Pour cela, trois types de modélisation ont été utilisés : un modèle type GLM (Generalized linear model), une méthode d'agrégation de modèle (RGLM) et une méthode de boosting (GBM).

Le choix de ces modèles répond à l'objectif secondaire de ce mémoire, qui est d'évaluer l'impact de l'introduction du machine learning dans des problématiques traditionnellement traitées par des modèles de régression plus classiques. Cette introduction s'est faite graduellement avec le RGLM, qui agrège les prédictions données de plusieurs GLM, puis avec une méthode d'apprentissage automatique, le gradient boosting model.

En ce qui concerne l'évaluation de l'influence d'une variable sur le coût du sinistre, le GLM se trouve être la solution la plus pertinente parmi les trois modèles testés. En effet, il estime un coefficient pour chaque variable, qui évalue directement l'impact de celle-ci sur le coût du sinistre. Les deux autres modèles disposent également d'outils permettant d'évaluer l'influence d'une variable, mais ils sont jugés soit trop instables, soit peu adéquat pour cette problématique.

En termes de qualité de prédiction, le GBM détient les meilleurs performances, suivi du RGLM et du GLM. L'écart entre ces trois modèles est assez important sur ce point, mais on remarque que plus le modèle est performant, plus les résultats sont difficilement interprétables. Pour prendre le cas du GBM, le nombre de calculs effectués est très important à travers la quantité d'arbres créés pour l'ajustement du modèle, ce qui gêne voire rend impossible la traçabilité des résultats. Là où le GLM calcul une fonction explicite, permettant de déterminer les prédictions, le GBM produit uniquement un résultat.

D'autre part, l'analyse des prédictions a révélé des écarts importants et répétés avec les observations. La question de la suffisance des variables à disposition pour la construction d'un modèle prédictif de qualité est alors posée. De nombreux éléments semblent montrer qu'il existe un manque d'information pour différencier des sinistres en apparence similaire mais dont le montant d'expertise est parfois très différent. Toutefois, ces écarts n'ont pas d'effets sur la cohérence de l'étude de performance des réseaux d'experts, à condition qu'on en précise le cadre et les limites.

Enfin, on constate que l'ajout de techniques propres au Machine Learning dans la modélisation semble complexifier les relations et produire des résultats de meilleures qualités, au détriment de leur interprétabilité. On note que le GBM est un modèle

permettant de gérer de manière efficace un grand volume de données, ce qui la rend particulièrement utile dans des problèmes où le nombre de variable est important. Etant donnée les problématiques du Big Data qui bousculent le monde de l'assurance automobile, les méthodes de modélisation présentées (le GBM et le RGLM) peuvent constituer une approche intéressante pour y répondre.

ANNEXE

A. PREPARATION DES DONNEES

Tab. 32. **Codage du type de sinistre**

Type de sinistre	Signification
ACP	accident de parking
ACT	accident avec tiers
APC	accident perte de contrôle
ACC	accident
BDG	bris de glace
CAT	catastrophe naturelle
GRE	grêle
INC	incendie
TEM	tempete
AUT	autre
VLP	vol partiel
VLT	vol total

B. GLM

Tab. 33. Estimation des coefficients du GLM

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0431  -0.6554  -0.2492   0.2314   7.1221

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)               7.353983   0.008591  855.995  < 2e-16 ***
cibleO                   -0.024427   0.001902  -12.844  < 2e-16 ***
marqueAUDI                0.092239   0.005081   18.155  < 2e-16 ***
marqueBMW                 0.150123   0.005333   28.148  < 2e-16 ***
marqueCITROEN            -0.168294   0.003478  -48.390  < 2e-16 ***
marqueFIAT               -0.175755   0.006053  -29.037  < 2e-16 ***
marqueFORD               -0.183383   0.004961  -36.963  < 2e-16 ***
marqueMERCEDES            0.168184   0.004976   33.801  < 2e-16 ***
marqueNISSAN              0.020447   0.006085    3.360 0.000779 ***
marqueOPEL               -0.162752   0.005534  -29.411  < 2e-16 ***
marquePEUGEOT            -0.184046   0.003180  -57.867  < 2e-16 ***
marqueRENAULT            -0.148155   0.003091  -47.932  < 2e-16 ***
marqueTOYOTA             -0.083262   0.005638  -14.767  < 2e-16 ***
marqueVOLKSWAGEN         -0.051996   0.004116  -12.632  < 2e-16 ***
kilometrageMoyen km       0.065846   0.002588   25.442  < 2e-16 ***
kilometrageFort km       -0.020203   0.003070   -6.580 4.70e-11 ***
densiteMoyenne densite   -0.005165   0.002285   -2.260 0.023800 *
densiteForte densite      0.021817   0.002909    7.501 6.35e-14 ***
typeSinistreACT           0.235809   0.004265   55.292  < 2e-16 ***
typeSinistreAPC           0.428019   0.004815   88.893  < 2e-16 ***
typeSinistreAUT           0.292872   0.005201   56.308  < 2e-16 ***
typeSinistreBDG          -0.125505   0.021384   -5.869 4.39e-09 ***
typeSinistreCAT           0.679000   0.013423   50.586  < 2e-16 ***
typeSinistreGRE           0.684775   0.009310   73.556  < 2e-16 ***
typeSinistreINC           0.561599   0.011012   50.999  < 2e-16 ***
typeSinistreTEM           0.508194   0.012501   40.653  < 2e-16 ***
typeSinistreVLP           0.266351   0.008228   32.372  < 2e-16 ***
typeSinistreVLT           0.637255   0.010528   60.527  < 2e-16 ***
ageVehiculeMiddle aged   -0.104096   0.002602  -40.004  < 2e-16 ***
ageVehiculeAged          -0.427806   0.003065 -139.578  < 2e-16 ***
tauxHoraireTH moyen       0.083421   0.002640   31.595  < 2e-16 ***
tauxHoraireTH fort        0.290564   0.002625  110.687  < 2e-16 ***
tauxHoraireNUL            1.102478   0.003171  347.622  < 2e-16 ***
ASSUREURAssureur_3        0.063919   0.002648   24.143  < 2e-16 ***
ASSUREURAssureur_7       -0.018519   0.002623   -7.059 1.68e-12 ***
RegionEST                -0.160893   0.007248  -22.199  < 2e-16 ***
RegionNORD               -0.205415   0.007230  -28.410  < 2e-16 ***
RegionOUEST              -0.240180   0.007355  -32.654  < 2e-16 ***
RegionREGION PARISIENNE  -0.178136   0.007439  -23.948  < 2e-16 ***
RegionSUD                -0.156037   0.007142  -21.847  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.6897202)

    Null deviance: 588375  on 835880  degrees of freedom
Residual deviance: 409147  on 835841  degrees of freedom
AIC: 13969284
```

C. RGLM

Tab. 34. Comparaison des prédictions RGLM et GLM sur un échantillon de 30 sinistres

Observations	RGLM amélioré	GLM
1380	1176	997
5333	5657	6514
1338	1325	1338
1449	1594	1781
1500	2662	3089
1871	1571	1807
628	1726	1610
5300	6608	4712
2000	2491	3387
1787	1961	2100
1169	2478	2385
850	1770	2739
4367	1770	1632
2339	1359	1426
481	991	971
1458	1573	2842
836	1727	2081
850	1183	1112
2573	1653	1601
1315	1915	1515
2495	1002	893
821	1208	1372
605	1501	1677
2575	1979	2062
1959	1953	1828
1624	1958	2355
849	1234	1249
2034	1732	1956
374	1236	1426
1434	1862	1942

D. BOOSTING

Nombre de modèles nécessaires afin de tomber sur un des 5 % meilleurs modèle avec une probabilité de 99,5 % :

Le calcul est relativement simple, sur le même principe qu'un calcul de Value at Risk (VaR). Il suffit de considérer qu'on a n variables uniformes sur [0,1], et le problème s'écrit alors :

$$P\left(\bigcup_{i=1}^{n}(X_i \geq 0.95)\right) \geq 99.5\%$$

$$\Leftrightarrow 1 - P\left(\bigcap_{i=1}^{n}(X_i \leq 0.95)\right) \geq 99.5\%$$

$\Leftrightarrow 1 - P(X_1 \leq 0.95)^n \geq 99.5\%$ car les variables sont i.i.d.

$\Leftrightarrow 1 - (0.95)^{\mathrm{n}} \geq 99.5\%$

$\Leftrightarrow \mathrm{e}^{nln(0.95)} \leq 0.05\%$

$\Leftrightarrow \mathrm{n} \geq \frac{ln(0.05\%)}{\ln(0.95)}$ car ln(0.95) est négatif.

$\Leftrightarrow \mathrm{n} \geq 149$ arrondi à l'entier supérieur, car n est entier.

Tab. 35. Résultat d'une recherche aléatoire après 1h de calcul

Hyper-Parameter Search Summary: ordered by increasing MSE

	col_sample_rate	learn_rate	learn_rate_annealing	max_depth	min_rows	nbins	sample_rate	model_ids	mse
1	0.4	0.062	0.992	8	20.0	1400	0.8	final_grid1_model_10	2417196.5032863957
2	0.4	0.092	0.996	8	5.0	1400	1.0	final_grid1_model_8	2418144.805845441
3	0.4	0.075	0.991	8	10.0	1000	0.7	final_grid1_model_22	2420738.74821195
4	0.6	0.057	1.0	7	5.0	600	0.9	final_grid1_model_29	2422783.162597588
5	0.7	0.07	0.999	8	20.0	1000	0.9	final_grid1_model_21	2423069.40356113

---

	col_sample_rate	learn_rate	learn_rate_annealing	max_depth	min_rows	nbins	sample_rate	model_ids	mse
39	1.0	0.074	1.0	4	10.0	500	0.9	final_grid1_model_1	2463762.0390505847
40	0.4	0.059	0.991	5	2.0	1100	0.9	final_grid1_model_14	2464679.993778976
41	0.7	0.064	0.992	5	10.0	1400	0.5	final_grid1_model_26	2465962.97148511
42	0.5	0.073	0.996	4	1.0	1200	0.6	final_grid1_model_0	2472599.4623382324
43	1.0	0.086	1.0	4	5.0	500	0.6	final_grid1_model_33	2473241.667312939
44	0.5	0.054	0.991	4	2.0	700	0.8	final_grid1_model_15	2520328.730098463

Tab. 36. Résultat d'une recherche aléatoire après 4h de calcul

Hyper-Parameter Search Summary: ordered by increasing MSE

	col_sample_rate	learn_rate	learn_rate_annealing	max_depth	min_rows	nbins	sample_rate	model_ids	mse
1	0.4	0.072	0.996	8	1.0	1500	0.8	final_grid4_model_122	2416349.1133353887
2	0.4	0.062	0.992	8	20.0	1400	0.8	final_grid4_model_10	2416515.959274707
3	0.5	0.094	0.997	8	1.0	1000	0.9	final_grid4_model_92	2417523.43129918
4	0.5	0.069	0.994	8	1.0	1500	1.0	final_grid4_model_100	2419422.0223530135
5	0.4	0.092	0.996	8	5.0	1400	1.0	final_grid4_model_8	2419812.1038665995

---

	col_sample_rate	learn_rate	learn_rate_annealing	max_depth	min_rows	nbins	sample_rate	model_ids	mse
169	0.7	0.052	0.993	4	5.0	1300	0.5	final_grid4_model_169	2502329.331989562
170	0.4	0.09	0.991	4	2.0	700	0.6	final_grid4_model_168	2505754.8830294735
171	0.4	0.085	0.99	4	2.0	800	0.7	final_grid4_model_68	2512066.7235918734
172	0.7	0.061	0.99	4	5.0	800	0.7	final_grid4_model_127	2514182.0699978843
173	0.6	0.065	0.99	4	20.0	700	0.7	final_grid4_model_84	2514552.159468062
174	0.5	0.054	0.991	4	2.0	700	0.8	final_grid4_model_15	2518231.4863458257

Tab. 37. Résultat d'une recherche aléatoire après 12h de calcul

Hyper-Parameter Search Summary: ordered by increasing MSE

	col_sample_rate	learn_rate	learn_rate_annealing	max_depth	min_rows	nbins	sample_rate	model_ids	mse
1	0.5	0.069	0.994	8	1.0	1500	1.0	final_grid2_model_100	2416794.6353701316
2	0.4	0.072	0.996	8	1.0	1500	0.8	final_grid2_model_122	2416837.6571178297
3	0.4	0.062	0.992	8	20.0	1400	0.8	final_grid2_model_10	2417196.5032863957
4	0.5	0.094	0.997	8	1.0	1000	0.9	final_grid2_model_92	2417438.739267008
5	0.4	0.092	0.996	8	5.0	1400	1.0	final_grid2_model_8	2418144.805845441

---

	col_sample_rate	learn_rate	learn_rate_annealing	max_depth	min_rows	nbins	sample_rate	model_ids	mse
230	0.4	0.067	0.992	4	1.0	700	0.8	final_grid2_model_216	2506400.886456834
231	0.7	0.064	0.99	4	10.0	1500	0.6	final_grid2_model_225	2510584.742660392
232	0.6	0.065	0.99	4	20.0	700	0.7	final_grid2_model_84	2512874.044325364
233	0.4	0.085	0.99	4	2.0	800	0.7	final_grid2_model_68	2513665.8268323876
234	0.7	0.061	0.99	4	5.0	800	0.7	final_grid2_model_127	2514260.4021267034
235	0.5	0.054	0.991	4	2.0	700	0.8	final_grid2_model_15	2520328.730098463

Tab. 38. Importance relative des variables explicatives du modèle après 1h de calcul

Variable Importances:

	variable	relative_importance	scaled_importance	percentage
1	tauxHoraire	4259945906176.000000	1.000000	0.297548
2	marque	2810348568576.000000	0.659715	0.196297
3	ageVehicule	2719275286528.000000	0.638336	0.189935
4	département	1993010708480.000000	0.467849	0.139207
5	kilometrage	1164783386624.000000	0.273427	0.081357
6	typeSinistre	702016389120.000000	0.164795	0.049034
7	ASSUREUR	290490318848.000000	0.068191	0.020290
8	TRIMESTRE	197447106560.000000	0.046350	0.013791
9	Mois	94359838720.000000	0.022150	0.006591
10	densite	59431550976.000000	0.013951	0.004151
11	cible	25744775168.000000	0.006043	0.001798

Tab. 39. Importance relative des variables explicatives du modèle après 4h de calcul

Variable Importances:

	variable	relative_importance	scaled_importance	percentage
1	tauxHoraire	3485173547008.000000	1.000000	0.354606
2	ageVehicule	2074475757568.000000	0.595229	0.211071
3	marque	1691301838848.000000	0.485285	0.172085
4	kilometrage	856844664832.000000	0.245854	0.087181
5	département	811077468160.000000	0.232722	0.082525
6	typeSinistre	502838263808.000000	0.144279	0.051162
7	ASSUREUR	210943524864.000000	0.060526	0.021463
8	TRIMESTRE	89978683392.000000	0.025818	0.009155
9	Mois	52609835008.000000	0.015095	0.005353
10	densite	39722582016.000000	0.011398	0.004042
11	cible	13342717952.000000	0.003828	0.001358

Tab. 40. Importance relative des variables explicatives du modèle après 12h de calcul

Variable Importances:

	variable	relative_importance	scaled_importance	percentage
1	tauxHoraire	3630648262656.000000	1.000000	0.288561
2	ageVehicule	3389205250048.000000	0.933499	0.269372
3	marque	2336063225856.000000	0.643429	0.185669
4	département	1037894352896.000000	0.285870	0.082491
5	kilometrage	1013063811072.000000	0.279031	0.080518
6	typeSinistre	684566183936.000000	0.188552	0.054409
7	ASSUREUR	299776081920.000000	0.082568	0.023826
8	TRIMESTRE	89086533632.000000	0.024537	0.007081
9	Mois	55771615232.000000	0.015361	0.004433
10	densite	33714778112.000000	0.009286	0.002680
11	cible	12107657216.000000	0.003335	0.000962

E. ANALYSE DES RESULTATS

Tab. 41.Comparaisons observations/prédictions sur un échantillon de 25 sinistres

Echantillon test	GLM	RGLM	GBM
348	1922	1571	1592
352	1051	1074	997
386	1903	1773	1727
411	894	906	1159
521	1016	983	709
590	1799	1548	1457
782	1123	1099	1213
936	1787	1557	1535
1051	1578	1360	1200
1063	903	886	786
1177	1806	1713	1606
1198	2234	2164	1893
1200	2618	1981	2289
1227	1139	1180	1040
1227	1941	1900	1955
1539	1405	1395	1456
1700	2835	2519	2974
1887	1417	1588	1661
2000	4098	3671	2471
2100	3890	4184	3608
2634	1940	1762	1725
3800	4050	2718	3943
6800	2859	2726	4752
8583	4223	9166	11878
20000	5133	10853	12019

F. SPECIFICITES TECHNIQUES

Les modèles ont été construit à l'aide d'un serveur de calcul optimisé sous Linux, mis à disposition par le cabinet Prim'Act.

C'est une caractéristique importante à prendre en compte, puisqu'il est probable que certains algorithmes présentés ici n'aboutissent pas sur des machines plus standard. La puissance de calcul est un élément clé dans le machine learning, qui couplé à l'optimisation de la complexité de l'algorithme constitue un élément primordial qu'il ne faut pas négliger.

Le serveur de calcul est muni de 64 go de mémoire RAM, et d'un processeur à 32 cœurs disposant d'une fréquence de fonctionnement de 2,6 GHz et d'une mémoire cache de 40 Mo.

On note ici les durées approximatives de calcul de chaque algorithme, dans la phase d'apprentissage et dans la phase d'optimisation des paramètres. Il est à noter que la phase d'optimisation des paramètres dépend des besoins de l'utilisateur, comme nous avons pu le montrer avec le GBM.

Tab. 42. **Temps de calcul en fonction de l'algorithme utilisé**

	Apprentissage	Optimisation des paramètres
GLM	5 min	
RGLM simple	1h	12h
RGLM amélioré	15 min	2h
RGLM méthode "meilleur modèle"	15 min	2h
GBM simple	2 min	1h
GBM optimisé	2 min	1h

Tous les algorithmes à l'exception du GLM sont parallélisés, et particulièrement le GBM qui utilise des techniques de parallélisation avancées.

Bibliographie

Bellina, R. 2014. Méthodes d'apprentissage appliquées à la tarification non-vie. 2014.
Bergstra, J. et Bengio, Y. 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research.* 2012, Vol. 13.
Boyer Chammard, R. 2007. Processus de surveillance et de majoration des contrats flottes d'entreprise d'AXA France. 2007.
Breihman. 1996. Bagging Predictors. *Machine Learning.* Springer, 1996, Vol. 26.
Breiman, Friedman, Olshen and Stone. 1983. *Classification and Regression Trees.* s.l. : Wadsworth, 1983.
Charpentier, A. et Denuit,M. 2005. *Mathématiques de l'assurance non-vie, tome 2.* s.l. : Economica, 2005.
FFA. 2016. L'assurance automobile des particuliers en 2015. *Fédération Française de l'Assurance.* [En ligne] 01 06 2016. https://www.ffa-assurance.fr/content/assurance-automobile-des-particuliers-en-2015.
François, N. 2014. Sujet d'étude : distance de Levenshtein. 2014.
Freund, Y. et Schapire, R. E. 1996. Experiments with a New Boosting Algorithm. 1996.
Friedman, J. H. 2001. Greedy Function Approximation : a gradient boosting machine. 2001.
Friedman, J. H.. 1999. Stochastic Gradient Boosting. 1999.
Ghemawat, S. et Dean, J. 2004. MapReduce : SImplified Data Processing on Large Clusters. *OSDI.* 2004.
INSEE. 2017. Estimation de la population au 1er Janvier 2016. [En ligne] 17 Janvier 2017. https://www.insee.fr/fr/statistiques/1893198.
KOLLI, Z. 2012. *Dynamique de renouvellement du parc automobile.* Paris : s.n., 2012.
LA POSTE. 2017. Base officielle des codes postaux. *data gouv.* [En ligne] 28 Mars 2017. https://www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux/.
Legoff, E. 2016. L'assurance automobile connectée représentera 12% du marché français en 2020. *Argus de l'assurance.* [En ligne] 15 11 2016. http://www.argusdelassurance.com/acteurs/l-assurance-automobile-connectee-representera-12-du-marche-francais-en-2020-etude-deloitte.113193.
Miller, H. 2015. A discussion on credibility and penalised regression, with implications for actuarial work. 2015.
Miseray, A. et Planchet, F. 2017. Tarification IARD : Introduction aux techniques avancées. 2017.
Nykodym, T. 2015. Genralized linear modeling with h2o. 2015.
Paglia, A. 2011. Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. 2011, Vol. 11, 22.
Ridgeway, G. 2007. Generalized Boosted Models : A guide to the gbm package. 2007.

Song L., Langfelder P. et Horvath, S. 2013. Random generalized model : a highly accurate and interpretable ensemble predictor. 2013.

Tufféry, S. 2015. *Modélisation prédictive et apprentissage statistique avec R.* s.l. : Technip, 2015.