Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires

le 2 4 / 0 1 / 2 0 2 2

Par : Mathilde CLEMENT
Titre : Utilisation d'arbres de régression pour la prédiction de coûts automobiles

Confidentialité : ☑ Non ☐ Oui (Durée : ☐ 1 an ☐ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : Prim'Act
Signature :

Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :

Directeur de Mémoire en entreprise :
Nom : Frédéric PLANCHET
et Maxime BEN-BRIK
Signature :

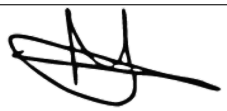
Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité)

Secrétariat :

Signature du responsable entreprise

Bibliothèque :

Signature du candidat



Résumé

Dans un contexte de concurrence forte, il est souhaitable pour un assureur de proposer des tarifs attractifs. Les assureurs automobiles étant couramment amenés à travailler avec des experts, un des aspects qu'ils peuvent piloter est le choix des réseaux d'experts avec lesquels ils travaillent. Ces derniers participent en effet à la détermination des coûts de sinistres survenus.

Ce mémoire se concentre sur ce dernier point : l'objectif est de trouver un modèle qui permette de prédire au mieux le coût des sinistres tout en permettant la mesure de l'influence du réseau d'experts intervenu. La base de l'étude regroupe des sinistres automobiles provenant de trois assureurs ayant eu recours à deux réseaux d'experts différents. Pour répondre à l'objectif de l'étude plusieurs méthodes de modélisation seront étudiées.

La première modélisation appliquée sera un modèle linéaire généralisé (GLM). Ce modèle, couramment utilisé en tarification automobile, permet une lecture aisée de l'influence des variables sur la prédiction. Il sera utilisé comme référentiel pour évaluer les améliorations fournies par les autres modèles. Des modèles à structures arborescentes seront ensuite étudiés : le modèle CART et le modèle MOB. Ces modèles d'arbres de régression sont réputés pour leur lisibilité et facilité de compréhension. L'étude sera complétée par l'application du modèle XGBoost. Ce modèle ne permettant pas une lecture directe de l'influence des variables sur la prédiction, son étude est principalement effectuée dans le but de comparer les résultats obtenus à ceux d'un modèle réputé pour sa qualité de prédiction. Ce modèle sera également analysé afin d'essayer d'améliorer les performances du modèle MOB, notamment par l'introduction de facteurs d'interactions.

Mots-clés : *Coûts automobiles*, *Arbres de régression*, *Machine learning*, *GLM*, *CART*, *MOB*, *XGBoost*, *Interprétabilité*.

Abstract

In a context of strong competition, it is desirable for an insurer to offer attractive rates. Since automobile insurers are often required to work with experts, one of the aspects they can control is the choice of the expert network they work with. The latter are involved in determining the costs of claims incurred.

This thesis focuses on this last point: the objective is to find a model which allows to predict the cost of the claims incurred as well as possible while allowing the measurement of the influence of the network of experts intervened. The database of the study gathers automobile claims from three insurers using two different networks of experts. To meet the objective of the study, several modeling methods will be studied.

The first model applied will be a generalized linear model (GLM). This model, commonly used in automobile pricing, allows an easy reading of the influence of variables on the prediction. It will be used as a reference to evaluate the improvements provided by the other models. Models with tree structures will then be studied: the CART model and the MOB model. These regression tree models are known for their readability and ease of understanding. The study will be completed by the application of the XGBoost model. As this model does not allow a direct reading of the influence of the variables on the prediction, its study in mainly carried out in order to compare the results obtained with a model known for its prediction quality. This model will also be analyzed in order to try to improve the performances of the MOB model, in particular by introducing interaction factors.

Keywords : *Automobile costs*, *Regression trees*, *Machine learning GLM*, *CART*, *MOB*, *XGBoost*, *Interpretability*.

Note de Synthèse

L'assurance automobile représentait en 2020 plus de 55% des cotisations des particuliers sur le marché de l'assurance de biens et de responsabilités. En ne prenant pas en compte l'année 2020, touchée par la crise sanitaire de la COVID-19, le secteur de l'assurance automobile possède un ratio combiné déficitaire depuis une décennie. De nombreux acteurs sont présents sur le marché de l'assurance automobile, impliquant l'existence d'une concurrence forte. Dans ce contexte, les assureurs automobiles recherchent des moyens d'être toujours plus compétitifs. Un des aspects qu'ils peuvent piloter est le choix des réseaux d'experts avec lesquels ils travaillent. Ces derniers jouent en effet un rôle important dans la détermination du montant final versé par l'assureur à ses assurés.
L'étude présentée dans ce mémoire se focalise sur ce dernier point. Plus précisément, l'objectif de ce mémoire est double : trouver un modèle permettant de prédire au mieux le coût des sinistres et mesurer l'influence du réseau d'experts intervenu sur ce coût.

Cadre de l'étude

L'étude ici présentée s'inscrit dans le cadre d'une mission du cabinet de conseil Prim'Act dont l'objectif était de comparer la performance de deux réseaux d'experts. La base de données initialement considérée résulte de la concaténation de données provenant sept assureurs et la période d'observation est de deux années glissantes : elle s'étend du 1er janvier 2019 au 31 décembre 2020. Au sein de cette base se trouve une variable représentant lequel des deux réseaux d'experts est intervenu : ces réseaux ont été anonymisés et seront appelés par la suite réseau d'experts A et réseau d'experts B.

Afin de pouvoir appliquer différents modèles à la base, des retraitements sont effectués, comme par exemple la suppression des variables associées à un nombre trop élevé de valeurs manquantes, la discrétisation des variables numériques ou bien la définition de seuils concernant les sinistres. À l'issue de ces retraitements la base de données finale contient 757 633 observations et 10 variables. Parmi ces variables se trouve la variable à prédire (le coût du sinistre) et des variables explicatives : le réseau d'experts intervenu (A ou B), l'assureur en charge du sinistre, le type de sinistre, le taux horaire du garage intervenu, l'âge, le kilométrage et la marque du véhicule sinistré. Le mois et le trimestre de survenance du sinistre sont également présents dans la base finale, mais dans le cadre d'une approche prospective l'intégration de variables temporelles de la sorte n'est pas pertinente : elles ne seront donc pas utilisées comme variables explicatives au sein des différents modèles.

Afin d'entraîner les différents modèles de prédiction et de limiter le phénomène de surapprentissage, la base de données finale est séparée en trois échantillons : un échantillon d'apprentissage, utilisé pour la construction des modèles, un échantillon de validation, utilisé pour l'optimisation des modèles, et un échantillon de test, utilisé pour leurs comparaisons. Les performances des modèles sont évaluées avec pour métriques la MSE (*Mean Square Error* en anglais) et la MAE (*Mean Absolute Error* en anglais).

Modèles linéaires généralisés

Le premier modèle étudié est le modèle linéaire généralisé, souvent appelé GLM (pour *Generalized Linear Model* en anglais). Ce modèle, usuellement utilisé en tarification automobile, permet une lecture aisée de l'influence des variables sur la prédiction via l'étude des coefficients retournés par ce dernier.

Présentation du modèle

Le GLM est la résultante de trois composantes :
— une composante aléatoire, les observations (ici le coût des sinistres), dont la loi est supposée appartenir à la famille exponentielle ;
— une composante déterministe, qui est une combinaison linéaire des variables explicatives avec des facteurs de régression ;
— une fonction de lien, permettant de lier les deux composantes précédentes.
Il est choisi d'utiliser la loi Gamma pour le coût des sinistres et pour fonction de lien la fonction log, ce qui permet de conduire à un modèle multiplicatif.

Avant d'appliquer le modèle pour effectuer des prédictions sur l'échantillon de validation, une sélection de variables spécifique à la régression est effectuée. Pour cela plusieurs méthodes peuvent être considérées, comme la sélection *forward*, la sélection *backward* et la sélection *stepwise*. L'application de ces trois méthodes conduit à la même conclusion : l'ensemble des sept variables explicatives sont conservées pour l'application du GLM.

Pouvoir prédictif du modèle

Afin de s'assurer que le modèle linéaire généralisé est bien adapté aux données de l'étude, un test de la déviance est effectué et les résidus de déviance sont étudiés. Ces analyses permettent de conclure de l'adéquation du modèle linéaire généralisé.
Une validation croisée sur l'échantillon d'apprentissage est ensuite effectuée afin de s'assurer que le modèle ne conduit pas au surapprentissage. L'idée est de séparer l'échantillon de validation en cinq blocs et d'utiliser à tour de rôle chacun des blocs comme échantillon de test pour évaluer le modèle GLM entraîné sur l'échantillon composé des quatre blocs restants. Si une forte variance des métriques est observée d'un bloc à l'autre cela signifie que le modèle n'a pas réussi à généraliser une règle de décision : le phénomène du surapprentissage est observé. En appliquant cette méthodologie, il est observé que le modèle linéaire généralisé n'effectue pas de surapprentissage et peut donc être entraîné sur l'ensemble de l'échantillon d'apprentissage pour effectuer des prédictions.

Les performances du modèle linéaire généralisé sont comparées à celles obtenues avec un tarif mutualisé, *i.e.* un modèle associant comme prédiction à chaque observation la moyenne des coûts de la base d'apprentissage, à savoir 2 294€. Les résultats des métriques obtenus pour le modèle GLM et le tarif mutualisé sont résumés dans le tableau (1) ci-dessous :

	MSE	MAE
Tarif mutualisé	6 588 897	1 560
GLM	4 581 209	1 293

Table 1 : Résultats de prédiction sur l'échantillon de validation pour un tarif mutualisé et le GLM

Le modèle GLM permet une amélioration de plus de 30% de la MSE et de plus de 17% de la MAE par rapport au tarif mutualisé. Au vu de l'objectif de l'étude, un autre aspect à prendre en compte est la mesure de l'influence du réseau d'experts.

Mesure de l'influence du réseau d'experts

La variable associée au réseau d'experts possède deux modalités : A ou B. Le modèle linéaire généralisé retourne alors un unique coefficient associé à cette variable, traduisant comment le fait d'appartenir à la modalité A plutôt qu'à la modalité B influe sur la prédiction du coût du sinistre. Ce coefficient est égal à -0.0287283. L'interprétation est alors immédiate : être expertisé par le réseau d'experts A plutôt que le réseau d'experts B fait baisser le coût de 2.9% en moyenne.

Arbres CART

Présentation du modèle

Le modèle CART fait partie de la famille des arbres de régression. Ce modèle est réputé pour sa facilité d'interprétation et de lisibilité des résultats. L'idée principale est de partir d'un nœud regroupant l'ensemble des individus, appelé racine, puis de construire une série de nœuds permettant à chaque étape le partitionnement binaire des individus en groupes les plus homogènes possible. La construction de l'arbre CART contient deux étapes principales :
— la construction d'un arbre binaire maximal ;
— l'élagage de cet arbre maximal selon un critère pré-spécifié pour obtenir l'arbre optimal.

Pour choisir l'arbre optimal, l'algorithme construit une séquence de sous-arbres emboîtés en partant de l'arbre maximal, puis en élaguant progressivement pour finir par l'arbre réduit à la racine. L'arbre optimal est choisi au sein de cette séquence selon un critère spécifié par l'utilisateur.
Au sein de chacun des nœuds terminaux de l'arbre optimal, appelés feuilles, une valeur est associée : il s'agit de la moyenne des coûts des sinistres présents dans la feuille.

Pouvoir prédictif du modèle

Un arbre optimal est construit selon la méthodologie décrite précédemment. Une restriction d'une profondeur maximale égale à 4 est imposée, ce qui signifie qu'au maximum l'arbre construit conduira à la création de $2^4 = 16$ groupes. Ce choix permet de ne pas conduire à une segmentation trop fine des individus et de conserver l'aspect facilité d'interprétation et lisibilité aisée des arbres CART.
Pour choisir l'arbre optimal lors de la procédure d'élagage, il est choisi d'utiliser le critère du 1 écart-type : l'arbre choisi est le plus petit arbre dont l'erreur de validation est inférieure à la somme entre la plus petite erreur de validation commise et l'écart-type estimé de cette erreur. L'arbre construit selon ces critères conduit à la création de 14 sous-groupes. Il est représenté en figure (1) ci-après.

Cet arbre, construit sur l'échantillon d'apprentissage, est utilisé pour effectuer des prédictions sur l'échantillon de validation. Les résultats obtenus sont présentés dans le tableau (2) ci-dessous :

MSE	MAE
4 137 226	1 266

Table 2 : Résultats de prédiction sur l'échantillon de validation avec l'arbre CART optimal

En termes de pouvoir prédictif, les résultats obtenus avec CART sont meilleurs que ceux obtenus avec le modèle GLM : une amélioration de l'ordre de 10% de la MSE et de 2% de la MAE est observée.

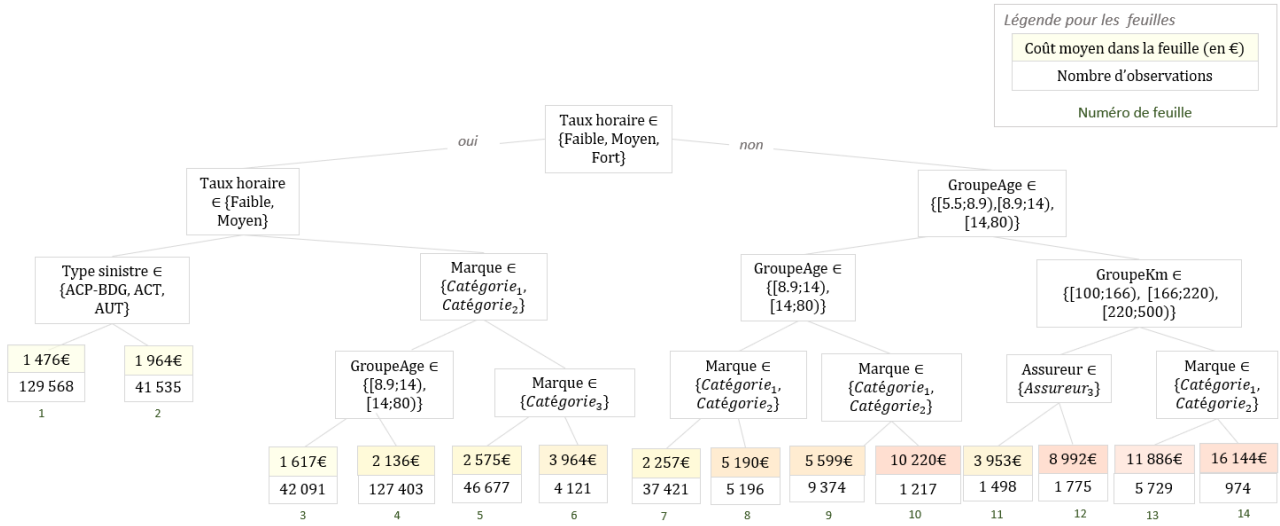


FIGURE 1 : Arbre CART obtenu après élagage selon la règle du 1 écart-type

Mesure de l'influence du réseau d'experts

Pour analyser l'influence d'une variable dans modèle CART, il suffit d'observer comment cette variable est utilisée pour partitionner les individus. Au vu des restrictions qui ont été imposées lors de la construction de l'arbre CART, la variable représentant le réseau d'experts intervenu n'apparaît pas dans l'arbre considéré. La mesure de l'influence de cette variable est alors limitée, le modèle CART considéré ne permet donc pas de répondre à cet objectif de l'étude. Un autre modèle à structure arborescente est alors étudié.

Arbres MOB

Présentation du modèle

Le modèle MOB, pour *MOdel Based recursive partitionning* en anglais, fait partie de la famille des algorithmes de partitionnement récursif basé sur les modèles. Ces modèles peuvent être considérés comme un compromis entre amélioration du pouvoir de prédiction des arbres et conservation de la facilité d'interprétation du modèle. Ils sont basés sur l'idée suivante : il peut être possible d'améliorer les résultats d'un modèle en partitionnant la population en sous-groupes et en ajustant dans chacun des groupes un modèle, plutôt qu'en appliquant un unique modèle à l'ensemble de la population. Le modèle considéré dans ce mémoire est le modèle MOB associé à des modèles GLM.
Pour construire un arbre MOB, une première étape consiste à diviser l'ensemble des variables explicatives en deux sous-groupes :
— les variables de partitionnement, permettant la création des groupes d'individus à chaque étape de l'algorithme ;
— les variables de régression, permettant l'ajustement des modèles.

L'algorithme part de l'ensemble des individus et construit une série de divisions binaires de façon à construire un arbre optimal. À chaque étape, au sein d'un nœud h, l'algorithme ajuste un modèle GLM en utilisant les variables de régression, puis, pour décider si le nœud est divisé, un test d'instabilité des paramètres par rapport aux variables de partitionnement est effectué. Si une instabilité significative est détectée, les individus du nœud h sont divisés en deux nœuds fils selon la variable de partitionnement associée à la plus grande instabilité. Sinon, la procédure dans le nœud s'arrête : il devient alors une

feuille dans laquelle un modèle GLM est ajusté.

Pouvoir prédictif du modèle

Il est choisi, tout comme pour l'arbre CART, d'imposer une restriction sur la profondeur maximale de l'arbre MOB. Pour construire cet arbre, il faut tout d'abord choisir quelles sont les variables de régression et de partitionnement. Pour effectuer ces choix, une validation croisée sur cinq blocs est effectuée : cette étape permet de s'assurer que le choix de la profondeur maximale et des sous-groupes de variables n'est pas fait pour convenir au mieux à l'échantillon de validation, mais bien pour être le meilleur choix de façon générale. Cette validation croisée conduit à imposer une profondeur maximale égale à 4 et à partitionner les individus selon la marque du véhicule sinistré, le taux horaire du garage intervenu et l'assureur en charge du sinistre.
L'arbre construit avec ces paramètres conduit à la création de huit sous-groupes. Une représentation graphique de cet arbre est donnée par la figure (2) ci-après.

Cet arbre est utilisé pour effectuer des prédictions sur l'échantillon de validation, les résultats obtenus sont présentés dans le tableau (3) ci-dessous :

MSE	MAE
3 799 924	1 202

Table 3 : Résultats de prédiction sur l'échantillon de validation pour le modèle MOB

Les résultats de prédiction obtenus avec MOB sont meilleurs que ceux obtenus avec les modèles CART et GLM. En effet, une amélioration de l'ordre de 8% de la MSE et de 5% de la MAE est observée par rapport à CART.

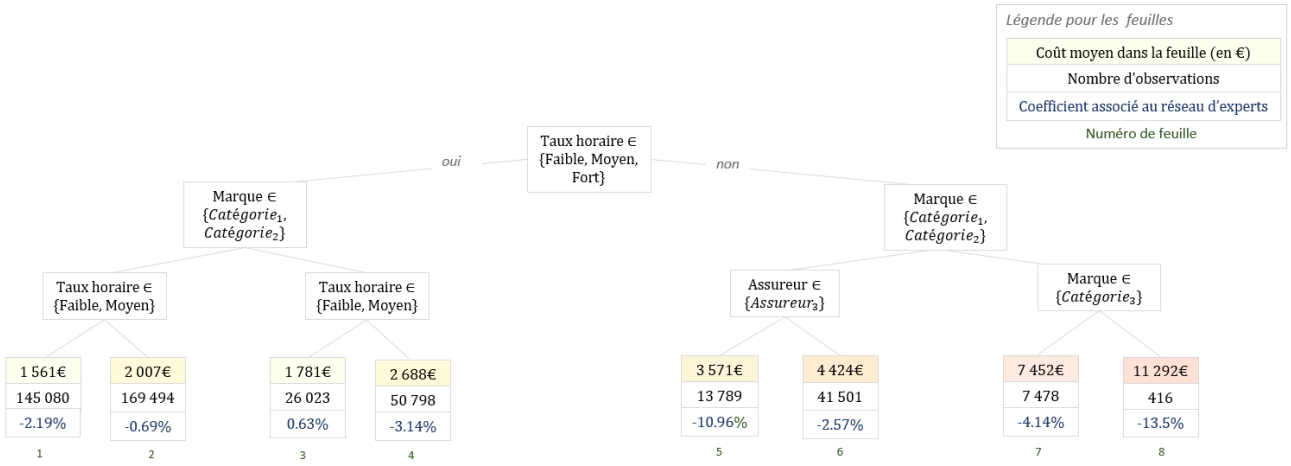


Figure 2 : Arbre construit par MOB

Mesure de l'influence du réseau d'experts

La variable spécifiant le réseau d'experts intervenu est utilisée comme variable de régression. Elle intervient donc dans chacun des huit modèles GLM associés aux feuilles de l'arbre. La mesure de l'influence se fait donc similairement à ce qui a été vu pour le modèle GLM précédemment : pour chacun des 8 sous-groupes un coefficient traduisant l'influence du réseau d'experts sur la prédiction est retourné. Les coefficients estimés dans chaque feuille ainsi que les intervalles de confiance à 95% qui leurs sont associés sont représentés dans la figure (3) ci-dessous :

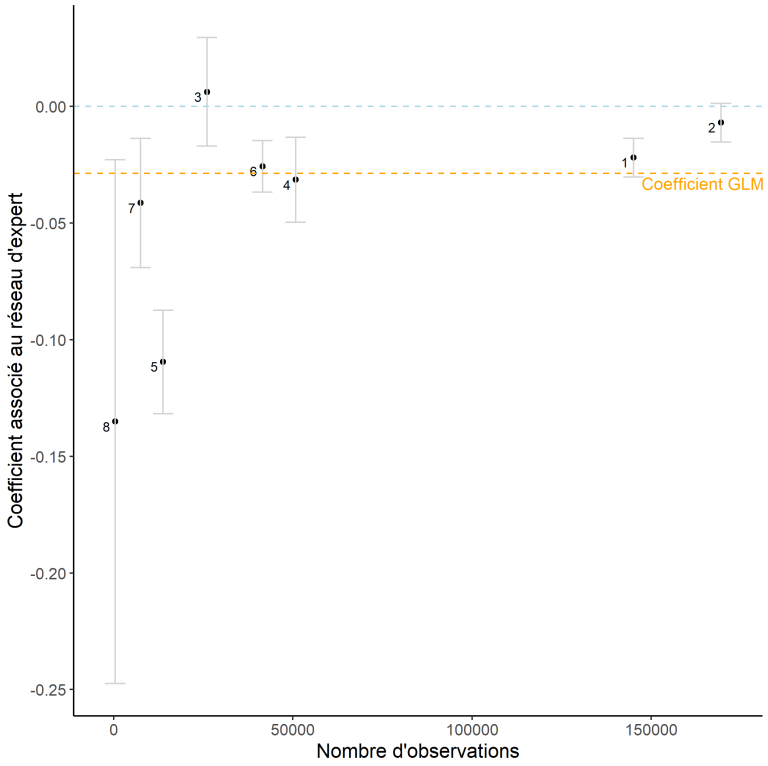


FIGURE 3 : Coefficient associé au réseau d'experts en fonction du nombre d'observations par feuille

Il est observé que pour sept des huit sous-groupes créés, le coefficient est négatif, signifiant qu' avoir recours au réseau d'experts A plutôt qu'au réseau d'experts B fait diminuer le coût des sinistres en moyenne. Concernant le sous-groupe pour lequel le coefficient est positif, *i.e.* le groupe des individus appartenant à la feuille n°3, la représentation des intervalles de confiance permet d'observer qu'il existe une incertitude sur le signe du coefficient.

Ainsi, la lecture de l'influence du réseau d'experts dans le modèle MOB est aussi aisée que dans un modèle GLM, tout en étant plus précise.

Amélioration du pouvoir prédictif des modèles : étude du modèle XGBoost

Le modèle XGBoost (pour *eXtreme Gradient Boosting* en anglais) est un modèle réputé pour ses performances de prédiction. Il fait partie des modèles basés sur des méthodes d'ensemble, c'est à dire utilisant plusieurs sous-modèles pour effectuer une prédiction finale. Cette multiplicité de sous-modèles permet d'obtenir un pouvoir prédictif élevé, en contrepartie d'une interprétabilité souvent limitée qui vaut au modèle XGBoost la caractérisation de modèle "boîte noire".

Le modèle XGBoost est optimisé par validation croisée puis appliqué pour prédire les coûts de l'échantillon de validation. Les résultats obtenus sont présentés dans le tableau (4) ci-dessous :

MSE	MAE
3 706 014	1 186

TABLE 4 : Évaluation des métriques pour XGBoost sur l'échantillon de validation

Le modèle XGBoost présente effectivement de meilleurs résultats de prédiction que les trois modèles étudiés précédemment.

Il est alors choisi d'analyser le modèle XGBoost afin d'améliorer le pouvoir prédictif de ces modèles. Le modèle MOB étant celui présentant les meilleurs résultats de prédiction, l'étude se concentrera sur l'amélioration de ce dernier. L'idée est d'analyser comment le modèle XGBoost utilise les variables explicatives et d'intégrer certains effets au sein du modèle MOB. Pour cela des outils d'analyse agnostiques sont utilisés :
— l'importance par permutation, pour comparer les importances des variables au sein des différents modèles ;
— les H-statistiques relatives et absolues de Friedman, pour mesurer les forces d'interactions entre variables prises en compte par les différents modèles.
L'analyse de ces outils a permis d'observer que l'ordre d'importance des variables était sensiblement le même entre les deux modèles. Concernant les forces d'interactions, des différences relativement importantes ont été observées.

La méthodologie choisie consiste alors à associer à chaque couple de variables explicatives (V_1, V_2) un score

$$score(V_1,V_2) = \underbrace{(int^{rel}_{XGB}(V_1,V_2) - int^{rel}_{MOB}(V_1,V_2))}_{\text{différence d'interactions relatives}} + \underbrace{(int^{abs}_{XGB}(V_1,V_2) - int^{abs}_{MOB}(V_1,V_2))}_{\text{différence d'interactions absolues}},$$

où $int^{rel}_{mod}(V_1,V_2)$ et $int^{abs}_{mod}(V_1,V_2)$ désignent respectivement les H-statistiques de Friedman relatives et absolues entre les variables V_1 et V_2 associées au modèle mod et où la notation XGB fait référence au modèle XGBoost.

L'ajout d'interaction(s) se fait par ordre décroissant du score jusqu'à ce que l'ajout d'interaction ne conduise plus à une amélioration des métriques. Cette méthodologie conduit à intégrer des facteurs d'interactions associés aux trois couples de variables ayant le score le plus élevé, à savoir :
— l'assureur et le kilométrage du véhicule ;
— l'âge et le kilométrage du véhicule ;
— l'assureur et l'âge du véhicule.

Le modèle MOB avec ajout de ces interactions, noté MOB_I, conduit à la même segmentation des individus que celle induite par l'arbre représenté en figure (2) ci-dessus. Au sein des huit modèles GLM associés à chacune des feuilles de cet arbre, des facteurs d'interactions ont été intégrés. Cet arbre est alors utilisé pour effectuer des prédictions sur l'échantillon de validation. Les résultats obtenus sont présentés dans le tableau (5) ci-dessous :

MSE	MAE
3 749 422	1 193

Table 5 : Évaluation des métriques pour le modèle MOB avec intégration d'interactions sur l'échantillon de validation

Ainsi, l'ajout d'interactions dans le modèle MOB permet d'en améliorer le pouvoir prédictif, avec une amélioration de la MSE de l'ordre de 1,3% et de la MAE de l'ordre de 0,7%.

Les modèles GLM, CART, MOB et XGBoost ont été entraînés et optimisés à partir des échantillons d'apprentissage et de validation. La comparaison de ces modèles sur l'échantillon de test peut alors être effectuée.

Comparaison des modèles

La comparaison finale des modèles se fait sur l'échantillon de test, jusqu'à présent inutilisé. Cette étape a pour but de déterminer quel modèle est le meilleur pour répondre aux objectifs de l'étude ici présentée.

Pouvoir prédictif

Un premier point de comparaison est le pouvoir prédictif associé à chaque modèle. Les modèles optimisés sont utilisés pour prédire les coûts de l'échantillon de test. Les résultats obtenus sont présentés dans le tableau (6) ci-dessous :

	MSE	MAE
GLM	4 656 265	1 293
CART	4 139 728	1 261
MOB	3 816 635	1 198
$\mathbf{MOB_I}$	3 764 333	1 189
XGB	3 714 200	1 182

Table 6 : Résultats de prédiction des modèles sur l'échantillon de test

Une hiérarchisation des performances similaire à ce qui a été observé sur l'échantillon de validation est observée. Le modèle XGBoost est à nouveau le modèle présentant le plus grand pouvoir de prédiction et l'intégration d'interactions au modèle MOB permet encore d'en améliorer le pouvoir prédictif. Au vu de l'objectif de l'étude, un autre aspect à prendre en compte est la mesure de l'influence du réseau d'experts au sein des différents modèles.

Mesure de l'influence du réseau d'experts

La lecture de l'impact d'une variable sur la prédiction diffère selon le modèle utilisé. La façon dont l'influence du réseau a pu être mesurée pour les différents modèles est résumée ci-après.

Dans un modèle GLM
Pour une variable qualitative, le modèle GLM retourne des coefficients traduisant comment le fait d'appartenir à une certaine modalité plutôt qu'à une modalité de référence influe sur la prédiction. La variable représentant le réseau d'experts possédant seulement deux modalités, un unique coefficient est retourné par le modèle et son interprétation est simple : si le coefficient est négatif, être expertisé par le réseau A plutôt que le réseau B fait baisser le coût en moyenne, et inversement. La valeur du coefficient permet de quantifier l'impact du choix d'un certain réseau d'experts.

Dans un modèle CART
Le modèle CART permet une lecture aisée et directe de l'influence des variables intervenant dans le partitionnement des individus. Dans le cadre de ce mémoire, les restrictions imposées sur la profondeur de l'arbre ont conduit à construire un arbre CART ne faisant pas apparaître la variable spécifiant le réseau d'experts intervenu. Cela limite la possibilité de mesure de son influence : le modèle CART ne permet pas ici une bonne lecture de l'influence du réseau d'experts sur le coût des sinistres.

Dans un modèle MOB
Dans ce mémoire la variable représentant le réseau d'experts a été utilisée comme variable de régression. La lecture de l'influence du réseau intervenu s'effectue alors en observant le coefficient estimé par le

modèle GLM dans chaque feuille de l'arbre. L'arbre construit dans cette étude contenant 8 feuilles, 8 coefficients ont été retournés. Le modèle MOB permet donc une lecture aussi aisée que dans un GLM, tout en étant plus précise.

Dans un modèle XGBoost
Bien que le modèle soit souvent caractérisé de modèle "boîte noire", la lecture de l'influence des variables sur la prédiction n'est pas impossible, notamment grâce à l'utilisation d'outils agnostiques tels que SHAP ou LIME. Cependant, au vu du rôle du modèle XGBoost dans ce mémoire, l'analyse de l'influence du réseau d'experts dans ce modèle n'a pas été étudiée.

Afin de terminer la comparaison des modèles, un dernier aspect à étudier est le temps de calcul qu'ils requièrent.

Temps de calcul

Les temps de calcul des modèles finaux, *i.e.* des modèles optimisés, ainsi que le temps d'optimisation qu'ils ont requis sont représentés dans le tableau (7) ci-dessous :

	Temps de calcul	
	Modèles finaux	Optimisation
GLM	$\simeq$ 15s	$\simeq$ 10mn
CART	$\simeq$ 6s	$\simeq$ 11mn
MOB	$\simeq$ 1mn	$\simeq$ 22h
MOB$_\mathbf{I}$	$\simeq$ 3mn	$\simeq$ +1h
XGBoost	$\simeq$ 2mn	$\simeq$ 18h

TABLE 7 : Temps d'apprentissage des modèles finaux et temps d'optimisation

Les temps de calcul des modèles finaux sont tous relativement faibles, puisqu'ils n'excèdent pas trois minutes. Cependant, afin de maximiser leurs performances, les modèles ont été optimisés : il est possible d'observer que les modèles MOB et XGBoost requièrent des temps d'optimisation relativement importants. Le choix des interactions ajoutées au modèle MOB a nécessité une heure de calcul supplémentaire.

Par ailleurs, le modèle MOB nécessite des retraitements non obligatoires pour les autres modèles afin d'en diminuer le temps de calcul et d'en permettre l'optimisation. Par exemple, si les variables numériques n'avaient pas été discrétisées, la simple application du modèle avec une variable numérique au sein des variables de partitionnement nécessite plus de dix heures de temps de calcul. Sans ce retraitement, l'optimisation du modèle MOB aurait été trop complexe pour pouvoir être mise en place.

Conclusion

L'étude des trois aspects précédents permet de conclure sur le modèle le plus adapté pour répondre aux objectifs de l'étude. Malgré les retraitements supplémentaires et le temps de calcul important qu'il requière, le modèle MOB avec ajout d'interactions est le modèle le plus adapté pour répondre aux objectifs de l'étude ici présentée. Ce modèle possède un pouvoir prédictif proche de celui du modèle XGBoost et permet une lecture aisée et précise de l'influence du réseau d'experts. Concernant ce dernier point, le réseau A apparaît comme le réseau le plus performant en moyenne, les différences de performances entre les deux réseaux pouvant varier selon la période considérée ou les caractéristiques

des individus étudiés.

L'étude présentée dans ce mémoire présente des limites : le nombre de variables explicatives à disposition était relativement faible, en partie car la fusion et l'harmonisation de bases de données provenant de différents assureurs ont conduit à l'exclusion de certaines variables. Il est probable que l'intégration de variables explicatives additionnelles dans les modèles aurait pu permettre d'en améliorer les pouvoirs prédictifs. De plus, l'étude de variables supplémentaires aurait pu permettre une analyse plus approfondie des différences de performances entre les réseaux. L'utilisation d'un nombre faible de variables explicatives a permis d'effectuer une optimisation du modèle MOB qui n'aurait pas pu être mise en place si ce nombre avait été relativement trop important. De plus, malgré le nombre faible de variables utilisées, l'optimisation du modèle MOB requière des retraitements non-nécessaires pour les autres modèles et un temps de calcul relativement important.

Dans ce mémoire, l'étude s'est concentrée principalement sur trois modèles. Il est néanmoins possible qu'il existe d'autres outils ou modèles permettant d'améliorer les résultats obtenus. Récemment, le développement d'outils agnostiques permettant d'éclaircir le fonctionnement des modèles "boîte noire", comme le modèle XGBoost, peuvent être considérés comme une solution pour répondre aux objectifs de l'étude. Il existe également des modèles de partitionnement récursif basés sur des modèles GAM (pour *Generalised Additive Model* en anglais) ou encore des arbres de régression basés sur la maximisation de la vraisemblance (*Maximum Likelihood Regression Tree* en anglais) dont l'étude n'a pas été menée dans ce mémoire, mais qui pourraient constituer une piste d'exploration pour des travaux futurs sur le même type de sujet.

La recherche d'un compromis entre pouvoir prédictif et interprétabilité des modèles constitue un sujet primordial en assurance automobile, puisqu'il est important pour un assureur de proposer les tarifs les plus attractifs, tout en ayant la capacité de comprendre et d'expliquer la construction de ces derniers. Le développement de différentes méthodes de machine learning à travers le temps offre de nombreux moyens permettant de répondre à ce type de problématique. L'avancée dans le monde de la Data Science permettra très certainement dans un futur proche de disposer de nouveaux outils pour répondre à la problématique de cette étude.

Synthesis note

Automobile insurance accounted for more than 55% of individuals' contributions in the property and liability insurance market in 2020. Excluding the year 2020, which was affected by the COVID-19 health crisis, the auto insurance industry has had a combined ratio in deficit for a decade. There are many players in the auto insurance market, implying the existence of strong competition. In this context, automobile insurers are looking for ways to be ever more competitive. One of the aspects they can control is the choice of the expert networks they work with. The latter play an important role in determining the final amount paid by the insurer to its policyholders.
The study presented in this thesis focuses on this last point. More precisely, the objective of this thesis is twofold: find a model that best predicts the cost of claims and measure the influence of the network of experts involved on this cost.

Study environment

The study presented here is part of a mission of the consulting firm Prim'Act whose objective was to compare the performances of two expert networks. The database initially considered results from the concatenation of data from seven insurers and the observation period is two rolling years: it extends from January 1, 2019 to December 31, 2020. Within this database is a variable representing which of the two expert networks intervened: these networks have been anonymized and will hereafter be referred to as expert network A and expert network B.

In order to be able to apply different models to the database, data processing is made, such as the deletion of variables associated with too many missing values, the discretization of numerical variables or the definition of thresholds for claims. After these adjustments, the final database contains 757 633 observations and 10 variables. These variables include the variable to be predicted (the cost of the claims) and explanatory variables: the network of experts involved (A or B), the insurer in charge of the claim, the type of claim, the hourly rate of the garage involved, the age, the mileage and the make of the damaged vehicle. The month and the quarter of the occurrence of the claims are also present in the final database, but in the framework of a prospective approach the integration of such time variables is not relevant: they will therefore not be used as explanatory variables in the different models.

In order to train the different prediction models and to limit the phenomenon of overfitting, the final database is separated into three samples: a learning sample, used for the construction of the models, a validation sample, used for the optimization of the models, and a test sample, used for their comparisons. The performance of the models is evaluated using the MSE (Mean Square Error) and MAE (Mean Absolute Error) as metrics.

Generalized Linear Model

The first model studied is the generalized linear model, often called by its acronym GLM. This model, usually used in automobile pricing, allows an easy reading of the influence of the variables on the prediction via the study of the coefficients returned by the latter.

Presentation of the model

The GLM is the result of three components:
— a random component, the observations (here the cost of claims), whose distribution is assumed to belong to the exponential family;
— a deterministic component, which is a linear combination of the explanatory variables with regression factors;
— a link function, which links the two previous components.
It is chosen to use the Gamma distribution for the cost of claims and the log function for the link function, which leads to a multiplicative model.

Before applying the model to make predictions on the validation sample, a selection of variables specific to the regression is made. For this, several methods can be considered, such as forward selection, backward selection and stepwise selection. The application of these three methods leads to the same conclusion: all seven explanatory variables are retained for the application of the GLM.

Predictive power of the model

In order to ensure that the generalized linear model is well adapted to the data of the study, a deviance test is performed and the deviance residuals are studied. These analyses allow to conclude the adequacy of the generalized linear model.
A cross-validation on the training sample is then performed to ensure that the model does not lead to overfitting. The idea is to separate the validation sample into five blocks and use each block in turn as a test sample to evaluate the trained GLM model on the sample consisting of the remaining four blocks. If a high variance of the metrics is observed from one block to another, it means that the model has failed to generalize a decision rule: the phenomenon of overfitting is observed. By applying this methodology, it is observed that the generalized linear model does not overfit and can therefore be trained on the whole training sample to make predictions.

The performance of the generalized linear model is compared to the results obtained with a mutualized rate, *i.e.* a model associating as a prediction to each observation the average of the costs of the learning sample, *i.e.* 2 294€. The results of the metrics obtained for the GLM model and the mutualized rate are summarized in the table (8) below:

	MSE	MAE
Mutualized rate	6 588 897	1 560
GLM	4 581 209	1 293

Table 8: Prediction results on the validation sample for a mutualized rate and the GLM

The GLM model provides an improvement of more than 30% in MSE and more than 17% in MAE compared to the mutualized rate. Given the purpose of the study, another aspect to consider is measuring the influence of the expert network.

Measuring the influence of the expert network

The variable associated with the network of experts can take two values: A or B. The generalized linear model then returns a single coefficient associated with this variable, reflecting how belonging to modality A rather than to modality B influences the prediction of the cost of the claim. This coefficient is equal to -0.0287283. The interpretation is then immediate: being appraised by the network of experts A rather than the network of experts B reduces the cost by 2.9% on average.

CART trees

Presentation of the model

The CART model belongs to the family of regression trees. This model is known for its ease of interpretation and readability of results. The main idea is to start from a node gathering all the individuals, called root, and then to build a series of nodes allowing at each step the binary partitioning of the individuals into the most homogeneous groups possible. The construction of the CART tree contains two main steps:
— the construction of a maximal binary tree;
— the pruning of this maximal tree according to a pre-specified criterion to obtain the optimal tree.

To choose the optimal tree, the algorithm builds a sequence of nested subtrees starting from the maximal tree, then progressively pruning to end with the root-reduced tree. The optimal tree is chosen from this sequence according to a criterion specified by the user.
Within each of the terminal nodes of the optimal tree, called leaves, a value is associated: it is the average of the costs of the claims present in the leaf.

Predictive power of the model

An optimal tree is constructed using the methodology described above. A restriction of a maximum depth of 4 is imposed, which means that at most the tree constructed will lead to the creation of $2^4 = 16$ groups. This choice allows to avoid a too fine segmentation of the individuals and to keep the easy interpretation and readability of the CART trees. To choose the optimal tree during the pruning procedure, the 1 standard deviation criterion is used: the tree chosen is the smallest tree whose validation error is less than the sum of the smallest validation error committed and the estimated standard deviation of this error. The tree built according to these criteria leads to the creation of 14 subgroups. It is represented in the figure (4) below.

This tree, built on the learning sample, is used to make predictions on the validation sample. The results obtained are presented in the table (9) below:

MSE	MAE
4 137 226	1 266

Table 9: Prediction results on the validation sample with the optimal CART tree

In terms of predictive power, the results obtained with CART are better than those obtained with the GLM model: an improvement of about 10% of the MSE and 2% of the MAE is observed.

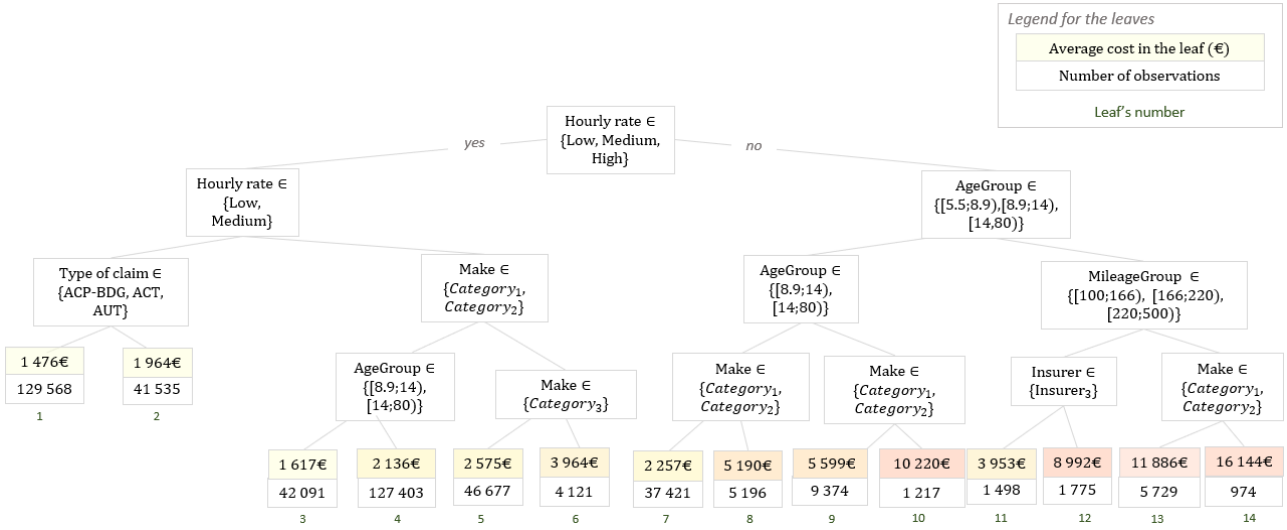


Figure 4: CART tree obtained after pruning according to the 1 standard deviation rule

Measuring the influence of the expert network

To analyze the influence of a variable in the CART model, it is possible to observe how this variable is used to partition the individuals. Given the restrictions that were imposed during the construction of the CART tree, the variable representing the network of experts involved does not appear in the tree considered. The measurement of the influence of this variable is then limited, and the CART model considered does not therefore make it possible to meet this objective of the study. Another model with a tree structure is then studied.

MOB trees

Presentation of the model

The MOB model, for MOdel Based recursive partitioning, belongs to the family of recursive partitioning algorithms based on models. These models can be considered as a compromise between improving the predictive power of the trees and keeping the ease of interpretation of the model. They are based on the idea that it may be possible to improve the results of a model by partitioning the population into subgroups and fitting a model to each of the groups, rather than applying a single model to the whole population. The model considered in this thesis is the MOB model associated with GLM models.
To build a MOB tree, a first step is to divide the set of explanatory variables into two subgroups:
— the partitioning variables, allowing the creation of groups of individuals at each stage of the algorithm;
— the regression variables, allowing the fitting of the models.

The algorithm starts from the set of individuals and builds a series of binary divisions in order to construct an optimal tree. At each step, within a h node, the algorithm fits a GLM model using the regression variables, and then, to decide whether the node is split, a test for instability of the parameters with respect to the partitioning variables is performed. If a significant instability is detected, the individuals in node h are split into two child nodes according to the partitioning variable associated with the greatest instability. Otherwise, the procedure in the node stops: it then becomes a leaf in which a GLM model is fitted.

Predictive power of the model

It is chosen, as for the CART tree, to impose a restriction on the maximum depth of the MOB tree. To build this tree, first the choice of the regression and partitioning variables has to be made. To make these choices, a cross-validation on five blocks is performed: this step ensures that the choice of the maximum depth and the subgroups of variables is not made to suit the validation sample, but to be the best choice in general. This cross-validation leads to impose a maximum depth equal to 4 and to partition the individuals according to the make of the damaged vehicle, the hourly rate of the intervening garage and the insurer in charge of the claim. The tree constructed with these parameters leads to the creation of eight subgroups. A graphical representation of this tree is given in the figure (5) below:

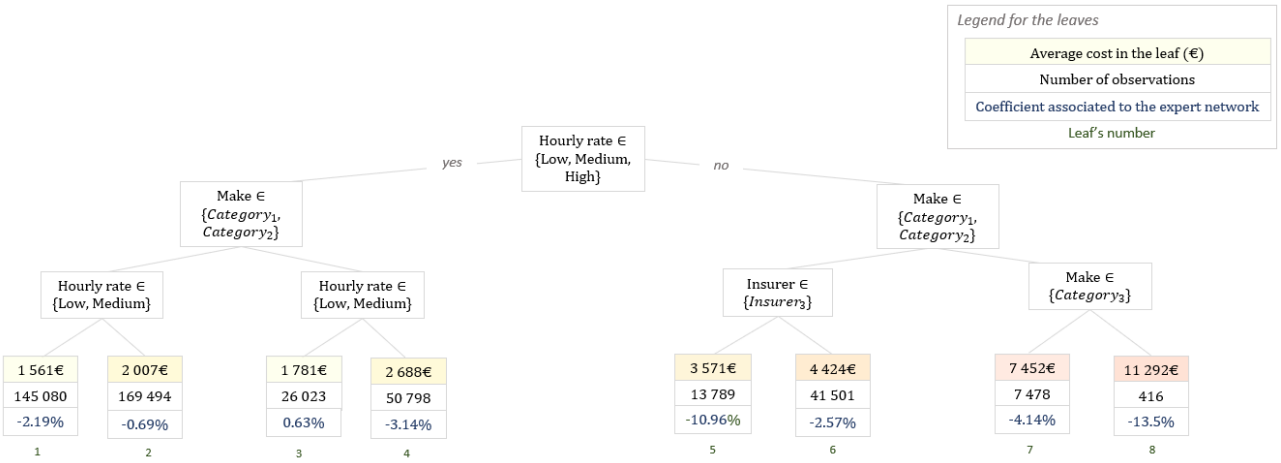


Figure 5: Tree built by MOB

This tree is used to make predictions on the validation sample, the results obtained are presented in the table (10) below:

MSE	MAE
3 799 924	1 202

Table 10: Prediction results on the validation sample for the MOB model

The prediction results obtained with MOB are better than those obtained with the CART and GLM models. Indeed, an improvement of the order of 8% of the MSE and 5% of the MAE is observed compared to CART.

Measuring the influence of the expert network

The variable specifying the network of experts involved is used as a regression variable. It is therefore used in each of the eight GLM models associated with the leaves of the tree. The influence is measured in a similar way to what was seen for the GLM model previously: for each of the 8 subgroups, a coefficient reflecting the influence of the expert network on the prediction is returned. The estimated coefficients in each leaf as well as the associated 95% confidence intervals are shown in the figure (6) below:

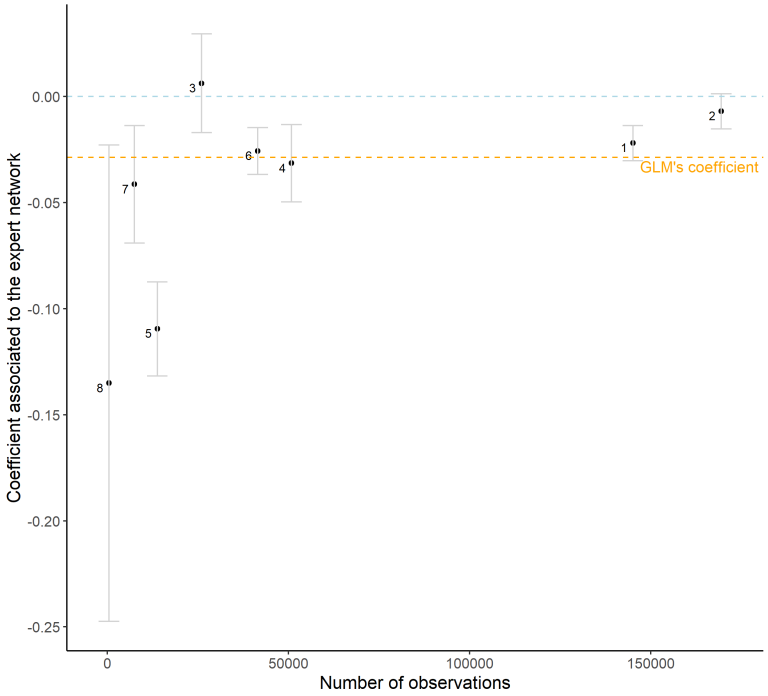


Figure 6: Coefficient associated with the expert network according to the number of observations per leaf

It is observed that for seven of the eight subgroups created, the coefficient is negative, meaning that using expert network A rather than expert network B decreases the cost of claims on average. For the subgroup for which the coefficient is positive, *i.e.* the group of individuals belonging to leaf n°3, the representation of the confidence intervals allows to observe that there is an uncertainty on the sign of the coefficient.
Thus, the reading of the influence of the expert network in the MOB model is as easy as in a GLM model, while being more precise.

Improving the predictive power of the models: study of the XGBoost model

The XGBoost model (for eXtreme Gradient Boosting) is a model known for its predictive performance. It is part of the models based on ensemble methods, *i.e.* using several sub-models to make a final prediction. This multiplicity of sub-models makes it possible to obtain a high predictive power, in exchange for an often limited interpretability which is why the XGBoost model is characterized as a "black box" model.

The XGBoost model is optimized by cross-validation and then applied to predict the costs of the validation sample. The results obtained are presented in the table (11) below:

MSE	MAE
3 706 014	1 186

Table 11: Evaluation of the metrics for XGBoost on the validation sample

The XGBoost model does indeed show better prediction results than the three models previously

studied.

It is then chosen to analyze the XGBoost model in order to improve the predictive power of these models. The MOB model being the one with the best prediction results, the study will focus on improving this model. The idea is to analyze how the XGBoost model uses the explanatory variables and to integrate some effects within the MOB model. For this, agnostic analysis tools are used:
— importance by permutation, to compare the importance of the variables within the different models;
— the relative and absolute Friedman's H-statistics, to measure the strength of the interactions between variables taken into account by the different models.
The analysis of these tools allowed to observe that the order of importance of the variables was approximately the same between the two models. Concerning the interaction forces, relatively important differences were observed.

The chosen methodology consists in associating a score to each pair of explanatory variables (V_1, V_2)

$$score(V_1, V_2) = \underbrace{(int^{rel}_{XGB}(V_1, V_2) - int^{rel}_{MOB}(V_1, V_2))}_{\text{relative interactions difference}} + \underbrace{(int^{abs}_{XGB}(V_1, V_2) - int^{abs}_{MOB}(V_1, V_2))}_{\text{absolute interactions difference}},$$

where $int^{rel}_{mod}(V_1, V_2)$ and $int^{abs}_{mod}(V_1, V_2)$ respectively denote the relative and absolute Friedman H-statistics between the variables V_1 and V_2 associated with the mod model and where the notation XGB refers to the XGBoost model.

The addition of interaction(s) is done in decreasing order of the score until the addition of interaction no longer leads to an improvement of the metrics. This methodology leads to the integration of interaction factors associated with the three pairs of variables with the highest score, namely:
— the insurer and the mileage of the vehicle;
— the age and the mileage of the vehicle;
— insurer and vehicle age.

The MOB model with the addition of these interactions, noted MOB_I, leads to the same segmentation of individuals as the one induced by the tree represented in the figure (5) above. Within the eight GLM models associated with each leaf of this tree, interaction factors have been integrated. This tree is then used to make predictions on the validation sample. The results obtained are presented in the table (12) below:

MSE	MAE
3 749 422	1 193

Table 12: Evaluation of the metrics for the MOB model with integration of interactions on the validation sample

Thus, the addition of interactions in the MOB model improves its predictive power, with an improvement of the MSE of the order of 1.3% and the MAE of the order of 0.7%.

The GLM, CART, MOB, and XGBoost models were trained and optimized from the training and validation samples. The comparison of these models on the test sample can then be performed.

Model comparison

The final comparison of the models is done on the previously unused test sample. The purpose of this step is to determine which model is the best to meet the objectives of the study presented here.

Predictive power

A first point of comparison is the predictive power associated with each model. The optimized models are used to predict the costs of the test sample. The results obtained are presented in the table (13) below:

	MSE	MAE
GLM	4 656 265	1 293
CART	4 139 728	1 261
MOB	3 816 635	1 198
$\mathbf{MOB_I}$	3 764 333	1 189
XGB	3 714 200	1 182

Table 13: Model prediction results on the test sample

A performance hierarchization similar to what was observed on the validation sample is observed. The XGBoost model is again the model with the highest predictive power and the integration of interactions to the MOB model still improves its predictive power. In view of the objective of the study, another aspect to consider is the measurement of the influence of the expert network within the different models.

Measure of the influence of the expert network

Reading the impact of a variable on prediction differs depending on the model used. The way in which the influence of the network could be measured for the different models is summarized hereinafter.

In a GLM model
For a qualitative variable, the GLM model returns coefficients reflecting how belonging to a certain modality rather than to a reference modality affects the prediction. Since the variable representing the expert network has only two modalities, a single coefficient is returned by the model and its interpretation is simple: if the coefficient is negative, being appraised by network A rather than network B lowers the cost on average, and vice versa. The value of the coefficient makes it possible to quantify the impact of choosing a certain network of experts.

In a CART model
The CART model allows an easy and direct reading of the influence of the variables intervening in the partitioning of the individuals. In the context of this thesis, the restrictions imposed on the depth of the tree have led to the construction of a CART tree which does not show the variable specifying the intervening network of experts. This limits the possibility of measuring its influence: the CART model does not allow a good reading of the influence of the expert network on the cost of claims.

In a MOB model
In this paper, the variable representing the expert network was used as a regression variable. The influence of the intervening network is then read by observing the coefficient estimated by the GLM model in each leaf of the tree. The tree constructed in this study contains 8 leaves, so 8 coefficients were returned. The MOB model thus allows a reading as easy as in a GLM, while being more precise.

In an XGBoost model
Although the model is often characterized as a "black box" model, it is not impossible to read the influence of the variables on the prediction, in particular thanks to the use of agnostic tools such as SHAP or LIME. However, in view of the role of the XGBoost model in this thesis, the analysis of the

influence of the network of experts in this model has not been studied

In order to finish the comparison of the models, a last aspect to study is the computation time they require.

Computing time

The computation times of the final models, *i.e.* the optimized models, as well as the optimization time they required are represented in the table (14) below:

	Computing time	
	Final models	Optimization
GLM	$\simeq$ 15s	$\simeq$ 10mn
CART	$\simeq$ 6s	$\simeq$ 11mn
MOB	$\simeq$ 1mn	$\simeq$ 22h
MOB_I	$\simeq$ 3mn	$\simeq$ +1h
XGBoost	$\simeq$ 2mn	$\simeq$ 18h

Table 14: Final models learning time and optimization time

The computation times of the final models are all relatively low, since they do not exceed three minutes. However, in order to maximize their performance, the models have been optimized: it is possible to observe that the MOB and XGBoost models require relatively long optimization times. The choice of the interactions added to the MOB model required an additional hour of computation.

Moreover, the MOB model requires some adjustments that are not mandatory for the other models in order to reduce its computation time and to allow its optimization. For example, if the numerical variables had not been discretized, the simple application of the model with a numerical variable within the partitioning variables would have required more than ten hours of computation time. Without this reprocessing, the optimization of the MOB model would have been too complex to implement.

Conclusion

The study of the three previous aspects allows to conclude which model is the most adapted to meet the objectives of the study. In spite of the additional reprocessing and the important computation time it requires, the MOB model with added interactions is the most adapted model to meet the objectives of the study presented here. This model has a predictive power close to the XGBoost model's one and allows an easy and precise reading of the influence of the expert network. Concerning this last point, network A appears to be the best performing network on average, although the differences in performance between the two networks may vary according to the period considered or the characteristics of the individuals studied.

The study presented in this paper has limitations: the number of explanatory variables available was relatively small, in part because the merging and harmonization of databases from different insurers led to the exclusion of some variables. It is likely that the inclusion of additional explanatory variables in the models could have improved their predictive powers. In addition, the study of additional variables could have allowed for a more in-depth analysis of the differences in performance between the expert networks. The use of a small number of explanatory variables allowed for an optimization of the MOB model that would not have been possible if the number of explanatory variables had been relatively too large. Moreover, despite the small number of variables used, the optimization

of the MOB model requires adjustments that are not necessary for the other models and a relatively long calculation time.

In this thesis, the study focused mainly on three models. However, it is possible that other tools or models exist to improve the results obtained. Recently, the development of agnostic tools allowing to clear up the functioning of "black box" models, such as the XGBoost model, can be considered as a solution to meet the objectives of the study. There are also recursive partitioning models based on GAM (Generalized Additive Model) models or regression trees based on the maximization of the likelihood (called Maximum Likelihood Regression Tree) which have not been studied in this thesis, but which could constitute an avenue of exploration for future work on the same type of subject.

The search for a compromise between predictive power and interpretability of the models is an essential subject in automobile insurance, since it is important for an insurer to propose the most attractive rates, while having the capacity to understand and explain the construction of the latter. The development of various machine learning methods over time offers many ways to address this type of issue. The progress in the world of Data Science will certainly allow in the near future to have new tools to answer the problematic of this study.

Remerciements

Je remercie en premier lieu mon tuteur de mémoire en entreprise : Frédéric Planchet, associé chez Prim'Act et membre agréé de l'institut des actuaires, pour la contribution qu'il a pu apporter à ce mémoire. Je tiens également à remercier Quentin Guibert, consultant chez Prim'Act et professeur à l'université Paris Dauphine, pour ses précieux conseils.

Je remercie également mon tuteur au sein de Prim'Act, Maxime Ben-Brik, pour son soutien et ses multiples conseils. Plus généralement, je remercie le cabinet de conseil Prim'Act et l'ensemble de ses consultants pour leur accueil et leur bienveillance durant toute la durée de mon stage.

Je témoigne également ma reconnaissance à Christophe Dutang, directeur du master actuariat à l'université Paris Dauphine, pour son investissement et ses conseils tout au long de l'écriture de ce mémoire.

Pour finir, j'adresse une attention particulière à mes amis et ma famille que je remercie pour leur soutien infaillible durant toutes mes années d'étude.

Table des matières

Introduction

L'assurance automobile hors flottes[1] représentait en 2020 plus de 55% des cotisations des particuliers sur le marché de l'assurance de biens et de responsabilités. En 2020, l'ensemble des cotisations affaires directes du secteur de l'assurance automobile s'élevait à 23,1 milliards d'euros, contre 22,8 milliards d'euros en 2019, soit une augmentation de 1,1%. Le ratio combiné[2] de ce secteur était de 96,1% en 2020 [FFA (2021b)]. À noter cependant que cette année a été affectée par la crise sanitaire mondiale de la COVID-19, la France ayant connu au cours de cette période deux confinements ainsi que des restrictions liées aux capacités de déplacement. Ces événements ont conduit à un changement dans le comportement des automobilistes et donc dans la sinistralité automobile.
En s'intéressant aux années précédentes, il ressort que le marché de l'assurance automobile possède un ratio combiné comptable (après réassurance) déficitaire depuis une décennie [FFA (2021a)]. Par ailleurs, les sinistres automobiles sont de moins en moins nombreux, mais de plus en plus coûteux, avec une hausse du coût moyen des accidents corporels de 5,7% par an en moyenne sur les dix dernières années [FFA (2021b)]. À ces difficultés viennent s'ajouter une concurrence importante, avec la présence de nombreux acteurs sur le marché français, la mise en place de règles de solvabilité et de nouvelles lois, comme la loi Hamon, entrée en vigueur le 1er janvier 2015, qui permet à un assuré de résilier son contrat d'assurance automobile à tout moment au bout d'un an de couverture, sans frais, ni pénalité.

Tous ces éléments conduisent les assureurs automobiles à chercher des moyens d'être toujours plus compétitifs. Pour cela, ils essayent de proposer des tarifs avantageux afin d'attirer et de fidéliser les clients, mais aussi de diversifier leurs offres et d'élargir leur réseau de distribution. Une caractéristique sur laquelle les assureurs peuvent influer est le choix des réseaux d'experts avec lesquels ils travaillent. Ces experts participent à l'évaluation des montant des sinistres en travaillant notamment auprès des garagistes. Pour un assureur, il est intéressant de mesurer la performance d'un réseau d'experts avec lequel il travaille puisque ces derniers jouent un rôle important dans la détermination du montant versé par l'assureur à son assuré.

Ce mémoire se focalise sur ce dernier point, plus précisément l'objectif de ce mémoire est double : trouver un modèle prédisant au mieux le montant des sinistres et mesurer l'influence de l'utilisation d'un certain réseau d'experts sur cette prédiction. Pour ce faire, ce mémoire est divisé selon le plan suivant :

— dans un premier chapitre le contexte et les enjeux de l'étude seront explicités, puis la base de

[1] L'assurance flotte automobile est une assurance permettant à une entreprise d'assurer l'ensemble son parc automobile avec un seul contrat.

[2] Le ratio combiné est une mesure de rentabilité technique, égale à $\frac{Charge\ de\ sinistres + Frais\ et\ commissions}{Primes\ acquises}$. Le seuil de rentabilité technique est 100% : si le ratio combiné d'une entreprise est inférieur à cette valeur, l'entreprise réalise un profit technique, à l'inverse s'il est supérieur à 100%, elle réalise une perte.

données utilisée et les retraitements qui y sont appliqués seront présentés ;

— dans un deuxième chapitre les liens entre variables seront étudiés, puis une première méthode de modélisation, le modèle linéaire généralisé, sera présentée ;

— dans un troisième chapitre les méthodes d'arbres de régression seront présentées, puis les modèles CART et MOB et les résultats qui leurs sont associés seront étudiés. Le modèle XGBoost sera également introduit, d'une part pour y comparer les performances prédictives des modèles, d'autre part pour essayer d'améliorer le pouvoir prédictif du modèle MOB tout en conservant sa facilité d'interprétation.

Chapitre 1

Contexte de l'étude et données

Le but de ce chapitre est de préciser le contexte et l'objectif de l'étude présentée dans ce mémoire. Dans un premier temps, le cadre de l'assurance automobile et la notion d'expertise seront présentés, puis l'objectif et les enjeux liés à l'étude seront détaillés. Enfin, une étude des variables explicatives pour chacun des réseaux sera présentée.

1.1 L'assurance automobile

Le secteur de l'assurance peut être divisé en deux branches principales :

— L'assurance de personnes, qui vise à couvrir des personnes physiques. Ce secteur inclut par exemple l'assurance vie, l'assurance santé, la retraite, la dépendance et la prévoyance. Sauf exception, cette branche de l'assurance repose sur un principe forfaitaire.

— L'assurance de biens et de responsabilités, qui vise à couvrir des particuliers, entreprises ou toute autre entité, des risques qui ne relèvent pas de la vie humaine. Ce secteur inclut par exemple l'assurance automobile, l'assurance habitation et l'assurance des entreprises. Ce type d'assurance est également appelé assurance IARD pour Incendie, Accidents et Risques Divers. Cette branche de l'assurance repose principalement sur un principe indemnitaire.

L'assurance automobile fait ainsi partie de l'assurance de biens et de responsabilités. Elle peut se définir comme l'assurance visant à couvrir les dommages matériels ou corporels causés ou subis par un véhicule. Elle représentait en 2020 plus de 55% des cotisations des particuliers pour l'assurance de biens et responsabilités, mais également plus de 58% des sinistres dommages [FFA (2021a)]. La fréquence des sinistres automobiles décroît depuis une décennie, en opposition à leur coût moyen qui lui augmente. L'évolution du ratio combiné de l'assurance automobile depuis 2010 est présentée en figure (1.1) ci-dessous.

Il est important de rappeler que l'année 2020 a été impactée par la crise sanitaire de la COVID-19. La France a connu durant cette année deux périodes de confinement, s'écoulant du 17 mars au 11 mai, puis du 30 octobre au 15 décembre, ainsi que des restrictions concernant les horaires et les zones de déplacement. Ces mesures ont conduit à une réduction du trafic automobile. En effet, le système de navigation Tom-Tom a par exemple relevé une baisse du trafic de 89% à Paris, 84% à Lyon et 83% à

Bordeaux par rapport à une semaine normale en 2019 [AutoPlus (2020)]. Par ailleurs, l'application Waze a enregistré une baisse de 83% du nombre de trajets quotidiens parcourus par ses utilisateurs sur cette période [Waze (2020)]. Cette baisse du trafic routier s'est accompagnée d'une baisse de la sinistralité automobile. Par exemple, concernant les accidents mortels en France métropolitaine, 2 541 personnes seraient décédées sur la route en 2020, soit une baisse de 22% par rapport à 2019 d'après l'ONISR[1]. Il s'agit de la plus faible mortalité routière enregistrée depuis 1924, alors qu'il est estimé que le nombre de véhicules en circulation depuis a été multiplié par 50 [ONISR (2021)]. Plus généralement, le nombre de sinistres automobiles enregistrés par jour a baissé de 19% entre les années 2019 et 2020 [FFA (2021b)]. Ainsi, l'année 2020 constitue une année très particulière dont les chiffres sont à contextualiser et interpréter avec prudence.

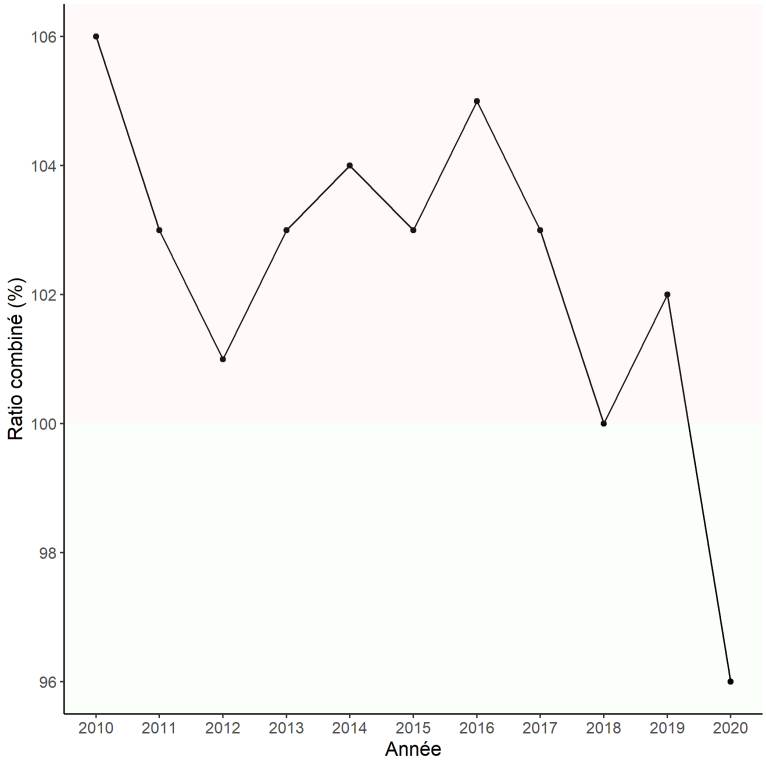


Figure 1.1 : Ratio combiné de l'assurance automobile

Concernant les années précédentes, un ratio combiné déficitaire (*i.e.* supérieur à 100) est observé pour le secteur de l'assurance automobile. Cela montre que la baisse de la fréquence des sinistres automobiles ne suffit pas à compenser la hausse de leurs coûts pour permettre à ce secteur de réaliser du profit dans son ensemble.

En assurance automobile, il existe une assurance obligatoire pour tout conducteur : l'assurance responsabilité civile (obligatoire depuis 1958) qui "permet l'indemnisation des dommages causés aux tiers par la faute du conducteur du véhicule ou d'un des passagers" [FFA AUTO (2021)]. D'autres assurances facultatives sont également proposées comme par exemple :

— la garantie dommage tous accidents, qui permet de couvrir les dommages matériels subis par le véhicule, pour tous types d'accident, quelle que soit la responsabilité du conducteur ;

[1]Observatoire National Interministériel de la Sécurité Routière.

— la garantie incendie et vol, qui permet l'indemnisation du propriétaire du véhicule si ce dernier est détruit ou endommagé par le feu ou bien volé ;

— la garantie bris de glace, qui permet l'indemnisation des dommages concernant le pare-brise du véhicule assuré.

1.1.1 Sinistres et expertise en assurance automobile

Sinistres

D'après le Code des assurances [Code des assurances (2003)], la définition d'un sinistre en assurance de biens et responsabilités est la suivante : "constitue un sinistre tout dommage ou ensemble de dommages causés à des tiers, engageant la responsabilité de l'assuré, résultant d'un fait dommageable ou d'un ensemble de faits dommageables ayant la même cause technique, imputable aux activités de l'assuré garanties par le contrat, et ayant donné lieu à une ou plusieurs réclamations." Plus concrètement cela concerne tout événement pouvant entraîner une prise en charge prévue par le contrat d'assurance souscrit. En assurance automobile, deux catégories de sinistres peuvent être distinguées :

— les sinistres corporels (dommages causés à un ou plusieurs individu(s)) ;

— les sinistres matériels (dommages causés à un ou plusieurs biens(s)).

En 2020, en moyenne 19 200 sinistres automobiles étaient comptabilisés par jour dont plus de 98% étaient des sinistres matériels [FFA (2021b)]. L'étude présentée dans ce mémoire ne concernera que des sinistres matériels.

Pour un sinistre automobile, il se peut que le montant des réparations soit supérieur à la valeur du véhicule, on parle alors de véhicule économiquement irréparable (noté VEI). Dans ce cas, l'assureur se doit de proposer la valeur de rachat du véhicule, cela s'appelle la procédure VEI. L'assuré peut accepter le montant d'indemnisation, appelé VRADE (valeur de remplacement à dire d'expert), mais il peut aussi refuser. En cas de refus, les travaux sont indemnisés dans la limite de la VRADE.

Expertise

Lors de la survenance d'un sinistre, une évaluation du montant des dommages doit être réalisée afin de pouvoir déterminer la valeur de l'indemnisation. L'assuré peut quantifier les dommages causés par le sinistre à l'aide de devis ou bien de factures, mais il peut également choisir de faire appel à un expert. Du côté de l'assureur, il peut choisir de ne pas avoir recours à un expert et de se baser sur les informations fournies par l'assuré, mais il peut aussi décider d'envoyer un expert de son choix pour évaluer le montant des dégâts causés par le sinistre.
Le rôle de l'expert est d'évaluer le montant des dommages, mais aussi de vérifier la cause du sinistre (notamment pour éviter la fraude) et de vérifier que les conditions d'application du contrat sont bien réunies. En assurance automobile, deux catégories d'expertise peuvent être distinguées ([lesfurets (2021)]) :

— L'expertise sur le terrain (expertise standard) :
L'expert se déplace chez le garagiste afin de déterminer avec lui quelles sont les réparations à effectuer, et quel montant représente une telle opération. Il envoie alors un rapport à l'entreprise d'assurance qui l'a mandaté et cette dernière décide ensuite de l'indemnisation proposée à son assuré. Une telle expertise nécessite en particulier un déplacement de l'expert et une disponibilité des garagistes pour convenir d'un rendez-vous. Afin de pouvoir gagner du temps un autre type d'expertise a été mis en place.

— L'expertise à distance :
Ce type d'expertise ne nécessite pas un déplacement de l'expert puisqu'elle consiste à évaluer à distance, sur la base de photographies et de devis, le montant des dommages causés. Cela permet un gain de temps pour les experts, et donc pour leurs clients, ainsi que pour les garagistes, qui n'ont plus à réserver des créneaux pour des rendez-vous avec les experts. Cependant, ce type d'expertise présente des limites car il ne permet pas une analyse profonde de tous les organes du véhicule. De plus, les photographies prises peuvent parfois cacher des détails qui ne sont observables que lors d'un contact direct avec le véhicule sinistré. Ainsi ce type d'expertise est généralement utilisé pour des sinistres peu graves, où les réparations ne concernent pas les organes de sécurité du véhicule. Dans tous les cas, l'expert doit informer l'assuré qu'il effectue une expertise à distance.

Pour ces deux types d'expertises, les experts sont amenés à travailler en collaboration avec différents garagistes, il existe notamment des partenariats entre experts et garagistes.

1.2 Contexte et objectif de l'étude

L'étude réalisée dans ce mémoire s'inscrit dans le contexte d'une mission du cabinet de conseil Prim'Act dont l'objectif était de comparer l'influence de deux réseaux d'experts sur les coûts de sinistres automobiles. En effet, dans un contexte de concurrence accrue, un assureur automobile se doit de mettre en place un certain nombre de stratégies pour tenter d'attirer et de conserver le plus de clients possible. Un des aspects que l'assureur peut piloter est le choix du réseau d'experts utilisé lors de l'évaluation d'un sinistre, il choisira alors le réseau d'experts le plus performant.

L'objectif est ainsi de trouver un modèle qui permet, d'une part, de prédire au mieux le coût d'un sinistre à l'aide de différentes variables explicatives et, d'autre part, d'étudier l'influence d'une variable en particulier, le réseau d'experts, sur le montant du sinistre prédit. Ainsi, lors de l'implémentation des différents modèles de prédiction présentés par la suite, leur pouvoir prédictif sera évalué, mais les capacités d'interprétation et d'étude de l'impact des variables associées au modèle seront également prises en compte. Ce mémoire s'inscrit dans la suite des mémoires de M. Lussac [Lussac (2018)], D. Khougea [Khougea (2019)] et de A. Wabo [Wabo (2020)].

Enjeux sur la mesure d'écart de performance

Au vu du rôle important de l'expert dans la détermination du montant final de l'indemnisation qui sera versée à l'assuré, il est naturel pour un assureur de se demander si le réseau d'experts auquel il fait appel est rentable pour son activité. Si un expert surestime le montant d'un sinistre cela représente un désavantage pour l'assureur, car il sera tenu d'indemniser un montant plus élevé que le coût du

sinistre réel. À l'inverse, si l'expert sous-estime le montant d'un sinistre, l'assureur sera avantagé mais le sinistré se trouvera en situation défavorable, puisqu'il lui sera indemnisé moins que le coût réel du sinistre. De ce fait, il est important pour un assureur, mais aussi pour ses assurés, qu'il puisse évaluer la performance de ses experts et déterminer lesquels sont les plus performants.

Dans le cadre de ce mémoire, deux réseaux d'experts sont comparés, ces réseaux ont été anonymisés et seront appelés réseau d'experts A et réseau d'experts B. Les règles appliquées aux sinistrés sont les mêmes pour ces deux réseau d'experts. Ainsi, il sera considéré que le réseau d'experts le plus performant est celui qui coûte le moins à l'assureur. Pour identifier ce réseau et répondre à la problématique de l'étude présentée, différentes méthodes de modélisation vont être comparées.

1.3 Notions de modélisation

Cette section présente quelques notions de modélisation, le principe de l'apprentissage supervisée et les méthodes d'évaluation de performance des modèles. Elle est inspirée du mémoire réalisé par M. de Lussac ([Lussac (2018)]), lui même inspiré du mémoire réalisé par R. Bellina ([Bellina (2014)]).

1.3.1 Pourquoi modéliser ?

Afin de comparer les performances entre différents réseaux d'experts, une méthode plutôt intuitive serait de comparer les coûts moyens des sinistres évalués par les différents réseaux et d'identifier le réseau le plus performant comme étant celui associé au coût moyen le plus faible. Il est assez aisé de comprendre que cette méthode intuitive n'est pas valide si les sinistres évalués par les différents réseaux ne sont pas de même nature. En effet, si un réseau d'experts est amené à évaluer plus de sinistres graves qu'un second réseau, le coût moyen du premier sera plus élevé que celui du second, sans pour autant que celui-ci soit nécessairement plus performant. Par exemple, si un réseau d'experts opère dans une zone géographique où le niveau de vie est plus élevé que celle où opère un second réseau, il est probable que les véhicules expertisés par le premier soit plus coûteux que ceux expertisés par le second, et donc que le coût moyen expertisé par le premier réseau soit plus élevé que celui du second, sans que cela implique nécessairement une différence de performance entre les deux réseaux.

Pour palier aux limites de cette méthode, des méthodes statistiques et de *machine learning* seront utilisées afin de prédire le coût des sinistres à l'aide de différentes variables explicatives. Au sein de ces variables explicatives se trouvera en particulier une variable représentant le réseau d'experts chargé de l'évaluation du sinistre, dont on cherchera à mesurer l'impact sur les coûts prédits par les modèles. Dans le cadre de ce mémoire, la base d'étude utilisée contient la variable qui spécifie le coût des sinistres que l'on cherche à prédire. Il s'agit donc ici d'un problème d'apprentissage supervisé dont le principe est présenté ci-après.

1.3.2 Principe de l'apprentissage supervisé

Les méthodes d'apprentissage automatique, plus connues sous leur nom anglais "méthodes de *machine learning*", peuvent être divisées en deux catégories : les méthodes d'apprentissage supervisé et les méthodes d'apprentissage non supervisé. Les méthodes d'apprentissage supervisé sont basées sur l'apprentissage des modèles sur des échantillons contenant la variable à prédire. Les données sont

dites "étiquetées", dans le sens où, pour chaque observation de la base d'étude, la variable d'intérêt est connue. Ainsi, dans le cadre de l'apprentissage supervisé, la base d'étude contient une variable de réponse, notée Y, que l'on cherche à prédire et un ensemble de p variables explicatives, notées $X = (X_1, ..., X_p)$. L'objectif est de prédire au mieux la variable Y lorsque X est connue. Le principe de l'apprentissage supervisé réside dans la recherche d'une fonction f telle que $Y = f(X)$. Cette fonction est appelée fonction de prédiction ou bien règle de prédiction.

En pratique, on dispose de n observations, ainsi $Y = (Y_1, ..., Y_n)^T \in \mathbb{R}^n$ et $X \in \mathbb{R}^{n \times p}$. Le problème réside dans le choix de la fonction f mais aussi dans le choix, ou bien la disponibilité, des variables explicatives X_j ($j \in \{1, ..., p\}$). En effet, il se peut que la relation qui permet d'expliquer Y soit de la forme $Y = h(X, Z)$ où h est une fonction et Z un ensemble de variables non observées dans l'étude. Il est également possible que la base de données considérée contienne des variables non pertinentes pour la modélisation. Il faut alors procéder à une sélection de variables afin de ne garder dans le modèle que les variables pertinentes pour pouvoir prédire Y. La recherche de la fonction f s'appuie sur un critère de minimisation d'une fonction de perte notée L, ainsi

$$f = \arg\min_{f \in \mathcal{F}} L(Y, f(X)).$$

où $\mathcal{F}$ désigne un ensemble de fonctions.

Une fonction de perte souvent utilisée est l'erreur quadratique moyenne, souvent notée MSE pour *Mean Square Error* en anglais. Elle est définie par

$$\forall (u, v) \in \mathbb{R}^2,\ MSE(u, v) = \mathbb{E}[(u - v)^2].$$

L'utilisation de cette fonction de perte permet en particulier d'utiliser le résultat suivant : si Y est de carré intégrable, alors il y a existence et unicité de la projection orthogonale dans L^2 et cette projection orthogonale est l'espérance conditionnelle de Y sachant X, notée $\mathbb{E}[Y|X]$.
Ainsi, en prenant $f(X) = \mathbb{E}[Y|X]$, f minimise l'erreur quadratique moyenne. Ce résultat sert de fondement pour certains types de modélisation, en particulier pour les modèles linéaires généralisés.

1.3.3 Pouvoir prédictif des modèles

On cherche à trouver une fonction f de façon à minimiser $\mathbb{E}[(Y - f(X))^2]$. On dispose en pratique de n observations, on cherche alors à résoudre

$$\arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

La fonction $f(x) = \sum_{i=1}^{n} y_i \mathbb{1}_{\{x = x_i\}}$ est solution du problème, mais elle ne présente aucun pouvoir prédictif puisqu'elle est adaptée seulement aux données sur lesquelles elle est construite. Pour tout nouvel individu k, muni de ses variables explicatives x_k, non présent dans la base sur laquelle f est construite, on aura $f(x_k) = 0$ systématiquement. Cet exemple est une illustration d'un phénomène appelé surapprentissage (*overfitting* en anglais). Ce phénomène se produit lorsque un modèle apprend "trop" à partir des données qui lui sont fournies, en apprenant par exemple le bruit produit par ces dernières. Le modèle se généralise alors mal lorsque de nouvelles données extérieures à la base d'apprentissage (sur laquelle il a été construit) lui sont présentées.

À noter qu'il existe également le phénomène de sous-apprentissage (*underfitting* en anglais). Cela correspond au cas où le modèle ne peut pas capturer de manière adéquate la structure sous-jacente des données. Le modèle possède alors également un faible pouvoir prédictif. Ainsi, on cherche à construire un modèle qui possède une qualité de prédiction suffisante et qui se généralise aisément à de nouvelles observations. Les modèles étudiés dans ce mémoire possèdent différents paramètres dont les valeurs seront choisies de façon à maximiser le pouvoir prédictif du modèle tout en évitant le phénomène de surapprentissage. Pour essayer de palier à ce phénomène, la base de l'étude sera découpée en trois échantillons distincts :

— l'échantillon d'apprentissage (60% de la base de données), utilisé pour la construction des modèles ;

— l'échantillon de validation (20% de la base de données) qui permet l'optimisation des paramètres et permet également une première comparaison entre les différents modèles ;

— l'échantillon de test (20% de la base de données), utilisé pour l'évaluation des prédictions associées aux modèles optimisés. Cet échantillon permet la comparaison des performances prédictives des différents modèles.

Au cours de la phase d'apprentissage, l'erreur commise par le modèle décroît lorsque la complexité du modèle augmente. En effet, il est possible d'adapter au maximum les paramètres du modèle à l'échantillon d'apprentissage afin que l'erreur de prédiction soit la plus faible possible. Mais, comme expliqué précédemment, une telle stratégie entraînerait la création d'un modèle trop complexe qui ne se généraliserait pas bien à de nouvelles données. Ainsi, l'augmentation de la complexité d'un modèle permet de faire baisser l'erreur sur l'échantillon de validation jusqu'à un certain seuil, puis au delà de ce seuil l'erreur de validation augmente avec la complexité à cause du surapprentissage. Ce phénomène est représenté par la figure (1.2) ci-dessous :

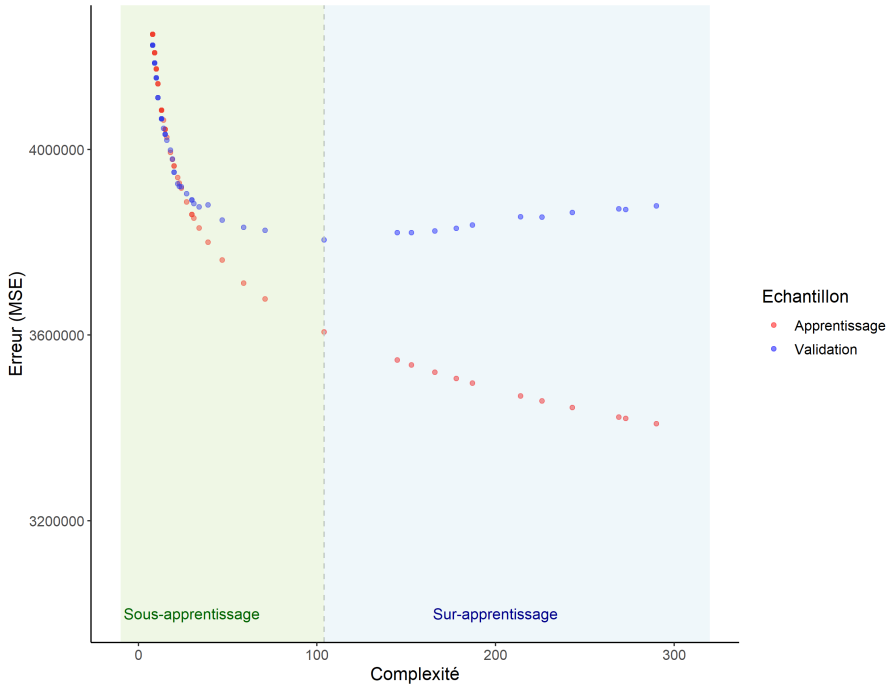


FIGURE 1.2 : Erreur de validation et d'apprentissage en fonction de la complexité du modèle

La figure (1.2) est une représentation de l'erreur des moindres carrés en fonction de la complexité du modèle. Le modèle appliqué ici est le modèle CART, présenté en section (3.1), et la complexité correspond à la taille de l'arbre construit, *i.e.* le nombre de feuilles qu'il contient. Le modèle a été entraîné sur l'échantillon d'apprentissage puis a été utilisé pour prédire les coûts de l'échantillon d'apprentissage (points rouges) et de l'échantillon de validation (points bleus). Deux zones peuvent alors être distinguées :
— une première zone, représentée en vert, où l'erreur décroît pour les deux échantillons, la décroissance forte pour des complexités faibles illustrant bien le phénomène du sous-apprentissage ;
— une seconde zone, représentée en bleue, où l'erreur sur l'échantillon d'apprentissage continue de décroître alors que celle sur l'échantillon de validation augmente, ce qui correspond au surapprentissage.

La mesure du pouvoir prédictif d'un modèle peut se faire à l'aide de différents indicateurs de performance, qui quantifient l'erreur commise par le modèle lors de la prédiction. Un indicateur souvent considéré est l'erreur quadratique moyenne, qui est calculée sur un échantillon $\mathcal{E}$ de taille n_e par

$$MSE = \frac{1}{n_e} \sum_{y_i \in \mathcal{E}} (y_i - \widehat{y}_i)^2.$$

où : y_i est la valeur observée dans l'échantillon $\mathcal{E}$ pour l'individu i et $\widehat{y}_i$ est la valeur prédite par le modèle pour l'individu i.

Plus la MSE d'un modèle est proche de 0, plus son pouvoir prédictif est élevé. D'autres mesures peuvent également être considérées comme par exemple la RMSE, qui n'est autre que la racine carré de la MSE, ou bien la MAE (pour *Mean Absolute Error* en anglais) définie par

$$MAE = \frac{1}{n_e} \sum_{y_i \in \mathcal{E}} |y_i - \widehat{y}_i|.$$

Du fait au passage au carré des erreurs, la MSE et la RMSE pénalisent plus fortement les grands écarts de prédiction que la MAE. En pratique, la fonction de prédiction f ne pourra quasi-jamais prédire parfaitement la réalité, on cherche alors à trouver la fonction qui s'en rapproche le mieux possible en utilisant des métriques telles que celles présentées ci-dessus.

Dans le cadre de ce mémoire, des modèles dits "modèles d'arbres", ou bien modèles à structure arborescente, vont être présentés. Le principe de ce type de modèles est de partitionner les individus afin de réduire l'hétérogénéité globale de la population. L'idée est de diviser la population afin de créer des groupes au sein desquels l'hétérogénéité est minimisée et pour lesquels un tarif propre au groupe va pouvoir être déterminé : c'est le principe de la segmentation présenté ci-après.

1.4 Mutualisation et segmentation en assurance

Qu'est ce que la mutualisation ?

L'assurance repose sur le principe de mutualisation des risques ([Charpentier (2015)]). L'idée est que chaque assuré paye une prime à l'assureur afin de s'assurer contre un ensemble de risque. Cependant,

la plupart du temps, le règlement de la prestation assurantielle n'intervient pas auprès de tous les assurés. Certains assurés payent donc une prime et ne recevront jamais de prestations, car ils ne connaîtront pas de sinistres. À l'inverse, certains assurés connaîtront un ou plusieurs sinistres et seront dédommagés par l'assureur grâce à l'ensemble des primes payées par tous les assurés, les risques sont donc partagés entre les assurés : c'est la mutualisation des risques. Ce principe repose sur des théorèmes dits "asymptotiques" ou bien "limites" comme la loi des grands nombres et le théorème central limite ([Charpentier (2011)]), énoncés ci-dessous :

Théorème 1 *Loi des grands nombres*
Soient $X_1, ..., X_n$ n variables indépendantes de même espérance finie $\mathbb{E}[X_i] = \mu$ $(i = 1, ..., n)$, alors

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \xrightarrow[n\to+\infty]{} 0.$$

Dans le cadre de l'assurance, les X_i représentent les charges de sinistres et le théorème s'interprète de la façon suivante : si les risques sont de même espérance et indépendants alors la charge moyenne tend vers la prime pure pour un nombre suffisamment élevés de risques.

Théorème 2 *Théorème Central Limite*
Soient $X_1, ..., X_n$ n variables indépendantes de même espérance finie $\mathbb{E}[X_i] = \mu$ et de même variance finie $\mathbb{V}(X_i) = \sigma^2$ $(i = 1, ..., n)$, alors

$$\sqrt{n}\frac{(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu)}{\sigma} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

En assurance, ce théorème s'interprète de la façon suivante : si les risques sont indépendants, homogènes et de variance finie, et que leur nombre n est suffisamment élevé alors la loi de la charge totale des sinistres est approximativement gaussienne. C'est sur ce théorème qu'est fondée l'idée que la mutualisation réduit le risque.

Ces théorèmes sont à la base de la tarification en assurance et du principe de mutualisation. Cependant ils reposent sur des hypothèses fortes. Pour que la mutualisation soit efficace, il faut disposer d'un nombre suffisamment élevé de risques indépendants et homogènes. Il se peut qu'au sein d'un groupe d'assurés couverts pour un même risque, l'homogénéité ne soit pas avérée et que la tarification sur le principe seul de la mutualisation ne soit pas efficace. Afin de créer des groupes au sein desquels l'homogénéité sera plus importante, l'assureur peut choisir de segmenter sa population d'assurés.

Qu'est ce que la segmentation ?

La segmentation des risques en assurance vise à diviser l'ensemble des assurés en plusieurs sous-groupes, au sein desquels le principe de mutualisation est appliqué afin de proposer pour chacun des groupes une tarification adaptée aux profils des individus qui le composent. Pour comprendre pourquoi l'application du principe de mutualisation ne s'applique par correctement à un portefeuille hétérogène, considérons un assureur et sa population d'assurés non-homogènes. Supposons qu'au sein de cette population, certains assurés ont une probabilité élevée de connaître un sinistre, ces assurés sont appelés "mauvais risques" en opposition aux autres assurés, appelés "bons risques", qui ont une probabilité

relativement faible de connaître un sinistre. Deux cas peuvent alors être distingués :

— *Cas n°1 : l'assureur n'effectue pas de segmentation*
Dans ce cas, l'assureur applique le principe de mutualisation à l'ensemble de son portefeuille, la prime payée est donc la même pour tous les assurés et est égale à l'espérance de la charge annuelle, notée μ. Cette prime μ est plus faible que celle qui aurait été calculée sur l'ensemble des mauvais risques, ces derniers réalisent donc un profit. À l'inverse, la prime μ est plus élevée que celle qui aurait été calculée sur l'ensemble des bons risques, ces derniers réalisent donc une perte et seront donc enclins à se rediriger vers un assureur qui segmente ses risques afin de payer une prime plus faible, adaptée à leurs profils. Finalement, l'assureur attirera les mauvais risques et ne sera pas attractif pour les bons risques. Cette situation illustre le phénomène d'antisélection, qui n'est pas souhaitable pour un assureur.

— *Cas n°2 : l'assureur effectue une segmentation*
Supposons qu'il est possible pour l'assureur de segmenter sa population d'assurés selon un ou plusieurs critères qui permettrait de réduire l'hétérogénéité de son portefeuille. Dans ce cas l'assureur est en capacité de proposer un tarif plus faible que μ à ses bons risques. À l'inverse, la prime proposée pour les mauvais risques est plus élevée que μ. Cette situation réduit le risque d'antisélection : les mauvais risques sont plus enclins à se rediriger vers un assureur qui n'effectue pas de segmentation, ou, s'ils restent, ils payeront une prime à la hauteur du risque qu'ils représentent.

L'exemple précédent permet d'illustrer les limites d'une tarification basée uniquement sur le principe de mutualisation en présence d'un portefeuille non-homogène. Les bons risques sont généralement attirés par les assureurs utilisant la segmentation pour leur tarification, à l'inverse des mauvais risques. La segmentation permet de limiter le phénomène d'antisélection, mais cette méthode présente des limites.

Limites de la segmentation et compromis avec la mutualisation

Au vu de l'exemple précédent, il pourrait être tentant pour un assureur de segmenter au maximum son portefeuille afin d'appliquer pour chacun de ses assurés le tarif le plus juste. Cependant, un nombre suffisamment élevé d'observations est nécessaire pour que le théorème central limite et la loi des grands nombres puissent être correctement utilisés, *i.e.* que le principe de mutualisation des risques puisse être correctement appliqué. De plus, effectuer une segmentation très fine conduit à créer des modèles complexes. Par exemple, dans le cas des modèles à structure arborescente une forte segmentation implique la création d'un arbre très grand, ce qui limite sa lisibilité et son interprétabilité. Il est également important de noter que la segmentation ne peut être effectuée que sur l'ensemble des variables dont dispose l'assureur. Il se peut qu'il existe une hétérogénéité latente au sein des assurés selon une variable que l'assureur ne peut observer.
Ainsi, il est souhaitable de trouver un équilibre entre mutualisation et segmentation afin de proposer des tarifs qui soient concurrentiels et qui permettent une limitation du risque pris par l'assureur. Cette idée est à garder en esprit lors de l'utilisation de modèles permettant une segmentation des individus.

Avant l'application de modèles statistiques ou de machine learning, une étape importante consiste à analyser et retraiter la base de données sur laquelle l'étude est réalisée. Cette étape permet en particulier de se familiariser avec les données manipulées au long de l'étude et ainsi de disposer d'une

meilleure capacité d'interprétation des résultats futurs.

1.5 Présentation et traitement de la base de données

1.5.1 Présentation des données

La base utilisée dans cette étude résulte de la concaténation de données provenant de sept assureurs différents. Ces assureurs ont été anonymisés et sont appelés dans le jeu de données `Assureur_1`, ..., `Assureur_7`. Parmi ces assureurs trois utilisent à la fois les réseaux d'experts A et B, les quatre autres utilisant seulement le réseau A. Au vu de l'objectif de l'étude, seuls les trois assureurs qui font appel aux deux réseaux d'experts seront considérés. L'étude est effectuée sur deux années glissantes, ainsi les données ont été collectées sur la période allant du premier trimestre de 2019 au dernier trimestre de 2020 inclus. Après suppression des doublons, la base initiale est constituée de 1 523 437 observations et de 19 variables dont le nom et la description sont explicités dans le tableau (1.1) ci-dessous :

Variable	Description
`ageVehicule`	Durée (en années) écoulée depuis la mise en circulation du véhicule jusqu'à la date du sinisitre
`assureur`	Nom de l'assureur en charge du sinistre
`cdpostexp`	Code postal du lieu de survenance du sinistre
`coutREP`	Montant de l'expertise en cas de réparation
`coutVEI`	Montant de l'expertise en cas de véhicule économiquement irréparable
`EAD`	Spécifie si l'expertise à été effectuée à distance ou non
`expertA`	Variable booléenne spécifiant si le réseau d'experts A a été chargé d'évaluer le sinistre ou non
`garage_agree`	Variable booléenne spécifiant si le garage est agréé ou non
`issue`	Spécifie s'il y a eu réparation (REP) ou non (VEI)
`kilometrage`	Kilométrage du véhicule sinistré
`mois`	Mois de survenance du sinistre
`marque`	Marque du véhicule sinistré
`nat_sin`	Nature du sinistre
`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
`trimestre`	Trimestre de survenance du sinistre
`typeSinistre`	Type d'accident survenu

TABLE 1.1 : Description des variables de la base initiale

Il est possible de noter l'absence de certaines variables explicatives qui se trouvent en général dans le portefeuille d'un assureur automobile. Par exemple, la base ne contient aucune information sur le conducteur (âge, ancienneté du permis, situation familiale, situation professionnelle...) ou sur les caractéristiques du véhicule (type de véhicule, type de carburant utilisé, gamme de prix...) ou encore sur la situation de l'accident (emplacement du choc, gravité du sinistre, météo au moment du sinistre...). Cette absence d'information peut s'expliquer par le fait que la base résulte de la fusion de plusieurs bases de données qui proviennent de différent assureurs. Il a donc fallu supprimer les variables spécifiques à chaque assureur par soucis d'harmonisation de la base finale. D'autre part, comme les données sont extraites trimestriellement, il faut prendre en compte le fait que, pour certains assureurs, extraire chaque trimestre les données demandées peut représenter une difficulté opérationnelle, ce qui a aussi conduit à la réduction du nombre de variables dans la base. Enfin, cette étude ayant pour

but la mesure de l'écart de performance entre deux réseaux d'experts, la sélection des variables a été effectuée selon cet objectif et certaines variables qui peuvent être retrouvées habituellement dans une base pour la tarification automobile ont été écartées car jugées non pertinentes pour cette étude.

Pour pouvoir être exploitée par les différents modèles présentés dans cette étude, la base nécessite des retraitements permettant par exemple la gestion des valeurs manquantes et des valeurs jugées aberrantes.

1.5.2 Retraitement de la base

Le retraitement appliqué dans cette section est en grande partie similaire à celui décrit dans le mémoire de M. de Lussac [Lussac (2018)] puisqu'il s'agit de la même base de données considérée sur une période différente. Notons cependant que le fait de considérer les données sur une autre période peut conduire à une base finale différente. En effet, il se peut qu'un assureur ait cessé de fournir une variable à travers le temps, ce qui implique que cette variable n'est plus commune à tous les assureurs et doit alors être retirée de l'étude. L'ensemble des travaux réalisés dans ce mémoire ont été réalisé sur R [R Core Team (2021)]. Les traitements qui ont été appliqués à la base de l'étude sont présentés ci-après.

Création et suppression de variables

— Variable à expliquer : le coût du sinistre
Cette variable est égale à `coutREP` si le véhicule à subi des réparations (*i.e.* si la variable `issue` est égale à `REP`) ou à `coutVEI` si le véhicule n'est pas économiquement réparable (*i.e.* si la variable `issue` est égale à `VEI`). Une fois cette variable créée, les variables `coutREP` et `coutVEI` sont supprimées de la base.

— Le taux horaire
La base contient quatre variables qui concernent le taux horaire, une seule variable `tauxHoraire` est créée en rassemblant au mieux ces informations. Cette variable prend initialement comme valeur la moyenne des taux horaires renseignés. Les taux horaires ayant une valeur inférieure à 15€ sont ensuite remplacés par le taux horaire moyen, car ces observations sont considérées comme anormalement faibles. À noter que plus de 99% des taux horaires inférieurs à 15€ correspondent à des taux horaires nuls. Les observations associées à des taux horaires supérieurs à 300€ sont supprimées de la base (soit 70 observations supprimées). Une fois ce retraitement effectué, la variable `tauxHoraire` est transformée en variable qualitative. Elle prend la valeur `NUL` lorsque l'issue est égale à `VEI`, puisque dans ce cas le véhicule n'a subi aucune réparation, et est à valeur dans {`Faible, Moyen, Fort`} sinon. Pour définir ces trois dernières catégories une discrétisation s'appuyant les trois premiers quartiles des taux horaires lorsque le véhicule est réparé est effectuée. Les variables `tauxHoraire1, tauxHoraire2, tauxHoraire3, tauxHoraireP` et `issue` sont ensuite supprimées de la base.

— Suppression des variables possédant trop de valeurs manquantes
Les variables `nat_sin, EAD, cdpostexp` et `garage_agree` sont supprimées car elles possèdent trop de valeurs manquantes pour être intégrées à l'étude. La présence d'un grand nombre de valeurs manquantes pour une variable peut être dû au fait qu'un assureur parmi les trois considérés dans cette étude n'ait pas fourni la variable. C'est notamment le cas pour la variable `garage_agree` qui n'a pas été fournie par l'un des assureurs ou bien pour la variable `EAD` qui n'a pas été fournie par l'un des assureurs pour un trimestre considéré dans l'étude. De tels cas

de figure soulignent les difficultés que peut impliquer la fusion de base de données provenant de différentes sources.

Suppression d'observations

On supprime les observations pour lesquelles :

— l'assureur n'utilise qu'un des deux réseaux d'experts. Cela concerne quatre des sept assureurs présents dans la base. Ce retraitement conduit à la suppression de 683 981 observations ;

— l'issue est mal renseignée, *i.e.* pour lesquelles l'issue n'est ni la réparation du véhicule (`issue = REP`) ou ni le cas où le véhicule est économiquement irréparable (`issue = VEI`). Ce retraitement conduit à la suppression de 44 942 observations ;

— la variable d'intérêt `coutSinistre` est mal renseignée ou est inférieure à 1. Ce retraitement conduit à la suppression de 673 observations ;

— l'âge du véhicule est négatif, nul ou bien supérieur à 80. Ce retraitement conduit à la suppression de 464 observations ;

— le kilométrage est nul ou bien supérieur à 500 000 km. Ce retraitement conduit à la suppression de 7 805 observations ;

— la variable `marque` possède des valeurs manquantes. Ce retraitement conduit à la suppression de 1 265 observations.

Une fois la gestion des valeurs manquantes et aberrantes effectuée, une étape habituelle lors de la gestion de coûts de sinistre est la définition de seuils concernant ces derniers.

1.5.3 Définition de seuils

Seuil à droite : sinistres atypiques

Dans le cadre d'une étude concernant les coûts des sinistres automobiles, il est courant de distinguer les sinistres attritionnels et les sinistres atypiques, qui sont généralement étudiés à part. Les sinistres dits attritionnels sont les sinistres dont les montants sont relativement faibles et qui surviennent de façon statistique sur le portefeuille considéré. À l'inverse les sinistres dits atypiques ou graves sont les sinistres qui surviennent rarement mais dont les montants sont particulièrement élevés. Afin de garantir l'efficacité des modèles, il est souhaitable de retirer ces sinistres de l'étude ou bien de les étudier à part.
Au sein de la base considérée dans ce mémoire les coûts des sinistres présentent les caractéristiques suivantes :

Minimum	1^{er} quartile	Médiane	Moyenne	$3^{ème}$ quartile	Maximum
1	836	1 457	2 302	2 621	900 000

TABLE 1.2 : Caractéristiques de la variable `coutSinistre`

L'écart important entre le coût maximal et le troisième quartile et la moyenne suggère l'existence de sinistres atypiques dans la base. Une représentation graphique des coûts de sinistre dans la base est donnée par la figure (1.3) ci-dessous :

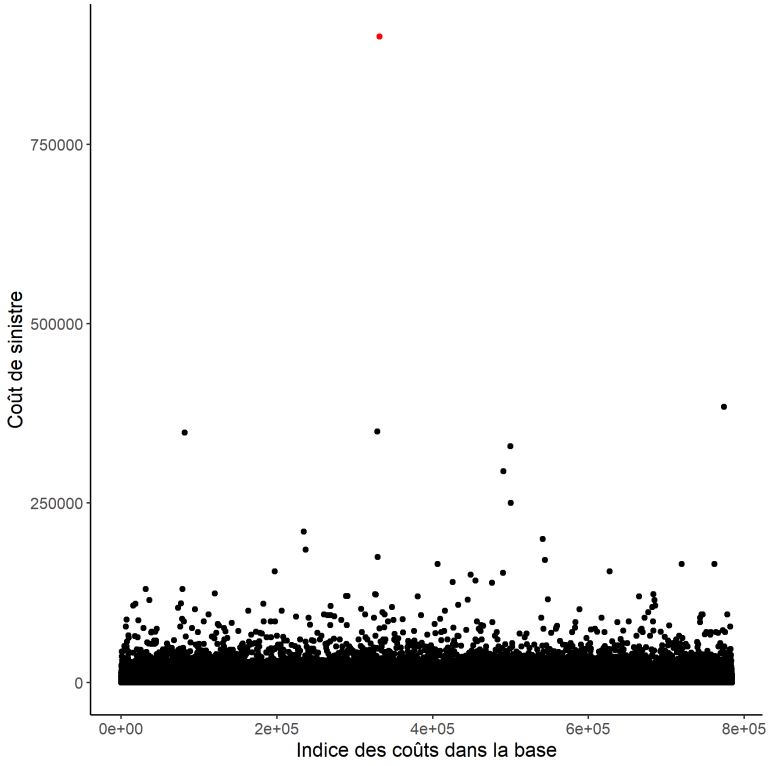


FIGURE 1.3 : Coût des sinistres dans la base

La figure (1.3) permet de mettre en évidence un sinistre au coût très élevé, représenté par la couleur rouge et correspondant à un coût de 900 000€. Sauf mention contraire ce sinistre sera exclu lors des représentations graphiques suivantes afin de permettre une meilleure visualisation. Cette première représentation graphique permet de mettre en évidence des sinistres aux coûts particulièrement élevés en comparaison avec le reste des sinistres de la base.

Afin d'analyser davantage les sinistres, une première étape consiste à analyser le graphique quantile-quantile, généralement appelé *QQ-plot*, des coûts avec une loi usuelle. Soit $Y_{(1)} \leq ... \leq Y_{(n)}$ la statistique d'ordre associée à l'échantillon de l'étude (dans le cadre de ce mémoire il s'agit des coûts ordonnés), le graphique représentant

$$\left\{ Y_{(i)}, F^{-1}\left(1 - \frac{i}{n}\right) : i = 1, ..., n \right\},$$

est le graphique quantile-quantile associé à la loi de fonction de répartition F, qui sera appelée dans la suite loi de référence.

Une loi de probabilité de fonction de répartition F est dite à queue légère, ou queue fine, s'il existe des constantes positives α et C et $x^* \in \mathbb{R}$ tels que

$$\forall x > x^*,\ 1 - F(x) \leq Ce^{-\alpha x},$$

Parmi les distributions usuelles, les lois normale, exponentielle et gamma sont à queue fine. Une loi qui n'est pas à queue fine est dite à queue épaisse, ou queue lourde. Les lois de Pareto, Weibull et log-normale sont des exemples de lois à queue épaisse.

L'analyse du QQ-plot d'un échantillon selon une loi de référence $\mathcal{L}$ permet d'avoir une caractérisation de sa queue de distribution : si la courbe est concave alors l'échantillon possède une queue de distribution plus fine que celle de la loi $\mathcal{L}$. À l'inverse si la courbe est convexe alors la queue de distribution de l'échantillon est plus lourde que celle de la loi $\mathcal{L}$. Enfin, si la courbe est linéaire l'échantillon suit une transformation linéaire de la loi de référence. Afin de comprendre ces conclusions d'analyse graphique, considérons l'exemple graphique suivant :

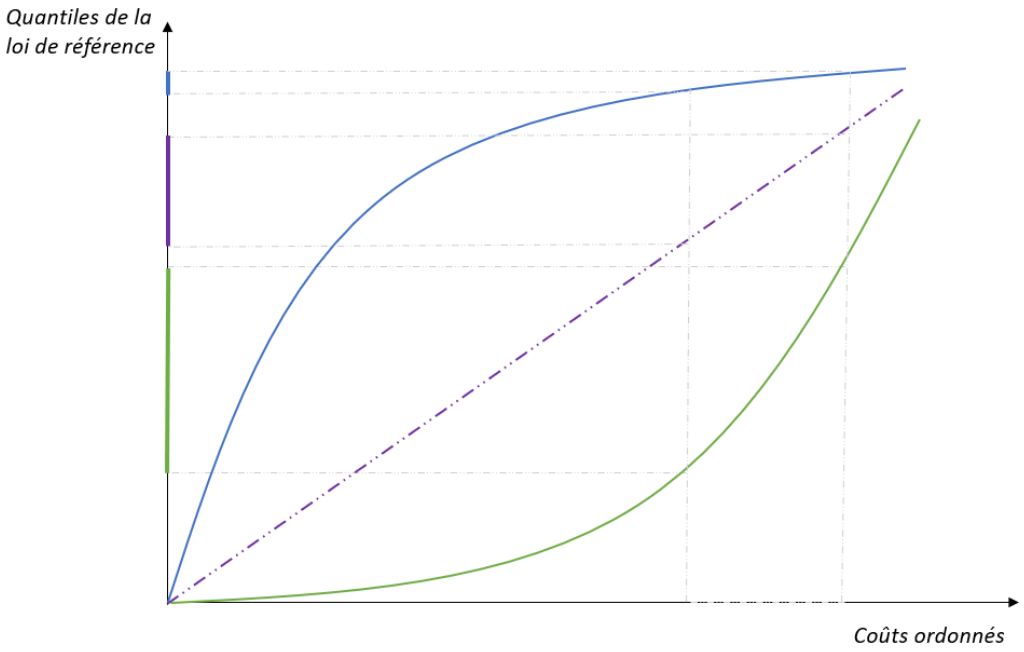


FIGURE 1.4 : Exemple d'allures de QQ-plot

La courbe verte correspond au cas convexe, la courbe bleue au cas concave et la courbe violette au cas linéaire, *i.e.* au cas où l'échantillon suit la loi de référence ou bien une transformation linéaire de cette dernière. Sur ce graphique il peut être observé que pour des coûts assez élevés, un même écart de coût (représenté par des pointillés sur l'axe des abscisses) conduit à des écarts de quantiles plus important lorsque la courbe est convexe que lorsqu'elle est linéaire. Cela montre que la queue de distribution de l'échantillon est plus épaisse que celle de la loi de référence. Le résultat inverse est observé dans le cas de concavité de la courbe : un même écart de coût conduit à un écart de quantiles plus faible que dans le cas linéaire, traduisant une queue de distribution plus fine que celle de la loi de référence.

Un QQ-plot habituellement effectué en analyse des valeurs extrêmes est le QQ-plot associé à la loi exponentielle. Le QQ-plot des coûts de la base avec cette loi est donné par la figure (1.5) ci-dessous :

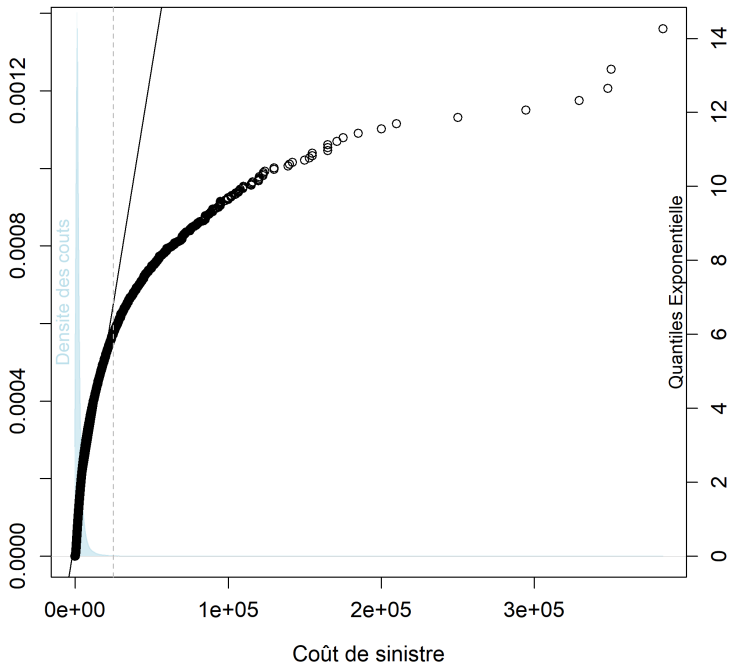


FIGURE 1.5 : QQ-plot des coûts avec la loi exponentielle

Deux zones peuvent être distinguées : une première partie, où la courbe a un comportement linéaire, puis, pour des coûts plus élevés, une partie où la courbe est concave. La séparation entre ces deux zones se fait aux alentours d'une valeur de 20 000€, qui a été représentée par une droite verticale pointillée sur la figure (1.5). La concavité de la courbe permet de conclure que la queue de distribution des observations est plus fine que celle de la loi exponentielle. En particulier, la loi exponentielle étant à queue fine, cela permet de conclure que la distribution des observations est à queue fine. La densité des coûts a également été représentée afin de souligner le nombre faible de coûts élevés dans la base. Cependant, bien que leur nombre soit faible, les sinistres correspondant à des coûts "trop" élevés sont à exclure de la base, car considérés comme atypiques. Un seuil au delà duquel les sinistres seront exclus de l'étude, dit "seuil de graves" doit être déterminé. Pour cela, plusieurs outils et méthodes peuvent être utilisés. Dans ce mémoire le choix du seuil s'est porté sur deux outils graphiques : le *Mean Excess Plot* et le *Hill plot*. Ces outils sont brièvement présentés avant d'exposer les résultats qui y sont associés.

— Le *Mean Excess Plot*
L'excès moyen par rapport au seuil t pour une variable de coût C est défini par

$$ME(t) = \mathbb{E}[C - t | C > t],$$

Pour un échantillon de coûts $C_1, ..., C_n$, la version empirique est

$$\widehat{ME}(t) = \frac{1}{\sum_{i=1}^{n} \mathbf{1}_{C_i > t}} \sum_{i=1}^{n} (C_i - t)\mathbf{1}_{C_i > t},$$

Le *Mean Excess Plot*, souvent noté ME plot, est la représentation graphique de $\{t, ME(t)\}$ avec t différentes valeurs de seuils. Le seuil choisi est tel que le *Mean Excess Plot* soit affine au delà de ce seuil.

— Le *Hill Plot*
L'estimateur empirique de Hill pour un seuil t est donné par

$$\widehat{\alpha}(t) = \frac{1}{\sum_{i=1}^{n} \mathbf{1}_{C_i>t}} \sum_{i=1}^{n} (\log(C_i) - \log(t)) \mathbf{1}_{C_i>t},$$

Le seuil optimal est choisi comme le plus grand seuil en dessous duquel la valeur de l'estimateur de Hill est stabilisée, ce qui graphiquement se traduit par une courbe horizontale à partir d'un certain rang.

Les résultats graphiques associés à ces outils sont présentés par les figures (1.6) et (1.7) *infra* :

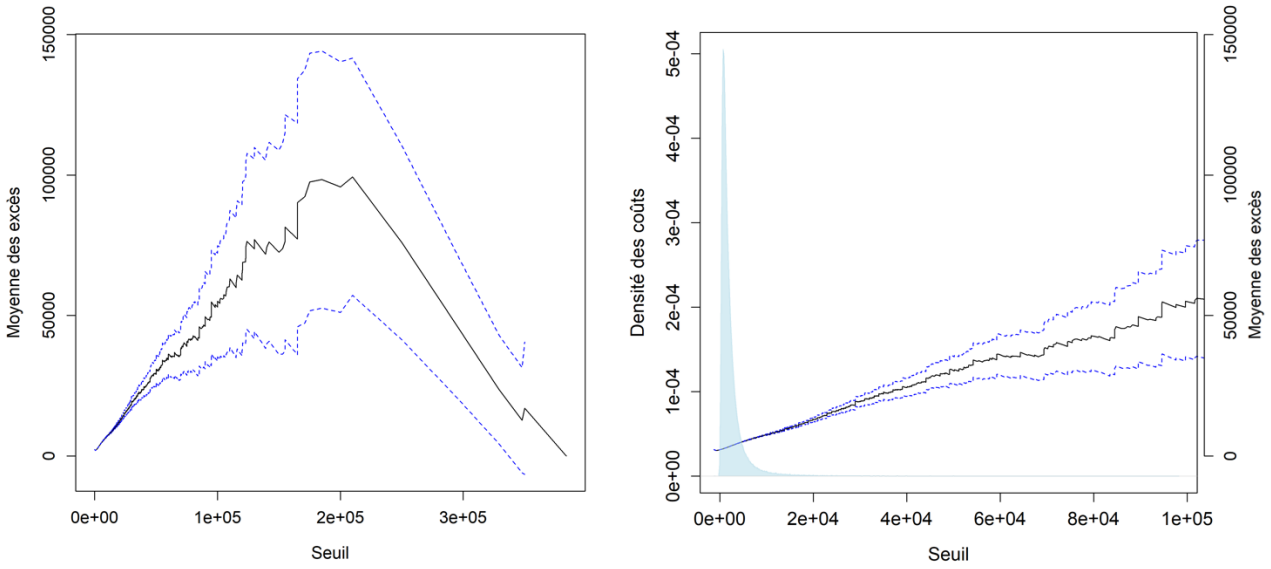


FIGURE 1.6 : *Mean Excess Plot* des coûts

La figure (1.6) présente le *Mean Excess Plot* des coûts de la base. Un zoom a été effectué sur la partie du graphique concernant les coûts inférieurs à 100 000€ et la densité des coûts a également été représentée en arrière plan. Les traits pointillés bleus représentent l'intervalle à 95% de l'estimation empirique de l'excès moyen. Plusieurs zones peuvent être distinguées : premièrement le ME plot croît linéairement jusqu'à un seuil aux alentours de 50 000€, puis le ME plot commence à présenter quelques irrégularités tout en continuant d'être croissant, enfin, à partir d'un seuil environ égal à 200 000€, le ME plot devient décroissant. Il est également observé que les intervalles de confiance deviennent relativement grands à partir d'un certain seuil situé aux alentours de 40 000€. Cette imprécision sur l'estimation s'explique par le nombre faible de sinistres dont le coût est supérieur à ce montant, comme il peut être vu sur la représentation de la densité des coûts. Ainsi, la représentation graphique du ME plot ne permet pas de tirer de conclusion rigoureuse pour le choix d'un seuil. Par conséquent, la recherche du seuil optimal est complétée par l'analyse du *Hill plot* donnée par la figure (1.7) ci-dessous :

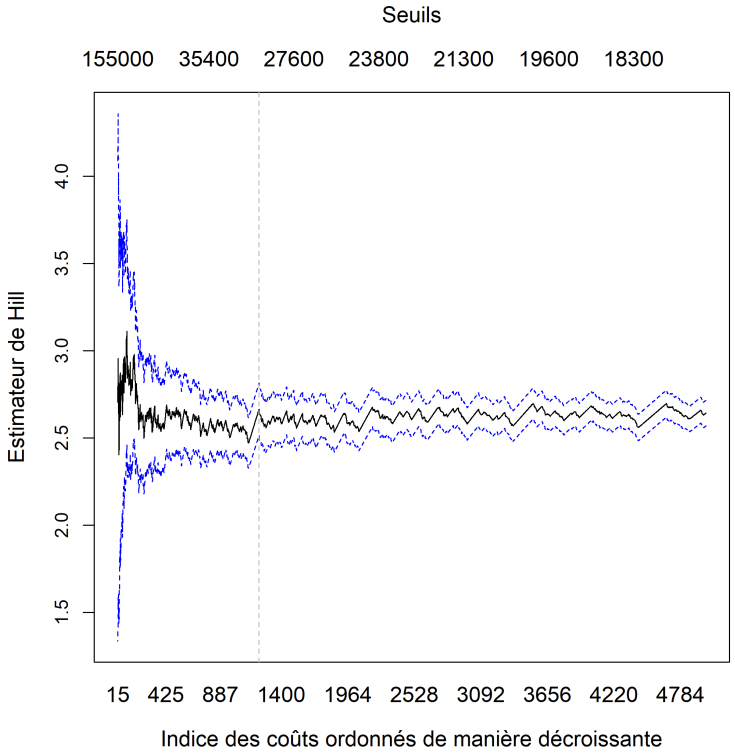


FIGURE 1.7 : *Hill plot* des coûts

En suivant le raisonnement expliqué précédemment, le seuil optimal est la plus grande valeur de seuil pour lequel l'estimateur empirique de Hill est stabilisé. Un seuil de 30 000€ est choisi, cela correspond à la ligne pointillée grise représentée sur la figure (1.7). Le choix de ce seuil conduit à la non-considération de 1 211 observations, ce qui correspond à écarter 0.15% des observations. Ces sinistres correspondent à plus de 3% du coût total des sinistres de la base. Une représentation graphique du choix de ce seuil est effectuée :

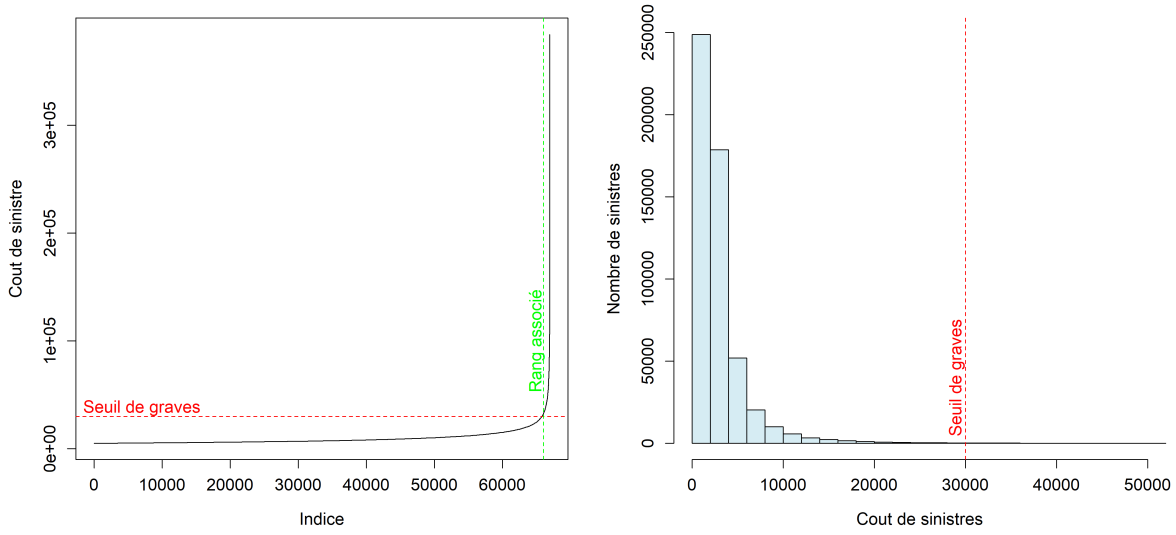


FIGURE 1.8 : Visualisation du choix d'un seuil égal à 30 000 €

Le premier graphique de la figure (1.8) est une représentation des coûts des sinistres présents dans la base triés par ordre croissant. Pour plus de clarté dans la visualisation, seuls les coûts des sinistres supérieurs à 5000€ ont été tracés, l'abscisse de ce graphique correspond au rang des sinistres au sein des sinistres supérieurs à 5000€. Ce graphique permet à nouveau d'attester de l'existence de sinistres atypiques, correspondant à des fréquences faibles et des coûts élevés, ce qui graphiquement se traduit par la présence d'une courbe quasi-verticale à partir d'un certain rang. La droite correspondant au seuil égal à 30 000€ coupe bien la courbe au niveau où celle-ci devient quasi-verticale et permet donc de distinguer les sinistres atypiques des sinistres attritionnels. Le second graphique représente un histogramme du nombre de sinistres en fonction du coût, la représentation du seuil de graves montre bien qu'un nombre faible de sinistres sont exclus de l'étude.

Seuil à gauche : franchise

Il est important de définir un seuil à gauche, *i.e.* un seuil en dessous duquel les sinistres ne seront pas pris en compte dans l'étude. En effet, la franchise est comprise dans le montant de l'expertise, ce qui signifie que les montants inférieurs à la franchise sont présents dans la base de données mais ne sont pas en prendre en compte, car il ne seront pas payés par l'assureur. La franchise appliquée vaut 337.5€. Ainsi, un seuil à gauche égal à 337.5€ est appliqué à la base, cela conduit à l'exclusion de 25 393 observations, soit 3.24% d'observations exclues, représentant 0.35% du coût total des sinistres (non graves) de la base.

Après application de ces retraitements la base finale comporte 757 633 observations et 10 variables : la variable à expliquer, *i.e.* le coût du sinistre, et 9 variables explicatives. Au sein de cette base 86% des sinistres correspondent à une réparation et les autres 14% correspondent au cas où le véhicule était économiquement irréparable. Le coût moyen des sinistres de la base est égal à 2 297€. Les variables de l'étude ainsi que leurs domaines de définition sont explicités dans le tableau (1.3) ci-dessous :

Variable	**à valeurs dans**
Variables qualitatives	
`assureur`	{Assureur_1, Assureur_3, Assureur_7}
`expertA`	{O, N}
`marque`	{RENAULT, CITROEN, MERCEDES, PEUGEOT, ...}
`mois`	{201901, 201902, ..., 202011, 202012}
`tauxHoraire`	{TH fort, TH moyen, TH faible, NUL}
`trimestre`	{T1 2019, T2 2019, T3 2019,T4 2019,T1 2020, T2 2020, T3 2020, T4 2020}
`typeSinistre`	{ACT, AUT, ACP, APC, INC, VLT, VLP, GRE, TEM, CAT, BDG}
Variables quantitatives	
`ageVehicule`	]0 ; 80[(ans)
`coutSinistre`	]337.5 ; 30 000[(€)
`kilometrage`	[1 ; 500 000[(km)

TABLE 1.3 : Variables de l'étude et leurs domaines de définition

La signification des modalités du type de sinistre est explicitée en annexe (A.1). Concernant la variable représentant le réseau d'experts intervenu, elle possède deux modalités : `O` si le réseau A a été chargé d'évaluer le sinistre et `N` si c'était le réseau B.

Le retraitement de la base est complété par une discrétisation des variables numériques.

1.5.4 Discrétisation des variables numériques

Pourquoi discrétiser ?

La discrétisation des variables numériques est une étape courante lors du retraitement d'une base d'étude. Une première motivation est la diminution du temps de calcul : la discrétisation peut en effet permettre d'accélérer le fonctionnement de certains algorithmes. Dans le cadre de ce mémoire c'est notamment le cas pour l'algorithme MOB, présenté en section (3.2) : comme il sera vu ultérieurement, la discrétisation des variables numériques a permis de réduire significativement le temps d'apprentissage de l'algorithme. Pour les modèles GLM et CART, présentés respectivement en section (2.1) et (3.1), cet aspect est moins fondé car les temps de calculs de ces modèles sont relativement faibles et donc n'ont pas nécessairement besoin d'être réduits. Cependant, dans le cadre du GLM, la discrétisation d'une variable numérique permet d'obtenir plusieurs coefficients qui traduisent l'influence de la variable sur la variable d'intérêt, plutôt qu'un unique coefficient traduisant un effet moyen unique. Ainsi, la discrétisation permet une meilleure analyse de l'influence de la variable dans le modèle. La discrétisation des variables numériques n'est cependant pas une étape obligatoire avant de procéder à une modélisation, mais, pour les raisons expliquées précédemment, il a été choisi d'y avoir recours dans le cadre de cette étude.

Comment discrétiser ?

La base de l'étude contient deux variables explicatives numériques : l'âge du véhicule et son kilométrage. Il n'existe pas de méthode unique pour discrétiser une variable numérique, dans le cadre de ce mémoire deux approches ont alors été étudiées :
— la première consiste à discrétiser les variables numériques en se basant sur un nombre pré-défini de quantiles. Les catégories créées via cette méthode contiennent alors approximativement le même nombre d'observations. Cependant, cette méthode ne prend pas en considération la variable d'intérêt ;
— la seconde approche prend en compte cet aspect en créant des catégories de façon à minimiser l'hétérogénéité intra-classes selon la variable d'intérêt, ici le coût du sinistre. Ce principe est celui mis en place par l'algorithme CART, présenté en section (3.1).

Afin de choisir entre ces deux méthodes, plusieurs discrétisations sont effectuées via chacune des méthodes pour l'âge et le kilométrage du véhicule. La qualité d'une discrétisation sera mesurée par l'hétérogénéité totale qui en résulte. Pour une discrétisation en n groupes, l'hétérogénéité totale est calculée de la façon suivante

$$H_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{y \in G_i} (y - \overline{y}_i)^2,$$

où G_i est le groupe numéro n°i issu de la discrétisation et $\overline{y}_i$ est la moyenne des coûts pour les observations au sein du groupe i. Ainsi chaque discrétisation est associée à une mesure d'hétérogénéité et à une complexité, correspondant au nombre de groupes qui résultent de la discrétisation. La figure (1.9) ci-dessous présente les résultats obtenus avec les deux méthodes pour différents choix de complexité :

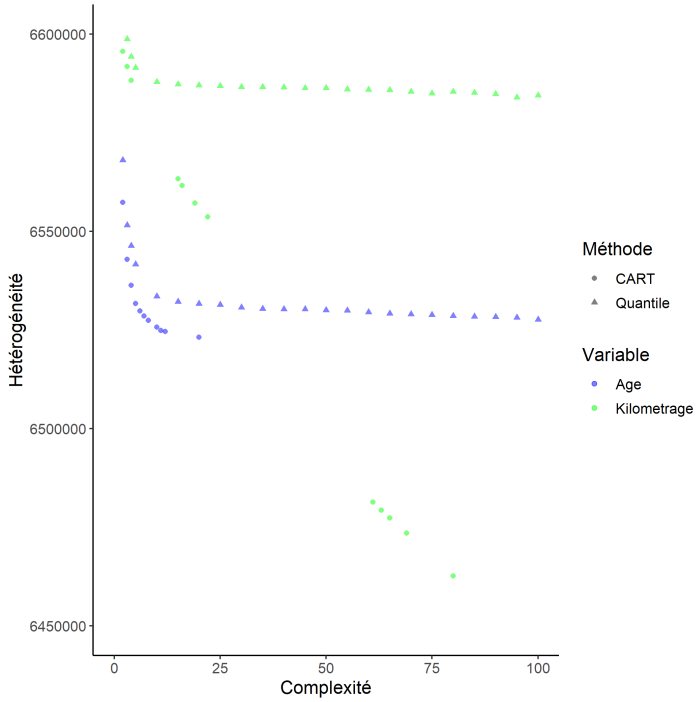


FIGURE 1.9 : Hétérogénéité selon la complexité pour les discrétisations via les quantiles et via CART pour l'âge et le kilométrage du véhicule

Il ressort de la figure (1.9) que la discrétisation via le modèle CART est de meilleure qualité que celle effectuée via la création de quantiles pour les deux variables étudiées. C'est donc cette méthode qui sera choisie.
Une fois le choix de la méthode effectué, il reste à déterminer le nombre de catégories souhaitées. Ce nombre doit être choisi en effectuant un compromis entre complexité et hétérogénéité. Une analyse des points représentés en figure (1.9) est alors effectuée. Pour une meilleure visualisation, les graphiques sont représentés séparément pour chaque variable et sont regroupés en figure (1.10) ci-dessous :

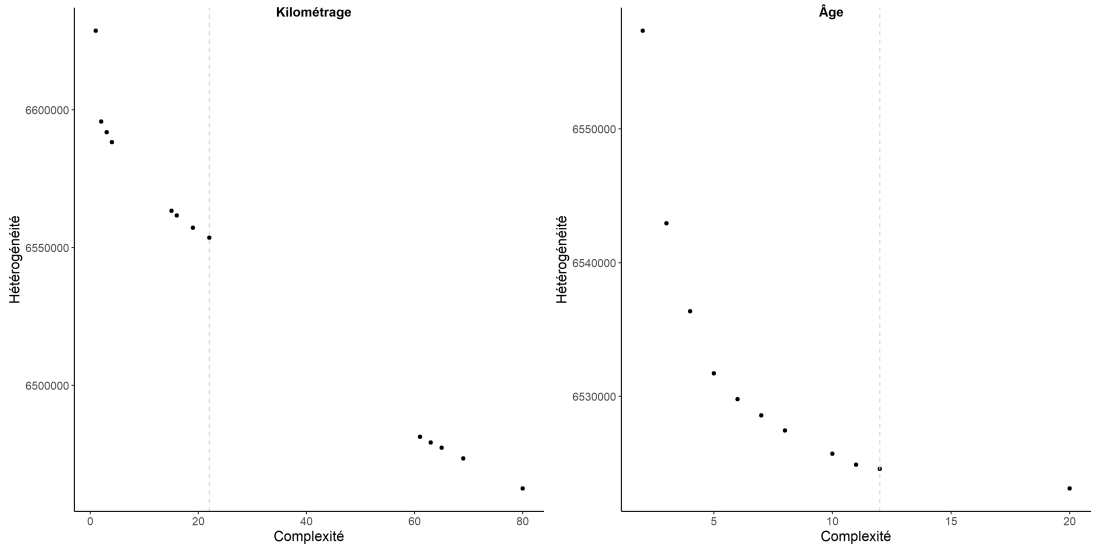


FIGURE 1.10 : Zoom sur l'hétérogénéité selon la complexité pour les discrétisations via CART pour l'âge et le kilométrage du véhicule

L'idée est de choisir un nombre de catégories qui ne soit pas trop élevé, afin de ne pas augmenter la complexité des modèles, mais pour lequel l'hétérogénéité induite est relativement faible. L'analyse de la figure (1.10) conduirait alors à choisir une discrétisation en 23 modalités pour le kilométrage, et 12 modalités pour l'âge. Ces choix sont représentés par les lignes pointillées sur la figure (1.10) ci-dessus. Cependant, il n'est pas souhaitable de créer des catégories contenant trop peu d'information : lors de la découpe en échantillon d'apprentissage, de validation et de test, chaque base doit idéalement contenir chaque modalité de chaque variable. Il faut en particulier que les modèles puissent être entraînés sur chacune de ces modalités. De plus, comme expliqué précédemment, il n'est pas souhaitable de créer une segmentation trop fine des individus. Le tableau (1.4) ci-dessous présente pour chaque discrétisation la complexité, *i.e.* la taille de la partition créée, et le nombre d'observations dans la plus petite catégorie :

Kilométrage	
Taille	Nombre d'observations dans la plus petite catégorie
91-88	12
80-65	22
64-62	30
63-61	58
22-15	72
4-3	71 993
2	166 281

Âge du véhicule	
Taille	Nombre d'observations dans la plus petite catégorie
20	425
12-5	988
4-3	91 619
2	254 055

TABLE 1.4 : Nombre d'observations dans la plus petite catégorie créée part CART pour différentes tailles de discrétisation pour l'âge et le kilométrage du véhicule

En choisissant d'effectuer une discrétisation en 23 catégories pour le kilométrage, la plus petite catégorie contiendrait 72 observations. Ce choix n'est donc pas adéquat pour les raisons explicitées précédemment. De même pour l'âge, la discrétisation en 12 catégories conduirait à la création d'un groupe trop petit. Il est alors choisi d'effectuer une discrétisation en quatre catégories pour les deux variables, ce choix permettant de créer des groupes de tailles suffisamment grandes pour pouvoir être correctement représentés dans la base d'étude. Le tableau (1.5) ci-dessous résume les catégories créées pour chacune des deux variables, ainsi que leurs caractéristiques :

Kilométrage		
Catégorie (km)	Nombre d'observations	Coût moyen
[1 ; 100 085)	410 122	2 438
[100 085 ; 165 601)	181 130	2 292
[165 601 ; 220 000)	94 388	2 071
[220 000 ; 500 000)	71 993	1 803

Âge du véhicule		
Catégorie (an)	Nombre d'observations	Coût moyen
]0 ; 5.5)	347 167	2 554
[5.5 ; 8.9)	156 411	2 339
[8.9 ; 14)	162 436	2 078
[14 ; 80)	91 619	1 645

TABLE 1.5 : Catégories crées via la discrétisation CART pour l'âge et le kilométrage du véhicule

L'importance de la bonne représentation de chaque catégorie dans la base ne s'applique pas uniquement aux variables numériques qui sont discrétisées, mais également aux variables qualitatives lorsque celles-ci possèdent des modalités faiblement représentée dans la base. C'est notamment souvent le cas lorsqu'une variable présente un nombre important de modalités.

1.5.5 Traitement des variables qualitatives possédant un nombre élevé de modalités

Une étape habituelle du retraitement d'une base avant de procéder à la modélisation consiste à diminuer le nombre de modalités des variables qualitatives lorsque celui-ci est considéré comme trop important.

Pourquoi diminuer le nombre de modalités ?

Lorsqu'une variable contient un nombre élevé de modalités, il est possible que certaines d'entre elles soient très peu représentées dans la base. Cela implique que les modèles seront entraînés sur un nombre très faible d'observations associées à ces modalités, ce qui risque d'impliquer de mauvais résultats de prédiction. Une seconde motivation de la réduction du nombre de modalités concerne le temps de calcul et la complexité du modèle. En effet, pour une variable possédant un nombre élevé de modalités, la plupart des modèles analyseront chaque modalité une par une, parfois plusieurs fois, par exemple dans les modèles à structure arborescente, ce qui augmentera le temps d'apprentissage des modèles. Dans le cadre de ce mémoire, le modèle GLM est étudié. Ce modèle calcule un coefficient par modalité de chaque variable, la réduction du nombre de modalités permet donc une réduction de la complexité du modèle. Il sera également vu par la suite que la réduction du nombre de modalités a permis une accélération relativement importante du modèle MOB.

Comment diminuer le nombre de modalités ?

Parmi les variables explicatives de l'étude, une seule variable qualitative présente un nombre trop élevé de modalités pour être directement exploitée : il s'agit de la marque du véhicule, qui possède plus de 640 modalités. Un premier retraitement est effectué : cette variable possède, par exemple, les modalités `PEUGEOT, PGT` et `PEUG`, qui correspondent toutes à la marque de voiture Peugeot. Les modalités de cette variable sont donc retraitées afin d'avoir une harmonisation du nom des marques dans la base. Cependant, même après ce retraitement, le nombre de modalités de cette variable reste trop élevé : la base contient 290 marques parmi lesquelles 272 sont représentées par moins de 1% des observations.

Plusieurs approches ont été étudiées afin de réduire le nombre de modalités de la variable marque. Une première approche consiste à conserver les n marques les plus représentées dans la base et à regrouper les marques restantes dans une modalité appelée `AUTRES`, avec n choisi préalablement. Cette approche présente cependant des limites car elle ne prend pas en considération le coût moyen associé à chaque marque, mais seulement le nombre de véhicules qui y sont associés. Afin de prendre en compte la variable d'intérêt, une seconde approche est étudiée. Cette approche est similaire à celle effectuée en section (1.5.4) pour la discrétisation de l'âge et du kilométrage du véhicule. Cette méthode, basée sur l'algorithme CART, permet de construire des groupes au sein desquels l'hétérogénéité selon le coût des sinistres est minimale. Pour savoir quelle méthode choisir, une comparaison de l'hétérogénéité totale en fonction de la complexité entre les deux méthodes a été effectuée. Les résultats sont présentés

graphiquement dans la figure (1.11) ci-dessous :

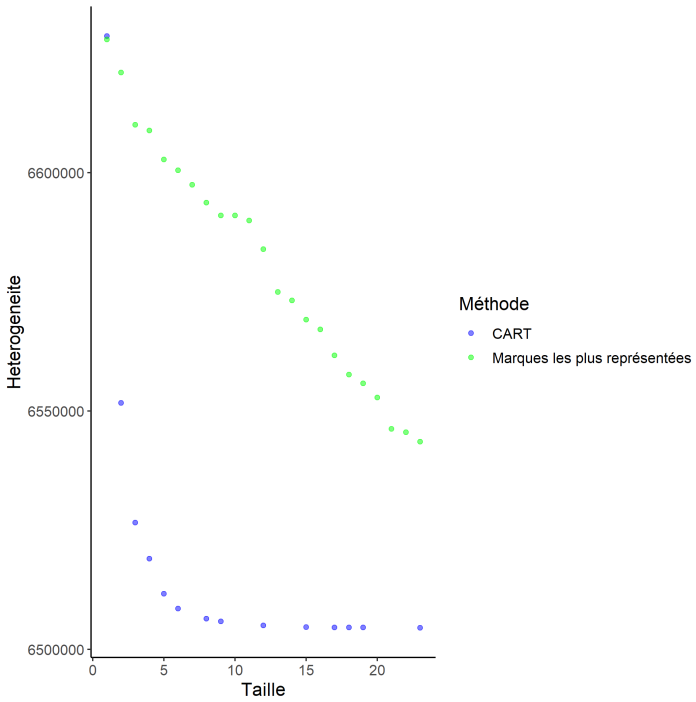


FIGURE 1.11 : Hétérogénéité selon la complexité pour différentes méthodes de réduction du nombre de modalités de la variable `marque`

La création de groupes de marques via la méthode CART donne de meilleur résultat en termes d'hétérogénéité et de complexité, c'est donc cette approche qui sera choisie. Il reste enfin à déterminer le nombre de catégories souhaité. Comme vu précédemment, ce choix doit être fait de façon à effectuer un compromis entre complexité et diminution d'hétérogénéité, mais aussi en prenant en compte le nombre d'observations contenues dans le plus petit groupe. Le tableau (1.6) ci-dessous répertorie le nombre d'observations dans le plus petit groupe créé par CART :

Marque	
Taille	Nombre d'observations dans la plus petite catégorie
23-17	1 012
15-4	1 058
3	9 094
2	141 047

TABLE 1.6 : Nombre d'observations dans la plus petite catégorie de marque créée par CART pour différentes tailles de discrétisation

Un groupe contenant seulement 1 058 observations a un risque élevé de ne pas être suffisamment représenté dans l'échantillon d'apprentissage. La méthode de regroupement par CART conduirait alors à choisir de créer 3 catégories. Cependant, en s'intéressant aux arbres créés par CART les résultats suivants ont pu être observés :

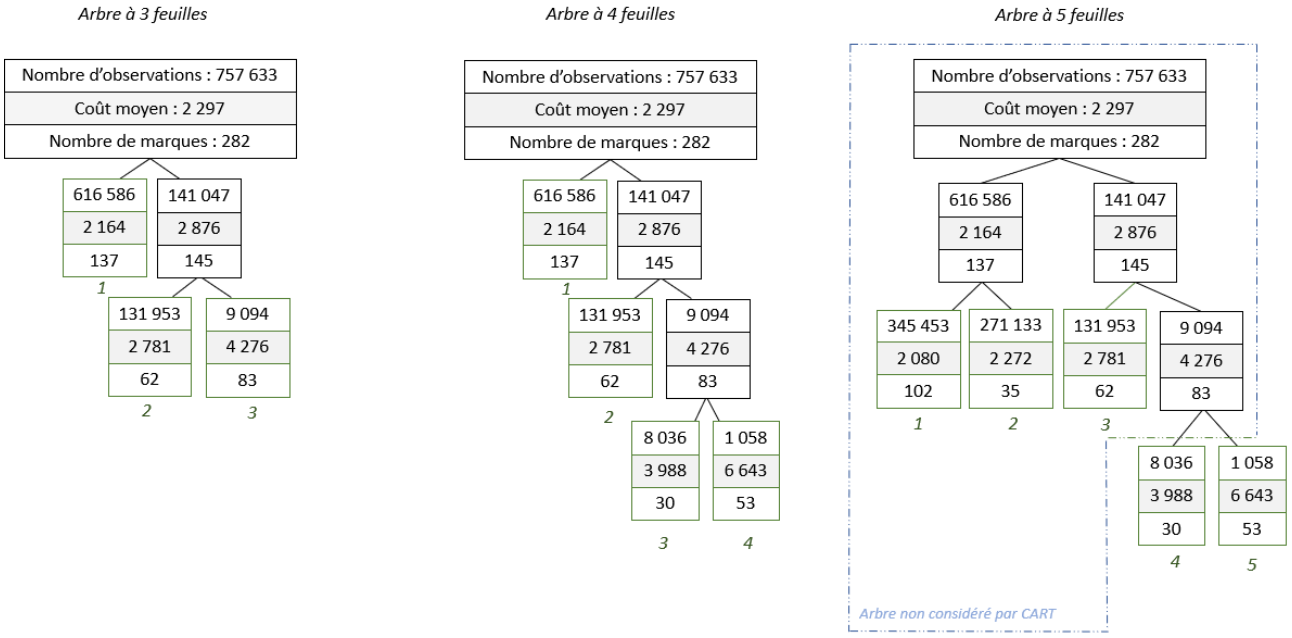


FIGURE 1.12 : Arbres à 3, 4 et 5 feuilles construit par CART selon la variable `marque`

La figure (1.12) présente les arbres à respectivement 3, 4 et 5 feuilles construits par CART selon la formule `coutSinistre` $\sim$ `marque`. La feuille qui contient le plus petit nombre d'observations pour les arbres à 4 et 5 feuilles est la feuille regroupant les marques associées au coût moyen le plus élevé. La figure (1.12) ci-dessus montre que la méthode de regroupement par CART présente des limites : en se restreignant à des groupes qui contiennent un nombre suffisamment élevé d'observations, l'arbre choisi serait l'arbre à 3 feuilles. Cependant, comme il sera vu ultérieurement, l'algorithme CART ne considère pas tous les arbres à 4 feuilles, en particulier il ne considère pas l'arbre à 4 feuille encadré sur la figure (1.12). Pourtant, cet arbre conduit à une meilleure hétérogénéité que l'arbre à 3 feuilles construit par CART et ne crée par de groupes trop "petits". Il est donc choisi de construire les groupes de marque selon cet arbre : en partant de l'arbre à 5 feuilles, les deux catégories de marques associées aux coûts les plus élevés, correspondant aux feuilles 4 et 5 de l'arbre, sont regroupées dans une seule catégorie qui contient alors 9 094 observations. Les caractéristiques du regroupement effectué sont résumées dans le tableau (1.7) ci-dessous :

Nom du groupe	Nombre d'observations	Coût moyen	Nombre de marques
`Categorie_marque1`	345 453	2 080	102
`Categorie_marque2`	271 133	2 272	35
`Categorie_marque3`	131 953	2 781	62
`Categorie_marque4`	9 094	4 276	83

TABLE 1.7 : Catégories créées après regroupement par CART pour la marque du véhicule

Les autres variables de la base présentent un nombre de modalités assez faible pour pouvoir être directement utilisées lors de la modélisation. Cependant, la variable spécifiant le type de sinistre possède 11 modalités, ce qui est un nombre suffisamment faible pour qu'elle puisse être directement exploitée, mais comme le montre la figure (1.13) ci-dessous, certains types de sinistres sont très peu représentés dans la base :

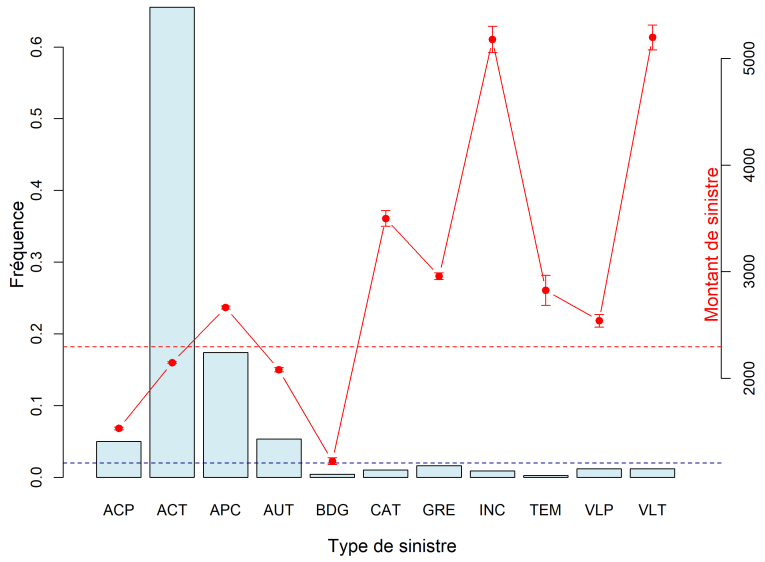


FIGURE 1.13 : Fréquence des modalités de la variable `typeSinistre`

La figure (1.13) présente la fréquence de chacune des modalités au sein de la base d'étude. Pour chaque modalité, le coût moyen des individus présentant la modalité (points rouges) et les intervalles de confiance à 95% associés ont été représentés. Le coût moyen dans l'ensemble de la base, à savoir 2 297€ a également été tracé (ligne pointillée rouge).
Les modalités `BDG, CAT, GRE, INC, TEM, VLP` et `VLT`[1] sont représentées par moins de 2% des observations de la base (ligne pointillée bleue). Comme il a été expliqué précédemment, cela implique un risque de mauvaise représentation de ces modalités au sein des échantillons d'apprentissage, de validation et de test. Un regroupement de ces modalités est alors effectué. Une première idée serait de rassembler l'ensemble de ces modalités dans un seul groupe, mais cette méthode ne prend pas en compte le coût des sinistres et conduirait par exemple à regrouper les sinistres "bris de glace" (`BDG`) et les sinistres "incendie" (`INC`) ensemble alors qu'il peut être observé sur la figure (1.13) que ces types de sinistres sont associés à des coûts moyens relativement différents. La méthode choisie pour effectuer le regroupement est similaire à la méthode utilisée pour la marque. Cette méthode est basée sur l'algorithme CART et conduit à effectuer les regroupements suivants :

Nom du groupe	Nombre d'observations	Coût moyen
`ACP-BDG`	41 332	1 505
`AUT`	40 518	2 081
`ACT`	496 360	2 148
`APC-VLP`	141 134	2 656
`GRE-TEM-CAT`	22 328	3 138
`INC-VLT`	15 991	5 190

TABLE 1.8 : Catégories créées après regroupement par CART pour le type de sinistre

Le tableau (1.8) montre que le regroupement des modalités via CART permet une meilleure représentation de chacune d'entre elles au sein de la base d'étude. Il sera vu par la suite que ce

[1]Pour rappel, les significations des différents type de sinistres sont explicitées en annexe (A.1).

regroupement a également permis une accélération des temps de calcul, notamment pour le modèle MOB. La variable `typeSinistre` ainsi retraitée sera dénotée `TS` dans la suite de ce mémoire.

Une fois la base retraitée, il est intéressant de procéder à une analyse des variables la composant.

1.6 Étude des variables

Dans le cadre de cette étude, on s'intéresse à la variable `coutSinistre`, que l'on cherche à prédire au mieux via différents modèles, et à la variable `expertA`, dont on cherche à mesurer l'influence sur le montant de sinistre prédit. Chacune des observations de la base contient un montant de sinistre et une variable spécifiant par quel réseau ce montant a été évalué. La figure (1.14) ci-dessous donne la fréquence des sinistres et le coût moyen (représenté par les points rouges) par réseau ainsi que les intervalles de confiance à 95% associés. La ligne rouge pointillée représente le coût moyen des sinistres dans toute la base, égal à 2 297€.

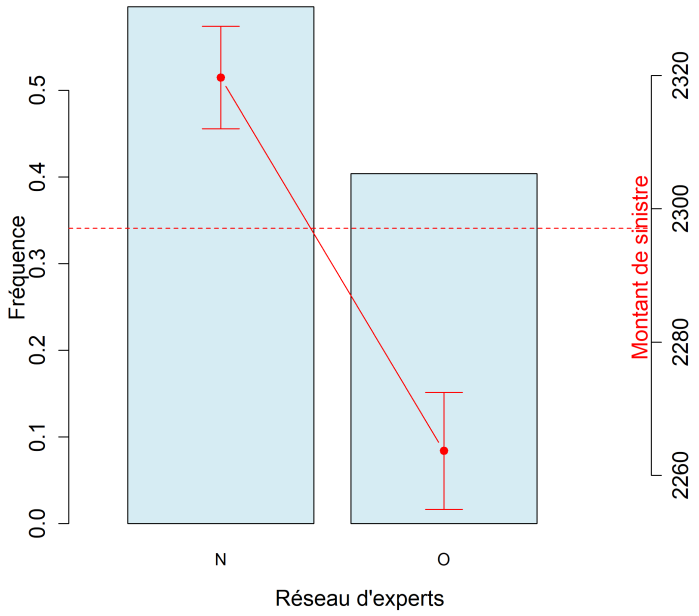


FIGURE 1.14 : Sinistralité en fonction du réseau d'expert

Dans la base d'étude, plus d'observations sont associées au réseau d'experts B (qui correspond au cas où `expertA = N`) qu'au réseau d'experts A : environ 40% des observations représentent des sinistres expertisés par ce dernier. D'autre part, le coût moyen des sinistres expertisés par le réseau A est plus faible que celui des sinistres expertisés par le réseau B. Plus précisément la différence de coût moyen entre les deux réseaux est égale à environ 56€, ce qui correspond à environ 2.5% du coût moyen associé au réseau A.
Ainsi, la considération seule de cette variable amènerait à conclure que le réseau d'experts A est plus performant que le réseau d'experts B. Cependant, comme expliqué précédemment, cette conclusion serait erronée car il suffit par exemple que le réseau A ait expertisé des sinistres à plus faible coût pour arriver à un tel résultat. Afin de s'assurer de la pertinence des comparaisons, une étude graphique de

la distribution des variables et des coûts de sinistres associés à chacune de leur modalités pour chacun des deux réseaux d'experts est effectuée.

1.6.1 Étude des variables par réseau d'expert

Selon le type de sinistre

La sinistralité selon le type de sinistre pour les deux réseaux d'experts est représentée par la figure (1.15) ci-dessous :

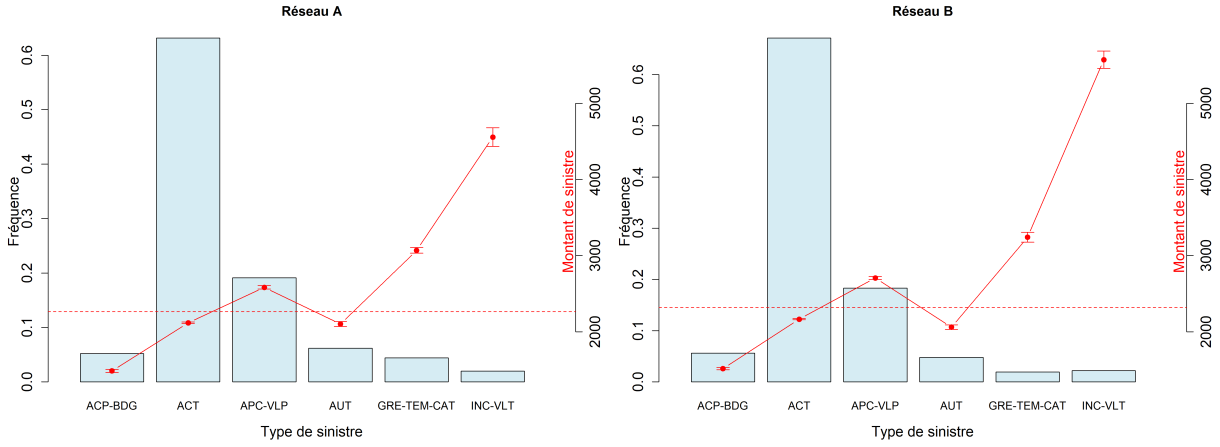


FIGURE 1.15 : Sinistralité en fonction du type de sinistre

Graphiquement, il apparaît que la fréquence et le coût des sinistres sont globalement similaires entre les deux réseaux d'experts. De plus, il est possible d'observer que les sinistres les plus fréquents dans la base sont les accidents avec tiers (`ACT`). Les sinistres les plus coûteux en moyenne sont les sinistres liés à un incendie ou à un vol total (`INC-VLT`). Il est possible de noter une différence de coût moyen relativement importante entre les deux réseaux d'experts pour les sinistres appartenant à cette catégorie. Cette différence est de l'ordre de 1 015€, ce qui correspond à environ 45% du coût moyen associé au réseau A.

Selon l'assureur

La sinistralité selon l'assureur en charge du sinistre pour les deux réseaux d'experts a été représentée sur la figure (1.16) ci-dessous :

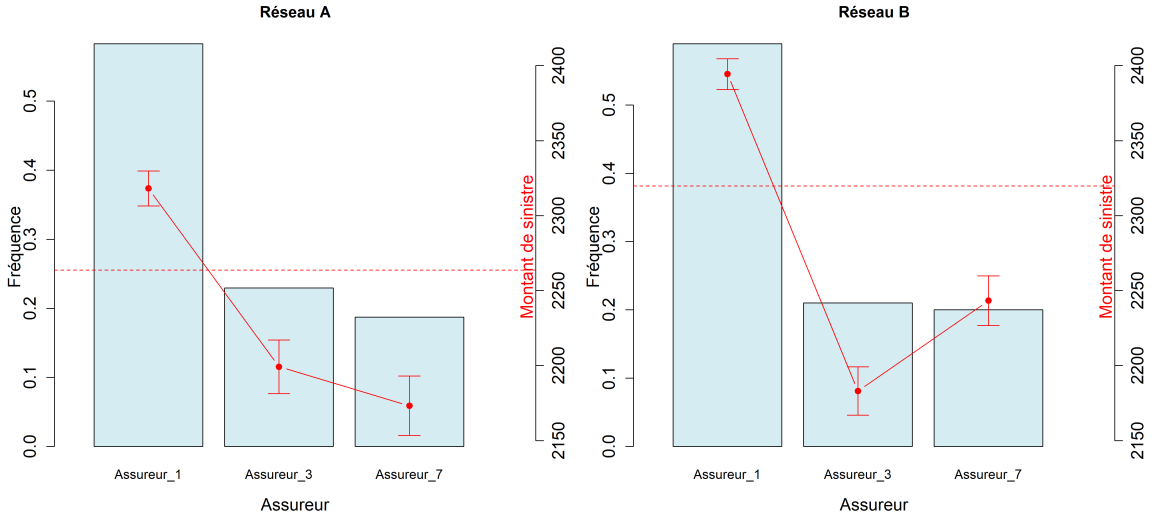


FIGURE 1.16 : Sinistralité en fonction de l'assureur

Il est possible d'observer la même fréquence de représentation des assureurs pour les deux réseaux d'experts. Il ressort de ces graphiques que l'assureur n°1 est le plus représenté dans la base, il est également le plus coûteux en moyenne. Des différences de coûts moyens entre les deux réseau d'experts sont observées pour chacun des assureurs, bien que ces différences soient relativement faibles : la différence la plus importante de coût moyen entre les deux réseaux est égale à environ 76€, ce qui représente environ 3% du coût moyen associé au réseau A.

Selon le taux horaire

La sinistralité en fonction du taux horaire associé au garage intervenu est représentée sur la figure (1.17) ci-dessous pour chacun des deux réseaux d'experts :

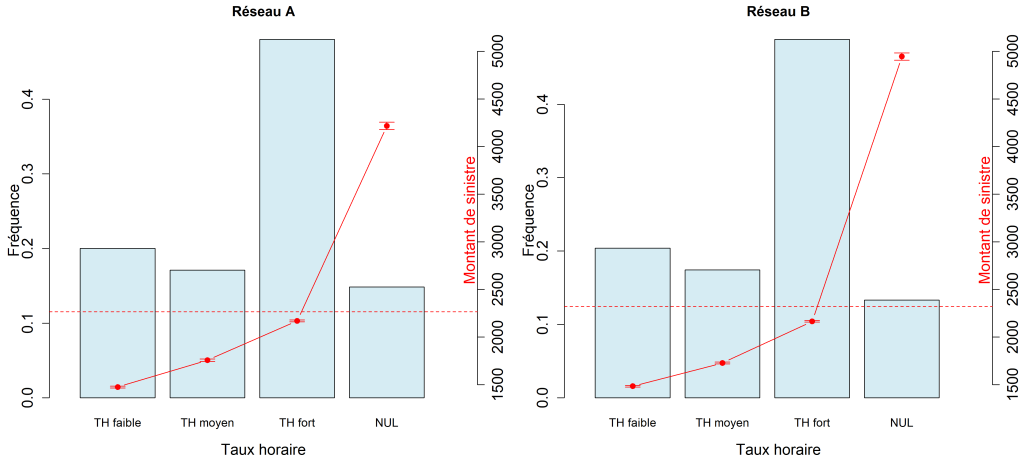


FIGURE 1.17 : Sinistralité en fonction du taux horaire

Un comportement de la sinistralité similaire d'un réseau à l'autre selon les différents taux horaires

peut être observé. Le taux horaire le plus présent dans la base est le taux horaire fort. Il ressort une croissance du coût du sinistre avec le taux horaire, ce qui est logique : plus un garage a un taux horaire élevé, plus important sera le montant des réparations et donc le coût du sinistre. C'est pour le cas ou le taux horaire est `NUL` que le coût du sinistre est le plus élevé. En effet, cela correspond au cas où le véhicule est économiquement irréparable et donc où le coût du sinistre correspond à la valeur du véhicule. Pour ce taux horaire une différence de coût moyen de plus de 728€ entre les deux réseaux est observée, ce qui représente plus de 32% du coût moyen associé au réseau d'experts A.

Selon le trimestre

L'évolution trimestrielle de la sinistralité pour les deux réseaux d'experts a été représenté ci-dessous en figure (1.18) :

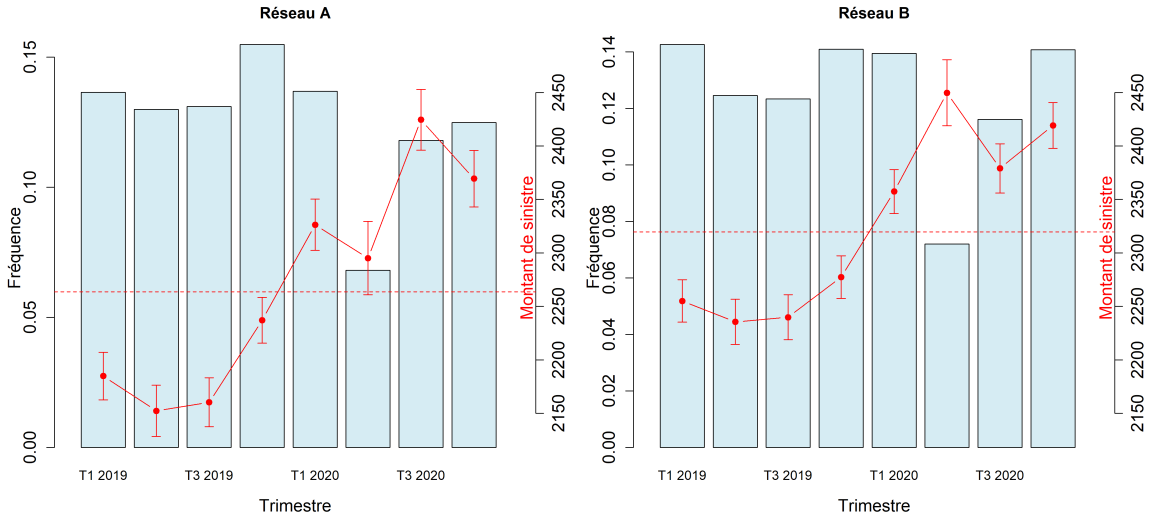


FIGURE 1.18 : Sinistralité en fonction du trimestre

La tendance d'évolution de la sinistralité d'un trimestre à l'autre est la même entre les deux réseaux d'experts pour l'année 2019, avec des coût de sinistres plus élevés en moyenne pour le dernier trimestre de cette année. Pour l'année 2020, quelques différences d'évolution entre les deux réseaux d'experts concernant le coût moyen sont observées. Les trimestres sont représentés similairement dans la base en termes de fréquence pour les deux réseaux d'experts. Le trimestre pour lequel la plus grande différence de coût moyen entre les deux réseaux d'experts est observée est le deuxième trimestre de 2020. Cette différence est de l'ordre de 155€, ce qui représente environ 7% du coût moyen associé au réseau A.
Pour mieux comprendre et comparer l'évolution de la sinistralité à travers le temps, un zoom sur la maille mensuelle est effectué.

Selon le Mois

L'évolution mensuelle de la sinistralité pour les deux réseaux d'experts a été représentée ci-dessous en figure (1.19) :

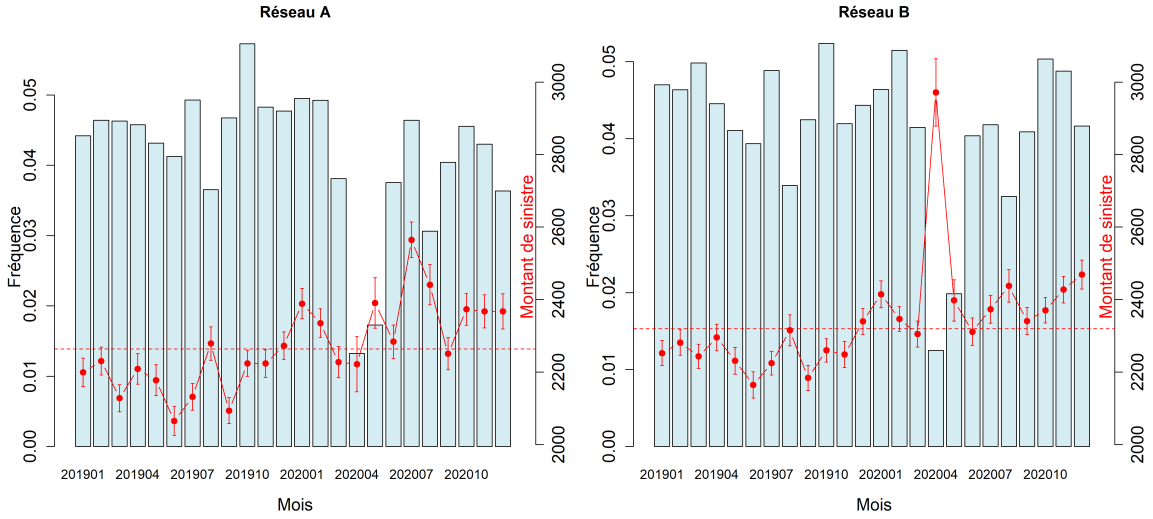


FIGURE 1.19 : Sinistralité en fonction du mois

Globalement une même évolution de la sinistralité à travers le temps pour les deux réseaux d'experts sur l'année 2019 est observée. Quelques différences d'évolution sont observées pour l'année 2020. La différence la plus importance de coût moyen entre les deux réseaux est observée pour le mois d'avril 2020 et est égale à environ 750€, soit plus de 33% du coût moyen associé au réseau A. Ce mois est également celui qui présente la fréquence la plus faible parmi les mois représentés dans la base. Cela peut notamment s'expliquer par le fait que la France a connu une période de confinement du 17 mars au 11 mai 2020, la population française était donc confinée durant tout le mois d'avril 2020.

Selon la marque

La sinistralité au sein des différentes catégories de marque de véhicule pour les deux réseaux d'experts est représentée sur la figure (1.20) ci-dessous :

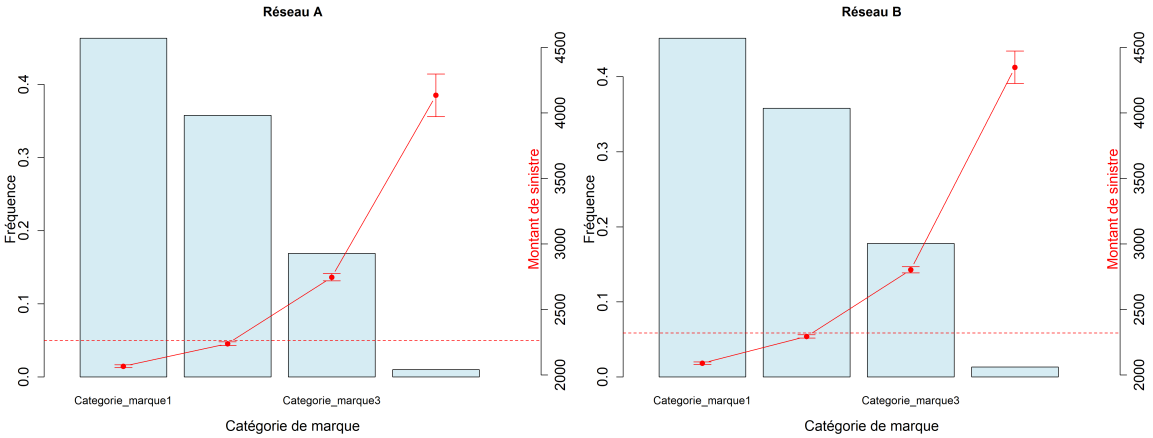


FIGURE 1.20 : Sinistralité en fonction de la marque du véhicule

Un comportement de la sinistralité selon la marque du véhicule similaire d'un réseau d'experts à

l'autre est observé. La catégorie de marque la plus représentée dans la base est celle qui regroupe les marques associées au coût moyen le plus faible. La plus grande différence de coût moyen entre les deux réseaux est égale à environ 214€ et est observée pour la catégorie de marque n°4. Cette différence est de l'ordre de 10% du coût moyen observé pour le réseau d'experts A.

Selon l'âge du véhicule

La sinistralité au sein des différentes catégories d'âge de véhicule pour les deux réseaux d'experts est représentée sur la figure (1.21) ci-dessous :

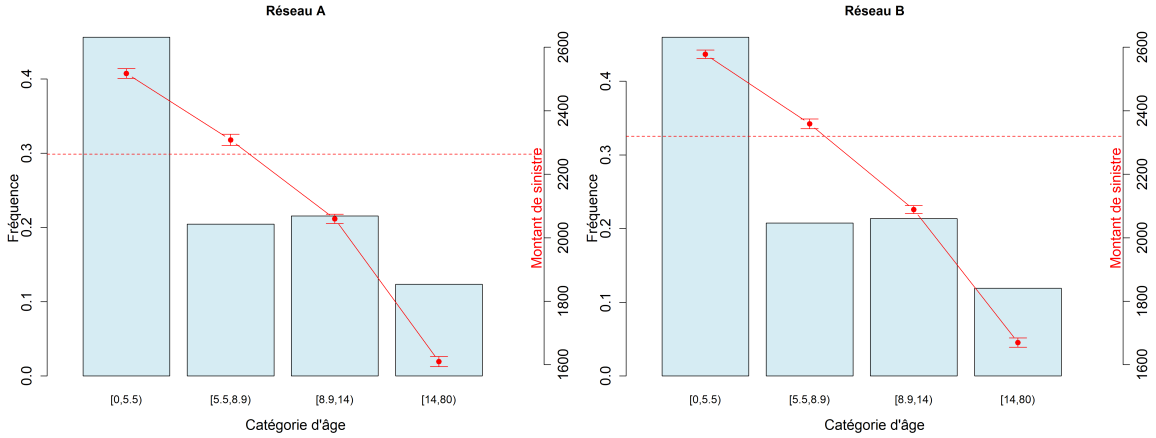


FIGURE 1.21 : Sinistralité en fonction de l'âge du véhicule

L'évolution de la sinistralité selon l'âge du véhicule est la même d'un réseau à l'autre. La figure (1.21) montre également que le coût du sinistre est une fonction décroissante de l'âge du véhicule, ce qui est logique car, en général, la valeur d'un véhicule diminue avec le temps. La catégorie de véhicule la plus représentée dans la base est celle qui contient les véhicules âgés de moins de 5 ans. Les différences de coût moyen entre les réseaux d'experts sont relativement faibles pour chacune des modalités, puisqu'elles n'excèdent pas 61€, ce qui représente moins de 3% du coût moyen associé au réseau d'experts A.

Selon le kilométrage

La sinistralité au sein des différentes catégories de kilométrage du véhicule pour les deux réseaux d'experts est représentée sur la figure (1.22) ci-dessous :

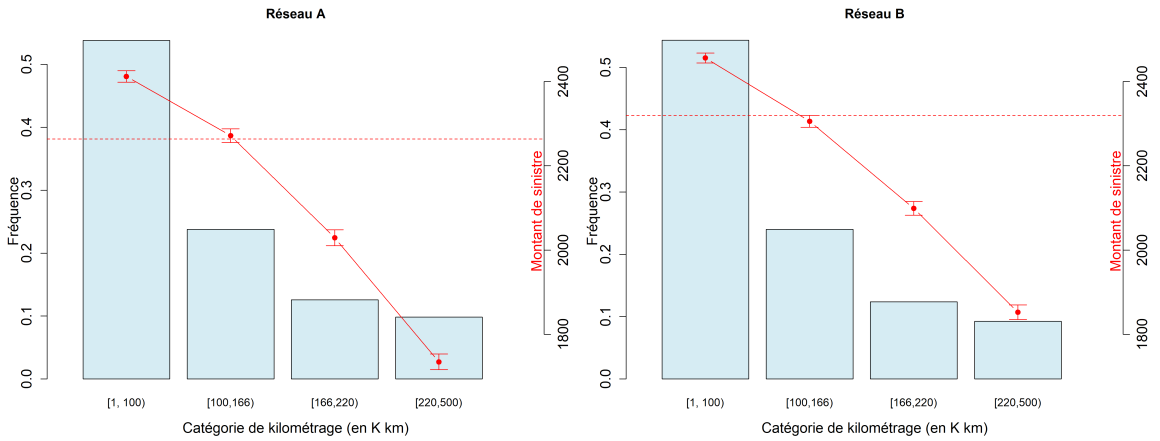


FIGURE 1.22 : Sinistralité en fonction du kilométrage

Une similitude du comportement de la sinistralité selon le kilométrage entre les deux réseaux d'experts est observée. Une décroissance du coût moyen avec le kilométrage du véhicule peut être soulignée : cela s'explique par le fait que lorsqu'un véhicule a un kilométrage élevé, il est souvent assez âgé et a donc perdu de la valeur. La figure (1.22) permet également d'observer que la catégorie de véhicule la plus représentée dans la base est celle des véhicules dont le kilométrage est inférieur à 100 085 kilomètres. La différence la plus importante du coût moyen entre les réseaux d'experts s'observe au sein de la catégorie regroupant les véhicules ayant parcouru le plus de kilomètres, elle est de l'ordre de 118€, ce qui représente environ 5% du coût moyen associé au réseau d'experts A.

Les études graphiques précédentes permettent d'observer que les sinistres traités par les réseaux d'experts A et B semblent bien être de même nature. Cependant, conclure que cela implique que le réseau d'experts A est plus performant que le réseau d'experts B ne serait pas rigoureux car il se peut que les représentations graphiques précédentes ne permettent pas d'observer quelques dissimilarités existantes entre les sinistres expertisés par les deux réseaux. L'accumulation de petites différences entre ces sinistres pourrait avoir un impact global assez important. Afin de garantir l'exactitude des conclusions sur la performance des deux réseaux des méthodes de modélisation sont implémentées.

Avant de procéder à une modélisation, une dernière étape consiste à vérifier que le découpage en échantillon d'apprentissage, de validation et de test n'a pas conduit à la création d'échantillons qui soient trop différents. Pour vérifier cela une représentation graphique de la fréquence de chaque modalité des variables explicatives de la base a été effectuée. Les résultats obtenus sont présentés en annexe (A.2) et permettent de conclure que le découpage effectué ne conduit pas à des différences trop importantes au niveau des variables explicatives.
Il est également important de vérifier que la variable d'intérêt, à savoir le coût du sinistre, est correctement représentée au sein des trois bases. La figure (1.23) ci-dessous présente un histogramme du coût des sinistres au sein de chacune des trois bases :

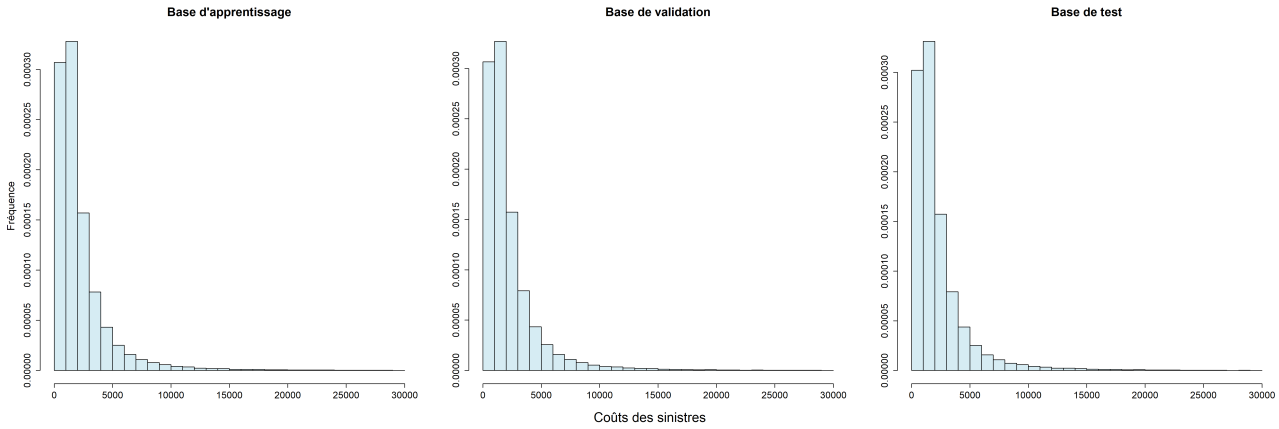


FIGURE 1.23 : Histogramme du coût des sinistres au sein des bases d'apprentissage, de validation et de test

L'allure de l'histogramme est la même pour chacune des trois bases. Ainsi la découpe en échantillons d'apprentissage, de validation et de test effectuée précédemment peut être conservée. Il est alors possible de procéder à une première modélisation.

Chapitre 2

Modèles linéaires généralisés

Une première modélisation est effectuée via un modèle linéaire généralisé [Denuit et Charpentier (2004)]. Les modèles linéaires généralisés (souvent notés GLM pour *Generalized Linear Models* en anglais) permettent d'étudier un lien non nécessairement linéaire entre une variable réponse et un ensemble de variables explicatives. Ces modèles sont apparus avec John Nelder et Robert Wedderburn en 1972 et sont couramment utilisés dans les domaines de la statistique et de l'actuariat.

Ce chapitre présente la théorie associée aux modèles linéaires généralisés et les résultats qui ont été obtenus après application de ce dernier. Ces résultats serviront de référentiel pour évaluer l'amélioration fournie par les modèles présentés par la suite. Ce chapitre décrit également les liens existants entre les variables de la base, ainsi que les traitements qui ont été appliqués en conséquence.

2.1 Théorie du modèle linéaire généralisé

Comme son nom l'indique le modèle linéaire généralisé est une extension d'un modèle plus simple : le modèle linéaire [Thomas (2014)]. Ce dernier permet de décrire une relation linéaire entre une variable d'intérêt et des covariables, mais il repose sur des hypothèses qui ne sont pas toujours adaptées en pratique. En particulier, il suppose l'existence d'une relation linéaire entre la variable à expliquer, notée Y, et des covariables, or cette hypothèse ne correspond pas toujours à la réalité et implique que Y puisse être à valeurs dans $\mathbb{R}$ tout entier, ce qui n'est pas toujours le cas. Par exemple, si Y représente le coût d'un sinistre, il serait aberrant que le modèle prédise des valeurs négatives pour cette variable. D'autre part, le modèle linéaire suppose que les observations $Y_1, ..., Y_n$ sont la réalisation d'une variable gaussienne et que les Y_i $(i = 1, ..., n)$ sont toutes de même variance. Ces hypothèses ne sont pas toujours adaptées en pratique, par exemple si la variable Y est une variable discrète ou si Y suit une loi de Poisson ou Gamma, pour lesquelles la variance varie selon la moyenne. L'utilisation du modèle linéaire généralisé permet de s'affranchir de ces hypothèses restrictives et de travailler avec un cadre plus général.

Dans ce mémoire la variable de réponse que l'on cherche à expliquer est le coût des sinistres. Soit X l'ensemble des variables explicatives, l'idée est de modéliser l'espérance de Y conditionnellement à X car, comme il a été expliqué précédemment, cette valeur correspond à la meilleure approximation de Y lorsque X est connue. En pratique, on dispose n observations et de p variables explicatives. On note $Y = (Y_1, ..., Y_n) \in \mathbb{R}^n$, $X = (X_0, X_1, ..., X_p) \in \mathbb{R}^{n \times p+1}$ où $X_0 = \mathbb{1}_n$ correspond à l'intercept.

Pour $j \in \{1, ..., p\}$, $X_j = (x_{1j}, ..., x_{nj})^T$ est l'ensemble des observations pour la $j^{\text{ème}}$ variable et pour $i \in \{1, ..., n\}$, $(x_{ij})_{0 \leq j \leq p}$, est l'ensemble des variables explicatives pour un individu i.

2.1.1 Les composantes d'un GLM

Un GLM est la résultante de 3 composantes :

— Une composante aléatoire
Il s'agit des observations du modèle. Le modèle linéaire généralisé s'appuie sur l'hypothèse que celles-ci suivent une distribution appartenant à la famille exponentielle de paramètre inconnu $\theta = (\theta_1, ..., \theta_n)^T$, appelé paramètre canonique ou bien paramètre naturel de la loi. Une variable aléatoire Y suit une loi appartenant à la famille exponentielle, si sa densité, sous la mesure adéquate (la mesure de Lebesgue si Y est continue, la mesure de comptage sur l'ensemble des entiers si Y est discrète) est de la forme

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

où :

— ϕ est un paramètre de dispersion, supposé connu ;
— $a(.)$ est une fonction non nulle dérivable sur $\mathbb{R}$;
— $b(.)$ est une fonction trois fois dérivable sur $\mathbb{R}$ et dont la dérivée première est inversible ;
— $c(.,.)$ est une fonction dérivable sur $\mathbb{R}^2$ qui ne dépend pas du paramètre canonique θ.

Les deux premiers moments sont donnés par la formule (2.1)

$$\mathbb{E}(Y) = b'(\theta) \text{ et } \mathbb{V}(Y) = a(\phi)b''(\theta), \tag{2.1}$$

Les lois usuelles suivantes font partie de la famille exponentielle : normale, Bernoulli, binomiale, Poisson, gamma, binomiale négative.

— Une composante déterministe
Il s'agit d'une fonction linéaire des variables explicatives faisant intervenir un nombre fini de paramètres à déterminer

$$\beta_0 + \beta_1 X_1 + ... + \beta_p X_p,$$

où les $\beta_j, (j = 0, ..., p)$ représentent les coefficients de régression. L'estimation de $\beta = (\beta_0, ..., \beta_p) \in \mathbb{R}^{p+1}$ se fait via la méthode du maximum de vraisemblance.

— Une fonction de lien
Il s'agit d'une fonction strictement monotone et dérivable, notée g, qui lie les deux composantes précédentes de telle sorte que l'on ait

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

en particulier, avec la formule (2.1), pour tout $i = 1, ..., n,\ \theta_i = (b')^{-1} \circ g^{-1}(x_i\beta)$.

Chacune des lois de distribution de la famille exponentielle possède une fonction de lien telle que $g(\mathbb{E}[Y_i]) = \theta_i$ pour tout $i \in \{1, ..., n\}$. Cette fonction de lien est appelée fonction de lien canonique et correspond au cas où $g = (b')^{-1}$.

Le modèle linéaire est un cas particulier du GLM pour lequel la loi est une distribution normale et la fonction de lien est la fonction de lien canonique associée, à savoir la fonction identité. La loi normale appartient bien à la famille exponentielle. En effet la densité d'une loi normale $\mathcal{N}(\mu, \sigma^2)$, est

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

ce qui peut être réécrit :

$$\begin{aligned} f(x; \mu, \sigma) &= \exp\{-\log(\sqrt{2\pi}\sigma) - \frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\} \\ &= \exp\{\frac{x\theta - b(\theta)}{a(\phi)} + c(x, \phi)\}, \end{aligned}$$

où,

$$\theta = \mu,\ a(\phi) = \phi,\ \phi = \sigma^2,\ b(\theta) = \frac{\theta^2}{2} \text{ et } c(x, \phi) = -\frac{1}{2}(\frac{x^2}{\sigma^2} + \log(2\pi\sigma^2)).$$

Afin d'assurer la pertinence de l'utilisation d'un modèle GLM, différents outils permettent sa validation.

2.1.2 Validation d'un modèle GLM

Pour assurer la performance d'un modèle linéaire généralisé il faut pouvoir vérifier sa bonne adéquation aux données. Plusieurs outils peuvent être utilisés pour ce faire, dont certains sont présentés ci-après :

Déviance

La déviance d'un modèle est définie par

$$D = 2 \times (\ln(\mathcal{L}(Y|Y)) - \ln(\mathcal{L}(Y|\widehat{\beta})).$$

Elle est basée sur la différence entre la log-vraisemblance du modèle saturé $\ln(\mathcal{L}(Y|Y))$ et celle du modèle estimé $\ln(\mathcal{L}(Y|\widehat{\beta}))$. Le modèle saturé est le modèle possédant autant de paramètres que d'observations : c'est le modèle qui s'ajuste le mieux aux données. Par conséquent, la déviance est toujours positive et plus la déviance est petite plus le modèle s'ajuste bien. Asymptotiquement, D suit une loi du Khi-Deux à $n - (p + 1)$ degrés de liberté, notée $\chi^2(n - (p + 1))$, $(p + 1)$ correspondant au nombre de paramètres à estimer dans le modèle. Il est alors possible d'effectuer un test permettant d'accepter ou de rejeter le modèle, en choisissant d'accepter le modèle à un niveau de risque α si $D \leq q_{1-\alpha}$, où $q_{1-\alpha}$ est tel que $\mathbb{P}(\chi^2(n - (p + 1)) > q_{1-\alpha}) = \alpha$.

Résidus

Il existe différents types de résidus, les résidus les plus couramment utilisés pour valider un modèle sont :

— Les résidus (normalisés) de Pearson, définis par

$$\epsilon_i = \frac{Y_i - \widehat{Y}_i}{\sqrt{\widehat{Y}_i}},$$

où Y_i est la i$^{\text{ème}}$ observation et $\widehat{Y}_i$ l'estimation associée.

— Les résidus de déviance, définis par

$$r_i = \text{sign}(Y_i - \widehat{Y}_i) \times \sqrt{d_i},$$

où d_i est la contribution de l'observation i à la déviance, son expression dépendant de la loi choisie. En particulier, on a $D = \sum_{i=1}^{n} d_i$, ainsi la déviance est la somme des résidus de déviance au carré.

De même que pour la déviance, il existe un test utilisant les résidus de Pearson permettant de valider le modèle. La statistique de test est la somme des résidus de Pearson au carré, à savoir

$$\mathcal{X}^2 = \sum_{i=1}^{n} \epsilon_i^2.$$

Asymptotiquement cette statistique suit une Khi-Deux à $n-(p+1)$ degrés de liberté. Un test similaire à celui présenté précédemment pour la déviance peut être effectué en choisissant d'accepter le modèle à un niveau de risque α si $\mathcal{X}^2 \leq q_{1-\alpha}$, où $q_{1-\alpha}$ est tel que $\mathbb{P}(\chi^2(n-(p+1)) > q_{1-\alpha}) = \alpha$.

Si le modèle est adéquat, les résidus sont supposés indépendants, et gaussiens centrés réduits. Ainsi, en représentant graphiquement les résidus d'un modèle, le modèle est considéré comme adéquat si les résidus ne présentent pas de structure particulière (indépendance des résidus), s'ils sont symétriques par rapports à 0 (résidus centrés) et s'ils sont principalement compris entre -2 et 2 (résidus gaussiens[1]). Ces résultats peuvent être utilisés pour valider graphiquement un modèle.

2.1.3 Estimation des paramètres

L'estimation des paramètres du modèle se fait via la méthode du maximum de vraisemblance.
Dans le cadre du GLM, les observations étant supposées suivre une loi appartenant à la famille exponentielle, la log-vraisemblance est donnée par

$$\mathcal{L}(Y, \theta) = \sum_{i=1}^{n} \frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi),$$

[1]Pour rappel, le quantile à 97,5% d'une $\mathcal{N}(0,1)$ est égal à 1,96.

et l'estimateur du maximum de vraisemblance (souvent noté EMV) $\widehat{\beta}$ est obtenu en résolvant

$$\widehat{\beta} = \arg\max_{\beta} \mathcal{L}(Y, \theta).$$

L'EMV est solution de l'équation du premier ordre

$$\frac{\partial \mathcal{L}(Y, \theta)}{\partial \beta} = 0, \tag{2.2}$$

Les équations résultant de la résolution du problème (2.2) sont explicitées en annexe (B.1). Il n'existe pas de solution explicite en général, l'estimateur du maximum de vraisemblance est alors approché via des procédures d'optimisation itératives, telles que l'algorithme de Newton-Raphson.

Le modèle GLM permet également de calculer les intervalles de confiance associés à chaque estimation $\widehat{\beta}_j,\ j = 1, ..., p$.

Intervalles de confiance

Soit $\widehat{\beta}_n$ l'estimateur du maximum de vraisemblance du vrai paramètre β^* calculé sur n observations. La construction des intervalles de confiance associés aux estimateurs est basée sur le résultat suivant

$$(\widehat{\beta}_n - \beta^*) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \mathcal{I}_n(\widehat{\beta}_n)^{-1}),$$

où $\mathcal{I}_n(\widehat{\beta}_n) = -\mathbb{E}[\nabla^2 \mathcal{L}(\widehat{\beta}_n)]$ est l'information de Fisher des n observations. Ce résultat permet de construire pour chacun des paramètres un intervalle de confiance de niveau asymptotique $1 - \alpha$ de la forme

$$\forall j \in \{1, ..., p\}, \quad I_n(\beta_j^*) = \left[\widehat{\beta}_{j,n} - q_{1-\frac{\alpha}{2}} \times \widehat{s}_j \ ;\ \widehat{\beta}_{j,n} + q_{1-\frac{\alpha}{2}} \times \widehat{s}_j\right].$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile de niveau $1 - \frac{\alpha}{2}$ d'une $\mathcal{N}(0, 1)$ et $\widehat{s}_j^2 = \mathcal{I}_n(\widehat{\beta}_n)^{-1}_{j,j}$ est la variance de $\widehat{\beta}_{j,n}$. Les intervalles de confiance permettent de quantifier l'incertitude associée à chaque estimation.

Le modèle GLM permet également une lecture aisée de l'influence des variables explicatives sur la prédiction.

2.1.4 Influence d'une variable dans un modèle GLM

La façon dont une variable influe sur la prédiction dépend de la fonction de lien choisie. Pour prédire des coûts automobiles, la fonction de lien log est généralement utilisée. En effet, elle permet de conduire à un modèle multiplicatif. Dans ce cas le modèle s'écrit

$$\begin{aligned} &\log(\mathbb{E}[Y|X]) = X\beta \\ &\Rightarrow \mathbb{E}[Y|X] = e^{X\beta} = e^{\beta_0} \times e^{\beta_1 X_1} \times ... \times e^{\beta_p X_p}. \end{aligned}$$

Ce modèle est utilisé pour la prédiction de coûts automobiles car il s'adapte mieux à ce cadre. En effet, considérons par exemple un contrat automobile, et supposons que le véhicule assuré soit remplacé par un véhicule dont le kilométrage est plus faible, toutes choses restant égales par ailleurs. Il est alors plus intuitif de modifier la prime de ce contrat en multipliant l'ancienne prime par un pourcentage plutôt que d'ajouter un montant fixe à cette ancienne prime [Thomas (2014)].

Dans le cadre de ce modèle l'étude de l'influence d'une variable est simple :

— Pour une variable qualitative
Si elle possède k modalités, le modèle GLM retourne pour cette variable $k-1$ coefficients. Ces coefficients décrivent le comportement par rapport à une modalité de référence (la $k^{\text{ème}}$ modalité). Pour comprendre comment s'interprètent ces coefficients, considérons par exemple un modèle GLM avec pour fonction de lien la fonction log visant à expliquer le coût des sinistres et une variable explicative nommée `VAR` possédant trois modalités notées A, B et C. Supposons que les coefficients suivants ont été retournés pour cette variable :

Variable	**Coefficients**
`VAR_B`	0.0234
`VAR_C`	-0.0451

Table 2.1 : Exemple de coefficients fournis par un GLM pour la variable `VAR`

Seulement deux coefficients sont retournés par le GLM, la modalité A étant la modalité de référence. Soit un individu faisant parti de la catégorie A et soit `coutA` le coût renvoyé par le GLM pour cet individu. Soient `coutB` et `coutC` les coûts renvoyés par le GLM pour des individus appartenant respectivement aux catégories B et C, toutes choses étant égales par ailleurs. Les coefficients du tableau (2.1) ci-dessus s'interprètent de la façon suivante

$$coutB = e^{0.0234} \times coutA = 1.0234 \times coutA \quad \text{et} \quad coutC = e^{-0.0451} \times coutA = 0.9559 \times coutA.$$

De plus, pour x assez proche de 0, $\exp(x) \simeq 1+x$. Ainsi les coefficients peuvent directement être lus de la façon suivante : appartenir à la catégorie B plutôt qu'à la catégorie A fait augmenter de le coût d'environ 2.3% et appartenir à la catégorie C plutôt qu'à la catégorie A fait baisser le côut d'environ 4.5%.

— Pour une variable quantitative
L'interprétation est également simple dans ce cas. Reprenons l'exemple du GLM précédent et considérons qu'il contient une variable explicative quantitative représentant le kilométrage du véhicule sinistré, appelée `km`. Supposons que le coefficient suivant ait été retourné pour cette variable :

Variable	**Coefficients**
`km`	-0.0134

Table 2.2 : Exemple de coefficient fourni par un GLM pour la variable `km`

L'interprétation de ce coefficient est immédiate : augmenter le kilométrage de 1 unité, toutes choses restant égales par ailleurs, revient à multiplier le coût initial par $e^{-0.0134} = 0.9867$, ou encore à diminuer le coût d'environ 1.3%.

Les exemples précédents permettent de voir pourquoi l'un des atouts des modèles linéaires généralisés est la facilité de lecture et d'interprétation de l'influence des variables sur les prédictions. Ces exemples permettent aussi de voir que pour une variable qualitative possédant m modalités le GLM calcule $m-1$ coefficients. Afin de limiter la complexité du modèle et d'éviter la redondance d'information, une étude des liens entre variables de la base doit être effectuée.

2.2 Étude des liens entre variables

Lors de la construction d'un modèle statistique une étape importante est la sélection des variables explicatives. En effet, il est préférable de ne conserver que les variables pertinentes pour la prédiction du modèle. La pertinence s'entend ici comme la capacité d'une variable à apporter de l'information sur la variable d'intérêt, sans pour autant apporter trop de redondance d'information. Ainsi il est souhaitable de conserver des variables explicatives liées à la variable d'intérêt qui soient le moins corrélées possible entre elles.

2.2.1 Lien avec la variable d'intérêt

L'intérêt d'étudier ce lien est de pouvoir identifier et supprimer les variables ne représentant pas d'intérêt pour la prédiction de la variable d'intérêt. En effet, conserver des variables non liées à cette dernière ne ferait qu'augmenter la complexité du modèle. Pour identifier ces variables les outils généralement utilisées sont :

— Le coefficient de corrélation
Ce coefficient permet d'étudier le lien entre deux variables quantitatives, sa formule est

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}},$$

Il est compris entre -1 et 1 : lorsque ce coefficient est proche de -1 les variables ont une relation linéaire négative, lorsque ce coefficient est proche de 1 les variables ont une relation linéaire positive, enfin lorsque ce coefficient est proche de 0 alors la relation linéaire entre les variables est faible. Attention cependant, une corrélation nulle entre deux variables n'implique pas forcément l'indépendance entre elles, de même un coefficient de corrélation proche de 1 en valeur absolue n'implique pas forcément une relation de causalité entre les deux variables.

— Le V de Cramer
Cette statistique permet d'étudier le lien entre les variables qualitatives, sa formule est la suivante

$$V\ de\ Cramer = \sqrt{\frac{\chi^2}{n \times (\min(m,r)-1)}}.$$

où :

— m est le nombre de modalités de la variable n°1,
— r est le nombre de modalités de la variable n°2,
— n est le nombre d'observations dans la base de données,
— χ^2 est la statistique du Chi-Deux à $\min(m,r)-1$ degrés de liberté.

Le V de Cramer est compris entre 0 et 1 et plus il est proche de 1 plus le lien entre les deux variables étudiées est fort.

— Le test d'indépendance du Chi-Deux
L'hypothèse nulle de ce test statistique est H_0 : "Les variables sont indépendantes". Sous cette hypothèse la statistique de test suit une loi du χ^2. Pour décider du rejet ou non de l'hypothèse nulle un niveau de confiance α est choisi et si la p-valeur du test est inférieure à α alors H_0 est rejetée, sinon elle ne l'est pas (cependant cela n'implique par nécessairement que H_0 est acceptée).

Au vu de la nature qualitative de l'ensemble des variables explicatives dans la base d'étude il est choisi d'utiliser le V de Cramer pour mesurer le lien entre les variables. Pour ce faire, la variable représentant le coût des sinistres est discrétisée en cinq catégories basées sur ses quantiles. Le calcul du V de Cramer et le test d'indépendance du Chi-Deux au niveau $\alpha = 5\%$ ont été effectués. Les résultats sont résumés dans le tableau (2.3) ci-dessous :

Variable	**V de Cramer**	**p-valeur du test du χ^2**
`tauxHoraire`	0.20410	<2.2e-16
`TS`	0.10730	<2.2e-16
`GroupeAge`	0.07336	<2.2e-16
`CategorieMarque`	0.05987	<2.2e-16
`GroupeKm`	0.04505	<2.2e-16
`assureur`	0.03111	<2.2e-16
`mois`	0.02191	<2.2e-16
`trimestre`	0.01890	<2.2e-16
`expertA`	0.00488	0.001173

TABLE 2.3 : V de cramer et test d'indépendance du Chi-Deux entre les variables explicatives et la variable d'intérêt

Une première information donnée par ces résultats est que toutes les p-valeurs du test d'indépendance du Chi-Deux sont inférieures à 5%, ainsi l'hypothèse d'indépendance peut être rejetée : toutes les variables explicatives semblent bien avoir un lien avec le coût du sinistre. De plus, les V de Cramer permettent de quantifier la puissance du lien entre ces variables et la variable d'intérêt, en soulignant notamment l'importance de la variable correspondant au taux horaire qui possède le V de Cramer le plus élevé avec le coût des sinistres.
Concernant la variable `expertA`, dont on cherche à mesurer l'influence, elle possède le V de Cramer le plus faible parmi toutes les variables explicatives, cela n'implique pas pour autant une indépendance entre cette variable et le coût du sinistre mais seulement un lien plus faible que celui avec les autres variables.

Ainsi, toutes les variables explicatives sont bien liées au coût du sinistre mais il se peut que certaines soient dépendantes entre elles. Pour identifier ces potentielles variables, une étude du lien entre les variables explicatives est effectuée.

2.2.2 Lien entre les variables explicatives

Il n'est pas souhaitable d'intégrer de la redondance d'information dans un modèle : conserver deux variables apportant les mêmes informations ne présente pas d'intérêt pour la prédiction et ne ferait

qu'augmenter la complexité du modèle. Le V de Cramer ou bien le coefficient de corrélation, en fonction de la nature des variables, peuvent être utilisés pour déterminer le niveau de lien entre deux variables explicatives. Un niveau à partir duquel il est considéré que deux variables sont "trop" corrélées est choisi et l'une des deux variables des couples "trop" corrélés est supprimée. Par exemple, pour le V de Cramer, on considère en général que le lien est intéressant lorsqu'il est supérieur à 0.1, fort lorsqu'il est supérieur à 0.4 et robuste lorsqu'il est supérieur à 0.7 [Bouret (2014)]. Cette étape permet également de vérifier la fiabilité des données, par exemple dans l'étude présentée dans ce mémoire si aucun lien n'est observé entre les variables `TRIMESTRE` et `Mois`, cela montrera qu'il existe très certainement une erreur dans la base de données.

Les V de Cramer entre les variables explicatives sont représentés sur la figure (2.1) ci-dessous :

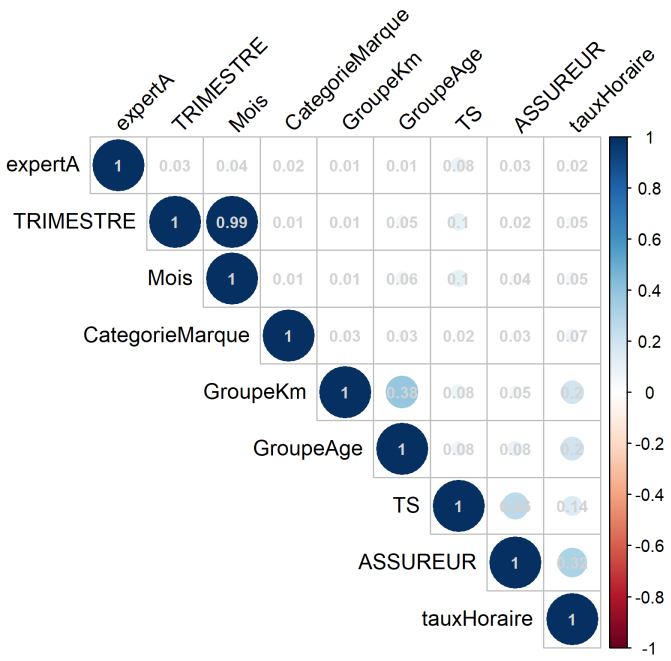


Figure 2.1 : V de Cramer entre les variables explicatives

Cette représentation permet d'observer un lien très fort entre les variables `Mois` et `TRIMESTRE` comme il pouvait être attendu. Le lien entre l'âge du véhicule et son kilométrage peut être également souligné et s'explique assez naturellement par le fait que plus un véhicule est âgé plus il a, en général, parcouru de kilomètres.
Concernant les variables représentant le mois et le trimestre de la survenance du sinistre, leur V de Cramer étant très élevé, il semblerait logique de supprimer l'une de ces deux variables de l'étude. Cependant, dans le cadre d'une approche prospective, l'intégration de variables temporelles de la sorte dans les modèles n'est pas pertinente. Ainsi les variables `Mois` et `TRIMESTRE` ne seront pas incluses parmi les variables explicatives des différents modèles de prédiction. Cependant, elles ne sont pas

supprimées de la base et elles permettront en particulier une étude de l'évolution à travers le temps de l'influence du réseau d'experts. Il est choisi de conserver l'ensemble des autres variables explicatives, les liens observés n'étant pas suffisamment élevés pour justifier la suppression de l'une d'entre elles. Cependant, il est possible que l'intégration de certaines variables ne soit pas pertinente pour le modèle linéaire généralisé. Une sélection de variables spécifique au GLM est alors effectuée.

2.3 Sélection des variables explicatives dans une régression

L'étude des liens entre les variables permet d'identifier les variables qui semblent pertinentes pour l'étude et celles qui sont à supprimer. Cette étape est suivie d'une sélection de variables pour le modèle utilisé, à savoir ici le modèle linéaire généralisé.

2.3.1 Méthode de sélection de variables

Pour sélectionner les variables explicatives les plus pertinentes, il existe trois méthodes principales :

— La sélection *forward*
Cette méthode part d'un modèle contenant seulement l'intercept. Les variables sont ajoutées une à une en choisissant à chaque étape la variable la plus optimale selon un critère défini en amont. Une fois qu'une variable est ajoutée, elle ne peut plus être retirée du modèle. La sélection s'arrête lorsqu'une règle d'arrêt est atteinte ou lorsqu'il n'y a plus de variable à ajouter dans le modèle.

— La sélection *backward*
Cette méthode part d'un modèle contenant toutes les variables explicatives. Les variables sont supprimées une par une en supprimant à chaque étape la variable la moins significative, selon un critère pré-défini pour le modèle de régression. Une fois une variable supprimée, elle ne peut plus être réintégrée au modèle. La sélection s'arrête lorsqu'une règle d'arrêt est atteinte ou lorsqu'il ne reste plus que l'intercept dans le modèle.

— La sélection *stepwise*
Cette approche est une combinaison des deux précédentes. Elle part soit du modèle contenant seulement l'intercept, il s'agit alors d'une sélection *stepwise forward*, soit du modèle contenant toutes les variables explicatives et il s'agit alors d'une sélection *stepwise backward*. À chaque étape, une variable explicative est supprimée ou ajoutée afin d'optimiser le modèle selon un critère pré-défini. La sélection s'arrête lorsqu'une règle d'arrêt est atteinte ou lorsqu'il n'y a plus de variable à ajouter ou supprimer dans le modèle.

Le critère choisi est généralement le critère d'information d'Akaike, souvent appelé AIC (pour *Akaike Information Criterion* en anglais), défini comme suit :

$$AIC = 2(p - \log(\tilde{L})).$$

où p est le nombre de paramètres du modèle et $\tilde{L}$ est la vraisemblance maximisée. Ce critère pénalise la log-vraisemblance du modèle par le nombre de paramètres retenus. Le meilleur modèle est celui qui présente l'AIC le plus faible, c'est le modèle qui réalise le meilleur compromis entre qualité d'ajustement et complexité. Il est également possible d'utiliser d'autres critères comme l'AIC_c, dont l'utilisation est

recommandée lorsque le nombre de paramètres est grand par rapport au nombre d'observations, ou bien le BIC, qui pénalise plus fortement le nombre de paramètres présents dans le modèle que l'AIC.

2.4 Application du modèle linéaire généralisé

Afin d'appliquer un modèle généralisé, il faut spécifier une loi et une fonction de lien. Lors de l'étude de coûts de sinistres automobiles, la loi Gamma est usuellement utilisée. Cette loi appartient bien à la famille exponentielle puisque sa densité s'écrit

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x),$$

qui peut être réécrit sous la forme

$$\begin{aligned} f(x;\alpha,\beta) &= \exp(-\beta x + (\alpha-1)\log(x) + \alpha\log(\beta) - \log(\Gamma(\alpha)) \\ &= \exp(\frac{x\theta - b(\theta)}{a(\phi)} + c(y,\phi)), \end{aligned}$$

avec $\theta = \frac{\beta}{\alpha},\ \phi = \frac{-1}{\alpha},\ a(\phi) = \phi,\ b(\theta) = \log(\theta)$ et $c(x,\phi) = (\alpha-1)\log(x) - \log(\Gamma(\alpha)) - \alpha\log(\alpha)$.

La fonction de lien canonique associée à la loi Gamma est telle que $g^{-1}(\theta) = b'(\theta) = \frac{1}{\theta}$, *i.e.* g est la fonction inverse : $g(\theta) = \frac{1}{\theta}$. Mais dans le cadre de ce mémoire la fonction de lien log sera utilisée puisqu'elle permet de conduire à un modèle multiplicatif.

Une sélection de variables a été effectuée afin de déterminer celles qui seront conservées dans le modèle linéaire généralisé. Les quatre méthodes précédemment décrites ont été appliquées avec pour critère l'AIC et toutes ont conduit au même résultat : c'est le modèle contenant l'ensemble des variables explicatives qui est le plus adapté. Ainsi le modèle sera appliqué avec les sept variables explicatives suivantes : le réseau d'experts intervenu, le type de sinistre, le taux horaire, l'assureur et les groupes d'âge, de marque et de kilométrage du véhicule.

2.4.1 Adéquation du modèle

L'analyse graphique des résidus permet de visualiser la bonne adéquation ou non du modèle aux données. Dans un modèle linéaire généralisé avec une loi Gamma, les résidus de déviance sont obtenus avec la formule (2.3)

$$\forall i \in \mathbb{N},\ r_i = \sqrt{2\times[\frac{Y_i - \widehat{Y}_i}{\widehat{Y}_i} - \log(\frac{Y_i}{\widehat{Y}_i})]}\times sign(Y_i - \widehat{Y}_i). \tag{2.3}$$

où $\widehat{Y}_i$ est la valeur prédite pour l'observation i et Y_i est la valeur réellement observée.

Une représentation graphique de ces résidus et de leur densité est donnée par la figure (2.2) ci-dessous. Le premier graphique représente les résidus de déviance qui ont été calculés pour chacune des observations de la base. Les résidus ne présentent pas de structure particulière et sont principalement compris entre -2 et 2. Plus précisément le pourcentage de résidus non compris entre -2 et 2 (ici cela ne concerne que les résidus supérieur à 2) est de 0.78%. Le second graphique présente la densité associée

aux résidus et permet d'observer que celle-ci est légèrement décalée à gauche par rapport à l'origine, ce qui indique que le modèle a tendance à plus surestimer les coûts qu'à les sous-estimer.

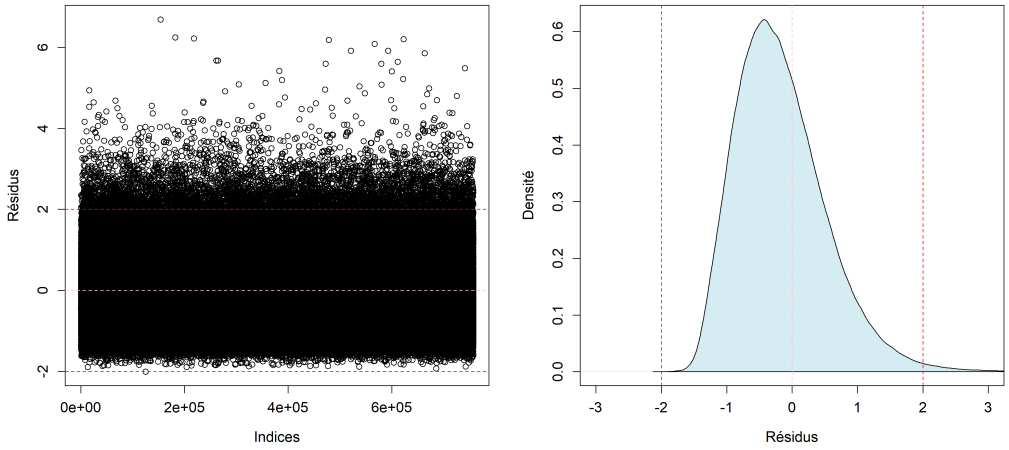


FIGURE 2.2 : Représentation des résidus et de leur densité

Malgré quelques irrégularités l'analyse graphique des résidus permet de conclure à la bonne adéquation du modèle. Afin de confirmer ce résultat le test sur la déviance présenté précédemment a été réalisé. La déviance du modèle est égale à 237 610, cette valeur est comparée au quantile de niveau 95% d'une $\chi^2(757\,633 - 7 - 1)$, et on a $D = 237\,610 < 764\,830 = q_{\chi^2,95\%}$. Ainsi le modèle est accepté à un niveau de risque égal à $\alpha = 5\%$.
Il est alors possible d'appliquer le modèle linéaire généralisé pour effectuer des prédictions.

2.4.2 Validation croisée

Avant d'utiliser le modèle GLM pour effectuer les prédictions sur l'échantillon de validation, une validation croisée est effectuée afin de vérifier que le modèle ne fait pas du surapprentissage. La validation croisée consiste à découper l'échantillon d'apprentissage en K blocs distincts de même taille, puis à sélectionner un des blocs comme échantillon de test et utiliser les blocs restants comme échantillon d'apprentissage. La procédure est répétée K fois de telle sorte que chacun des blocs soit utilisé une fois en échantillon de test. Il en ressort K évaluations des métriques sur différents échantillons de test, avec un modèle entraîné sur différents échantillons d'apprentissage. Le principe de la validation croisée est résumé par la figure (2.3) ci-dessous.
Si les métriques varient fortement d'un bloc à l'autre cela montre que le modèle n'a pas réussi à généraliser une règle de décision. Une validation croisée avec 5 blocs a été effectuée sur l'échantillon d'apprentissage, les résultats obtenus sont présentés dans le tableau (2.4) ci-dessous :

Bloc	**MSE**	**MAE**
1	4 560 959	1 284
2	4 450 270	1 278
3	4 615 580	1 296
4	4 532 371	1 286
5	4 614 634	1 291

TABLE 2.4 : Résultats de la validation croisée pour le GLM

La colonne “Bloc” fait référence au numéro du bloc utilisé comme échantillon de test. La MSE et la MAE restent globalement du même ordre de grandeur d'un bloc à l'autre. Ainsi, le modèle GLM a réussi à généraliser une règle de décision et ne fait pas de surapprentissage. Il peut alors être entraîné sur l'ensemble de l'échantillon d'apprentissage puis appliqué à l'échantillon de validation pour évaluer son pouvoir prédictif.

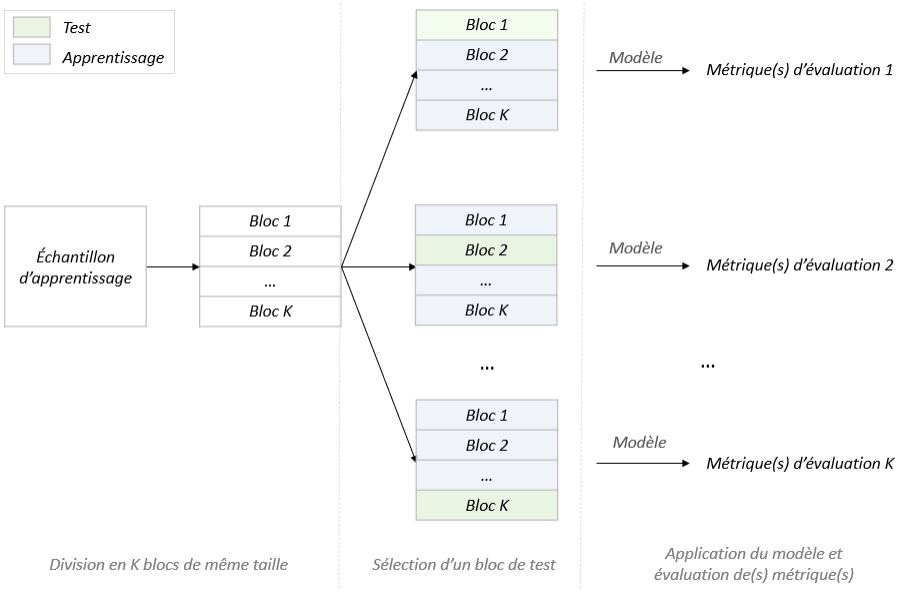


FIGURE 2.3 : Schéma simplifié du principe de la validation croisée

2.4.3 Résultats du GLM

Pouvoir prédictif du modèle

Dans le cadre du GLM la division entre l'échantillon de validation et de test n'est pas réellement nécessaire puisqu'aucune optimisation de paramètres n'est réalisée. Cependant, il est choisi d'utiliser la même décomposition en trois échantillons que pour les autres modèles. Si ce n'était pas le cas, la comparaison entre les performances des différents modèles serait biaisée. Le modèle linéaire généralisé est donc entraîné sur l'échantillon d'apprentissage et les performances sont évaluées sur l'échantillon de validation.
Afin de souligner l'amélioration fournie par le GLM, les performances du modèle sont comparées à celles obtenues avec un tarif mutualisé, *i.e.* un modèle associant comme prédiction à chaque observation la moyenne des coûts de la base d'apprentissage, à savoir 2 294€. Les résultats des métriques obtenus pour le modèle GLM et le tarif mutualisé sont résumés dans le tableau (2.5) ci-dessous :

	MSE	MAE
Tarif mutualisé	6 588 897	1 560
GLM	4 581 209	1 293

TABLE 2.5 : Résultats de prédiction sur l'échantillon de validation pour un tarif mutualisé et le GLM

Le modèle GLM permet une amélioration de plus de 30% de la MSE et de plus de 17% de la MAE par rapport à la prédiction simple qui consisterait à prédire la moyenne des coûts de la base

d'apprentissage pour tous les individus. Les résultats obtenus avec le GLM serviront de référentiels par la suite pour évaluer l'amélioration du pouvoir prédictif fournie par les modèles présentés en chapitre 3.

Bien que le modèle linéaire généralisé permette une amélioration des métriques par rapport au tarif mutualisé, il commet encore des erreurs. Afin de mieux comprendre d'où proviennent ces erreurs, il pourrait être intéressant de comparer les caractéristiques des vecteurs de sinistres observés et prédits. Cependant, les sinistres observés sont la réalisation d'une variable aléatoire, tandis que les prédictions sont obtenues via le calcul d'une espérance de cette même variable aléatoire. L'étude comparative des deux vecteurs ne serait donc pas rigoureuse, puisqu'elle consisterait à comparer deux objets de nature différente. Afin d'effectuer une comparaison entre deux éléments de même type, il est choisi de s'intéresser aux différences entre les moyennes des sinistres observés et prédits. Tout d'abord, la moyenne des coûts au sein de la base de validation est égale à 2 296€, tandis que la moyenne des prédictions issues du GLM est égale à environ 2 259€. À noter que cela n'est pas en contradiction avec ce qui avait été observé sur la densité des résidus en figure (2.2) *supra*. En effet, le léger décalage de la densité sur la gauche par rapport à l'origine montre que le modèle a tendance à plus surestimer les coûts qu'à les sous-estimer, cependant cela n'implique par forcément que la moyenne des coûts prédits soit supérieure à celle des coûts observés. En effet, il suffit par exemple que les sous-estimations induites par le modèle soit plus grandes que les surestimations effectuées par ce dernier.
Afin de mieux observer d'où vient la différence entre les moyennes des coûts prédits et observés, une représentation graphique de ces moyennes au sein de chacune des modalités des variables explicatives a été effectuée. Les résultats sont présentés en figure (2.4) ci-dessous :

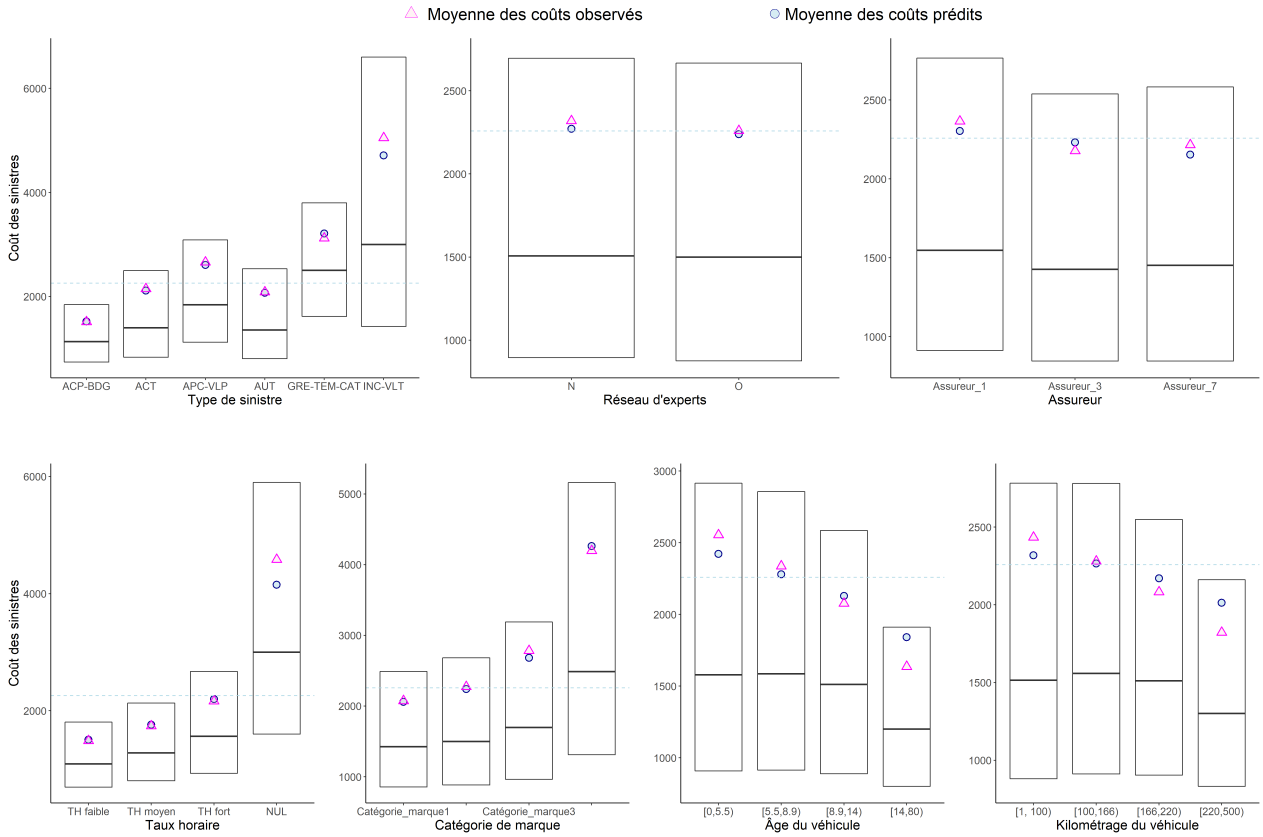


FIGURE 2.4 : Moyennes des sinistres observés et prédits par le GLM pour chaque modalité des variables explicatives

Pour chaque variable explicative de la base, les médianes, premiers et troisièmes quartiles ont également été représentés sous la forme de *box-plots* pour chacune des modalités. Sur chacun des graphiques de la figure (2.4) ci-dessus la moyenne des coûts prédits par le modèle GLM (environ 2 259€) est représentée par une ligne pointillée bleue.
L'analyse de ces graphiques permet d'observer que pour certaines variables explicatives il existe une ou plusieurs modalités pour lesquelles les écarts entre les moyennes des sinistres observés et prédits sont plus importants que pour les autres modalités de la variable. Par exemple, pour les sinistres incendies ou vols totaux (`INC-VLT`), l'écart entre les moyennes des sinistres observés et prédits est plus important que pour les autres types de sinistres. Ces sinistres sont ceux présentant les coûts les plus élevés en moyenne, comme il peut être vu sur la figure (2.4) ci-dessus. Un écart du même type peut être observé pour le cas où le véhicule était économiquement irréparable, *i.e.* lorsque `tauxHoraire = NUL`, tandis que pour les autres taux horaires les écarts de moyennes sont relativement faibles.
Il existe également des modalités pour lesquelles l'écart entre les moyennes est inverse à celui observé sur l'ensemble des observations, *i.e.* des modalités pour lesquelles la moyenne des coûts prédits est supérieure à celle des coûts observés. C'est notamment le cas pour les variables représentant les catégories d'âge et de kilométrage du véhicule sinistré : les coûts sont surestimés en moyenne par le modèle pour les véhicules d'âges supérieurs à 8.9 ou bien possédant un kilométrage supérieur à 165 601 kilomètres.

Au vu de l'objectif de l'étude, afin d'évaluer le modèle GLM un autre aspect à prendre en compte est la lisibilité de l'influence des variables sur la prédiction.

2.4.4 Mesure de l'influence du réseau d'expert

Le résumé des coefficients fournis par le GLM est donné en annexe (B.2). En particulier, le coefficient associé à la variable spécifiant l'utilisation du réseau d'experts A est égal à -0.0287283. Cela signifie qu'avoir recours au réseau d'experts A permet de faire baisser le coût d'environ 2.9% en moyenne. L'intervalle de confiance associé à cette estimation est [-0.0326 ; -0.0249].

Il est possible d'analyser l'évolution de ce coefficient à travers le temps grâce aux variables `TRIMESTRE` et `Mois`. Une analyse de ce coefficient par trimestre est donnée par la figure (2.5) ci-dessous. Chaque point de la figure (2.5) correspond au coefficient associé au réseau d'experts, calculé sur l'ensemble des sinistres survenu durant le trimestre spécifié en abscisse. Les intervalles de confiance associés à ces estimations ont également été représentés. La zone rosée du graphique représente le cas où le coefficient est positif, ce qui correspond à la situation où le réseau d'experts B a été plus performant que le réseau d'experts A en moyenne, à l'inverse de la zone bleutée qui représente la situation où le réseau d'experts A a été le plus performant en moyenne. Il est possible d'observer que pour l'ensemble des trimestres le coefficient est négatif, bien qu'il existe une incertitude pour les deux derniers trimestres de 2020.

Il est alors intéressant de zoomer cette maille temporelle et d'observer l'évolution de ce coefficient par mois. La figure (2.6) ci-dessous représente l'évolution mensuelle du coefficient associé au réseau d'experts. Cette représentation permet d'observer qu'il existe des mois (août et octobre 2020) au cours desquels le réseau d'experts B a été plus performant en moyenne, bien qu'il existe une incertitude sur le signe des estimations. Un intervalle de confiance plus large que les autres peut être observé pour le mois d'avril 2020. Cela s'explique par le nombre plus faible d'observations associées à ce mois, comme il a été vu sur la figure (1.19) *supra*.

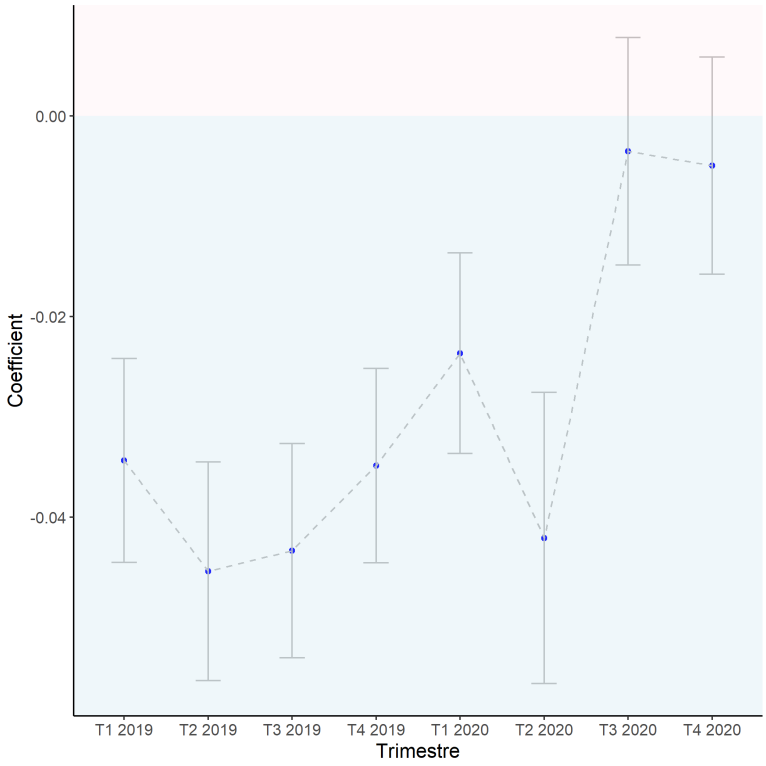


FIGURE 2.5 : Évolution de l'influence du réseau d'experts par trimestre

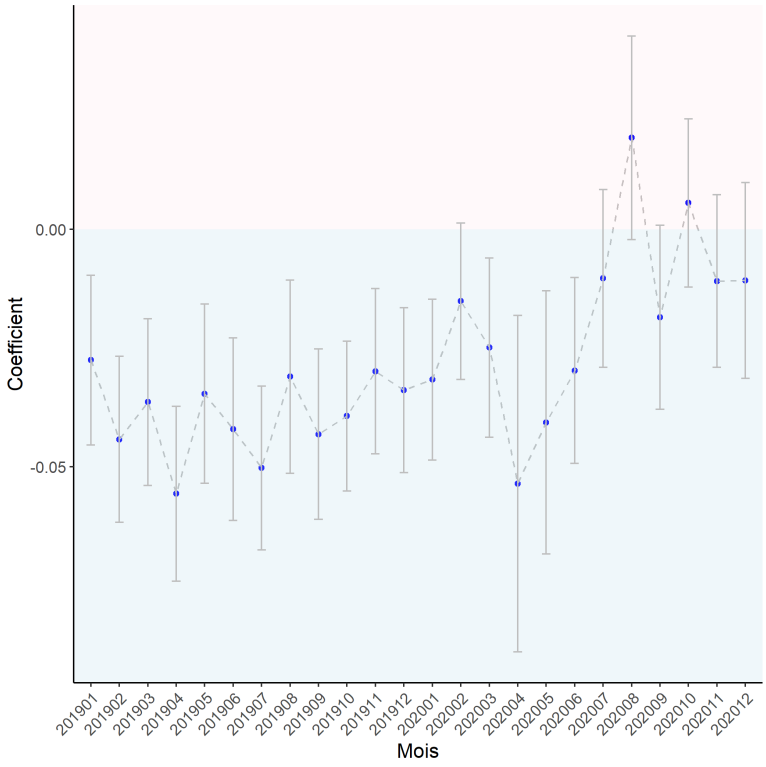


FIGURE 2.6 : Évolution de l'influence du réseau d'experts par mois

Le modèle linéaire généralisé permet donc une lecture aisée et compréhensible de l'influence des réseaux d'experts sur la prédiction. Cependant, son pouvoir prédictif présente quelques limites. Pour tenter de palier à ces défauts, des modèles d'arbres de régression vont être appliqués. Ces modèles sont présentés dans le chapitre suivant.

Chapitre 3

Arbres de régression

Les arbres de décision désignent un ensemble de méthodes permettant de prédire une variable (qualitative ou quantitative) à l'aide d'autres variables via la construction d'un ou plusieurs arbres. Lorsque la variable à prédire est quantitative ils sont appelés arbres de régression. Les méthodes d'arbres de régression figurent parmi les méthodes d'apprentissage supervisé les plus populaires. Le premier algorithme d'arbre de régression date de plus de 50 ans, il s'agit de l'algorithme AID (pour *Automatic Interaction Detection* en anglais) développé par Morgan & Sonquist en 1963. Le principe de ces méthodes est simple et intuitif, de plus elles offrent une lisibilité aisée des résultats. Ce chapitre présente deux modèles d'arbre de régression : tout d'abord le modèle CART (pour *Classification and Regression Tree* en anglais), puis le modèle MOB (pour *MOdel Based recursive partitionning* en anglais). Les pouvoirs prédictifs de ces modèles seront confrontés à celui du modèle XGBoost, modèle réputé pour ses performances de prédiction. Ce dernier sera également analysé pour permettre une amélioration du modèle MOB.

3.1 Algorithme CART

L'algorithme CART ([Breiman et al. (1984)], [Genuer et Poggi (2017)]) est une méthode permettant la construction d'un arbre de décision. Il a été introduit en 1984 par Breiman, Friedman, Olshen et Stone. L'idée générale est de regrouper l'ensemble des individus présents dans la base de données dans un premier nœud, appelé racine, puis d'effectuer une première division binaire selon un critère pré-défini. Cette première division entraîne la création de deux nœuds fils et la procédure est répétée dans chacun des nœuds construits jusqu'à ce qu'un critère d'arrêt pré-défini soit atteint. Un grand arbre, appelé arbre maximal, est alors construit. Dans un second temps l'élagage de cet arbre est effectué afin d'obtenir l'arbre optimal selon un critère pré-défini. À chaque nœud terminal de cet arbre est associée une valeur numérique, si la variable à prédire est quantitative, ou une modalité de cette dernière, si elle est qualitative. La description de ces différentes étapes ainsi que les résultats obtenus via l'application de l'algorithme CART sont explicités dans cette première section.

3.1.1 Construction d'un arbre binaire maximal

Soit Y la variable à expliquer et $X_1, ..., X_p$ les variables explicatives qui peuvent être qualitatives ou quantitatives. L'algorithme part de la racine qui contient l'ensemble des individus, puis construit une série de nœuds permettant à chaque étape le partitionnement binaire des individus en groupes les plus

homogènes possible.

Pour effectuer une division il faut tout d'abord choisir la variable explicative sur laquelle la division va s'effectuer (appelée *split variable* en anglais). Si cette dernière est quantitative alors le critère de division est de la forme $X_j \leq \alpha$, où X_j est la variable de division sélectionnée et α est une valeur choisie par l'algorithme, sinon le critère de division est de la forme $X_j \in A$, où A représente un sous-ensemble des modalités de la variable X_j. La division conduit alors à la création de deux nœuds fils sur lesquels la même procédure est appliquée jusqu'à l'atteinte d'un critère d'arrêt pré-défini.

Critère de division

Lorsqu'une division est effectuée l'objectif est de séparer les individus en deux groupes qui soient le plus homogènes possible au sens de la variable à expliquer Y. Le critère de division se base sur une fonction d'hétérogénéité. Cette fonction doit en particulier être nulle si et seulement si le nœud est homogène, *i.e.* si tous les individus du nœud possèdent la même valeur pour la variable Y. L'algorithme recherche la meilleure division parmi l'ensemble des divisions admissibles, *i.e.* ne créant pas de nœud fils vide. Lorsque la variable de séparation choisie est qualitative avec m modalités, il existe $2^{m-1}-1$ divisions binaires admissibles, si, de plus, elle est ordinale alors ce nombre est réduit à $(m-1)$ divisions binaires admissibles. Enfin, pour une variable quantitative, ce cas est similaire à une variable qualitative ordinale où les modalités seraient alors l'ensemble des valeurs prises par la variable quantitative.

Soient h un nœud et h_g et h_d les deux nœuds fils (respectivement gauche et droit) issus de sa division. Pour effectuer cette division l'ensemble des divisions admissibles est étudié et l'algorithme choisit celle qui minimise la somme des hétérogénéités de chacun des nœuds fils, ou encore celle qui résout

$$\max_{j=1,\dots,p} \{ \max_{divisions\ de\ X_j} D_h - (p_{h_g} D_{h_g} + p_{h_d} D_{h_d})\}, \tag{3.1}$$

où $p_{h_g} = \frac{|h_g|}{|h|}$ et $p_{h_d} = \frac{|h_d|}{|h|}$ avec $|h|$ le cardinal du nœud h. Ainsi p_{h_g} et p_{h_d} sont respectivement les proportions d'individus envoyés dans les nœuds gauche et droit lors de la division du nœud h. D_h représente l'hétérogénéité au sein du nœud h. Pour le calcul de cette dernière deux cas sont à distinguer :

— Si Y est quantitative
Dans ce cas l'hétérogénéité du nœud est définie par sa variance

$$D_h = \frac{1}{|h|} \sum_{y_i \in h} (y_i - \overline{y}_h)^2,$$

où $\overline{y}_h$ est la moyenne de la variable Y pour les individus présents dans le nœud h. L'optimisation du critère (3.1) revient à minimiser la variance intra-classes, *i.e.* minimiser

$$\frac{1}{|h|} \sum_{y_i \in h_g} (y_i - \overline{y}_{h_g})^2 + \frac{1}{|h|} \sum_{y_i \in h_d} (y_i - \overline{y}_{h_d})^2.$$

— Si Y est qualitative
Soit $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_m\}$ l'ensemble de des modalités de la variable Y. Plusieurs critères peuvent être choisis pour la division, les plus communs sont :

— Critère de l'entropie

$$D_h = -\sum_{l=1}^{m} p_h^l log(p_h^l),$$

où p_h^l est la proportion de la modalité $\mathcal{C}_l$ dans le nœud h, avec la convention $0\log(0) = 0$.

— Critère de Gini

$$D_h = \sum_{l=1}^{m} p_h^l(1 - p_h^l).$$

Dans tous les cas l'algorithme choisit la division qui maximise la réduction d'impureté entre le nœud père h et ses deux nœuds fils h_g et h_d, notée $\Delta i_r(h) = D_h - (p_{h_g} D_{h_g} + p_{h_d} D_{h_d})$ où h_g et h_d sont les nœuds fils de h construits selon la règle de décision r.

Critère d'arrêt et affectation

La procédure décrite ci-dessus est arrêtée dans un nœud lorsque celui-ci est homogène ou lorsqu'il n'existe plus de partition admissible à créer à partir de ce nœud, ou encore si le nombre d'individus dans ce nœud est inférieur à une valeur seuil, choisie en général entre 1 et 5. Le nœud devient alors un nœud terminal, également appelé feuille. Une valeur est affectée à chaque nœud terminal : si Y est quantitative la valeur affectée est la moyenne de Y pour les individus présents dans le nœud, sinon c'est la modalité la plus représentée dans le nœud.

3.1.2 Élagage

La procédure décrite dans la section précédente permet la construction d'un arbre, appelé arbre maximal, noté T_{max} et possédant K feuilles. Du fait de sa grande taille, cet arbre est très dépendant de l'échantillon avec lequel il a été construit. L'utilisation de cet arbre pour effectuer des prédictions risque de conduire au phénomène de surapprentissage. Pour éviter cette situation un élagage (*pruning* en anglais) de l'arbre maximal est effectué, afin de trouver un arbre optimal. L'algorithme va, en élaguant, construire une séquence de sous-arbres emboîtés en partant de l'arbre maximal. À noter que d'autres arbres candidats pourraient être étudiés mais dans ce cas la recherche de l'arbre optimal serait trop exhaustive. L'algorithme se concentre uniquement sur une suite de sous-arbres de l'arbre maximal. En particulier, comme il a pu être vu en (1.5.5) lors de la création des catégories de marque, pour un nombre k donné, tous les arbres à k feuilles ne sont pas considérés par CART.

Construction de la séquence de sous-arbres

Cette séquence est construite en partant de l'arbre maximal T_{max} et puis en créant une séquence de la forme $T_1 \subset ... \subset T_{K-1} \subset T_K = T_{max}$ où $T_j,\ j = 1, ..., K$, représente l'arbre possédant j feuilles (en particulier T_1 est la racine de l'arbre T_{max}) et où la notation $\subset$ signifie "est un sous arbre de". Un élagage de type "coût-complexité" est appliqué à T_{max} en cherchant à minimiser le critère d'élagage

défini par la formule (3.2) ci-dessous

$$crit_\delta(T) = D(T) + \delta|\tilde{T}| = \underbrace{\sum_{h\in\tilde{T}} |h|D_h}_{coût} + \underbrace{\delta|\tilde{T}|}_{complexité}, \tag{3.2}$$

où $\tilde{T}$ représente l'ensemble des feuilles de l'arbre et $|\tilde{T}|$ désigne le cardinal de $\tilde{T}$, *i.e.* le nombre de feuilles de l'arbre. Ce critère introduit un paramètre de pénalisation $\delta \geq 0$. Ce paramètre règle le compromis entre la taille de l'arbre et la qualité de l'ajustement aux données : les grandes valeurs de δ déterminent des arbres de petite taille et inversement. En particulier, lorsque $\delta = 0$ la solution est l'arbre maximal T_{max}. L'élagage de l'arbre s'appuie sur le résultat suivant : pour chaque valeur de δ, il existe un unique sous-arbre de T_{max}, noté T_δ tel que

$$T_\delta = \arg\min_{T\subset T_{max}} crit_\delta(T). \tag{3.3}$$

Pour déterminer T_δ l'élagage du plus faible lien est utilisé, c'est à dire que l'algorithme supprime le nœud interne qui produit le plus petit accroissement du critère de coût $\sum_{h\in\tilde{T}} |h|D_h$. Ce processus d'élimination de nœuds internes est itéré jusqu'à produire l'arbre réduit à la racine de T_{max}. Une séquence d'arbres emboîtés est ainsi obtenue et il est possible de montrer que cette suite d'arbre contient nécessairement l'arbre T_δ pour tout $\delta \geq 0$. Plus précisément,

$$\begin{aligned}\forall l \in [1;K], \forall \delta \in [\delta_l;\delta_{l+1}[,\ T_l &= \arg\min_{T\subset T_{max}} crit_\delta(T)\\ &= \arg\min_{T\subset T_{max}} crit_{\delta_l}(T).\end{aligned}$$

Ainsi la séquence de sous arbres emboîtés créée par CART contient l'arbre optimal défini par l'équation (3.3) pour toute valeur de $\delta \geq 0$.

3.1.3 Mesure de l'importance des variables

L'algorithme CART fournit une mesure de l'importance des variables explicatives dans la prédiction ([Ghattas (2010)]). Une approche qui pourrait être utilisée pour mesurer l'importance d'une variable explicative serait par exemple de compter le nombre de fois où cette variable est utilisée pour séparer un nœud dans l'arbre produit. Les variables les plus importantes seraient alors les variables séparant le plus grand nombre de nœuds. Cependant cette approche conduirait à conclure qu'une variable n'apparaissant pas dans l'arbre final n'est pas importante pour la prédiction de la variable d'intérêt. Or, la variable choisie à chaque étape pour séparer un nœud est la variable dont la division maximise le critère défini par le critère (3.1) *supra*. Mais il se peut qu'une autre division sur cette variable ou bien une division sur une autre variable conduise à une valeur du critère très proche de celle fournie par la variable choisie. Utiliser l'approche décrite précédemment conduirait à considérer que la variable réalisant le deuxième maximum du critère (3.1) n'est pas discriminante pour la prédiction, seulement car une autre variable l'est encore plus.

L'approche proposée par l'algorithme CART pour calculer l'importance d'une variable explicative est basée sur la notion de division de substitution (*surrogate splits* en anglais). Soit r^* la règle de division optimale au nœud h, *i.e.* la règle satisfaisant le critère (3.1). La division de substitution au nœud h est définie comme la division alternative minimisant le nombre de désaccord avec la règle r^*. La division de substitution de r^* sera notée $\tilde{r}^*$ et la division de substitution basée sur la variable X_j

sera notée $\tilde{r}_j{}^*$. Avec ces notations, l'importance fournie par CART pour une variable X_j est donnée par

$$I(X_j) = \sum_{h \in T} \Delta i_{\tilde{r}_j{}^*}(h), \tag{3.4}$$

Ainsi, l'importance d'une variable X_j est donnée par la somme des diminutions d'hétérogénéité en chaque nœud h de l'arbre si l'on remplaçait la division optimale r^* par la division de substitution basée sur X_j. Avec le critère (3.4) une variable peut être considérée comme importante pour un modèle même si elle n'apparaît pas dans l'arbre associé. L'importance des variables est souvent ramenée à un pourcentage en divisant les valeurs par la somme des importances, puis en multipliant par 100.
À noter cependant qu'il existe un biais dans la sélection des variables par l'algorithme CART et donc dans l'importance des variables fournie par l'algorithme. Cela sera explicité dans la partie suivante présentant les avantages et les limites de ce modèle.

3.1.4 Avantages et limites de l'algorithme

Comme toute méthode de prédiction l'algorithme CART possède ses avantages et ses inconvénients. Il est important de les connaître et de les garder en mémoire lors de l'application du modèle afin de garantir une bonne interprétabilité des résultats.

Avantages

— L'algorithme CART permet de construire de façon rapide des estimateurs constants par morceaux.
— Concernant la présence d'éventuelles valeurs manquantes au sein des variables explicatives, CART permet un traitement efficace de ces dernières. Si une variable comportant des valeurs manquantes est choisie pour effectuer la division d'un nœud, l'algorithme utilise la série des divisions de substitution en ce nœud, ce qui permet de minimiser le nombre de différences avec la règle d'affectation définie par la variable optimale, inutilisable en raison de la présence de valeurs manquantes.
— Un autre avantage de l'algorithme CART est qu'il s'agit d'une approche non paramétrique : aucune loi de probabilité n'est imposée sur la variable à expliquer.
— Par ailleurs, la présence d'une donnée aberrante dans l'ensemble d'apprentissage ne contaminera essentiellement que la feuille à laquelle elle appartient et aura ainsi un impact assez faible sur les autres feuilles de l'arbre.
— Le principe de l'algorithme CART est simple et intuitif et ses résultats sont aisément lisibles et interprétables. Il est en effet possible de visualiser les arbres construits et il est alors facile de comprendre pourquoi, pour un ensemble de variables explicatives donné, une certaine valeur de Y est obtenue en sortie.

Limites

L'algorithme CART présente malheureusement également des inconvénients :

— Une des limites de cet algorithme est son instabilité : lorsque l'échantillon d'apprentissage est très légèrement modifié, les résultats de l'algorithme peuvent être très différents.

— Un autre inconvénient de l'algorithme est l'existence d'un biais dans la sélection des variables ([Strobl et al. (2010)]). En effet, pour X_1 et X_2 deux variables explicatives présentant respectivement n_1 et n_2 valeurs distinctes, si $n_1 < n_2$ alors X_2 présente plus de points de séparation que X_1 et donc a plus de chance d'être sélectionnée par hasard comme variable de séparation. Au vu de la définition de l'importance des variables dans CART, celle-ci est aussi biaisée.
— L'arbre construit par CART ne peut prédire que des valeurs qui ont été choisies pour convenir à la base d'apprentissage. L'algorithme ne peut pas s'adapter à des individus pour lesquels les valeurs de la variable d'intérêt diffèrent fortement de celles de la base d'apprentissage. Afin d'illustrer ce propos, considérons $n = 10000$ observations, dont $n_1 = 7000$ sont issues d'un tirage de variable uniforme sur [-5 ;5], ces observations sont stockées dans un vecteur noté `x_train`, et $n_2 = 3000$ observations sont issues d'un tirage de variable uniforme sur [-10 ;10] et stockées dans un vecteur noté `x_test`. Soient `y_train` et `y_test` les variables définies par

$$\begin{aligned} \texttt{y\_train}[i] &= 5 \times \texttt{x\_train}[i] + \epsilon_i, & i &= 1, ..., n_1 \\ \texttt{y\_test}[i] &= 5 \times \texttt{x\_test}[i] + \epsilon_i, & i &= 1, ..., n_2 \end{aligned}$$

où $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1), i = 1, ..., n$.
Ainsi `y_train` et `y_test` sont des transformations linéaires de respectivement `x_train` et `x_test`, auxquelles sont ajoutés des termes d'erreur gaussiens centrés réduits. Un modèle GLM avec une loi normale et la fonction de lien identité et un modèle CART sont appliqués à la base d'apprentissage, constituée de `y_train` et `x_train`, avec la formule `y_train` $\sim$ `x_train`. Les modèles entraînés sont utilisés pour prédire `y_test` à partir de `x_test`. Une erreur de prédiction est calculée pour chaque observation pour chacun des deux modèles, elle est égale au carré de l'écart entre la prédiction et la vraie valeur de `y_test`. Une représentation graphique de la valeur de cette erreur quadratique en fonction de `y_test` pour les deux modèles est donnée par la figure (3.1) ci-dessous :

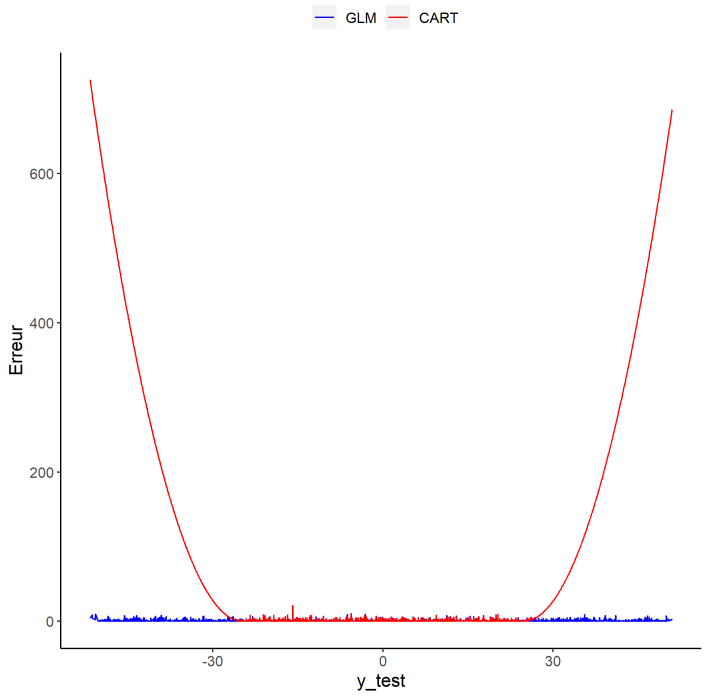


Figure 3.1 : Erreur de prédiction pour les modèles GLM et CART sur un échantillon gaussien

La figure (3.1) permet d'observer que pour les valeurs de `y_test` sur lesquelles les algorithmes n'ont

pas été entraînés, *i.e.* pour `y_test` < -25 ou `y_test` > 25, l'erreur est plus importante pour le modèle CART, alors que le modèle GLM a réussi à généraliser la structure sous-jacente des données : l'erreur est homogène pour toutes les valeurs de `y_test`. Les minimums et maximums prédits par chacun des deux modèles sont présentés dans le tableau (3.1) ci-dessous :

	Minimum	Maximum
$y_{pred_{GLM}}$	-50.03	50.03
$y_{pred_{CART}}$	-24.38	24.45
`y_test`	-50.82	51.47

TABLE 3.1 : Caractéristiques des prédictions des modèles GLM et CART sur un échantillon gaussien

Le minimum et le maximum du vecteur `y_test` ont également été représentés et la notation $y_{pred_{mod}}$ fait référence au vecteur des prédictions associées au modèle *mod*. Le tableau (3.1) montre que le modèle CART n'est pas capable de prédire des valeurs sortant du cadre de celles sur lesquelles il a été construit. Cet aspect est à garder en esprit lors du choix de modèles pour résoudre une problématique. Par exemple, dans ce mémoire une franchise de 337.5€ est appliquée, ainsi aucun coût de sinistre n'est inférieur à ce seuil dans la base d'apprentissage. Considérons, par exemple, une nouvelle base de sinistres automobiles, possédant les mêmes variables que la base étudiée dans ce mémoire, mais dans laquelle la franchise appliquée est de 200€. S'il était souhaité d'appliquer le modèle CART entraîné sur la base d'étude de ce mémoire sur cette nouvelle base, ce dernier ne serait pas capable de prédire des coûts inférieurs à 337.5€ (ou plus) et donc les résultats seraient moins bons que ceux obtenus dans le cadre de ce mémoire.

Ce dernier exemple souligne l'intérêt et l'importance de connaître les avantages et les inconvénients d'un modèle avant de pouvoir interpréter les résultats qui y sont associés.

3.1.5 Application du modèle CART

Construction et élagage de l'arbre maximal

Pour élaguer l'arbre maximal, il a été choisi d'utiliser la règle du 1 écart-type. Pour cela une suite d'arbre emboîtés est construite conformément à ce qui a été expliqué précédemment et pour chaque arbre l'erreur de validation et l'écart-type estimé de cette erreur sont calculés. La règle du 1 écart-type consiste à choisir l'arbre le plus petit pour lequel l'erreur de validation est inférieure à la somme entre la plus petite erreur de validation commise et l'écart-type estimé de cette erreur.

L'arbre maximal construit sans aucune restriction contient 3 680 nœuds internes et 3 681 feuilles. Cependant, il est souhaité obtenir des arbres élagués qui ne soient pas trop "grands", afin de pouvoir être facilement interprétables et de ne pas induire une segmentation trop fine. Pour empêcher la construction d'un arbre trop "grand", CART propose un paramètre permettant de spécifier la profondeur maximale de l'arbre construit. Afin de choisir quelle profondeur maximale spécifier, une validation croisée en $K = 5$ blocs est réalisée. Cette validation croisée est réalisée selon les étapes suivantes :

(1) : Division de l'échantillon d'apprentissage en K blocs

(2) : Pour $k = 1, ..., K$,

— Le bloc k est choisi comme bloc de test. Les $K-1$ blocs restants constituent l'échantillon d'apprentissage.

(3) : Pour $m \in \{m_1, ..., m_M\}$,

— L'arbre maximal est construit sur l'échantillon d'apprentissage avec pour condition que la profondeur de l'arbre ne peut être supérieure à m, puis l'arbre est élagué selon la règle de 1 écart-type.

(4) : Pour chacun des M modèles construits précédemment une métrique est évaluée. Il est choisi ici d'utiliser l'erreur des moindres carrés.

(5) : Pour chaque valeur m de profondeur maximale, une métrique est associée : c'est la moyenne des métriques pour les modèles CART évaluées sur chacun des K blocs, avec pour restriction d'avoir une profondeur inférieure ou égale à m.

Ces étapes sont illustrées schématiquement dans la figure (3.2) ci-dessous :

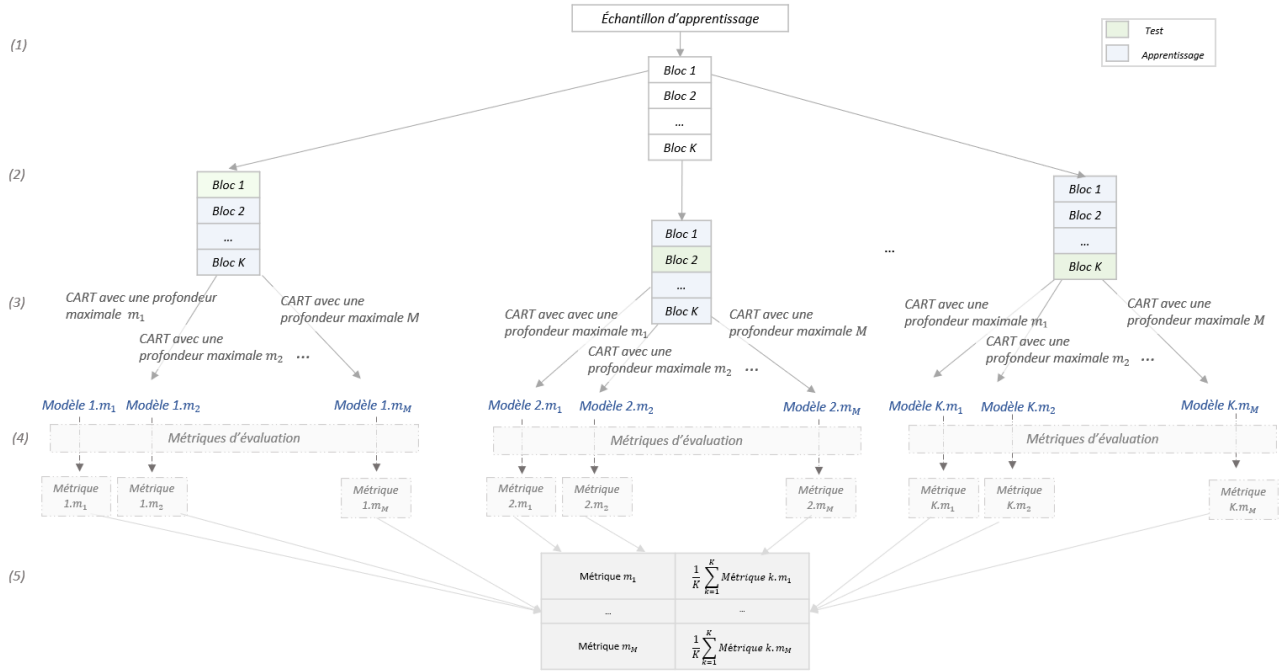


FIGURE 3.2 : Schéma simplifié du fonctionnement de la validation croisée pour déterminer la profondeur maximale de l'arbre à spécifier

L'intérêt d'effectuer une telle validation croisée est de s'assurer que le choix de la profondeur maximale de l'arbre n'est pas fait pour coller parfaitement à l'échantillon de validation mais bien pour être le meilleur choix de façon générale. En utilisant K échantillons de test différents, puis en prenant la moyenne des métriques pour chacun des découpages, le résultat obtenu est plus général et permet d'assurer la pertinence du choix d'une certaine profondeur. Les valeurs {2, 3, 4, 5, 6, 7, 8, 9, 10} ont été testées et les résultats obtenus sont donnés par la figure (3.3) ci-dessous :

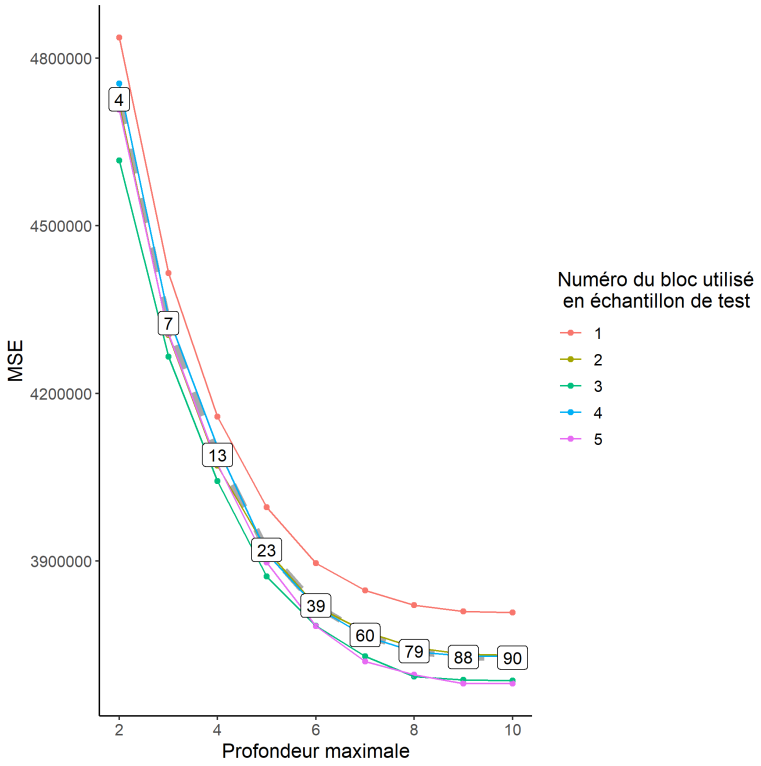


FIGURE 3.3 : Erreur des moindres carrés selon la profondeur maximale après une validation croisée sur $K = 5$ blocs

Chaque point de la figure (3.3) correspond à un choix de profondeur maximale et chaque couleur correspond à un bloc, *i.e.* aux résultats de la validation croisée lorsque le $k^{\text{ème}}$ bloc ($k = 1, ..., 5$) était utilisé comme échantillon de test. La moyenne des résultats est représentée par la ligne pointillée grise et la taille moyenne des arbres construits (après élagage) a également été représentée. Comme il pouvait être attendu, une décroissance de l'erreur des moindres carrés avec la profondeur maximale de l'arbre est observée. En effet, en restreignant la profondeur de l'arbre, la création de groupes qui permettraient potentiellement une baisse de l'hétérogénéité est empêchée. Cependant, si l'arbre possède trop de feuilles, son analyse peut devenir fastidieuse et l'aspect "facilité d'interprétation" des arbres CART est alors atténué.
Il est choisi dans ce mémoire de construire des arbres dont la profondeur ne peut excéder 4. Ainsi, au maximum, l'arbre construit pourra contenir $2^4 = 16$ feuilles, ce qui est un nombre convenable de groupes créés pour pouvoir en permettre une bonne analyse et permettre de conduire à une segmentation des individus qui ne soit pas trop fine.

L'arbre maximal est construit selon les paramètres décrits précédemment : il contient 15 nœuds internes et 16 feuilles. Une représentation graphique de cet arbre est donnée par la figure (3.4) ci-dessous. Chacune des feuilles de l'arbre contient un certain nombre d'observations et est associée à une valeur qui n'est autre que la moyenne des coûts des sinistres contenus dans la feuille. Cet arbre maximal est ensuite élagué selon la règle du 1 écart-type.

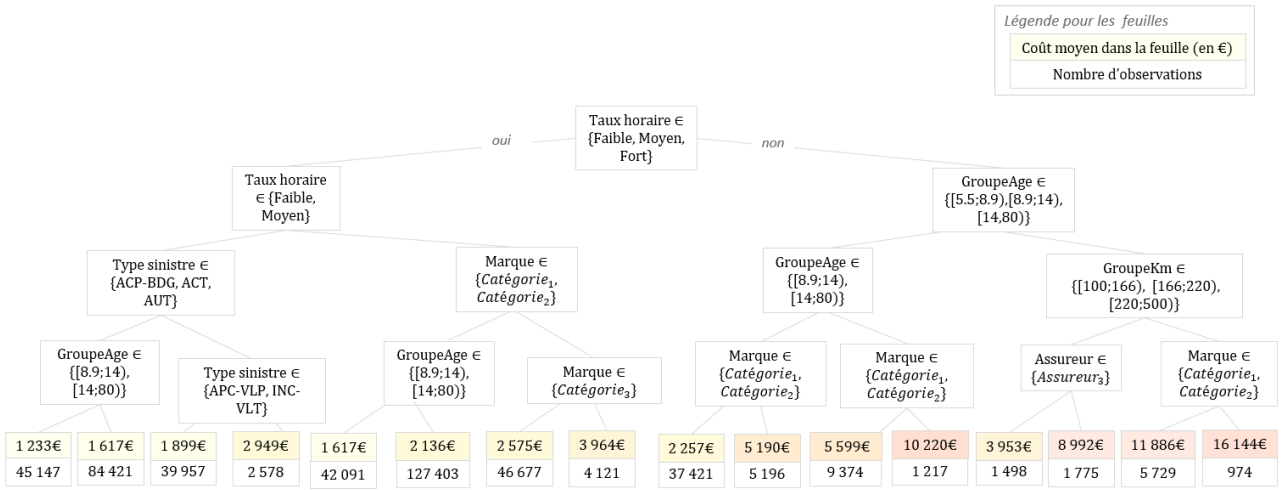


FIGURE 3.4 : Arbre maximal obtenu avec une profondeur maximale égale à 4

Une représentation graphique de l'erreur de validation commise par l'arbre en fonction du critère de pénalisation choisi est donnée par la figure (3.5) ci-dessous :

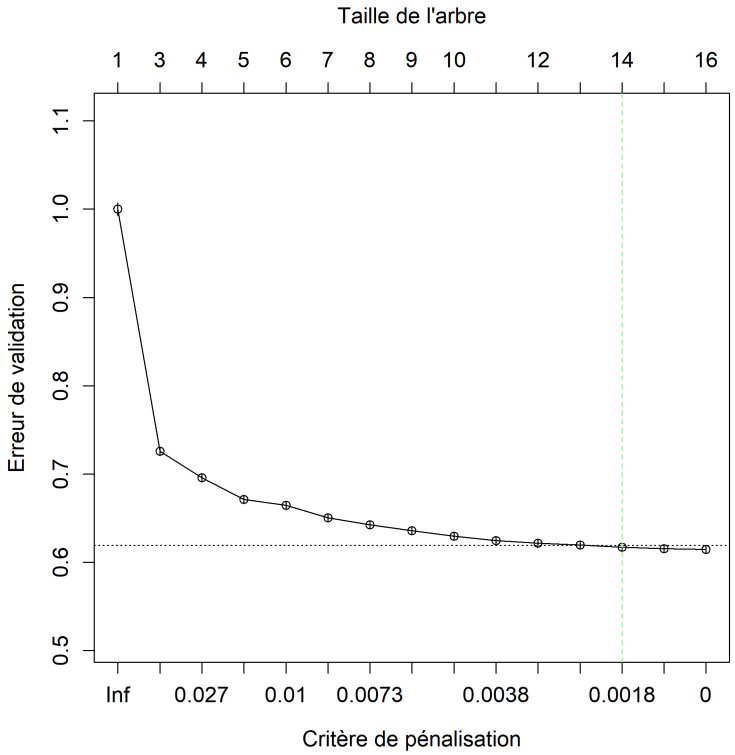


FIGURE 3.5 : Erreur de validation et taille de l'arbre en fonction du critère de pénalité choisi

La ligne pointillée noire représente l'erreur de validation la plus faible sur l'ensemble des arbres à laquelle il a été ajouté son écart-type estimé. L'arbre considéré comme optimal est celui associé au critère de pénalisation le plus élevé pour lequel l'erreur de validation commise est inférieure à ce seuil. Il s'agit ici de l'arbre possédant 13 nœuds internes et 14 feuilles, dont une représentation graphique est donnée ci-dessous par la figure (3.6) :

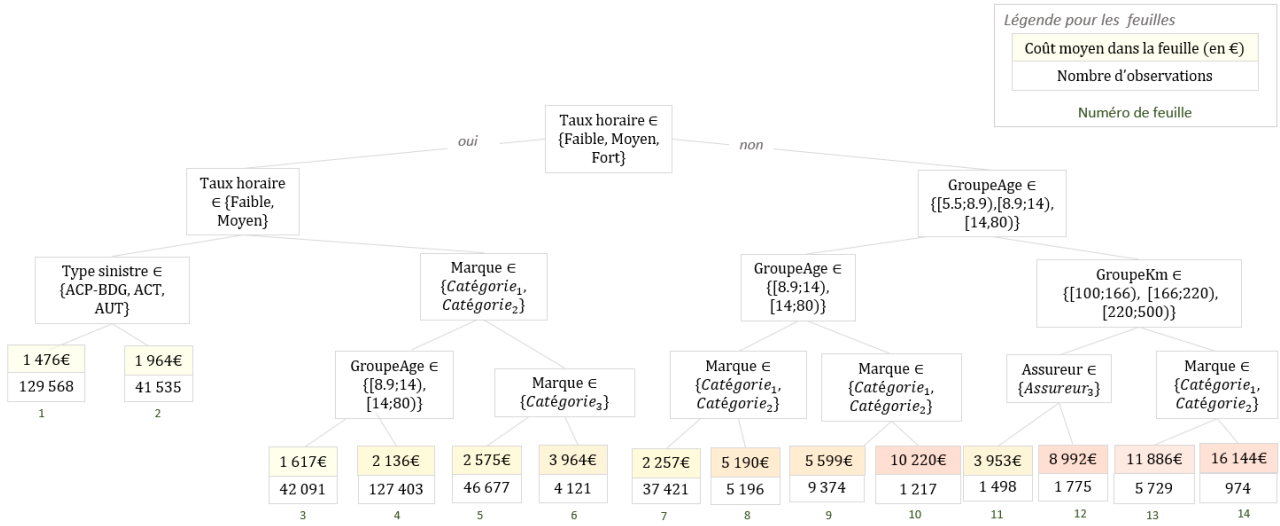


FIGURE 3.6 : Arbre obtenu après élagage selon la règle du 1 écart-type

Les feuilles de l'arbre élagué ont été numérotées de 1 à 14. À noter que s'il n'avait pas été spécifié de profondeur maximale, l'arbre optimal aurait contenu 95 feuilles et la plus petite feuille de cet arbre aurait regroupé 10 observations. Cela montre que construire l'arbre CART sans restriction aurait conduit à créer une segmentation trop fine.

Pouvoir prédictif du modèle

L'arbre optimal représenté ci-dessus en figure (3.6) est utilisé pour évaluer les prédictions sur l'échantillon de validation. Les résultats suivants ont été obtenus :

MSE	MAE
4 137 226	1 266

TABLE 3.2 : Résultats de prédiction sur l'échantillon de validation avec l'arbre CART optimal

Les résultats obtenus avec CART sont meilleurs que ceux obtenus avec le modèle GLM, bien que l'algorithme commette encore des erreurs.
La moyenne des sinistres prédits par CART pour l'échantillon de validation est égale à 2 294€. Pour rappel, la moyenne des sinistres dans l'échantillon de validation est égale à 2 296€, et la moyenne des sinistres prédits par le modèle GLM était de 2 259€. L'écart entre les moyennes des sinistres prédits et observés est donc plus faible pour le modèle CART que pour le modèle GLM. Afin de mieux observer d'où provient l'écart de moyennes pour le modèle CART, une analyse des moyennes des sinistres prédits et observés par modalités est effectuée, similairement à ce qui avait été fait précédemment pour le modèle GLM. Les résultats obtenus sont présentés en figure (3.7) ci-dessous :

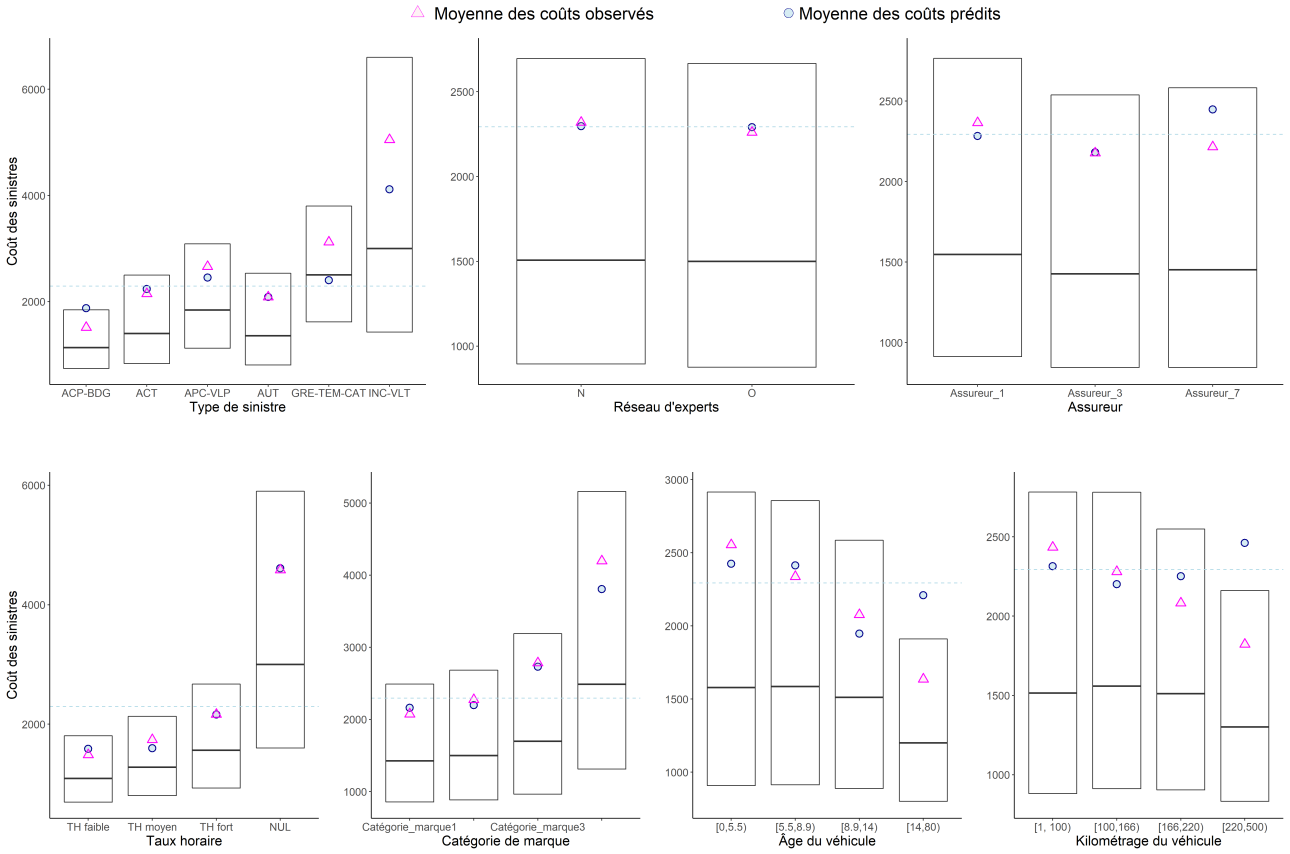


FIGURE 3.7 : Moyennes des sinistres observés et prédits par CART pour chaque modalité des variables explicatives

La ligne pointillée bleue représente sur chacun des graphiques la moyenne des coûts prédits par CART (égale à 2 294€). L'analyse de la figure (3.7) permet d'observer que l'amélioration fournie par le modèle CART par rapport au modèle GLM ne provient pas seulement d'une réduction des écarts de moyennes pour chacune des modalités. En effet, des différences avec ce qui avait été observé pour le GLM sur la figure (2.4) *supra* peuvent être soulignées, comme par exemple :
— une sous-estimation moyenne relativement importante pour la catégorie de marque de véhicule n°4. Dans le cadre du GLM, un écart plus faible et inverse était observé : en moyenne le modèle GLM surestimait légèrement les coûts des sinistres pour cette catégorie de marque ;
— une sous-estimation en moyenne des sinistres grêle (`GRE`), tempête (`TEM`) ou catastrophe naturelle (`CAT`), alors qu'ils étaient légèrement surestimés par le modèle GLM en moyenne ;
— une surestimation en moyenne des coûts associés à l'assureur n°7, alors qu'ils étaient sous-estimés par le modèle GLM en moyenne ;
— une sous-estimation en moyenne des coûts des sinistres pour les véhicules dont l'âge est compris entre 8.9 et 14 ans, alors que le phénomène inverse était observé avec le modèle GLM.
— l'écart relativement important qui avait été observé au sein des taux horaires nuls (*i.e.* lorsque le véhicule était économiquement irréparable) n'est plus observé lorsque l'on considère les prédictions associées à CART ;

Néanmoins, des phénomènes similaires à ce qui avait été observé avec les prédictions du modèle GLM sont également retrouvés, comme par exemple :
— une surestimation en moyenne des coûts pour les véhicules possédant un kilométrage supérieur à

165 601 kilomètres, avec un écart de moyennes encore plus important que celui qui avait été observé avec le GLM ;
— une sous-estimation des coûts pour les sinistres incendies ou vols totaux, avec une sous-estimation encore plus importante que celle observée dans le cadre du GLM ;
— des écarts de moyennes relativement faibles pour les deux réseaux d'experts étudiés dans ce mémoire.

Le modèle CART commettant encore des erreurs, il peut être intéressant de s'intéresser aux caractéristiques de chaque feuille créée par CART, afin d'observer si certaines feuilles conduisent à des erreurs particulièrement importantes. Une représentation graphique de l'erreur dans chaque feuille en fonction du nombre d'observations qu'elle contient est donnée ci-dessous par la figure (3.8) :

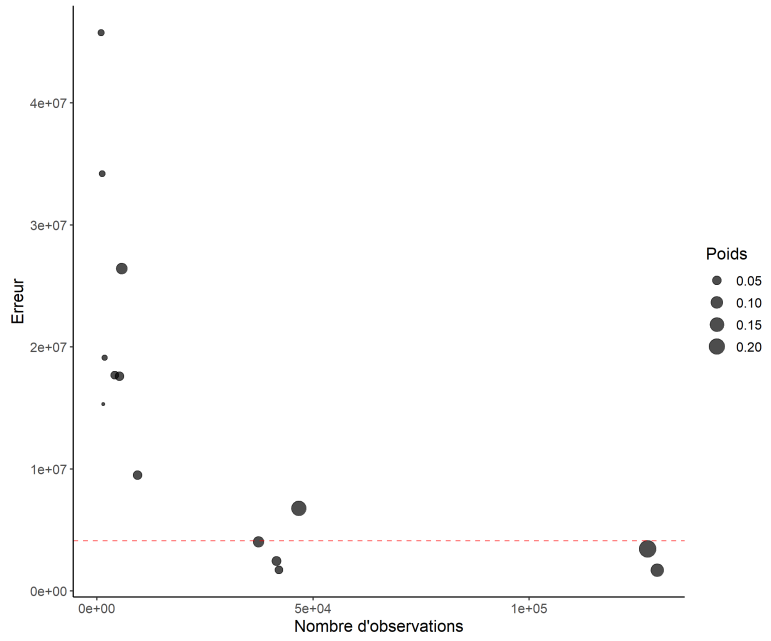


FIGURE 3.8 : Erreur en fonction du nombre d'observations dans chaque feuille issue de CART

Chaque point de la figure (3.8) représente une feuille et l'erreur des moindres carrés commise par le modèle sur l'échantillon de validation a été représentée par une ligne pointillée rouge. La taille du point est associée au poids que la feuille a dans l'erreur quadratique moyenne, défini pour une feuille $F_k,\ k \in \{1, ..., 14\}$, par

$$Poids(F_k) = \frac{\sum_{y_i \in F_k} (y_i - \overline{y}_k)^2}{\sum_{k=1}^{14} \sum_{y_i \in F_k} (y_i - \overline{y}_k)^2}, \tag{3.5}$$

où $\overline{y}_k$ est la moyenne du coût des sinistres pour les observations au sein de la feuille k.
La figure (3.8) permet d'observer que les erreurs les plus importantes sont produites dans les "petites" feuilles. Cependant, du fait du nombre faible d'observations qu'elles contiennent, le poids de ces erreurs dans l'erreur totale est assez faible.
L'algorithme CART permet de spécifier le nombre minimum d'observations souhaité dans chaque feuille de l'arbre. Afin d'observer si spécifier un nombre minimum d'observations par feuille permettrait d'améliorer les résultats, une validation croisée est effectuée sur l'échantillon d'apprentissage pour différentes valeurs de minimum. Cette validation croisée est réalisée selon le même principe que celui illustré par la figure (3.2) *supra* avec les valeurs candidates de minimum {7, 1000, 2500, 5000, 7500, 10000}, la valeur 7 correspondant à la valeur par défaut. Les résultats obtenus sont présentés en figure (3.9) ci-dessous :

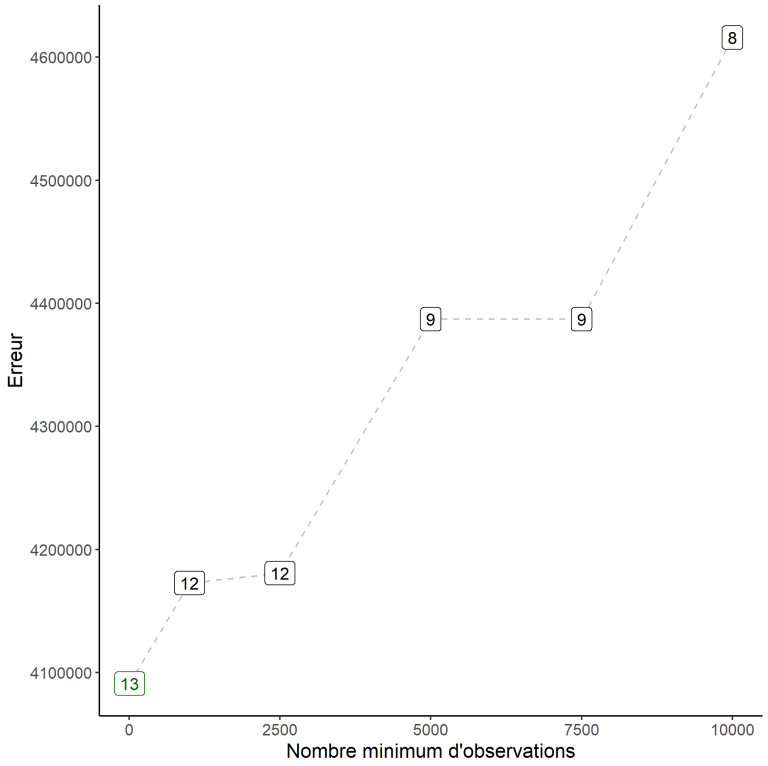


FIGURE 3.9 : Erreur des moindres carrés pour différentes valeurs de minimum d'observations à avoir dans chaque feuille de l'arbre CART

Chaque point de la figure (3.9) correspond à la moyenne de l'erreur des moindres carrés sur les 5 blocs de la validation croisée pour une valeur minimum d'observations par feuille. La taille moyenne des arbres en résultant à également été représentée. Au vu des résultats obtenus, il est choisi de ne pas modifier le paramètre du nombre minimum d'observations et de conserver sa valeur par défaut. Il est alors possible de procéder à une étude de l'importance des variables au sein de l'arbre optimal.

Importance des variables

L'importance des variables définies en (3.1.3) a été calculée et les résultats obtenus sont présentés en pourcentage dans le tableau (3.3) ci-dessous :

Variable	**Importance (%)**
`GroupeAge`	43
`tauxHoraire`	37
`CategorieMarque`	8
`GroupeKm`	7
`ASSUREUR`	3
`typeSinistre`	2
`expertA`	0.03

TABLE 3.3 : Importance des variables explicatives dans CART

Les variables représentant le groupe d'âge du véhicule sinistré et le taux horaire du garage intervenu

sont les plus importantes pour le modèle CART. Concernant la variable représentant le réseau d'experts intervenu, cette variable possède l'importance la plus faible parmi toutes les variables explicatives. Bien que la lecture de l'importance des variables dans le modèle CART soit aisée, cette dernière ne permet pas de mesurer l'influence des variables sur la prédiction.

Mesure de l'influence du réseau d'expert

Pour analyser l'influence d'une variable, il est intéressant d'observer comment la variable est utilisée pour séparer des groupes de sinistres. Concernant la variable représentant le réseau d'experts intervenu, cette dernière n'intervient pas dans la séparation des individus, comme il peut être vu sur la figure (3.6) *supra*. Il est tout de même possible d'observer les caractéristiques de cette variable au sein de chacun des 14 groupes créés par CART :

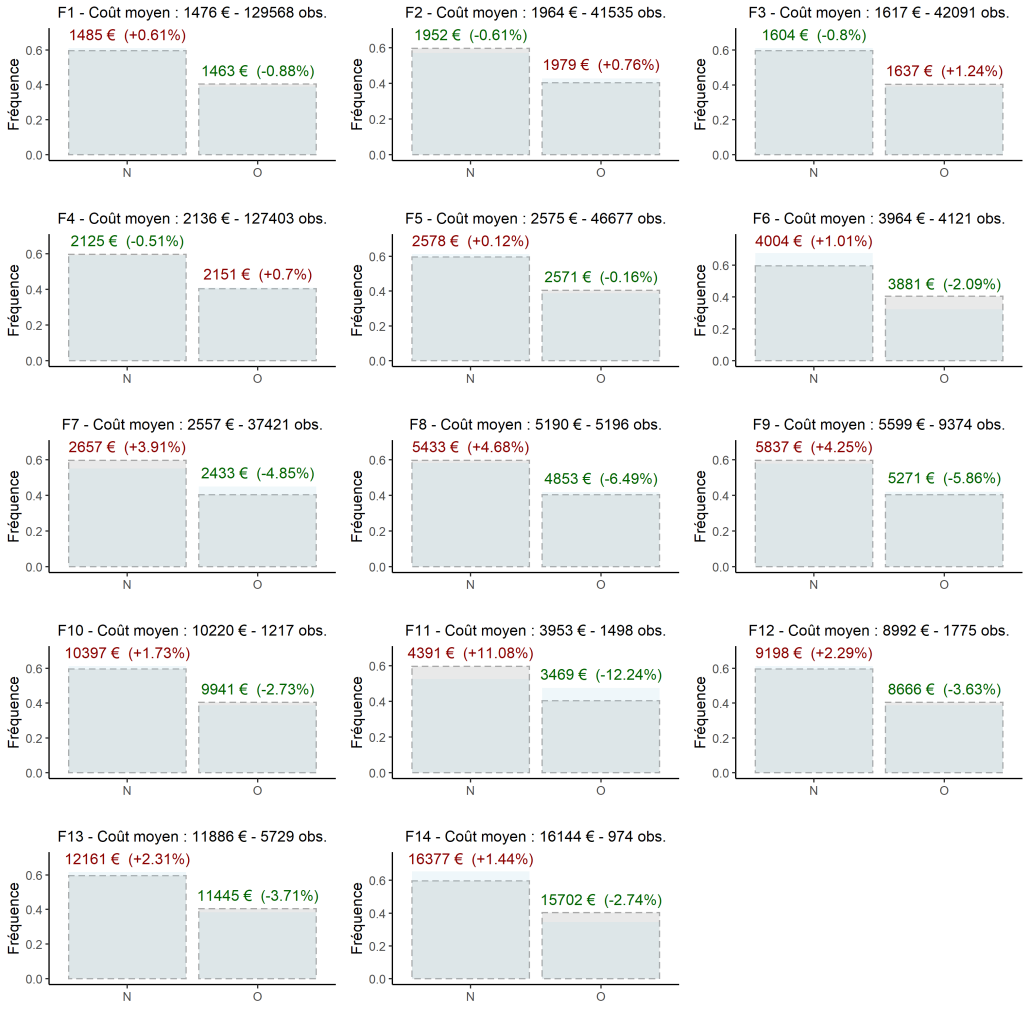


FIGURE 3.10 : Caractéristiques du réseau d'experts au sein de chaque feuille de l'arbre CART

La figure (3.10) ci-dessus présente les caractéristiques de la variable `expertA` dans chacune des feuilles de l'arbre construit sur l'échantillon d'apprentissage. Au sein de chaque feuille les individus peuvent être séparés en deux groupes : ceux ayant eu recours au réseau d'experts A et ceux ayant eu recours au réseau d'experts B. La fréquence de chacun de ces groupes au sein de la feuille a été

représentée (couleur bleue) avec en arrière plan la fréquence des deux réseaux au sein de l'ensemble de l'échantillon d'apprentissage (couleur grise), à savoir environ 60% d'observations associées au réseau B. Le coût moyen au sein des deux groupes a également été représenté. Par exemple, pour la feuille numéro 1, dénotée F1, le coût moyen au sein de la feuille est de 1 476€ et cette feuille contient 129 568 observations. Au sein des individus de la feuille, le coût moyen des individus ayant eu recours au réseau d'experts B est de 1 485€, soit une augmentation de 0.61% par rapport au coût moyen. À l'inverse, pour les individus ayant eu recours au réseau d'experts A une diminution du coût moyen égale à -0.88% est observée, puisque le coût moyen pour ces individus est égal à 1 463€.
Pour 11 des 14 feuilles, regroupant plus de 70% des observations de l'échantillon d'apprentissage, le coût moyen des individus ayant eu recours au réseau d'experts A est plus faible que celui des individus ayant eu recours au réseau d'experts B. Cependant, comme expliqué précédemment, la comparaison des coûts moyens associés à chaque réseau d'experts pour conclure à une hiérarchisation de performances ne serait pas rigoureuse. Ainsi, le modèle CART ne permet pas de tirer de conclusion sur la différence de performance entre les deux réseau d'experts.

L'algorithme CART fournit des meilleures performances de prédiction que le GLM, mais il commet encore des erreurs. Afin d'essayer d'améliorer les résultats obtenus précédemment l'algorithme MOB va être implémenté. Cet algorithme fait partie des algorithmes de partitionnement récursif basé sur des modèles, présentés ci-après.

3.2 Partitionnement récursif basé sur des modèles

Comme il a été vu dans la section précédente, les arbres de régression sont caractérisés par deux critères qui expliquent leur succès : leur facilité d'interprétation et leur pouvoir prédictif. Cependant, cette deuxième caractéristique est de moins en moins fondée, puisque les arbres de régression sont concurrencés par des modèles de *machine learning* tels que le *Boosting*, les forêts aléatoires ou encore les machines à support vecteur. Les pouvoirs prédictifs de ces modèles s'avèrent en effet souvent supérieurs à ceux des arbres de régression. Ces derniers sont néanmoins souvent caractérisés de modèles "boite noire" en référence à leur interprétabilité généralement limitée.

Dans ce contexte, les algorithmes de partitionnement récursif basé sur des modèles ([Zeileis et al. (2008)]), souvent appelés MOB pour *MOdel Based recursive partitionning*, ont été introduits. Ces modèles peuvent être considérés comme un compromis entre amélioration du pouvoir de prédiction des arbres et conservation de la facilité d'interprétation du modèle. Ils reposent sur l'idée suivante : il peut être possible d'améliorer les résultats d'un modèle en partitionnant la population en sous groupes et en ajustant dans chacun des groupes un modèle, plutôt qu'en appliquant un unique modèle à l'ensemble de la population. Les modèles de ce type sont appelés "modèles segmentés", ils sont introduits ci-après.

3.2.1 Modèles segmentés

Considérons un ensemble d'observations $(Y_i, X_i)_{i=1,\dots,n}$, où $Y \in \mathbb{R}^n$ est la variable à expliquer et $X \in \mathbb{R}^{n\times p}$ constitue l'ensemble des variables explicatives. Il est souhaité d'ajuster sur ces observations un modèle paramétrique $\mathcal{M}(Y, X, \theta)$ associé à un paramètre $\theta \in \Theta$, vecteur de dimension p. Ce modèle peut être ajusté en minimisant une certaine fonction objectif $\Psi(Y, X, \theta)$ donnant alors l'estimateur de θ suivant

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \Psi(Y_i, X_i, \theta).$$

Par exemple, si $\Psi(Y_i, X_i, \theta) = (Y_i - X_i\theta)^2$ est l'erreur au carré, il s'agit de l'estimateur des moindres carrés ordinaires, et lorsque Ψ est la log-vraisemblance négative cela correspond à l'estimateur du maximum de vraisemblance. Dans le cadre du GLM, θ représente les coefficients de régression.

Il se peut que le modèle $\mathcal{M}$ ne s'ajuste pas bien à toutes les observations. L'idée des modèles segmentés est de partitionner les observations selon certaines variables de façon à créer des groupes dans lesquels différents modèles pourront être ajustés, afin d'obtenir un meilleur ajustement au global. Plus précisément, un nombre l de variables de partitionnement est choisi, puis l'ensemble des variables explicatives X est séparé en deux groupes : $X = (Z, V)$, où $Z \in \mathbb{R}^{n\times l}$ constitue l'ensemble des variables de partitionnement qui seront utilisées pour créer les groupes et $V \in \mathbb{R}^{n\times(p-l)}$ constitue l'ensemble des variables permettant l'ajustement des modèles au sein de ces groupes. Dans le cadre de ce mémoire, l'objectif étant de prédire une variable quantitative, les variables V seront appelées variables de régression.

L'idée est alors de définir une partition $\mathcal{B} = \{\mathcal{B}_b\}_{b=1,\dots,B}$ de l'espace $\mathcal{Z} = Z_1 \times \ldots \times Z_l$ de B groupes, telle que dans chaque groupe $\mathcal{B}_b$ un modèle $\mathcal{M}(Y, \theta_b)$ associé à un paramètre θ_b spécifique au groupe soit ajusté. Le modèle segmenté, noté $\mathcal{M}_{\mathcal{B}}(Y, Z, V, \theta)$, regroupe alors B sous-modèles et est associé à un paramètre de dimension B noté $\theta_{\mathcal{B}} = (\theta_1, ..., \theta_B)^T$. L'idée générale des modèles segmentés est illustrée par la figure (3.11) ci-dessous :

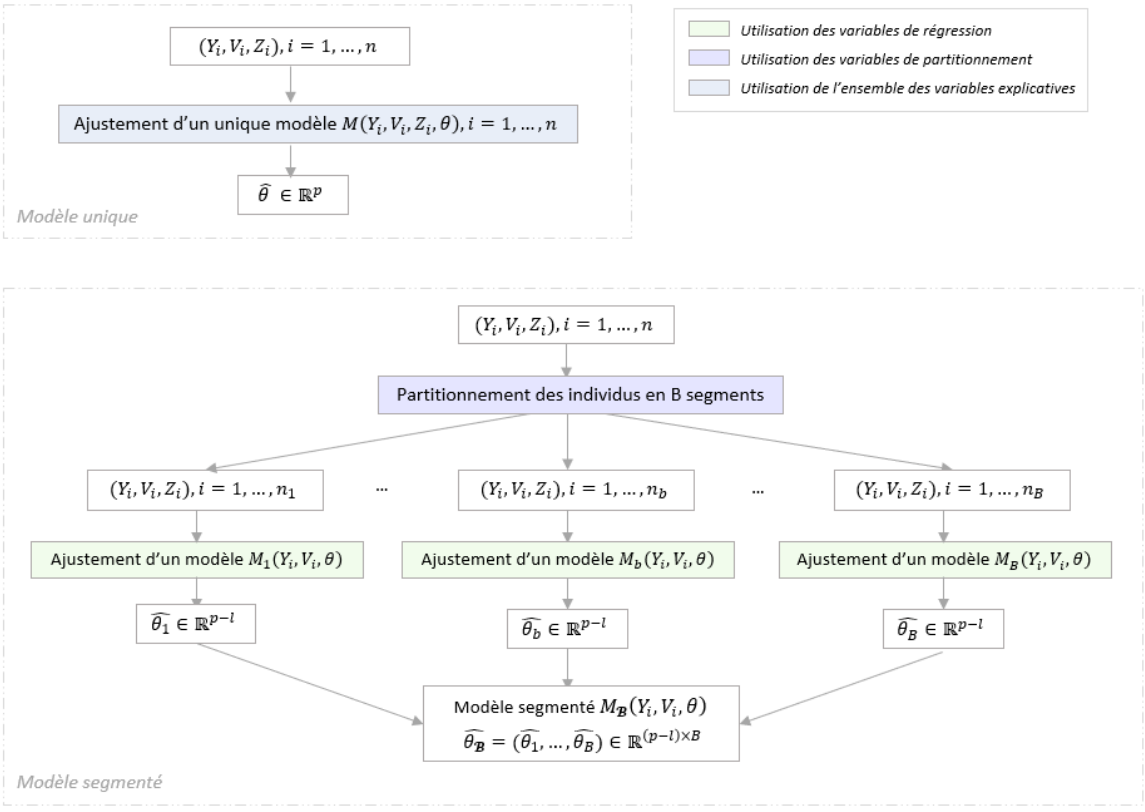


FIGURE 3.11 : Schéma simplifié du principe des modèles segmentés

L'ajustement du modèle segmenté repose une fois de plus sur la minimisation d'une fonction objectif

$$\Psi_{\{\mathcal{B}_b\}_b}(\theta) = \sum_{b=1}^{B} \sum_{Y_i \in \mathcal{B}_b} \Psi(Y_i, V_i, \theta_b). \tag{3.6}$$

La minimisation de cette fonction peut être réalisée en estimant localement chaque paramètre θ_b au sein de chacun des B groupes.

Cependant, minimiser la fonction objectif définie en (3.6) nécessite de connaître la partition $\{\mathcal{B}_b\}_{b=1,...,B}$. Même si le nombre de partition B est connu, dès qu'il y a plus d'une variable de partitionnement le nombre de partitions candidates devient rapidement trop grand pour procéder à une recherche exhaustive de la partition optimale. Lorsque ce nombre est inconnu la complexité du problème est encore plus importante. Afin de trouver une partition proche de la partition optimale sans introduire trop de complexité dans la recherche, une méthode de recherche dite "gloutonne" (appelée *greedy search* en anglais) est utilisée. Le principe de ce type de méthodes est d'effectuer à chaque étape un choix localement optimal dans le but d'obtenir un résultat globalement optimal. Cette recherche est mise en place au sein de l'algorithme de partitionnement récursif.

3.2.2 Algorithme de partitionnement récursif

L'objectif de l'algorithme de partitionnement récursif est de construire une partition de façon à ajuster un modèle dans chacun des groupes créés. De même que pour l'algorithme CART, l'ensemble des observations sont initialement regroupées dans un premier nœud, appelé racine. Ce premier nœud est ensuite divisé en B nœuds fils et la procédure se répète ainsi au sein de chaque nœud. Pour déterminer si la division d'un nœud est réalisée un test d'instabilité des paramètres est effectué. S'il y a une instabilité significative par rapport à l'une des variables de partitionnement, le nœud est divisé en B segments localement optimaux et la procédure est ensuite répétée. À l'inverse, s'il est décidé de ne pas partitionner les observations au sein d'un nœud, ce nœud devient un nœud terminal, appelé feuille, et un modèle est ajusté dans cette feuille. Plus précisément, l'algorithme de partitionnement récursif suit les étapes suivantes :

Au sein d'un nœud h, l'algorithme :

— ajuste le modèle sur l'ensemble des observations du nœud, *i.e.* calcule le paramètre $\theta \in \Theta$ via la minimisation de la fonction objectif Ψ

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \sum_{Y_i \in h} \Psi(Y_i, V_i, \theta). \tag{3.7}$$

— évalue si les estimations des paramètres du modèle sont stables par rapport aux variables de partitionnement $Z_1, ..., Z_l$. S'il existe une instabilité significative pour une ou plusieurs variables Z_j, la variable associée à la plus grande instabilité, notée Z_{j^*}, est sélectionnée, sinon l'algorithme s'arrête dans ce nœud, qui devient alors une feuille associée au modèle ajusté via (3.7).

— calcule le(s) point(s) de séparation (appelés *splits points* en anglais) de la variable Z_{j^*} qui optimise(nt) localement Ψ, pour un nombre de séparations fixe ou choisi de manière adaptative.

— Le nœud est divisé en B nœuds fils et la procédure est répétée jusqu'à ce qu'aucune instabilité significative ne soit détectée, ou qu'un autre critère d'arrêt pré-défini soit atteint.

Les arbres de régression constituent un cas particulier de l'algorithme de partitionnement récursif dans lequel le nombre de segments est égal à $B = 2$ et l'ensemble des variables explicatives sont utilisées comme variables de partitionnement. Chaque groupe créé est associé à modèle simple qui n'est autre que la moyenne des observations dans chaque groupe.

Les étapes 1 à 3 de l'algorithme de partitionnement récursif sont détaillées ci-après.

Ajustement du modèle

L'ajustement du modèle dans un nœud h se fait via l'estimation du paramètre θ donnée par l'équation (3.7). Cette estimation peut être calculée en résolvant les conditions du premier ordre

$$\sum_{Y_i \in h} \frac{\partial \Psi(Y_i, V_i, \theta)}{\partial \theta} = \sum_{Y_i \in h} \psi(Y_i, V_i, \theta) = 0,$$

où : $\psi(Y_i, V_i, \theta) = \frac{\partial \Psi(Y_i, V_i, \theta)}{\partial \theta}$ est appelée fonction score. Cette fonction évaluée en les paramètres estimés est notée $\widehat{\psi_i} = \psi(Y_i, V_i, \widehat{\theta})$.

Test de l'instabilité des paramètres

L'objectif de cette étape est de déterminer si les paramètres du modèle ajusté sont stables sur chaque ordre particulier impliqué par les variables de partitionnement $Z_j, j = 1, ..., l$, ou si la division de l'échantillon par rapport à l'une des variables Z_j pourrait capturer des instabilités dans les paramètres et alors améliorer l'ajustement du modèle.

Afin d'évaluer l'instabilité des paramètres selon une variable de partitionnement Z_j, une idée est de vérifier si les scores $\widehat{\psi_i}$ fluctuent aléatoirement autour de leur moyenne 0, ou s'ils présentent des déviations systématiques de 0 sur Z_j. Pour capturer ces déviations, le processus de fluctuation empirique suivant est utilisé

$$\forall t \in [0, 1],\ W_j(t) = \frac{1}{\sqrt{\widehat{J}}} \frac{1}{\sqrt{n_h}} \sum_{i=1}^{\lfloor n_h t \rfloor} \widehat{\psi}_{\sigma(Z_{ij})},$$

où $\sigma(Z_{ij})$ est la permutation d'ordonnancement donnant l'anti-rang[1] de l'observation Z_{ij} dans le vecteur $Z_j = (Z_{1j}, ..., Z_{n_h j})$. $W_j(t)$ est le processus de la somme partielle des scores ordonnés par la variable Z_j, mis à l'échelle par le nombre d'observations présentes dans le nœud n_h et une estimation $\widehat{J}$ de la matrice de covariance de $\psi(Y, \widehat{\theta})$. Par exemple, l'estimateur de la matrice de covariance suivant peut être utilisé

$$\widehat{J} = \frac{1}{n_h} \sum_{i=1}^{n_h} \psi(Y_i, \widehat{\theta}) \psi(Y_i, \widehat{\theta})^T,$$

mais d'autres estimateurs plus robustes sont également applicables.

[1]Pour $v = (v_1, ..., v_n)$ l'anti rang de la valeur v_i $(i = 1, ..., n)$ est l'indice de cette valeur dans le vecteur v trié de façon décroissante.

Afin de tester l'instabilité des paramètres un test dit test de M-fluctuation ([Zeileis et Hornik (2003)]) est utilisé. L'hypothèse nulle de ce test est la stabilité des paramètres et il repose sur un théorème central limite fonctionnel : sous l'hypothèse nulle le processus empirique $W_j(t)$ converge vers un pont brownien W^0. Ce résultat permet la construction d'une statistique de test permettant d'affirmer ou non l'instabilité des paramètres. Deux statistiques de test peuvent être utilisées en fonction de la nature de la variable de partitionnement :

Si la variable de partitionnement est quantitative
Pour capturer les instabilités éventuelles sur une variable numérique Z_j la statistique de test suivante est utilisée

$$\lambda_{supLM}(W_j) = \max_{i=\underline{i},\ldots,\bar{i}} \left(\frac{i}{n_h} \cdot \frac{n_h - i}{n_h} \right)^{-1} \left\| W_j \left(\frac{i}{n_h} \right) \right\|_2^2 .$$

Il s'agit du maximum du carré de la norme L^2 du processus de fluctuation empirique normalisé par sa fonction de variance. C'est la statistique *supLM* d'Andrews (1993) ([Andrews (1993)]). L'intervalle $[\underline{i}, \bar{i}]$ est défini en exigeant une certaine taille minimale de segment $\underline{i}$, puis en définissant $\bar{i} = n_h - \underline{i}$. C'est l'intervalle associé aux point de rupture potentiels de la variable Z_j. La distribution limite de cette statistique est donnée par le supremum d'un processus de Bessel lié à k dimensions, au carré : $sup_t(t(1-t))^{-1}\|W^0(t)\|_2^2$. La p-valeur associée p_j peut alors être calculée.

Si la variable de partitionnement est qualitative
Pour capturer l'instabilité par rapport à une variable catégorielle Z_j avec C catégories différentes, la statistique utilisée est donnée par

$$\lambda_{\chi^2}(W_j) = \sum_{c=1}^{C} \frac{|I_c|^{-1}}{n_h} \left\| \Delta_{I_c} W_j \left(\frac{i}{n_h} \right) \right\|_2^2 , \tag{3.8}$$

où $\Delta_{I_c} W_j$ est l'incrément du processus de fluctuation empirique sur les observations de la modalité $c = 1, ..., C$, et I_c est l'ensemble des indices des observations appartenant à la catégorie c. La statistique de test a une distribution asymptotique du χ^2 avec $k \cdot (C-1)$ degrés de liberté. La p-valeur correspondante p_j peut alors être calculée ([Hjort et Koning (2002)]).

Pour tester s'il existe une certaine instabilité globale dans un nœud h, il suffit de vérifier si la p-valeur minimale $p_{j^*} = \min_{j=1,\ldots,l} p_j$ est inférieure à un certain seuil de significativité pré-spécifié α. Si c'est le cas, la variable Z_{j^*} associée à la p-valeur minimale est choisie pour diviser le modèle dans l'étape suivante de l'algorithme.

Segmentation

Cette étape est effectuée par l'algorithme dans un nœud h lorsqu'une instabilité par rapport à au moins une variable de partitionnement a été détectée lors de l'étape précédente. L'ensemble des observations présentes dans le nœud h doit être divisé en B segments par rapport à la variable Z_{j^*} associée à la plus grande instabilité. Le but est de déterminer le critère de division optimal par rapport à cette variable (par exemple, si Z_{j^*} est une variable numérique des seuils de division optimaux sont recherchés, si elle est catégorielle la construction de groupes de modalités optimaux est recherchée). L'algorithme CART

fournissant exclusivement des divisions binaires, dans le cadre de ce mémoire le nombre de segments choisi sera toujours égal à $B = 2$.

Pour choisir quelle segmentation est la plus optimale, les segmentations candidates sont comparées à l'aide de la fonction objectif définie en (3.6) *supra*. À noter qu'une recherche exhaustive sur l'ensemble des segmentations possibles avec B segments garantit de trouver la segmentation optimale, mais cette recherche peut rapidement devenir fastidieuse. Cependant, la restriction à des divisions binaires permet de réduire la complexité de cette recherche :
— si la variable de partitionnement est numérique, la recherche de la division optimale est basée sur des méthodes de détection de points de ruptures et de changement structurels. Lorsque les divisions candidates sont binaires, cette recherche est d'ordre $O(n)$ et, pour $B > 2$, la recherche exhaustive est d'ordre $O(n^{B-1})$.
— si la variable de partitionnement est catégorielle et possède C catégories, la recherche d'une division binaire optimale consiste à tester toutes les combinaisons formant deux groupes de modalités, cette recherche est alors d'ordre $O(2^{C-1})$. Si, de plus, la variable est ordonnée, cette complexité est réduite à $O(C)$.

Les étapes décrites ci-dessus sont exécutées à nouveau dans chacun des deux nœuds fils jusqu'à ce qu'aucune instabilité significative ne soit détectée ou qu'un critère d'arrêt pré-défini soit atteint. Un arbre est alors construit et chaque feuille est associée à un modèle. Une illustration du fonctionnement de l'algorithme au sein d'un nœud h est donnée par la figure (3.12) ci-dessous :

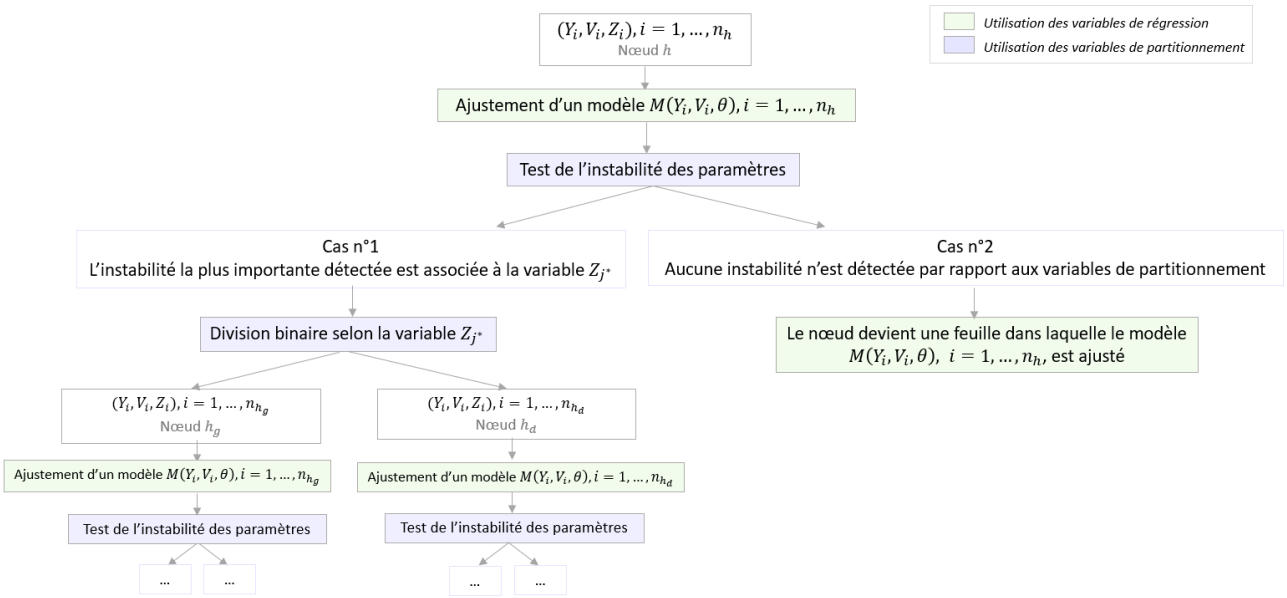


FIGURE 3.12 : Schéma simplifié du fonctionnement de l'algorithme de partitionnement récursif au sein d'un nœud h

Dans le cadre de ce mémoire il est choisi d'étudier les arbres pour lesquels le modèle ajusté dans chaque nœud est un modèle GLM.

3.2.3 Arbres de régression GLM

Cette partie décrit un cas particulier de l'algorithme présenté en section précédente : celui où le modèle ajusté dans chacun des nœuds terminaux est un modèle linéaire généralisé [Rusch et Zeileis (2013)]. Dans ce cas, la fonction objectif Ψ est la log-vraisemblance associées aux observations

$$\Psi = -\mathcal{L}, \qquad \text{avec } \mathcal{L}(Y,\theta) = \sum_{i=1}^{n} \frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i,\phi).$$

Un des avantages de cette méthode est qu'elle permet de conserver la simplicité d'interprétation des modèles linéaires généralisés, tout en permettant, via la structure arborescente, d'intégrer des effets non-linéaires pour les variables explicatives. Cela peut en particulier permettre de révéler des structures (ou *pattern* en anglais) sous-jacentes dans les données, qui ne peuvent pas être détectées par un simple modèle GLM. D'autre part, la spécification d'un modèle paramétrique dans les feuilles de l'arbre plutôt qu'une constante permet d'obtenir une meilleure stabilité de l'arbre. Cela permet en particulier de palier à l'une des limites des arbres CART vue précédemment : les arbres CART ne peuvent pas prédire d'autres valeurs que les constantes ajustées dans chaque feuille lors de la phase d'apprentissage. L'intégration d'un modèle GLM dans chaque feuille permet de palier à cette limite, puisque chaque feuille n'est plus associée à une constante mais à un ensemble de coefficients permettant l'application d'un modèle GLM. Ce type de modèle bénéficie toujours d'une bonne capacité de visualisation et d'interprétation, ce qui est un atout non-négligeable compte tenu du développement de nombreux modèles "boites noires" au cours des dernières années.

3.2.4 Application du modèle MOB

Pour appliquer le modèle MOB, il faut tout d'abord déterminer la formule avec laquelle va être appliqué le modèle, *i.e.* il faut choisir les variables de partitionnement et de régression du modèle. Avec M variables explicatives, il existe

$$\sum_{m=1}^{M-1} \binom{M}{m} = \sum_{m=1}^{M-1} \frac{M!}{(M-m)!m!},$$

combinaisons possibles. Le cas où $m = 1$ correspond au cas où il est choisi d'utiliser une seule variable de partitionnement et le reste en variables de régression. En utilisant la formule du binôme de Newton, la simplification suivante peut être obtenue

$$\sum_{m=1}^{M-1} \binom{M}{m} = \sum_{m=0}^{M} \binom{M}{m} 1^m 1^{M-m} - \binom{M}{0} - \binom{M}{M} = (1+1)^M - 1 - 1 = 2^M - 2,$$

Ainsi, avec M variables explicatives, il existe $2^M - 2$ combinaisons possibles. Dans le cadre de ce mémoire, les modèles sont entraînés sur $M = 7$ variables explicatives, ainsi 126 formules sont à tester pour déterminer quelles variables seront utilisées pour le partitionnement et la régression.
Tout comme pour le modèle CART, il est possible pour le modèle MOB de spécifier une profondeur maximale de l'arbre construit. En effet, comme expliqué précédemment, il n'est pas souhaitable d'obtenir un nombre de groupes créés par le modèle qui soit trop important.
Afin de déterminer la formule à utiliser, ainsi que la profondeur maximale à spécifier, une validation croisée sur 5 blocs est effectuée. L'idée est à nouveau de choisir une formule et une profondeur maximale qui soient optimales de façon générale et pas seulement pour convenir au mieux à l'échantillon

de validation. Les valeurs de profondeur maximale testées sont 4 et 5, permettant respectivement de créer au maximum $2^4 = 16$ et $2^5 = 32$ groupes. Ainsi, pour un bloc de test $k \in \{1, ..., 5\}$, le modèle MOB est entraîné sur les blocs d'apprentissage pour chacune des 126 formules en spécifiant l'une des deux profondeurs maximales, puis les métriques sont évaluées sur le bloc k. La recherche de la formule et de la profondeur maximale optimales par validation croisée implique donc d'appliquer le modèle MOB $5 \times 2 \times 126 = 1260$ fois. Les métriques sont ensuite agrégées en prenant pour chaque couple $(formule,\ profondeur\ maximale)$ la moyenne des métriques obtenues sur chaque bloc. À chaque étape le modèle MOB est entraîné sur les blocs d'apprentissage en prenant pour le GLM les mêmes paramètres qu'en section (2.1), à savoir une famille Gamma avec la fonction de lien log.

Une restriction doit être imposée sur les arbres construits : lors de la phase d'apprentissage chaque feuille de l'arbre doit contenir l'ensemble des modalités présentées par chacune des variables de régression. En effet, considérons une variable V qualitative possédant M modalités notées $\{m_1, ..., m_M\}$ utilisée comme variable de régression et supposons que l'arbre construit possède une feuille F dans laquelle aucune observation de l'échantillon d'apprentissage possédant la modalité m_j pour un $j \in \{1, ..., M\}$ ne soit présente. Un modèle GLM est entraîné dans la feuille F sur les observations de l'échantillon d'apprentissage associées à cette feuille. Le modèle GLM ainsi entraîné ne présentera aucun coefficient pour la modalité m_j de la variable V. Or, il se peut qu'en appliquant l'arbre construit à l'échantillon de validation une observation possédant la modalité m_j pour la variable V soit envoyée dans la feuille F. Le modèle GLM fournira alors des prédictions erronées pour cette observation, ce qui n'est pas souhaitable.

La validation croisée présentée précédemment a été effectuée, avec un temps de calcul d'environ 21 heures, et a conduit à choisir une profondeur maximale égale à 4, avec pour formule optimale

$$\texttt{coutSinistre} \sim \underbrace{\texttt{GroupeKm + GroupeAge + typeSinistre + expertA}}_{\text{Variables de régression}} \Big| \underbrace{\begin{matrix}\texttt{CategorieMarque} \\ \texttt{tauxHoraire + ASSUREUR}\end{matrix}}_{\text{Variables de partitionnement}} \tag{3.9}$$

Ainsi, les observations sont partitionnées selon la catégorie de marque du véhicule sinistré, le taux horaire du garage intervenu et l'assureur en charge du sinistre, puis, dans chacune des feuilles de l'arbre, un modèle GLM est entraîné selon le réseau d'experts intervenu, les groupes d'âge et de kilométrage du véhicule et le type de sinistre. En particulier, la variable représentant le réseau d'experts n'intervient pas dans le partitionnement des individus mais en tant que variable de régression dans chacune des feuilles de l'arbre construit.

L'arbre construit par MOB est représenté sur la figure (3.13) ci-dessous. Pour chaque feuille, le nombre d'observations contenues dans la feuille, le coût moyen associé et le coefficient associé au réseau d'experts, calculé par le modèle GLM entraîné au sein de la feuille, ont été représentés.
L'arbre construit possède 7 nœuds internes et 8 feuilles qui ont été numérotées de 1 à 8. Ainsi, le modèle contient 8 sous-modèles GLM ajustés selon les variables suivantes : le réseau d'experts, le groupe de kilométrage et d'âge du véhicule et le type de sinistre. Ces quatre variables regroupent un total de 16 modalités (4 pour le kilométrage, 4 pour l'âge du véhicule, 6 pour le type de sinistre et 2 pour le réseau d'expert), le GLM estimant pour une variable à m modalités $m - 1$ coefficients, un total de 13 coefficients sont estimés en chaque feuille de l'arbre (12 coefficients pour les variables de régression, plus un coefficient associé à l'intercept). Cet arbre est alors utilisé pour effectuer des prédictions.

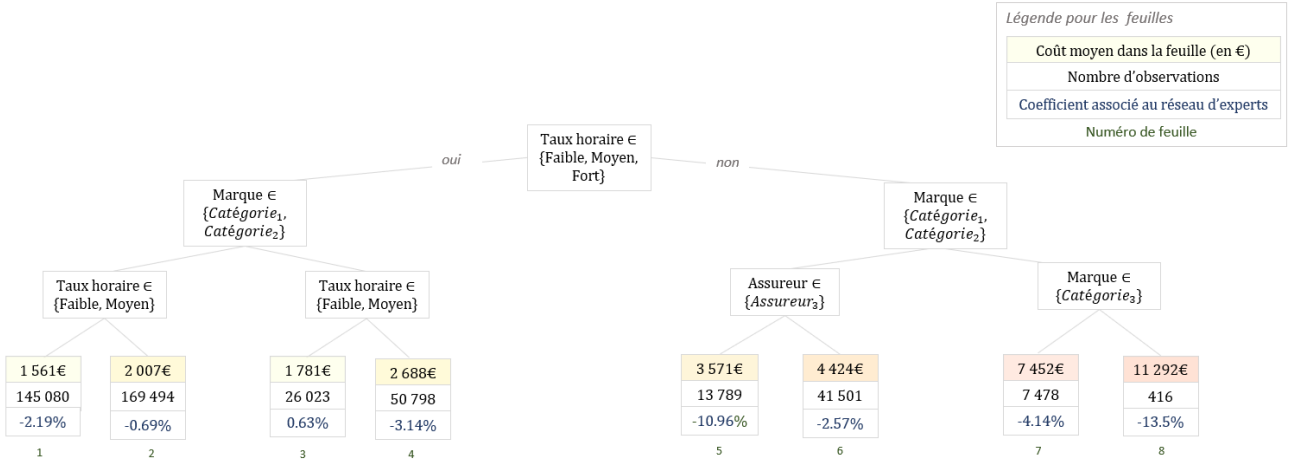


FIGURE 3.13 : Arbre construit par MOB

Pouvoir prédictif du modèle

Le modèle MOB est utilisé pour effectuer des prédictions sur l'échantillon de validation. Les métriques d'évaluation suivantes ont été obtenues :

MSE	MAE
3 799 924	1 202

TABLE 3.4 : Résultats de prédiction sur l'échantillon de validation pour le modèle MOB

Les résultats de prédiction obtenus avec MOB sont meilleurs que ceux obtenus avec les modèles CART et GLM.
Afin de s'assurer que le modèle MOB n'est pas plus performant que le modèle GLM seulement parce qu'il utilise un ensemble différent de variables de régression, un modèle GLM avec pour variables explicatives les variables de régression du modèle MOB a été étudié. Les résultats de prédiction issus de ce modèle GLM, associé à la formule `coutSinistre` $\sim$ `GroupeKm + GroupeAge + typeSinistre + expertA`, sont présentés dans le tableau (3.5) ci-dessous :

MSE	MAE
6 212 430	1 508

TABLE 3.5 : Résultats de prédiction sur l'échantillon de validation pour le modèle GLM utilisant pour variables explicatives les variables de régression du modèle MOB

Ces résultats permettent d'observer que ce n'est pas seulement l'ensemble des variables de régression sélectionnées qui permet au modèle MOB d'obtenir de meilleurs résultats de prédiction : le partitionnement des individus est nécessaire pour garantir une telle performance prédictive.

Tout comme pour les modèles GLM et CART une analyse des moyennes des coûts observés et prédits est effectuée. Tout d'abord la moyenne des coûts prédits par le modèle MOB est égale à environ 2 294€, ce qui est similaire à ce qui avait été observé pour CART. Cela montre notamment que l'écart entre les moyennes des sinistres prédits et observés n'est pas une mesure suffisante pour hiérarchiser les performances des modèles car, comme vu précédemment, le modèle MOB possède un

plus grand pouvoir de prédictif que le modèle CART. Cependant, l'étude de ce type d'écart au sein des différentes modalités des variables explicatives permet de donner une idée des erreurs commises par le modèle, tout en garantissant d'effectuer des comparaisons entre des objets de même nature. Une représentation graphique des moyennes des sinistres observés et prédits par MOB pour chaque modalité des variables explicatives est donnée par la figure (3.14) ci-dessous :

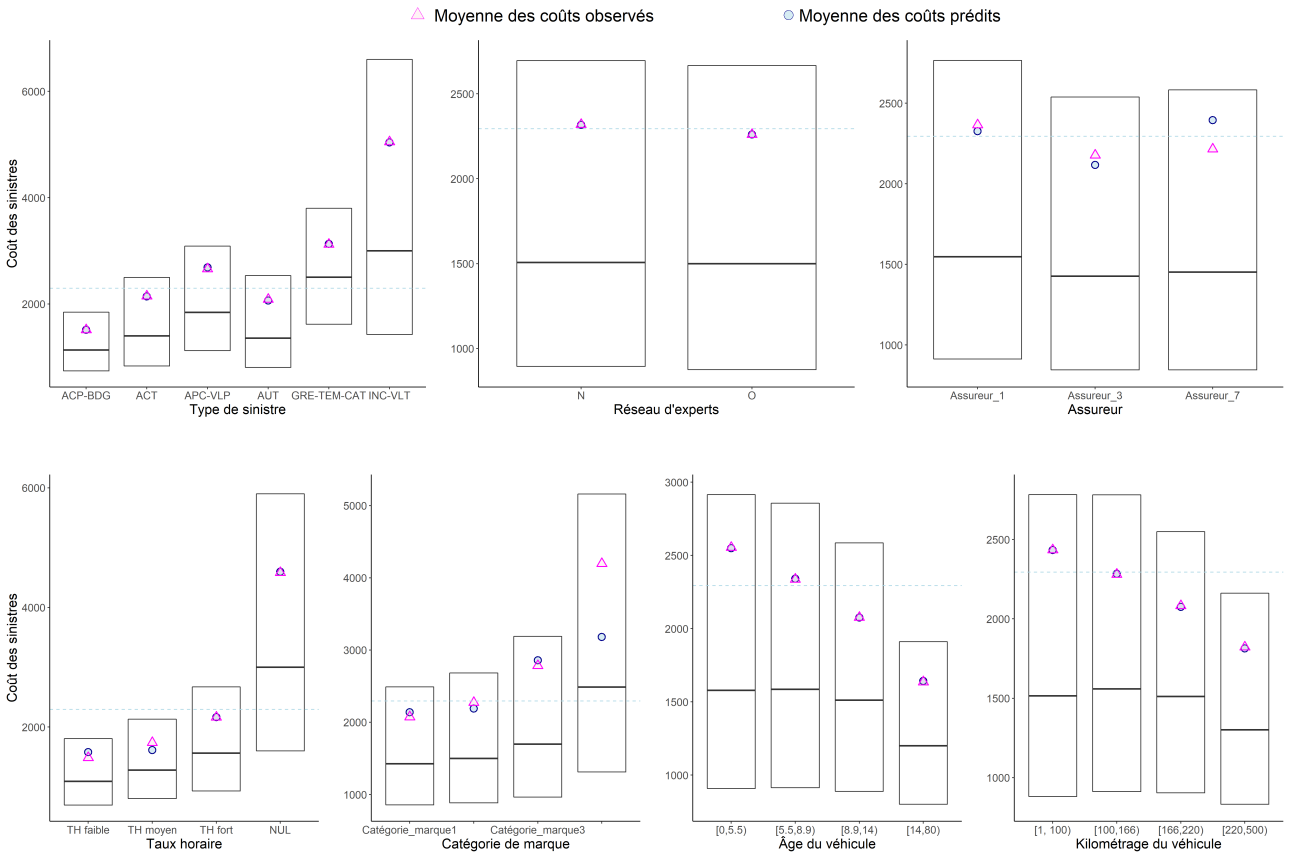


FIGURE 3.14 : Moyennes des sinistres observés et prédits par MOB pour chaque modalité des variables explicatives

Des diminutions d'écarts entre les moyennes des sinistres observés et prédits peuvent être soulignées par rapport à ce qui avait été observé pour CART en figure (3.7) *supra*. En effet, pour les variables représentant le type de sinistre, les groupes d'âge ou de kilométrage du véhicule, les écarts qui avaient pu être observés précédemment avec le modèle CART sont très fortement réduits. Cependant, la sous-estimation moyenne relativement importante des coûts pour la catégorie de marque n°4 est toujours existante et est même plus importante qu'avec le modèle CART. De même, la surestimation moyenne des coûts pour l'assureur n°7 est encore présente avec le modèle MOB, bien qu'elle soit moins importante que la surestimation induite par le modèle CART pour cet assureur.

Tout comme pour CART, il peut être intéressant de regarder l'erreur commise dans chacune des feuilles afin d'identifier d'éventuels groupes pour lesquels cette erreur est particulièrement importante. Une représentation graphique de l'erreur par feuille, en fonction du nombre d'observations contenues dans la feuille est donnée par la figure (3.15) ci-dessous. Chaque point de la figure (3.15) représente une feuille et la taille de chaque point correspond au poids de la feuille dans l'erreur totale tel que défini par la formule (3.5) *supra*. L'erreur totale commise par le modèle sur l'échantillon de validation est représentée par la ligne pointillée rouge.

Les erreurs les plus importantes sont produites au sein des “petites” feuilles, cependant, du fait du nombre faible d'observations présentes dans ces feuilles, le poids de ces erreurs dans l'erreur totale est assez faible.

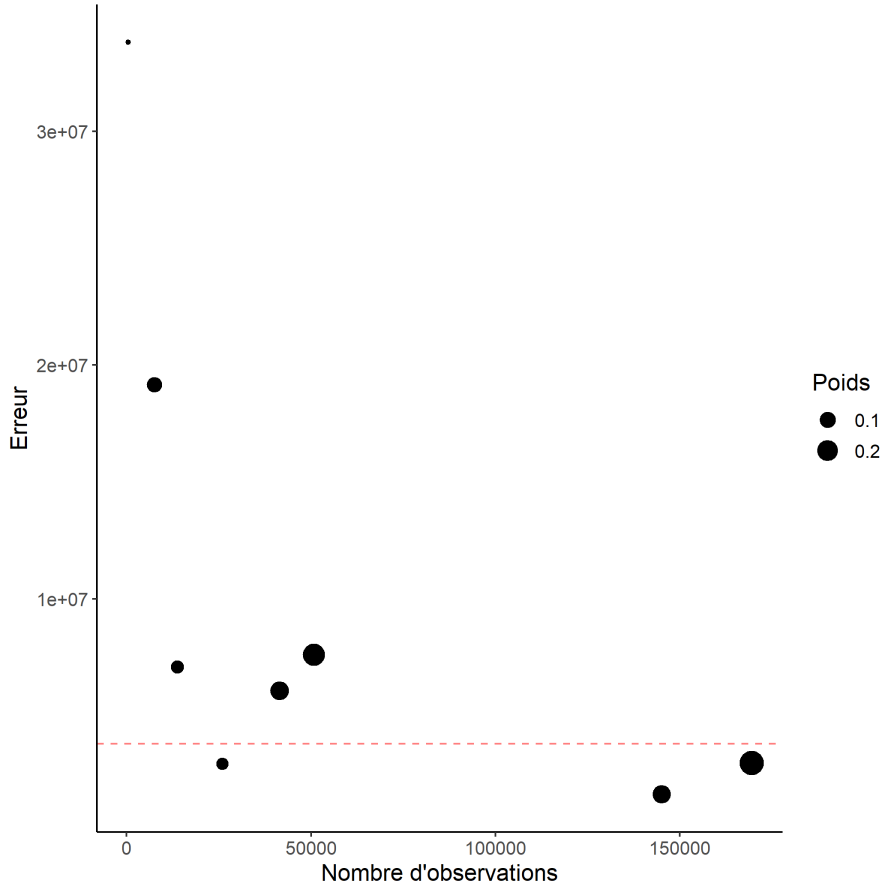


FIGURE 3.15 : Erreur en fonction du nombre d'observations dans chaque feuille issue de MOB

Tout comme CART, MOB permet de spécifier le nombre minimum d'observations souhaité dans chaque feuille de l'arbre. La valeur par défaut de ce paramètre est le nombre de paramètres à estimer par feuilles multiplié par 10. Elle est donc égale à $10 \times 13 = 130$ pour le modèle MOB utilisé. Afin de déterminer un nombre minimum d'observations optimal une validation croisée en 5 blocs est effectuée sur l'échantillon d'apprentissage pour différentes valeurs de minimum. Cette validation croisée est réalisée selon le même principe que celui illustré en figure (3.2) *supra*. Les valeurs de minimum $\{130, 500, 5000, 10000, 25000, 50000\}$ ont été testées et les résultats obtenus sont présentés ci-dessous en figure (3.16). La taille des arbres construits a également été représentée.
Il est observé que spécifier un nombre minimum d'observations supérieur à la valeur par défaut (représentée en vert) ne permet pas d'améliorer les résultats de prédiction. Il est donc choisi de ne pas modifier la valeur de ce paramètre.

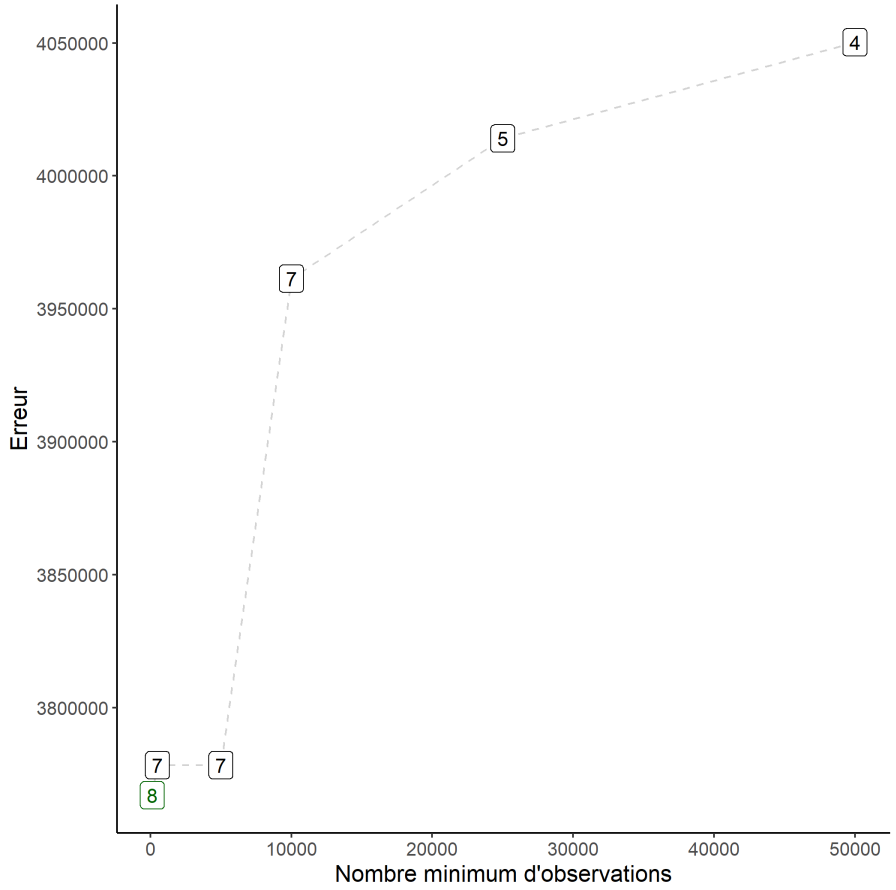


FIGURE 3.16 : Erreur des moindres carrés pour différentes valeurs de minimum d'observations à avoir dans chaque feuille de l'arbre MOB

Le nombre faible de groupes construits par l'arbre MOB permet une lecture aisée de l'influence des variables de régression au sein de chacun d'entre eux.

Mesure de l'influence du réseau d'expert

La variable spécifiant le réseau d'experts intervenu est utilisée comme variable de régression. Ainsi, huit estimations du coefficient de l'influence du réseau d'experts sur le coût du sinistre sont données par l'arbre MOB. La figure (3.17) ci-dessous représente les coefficients estimés dans chaque feuille (sur l'échantillon d'apprentissage) ainsi que les intervalles de confiance qui leurs sont associés. Le numéro de la feuille associée au coefficient et le coefficient associé au modèle GLM présenté dans le chapitre 2, égal à -2.9%, ont également été représentés. Le coefficient estimé est négatif dans sept des huit feuilles créées par l'arbre et l'ensemble des individus pour lesquels le coefficient estimé est positif, *i.e.* les individus de l'échantillon d'apprentissage appartenant à la feuille numéro 3, représente moins de 6% de l'ensemble des individus de l'échantillon d'apprentissage. La représentation des intervalles de confiance permet d'observer que pour ce coefficient négatif, il existe une incertitude sur le signe de l'estimation. Il est également possible d'observer une croissance de la taille des intervalles de confiance lorsque le nombre d'observations diminue. En particulier, il existe une incertitude assez importante sur la valeur du coefficient estimé au sein de la feuille numéro 8, ce qui est dû au nombre relativement faible d'observations contenues dans la feuille (416 observations).

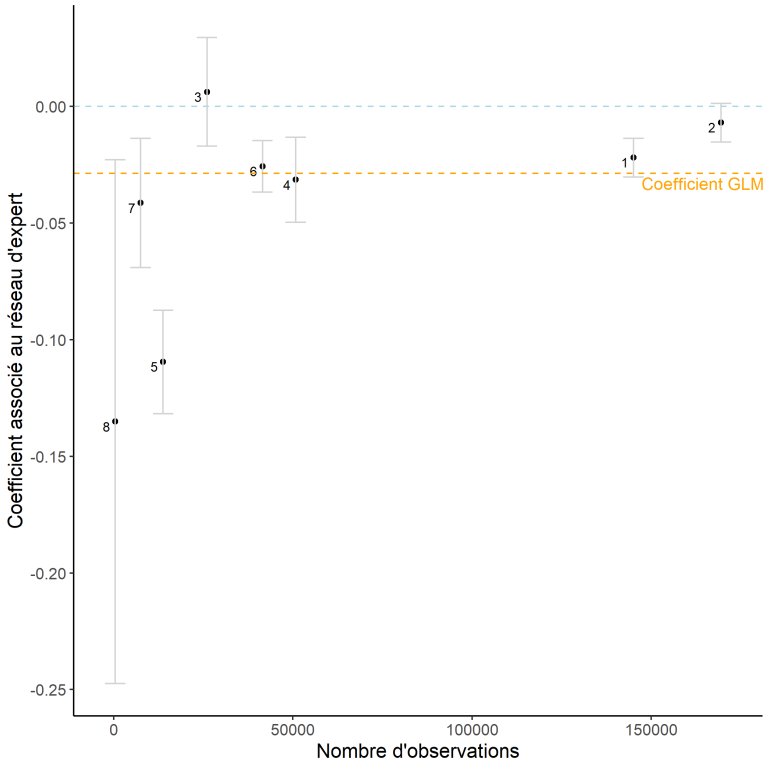


FIGURE 3.17 : Coefficient associé au réseau d'experts en fonction du nombre d'observations par feuille

Ainsi le modèle MOB permet une étude aisée de l'influence du réseau d'experts sur le montant des coûts de sinistres. De plus, ce modèle permet d'avoir plusieurs effets différents selon les caractéristiques des individus, plutôt qu'un unique effet moyen comme avec un GLM.

Ainsi, parmi les trois modèles étudiés précédemment, le modèle MOB présente les meilleurs résultats en termes de prédiction sur l'échantillon de validation. Avant de procéder à une comparaison de ces modèles sur l'échantillon de test, le modèle XGBoost est introduit afin de confronter les performances obtenues précédemment à celles d'un modèle connu pour son pouvoir de prédiction, mais aussi afin d'essayer d'améliorer le pouvoir prédictif des modèles précédents en s'inspirant du modèle XGBoost.

3.3 Utilisation d'un modèle "boîte noire" pour améliorer les performances : le modèle XGBoost

Les modèles introduits précédemment possèdent l'atout non négligeable d'être facilement interprétables. Leur pouvoir de prédiction est cependant concurrencé par certains modèles de *machine learning* tels que les forêt aléatoires, les machines à vecteurs supports ou encore le *Boosting*. Ces modèles possèdent généralement des pouvoirs de prédiction importants, en contrepartie d'une perte d'interprétabilité qui leur vaut la caractérisation de modèles "boîte noire".
Cette section présente le modèle XGBoost (pour *eXtreme Gradient Boosting* en anglais) introduit par Tianqi Chen et Carlos Guestrin lors d'un projet de recherche en 2016 ([CHEN et GUESTRIN (2016)]). Une présentation du modèle XGBoost est donnée en annexe (C). Ce modèle a fait preuve de succès

lors de nombreuses compétitions de *machine learning* grâce à son pouvoir prédictif élevé. Cependant, cette puissance de prédiction a un coût : le modèle XGBoost est difficile à comprendre et à interpréter. Cela est notamment dû au fait que le modèle XGBoost est basé sur une méthode d'ensemble : plusieurs sous-modèles sont combinés pour obtenir une prédiction finale. Cela rend la lisibilité et l'interprétation des résultats complexes, ce qui est un inconvénient non négligeable du modèle.
L'objectif de cette section est de comparer les résultats des modèles "boite blanche" présentés précédemment avec ceux du modèle XGBoost. L'idée est alors d'analyser au mieux ce modèle dans le but de trouver des moyens d'améliorer les prédictions des modèles "boite blanche", en particulier les résultats du modèle présentant le plus grand pouvoir prédictif, à savoir le modèle MOB.

3.3.1 Application du modèle XGBoost

Avant d'appliquer le modèle XGBoost à l'ensemble de la base d'apprentissage, une validation croisée à été effectuée afin d'optimiser les performances du modèle. Pour cela différents paramètres ont été étudiés :
— le nombre d'arbres construits n_{arbres}, *i.e.* le nombre d'apprenants faibles utilisés ;
— la profondeur maximale des arbres construits max_{prof} ;
— le taux d'apprentissage η, permettant de régler le poids associé aux apprenants faibles ;
— le paramètre de pénalisation γ, permettant de régler la pénalisation des arbres de grandes tailles ;
— le paramètre de régularisation λ.

Les valeurs testées pour ces paramètres sont explicitées dans le tableau (3.6) ci-dessous :

n_{arbres}	10	50	100	500	**1000**
max_{prof}	4	**6**	10		
η	**0.01**	0.2	0.3		
γ	**0**	0.5	1		
λ	**0**	0.5	1		

TABLE 3.6 : Valeurs testées lors de la validation croisée pour optimiser le modèle XGBoost

Les valeurs des paramètres sélectionnées après la validation croisée sont celles écrites en gras dans le tableau (3.6). Le modèle est appliqué à l'ensemble de l'échantillon d'apprentissage avec ces paramètres, puis est utilisé pour prédire les coûts de l'échantillon de validation. Les résultats obtenus sont présentés dans le tableau (3.7) ci-dessous :

MSE	MAE
3 706 014	1 186

TABLE 3.7 : Évaluation des métriques pour XGBoost sur l'échantillon de validation

Les résultats de prédiction sont meilleurs que ceux obtenus avec les modèles GLM, CART et MOB. Plus précisément, les améliorations suivantes sont observées :

	Amélioration fournie par XGBoost	
	MSE	MAE
par rapport au GLM	24%	9%
par rapport à CART	12%	7%
par rapport à MOB	3%	1%

TABLE 3.8 : Amélioration des métriques fournie par le modèle XGBoost par rapport aux modèles GLM, CART et MOB

Bien que l'amélioration des métriques entre les modèles XGBoost et MOB puisse paraître relativement faible, elle n'est pas sans conséquence. En effet, dans un contexte de concurrence accrue toute optimisation du résultat a son importance et n'est donc pas à négliger.
La moyenne des prédictions associées au modèle XGBoost est égale à environ 2 294€, tout comme pour les modèles CART et MOB. Une représentation graphique des moyennes des sinistres observés et prédits par le modèle XGBoost est donnée par la figure (3.18) ci-dessous :

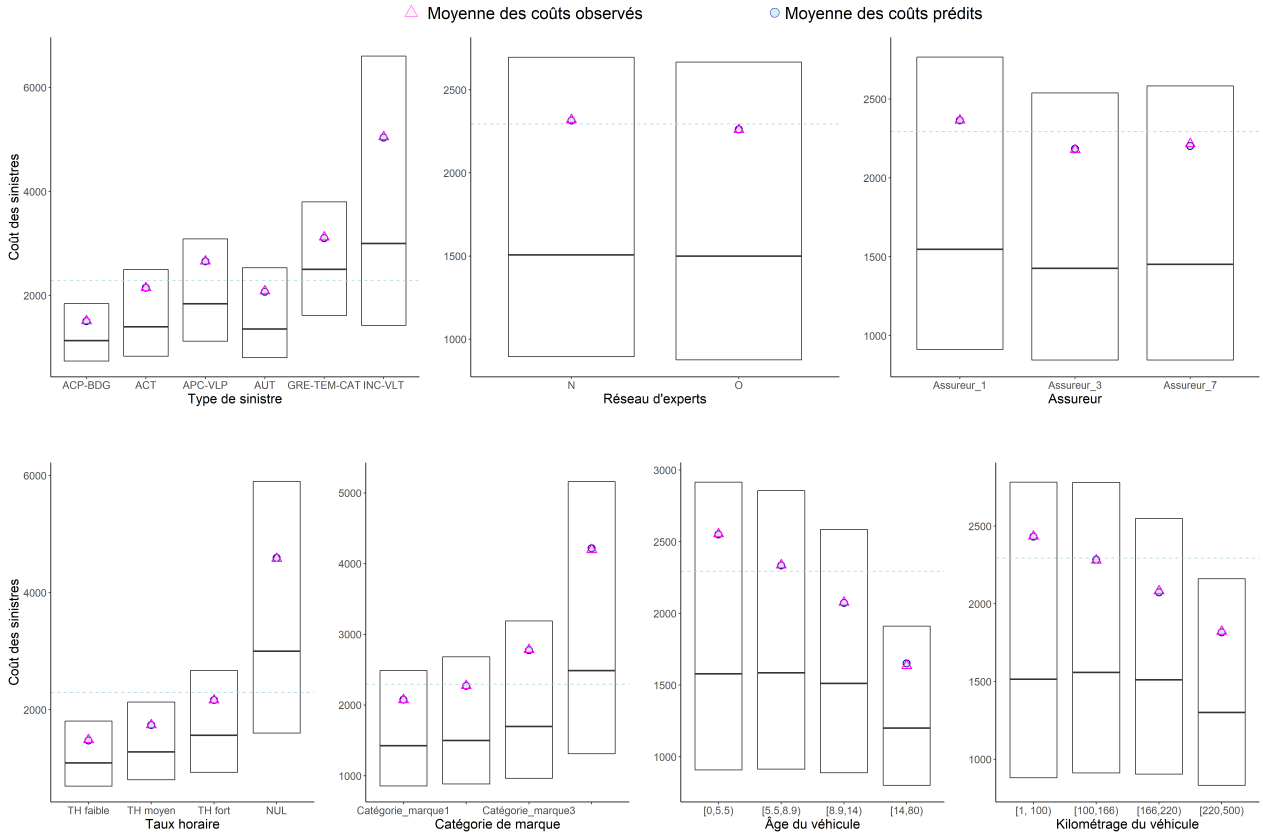


FIGURE 3.18 : Moyennes des sinistres observés et prédits par XGBoost pour chaque modalité des variables explicatives

La représentation graphique (3.18) ci-dessus témoigne d'une amélioration fournie par le modèle XGBoost sur la réduction des écarts entre les moyennes des sinistres observés et prédits par rapport aux modèles étudiés précédemment. Les écarts sont relativement faibles pour toutes les modalités de l'ensemble des variables explicatives de la base.

Au vu de l'objectif de ce mémoire, le modèle XGBoost ne semble pas particulièrement adapté

puisque la mesure de l'influence du réseau d'experts intervenu est plus complexe que dans les modèles "boîte blanche" présentés précédemment. Cependant, l'étude du modèle XGBoost peut être intéressante afin d'améliorer les résultats de ces modèles tout en conservant leur facilité d'interprétation. En particulier, une analyse du modèle XGBoost est effectuée afin d'essayer d'améliorer les résultats de prédiction du modèle MOB. L'idée de cette section est d'analyser comment le modèle XGBoost utilise les effets des variables explicatives pour la prédiction et d'essayer d'incorporer ces effets dans des modèles de type "boîte blanche", *i.e.* des modèles facilement interprétables. Pour ce faire, des outils d'analyse agnostiques sont utilisés.

3.3.2 Présentation des outils d'analyse agnostiques

Afin de pouvoir comparer comment les modèles utilisent les variables explicatives, des outils d'analyse agnostiques vont être utilisés afin d'assurer la pertinence des comparaisons. Les outils d'analyse agnostiques s'opposent aux outils d'analyse intrinsèques, qui sont inhérents aux modèles. Par exemple, pour calculer l'importance des variables les modèles CART et XGBoost possèdent des outils de calcul qui sont intégrés aux modèles et propres à ces derniers. Pour pouvoir comparer les importances de variables dans chacun des modèles, il est plus pertinent d'utiliser un outil agnostique, *i.e.* extérieur aux modèles, qui utilisera la même mesure d'importance pour tous les modèles. Les outils agnostiques utilisés dans cette partie sont issus du package flashlight ([Mayer (2021)], [Lorentzen et Mayer (2020)]) du logiciel R.

Importance des variables

Une façon agnostique d'obtenir l'importance des variables est de calculer l'importance par permutation (*permutation importance* en anglais). Pour calculer l'importance par permutation d'une variable explicative X, l'idée est de mélanger aléatoirement les valeurs de la variable X au sein de la base d'étude. Cette permutation aléatoire permet de casser le lien potentiel existant entre la variable explicative et la variable à expliquer. L'importance de la variable se calcule en observant la perte de performance de prédiction du modèle lorsque les valeurs de la variable ont été mélangées aléatoirement. Plus la décroissance de performance est importante, plus l'importance de la variable est haute pour le modèle. Afin de mieux comprendre comment les variables sont utilisées, il est également intéressant d'étudier les interactions entre variables explicatives prises en compte au sein des différents modèles.

Interactions entre les variables

Une façon agnostique de mesurer l'interaction entre les variables est donnée par la H-statistique de Friedman. Avant de l'introduire les notions de dépendance partielle et d'espérance conditionnelle individuelle sont définies.

Espérance conditionnelle individuelle

Cette quantité est souvent appelée ICE pour *Individual Conditional Expectation* en anglais. L'idée est d'étudier les profils ICE de différentes observations. Pour une variable explicative X les profils ICE montrent comment la prédiction pour une observation i est modifiée lorsque la variable X change de valeur. La figure (3.19) ci-dessous montre comment le profil ICE pour une variable X et une observation i est construit :

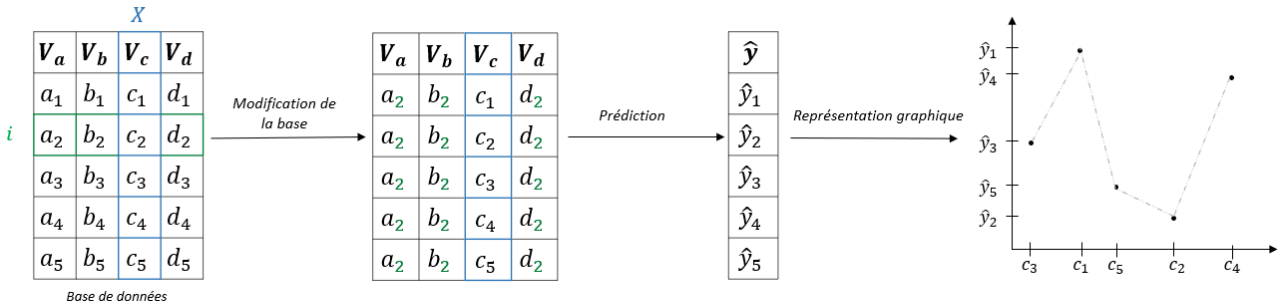


FIGURE 3.19 : Schéma simplifié des étapes de la construction d'un profil ICE pour une variable X et une observation i

La représentation graphique des profils ICE permet de donner une indication de si la variable X possèdent des effets d'interaction avec d'autres variables de la base d'étude : plus il y a un effet d'interaction fort entre la variable X et une ou plusieurs variables, plus les formes des représentations graphiques des profils ICE de différentes observations diffèrent. Cependant, une telle représentation graphique ne permet pas de mettre en avant avec quelle(s) variable(s) l'interaction se produit.

Les profils de dépendance partielle
Ces profils sont souvent notés PDP pour *Partial Dependance Profile* en anglais. Ils sont obtenus en prenant la moyenne de plusieurs profils ICE et peuvent être vus comme l'effet principal (*main effect* en anglais) de la variable X. Les profils de dépendance partiels sont utilisés pour calculer la H-statistique de Friedman.

La H-statistique de Friedman est utilisée pour mesurer la force d'interaction entre deux variables explicatives X_j et X_k. Elle est définie par

$$H_{jk}^2 = \frac{\sum_{i=1}^{n}[PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)})]^2}{\sum_{i=1}^{n} PD_{jk}^2(x_j^{(i)}, x_k^{(i)})},$$

où PD_j, PD_k, PD_{jk} désignent les profils de dépendance partielle centrés et réduits de dimension 1 et 2 pour les variables X_j et X_k. La dénotation (i) fait référence à la valeur du profil de dépendance partielle de l'observation i. La statistique H^2 mesure la proportion de variabilité dans l'effet commun de X_j et X_k qui n'est pas expliqué par leur effet principal. Si H^2 est proche de 0 alors cela signifie qu'il n'y a presque pas d'interaction entre X_j et X_k. À l'inverse, si H^2 est proche de 1 alors cela signifie que la plupart des effets des variables proviennent de leur interaction. La statistique H^2 mesure la force d'interaction relative à l'effet joint total de X_j et X_k.
Il est possible d'obtenir de l'information supplémentaire en considérant la racine carré du numérateur de H^2

$$\tilde{H}_{jk} = \sqrt{\sum_{i=1}^{n}[PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)})]^2},$$

qui donne une mesure absolue de la force d'interaction. Cette statistique peut être utilisée pour trouver quelles sont les variables possédant les plus grandes forces d'interaction absolues.

Ces outils agnostiques sont appliqués aux quatre modèles étudiés dans ce mémoire, afin de trouver des moyens d'améliorer le pouvoir de prédiction des modèles "boîte blanche" tout en conservant leur facilité d'interprétation.

3.3.3 Amélioration des modèles "boîte blanche"

L'idée est d'utiliser les outils présentés précédemment et de les comparer entre le modèle XGBoost et les modèles "boîte blanche" GLM, CART et MOB. En particulier, le modèle MOB présentant les meilleurs résultats en termes de prédiction parmi ces trois modèles, l'analyse se concentrera principalement sur une comparaison entre ce modèle et le modèle XGBoost.

L'importance par permutation a été calculée pour chacune des sept variables explicatives au sein des quatre modèle étudiés dans ce mémoire. Les importances obtenues sont représentées sur la figure (3.20) ci-dessous :

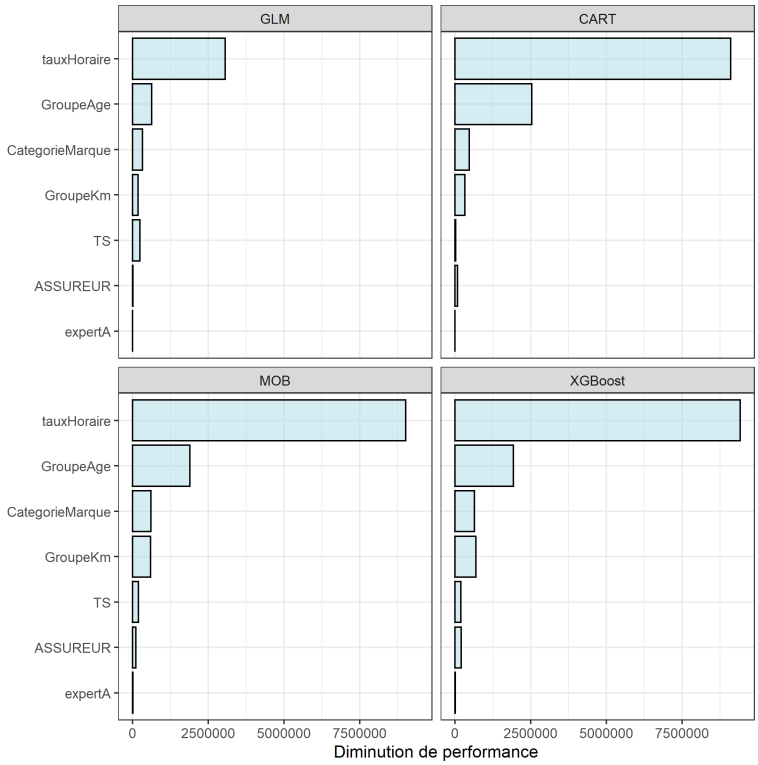


FIGURE 3.20 : Importance par permutation des variables explicatives pour les modèles GLM, CART, MOB et XGBoost

L'ordre d'importance des variables[1] est globalement le même pour tous les modèles avec une domination de la variable `tauxHoraire` pour l'ensemble des modèles. Cependant il est possible d'observer que cette variable est moins importante dans le modèle GLM qu'elle ne l'est dans les autres modèles. Des différences d'ordre d'importance peuvent être soulignées par rapport à ce qui avait été observé dans le tableau (3.3) *supra* pour le modèle CART. Cela souligne l'intérêt d'utiliser des outils agnostiques afin de garantir la pertinence des comparaisons, puisque selon la formule choisie pour caractériser l'importance d'une variable des résultats différents peuvent être obtenus.
L'importance des variables est sensiblement la même pour les modèles MOB et XGBoost. Cet outil ne permet donc pas d'identifier une façon d'améliorer les résultats de prédiction du modèle MOB. Une étude des interactions entre les variables est alors effectuée.

[1]Pour rappel, la notation `TS` fait référence à la variable représentant le type de sinistre.

La figure (3.21) ci-dessous présente les H-statistiques de Friedman relatives et absolues pour les modèles CART, MOB et XGBoost :

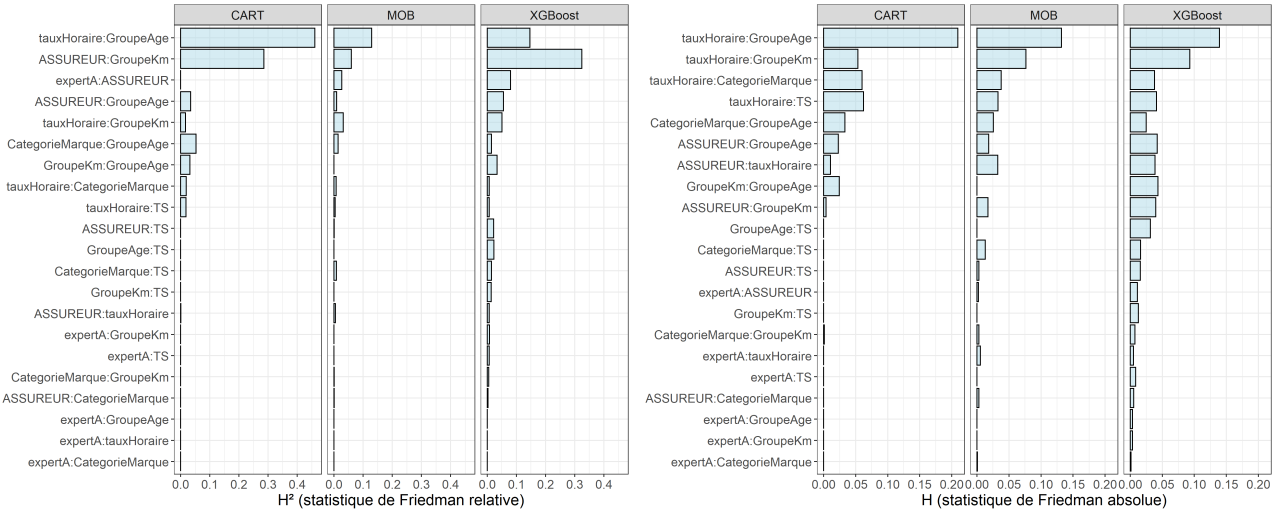


FIGURE 3.21 : Interactions entre les variables explicatives pour les modèles CART, MOB et XGBoost

Les statistiques ont été calculées à partir des profils de dépendance partiels sur l'échelle de la fonction de lien, *i.e.* l'échelle logarithmique. Le modèle GLM n'a pas été représenté car, ce modèle n'intégrant aucune interaction entre variables, les H-statistiques de Friedman relatives et absolues sont nulles pour ce modèle. Des différences de forces d'interactions peuvent être observées entre les modèles MOB et XGBoost. Ces différences peuvent en particulier être source des écarts de performances observés entre les deux modèles. Par exemple, il est possible d'observer une différence de forces d'interactions relatives entre XGBoost et MOB pour les variables `ASSUREUR` et `GroupeKm` relativement importante. En effet, même si la variable `ASSUREUR` est utilisée comme variable de partitionnement dans le modèle MOB, le fait de s'être restreint à des arbres de profondeur maximale égale à 4 empêche la création de partition(s) supplémentaire(s) selon cette variable. Plus précisément, les modalités possibles pour chacune des variables de partitionnement dans chaque feuille de l'arbre MOB sont résumées dans le tableau (3.9) ci-dessous :

	F1	F2	F3	F4	F5	F6	F7	F8
`ASSUREUR`	{1,3,7}	{1,3,7}	{1,3,7}	{1,3,7}	{3}	{1,7}	{1,3,7}	{1,3,7}
`CategorieMarque`	{1,2}	{1,2}	{3,4}	{3,4}	{1,2}	{1,2}	{3}	{4}
`tauxHoraire`	{Faible, Moyen}	{Fort}	{Faible, Moyen}	{Fort}	{NUL}	{NUL}	{NUL}	{NUL}

TABLE 3.9 : Modalités possibles pour les variables de partitionnement dans chacune des feuilles de l'arbre MOB

Le tableau (3.9) permet d'observer que pour de nombreuses feuilles les restrictions sur les variables de partitionnement ne sont pas réduites à une unique modalité de ces dernières. Il est alors intéressant d'observer si l'ajout d'interaction(s) au sein du modèle MOB entre les variables de régression ou entre les variables de partitionnement et de régression permettrait d'en améliorer les performances prédictives.

Pour décider quelle(s) interaction(s) ajouter au modèle MOB, il est calculé pour chaque couple de variables les différences de forces d'interactions absolues et relatives entre le modèle XGBoost et le modèle MOB. Il a ensuite été choisi d'associer à chaque couple de variables un score égal à la somme entre les différences d'interactions absolues et relatives des variables du couple. Plus précisément, le score calculé pour deux variables V_1 et V_2 est donné par

$$score(V_1, V_2) = \underbrace{(int^{rel}_{XGB}(V_1, V_2) - int^{rel}_{MOB}(V_1, V_2))}_{\text{différence d'interactions relatives}} + \underbrace{(int^{abs}_{XGB}(V_1, V_2) - int^{abs}_{MOB}(V_1, V_2))}_{\text{différence d'interactions absolues}}, \quad (3.10)$$

où $int^{rel}_{mod}(V_1, V_2)$ et $int^{abs}_{mod}(V_1, V_2)$ désignent respectivement les H-statistiques de Friedman relatives et absolues entre les variables V_1 et V_2 associées au modèle *mod* et où la notation XGB fait référence au modèle XGBoost.
Les cinq couples de variables possédant le score ainsi calculé le plus élevé sont présentés ci-dessous dans le tableau (3.10) :

	Score
`ASSUREUR:GroupeKm`	0.29
`GroupeKm:GroupeAge`	0.08
`ASSUREUR:GroupeAge`	0.07
`ASSUREUR:expertA`	0.06
`GroupeAge:typeSinistre`	0.05

TABLE 3.10 : Les cinq couples de variables possédant le score le plus élevé

L'ajout d'interaction(s) se fait par ordre décroissant du score : l'ajout du couple de variables possédant le score plus élevé est étudié et si celui-ci fait diminuer les métriques d'évaluation, l'interaction est ajoutée au modèle. La procédure est répétée en testant l'ajout du couple de variables présentant le deuxième score le plus élevé, et ainsi de suite jusqu'à ce que l'ajout d'interaction ne conduise plus à une amélioration des métriques.
L'ajout d'un terme d'interaction entre deux variables qualitatives V_1 et V_2 possédant respectivement n_1 et n_2 modalités dans un modèle MOB implique que dans chaque modèle GLM des feuilles de l'arbre $(n_1 - 1) \times (n_2 - 1)$ coefficients sont calculés.

Afin de garantir la robustesse des résultats, une validation croisée sur cinq blocs a été effectuée : pour un bloc de test $k \in \{1, ..., 5\}$, les quatre autres blocs forment l'échantillon d'apprentissage sur lequel le modèle MOB est entraîné en testant l'ajout d'une interaction. Pour savoir si cette interaction est ajoutée au modèle, les métriques sont évaluées sur le bloc k et, si une amélioration est constatée, l'interaction est ajoutée au modèle, sinon la procédure s'arrête.

Cette validation croisée a permis de conduire au résultat suivant : l'ajout des trois interactions possédant le score le plus élevé permet d'améliorer les performances du modèle. La formule utilisée est donc

```
coutSinistre ~ GroupeKm + GroupeAge + typeSinistre + expertA +
               ASSUREUR:GroupeKm + GroupeKm:GroupeAge + ASSUREUR:GroupeAge
             | CategorieMarque + tauxHoraire + ASSUREUR                    (3.11)
```

La formule (3.11) est alors utilisée pour entraîner le modèle MOB sur l'ensemble de l'échantillon d'apprentissage, puis les métriques sont évaluées sur l'échantillon de validation. L'arbre construit avec cette formule conduit à la même segmentation que celle induite par l'arbre représenté en (3.13) *supra*,

mais les modèles GLM au sein de chacune des huit feuilles de l'arbre contiennent des coefficients supplémentaires associés aux interactions qui ont été ajoutées. Les résultats de prédiction obtenus sont présentés dans le tableau (3.11) ci-dessous :

MSE	MAE
3 749 422	1 193

TABLE 3.11 : Évaluation des métriques pour le modèle MOB avec intégration d'interactions sur l'échantillon de validation

Ainsi, l'ajout d'interactions dans le modèle MOB permet d'améliorer la MSE de 1.3% et la MAE de 0.72%. Une représentation graphique des coûts observés et prédits par le modèle MOB avec ajout d'interactions pour chaque modalité des variables explicatives est donnée par la figure (3.22) ci-dessous :

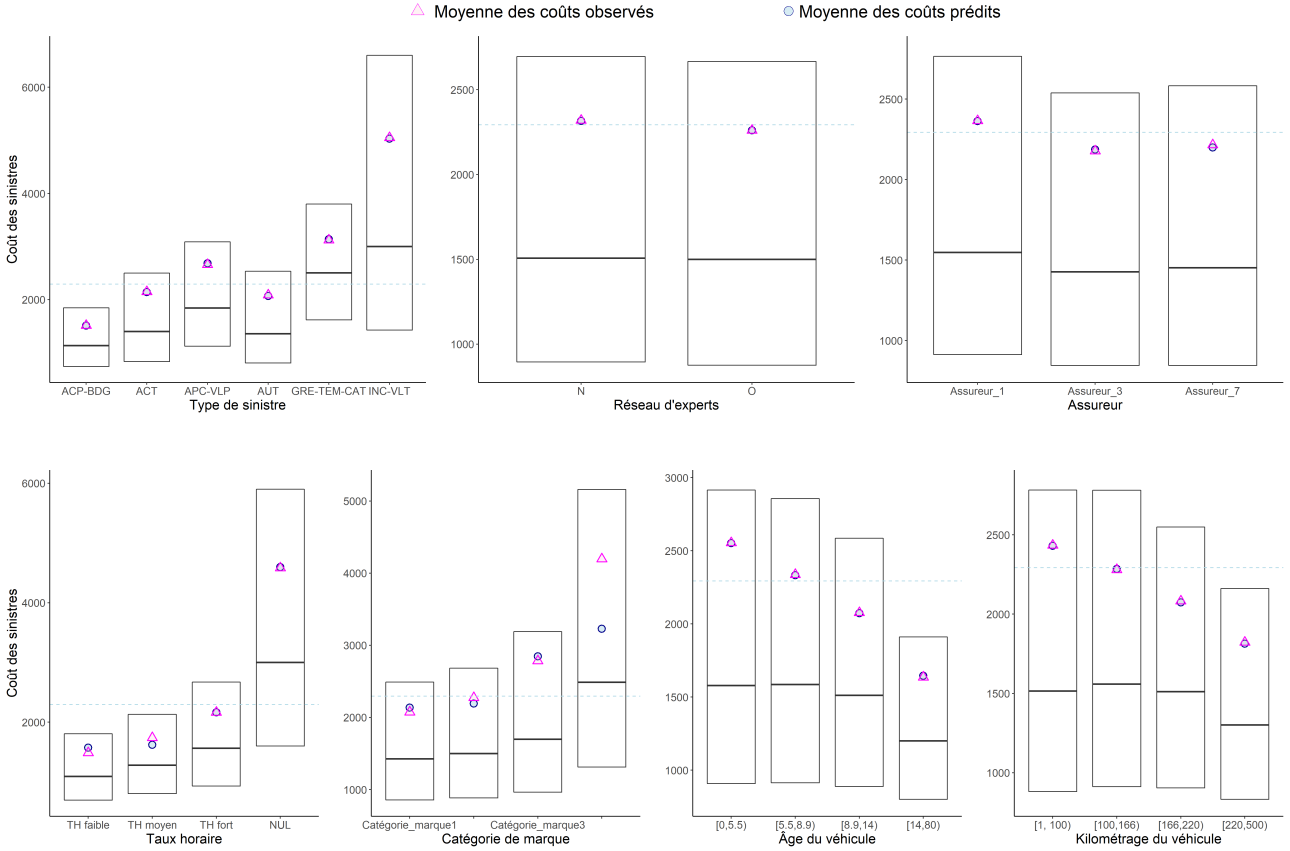


FIGURE 3.22 : Moyennes des sinistres observés et prédits par le modèle MOB avec ajout d'interactions pour chaque modalité des variables explicatives

La moyenne des prédictions associées à ce nouveau modèle MOB est égale à environ 2 294€, similairement aux modèles CART, MOB et XGBoost étudiés précédemment. Globalement, les écarts entre les moyennes pour chaque modalité sont similaires à ce qui avait été observé pour le modèle MOB simple, *i.e.* sans ajout d'interaction. Néanmoins une diminution de l'écart des moyennes pour les sinistres associés à l'assureur n°7 est observée. L'ajout d'interactions au modèle MOB a donc permis d'améliorer les prédictions en moyenne pour les sinistres rattachés à cet assureur. Un écart relativement important de moyenne reste observé pour la catégorie de véhicule n°4. La représentation

graphique (3.22) permet de mieux comprendre comment l'ajout d'interactions a permis l'amélioration du pouvoir prédictif du modèle.
Les résultats de prédiction obtenus sont toujours moins bons que ceux associés au modèle XGBoost, mais une amélioration des prédictions a été réalisée tout en conservant la facilité d'interprétation des modèles MOB. L'étude du modèle XGBoost peut également être effectuée pour essayer d'améliorer les résultats du modèle GLM par l'ajout d'interaction(s). Les résultats associés à cette étude sont présentés en annexe (B.3).

Les modèles GLM, CART, MOB et XGBoost ont été entraînés et optimisés à partir des échantillons d'apprentissage et de validation. Les modèles optimisés peuvent maintenant être appliqués sur l'échantillon de test afin de comparer les modèles.

3.4 Comparaison des modèles

La comparaison finale des modèles se fait sur l'échantillon de test, jusqu'à présent inutilisé. Cette étape a pour but de déterminer quel modèle est le meilleur pour répondre aux objectifs de l'étude ici présentée. Un première étape de la comparaison des modèles est l'évaluation de leurs pouvoirs de prédiction.

3.4.1 Pouvoir prédictif des modèles

Un premier aspect permettant d'évaluer les modèles est leur pouvoir de prédiction des coûts de sinistres. Pour cela deux métriques ont été utilisées : la MSE et la MAE. Les résultats obtenus sur l'échantillon de test sont donnés dans le tableau (3.12) ci-dessous :

	MSE	MAE
GLM	4 656 265	1 293
CART	4 139 728	1 261
MOB	3 816 635	1 198
$\mathbf{MOB_I}$	3 764 333	1 189
XGB	3 714 200	1 182

	Amélioration MSE					Amélioration MAE				
	GLM	CART	MOB	MOB_I	XGB	GLM	CART	MOB	MOB_I	XGB
GLM	×	×	×	×	×	×	×	×	×	×
CART	11.1%	×	×	×	×	2.4%	×	×	×	×
MOB	18.0%	7.8%	×	×	×	7.3%	5.0%	×	×	×
$\mathbf{MOB_I}$	19.2%	9.1%	1.4%	×	×	8.0%	5.7%	0.8%	×	×
XGB	20.2%	10.3%	2.7%	1.3%	×	8.6%	6.3%	1.4%	0.6%	×

TABLE 3.12 : Résultats de prédiction des modèles sur l'échantillon de test et présentation des améliorations fournies

Le modèle MOB auquel des interactions ont été intégrées est noté MOB_I et le modèle XGBoost est noté XGB. Les améliorations des métriques ont également été représentées : par exemple, le modèle MOB simple (*i.e.* sans ajout d'interaction) permet une amélioration de la MSE par rapport au modèle GLM de $\frac{4\ 656\ 265-3\ 816\ 635}{4\ 656\ 265} = 18\%$.

L'application des métriques sur l'échantillon de test permet d'observer que le modèle XGBoost est celui donnant, à nouveau, les meilleurs résultats de prédiction. L'intégration d'interactions dans le modèle MOB permet d'améliorer les résultats de prédiction du modèle, même si ce dernier reste moins performant que le modèle XGBoost.

Cette analyse des métriques est complétée par une analyse des moyennes des sinistres observés et prédits par les différents modèles.

Moyennes des sinistres observés et prédits

De même que sur l'échantillon de validation, il est choisi de comparer les moyennes des sinistres observés et prédits par les différents modèles. Les moyennes des coûts prédits par les différents modèles sur l'échantillon de test sont résumées dans le tableau (3.13) :

	Moyenne prédite	Ecart (%)
GLM	2 261	-2.03%
CART	2 297	-0.45%
MOB	2 299	-0.38%
$\mathbf{MOB_I}$	2 298	-0.41%
XGB	2 298	-0.40%

TABLE 3.13 : Moyenne des prédictions associées aux différents modèles

L'écart par rapport à la moyenne des coûts dans l'échantillon de test a également été renseigné : il est observé que, tout comme sur l'échantillon de validation, l'ensemble des modèles étudiés sous-estiment les coûts en moyenne. Pour rappel, la mesure d'un tel écart ne permet pas la hiérarchisation des performances des modèles, car il a été vu précédemment que deux modèles peuvent présenter le même écart de moyennes tout en ayant des pouvoirs prédictifs différents. Cette mesure a été choisie afin de s'assurer de comparer des objets de même nature. Il est alors intéressant d'étudier les différents écarts de moyennes au sein de chaque modalité des variables explicatives de la base. Une représentation graphique des moyennes des coûts observés et prédits par les différents modèles sur la base de test est donnée ci-dessous en figure (3.23).
Les points associés à chaque modèle ont été volontairement été décalés horizontalement les uns par rapport aux autres afin de permettre une meilleure visualisation. L'écart par rapport à la moyenne observée, représentée en noir, se lit uniquement de façon verticale sur la figure (3.23) ci-dessous.

Cette représentation graphique permet globalement d'observer les mêmes écarts de moyennes que ceux observés sur l'échantillon de validation. En particulier, il est possible d'identifier des modalités pour lesquelles des écarts relativement importants de moyennes peuvent être observés pour certains modèles, comme par exemple :
— la catégorie de marque de véhicule n°4 : pour cette catégorie il est possible d'observer une sous-estimation moyenne relativement importante pour les deux modèles MOB (avec et sans ajout d'interaction). Le modèle CART sous-estime également les coûts de cette catégorie en moyenne, mais l'écart est moins important ;
— le taux horaire nul : en moyenne le modèle GLM sous-estime plus fortement les coûts associés à des taux horaires nuls que les autres modèles ;
— les groupes de véhicules les plus âgés ou ayant parcouru le plus de kilomètres : en moyenne les modèles GLM et CART surestiment les coûts associés à ces catégories, la surestimation étant plus

importante avec le modèle CART ;
— les sinistres incendies ou vol totaux : en moyenne les modèles GLM et CART sous-estiment relativement fortement le coût de ces types de sinistres, la sous-estimation étant plus importante avec le modèle CART ;
— l'assureur n°7 : en moyenne les modèles CART et MOB surestiment les coûts associés à cet assureur. De même que sur l'échantillon de validation, l'ajout d'interactions dans le modèle MOB permet de diminuer cet effet.

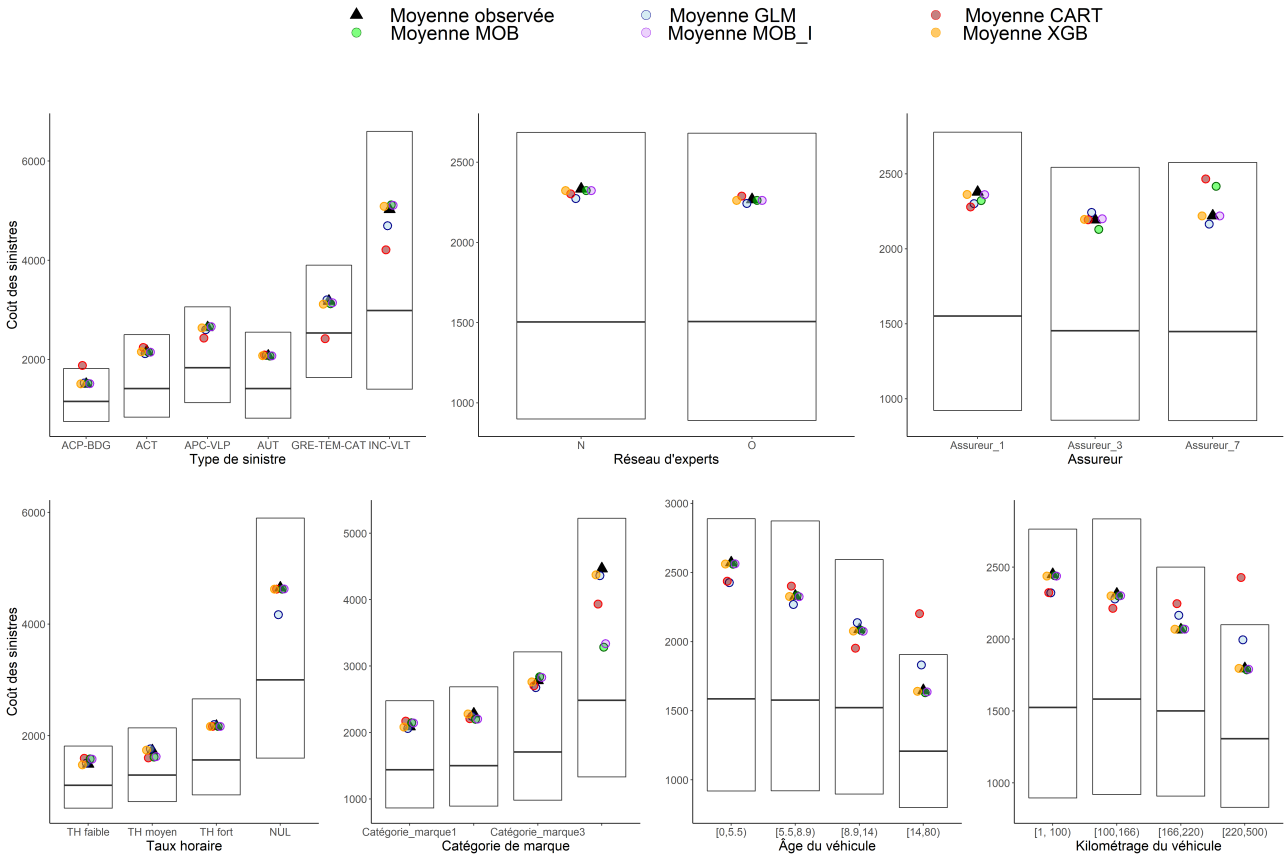


FIGURE 3.23 : Moyennes des sinistres observés et prédits par les différents modèles pour chaque modalité des variables explicatives

Il ressort de l'évaluation des métriques que le modèle XGBoost possède le meilleur pouvoir de prédiction. Au vu de l'objectif de l'étude, une autre caractéristique à prendre en compte pour comparer les modèles est la mesure de l'influence du réseau d'experts sur la prédiction.

3.4.2 Mesure de l'influence du réseau d'expert

Dans le cadre de ce mémoire, l'objectif est double : prédire au mieux le coût des sinistres et pouvoir mesurer l'influence du réseau d'experts sur cette prédiction. Ainsi, il est important qu'au delà du pouvoir de prédiction, un modèle puisse permettre une lecture aisée de cette influence. La lecture de l'impact d'une variable sur la prédiction diffère selon le modèle utilisé. La façon dont l'influence du réseau a pu être mesurée pour les différents modèles est résumée ci-après.

Mesure de l'influence d'une variable dans un modèle GLM

Comme vu précédemment, le GLM retourne pour chaque variable un ou plusieurs coefficients quantifiant l'influence d'une variable sur la prédiction. Plus précisément, pour une variable qualitative les coefficients traduisent comment le fait d'appartenir à une certaine modalité plutôt qu'à une modalité de référence influe sur la prédiction. Le modèle GLM fournit également un intervalle de confiance associé à l'estimation de chaque coefficient. Ce modèle permet ainsi une lecture aisée et facilement compréhensible de l'influence des variables sur la prédiction.

Pour le réseau d'experts, cette variable possédant deux modalités, un unique coefficient est retourné par le modèle et son interprétation est simple : si le coefficient est négatif, être expertisé par le réseau A plutôt que le réseau B fait baisser le coût en moyenne, et inversement. La valeur du coefficient permet de quantifier l'impact du choix d'un certain réseau d'experts. Le coefficient a été estimé sur l'ensemble de la base d'étude et a permis de conduire à la conclusion suivante : avoir recours au réseau d'experts A plutôt qu'au réseau d'experts B fait baisser le coût du sinistre de 2.9% en moyenne.

Mesure de l'influence d'une variable dans un modèle CART

La capacité de mesure de l'influence d'une variable explicative dans un modèle CART dépend de si la variable apparaît ou non dans l'arbre construit par le modèle. En effet, si la variable apparaît dans l'arbre il est alors possible d'observer comment elle est utilisée pour séparer l'ensemble des individus, et quelles sont les caractéristiques de la variable d'intérêt au sein des groupes créés. Cependant, si la variable n'apparaît pas dans l'arbre ce raisonnement ne peut plus être appliqué et la mesure de l'influence de la variable sur la prédiction devient assez limitée. À noter que le modèle CART fournit une mesure d'importance des variables. Cette valeur ne permet cependant pas de comprendre comment la variable influe sur la prédiction mais quantifie seulement son importance au sein du modèle.

Dans le cadre de ce mémoire, l'arbre construit par CART selon les restrictions imposées, à savoir une profondeur maximale égale à 4, ne fait pas apparaître la variable spécifiant le réseau d'experts intervenu. Cela limite la possibilité de mesure de son influence. Il est possible de comparer dans chacun des groupes construits par CART le coût moyen des sinistres expertisés respectivement par les réseaux d'experts A et B. Cependant, il n'est pas possible de tirer de conclusion rigoureuse de ce type de comparaison car, comme expliqué précédemment, il suffit par exemple que le réseau d'experts B ait expertisé des véhicules plus coûteux que le réseau d'experts A pour que son coût moyen de sinistres expertisés soit plus élevé que celui du réseau A. Ainsi, le modèle CART ne permet pas ici une bonne lecture de l'influence du réseau d'experts sur le coût des sinistres.

Mesure de l'influence d'une variable dans un modèle MOB

Pour le modèle MOB, deux cas sont à distinguer : si la variable dont on cherche à mesurer l'influence est utilisée comme variable de partitionnement, dans ce cas la mesure de son influence est similaire à celle dans CART, ou, si elle est utilisée comme variable de régression, dans ce cas la mesure de son influence est similaire à celle dans un GLM. Cependant, dans ce dernier cas, pour chaque variable le modèle retourne autant de coefficients qu'il existe de sous-modèles GLM créés par MOB. Ainsi le modèle permet une lecture de l'influence des variables de régression similaire au GLM, tout en permettant d'obtenir plusieurs coefficients selon les différents groupes d'individus créés par l'arbre. Cela permet d'avoir une mesure plus précise de l'influence de la variable, plutôt qu'un seul effet par modalité, comme dans un GLM.

Dans ce mémoire la variable représentant le réseau d'experts a été utilisée comme variable de régression. La lecture de l'influence du réseau intervenu s'effectue alors en observant le coefficient estimé par le modèle GLM dans chaque feuille de l'arbre. L'arbre construit dans cette étude contenant 8 feuilles, 8 coefficients ont été retournés, permettant de lire l'influence du réseau pour les différents groupes d'individus créés lors de la construction de l'arbre.

Mesure de l'influence d'une variable dans un modèle XGBoost
Le modèle XGBoost étant basé sur une méthode d'ensemble, il regroupe plusieurs sous-arbres permettant d'aboutir à une unique prédiction. Cette multiplicité de sous-modèles rend la visualisation et l'interprétation du modèle XGBoost relativement complexe.
Dans le cadre de ce mémoire le modèle XGBoost est basé sur la construction de 1000 sous-arbres, son interprétation directe est alors assez limitée. Ce modèle a été étudié dans le but de comparer les performances des modèles GLM, CART et MOB à un modèle connu pour ses performances de prédiction. L'étude de l'influence du réseau d'experts dans le modèle XGBoost n'a donc pas été traitée dans ce mémoire. Cependant, bien que ce modèle soit qualifié de "boîte noire", l'étude de l'influence des variables sur la prédiction n'est pas impossible. En particulier, il existe des outils d'analyse agnostiques, tels que LIME (pour *Local Interpretable Model-agnostic Explanations* en anglais) [Ribeiro et al. (2016)] ou bien SHAP (pour *SHapley Additive exPlanations* en anglais) [Lundberg et Lee (2017)], permettant l'analyse des variables sur la prédiction. Ces outils sont notamment présentés dans le mémoire de A. Wabo [Wabo (2020)].
Au vu du rôle du modèle XGBoost dans ce mémoire ces outils n'ont pas été étudiés mais ils présentent un moyen de rendre plus interprétables les modèles qualifiés de "boîtes noires".

Afin de terminer la comparaison des modèles, un dernier aspect à étudier est le temps de calcul qu'ils induisent.

3.4.3 Temps de calcul

Lors de la comparaison de différents modèles, il est important de prendre en compte les temps de calcul associés à ces derniers. En effet, comme il a été vu précédemment, certains modèles possèdent des paramètres dont l'optimisation peut être recherchée. La recherche des meilleurs paramètres pour un modèle peut être particulièrement longue lorsque le modèle possède un temps de calcul important. Les temps de calcul sur l'échantillon d'apprentissage des modèles finaux, *i.e.* des modèles avec les paramètres optimisés, sont résumés dans le tableau (3.14) ci-dessous :

	Temps de calcul
GLM	$\simeq$ 15s
CART	$\simeq$ 6s
MOB	$\simeq$ 1mn
MOB$_\mathbf{I}$	$\simeq$ 3mn
XGBoost	$\simeq$ 2mn

Table 3.14 : Temps d'apprentissage des modèles finaux

Le tableau (3.14) montre que les modèles utilisés dans ce mémoire ont des temps de calcul relativement faibles. Cependant, ces temps de calcul ont été réduits grâce aux traitements effectués sur la base d'étude explicités en section (1.5). Pour souligner l'impact des retraitements effectués quelques exemples de temps de calcul des différents modèles sont explicités ci-après.

Diminution du nombre de modalités
Les variables `marque` et `typeSinistre` ont été retraitées afin de réduire leurs nombres de modalités. Afin de souligner l'impact de ce retraitement les temps de calcul des modèles si la variable `typeSinistre` n'avait pas été retraitée sont explicités ci-dessous dans le tableau (3.15) :

	Temps de calcul	
GLM	$\simeq$ 24s	
CART	$\simeq$ 6s	
MOB	*Cas 1*	*Cas 2*
	$\simeq$ 2mn	$\simeq$ 48mn
MOB$_\mathbf{I}$	*Cas 1*	*Cas 2*
	$\simeq$ 3mn	$\simeq$ 1h
XGBoost	$\simeq$ 2mn	

TABLE 3.15 : Temps d'apprentissage des modèles si la variable `typeSinistre` n'avait pas été retraitée

Pour rappel, avant diminution de son nombre de modalités la variable `typeSinistre` possédait 11 modalités. Ce nombre a été réduit à 6 après retraitement. Le tableau (3.15) montre que la diminution de ce nombre a très peu d'impact sur les temps de calcul des modèles GLM, CART et XGBoost.

Concernant le modèle MOB deux cas sont à distinguer, en fonction de si la variable `typeSinistre` était utilisée comme variable de régression (cas n°1) ou comme variable de partitionnement (cas n°2). Une forte augmentation du temps de calcul est observée si la variable `typeSinistre` est utilisée en variable de partitionnement. Bien que le temps de calcul du modèle reste relativement faible, l'optimisation du modèle MOB si la variable `typeSinistre` n'est pas retraitée implique un temps de calcul très important. Par exemple, pour la recherche de la formule optimale, parmi les 126 formules testées chaque variable est utilisée 64 fois en variable de partitionnement. La simple recherche de la formule optimale sur l'échantillon de validation, sans validation croisée et sans essayer plusieurs valeurs de profondeur maximale prend plus de 38 heures. L'augmentation du temps de calcul du modèle rend son optimisation très complexe. Cet exemple souligne l'importance du retraitement des variables possédant un nombre élevé de modalités pour diminuer le temps de calcul de certains algorithmes.

Discrétisation des variables numériques

Les variables `ageVehicule` et `kilométrage` ont été discrétisées et ont respectivement conduit à la création des variables qualitatives `GroupeAge` et `GroupeKm`. Afin de souligner l'impact de ce retraitement, les temps de calcul des modèles si ces variables n'avaient pas été discrétisées sont explicités dans le tableau (3.16) ci-dessous :

	Temps de calcul	
GLM	$\simeq$ 16s	
CART	$\simeq$ 7s	
MOB	*Cas 1*	*Cas 2*
	$\simeq$ 1mn	$>$ 10h
MOB$_\mathbf{I}$	*Cas 1*	*Cas 2*
	$\simeq$ 3mn	$>$ 10h
XGBoost	$\simeq$ 3mn	

TABLE 3.16 : Temps d'apprentissage des modèles si les variables numériques n'avaient pas été discrétisées

Les temps de calcul des modèles GLM, CART et XGBoost sont très peu modifiés lorsque les variables numériques ne sont pas discrétisées. Pour le modèle MOB, il est possible d'observer une très forte augmentation du temps de calcul si l'une des variables numériques est utilisée comme variable de partitionnement (cas n°2). Si les variables numériques n'avaient pas été discrétisées la recherche

de la formule optimale aurait été trop longue pour pouvoir être mise en place et l'optimisation de l'algorithme aurait été très complexe.

Ainsi, le modèle MOB présente des meilleurs résultats en termes de prédiction que les modèles GLM et CART mais l'application du modèle nécessite des retraitements qui ne sont pas essentiels pour l'application des modèles GLM, CART et XGBoost.

Avant d'appliquer les modèles pour effectuer des prédictions, une optimisation de ces derniers a été réalisée. Les temps de calcul associés à l'optimisation de chacun des modèles sont donnés ci-dessous dans le tableau (3.17) :

	Temps d'optimisation
GLM	$\simeq$ 10mn
CART	$\simeq$ 11mn
MOB	$\simeq$ 22h
MOB$_\mathbf{I}$	$\simeq$ +1h
XGBoost	$\simeq$ 18h

TABLE 3.17 : Temps d'optimisation des modèles

Les optimisations qui ont été effectuées pour les différents modèles sont rappelées ci-dessous :
— l'optimisation du modèle GLM consistait en la procédure de sélection des variables ;
— l'optimisation du modèle CART consistait en la recherche de la profondeur maximale, puis du nombre minimum d'observations à spécifier ;
— l'optimisation du modèle MOB consistait en la recherche de la formule et de la profondeur maximale optimales, puis la recherche d'un nombre minimum d'observations par feuille à spécifier. Cette recherche a été complétée par l'ajout d'un certain nombre d'interactions au modèle, ce qui a nécessité environ une heure de temps de calcul supplémentaire ;
— l'optimisation du modèle XGBoost consistait en la recherche de la meilleure combinaison de valeurs des paramètres explicités dans le tableau (3.6) *supra*.

Le tableau (3.17) ci-dessus permet d'observer que les temps d'optimisation des modèles GLM et CART sont relativement faibles. À l'inverse, les optimisations des modèles MOB et XGBoost requièrent des temps de calcul assez importants.

La comparaison des temps d'apprentissage et d'optimisation des modèles permet de mettre en avant un inconvénient du modèle MOB : ce modèle peut nécessiter des retraitements sur les variables explicatives afin de réduire le temps de calcul du modèle, relativement important. L'application du modèle nécessite la détermination d'une formule, *i.e.* d'ensembles de variables de partitionnement et de régression : la recherche de la formule optimale peut devenir très complexe lorsque le nombre de variables explicatives est important. Ainsi, la bonne performance de prédiction du modèle MOB est conditionnée à un temps de calcul relativement important.
Malgré cet inconvénient, le modèle MOB semble être le modèle le plus adapté pour répondre aux objectifs de l'étude parmi les modèles qui ont été étudiés dans ce mémoire. Son pouvoir prédictif se rapproche de celui d'un modèle connu pour ses performances de prédiction, tout en permettant une bonne lisibilité et une bonne interprétabilité des résultats de prédiction. De plus, l'étude d'un modèle plus performant, tel que le modèle XGBoost, permet d'obtenir des moyens d'améliorer les performances du modèle MOB, notamment via l'ajout d'interactions. Le modèle MOB ainsi construit possède un pouvoir prédictif relativement important et son interprétabilité est conservée, ce qui en fait le meilleur modèle pour répondre aux objectifs de l'étude présentée dans ce mémoire.

Conclusion

L'étude présentée dans ce mémoire avait pour objectif de trouver un modèle permettant de prédire au mieux les coûts de sinistres tout en permettant une mesure de l'influence du réseau d'experts intervenu sur ces coûts. Pour cela, trois modèles ont été étudiés.
Dans un premier temps, un modèle linéaire généralisé a été appliqué : ce modèle, usuellement utilisé en tarification automobile, permet une lecture aisée de l'influence des variables explicatives sur les prédictions. Son pouvoir prédictif est cependant relativement limité.
Afin d'améliorer les performances de prédiction, des modèles à structure arborescente ont été introduits. Tout d'abord, le modèle CART, connu pour sa lisibilité et son interprétabilité, a été appliqué. Ce modèle a permis d'obtenir de meilleurs résultats de prédiction que ceux du modèle GLM. Cependant, les restrictions sur la taille des arbres imposées dans cette étude ont conduit à construire un arbre qui ne fait pas intervenir la variable représentant le réseau d'experts pour séparer les individus. Cela a rendu la mesure de l'influence du réseau d'experts difficile pour ce modèle.
L'étude a été complétée par l'application d'un autre modèle d'arbres de régression : le modèle MOB. Ce modèle, utilisant à la fois une structure arborescente et des modèles linéaires généralisés, a permis d'obtenir des meilleurs résultats de prédiction. De plus, il permet une lisibilité de l'influence des variables aussi aisée que dans un GLM tout en étant plus précise, puisque l'influence est fournie pour chaque catégorie d'observations créée par l'arbre MOB.
Les résultats du modèle MOB ont été confrontés à ceux d'un modèle connu pour ses performances de prédiction : le modèle XGBoost. Ce dernier a effectivement présenté des meilleurs résultats prédictifs, mais la lisibilité de l'influence des variables n'est pas immédiate et nécessite l'utilisation d'outils agnostiques. Ce modèle a été analysé afin d'identifier les interactions qu'il exploite, et de les ajouter au modèle MOB. Cette modification du modèle MOB a permis une amélioration du pouvoir de prédiction du modèle, tout en permettant de conserver sa facilité de lecture d'influence des variables. Ce modèle ainsi modifié est le modèle le plus adapté pour répondre aux objectifs de l'étude.
Tout comme les travaux qui avaient été effectués auparavant sur la comparaison entre les réseaux d'experts A et B, ce mémoire conduit à conclure que le réseau d'experts A est plus performant en moyenne que le réseau B, les différences de performances entre les deux réseaux pouvant varier selon la période considérée ou les caractéristiques des individus étudiés.

Il est important de noter une limite de l'étude effectuée dans ce mémoire : les variables explicatives à disposition. En effet, certaines variables explicatives n'ont pas pu être intégrées au modèle, faute de disponibilité des données. Il est probable que l'intégration de certaines variables explicatives dans les modèles permettrait d'en améliorer les pouvoirs prédictifs. De plus, l'étude de variables supplémentaires aurait pu permettre une analyse plus approfondie des différences de performances entre les réseaux. Il se peut par exemple qu'il existe des zones géographiques dans lesquelles le réseau B soit plus performant en moyenne, mais comme la variable représentant le code postal du sinistre contenait trop de valeurs manquantes, il n'a pas été possible d'effectuer des études de la sorte. Ce type d'exemple souligne les difficultés que peuvent entraîner la fusion et l'harmonisation de bases de

données provenant de différentes sources externes.

Le modèle MOB, bien qu'adapté aux objectifs de l'étude, a nécessité des temps d'optimisation importants. Ce type d'optimisation n'aurait pas pu être mis en place si le nombre de variables explicatives utilisées avait été relativement trop important. De plus, pour obtenir des temps de calculs suffisamment faibles pour pouvoir être appliqué et optimisé, le modèle MOB a nécessité des retraitements des variables explicatives qui n'avaient pas besoin d'être mis en place pour les autres modèles. Bien que la performance et l'interprétabilité du modèle MOB le positionnent comme le meilleur modèle dans le cadre de cette étude, ces inconvénients ne sont pas à négliger et sont à garder en esprit.

Dans ce mémoire, l'étude s'est concentrée principalement sur trois modèles. Il est néanmoins possible qu'il existe d'autres outils ou modèles permettant d'améliorer les résultats obtenus. Récemment, le développement d'outils agnostiques permettant d'éclaircir le fonctionnement des modèles "boîte noire", comme le modèle XGBoost, peuvent être considérés comme une solution pour répondre aux objectifs de l'étude. Ces outils n'ont pas été étudiés dans ce mémoire, le lecteur intéressé peut se référer au mémoire de A. Wabo [Wabo (2020)]. Il existe également des modèles de partitionnement récursif basés sur des modèles GAM (pour *Generalised Additive Model* en anglais) ou encore des arbres de régression basés sur la maximisation de la vraisemblance (*Maximum Likelihood Regression Tree* en anglais) dont l'étude n'a pas été menée dans ce mémoire, mais qui pourraient constituer une piste d'exploration pour des travaux futurs sur le même type de sujet.

La recherche d'un compromis entre pouvoir prédictif et interprétabilité des modèles constitue un sujet primordial en assurance automobile, puisqu'il est important pour un assureur de proposer les tarifs les plus attractifs, tout en ayant la capacité de comprendre et d'expliquer la construction de ces derniers. Le développement de différentes méthodes de machine learning à travers le temps offre de nombreux moyens permettant de répondre à ce type de problématique. L'avancée dans le monde de la Data Science permettra très certainement dans un futur proche de disposer de nouveaux outils pour répondre à la problématique de cette étude.

Bibliographie

Andrews, D. W. K. (1993). Tests for Parameter Instability and Structural Change With Unknown Change Point. *Econometrica* 61, p. 821-856.

AutoPlus (2020). Comment le trafic a-t-il évolué pendant le confinement ? url : https://www.autoplus.fr/securite-routiere/comment-le-trafic-a-t-il-evolue-pendant-le-confinement-320128.html#item=1.

Bellina, R. (2014). Méthodes d'apprentissage appliquées à la tarification non-vie. Mém. de mast. ISFA.

Bouret, R. (2014). Méthode quantitatives et Sciences Humaines, 2ème édition. Renaud Bouret Éditeur. Chap. 9.

Breiman, L., Friedman, J. H., Stone, C. J. et Olshen, R. A. (1984). Classification and Regression Trees. Chapman et Hall/CRC.

Charpentier, A. (juin 2011). La loi des grands nombres et le théorème central limite comme base de l'assurabilité. *Risques* 86.

Charpentier, A. (jan. 2015). Segmentation et Mutualisation, les deux faces d'une même pièce. *Risques* 103.

Chen, T. et Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*. ACM, p. 785-794.

Code des assurances (2003). Article L251-2.

Denuit, M. et Charpentier, A. (jan. 2004). Mathématiques de l'assurance non-vie : principes fondamentaux de théorie du risque. Tome 1. Economica.

FFA (2021a). L'assurance française : données clés 2020. *Fédération française de l'assurance*, p. 54,57,65.

FFA (2021b). Les assureurs, acteurs de la relance durable : les chiffres de l'assurance en 2020. *Dossier de presse, Fédération Française de l'assurance*, p. 19-25.

FFA AUTO (2021). L'assurance auto. *Fédération française de l'assurance, infos assurés*.

Freund, Y. et Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55, p. 119-139.

Friedman, J. H. (mars 1999). Stochastic Gradient Boosting. *Computational Statistics and Data Analysis* 38, p. 367-378.

Genuer, R. et Poggi, J.-M. (2017). Arbres CART et Forêts aléatoires, Importance et sélection de variables.

Ghattas, B. (jan. 2010). Importance des variables dans les méthodes CART. *Universite Aix-Marseille III, G.R.E.Q.A.M.*

Hjort, N. L. et Koning, A. (2002). Tests for Constancy of Model Parameters Over Time. *Journal of nonparametric Statistic* 14.1-2, p. 113-132.

Khougea, D. (2019). Tarification IARD avec des modèles de régression avancés. Mém. de mast. Université de Strasbourg.

lesfurets (2021). Assurance auto : tout savoir sur l'expertise à distance. url : https://www.lesfurets.com/assurance-auto/guide/tout-savoir-sur-lexpertise-a-distance-ead.

Li, C. (2016). A Gentle Introduction to Gradient Boosting.

Lorentzen, C. et Mayer, M. (mai 2020). Peeking into the Black Box : An Actuarial Case Study for Interpretable Machine Learning.

Lundberg, S. M. et Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. T. 30. Curran Associates, Inc. URL : https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Lussac, M. D. (2018). Comparaison de modèles prédictifs pour l'évaluation des coûts matériels automobiles. Mém. de mast. Université Paris Dauphine.

Mayer, M. (2021). flashlight: Shed Light on Black Box Machine Learning Models. URL : https://CRAN.R-project.org/package=flashlight.

ONISR (2021). Bilan 2020 de la sécurité routière. *Accidentalité routière 2020 : données définitives*, p. 6. URL : https://www.onisr.securite-routiere.gouv.fr/etat-de-l-insecurite-routiere/bilans-annuels-de-la-securite-routiere/bilan-2020-de-la-securite-routiere.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL : https://www.R-project.org/.

Ribeiro, M. T., Singh, S. et Gestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier . *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 1135–1144. URL : https://doi.org/10.1145/2939672.2939778.

Rusch, T. et Zeileis, A. (2013). Gaining Insight with Recursive Partitioning of Generalized Linear Models. *Journal of Statistical Computation and Simulation* 83.7, p. 1301-1315.

Strobl, C., Zeileis, A., Boulesteix, A.-L. et Hothorn, T. (déc. 2010). Variable Selection Bias in Classification Trees and Ensemble Methods.

Thomas, J. (2014). Modèles linéaires & GLM Analyse logit & Régression de Poisson Analyse d'un portefeuille d'assurance algorithme IRWLS avec R. ISFA.

Wabo, A. (2020). Mesure de l'écart de performance entre deux réseaux d'experts en assurance automobile. Mém. de mast. Université Paris Dauphine.

Waze (2020). Confinement : Waze confirme la chute spectaculaire du trafic routier en France. *01net.*

Zeileis, A. et Hornik, K. (juill. 2003). Generalized M-Fluctuation Tests for Parameter Instability. *Stat. Neerland* 61.

Zeileis, A., Hothorn, T. et Hornik, K. (2008). Model-based Recursive Partitioning. *Journal of Computational and Graphical Statistics* 17.2, p. 492-514.

Annexe A

Compléments sur la base d'étude

A.1 Les différents types de sinistres

Les significations des modalités de la variable `typeSinistre` sont résumées dans le tableau (A.1) ci-dessous :

Modalité	Description
ACP	Accident de parking
ACT	Accident avec tiers
APC	Accident perte de contrôle
AUT	Autres
BDG	Bris de glace
CAT	Catastrophe naturelle
GRE	Grêle
INC	Incendie
TEM	Tempête
VLP	Vol partiel
VLT	Vol total

Table A.1 : Les différents types de sinistre

A.2 Étude des échantillons d'apprentissage, de validation et de test

L'objectif est ici de vérifier graphiquement que le découpage en échantillons d'apprentissage, de validation et de test n'a pas conduit à la création d'échantillons qui soient trop différents. Une représentation graphique de la fréquence de chaque modalité des variables au sein des trois échantillons a été effectuée :

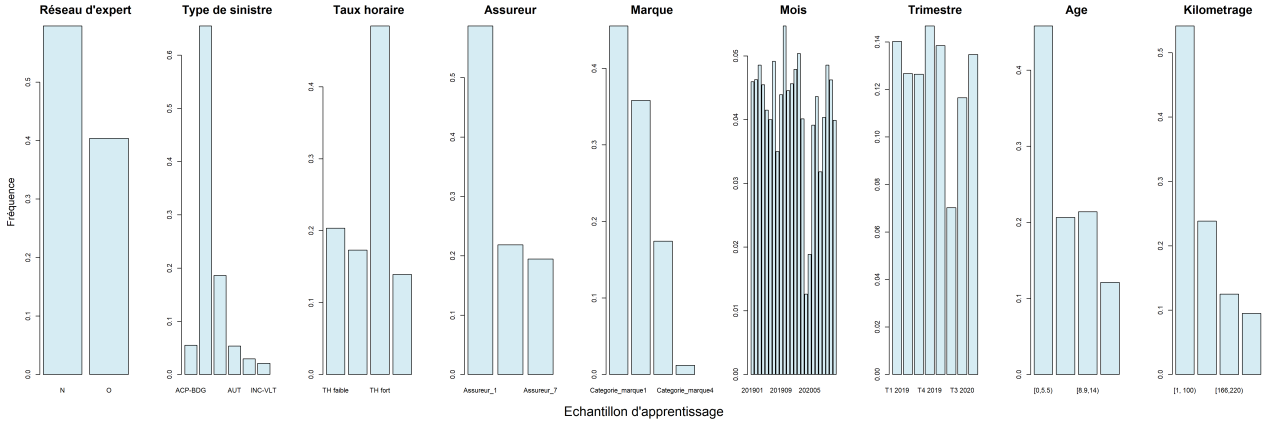


FIGURE A.1 : Caractéristiques des variables au sein de la base d'apprentissage

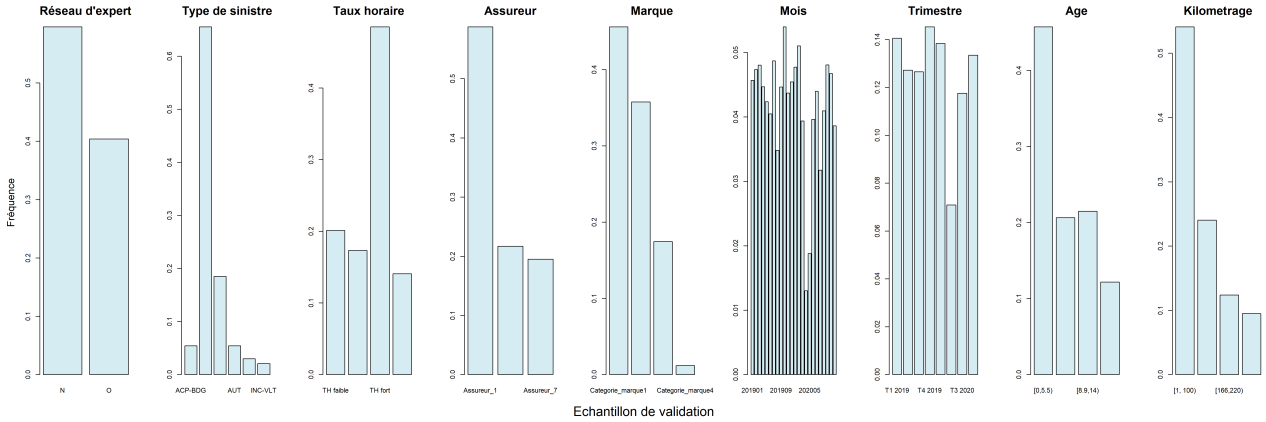


FIGURE A.2 : Caractéristiques des variables au sein de la base de validation

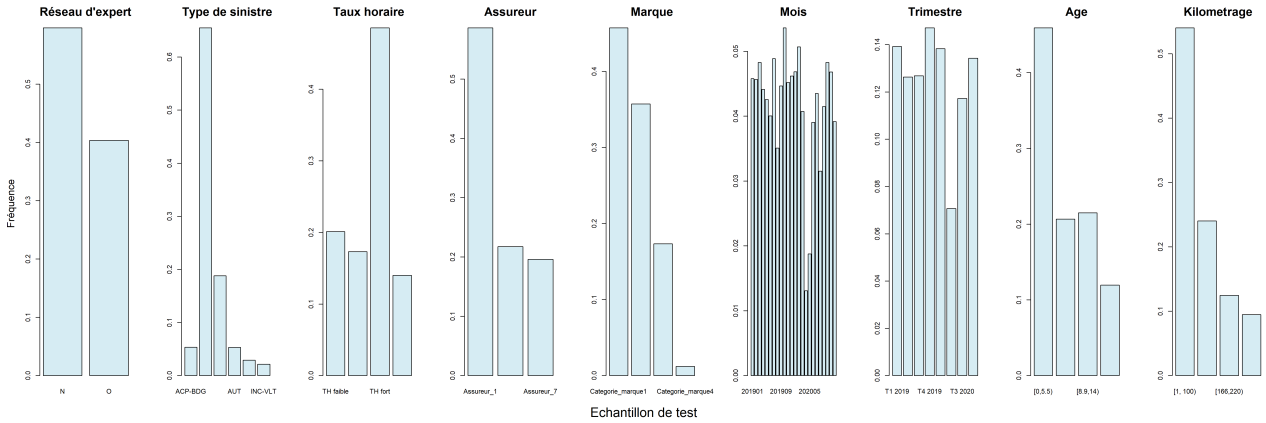


FIGURE A.3 : Caractéristiques des variables au sein de la base de test

Les figures (A.1), (A.2), (A.3) permettent d'observer que pour toutes les variables la fréquence de chacune des modalités est du même ordre de grandeur au sein des trois échantillons construits.

Annexe B

Compléments sur le modèle linéaire généralisé

B.1 Équation du premier ordre associée aux paramètres du GLM

L'estimation des paramètres d'un modèle linéaire généralisé se fait par maximum de vraisemblance, l'objectif est de résoudre :

$$\frac{\partial \mathcal{L}(Y,\theta)}{\partial \beta} = 0,$$

Les observations étant supposées suivre une loi appartenant à la famille exponentielle, la log-vraisemblance du modèle est donnée par :

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i,\phi),$$

de plus,

$$\forall i \in \{1,...,n\},\ \theta_i = (b')^{-1} \circ g^{-1}(x_i\beta) = h(x_i\beta), \qquad \text{où } h = (b')^{-1} \circ g^{-1},$$

d'où,

$$\begin{aligned}\mathcal{L}(\beta) &= \sum_{i=1}^{n} \frac{Y_i h(x_i\beta) - b(h(x_i\beta))}{a(\phi)} + c(Y_i,\phi) \\ &= \sum_{i=1}^{n} \mathcal{L}_i(\beta) \qquad \text{avec } \mathcal{L}_i(\beta) = \frac{Y_i h(x_i\beta) - b(h(x_i\beta))}{a(\phi)} + c(Y_i,\phi).\end{aligned}$$

Il faut résoudre les équations de vraisemblance, *i.e.* $\forall j \in \{1,...,p\}$, il faut résoudre :

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \mathcal{L}_i(\beta)}{\partial \beta_j} = 0,$$

or,

$$\begin{aligned}\frac{\partial \mathcal{L}_i(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{Y_i x_i^j h'(x_i\beta) - x_i^j (b \circ h)'(x_i\beta))}{a(\phi)} \\ &= \sum_{i=1}^{n} \frac{Y_i x_i^j h'(x_i\beta) - x_i^j h'(x_i\beta) b' \circ h(x_i\beta)}{a(\phi)} \\ &= \sum_{i=1}^{n} \frac{Y_i x_i^j h'(x_i\beta) - x_i^j h'(x_i\beta) g^{-1}(x_i\beta)}{a(\phi)} \\ &= \sum_{i=1}^{n} \frac{x_i^j h'(x_i\beta)(Y_i - g^{-1}(x_i\beta)}{a(\phi)},\end{aligned}$$

de plus,

$$\begin{aligned}h'(x_i\beta) &= ((b')^{-1} \circ g^{-1})'(x_i\beta) \\ &= (g^{-1})' \times ((b')^{-1})' \circ g^{-1}(x_i\beta) \\ &= (g^{-1})'(x_i\beta) \times \frac{1}{b'' \circ (b')^{-1} \circ g^{-1}(x_i\beta)} \\ &= (g^{-1})'(x_i\beta) \times \frac{1}{b'' \circ h(x_i\beta)} \\ &= (g^{-1})'(x_i\beta) \times \frac{1}{b''(\theta_i)} \\ &= (g^{-1})'(x_i\beta) \times \frac{a(\phi)}{\mathbb{V}(Y_i)}, \qquad \text{avec la formule (2.1)}\end{aligned}$$

ainsi,

$$\frac{h'(x_i\beta)}{a(\phi)} = \frac{(g^{-1})'(x_i\beta)}{\mathbb{V}(Y_i)},$$

Les équations de vraisemblance obtenues sont données par la formule suivante :

$$\sum_{i=1}^{n} \frac{x_i^j h'(x_i\beta)(Y_i - g^{-1}(x_i\beta))}{a(\phi)} = \sum_{i=1}^{n} (g^{-1})'(x_i\beta) \frac{x_i^j (Y_i - g^{-1}(x_i\beta))}{\mathbb{V}(Y_i)} = 0 \qquad \forall j = 1, ..., p. \tag{B.1}$$

Les équations de vraisemblance n'ont pas de solution explicite en général, sauf lorsque $b' = id$. Des méthodes itératives telles que l'algorithme de Newton-Raphson peuvent être utilisées pour approcher la solution.

B.2 Coefficients du GLM

Les coefficients obtenus avec le modèle généralisé entraîné sur l'ensemble de la base d'étude sont résumés dans le tableau (B.1) ci-dessous :

	Estimation	Erreur	t-valeur	p-valeur
(Intercept)	7.0703909	0.0052258	1352.968893	$< 2e-16$
expertAO	**-0.0287283**	0.0019544	-14.699191	$< 2e-16$
TSACT	0.2534287	0.0043317	58.505471	$< 2e-16$
TSAPC-VLP	0.4085261	0.0046687	87.503284	$< 2e-16$
TSAUT	0.2573804	0.0058865	43.723881	$< 2e-16$
TSGRE-TEM-CAT	0.6254468	0.0069624	89.832181	$< 2e-16$
TSINC-VLT	0.5866210	0.0079901	73.418863	$< 2e-16$
GroupeAge.L	-0.4129210	0.0026723	-154.520066	$< 2e-16$
GroupeAge.Q	-0.1141382	0.0022254	-51.287999	$< 2e-16$
GroupeAge.C	-0.0291792	0.0021016	-13.884530	$< 2e-16$
ASSUREURAssureur_3	0.0241388	0.0025221	9.570808	$< 2e-16$
ASSUREURAssureur_7	-0.1124055	0.0028495	-39.447527	$< 2e-16$
CategorieMarque.L	0.5529666	0.0059842	92.404366	$< 2e-16$
CategorieMarque.Q	0.1935174	0.0046364	41.738901	$< 2e-16$
CategorieMarque.C	0.0179178	0.0027262	6.572374	$< 2e-16$
GroupeKm.L	-0.2081806	0.0028374	-73.369501	$< 2e-16$
GroupeKm.Q	-0.0756562	0.0024572	-30.790045	$< 2e-16$
GroupeKm.C	0.0050254	0.0023719	2.118696	0.0341165
tauxHoraireTH moyen	0.1322110	0.0031814	41.557439	$< 2e-16$
tauxHoraireTH fort	0.3317477	0.0027457	120.825482	$< 2e-16$
tauxHoraireNUL	1.2033285	0.0035726	336.822891	$< 2e-16$

TABLE B.1 : Résumé des coefficients du modèle GLM appliqué à l'ensemble de la base d'étude

B.3 Ajout d'interactions au modèle GLM

Cette section présente l'étude de l'amélioration du modèle GLM via l'ajout d'interactions après analyse du modèle XGBoost, selon le même principe que celui présenté en sous-section (3.3.3).

L'ajout d'interactions au sein du modèle GLM peut être étudié pour essayer d'en améliorer le pouvoir prédictif. Le modèle GLM présenté en chapitre 2 n'intègre aucun interaction, dans la suite de cette section il sera appelé "modèle GLM simple". Pour savoir quelles interactions ajouter au modèle, les interactions associées au modèle XGBoost sont étudiées. Pour chaque couple de variables (V_1, V_2) un score est calculé en sommant les H-statistiques de Friedman relatives et absolues du couple (V_1, V_2) associées au modèle XGBoost.
Les cinq couples de variables possédant le score ainsi calculé le plus élevé sont présentés ci-dessous dans le tableau (B.2) :

	Score
ASSUREUR:GroupeKm	0.36
tauxHoraire:GroupeAge	0.28
tauxHoraire:GroupeKm	0.14
ASSUREUR:GroupeAge	0.10
ASSUREUR:expertA	0.09

TABLE B.2 : Les cinq couples de variables possédant le score le plus élevé pour le modèle XGBoost

L'ajout d'interaction(s) se fait alors par ordre décroissant du score : l'ajout au modèle GLM du couple de variables possédant le score plus élevé est étudié et si celui-ci fait diminuer les métriques

d'évaluation, l'interaction est ajoutée au modèle. La procédure est répétée en testant l'ajout du couple de variables présentant le deuxième score le plus élevé, et ainsi de suite jusqu'à ce que l'ajout d'interaction ne conduise plus à une amélioration des métriques.
L'ajout d'un terme d'interaction entre deux variables qualitatives V_1 et V_2 possédant respectivement n_1 et n_2 modalités dans un modèle GLM implique que $(n_1 - 1) \times (n_2 - 1)$ coefficients sont calculés.

Afin de garantir la robustesse des résultats, une validation croisée sur cinq blocs a été effectuée : pour un bloc de test $k \in \{1, ..., 5\}$, les quatre autres blocs forment l'échantillon d'apprentissage sur lequel le modèle GLM est entraîné en testant l'ajout d'une interaction. Pour savoir si cette interaction est ajoutée au modèle, les métriques sont évaluées sur le bloc k et, si une amélioration est constatée, l'interaction est ajoutée au modèle, sinon la procédure s'arrête.

Cette validation croisée a été effectuée, avec un temps de calcul d'environ 10 minutes, et a permis de conduire au résultat suivant : l'ajout des quatre interactions possédant le score le plus élevé permet d'améliorer les performances du modèle GLM. La formule utilisée est donc :

```
coutSinistre ~ GroupeKm + GroupeAge + typeSinistre + expertA +CategorieMarque
             + tauxHoraire + ASSUREUR + ASSUREUR:GroupeKm + tauxHoraire:GroupeAge
             + tauxHoraire:GroupeKm + ASSUREUR:GroupeAge                    (B.2)
```

La formule (B.2) est alors utilisée pour entraîner le modèle GLM sur l'ensemble de l'échantillon d'apprentissage, puis les métriques sont évaluées sur l'échantillon de validation. Les résultats de prédiction obtenus sont présentés dans le tableau (B.3) ci-dessous :

MSE	MAE
3 890 173	1 212

TABLE B.3 : Évaluation des métriques sur l'échantillon de validation pour le modèle GLM avec intégration d'interactions

L'ajout d'interactions dans le modèle GLM permet d'améliorer la MSE de 15% et la MAE de 6%. La moyenne des prédictions associées à ce nouveau modèle GLM est égale à environ 2 289€, contre 2 256€ lorsqu'aucune interaction n'était intégrée au modèle. La sous-estimation moyenne des coûts par le GLM a donc été réduite via l'ajout d'interactions. Afin de mieux comprendre ce résultat, une représentation graphique des coûts observés et prédits par le modèle GLM avec ajout d'interactions pour chaque modalité des variables explicatives est donnée par la figure (B.1) ci-dessous.

Il est observé que la plupart des écarts de moyennes relativement importants qui avaient été observé avec le modèle GLM simple ont été réduits, comme par exemple :
— la sous-estimation moyenne pour les sinistres incendies ou vols totaux (`INC-VLT`) ;
— la sous-estimation moyenne pour les taux horaires nuls ;
— la sous-estimation moyenne des véhicules d'âge inférieur à 8.9 ans ou ayant un kilométrage inférieur à 165 601 kilomètres ;
— la surestimation moyenne des véhicules d'âge supérieur à 8.9 ans ou ayant un kilométrage supérieur à 165 601 kilomètres.

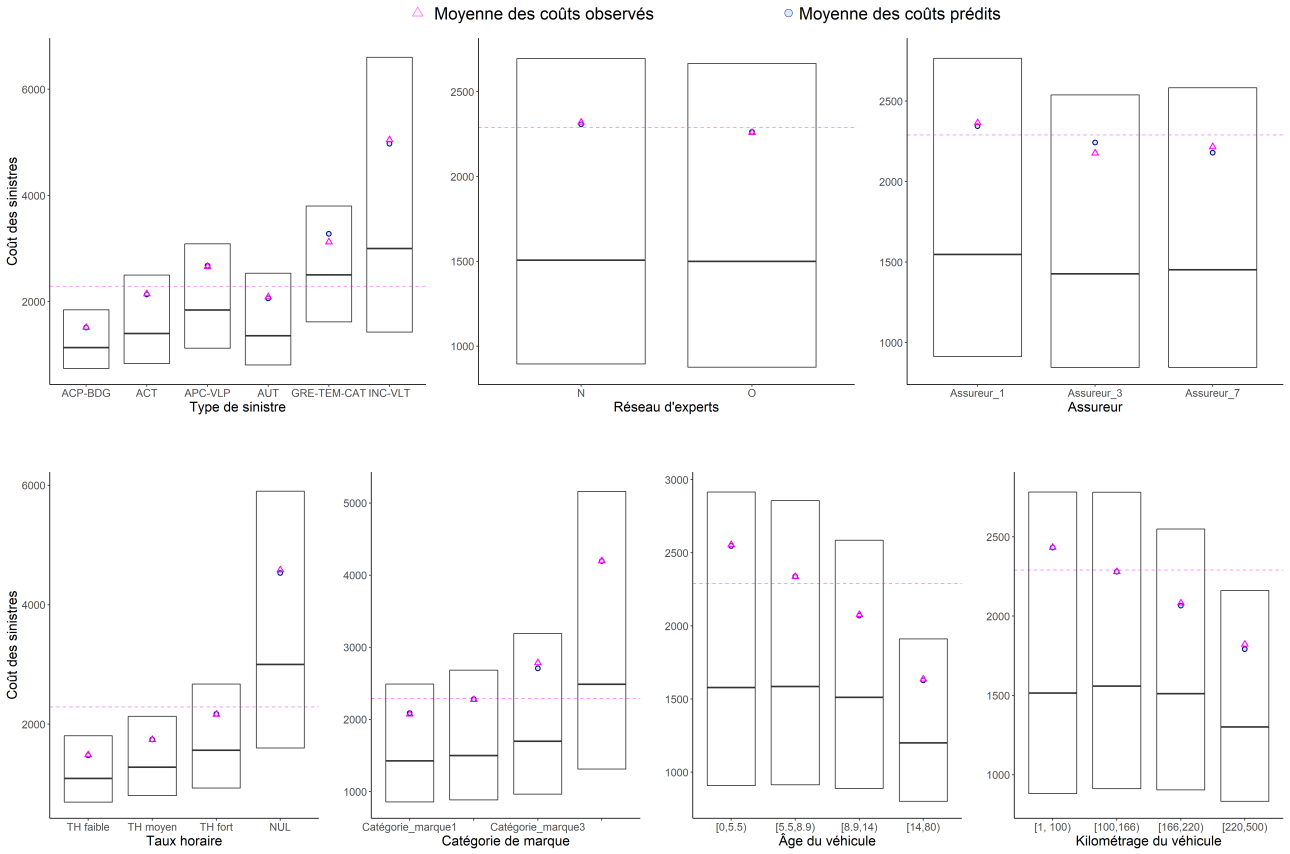


FIGURE B.1 : Moyennes des sinistres observés et prédits par le modèle GLM avec ajout d'interactions pour chaque modalité des variables explicatives

L'ajout d'interactions au modèle GLM a permis une amélioration de son pouvoir prédictif. En particulier, ce modèle possède un meilleur pouvoir de prédiction que le modèle CART, avec une amélioration de l'ordre de 6% de la MSE et de 4% de la MAE. De plus, le modèle GLM avec ajout d'interactions permet toujours une lecture aisée de l'influence du réseau d'experts sur la prédiction. Cependant, ce modèle possède encore un pouvoir prédictif plus faible que le modèle XGBoost et les modèles MOB avec et sans ajout d'interactions. Les écarts de performances de prédiction par rapport à ces trois modèles sont résumés dans le tableau (B.4) ci-dessous :

	Écart (%)	
	MSE	**MAE**
MOB	-2%	-1%
MOBI	-4%	-2%
XGB	-5%	-2%

TABLE B.4 : Écart des métriques d'évaluation des modèles MOB, MOB avec ajout d'interactions et XGBoost par rapport au modèle GLM avec ajout d'interactions

Bien que ces écarts puissent sembler relativement faibles, ils ne sont pas à négliger : dans un contexte de concurrence accrue, chaque optimisation de la tarification peut être intéressante pour un assureur, afin de proposer des tarifs toujours plus attractifs et compétitifs. Cet exemple d'utilisation du modèle XGBoost pour améliorer le pouvoir prédictif d'un modèle GLM permet tout de même d'illustrer que l'étude d'un modèle "boîte noire", via l'utilisation d'outils agnostiques, peut être un moyen efficace

d'améliorer les performances d'un modèle plus simple, tel que le modèle linéaire généralisé.

Annexe C

Présentation du modèle XGBoost

Cette annexe présente le modèle XGBoost et quelques éléments théoriques qui y sont associés.

C.1 Méthodes d'ensemble

Le principe de ces méthodes est d'utiliser les prédictions de plusieurs modèles, appelés "modèles faibles" ou bien "apprenants faibles" (pour *weak learner* en anglais) pour décider d'une prédiction finale. Un modèle (ou apprenant) faible se définit comme un modèle simple dont les prédictions sont légèrement meilleures que des prédictions purement aléatoires. L'idée est d'utiliser un ensemble d'apprenants faibles pour obtenir un modèle "fort" permettant d'obtenir une bonne qualité de prédiction. Les méthodes d'ensemble peuvent être divisées en deux catégories : les méthodes ensemblistes séquentielles et les méthodes ensemblistes parallèles. Lors de l'utilisation d'une méthode d'ensemble séquentielle les modèles sont entraînes les uns à la suite des autres en prenant en compte les erreurs commises par les modèles précédents. Un exemple de méthode d'ensemble séquentielle est l'algorithme Boosting. Dans le cadre d'une méthode ensembliste parallèle les modèles sont tous entraînés simultanément et la prédiction finale est obtenue par agrégation. Le Bagging, ou bien les forêts aléatoires sont des exemples de méthodes ensemblistes parallèles. Comme il peut être vu dans son nom, le modèle XGBoost repose sur le principe du Boosting présenté ci-après.

C.2 Boosting

Le Boosting est une méthode d'ensemble séquentielle. L'idée est d'utiliser une suite de modèles entraînés les uns à la suite des autres en prenant en compte l'erreur commise par le modèle précédent à chaque étape. La fonction de prédiction est de la forme :

$$f(x) = \sum_{k=1}^{K} f_k(x),$$

où K désigne le nombre d'apprenants faibles et f_k représente la règle de décision associée à l'apprenant faible k.

L'algorithme Boosting peut être utilisé aussi bien pour la classification que pour la régression. Le premier algorithme de Boosting a été introduit par Freund et Schapire, il s'agit du modèle Ada-

Boost ([Freund et Schapire (1997)]). De nombreuses versions de l'algorithme du Boosting ont été développées. Le modèle XGBoost s'appuie sur l'algorithme du Gradient Boosting présenté brièvement ci-après.

C.3 Gradient Boosting

Le Gradient Boosting ([Friedman (1999)], [Li (2016)]) désigne une famille d'algorithmes basés sur une fonction de perte convexe différentiable L. C'est un cas particulier de la méthode Boosting où le gradient de la fonction de perte est utilisé pour améliorer la prédiction. Dans le cas du Gradient Boosting les modèles "faibles" sont des arbres de décision. L'idée est toujours de trouver une fonction f telle que :

$$f = \arg\min_{f \in \mathcal{F}} L(y, f(x)),$$

Le Gradient Boosting s'inspire de la méthode de descente de gradient. Pour rappel, pour une fonction $g : \mathbb{R} \to \mathbb{R}$ convexe différentiable que l'on cherche à minimiser, cette méthode consiste à construire une suite $(x_k)_k$ convergeant vers x^*, où x^* est solution du problème de minimisation, *i.e.* $g(x^*) = \min_{x \in \mathbb{R}} g(x)$. Pour ce faire, l'algorithme part d'une valeur initiale x_0, puis à l'étape k de l'algorithme x_k est défini par la formule de récurrence suivante :

$$x_k = x_{k-1} - \lambda_k f'(x_{k-1}),$$

où λ_k est le pas calculé à l'étape k. Ce principe est utilisé dans l'algorithme du Gradient Boosting présenté ci-après.

Algorithm 1 Gradient Boosting

Initialisation Le modèle est initialisé avec une valeur constante $F_0(x) = \arg\min_z \sum_{i=1}^{n} L(y_i, z)$.

for k=1,...,K **do**

— Calculer les pseudos-résidus :

$$r_{i,k} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{k-1}(x)}, \qquad i = 1, ..., n$$

— Ajuster un arbre de régression A_k (apprenant faible) aux résidus. L'arbre A_k conduit à la création de n_{A_k} segments notés $R_{j,k},\ j = 1, ..., n_{A_k}$.

— Pour $j = 1, ..., n_{A_k}$, la valeur associée à la $j^{\text{ème}}$ feuille de l'arbre A_k est donnée par :

$$z_{j,k} = \arg\min_z \sum_{x_i \in R_{j,k}} L(y_i, f_{k-1}(x_i) + z),$$

— Mise à jour de la prédiction :

$$f_k(x) = f_{k-1}(x) + \eta A_k(x),$$

où $A_k(x) = \sum_{j=1}^{n_{A_k}} z_{j,k} \mathbf{1}_{\{x \in R_{j,k}\}}$ est la valeur retournée pour x par l'arbre A_k et η est un paramètre appelé taux d'apprentissage (*learning rate* en anglais) spécifié par l'utilisateur.

end for

Sortie La valeur finale de l'algorithme est donnée par $f_K(x)$.

La fonction de perte généralement utilisée est la fonction quadratique suivante :

$$L(y, f(x)) = \frac{1}{2}(y - f(x))^2, \tag{C.1}$$

En particulier, la dérivée de cette fonction est égale à :

$$\frac{\partial L(y, f(x))}{\partial f(x)} = -(y - f(x)),$$

ce qui implique en particulier :

$$\begin{aligned}
&r_{i,k} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{k-1}(x)} \Rightarrow\ r_{i,k} = y_i - f_{k-1}(x_i),\\
&\frac{\partial \sum_{i=1}^{n} L(y_i, z)}{\partial z} = 0 \ \Rightarrow\ n \times z - \sum_{i=1}^{n} y_i = 0 \ \Rightarrow\ z = \frac{1}{n}\sum_{i=1}^{n} y_i,\\
&\frac{\partial \sum_{x_i \in R_{j,k}} L(y_i, f_{k-1}(x_i) + z)}{\partial z} = 0 \ \Rightarrow\ n \times z - \sum_{i=1}^{n}(y_i - f_{k-1}(x_i)) = 0 \ \Rightarrow\ z = \frac{1}{n}\sum_{i=1}^{n} r_{i,k}.
\end{aligned}$$

Ainsi, lorsque la fonction de perte utilisée est celle définie par (C.1), l'algorithme est initialisé en prenant pour valeur la moyenne de la variable d'intérêt. À l'étape k les résidus sont calculés comme la différence entre la valeur observée et la valeur prédite à l'étape $k-1$ et les feuilles de l'arbre construit en k sont associées à la moyenne des résidus compris dans la feuille. La présentation de cet algorithme permet de mieux comprendre le fonctionnement de l'algorithme XGBoost.

C.4 Extreme Gradient Boosting : le modèle XGBoost

Le modèle XGBoost ([Chen et Guestrin (2016)]) peut être vu comme une complexification du Gradient Boosting. L'initialisation n'est plus la même puisque la première prédiction donnée par XGBoost est par défaut égale à 0,5. D'autre part, la fonction de perte L est complétée par un terme de régularisation pour obtenir la fonction objectif suivante :

$$\Psi(y, f(x)) = \underbrace{\sum_{i=1}^{n} L(y_i, f(x_i))}_{fonction\ de\ perte} + \underbrace{\sum_{k=1}^{K} \Omega(f_k)}_{régularisation},$$

où K désigne le nombre d'apprenants faibles utilisés dans le modèle et Ω est une fonction de la forme :

$$\Omega(f) = \gamma|T| + \frac{1}{2}\lambda\|w\|^2.$$

avec $|T|$ le nombre de feuilles de l'arbre associé à la règle de décision f,γ un paramètre permettant de régler la pénalisation des arbres de grande taille et λ un paramètre de régularisation. La variable $w \in \mathbb{R}^{|T|}$ est un vecteur représentant l'ensemble des valeurs associées aux feuilles de l'arbre étudié. La fonction de perte permet toujours de mesurer la qualité de l'ajustement du modèle aux données et le terme de régularisation est introduit afin de permettre un contrôle de la complexité du modèle, l'introduction de ce terme permettant notamment de limiter le risque de surapprentissage.

À l'étape k de l'algorithme, l'objectif est de minimiser

$$L_k = \sum_{i=1}^{n} L(y_i, \widehat{y}_i^{k-1} + f_k(x_i)) + \Omega(f_k), \tag{C.2}$$

par rapport à w, avec $\widehat{y}_i^{k-1}$ l'estimation associée à l'observation x_i donnée à l'étape $k-1$. Ainsi à l'étape k l'objectif est de résoudre le problème suivant :

$$\begin{aligned}\arg\min_w L_k &= \arg\min_w \sum_{i=1}^{n} L(y_i, \widehat{y}_i^{k-1} + f_k(x_i))) + \gamma|T_k| + \frac{1}{2}\lambda\|w\|^2 \\ &= \arg\min_w \sum_{i=1}^{n} L(y_i, \widehat{y}_i^{k-1} + f_k(x_i)) + \frac{1}{2}\lambda\|w\|^2,\end{aligned}$$

Afin de résoudre ce problème de minimisation, XGBoost utilise une approximation de Taylor du second dégré :

$$\arg\min_w L_k \simeq \arg\min_w \sum_{i=1}^{n} \left[L(y_i, \widehat{y}_i^{k-1}) + \underbrace{\frac{\partial L(y_i, \widehat{y}_i^{k-1})}{\partial \widehat{y}_i^{k-1}}}_{g_i} f_k(x_i) + \frac{1}{2} \underbrace{\frac{\partial^2 L(y_i, \widehat{y}_i^{k-1})}{\partial^2 \widehat{y}_i^{k-1}}}_{h_i} f_k(x_i)^2 \right] + \frac{1}{2}\lambda\|w\|^2,$$

où g_i et h_i désignent respectivement le gradient et la hessienne de la fonction de perte L évaluée en $(y_i, \widehat{y}_i^{k-1})$.

$$\arg\min_w L_k \simeq \arg\min_w \sum_{i=1}^{n} \left[g_i f_k(x_i) + \frac{1}{2} h_i f_k(x_i)^2 \right] + \frac{1}{2}\lambda\|w\|^2, \text{ car } L(y_i, \widehat{y}_i^{k-1}) \text{ ne dépend pas de } w.$$

L'arbre construit en k possède $|T_k|$ feuilles associées aux valeurs $w = (w_1, ..., w_{|T_k|})^T$. Cet arbre peut être décrit par un ensemble de règles d'affectation $q : \mathbb{R}^p \to \{1, ..., |T_k|\}$. Pour un individu i la fonction q prend en entrée l'ensemble des variables explicatives x_i de l'individu i et retourne le numéro de la feuille à laquelle l'individu est associé. Il est alors possible de définir pour $t = 1, ..., |T_k|$, $I_t = \{i | q(x_i) = t\}$, l'ensemble des indices des observations appartenant à la feuille t. Le problème de minimisation de la fonction L_k peut alors se réécrire :

$$\begin{aligned}&\arg\min_w \sum_{t=1}^{|T_k|} \left[\sum_{i \in I_t} g_i f_k(x_i) + \frac{1}{2} \sum_{i \in I_t} h_i f_k(x_i)^2 \right] + \frac{1}{2}\lambda \sum_{t=1}^{|T_k|} w_t^2 \\ = &\arg\min_w \underbrace{\sum_{t=1}^{|T_k|} \left[(\sum_{i \in I_t} g_i) w_t + \frac{1}{2} (\sum_{i \in I_t} h_i + \lambda) w_t^2 \right]}_{s(w)},\end{aligned}$$

La solution w^* est obtenue en résolvant pour tout $t \in \{1, ..., |T_k|\}$:

$$\begin{aligned} & s'(w_t^*) = 0 \\ \iff & \sum_{i \in I_t} g_i + (\sum_{i \in I_t} h_i + \lambda) w_t^* = 0 \\ \iff & w_t^* = -\frac{\sum_{i \in I_t} g_i}{\sum_{i \in I_t} h_i + \lambda}, \end{aligned}$$

Lorsque la fonction de perte utilisée est celle définie par (C.1), on a :

$$\begin{aligned} g(y, f(x)) &= \frac{\partial}{\partial f(x)} \left[\frac{1}{2} (y - f(x))^2 \right] = -(y - f(x)) \\ h(y, f(x)) &= \frac{\partial^2}{\partial^2 f(x)} \left[\frac{1}{2} (y - f(x))^2 \right] = \frac{\partial}{\partial f(x)} g(y, f(x)) = 1. \end{aligned}$$

D'où,

$$\forall t \in \{1, ..., T\}, w_t^* = \frac{\sum_{i \in I_t} (y_i - \widehat{y}_i^{k-1})}{|I_t| + \lambda}.$$

Ainsi la valeur associée à la feuille t est la somme des résidus présents dans la feuille, divisé par le nombre d'observations au sein de la feuille augmenté de λ.

Afin de savoir si la séparation d'un nœud h est effectuée, le modèle XGBoost s'appuie sur un score de similarité (*similarity score* en anglais) défini selon la formule suivante :

$$sim_h = \frac{(\sum_{x_i \in h} g_i)^2}{\sum_{x_i \in h} h_i + \lambda}.$$

Pour chaque séparation possible du nœud h, le modèle calcule le gain effectué selon la formule suivante :

$$Gain_h^r = sim_{h_g} + sim_{h_d} - sim_h,$$

où h_g et h_d sont les nœuds fils du nœud h créés selon la règle de décision r. La séparation choisie est celle maximisant le gain, elle sera notée r^*. Pour savoir si la séparation du nœud h selon la règle r^* est effectuée le gain est comparé à γ :

— si $Gain_h^{r^*} - \gamma < 0$, alors le nœud h n'est pas divisé et devient une feuille ;

— si $Gain_h^{r^*} - \gamma > 0$, alors le nœud h est divisé selon la règle de division r^*.

Ainsi, le paramètre γ est permet de plus ou moins pénaliser les arbres de grandes tailles.