Assurances et gestion des risques
Insurance and Risk Management

Mesure d'impact d'une variable binaire sur une réponse quantitative dans un cadre non paramétrique

Auriol Wabo and Frédéric Planchet

Volume 87, Number 1-2, July 2020

URI: https://id.erudit.org/iderudit/1070751ar
DOI: https://doi.org/10.7202/1070751ar

See table of contents

Publisher(s)

Faculté des sciences de l'administration, Université Laval

ISSN

1705-7299 (print)
2371-4913 (digital)

Explore this journal

Cite this article

Wabo, A. & Planchet, F. (2020). Mesure d'impact d'une variable binaire sur une réponse quantitative dans un cadre non paramétrique. *Assurances et gestion des risques / Insurance and Risk Management*, *87*(1-2), 33–68. https://doi.org/10.7202/1070751ar

Article abstract

On propose dans cet article une méthode de quantification de l'effet d'une variable explicative qualitative sur une réponse quantitative plus souple à utiliser que le simple coefficient d'un modèle GLM multiplicatif ; l'objectif est de disposer d'une mesure ne nécessitant pas l'hypothèse de proportionnalité du GLM et complètement décorrélée de l'effet des autres variables explicatives incluses dans le modèle. L'approche est illustrée à l'aide de données de coûts de sinistres matériels automobiles, pour lesquelles on cherche à quantifier l'impact de l'expert qui a évalué le sinistre sur le montant de l'évaluation.

Assurances et gestion des risques, Vol. 87 (1-2)
Insurance and Risk Management, Vol. 87 (1-2)

MESURE D'IMPACT D'UNE VARIABLE BINAIRE SUR UNE RÉPONSE QUANTITATIVE DANS UN CADRE NON PARAMÉTRIQUE

Auriol WABO[1] et Frédéric PLANCHET[2]

ISFA – Laboratoire SAF[3]
Université de Lyon – Université Claude Bernard Lyon 1

RÉSUMÉ

On propose dans cet article une méthode de quantification de l'effet d'une variable explicative qualitative sur une réponse quantitative plus souple à utiliser que le simple coefficient d'un modèle GLM multiplicatif ; l'objectif est de disposer d'une mesure ne nécessitant pas l'hypothèse de proportionnalité du GLM et complètement décorrélée de l'effet des autres variables explicatives incluses dans le modèle. L'approche est illustrée à l'aide de données de coûts de sinistres matériels automobiles, pour lesquelles on cherche à quantifier l'impact de l'expert qui a évalué le sinistre sur le montant de l'évaluation.

1. Introduction

L'assureur est souvent amené à devoir mesurer l'impact d'une variable explicative binaire sur une réponse quantitative : les assurés arrivés par un canal donné présentent-ils une meilleure sinistralité que les autres ? Les possesseurs d'un deux-roues ont-ils une valeur client inférieure à ceux n'ayant assuré qu'une voiture ? Fournir une réponse à ce type de questions nécessite de corriger des effets d'autres variables influençant le risque, dès lors que les typologies d'assurés ou de sinistres diffèrent pour les deux modalités de la caractéristique considérée.

En assurance IARD, l'utilisation de modèles de régression de type GLM (cf. Besson et Partrat [2004]) permet de décorréler les différentes variables et fournit donc une réponse simple, sous la forme d'un unique coefficient synthétisant l'effet de la variable, dès lors que l'on retient une fonction de lien logarithme.

La validité de ce coefficient suppose toutefois que l'hypothèse de proportionnalité des effets soit vérifiée, l'espérance conditionnelle de la réponse s'écrivant comme un produit de coefficients attachés chacun aux modalités des variables, ce qui peut s'avérer assez contraignant. Lorsque cette hypothèse n'est pas vérifiée et que l'effet relatif de la variable d'intérêt varie d'un segment à l'autre, il est nécessaire de construire d'autres mesures et la réalisation d'un GLM par segment conduit à augmenter la volatilité des estimateurs, ce qui en fait une solution peu efficace lorsque le nombre de segments augmente.

On propose dans le présent travail deux approches non paramétriques, basées sur le *gradient boosting* (Ridegway [1999]) et les valeurs SHAP (*cf.* Shapley [1953] et Lundberg et Lee [2017]), pour construire cette mesure d'influence.

Afin de pouvoir construire la mesure d'influence, on doit préalablement reconstruire le modèle de prédiction des coûts des sinistres considérés ici avec des techniques adaptées, en pratique le *gradient boosting*. Pour cela, on se place dans le contexte de l'étude réalisée par De Lussac [2018] et prolongée par Khougea [2019], consistant à mesurer l'impact d'un réseau d'experts sur le coût d'un sinistre matériel automobile.

Le présent papier est organisé de la manière suivante : dans une première partie, on rappelle le contexte de travail (section 2) et les algorithmes de construction des prédicteurs servant de base aux mesures d'influence sont décrits (section 3). La section 4 est consacrée à la description des deux mesures d'influence proposées. Une comparaison des différentes approches avec les résultats issus de l'utilisation

d'un GLM est proposée à la section 5, avant de conclure et de formuler une préconisation (section 6). Le lecteur intéressé par une présentation plus large de ces travaux peut se reporter à Wabo [2019].

2. Notations et contexte de l'étude

La présente étude s'inscrit dans le contexte suivant, aisément transposable à de nombreuses autres situations : on cherche à mesurer l'influence de deux réseaux d'experts, A et B, sur les coûts de sinistres matériels automobiles, pour déterminer le réseau conduisant, toutes choses égales par ailleurs, au coût le plus faible, que l'on qualifiera de « plus performant[4] ». Dans la suite, on privilégiera arbitrairement l'écart de coût de A par rapport à B.

L'objectif de la modélisation est de mesurer l'écart de performance entre deux réseaux d'experts lors de l'évaluation des coûts de sinistres matériels. Cette évaluation consiste à reconstituer les coûts de sinistres (variable réponse) à l'aide de neuf caractéristiques (variables explicatives) qui fournissent des informations concernant le véhicule (kilométrage, âge et marque), le type d'accident, le taux horaire du garage, le lieu et la date de survenance de l'accident, l'assureur ayant pris en charge le sinistre et le réseau d'experts qui est intervenu, dont la variable a été nommée « cible ».

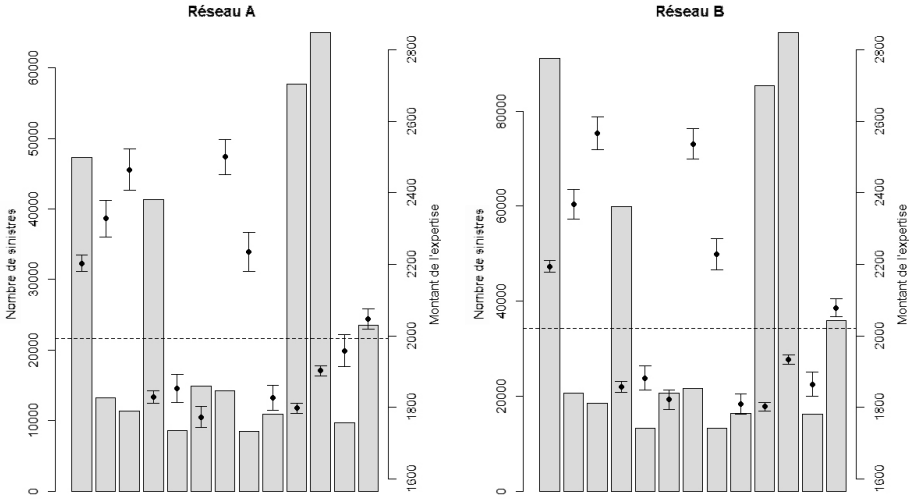
On présente ci-après quelques résumés statistiques des données utilisées pour les illustrations numériques[5].

- lignes discontinues jaunes : coût moyen des sinistres expertisés par le réseau A (ou B) ;
- points en jaunes : coûts moyens pour une modalité considérée, les barres d'erreur associées représentant l'intervalle de confiance[6] associé à 95 % ;
- diagrammes en gris : nombre de sinistres par modalité.

Pour chaque réseau, la figure 1 renseigne la composition du portefeuille de sinistres pour chaque marque de véhicule et le coût moyen associé. On compte en tout treize marques dont douze marques principales (AUDI, PEUGEOT, BMW, *etc.*), les autres marques ayant été regroupées dans une modalité nommée AUTRE.

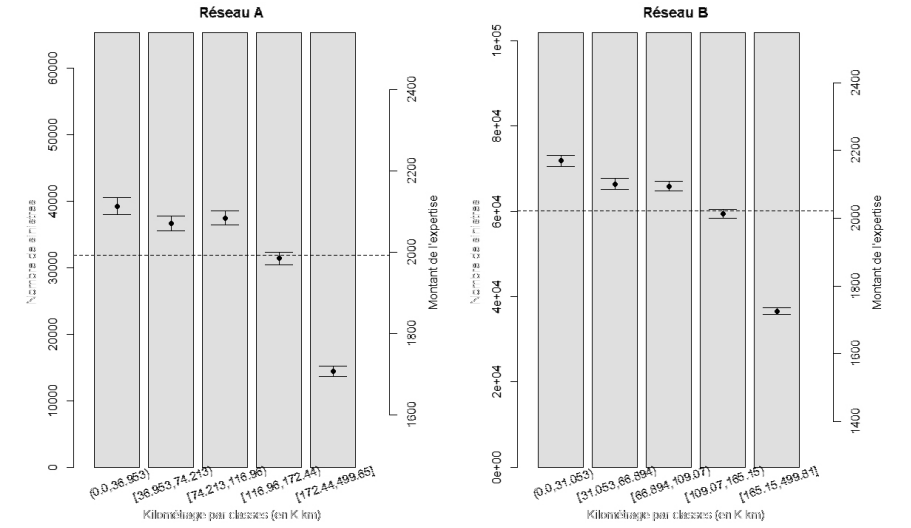
On observe une similitude entre les deux réseaux.

FIGURE 1 *Distribution du portefeuille des réseaux en fonction de la marque du véhicule*



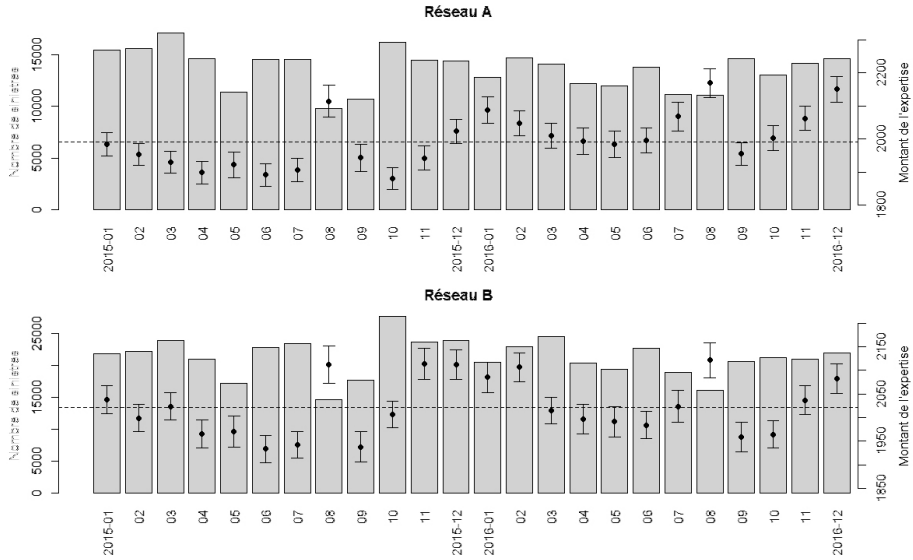
Le kilométrage a été catégorisé en cinq classes à l'aide des quantiles. Cette discrétisation ne concerne que la représentation graphique précédente. En général, plus un véhicule est ancien, plus il a probablement parcouru une longue distance, d'où la décroissance du coût moyen observée ci-dessous (figure 2).

FIGURE 2 *Distribution du portefeuille des réseaux en fonction du kilométrage*



La figure 3 renseigne une évolution des sinistres assez similaire pour les deux réseaux.

FIGURE 3 *Distribution du portefeuille des réseaux en fonction de la période de survenance de l'accident*



Mesurer l'écart de performance en comparant leurs coûts moyens de sinistres n'est valide que si les sinistres expertisés par les réseaux sont homogènes, ce qui n'est pas le cas en pratique. C'est pourquoi il est nécessaire de prendre en compte les dissimilarités relatives à la nature des sinistres expertisés par chaque réseau. Pour cela, on construit un modèle de prédiction des coûts de sinistres en fonction de leurs caractéristiques.

Une première approche possible consiste à utiliser un GLM à fonction de lien logarithmique. L'écart de performance obtenu est alors de –2,4% en faveur du réseau A (cf. De Lussac [2018]). Cet auteur a toutefois mis en évidence le fait que l'hypothèse de proportionnalité des effets du GLM était discutable et qu'il était préférable, pour modéliser le coût moyen, de recourir au *Gradient Boosting Model*.

3. Le modèle de prédiction : *Gradient Boosting Models* (GBM)

Les études réalisées par De Lussac [2018] et Khougea [2019] avaient principalement pour objectif de reconstituer des coûts de sinistres dans un contexte IARD (automobile) avec un modèle prédictif. Afin de trouver le modèle prédictif le mieux adapté, ces derniers ont comparé plusieurs modèles tels que le *Generalized Linear Model* (GLM), le *Random Generalized Linear Model* (RGLM), le *Gradient Boosting Model* et le réseau de neurones RVFL (*Random Vector Functional Link*).

La comparaison de ces modèles s'est faite sur la base de l'erreur quadratique moyenne (*Mean Square Error*) et de l'erreur L1 obtenues sur l'échantillon test :

$$MSE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2 \quad \text{Erreur } L_1 = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left| y_i - \hat{y}_i \right|$$

n_{test} est le nombre d'observations de l'échantillon test. Pour $i \in \{1, \ldots, n_{\text{test}}\}$ y_i et $\hat{y}_i$ représentent respectivement les valeurs observées et prédites.

Le modèle prédictif le plus performant est le GBM (cf. Ridgeway [1999]) avec une erreur quadratique moyenne de 2 411 535 et une erreur L1 de 651. En comparaison aux autres modèles, le GBM est celui qui a obtenu les erreurs les plus faibles. C'est pour cette raison que ce modèle a été retenu pour la modélisation de cette étude.

Néanmoins, des écarts significatifs entre les valeurs prédites et observées restent relevés. Ceux-ci s'expliquent en partie par l'insuffisance des variables à disposition pour la construction d'un modèle prédictif de qualité, qui conduit à un manque d'informations pour différencier certains sinistres en apparence similaire mais dont les montants d'expertise sont sensiblement différents. Ces écarts ne remettent toutefois pas en question la cohérence et l'intérêt de l'étude, basée sur la mesure d'un écart relatif.

Avant de détailler le paramétrage effectué pour construire le GBM optimal, commençons par présenter le cadre de travail.

a. Construction des prédicteurs

Le *Gradient Boosting Models* (GBM) est une famille d'algorithmes basée sur le *boosting*[7] et le gradient d'une fonction de perte supposée convexe et différentiable. Leur principe de base est de construire une séquence de modèles adaptatifs de sorte qu'à chaque étape, chaque modèle ajouté à la combinaison des modèles précédents apparaisse comme un pas vers une meilleure solution. Ce pas est franchi dans la direction du gradient de la fonction de perte afin d'améliorer les propriétés de convergence.

L'objectif du GBM est de construire un modèle agrégé f minimisant la fonction de perte quadratique L.

$$f(x) = \underset{\psi}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(y_i - \psi(x_i) \right)^2 = \underset{\psi}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} L\left(y_i, \psi(x_i) \right)$$

Parmi les nombreuses versions du *boosting*, celle utilisée dans cette étude est le *Stochastic Gradient Boosting* qui diffère du *Gradient Boosting* classique par l'ajout d'un tirage aléatoire sans remise de taille $\tilde{n}(< n)$ à chaque itération pour la création d'un modèle adaptatif. Dans Friedman [2002], plusieurs cas traités ont montré la pertinence de ce paramètre. L'estimation $\hat{f}$ de f est obtenue selon l'algorithme standard figurant en annexe.

Les algorithmes du GBM implémentés sous R (*gbm*) et sous Python (*scikit-learn*) sont relativement faciles à utiliser lorsqu'on s'en tient à des données non massives ou à une analyse simple basée sur les paramètres par défaut (le nombre d'itérations maximal M, la proportion d'observations sélectionnées aléatoirement p, la profondeur maximale des arbres de régression K et le taux d'apprentissage ou *shrinkage* λ). Néanmoins, ces deux cas sont rares dans la réalité. C'est pour cette raison qu'il est nécessaire de trouver une alternative avec un algorithme optimisé en termes de gestion de la mémoire et permettant une parallélisation poussée des calculs : le package GBM du module H2O répond à ces contraintes et, dans cette étude, tous les algorithmes du GBM ont été exécutés à l'aide de ce module.

Par la suite, nous résumerons la méthodologie utilisée et les résultats obtenus (*cf.* De Lussac [2018]). L'idée est d'appliquer le modèle une première fois avec des paramètres par défaut, ensuite les optimiser pour ainsi obtenir le modèle prédictif optimal.

b. Le modèle prédictif utilisé

Le paramétrage d'un modèle dépend de deux principales contraintes :

- la complexité du modèle : l'algorithme doit s'exécuter en un temps raisonnable tout en fournissant de bons résultats. La complexité est majoritairement contrôlée par le nombre d'itérations de l'algorithme.
- le surapprentissage : l'algorithme va naturellement se mettre en adéquation avec l'échantillon d'apprentissage au fur et à mesure des itérations, d'où la nécessité de définir un critère d'arrêt associé à un apprentissage lent afin que chaque itération n'apporte qu'une légère modification du modèle précédent.

Dans cette étude, le critère d'arrêt choisi est le contrôle du nombre d'itérations par l'erreur moyenne quadratique sur l'échantillon de validation. Plus précisément, l'algorithme s'arrête dès lors que la MSE sur l'échantillon de validation n'augmente plus après 50 itérations. Il est également possible de ralentir l'apprentissage du modèle en diminuant le taux d'apprentissage et/ou la profondeur maximale des arbres de régression.

La première application de l'algorithme du GBM a été faite sous R avec les paramètres suivants : $M = 10\,000$, $p = 50\%$, $K = 5$ et $\lambda = 0{,}1$. Les erreurs de prédiction obtenues sur l'échantillon test sont les suivantes : MSE = 2 477 139 et Erreur L1 = 661. Pour information, le nombre d'itérations qui a été effectué est 2 822 et le temps d'exécution[8] est de 1 minute et 31 secondes.

Les résultats obtenus sont issus d'un paramétrage par défaut. Or, une bonne qualité de prévision dépend d'une configuration optimale des paramètres, réalisée à l'aide d'une méthode dite « recherche par grille » : des valeurs sont proposées pour chaque paramètre, créant ainsi de nombreuses configurations possibles qui seront toutes parcourues pour ainsi obtenir la meilleure configuration. Cette optimisation est assez délicate pour deux raisons :

- l'existence de nombreux paramètres : en plus des quatre paramètres présentés précédemment, il en existe d'autres tels que le nombre de variables à sélectionner aléatoirement, le nombre minimal d'observations par feuille, etc. ;
- la majorité des paramètres sont liés : ajuster la valeur de l'un modifie la valeur de l'autre.

Le nombre de paramètres étant élevé et certains d'entre eux jouant des rôles redondants sur le contrôle du surapprentissage, il n'est pas nécessaire de tous les optimiser. Seuls sept paramètres furent configurés (cf. annexe f).

Afin de trouver les valeurs optimales des paramètres, il faudrait d'abord au préalable avoir une idée de l'ensemble fini de valeurs où choisir la valeur adaptée du modèle. Pour chaque paramètre, l'intervalle de recherche associé a été obtenu grâce à de nombreux tests effectués. Ces derniers consistent à calculer la MSE du modèle sur l'échantillon de validation suivant plusieurs valeurs suffisamment distantes afin de repérer (à dire d'expert) un ensemble de valeurs où la moyenne des MSE calculées est la plus faible. L'ensemble résultant constitue l'ensemble idéal pour la recherche par grille.

Le modèle prédictif utilisé dans cette étude étant rappelé, il est maintenant temps de répondre à la problématique à travers les deux approches présentées ci-après.

4. Mesures d'influence

Les mesures d'influence proposées sont d'abord construites de manière statique, pour fournir une estimation, sur une période donnée, de l'impact sur le niveau de la réponse de la variable binaire d'intérêt. Dans un second temps, on s'intéresse à l'utilisation de ces estimations dans un cadre dynamique, en les réalisant chaque trimestre sur une période de deux ans, afin de voir si des tendances se dégagent.

Les écarts entre les modélisations proposées seront donc analysés en niveau et en tendance.

c. Première approche : imputation des réponses manquantes

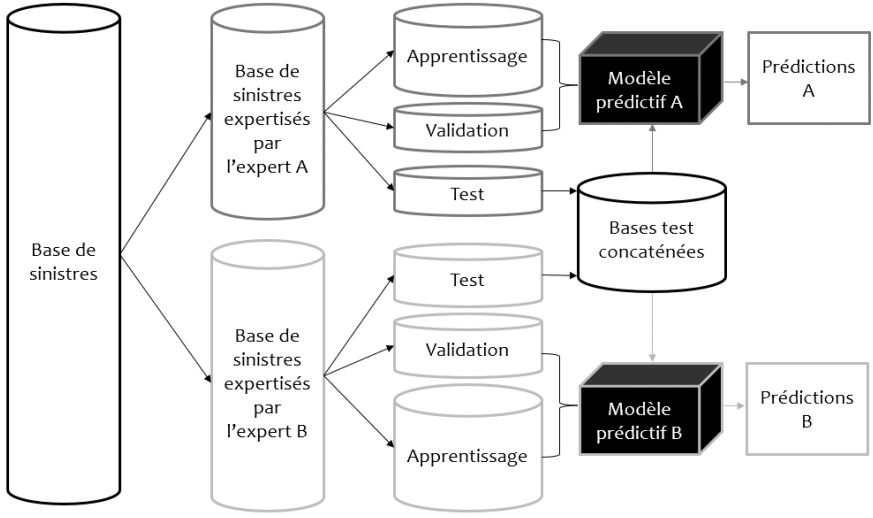
Cette approche consiste à comparer les deux réseaux sur une base de sinistres unique, pour éliminer les biais d'hétérogénéité. On ne dispose toutefois pas d'une base de sinistres où les deux réseaux sont intervenus, un sinistre étant expertisé par A ou B mais jamais par A et B. Il convient donc d'estimer le montant du coût manquant pour chaque sinistre de la base considérée. Pour cela, on choisit de faire correspondre à chaque réseau un modèle construit sur la base des sinistres

expertisés par ce dernier. À l'aide de ces modèles prédictifs, on disposera en tout de 3 montants par sinistre : le montant observé, le montant prédit par le réseau A, et celui prédit par le réseau B.

La construction des deux modèles prédictifs est effectuée en séparant les sinistres du réseau A et ceux du réseau B, pour ainsi obtenir des modèles construits sans la variable « cible[9] » (comme effectué ci-dessous). Ce processus ne fonctionne que si les sinistres expertisés présentent des similitudes. Il aurait été également possible de réaliser la construction des deux modèles avec un seul modèle prédictif intégrant la variable « cible ». Ce processus a pour avantage d'être utilisable même si les sinistres expertisés sont totalement différents pour chaque réseau. Dans le présent travail, on retient l'approche à deux modèles, mais le choix de l'une ou l'autre approche est indifférent.

La prédiction des coûts par les deux réseaux s'est déroulée comme suit :

FIGURE 4 *Étapes de la prédiction des coûts de sinistres des réseaux A et B*



La division des échantillons s'est faite suivant les proportions 3/5 (pour l'échantillon d'apprentissage), 1/5 (échantillon de validation) et 1/5 (échantillon de test). Les bases de test concaténées constituent la base de sinistres sur laquelle la comparaison a lieu.

Deux possibilités de comparaison sont proposées :

- la première consiste à comparer les réseaux sur la base des coûts prédits pour chaque réseau, le coût observé étant pris pour une référence sans intervenir directement dans l'appréciation de l'écart de coût ;
- la seconde consiste à comparer les deux réseaux sur la base de leurs coûts observés et prédits, la valeur prédite venant simplement compléter les données lorsqu'elle n'est pas fournie.

Comparaison sur une même sinistralité – méthode n° 1

Ici, l'évaluation est faite à partir des prédictions effectuées et sur la base des mesures *M* et *Err* définies ci-dessous. Cette méthode nécessite *a priori* de définir des montants de référence, c'est-à-dire des montants auxquels s'attend l'assureur. Ainsi on regarde le réseau (à travers son modèle prédictif construit) ayant obtenu en moyenne le coût de sinistres le plus faible tout en ayant les coûts reconstitués les plus proches de ceux attendus.

$$M(\hat{y}_1, \ldots, \hat{y}_n) = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i,$$

$$\mathrm{Err}\left((y_1, \hat{y}_1), \ldots, (y_n, \hat{y}_n)\right) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (4.1)$$

où y_i et $\hat{y}_i$ pour $i \in \{1, \ldots, n\}$ correspondent respectivement aux montants observés et prédits. n est le nombre de sinistres contenus dans la base test résultante.

En effectuant une simple application numérique des mesures précédemment définies aux coûts de sinistres, on obtient :

$$M(\hat{y}_1^A, \ldots, \hat{y}_n^A) = 2\,002{,}43\ € \qquad Err\left((y_1, \hat{y}_1^A), \ldots, (y_n, \hat{y}_n^A)\right) = 2\,453\,265$$

et

$$M(\hat{y}_1^B, \ldots, \hat{y}_n^B) = 2\,016{,}95\ € \qquad Err\left((y_1, \hat{y}_1^B), \ldots, (y_n, \hat{y}_n^B)\right) = 2\,417\,710$$

On observe une valeur moyenne des coûts de sinistres prédits par le modèle du réseau A inférieure à celle des coûts prédits par le réseau B. Cette observation permet de conclure que la présence du réseau A diminue en moyenne les coûts d'environ 0,72 % (il s'agit là d'une interprétation similaire à celle des modèles linéaires) ou encore, l'écart de performance est de –0,72 %. Le calcul de ce coefficient est détaillé à la section 4.

Toutefois, la mesure d'erreur obtenue pour le réseau A est supérieure à celle du réseau B. En prenant en compte le fait que le coût moyen des sinistres observés est de 2 006,93 €, on peut dire que le réseau A a été plus performant que le réseau B, car le coût moyen de ce dernier est plus éloigné du coût moyen observé comparativement au réseau A.

Néanmoins, on ne peut pas négliger le fait que ces calculs ont été effectués à partir de toutes les prédictions des modèles (bonnes et mauvaises). C'est pour cela que la comparaison est à nouveau faite, mais cette fois-ci en fonction de la qualité des prédictions. Pour ce faire, on applique un *K-means* pour construire des classes homogènes de sinistres vis-à-vis de la mesure d_1 définie ci-dessous :

$$d_1(y_i, \hat{y}_i^A, \hat{y}_i^B) = \max\left(\left|\frac{y_i - \hat{y}_i^A}{\hat{y}_i^A}\right|, \left|\frac{y_i - \hat{y}_i^B}{\hat{y}_i^B}\right|\right) \qquad (4.2)$$

où y_i, $\hat{y}_i^A$ et $\hat{y}_i^B$ correspondent respectivement aux coûts de sinistres observés, prédits par le modèle A et prédits par le modèle B.

La classification *via* l'algorithme des *K-means* nécessite un nombre de classes à définir avant l'exécution de l'algorithme. Nous avons déterminé le nombre de classes de sorte que l'inertie expliquée[10] soit supérieure à 95 %, ce qui nous conduit à en retenir 10, homogènes pour la mesure d_1. Ces classes sont numérotées de 1 à 10 par ordre croissant des centres de classe.

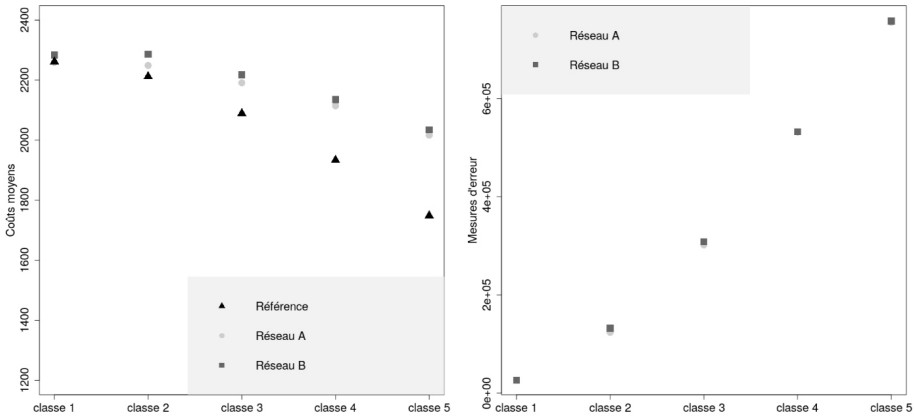
TABLEAU 1 *Valeurs des centres des classes construites*

	NUMÉROS DES CLASSES									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Valeurs des centres	0,06	0,13	0,21	0,28	0,35	0,43	0,50	0,58	0,91	3,69

Compte tenu du fait que dans cette première méthode, on considère que les estimations fournies par chacun des deux modèles prédictifs correspondent aux montants qu'auraient fournis les réseaux après expertise, il a été jugé raisonnable de s'intéresser uniquement aux sinistres contenus dans les cinq premières classes. Pour ces classes, les applications numériques des mesures *M* et *Err* ont été effectuées. Cette procédure permet de limiter les biais associés à la qualité de prédiction moyenne du modèle considéré globalement du fait du nombre limité (10) de variables explicatives du modèle.

Les résultats obtenus sont les suivants :

FIGURE 5 *Mesures* M *et* Err *calculées dans les cinq meilleures classes*



Le graphe de gauche ci-dessus renseigne par classe trois coûts moyens, celui des sinistres observés et celui des sinistres prédits par les modèles A et B. On observe que les estimations fournies par le réseau A sont en moyennes inférieures à celles fournies par le réseau B dans chaque classe. De plus, le graphe de droite indique un écart croissant entre les coûts moyens des réseaux et le coût moyen attendu. Cet écart croissant peut s'expliquer par la moins bonne qualité des prévisions au fil des classes. Seul dans la classe 1, le réseau B a une mesure d'erreur plus faible que celle du réseau A.

De manière équivalente à celle utilisée pour obtenir l'écart de performance précédente de –0,72% , ici on obtient –1,26% en ne prenant en compte que les classes 1 à 5. C'est-à-dire que le réseau A fait baisser en moyenne les coûts de sinistres de 1,26% en comparaison au réseau B. Les analyses globale (en considérant toutes les prédictions) ou restreinte (en considérant les meilleures prédictions) conduisent à la même conclusion, avec une différence d'appréciation de l'ampleur de l'écart.

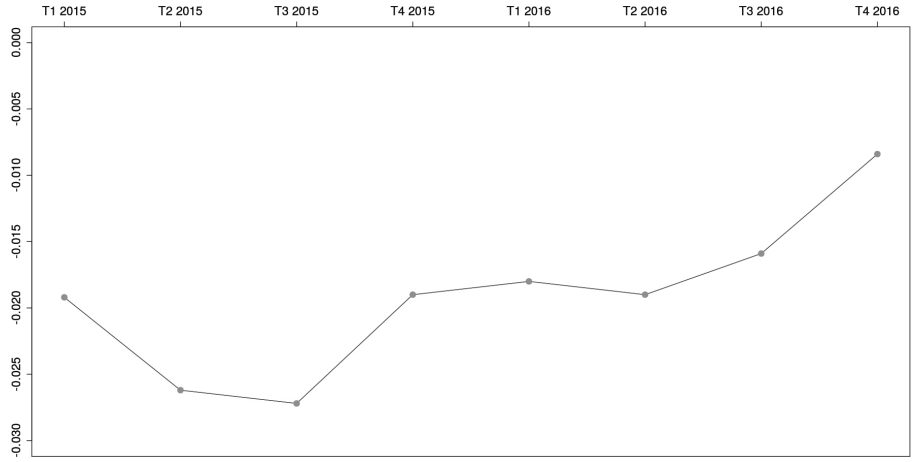
On étudie maintenant l'évolution temporelle de l'écart de performance pour voir si une tendance en ressort. L'étude temporelle a lieu sur les trimestres des années 2015 et 2016, et ne concerne que les meilleures prédictions. Le tableau suivant résume l'écart de performance par trimestre.

TABLEAU 2 *Mesure de l'écart de performance par trimestre dans les cinq meilleurs clusters*

	2015				2016			
	T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**
Écart (en %)	–1,92	–2,62	–2,72	–1,90	–1,80	–1,90	–1,59	–0,84

Le graphique correspondant est le suivant :

FIGURE 6 *Évolution de l'écart de performance au fil des trimestres dans les cinq meilleurs clusters*



On observe une tendance croissante, synonyme de dégradation de la performance, ce point sera commenté plus avant *infra*.

Comparaison sur une même sinistralité – méthode n° 2

Contrairement à la méthode précédente qui nécessite *a priori* un montant de référence, dans celle-ci nous comparons les deux réseaux sur la base des coûts de sinistres observés et des coûts de sinistres prédits. Par exemple, pour un sinistre expertisé par le réseau A, le montant observé est celui qui figure dans la base de données et le montant prédit est le montant issu du modèle B.

On calcule donc le coût qu'aurait obtenu l'autre réseau s'il était intervenu à la place du réseau intervenu en réalité. Ce calcul s'effectue *via* le modèle prédictif du réseau correspondant. Cette approche présente par construction l'intérêt d'éliminer le biais pour les valeurs observées.

La comparaison ici se fait à l'aide de la mesure suivante :

$$M'(y_1, \ldots, y_m, \hat{y}_{m+1}, \ldots, \hat{y}_n) = \frac{1}{n}(\sum_{k=1}^{m} y_k + \sum_{k=m+1}^{n} \hat{y}_k) \qquad (4.3)$$

où y_i pour $i \in \{1, \ldots, m\}$ correspondent aux coûts observés et les $\hat{y}_i$ pour $i \in \{m + 1, \ldots, n\}$ correspondent aux coûts de sinistres prédits. Avec m qui est égal au nombre de sinistres expertisés par un réseau (ceux observés) et n est le nombre de sinistres contenus dans la base test (celle sur laquelle les réseaux sont évalués).

On obtient :

$$M'(y_1^A, \ldots, y_m^A, \hat{y}_{m+1}^A, \ldots, \hat{y}_n^A) = 1\,998{,}88\ €$$

$$M'(y_1^B, \ldots, y_m^B, \hat{y}_{m+1}^B, \ldots, \hat{y}_n^B) = 2\,016{,}98\ €$$

Les résultats montrent que le réseau A présente en moyenne un coût plus faible que celui du réseau B. Ce qui fait du réseau A le plus performant des deux. Plus précisément, nous pouvons conclure que la présence de ce dernier diminue en moyenne le coût d'environ 0,90 %, soit 0,18 point de plus que la mesure obtenue dans la première méthode.

Ce résultat est obtenu sur la base des bonnes et des mauvaises prédictions. Mais on ne dispose pas d'une technique suffisamment fiable et pertinente pour sélectionner les meilleures prédictions. Cette limite peut être assimilable aux prédictions *via* un modèle linéaire, notamment le calcul de toutes les prédictions est effectué à partir des mêmes coefficients de régression.

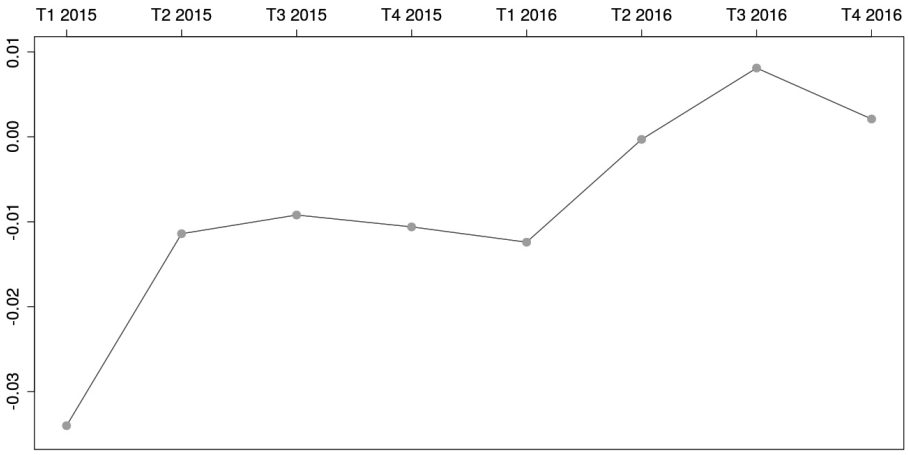
Comme dans la précédente méthode, il est utile d'étudier l'évolution trimestrielle de l'écart de performance. Dans ce but, la mesure définie en (4.3) a été appliquée aux coûts des réseaux trimestre par trimestre :

TABLEAU 3 *Mesure de l'écart de performance par trimestre*

	2015				2016			
	T1	T2	T3	T4	T1	T2	T3	T4
Écart (en %)	–3,40	–1,14	–0,92	–1,06	–1,24	–0,03	0,81	0,21

Une vision graphique de ces écarts donne :

FIGURE 7 *Évolution de l'écart de performance au fil des trimestres*



Les résultats ci-dessus indiquent une tendance à un resserrement de l'écart entre les deux réseaux, cette fois-ci pas toujours en faveur du réseau A comme dans la première méthode.

Les deux méthodes conduisent à une même conclusion globale qui est en faveur du réseau A. La deuxième méthode, elle présente pour principale limite de ne pas pouvoir se restreindre aux bonnes prédictions faites par le modèle. Les deux méthodes fournissent de meilleures prédictions que celles d'un modèle linéaire et renseignent (à l'aide d'un calcul simple qui est détaillé à la section 4) l'écart de performance entre les réseaux, cet écart pouvant être calculé globalement ou sur n'importe quel segment, ce que le GLM ne permet pas de réaliser.

d. Seconde approche : utilisation des valeurs SHAP

La démarche pragmatique proposée ci-dessus, si elle est simple à mettre en œuvre, présente toutefois une limite importante, qui est que la mesure d'écart qui en résulte dépend de la structure d'hétérogénéité des ensembles de sinistres sur lesquels elle est calculée. Pour contourner cette difficulté, on va ici calculer les contributions des variables pour une valeur prédite, ceci afin d'isoler mieux celle de la variable « cible ».

Les contributions sont calculées à l'aide du *Kernel* SHAP, méthode issue de la valeur de Shapley (*cf.* Shapley [1953]). Cette dernière offre une répartition équitable entre les contributions des variables et la théorie mathématique solide du *Kernel* SHAP.

Le *Kernel* SHAP est une technique qui permet d'expliquer une prédiction d'un modèle d'apprentissage automatique, en construisant localement un modèle linéaire autour de cette prédiction (c*f.* Ribeiro et *al.* [2016]). Son principe est de trouver le meilleur modèle linéaire qui traduit le comportement du modèle prédictif pour une prédiction donnée (c*f.* Lundberg et Lee [2017]) :

$$g(x') = \text{argmin}_{\varepsilon \in G_L} \quad L(f, \varepsilon, \pi_{x'}) \tag{4.4}$$

où f et g sont respectivement le modèle à expliquer et le modèle explicatif ; G_L est la classe des modèles linéaires ;

x' est le vecteur correspondant à l'absence/présence[11] des variables explicatives de l'instance d'intérêt x dans la prédiction obtenue ;

Et les paramètres $\pi_{x'}$ (mesure de proximité[12]) et L (fonction de perte) sont donnés ci-dessous.

$$\begin{cases} L(f, \varepsilon, \pi_{x'}) = \sum_{z' \in Z} \pi_{x'}(z') \left(f\left(h_x(z')\right) - \varepsilon(z') \right)^2, \\ \pi_{x'}(z') = \dfrac{m-1}{|z'|(m-|z'|)} \left(\dfrac{m}{|z'|} \right)^{-1}. \end{cases}$$

z' est le vecteur correspondant à l'absence/présence des m variables explicatives de l'instance voisine créée z, dans la prédiction obtenue. z' est à valeurs dans $\{0,1\}^m$;

Z est l'ensemble des vecteurs lignes de la matrice binaire représentant toutes les combinaisons possibles[13] de z' ;

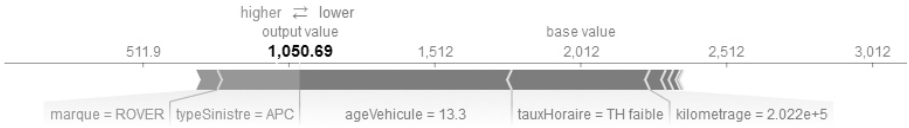
h_x est une fonction telle que $z = h_x(z')$ et $|z'|$ correspond au nombre d'éléments non nuls de z'.

Exemple d'application

Le *Kernel* SHAP consiste à expliquer l'écart existant entre une valeur prédite et la valeur moyenne prédite. Plus précisément, il affecte à chaque variable explicative une valeur SHAP, valeur correspondant à sa contribution dans l'écart obtenu entre la valeur prédite et la valeur moyenne prédite.

Le schéma ci-dessous l'illustre bien, il s'agit d'une prédiction effectuée sur un sinistre dont le coût observé vaut 1 057,16 €.

FIGURE 8 *Décomposition pour une observation*



Les contributions (valeurs SHAP) positives sont représentées en rouge, et celles négatives le sont en bleu. De plus, on observe que la valeur prédite (notée *output value* sur le schéma) est de 1 050,69 € et la valeur moyenne prédite (notée *base value* sur le schéma) vaut 2 012 €. Toutes les contributions ont bien été représentées sur le schéma ci-dessus mais elles ne sont pas toutes écrites à cause d'un manque d'espace.

Ces valeurs SHAP ont été calculées à l'aide de la fonction Python du module H2O *predict_contributions*. Cette fonction est uniquement utilisable pour les modèles prédictifs type GBM, XGBoost obtenus à l'aide du package H2O. Sa syntaxe est la suivante :

$$\phi = model.predict\_contributions(data)$$

où *model* et *data* correspondent respectivement au modèle prédictif construit[14] sous H2O et à l'échantillon test et $\phi = (\phi_0, \ldots, \phi_m)$ (avec ϕ_j vecteur colonne à n_{test} valeurs) correspond à une matrice de taille n_{test} lignes et m colonnes (m étant le nombre de variables explicatives), dont les $m - 1$ dernières colonnes correspondent aux valeurs SHAP des valeurs prédites. ϕ_0 correspond à la valeur prédite moyenne, il s'agit de la même valeur dans toute la colonne, nous pouvons l'assimiler à l'*intercept* dans un modèle linéaire.

Modélisation

L'intérêt principal de cette approche est la contribution de la variable « cible » (variable représentant le réseau d'experts intervenu). Plus précisément, on cherche à connaître quel réseau a la contribution moyenne la plus faible. Le réseau avec la contribution moyenne la plus faible sera le réseau le plus performant.

Toutefois, pour rester en adéquation avec les écarts de performance calculés jusqu'ici (en pourcentage), il convient de diviser ces contributions moyennes par le coût moyen prédit :

$$sH_A = \frac{1/n_{test}^A \sum_{i=1}^{n_{test}^A} \hat{\phi}_{ip}^A}{1/n_{test} \sum_{i=1}^{n_{test}} \sum_{j=0}^{m} \hat{\phi}_{ij}} = -0{,}79\,\% \qquad sH_B = \frac{1/n_{test}^B \sum_{i=1}^{n_{test}^B} \hat{\phi}_{ip}^B}{1/n_{test} \sum_{i=1}^{n_{test}} \sum_{j=0}^{m} \hat{\phi}_{ij}} = 0{,}53\,\%$$

avec $p \in \{1, \ldots, m\}$ indice colonne de la variable « cible ». Les $\hat{\phi}_{1p}^A, \ldots, \hat{\phi}_{n_{test}^A p}^A$ représentent précisément les valeurs approximées des contributions du réseau A et les $\hat{\phi}_{1p}^B, \ldots, \hat{\phi}_{n_{test}^B p}^B$ sont celles des contributions du réseau B.

$(\hat{\phi}_{ij})_{1 \le i \le n_{test};\ 1 \le j \le m}$ est la matrice des valeurs SHAP obtenues sur l'échantillon de test. Ces contributions sont des valeurs approximées, ceci pour éviter des temps de calcul trop longs.

Les résultats ci-dessus permettent de conclure que le réseau A est plus performant que le réseau B. De manière équivalente à l'interprétation effectuée dans les modèles linéaires, on peut dire que la présence du réseau A diminue en moyenne les coûts de sinistres de 1,32 % ou encore que l'écart de performance est de –1,32 %. Ce coefficient correspond à la différence entre sH_A et sH_B, il sera expliqué de manière détaillée le calcul de coefficient à la section 4.

Bien que les contributions des variables explicatives soient réparties de manière équitable, il n'est pas inutile de se demander si l'on obtient la même conclusion en s'intéressant uniquement aux meilleures prédictions. En effet, il se pourrait que les contributions de la variable « cible » issues des mauvaises prédictions soient plus ou moins importantes que prévu.

On calcule à nouveau ces scores, mais cette fois-ci en fonction de la qualité des prédictions. Pour ce faire, on construit des classes homogènes vis-à-vis de la mesure d_2 :

$$d_2(y_i, \hat{y}_i) = \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right|$$

où y_i et $\hat{y}_i$ correspondent respectivement aux coûts de sinistres observés et prédits.

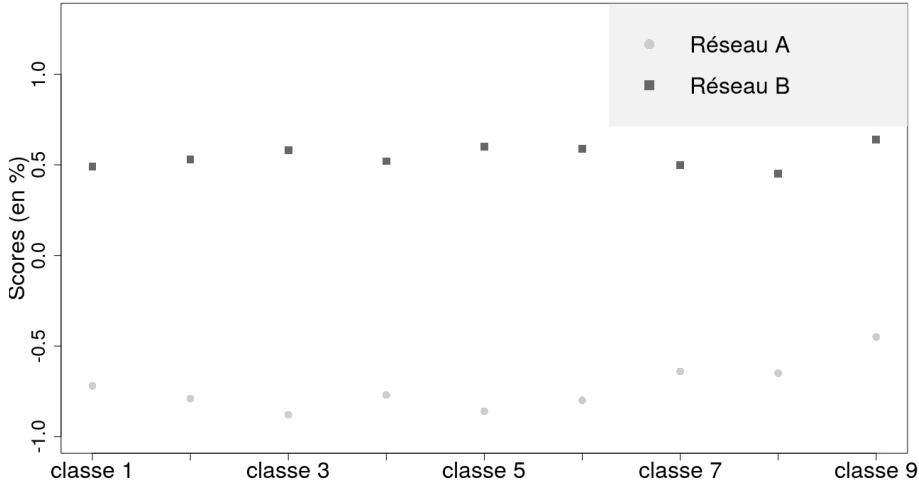
La classification s'est faite à l'aide de l'algorithme des *K-means*. La méthode utilisée pour trouver le nombre de classes est la même que celle utilisée dans la première approche : on prend un nombre de *clusters* tel que l'inertie expliquée soit supérieure à 95 %. L'algorithme des *K-means* est appliqué avec $K = 9$. Les neuf classes homogènes vis-à-vis de la mesure d_2 ont été numérotées de 1 à 9 par ordre croissant des valeurs des centres des classes (voir ci-dessous).

TABLEAU 4 *Valeurs des centres des classes construites*

	NUMÉROS DES CLASSES								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Valeurs des centres	0,10	0,31	0,53	0,75	1,33	2,19	3,54	5,89	10,95

On compare à nouveau les deux réseaux non pas sur la base de test entière, mais dans chaque groupe construit. Par conséquent, on obtient les résultats suivants :

FIGURE 9 *Représentation graphique des scores calculés par classe, pour chaque réseau*



En observant le graphe ci-dessus, on remarque que le réseau A a un meilleur score que le réseau B dans chaque groupe. De plus, on remarque une certaine fluctuation des scores probablement due aux erreurs de prédiction. Comme la répartition des contributions dépend des valeurs prédites, si ces dernières sont mauvaises probablement les

contributions de la variable « cible » le seront également. Pour cette raison, il a été jugé raisonnable de ne s'intéresser qu'aux résultats de la première classe, pour ainsi se rapprocher le plus possible d'une répartition fiable. Dans cette classe, on obtient :

$$sH_A = -0{,}72\% \quad sH_B = 0{,}49\%$$

Dans ce cas, l'écart de performance entre les deux réseaux est de –1,21 %, soit une augmentation de 0,11 point sur l'écart de performance calculé sur la base de test entière.

Toutefois, il n'est pas utile de se demander si en comparant de nouveau les deux réseaux sur un nombre plus conséquent de sinistres bien prédits, la même conclusion sera obtenue. C'est pour cette raison que nous avons appliqué le *Kernel* SHAP à la base de données entière, avec pour modèle prédictif le même que celui utilisé lors de l'application à la base test. Les contributions obtenues ont été classées en fonction de la qualité des prédictions.

La classe des meilleures prédictions comporte ici en plus du nombre de valeurs bien prédites précédemment, environ cent cinquante mille autres valeurs bien prédites. Une application numérique des contributions résultantes donne :

$$sH_A = -0{,}71\% \quad sH_B = 0{,}48\%$$

On obtient à une différence près les mêmes valeurs que celles précédemment obtenues. Malgré un nombre plus conséquent de valeurs bien prédites, la conclusion reste inchangée, la démarche présente donc une certaine robustesse[15].

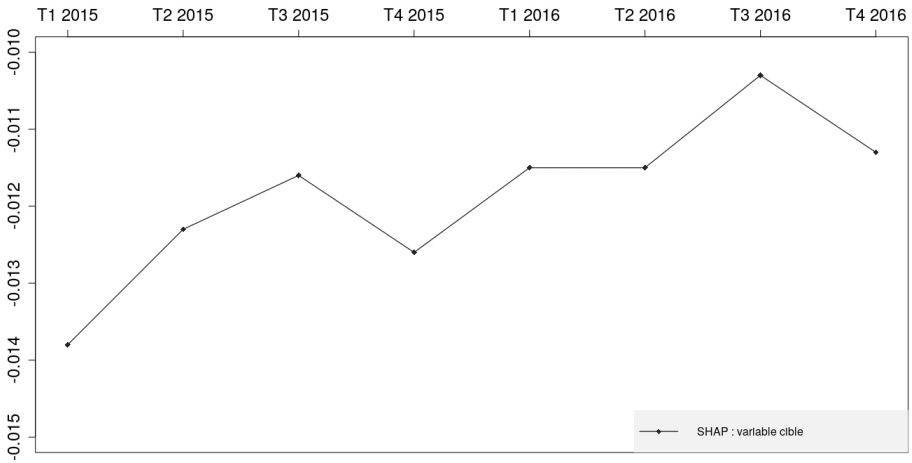
Comme dans l'approche précédente, il est présenté l'évolution trimestrielle de l'écart de performance. En d'autres termes, les scores précédemment obtenus ont été à nouveau calculés, mais cette fois-ci trimestre par trimestre. Par application numérique, on obtient :

TABLEAU 5 *Mesure de l'écart de performance par trimestre dans l'approche SHAP*

	2015				2016			
	T1	T2	T3	T4	T1	T2	T3	T4
Écart (en %)	–1,21	–0,79	–0,74	–1,07	–0,77	–0,63	–0,85	–0,99

La représentation graphique correspondante est la suivante :

FIGURE 10 *Évolution de l'écart de performance au fil des trimestres dans l'approche SHAP*



Ce graphe traduit la même tendance que chacun des graphes effectués dans les méthodes précédentes. De plus, on observe une certaine stabilité des écarts au fil du temps. Nous y reviendrons à la section 4.

On compare les résultats obtenus dans les approches proposées à ceux obtenus avec les modèles linéaires généralisés.

5. Synthèse des approches et rapprochement avec le GLM

Comme énoncé à la première partie de la section 2, la problématique de l'étude avait été traitée par De Lussac [2018] dans le cadre des prédictions issues des modèles linéaires généralisés.

Tout comme le GLM, les deux approches précédemment présentées nous ont permis de déterminer le réseau d'experts le plus performant. Avant de comparer les résultats des différentes approches, il est rappelé ci-après la modélisation effectuée dans le GLM.

Écart de performance à partir du GLM

Cette partie fournit le procédé de modélisation effectué et les résultats obtenus (*cf.* De Lussac [2018]).

Comme distribution, il a été choisi la loi gamma du fait de sa bonne adéquation aux distributions asymétriques et positives. La fonction logarithmique a été utilisée pour ainsi avoir des tarifs multiplicatifs. En effet :

$$\log\left(\mathbb{E}[Y|X]\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$$
$$\mathbb{E}[Y|X] = \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m)$$
$$\mathbb{E}[Y|X] = \exp(\beta_0) \times \left(\exp(\beta_1)\right)^{X_1} \times \ldots \times \left(\exp(\beta_m)\right)^{X_m}$$

La fonction de lien logarithmique permet en plus de raisonner en pourcentage.

La mise en place du GLM ne nécessite pas uniquement de définir la distribution et la fonction de lien. En effet, il est important de réduire la complexité des calculs. La base de données utilisée comprend des variables discrètes avec un nombre assez élevé de modalités. Étant donné que le GLM calcule un coefficient de régression par modalité pour chaque variable discrète, il a été jugé nécessaire de catégoriser ces variables.

On peut également noter le choix d'autres paramétrages effectués tels que le choix de la procédure de sélection de variables ou encore le critère de pénalisation de la log-vraisemblance du modèle. Pour plus d'informations, *cf.* De Lussac [2018].

La modélisation faite à partir du GLM permet de conclure que le réseau A diminue en moyenne les valeurs prédites d'environ 2,4% comparativement au réseau B. En d'autres termes, on a un écart de performance de −2,4% (valeur estimée du coefficient de régression). Cette mesure se situe dans un intervalle de confiance à 95% de plus ou moins 0,4%.

Écart de performance à partir du GBM

Dans la modélisation faite avec le GLM à fonction de lien logarithmique, l'écart de performance est obtenu par lecture du coefficient de régression de la variable correspondant au réseau d'experts intervenu.

Cependant, la modélisation faite avec le *Gradient Boosting Model* ne fournit pas de manière explicite cet écart de performance. Ce dernier a été obtenu comme suit :

- **Écart de performance dans la première approche** : ici, l'écart de performance entre les deux réseaux est obtenu à partir de la formule suivante :

$$écart_1 = \frac{\overline{y}_A - \overline{y}_B}{\overline{y}} \tag{4.5}$$

Où $\overline{y}_A$ et $\overline{y}_B$ correspondent respectivement à la valeur prédite moyenne du réseau A et celle du réseau B. Et $\overline{y} = 1/n_{test} \sum_{i=1}^{n_{test}} y_i$, les y_i sont les coûts observés.

Le but étant de mesurer l'écart de performance entre les deux réseaux, le coefficient de variation propre à chaque réseau est d'abord calculé, soit les coefficients k_A et k_B tels que $\overline{y}_A = \overline{y}(1 + k_A)$ et $\overline{y}_B = \overline{y}(1 + k_B)$. Ensuite, la différence des deux coefficients a été effectuée, d'où l'expression en (4.4).

Par application numérique de la formule (4.4) dans la première méthode de l'approche 1, on obtient :

$$écart_1 = \frac{2\,002{,}43 - 2\,016{,}95}{2\,006{,}93} = -0{,}72\,\%$$

Dans la seconde méthode de la même approche, on obtient :

$$écart_1 = \frac{1\,998{,}88 - 2\,016{,}98}{2\,006{,}93} = -0{,}90\,\%$$

- **Écart de performance dans la deuxième approche** : ici, la formule construite pour mesurer la performance entre les deux réseaux est :

$$écart_2 = sH_A - sH_B \tag{4.6}$$

où sH_A et sH_B correspondent aux scores des réseaux A et B, calculés à la section 3. Il s'agit de la variation entre la contribution moyenne d'un réseau et la contribution moyenne totale (ou encore valeur moyenne prédite). L'idée est la même que celle de l'approche précédente : on commence par mesurer en pourcentage (pour chaque réseau) la variation entre la contribution moyenne du réseau et la contribution moyenne totale. C'est-à-dire les coefficients k_A et k_B tels que $C_A = \overline{y}(1 + k_A)$ et $C_B = \overline{y}(1 + k_B)$ (avec C_A et C_B les

contributions moyennes des réseaux A et B). Ensuite, la différence entre ces deux coefficients est effectuée, d'où le résultat en (4.5). Par application numérique, on obtient :

$$écart_2 = -0{,}79\% - 0{,}53\% = -1{,}32\%$$

Récapitulatif des résultats

Le tableau ci-dessous résume les écarts de performance obtenus :

TABLEAU 6 *Tableau récapitulatif de l' écart de performance dans les quatre méthodes*

	ÉCART DE PERFORMANCE
GLM	–2,40 %
Sinistralité identique 1	–0,72 %
Sinistralité identique 2	–0,90 %
SHAP	–1,32 %

Les écarts renseignés dans le tableau ci-dessus s'interprètent comme étant la réduction (en pourcentage) sur le coût moyen qu'influe le réseau A comparativement au réseau B.

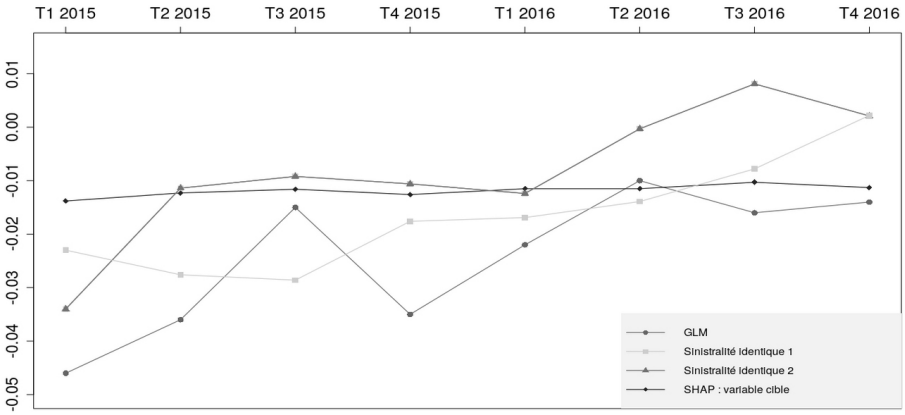
Il s'agit ci-dessus d'une comparaison statique. Il serait intéressant d'effectuer une comparaison dynamique, plus précisément une comparaison trimestre par trimestre. Avant de représenter sur un même graphique l'évolution temporelle des écarts des quatre méthodes, on rappelle les mesures d'écart obtenues à l'aide du GLM :

TABLEAU 7 *Mesure de l'écart de performance par trimestre dans un GLM*

	2015				2016			
	T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**
Écart (en %)	–4,6	–3,6	–1,5	–3,5	–2,2	–1,0	–1,6	–1,4
Écart-type (en %)	0,50	0,54	0,59	0,51	0,52	0,55	0,57	0,54

Le graphe suivant montre l'évolution trimestrielle des écarts de performance calculés dans les quatre méthodes :

FIGURE 11 *Évolution trimestrielle de l'écart de performance dans les quatre méthodes*



On observe une tendance similaire pour les écarts calculés *via* l'approche GLM et l'approche 1. De plus, on observe une certaine stabilité sur les écarts calculés *via* l'approche SHAP. Mais cette approche traduit également la même tendance que les trois autres courbes, comme l'illustre la figure 6.

La stabilité observée dans l'approche SHAP s'explique par le fait que cette dernière parvient à dissocier les effets de corrélation (minimes) entre les variables explicatives. Chose que ne peut faire le GLM car il suppose une indépendance totale entre les variables. Pourtant, dans certains cas pratiques on retrouve une faible corrélation entre les variables et le *Kernel* SHAP permet grâce à la valeur de Shapley de mieux isoler la contribution d'une variable. L'approche 1 également ne permet pas de dissocier l'effet de corrélation entre les variables.

Maintenant, en tenant compte de cet effet de corrélation dans l'approche 2, c'est-à-dire en appliquant le calcul de l'écart effectué dans l'approche 1 à l'approche 2, on obtient la formule suivante :

$$\frac{\bar{y}_A - \bar{y}_B}{\bar{y}} = \frac{\frac{1}{n^A_{test}} \sum_{i=1}^{n^A_{test}} \sum_{j=0}^{m} \hat{\phi}^A_{ij} - \frac{1}{n^B_{\text{test}}} \sum_{i=1}^{n^B_{test}} \sum_{j=0}^{m} \hat{\phi}^B_{ij}}{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \sum_{j=0}^{m} \hat{\phi}_{ij}} \qquad (4.6)$$

Dans l'approche 1, on évalue les réseaux sur la base d'une sinistralité identique, contrairement à l'approche 2 qui évalue les deux réseaux à travers leur base de données. Toutefois, le fait d'utiliser la même formule que celle de l'approche 1 n'est pas aberrant car on retrouve des similarités entre les données des deux réseaux pour chaque variable explicative.

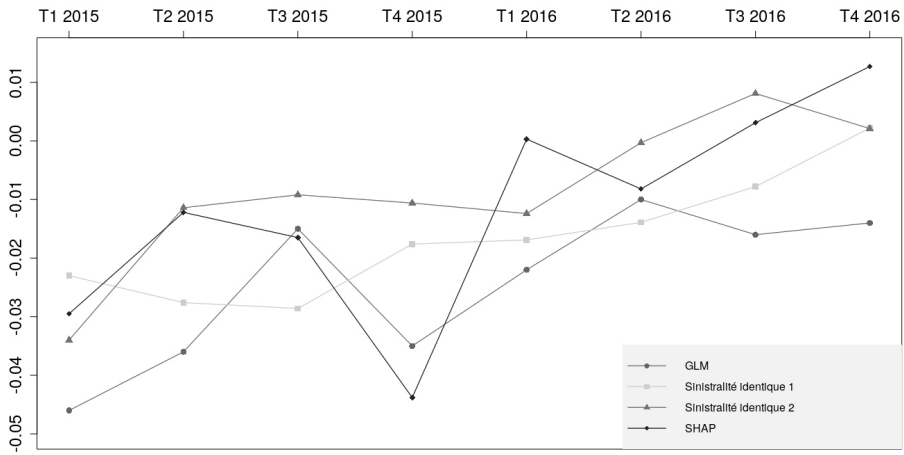
Par application numérique de la formule (4.6) aux valeurs SHAP, on retrouve un écart de performance de –1,30 % (en faveur du réseau A). En effectuant de nouveau ce calcul, mais cette fois-ci trimestre par trimestre, on obtient :

TABLEAU 8 *Mesure de l'écart de performance avec effet de corrélation, par trimestre dans l'approche SHAP*

	2015				2016			
	T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**
Écart (en %)	–2,95	–1,22	–1,65	–4,38	0,03	–0,82	0,31	1,27

En comparant les écarts obtenus à ceux des trois autres approches, on trouve :

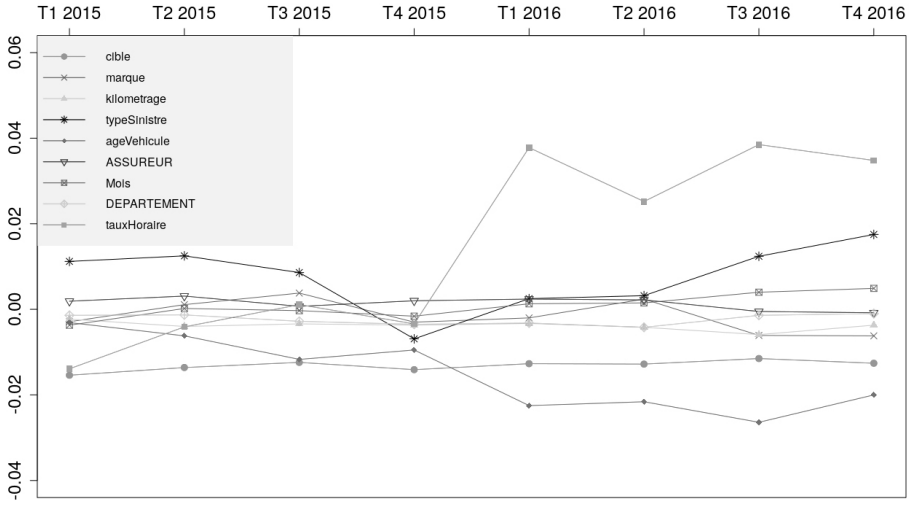
FIGURE 12 *Évolution trimestrielle de l'écart de performance dans les quatre approches, toutes prenant en compte la corrélation entre variables*



Les courbes des quatre méthodes décrivent toutes une dégradation de l'écart de performance, mais néanmoins en faveur du réseau A. On observe bien une cohérence et une certaine adéquation entre les résultats. Ici, la tendance est la même pour les quatre approches, ce qui n'était pas le cas à la figure 7. Dans cette figure, les corrélations entre variables (aussi faibles qu'elles soient) n'étaient pas prises en compte.

À l'aide du *Kernel* SHAP, il est possible d'isoler l'effet des autres variables sur l'écart de performance. Le graphique suivant renseigne l'effet de chaque variable dans l'évolution trimestrielle de l'écart de performance.

FIGURE 13 *Évolution trimestrielle des contributions des variables sur l'écart de performance avec corrélation entre variables dans l'approche SHAP*



La figure ci-dessus a été obtenue en calculant pour chaque trimestre l'écart de chaque variable à partir de l'écart global défini en (4.6). Plus précisément, on raisonne comme suit :

$$\frac{\bar{y}_A - \bar{y}_B}{\bar{y}} = \frac{1/n_{test}^A \sum_{i=1}^{n_{test}^A} \sum_{j=0}^{m} \hat{\phi}_{ij}^A - 1/n_{test}^B \sum_{i=1}^{n_{test}^B} \sum_{j=0}^{m} \hat{\phi}_{ij}^B}{1/n_{test} \sum_{i=1}^{n_{test}} \sum_{j=0}^{m} \hat{\phi}_{ij}}$$

$$= \frac{1/n_{test}^A \sum_{i=1}^{n_{test}^A} \hat{\phi}_{i1}^A - 1/n_{test}^B \sum_{i=1}^{n_{test}^B} \hat{\phi}_{i1}^B}{1/n_{test} \sum_{i=1}^{n_{test}} \sum_{j=0}^{m} \hat{\phi}_{ij}} + \ldots$$

$$+ \underbrace{\frac{1/n_{test}^A \sum_{i=1}^{n_{test}^A} \hat{\phi}_{ip}^A - 1/n_{test}^B \sum_{i=1}^{n_{test}^B} \hat{\phi}_{ip}^B}{1/n_{test} \sum_{i=1}^{n_{test}} \sum_{j=0}^{m} \hat{\phi}_{ij}}}_{\textbf{Effet de la variable « cible »}} + \ldots$$

$$+ \frac{1/n_{test}^A \sum_{i=1}^{n_{test}^A} \hat{\phi}_{im}^A - 1/n_{test}^B \sum_{i=1}^{n_{test}^B} \hat{\phi}_{im}^B}{1/n_{test} \sum_{i=1}^{n_{test}} \sum_{j=0}^{m} \hat{\phi}_{ij}}$$

En effet, c'est la somme des écarts de chaque variable qui explique l'écart de performance obtenu. Le *Kernel* SHAP parvient à dissocier les effets existants entre variables, pour ainsi attribuer à chaque variable sa contribution marginale.

Les approches développées dans cette étude offrent de meilleures prédictions que le GLM à fonction de lien logarithmique, soit une diminution moyenne d'environ 34 % de la MSE du GLM a été constatée sur la MSE du GBM utilisé dans ces approches. Ceci est dû à l'utilisation du GBM, modèle d'apprentissage automatique. Toutefois, relevons le fait que ces deux approches, contrairement aux GLM, ne renvoient pas des mesures d'erreur associées aux mesures d'écart.

Un point important à relever est la modélisation effectuée pour le calcul de l'évolution temporelle des écarts de performance : le GLM a pour avantage d'être applicable au fur et à mesure que les données sont à disposition sans toutefois avoir à relancer les précédents calculs. En d'autres termes, l'écart de performance calculé pour un trimestre n'influe pas sur les écarts calculés pour les trimestres précédents. Les deux autres approches, basées sur l'utilisation du GBM, nécessitent de relancer tous les calculs une fois la prise en compte de nouvelles observations. Ceci peut entraîner une légère variation des mesures d'écart précédemment calculées. Il est en effet peu prudent d'envisager une modélisation GBM par trimestre, du fait d'un trop faible nombre d'observations et donc de s'exposer à des mesures d'erreur plus grandes que celles obtenues avec la base globale.

Un point essentiel qui justifie l'intérêt des approches proposées est de pouvoir moduler la mesure d'impact sur des sous-segments lorsque l'hypothèse de proportionnalité n'est pas vérifiée.

6. Conclusion et discussion

Au terme de cette étude qui avait pour objectif de mesurer l'influence du réseau d'experts retenu sur les coûts de sinistres matériels automobile, deux approches ont été développées. La première consiste à comparer les deux réseaux sur une base de sinistres identiques et la seconde consiste à mesurer (à l'aide du *Kernel* SHAP) la contribution marginale d'un réseau dans les coûts prédits. Ces deux approches ont été mises en œuvre avec un modèle d'apprentissage automatique : le *Stochastic Gradient Boosting Model*. Les résultats sont confrontés à ceux obtenus par De Lussac [2018] avec un modèle GLM multiplicatif.

Toutes ces approches conduisent qualitativement au même résultat, tant en vision instantanée (surperformance de A par rapport à B) que dynamique (dégradation de la performance au fil des trimestres de l'historique de deux ans considéré).

Le GLM, malgré ses hypothèses limitées ou encore sa qualité de prédiction moindre comparée à celle d'un modèle d'apprentissage automatique, demeure un modèle suffisamment pertinent pour se faire une idée sur la comparaison entre deux réseaux. D'autant plus que sa modélisation est relativement simple et son interprétation compréhensible et facile.

En termes quantitatifs, on obtient des niveaux de performance différents et il est important de quantifier avec le plus de précision possible la mesure d'écart. Dans ce contexte, si nous devions choisir une mesure parmi les quatre méthodes proposées, celle obtenue avec les valeurs SHAP serait privilégiée. Elle offre en effet de meilleures prédictions comparativement aux GLM, avec notamment une diminution de la MSE d'environ 34 % entre le GLM et le GBM. On peut également relever le fait qu'elle repose sur un cadre théorique de la valeur de Shapley. Toutefois, cette dernière ne renvoie pas de mesures d'erreur associées comme le fait la modélisation *via* le GLM.

Ayant choisi l'approche SHAP comme celle la plus représentative de la réalité, il s'ensuit la problématique du choix de la mesure d'écart de performance. Doit-on privilégier la mesure d'écart qui isole la

contribution d'un réseau, donc qui parvient à dissocier la liaison entre les variables? Ou celle qui prend en compte en plus de la contribution d'un réseau, la corrélation (aussi faible qu'elle soit) entre les variables?

C'est une question assez difficile. Toutefois, nous préconisons de choisir la mesure d'écart qui isole complètement la contribution d'un réseau. Car, il semble de base que c'est l'objectif recherché par le GLM. Sauf que, dû au fait que ce dernier néglige les corrélations minimes existant entre variables (en supposant une indépendance totale entre celles-ci), il renvoie au final une mesure d'écart qui renseigne la contribution d'un réseau qui prend en compte la liaison existante entre variables. Chose que permet d'éviter l'approche SHAP.

Pour finir, notons que toutes les méthodes proposées peuvent s'appliquer dans un grand nombre de situations, dès qu'il s'agit de mesurer l'impact d'une variable binaire sur une variable réponse quantitative.

7. Références

[1] Bergstra J., Bengio Y. [2012] Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305.

[2] Besson J.L., Partrat C. [2004] Assurance non-vie – Modélisation, simulation, Collection : Assurance – Audit – Actuariat, Paris : Economica.

[3] De Lussac M. [2018] Comparaison de modèles prédictifs pour l'évaluation des coûts matériels automobiles, Dauphine, Mémoire d'actuaire.

[4] Friedman J.H. [2002] Stochastic gradient boosting. *Computational Statistics & Data Analysis*, Volume 38, Issue 4, 28, Pages 367-378.

[5] Khougea D. [2019] Tarification IARD avec des modèles de régression avancés, Université de Strasbourg, Mémoire d'actuaire.

[6] Lundberg S.M., Erion G.G., Lee S.U [2019] Consistent Individualized Feature Attribution for Tree Ensembles, University of Washington, Working Paper.

[7] Lundberg S.M., Lee S.I. [2017] A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems*.

[8] Ribeiro m. T., Singh S., Guestrin 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, Pages 1135-1144

[9] Ridgeway G. [1999] The State of Boosting, *Computing Science and Statistics*, 31, 172–181.

[10] Ridgeway G. [2007] Generalized boosted models : A guide to the gbm package. University of Pennsylvania, Working Paper.

[11] Shapley L.S. [1953] A value for n-person games. *Contributions to the Theory of Games.* 2.28 : 307-317.

[12] Wabo A. [2019] Mesure de l'écart de performance entre deux réseaux d'experts en assurance automobile, Dauphine, Mémoire d'actuaire.

NOTES

1. Auriol Wabo est consultant chez PRIM'ACT. Contact : auriol.wabo@gmail.com.

2. Frédéric Planchet est Professeur à l'ISFA et actuaire associé chez PRIM'ACT. Contact : frederic@planchet.net.

3. Univ Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances (ISFA), Laboratoire SAF EA2429, F-69366, LYON, France.

4. La justification du lien entre une notion de performance et des coûts moins élevés est hors du périmètre de la présente étude. On peut juste indiquer ici que les experts respectent un cahier des charges et des normes professionnelles devant assurer un service rendu à l'assuré identique, que ce soit au travers du réseau A ou du réseau B.

5. Une présentation détaillée des données est fournie dans Wabo [2019].

6. L'intervalle de confiance a été obtenu à l'aide d'une loi de *Student* à n-1 degrés de liberté.

7. Le *boosting* est une méthode d'apprentissage automatique qui consiste à appliquer itérativement un algorithme d'apprentissage faible afin d'optimiser ses performances.

8. Les calculs précédents et suivants ont été effectués à l'aide d'un serveur de calcul optimisé sous Linux, mis à disposition par le cabinet Prim'Act. Ce serveur est muni de 64 Go de mémoire RAM, et d'un processeur à 32 cœurs qui dispose d'une fréquence de fonctionnement de 2,6 GHz et d'une mémoire cache de 40 Mo. Les questions de performance ne sont pas centrales ici, lorsqu'elles le deviennent, l'approche peut être adaptée en considérant l'approche de Lundberg et al. [2019].

9. Pour rappel, il s'agit du nom de la variable correspondant au réseau d'experts intervenu.

10. L'inertie expliquée est égale à la variance interclasses divisée par la variance totale.

11. Si x est de taille m, alors x' est à valeurs dans $\{0,1\}^m$.

12. Elle permet de créer des instances voisines à l'instance d'intérêt x.

13. Par exemple, si $m = 2$, alors $Z = \{(0,0) ; (0,1) ; (1,0) ; (1,1)\}$.

14. La construction du modèle a été effectuée à l'aide des échantillons d'apprentissage et de validation.

15. Le *Kernel* SHAP a été appliqué à la base de test et à la base entière, et dans chaque cas les contributions obtenues sont des valeurs approximées. Le temps de calcul des valeurs SHAP de la base de test (environ cent soixante mille observations) a été d'environ 7 minutes et 20 secondes, tandis que celui correspondant à la base entière (comprenant environ huit cent cinquante mille observations) a été d'environ 8 minutes et 30 secondes.

e. Algorithme GBM

Algorithme 1 : Stochastic Gradient Boosting pour la régression

Entrée(s) x: observation à prévoir; $(y_i, x_i)_{1\leq i\leq n}$: échantillon de données; M: nombre d'itérations maximal; p: proportion d'observations sélectionnées aléatoirement; K: profondeur maximale des arbres de régression; λ: taux d'apprentissage ou *shrinkage*.

Initialisation: $\hat{f}(x) = \frac{1}{n}\sum_{j=1}^{n} y_j$

Calcul: $\tilde{n} = p \times n$

Pour m allant de 1 **à** M **faire**

Tirage aléatoire et sans remise de $\tilde{n}$ indices dans {1, …, n}. Les indices tirés sont stockés dans $\mathbb{O}$.

Calcul de $z_i = -\left.\frac{\partial L(y_i, \rho)}{\partial \rho}\right|_{\rho=\hat{f}(x_i)}$ pour $i \in \mathbb{O}$

Ajustement d'un arbre de régression δ, de profondeur maximale K, aux couples $(x_i, z_i)_{i\in\mathbb{O}}$

Calcul de $\gamma = \text{argmin}_w \sum_{i\in\mathbb{O}} L\left(y_i, f(x_i) + w\delta(x_i)\right)$

Mise à jour: $\hat{f}(x) = \hat{f}(x) + \lambda\gamma\delta(x)$

Fin Pour

Sortie(s)

f. Paramètres du modèle de prédiction

Les paramètres concernés et leurs bornes de recherche associées sont les suivants:

- **Les nombres de subdivision minimal** (*nbis*) **et maximal** (*nbis_top_level*) **pour les variables continues**: lors de la construction de l'arbre de régression, chaque variable quantitative est transformée en variable qualitative dont le nombre de facteurs de subdivision est un paramètre du modèle. Plus le nombre de facteurs est grand, plus il y aura de possibilités à étudier, donc une complexité accrue, et plus la discrétisation sera précise. Par défaut, *nbis* est fixé à 20 et

nbis_top_level à 1 024. L'étude réalisée par De Lussac [2018] montre une faible influence de ces paramètres sur la MSE (sur l'échantillon de validation) malgré une forte variation. C'est pour cette raison que la valeur de *nbis* est recherchée dans l'intervalle [500 ; 1 500] et celle de *nbis_top_level* est fixée à 3 000.

- **Le nombre de variables à sélectionner aléatoirement** *col_sample_rate* : ce paramètre permet de sélectionner de manière aléatoire et sans remise les variables dans chaque division. Par défaut, il est égal à 1. Il a été montré à l'aide de nombreux tests sur plusieurs valeurs que la qualité de prévision du modèle dépend également de ce paramètre. Ainsi, l'intervalle de recherche définie est de [0,4 ; 1] par pas de 0,1. Cet intervalle est motivé par les résultats des erreurs de prédiction (sur l'échantillon de validation) qui étaient intéressantes à partir de 0,4.

- **Le minimum d'observations par feuille** *min_rows* : il est par défaut égal à 10. Après plusieurs tests, l'ensemble des valeurs possibles défini est {1, 2, 5, 10, 20}.

- **La profondeur maximale des arbres** K : généralement égale à 5 , il a été décidé de rechercher sa valeur optimale entre 4 et 8.

- **La proportion d'observations sélectionnées aléatoirement** p : tout comme le paramètre *ncol_sample_rate*, le tirage aléatoire s'effectue sans remise. Après de nombreux tests effectués, p est recherchée dans l'intervalle [0,5 ; 1] par pas de 0,1.

- **Le taux d'apprentissage** λ : il permet de pénaliser l'ajout d'un nouveau modèle dans l'agrégation et de ralentir la convergence. Sa valeur optimale est recherchée dans l'intervalle [0,05 ; 0,1] par pas de 0,001.

- **Le facteur dégressif** λ_m : il permet de mieux contrôler la convergence en la rendant suffisamment lente. Il intervient en remplaçant le taux d'apprentissage λ par $\lambda \times \lambda_m$ dans l'algorithme du GBM (en annexe), avec $\lambda_m = 0{,}999^m$ par exemple, m étant le compteur de la boucle. L'idée de cet ajout a été motivée par la relation entre le taux d'apprentissage et le nombre d'itérations nécessaire pour obtenir la solution optimale. En effet, un faible taux conduit à accroître le nombre d'arbres mais entraîne généralement une amélioration de la qualité de prédiction. Ainsi, le facteur dégressif λ_m permet au modèle d'avoir un coefficient d'apprentissage élevé pour les premières itérations pour ensuite lui permettre un apprentissage lent au fur et à mesure des itérations. Pour cette raison, λ_m est recherchée dans l'intervalle [0,99 ; 1] par pas de 0,001.

Les paramètres à optimiser et l'ensemble des valeurs où leur valeur optimale a été recherchée, sont résumés.

TABLEAU 9 *Paramètres à optimiser et leurs domaines d'existence*

Nombre de subdivisions pour les variables continues	*nbins*	[500 ; 1 500]
Sélection aléatoire des variables	*col_sample_rate*	[40 % ; 100 %]
Minimum d'observation par feuille	*min_rows*	{1, 2, 5, 10, 20}
Profondeur maximale des arbres	*K*	[4 ; 8]
Observations sélectionnées aléatoirement	*p*	[50 % ; 100 %]
Coefficient d'apprentissage	λ	[0,05 ; 0,1]
Facteur dégressif	λ_m	[0,99 ; 1]

À l'aide d'un calcul simple et du nombre de valeurs possibles pour chaque paramètre ci-dessus, il existe $11 \times 7 \times 5 \times 5 \times 6 \times 51 \times 11 = 6\,479\,550$ configurations possibles. Parcourir toutes ces combinaisons n'est pas envisageable. La recherche par grille sera alors délaissée au profit d'une autre méthode dite « recherche aléatoire par grille ». Ce procédé a été théorisé par Bergstra et Bengio [2012]. L'argument de base est qu'à temps de calcul équivalent, la recherche aléatoire par grille permet de trouver des modèles performants sur de plus grands espaces que ne le fait la recherche par grille. De plus, Bergstra et Bengio [2012] ont constaté que seuls certains paramètres ont un réel impact sur les performances de l'algorithme. Ces observations justifient la non-nécessité de tester toutes les valeurs des paramètres pas assez influents, ce qui permet un énorme gain de temps de calcul.

Il est indispensable de définir un critère d'arrêt pour la méthode de la recherche aléatoire. Celui-ci est soit fonction du temps de calcul, soit fonction du nombre de modèles calculés. Dans le cadre de cette étude, il a été choisi de fixer un temps de calcul maximal. En effet, plus le temps est important, plus il y aura de modèles calculés et plus on se rapproche de la solution optimale.

Néanmoins, une recherche aléatoire d'une heure peut obtenir une solution optimale meilleure (au sens de la MSE sur l'échantillon de test) que celle d'une recherche aléatoire de 4 heures. À titre d'illustration, voir le tableau ci-dessous.

TABLEAU 10 *Résultats de la recherche aléatoire par grille en fonction du temps de calcul*

	PARAMÈTRES								
TEMPS DE CALCUL	*col_sample_rate*	λ	λ_m	*K*	*nbins*	*min_rows*	*p*	Nb arbres	MSE TEST
1 heure	0,4	0,062	0,992	8	1 400	20	0,8	1 375	2 422 460
4 heures	0,4	0,072	0,996	8	1 500	1	0,8	2 792	2 428 957
12 heures	0,5	0,069	0,994	8	1 500	1	1,0	1 853	2 421 259

Le tableau ci-dessus renseigne les paramètres des meilleurs modèles obtenus en fonction du nombre d'heures effectuées par recherche aléatoire. Le modèle retenu est celui obtenu après 12 heures de recherche aléatoire.

Afin de parvenir à de meilleurs résultats, le modèle obtenu a subi une modification. Cette dernière est motivée par Ridegway [2007], qui annonce un résultat significativement meilleur lorsqu'une validation croisée à 5 échantillons est effectuée. Le principe est le suivant : tout d'abord agréger les échantillons d'apprentissage et de validation, ensuite diviser l'échantillon obtenu en 5 échantillons de même taille, enfin chacun des 5 échantillons servira d'échantillon de validation tandis que les 4 autres agrégés serviront d'échantillon d'apprentissage. Le critère d'arrêt utilisé est le même que précédemment.

La validation croisée se révèle plus efficace en abaissant* la MSE du boosting sur l'échantillon de test à 2 411 535.

En somme, comparativement au GBM simple qui dispose d'une MSE sur l'échantillon de test de 2 477 139, le GBM optimisé dispose désormais d'une MSE de 2 411 535. Soit une diminution de 2,7% qui est un gain non négligeable. De plus, une comparaison a également été faite par rapport à la norme L1 et il s'est avéré que les prédictions issues du GBM optimisé sont d'environ 1,51% plus précises que celles du GBM simple. Ceci atteste bien qu'un modèle de meilleure qualité a été obtenu et donc de l'utilité de tous les procédés appliqués pour l'obtenir.

* Toutefois, cette méthode présente un désavantage majeur qui est son temps de calcul. À titre d'exemple, elle a mis 36 minutes et 7 secondes de plus que le GBM simple (qui a pris 1 minute et 31 secondes d'exécution).