Universite Claude Bernard Lyon 1 Institut de Science Financière et d'Assurances

DATA SCIENCE

✧✧✧

P R O J E T :

Tarification des assurances automobiles par l'utilisation des techniques d'apprentissage statistique

✧✧✧

NGUYEN Tuan Anh

Lyon, Janvier 2023

Contents

Description des données

Nos données concluent les caractéristiques principales des assurés dans le portefeuille, ainsi la historique des assurés observées pendant au moins 3 mois lors de l'année précédente. On va commencer "Gestion des données" par l'étape de lecture et de compréhension des données, et puis d'analyse des coefficients de corrélation entre les variables et enfin de traitement des données. Cette étape consiste d'une part aux traitements des données aberantes, des valeurs manquantes, des valeurs anormales,... ensuite d'autre part à des analyses univariés et multivariées des variables.

1.1 Préparation des données

Nos données consistent 60.000 observations sous 16 attributs,qui consistent en 6 variables qualitatives et 10 variables quantitatives. Il n'y aucune manquantes dans nos données.

À propos des coefficients de corrélation entre des variables quantitatives, on les représente sous la forme d'une matrice de corrélation.

On peut voir que les variables "claimValue" et "carValue" sont indépendantes - ni une relation

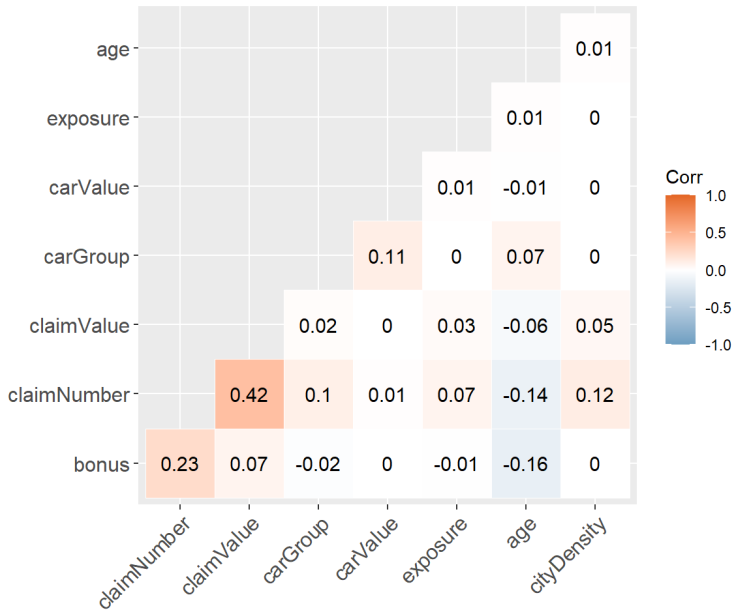

Figure 1: Carte thermique de la matrice de corrélation

linéaire ni un autre type de relation. Mais si on abandonne le prédicteur "carValue" simplement parce qu'il n'a pas de relation linéaire avec la réponse "claimValue", on peut obtenir un ajustement bien pire ou on manquera l'information quand on appliquera d'autres modèles. Donc, la matrice des coefficients de corrélation nous donne juste un aperçu de la référence sur la corrélation linéaire entre une paire de variables.

Représenter graphiquement le nombre des sinistres en fonction de différents indicateurs dans les données `train`. Un histogramme est une description graphique de points de données or-

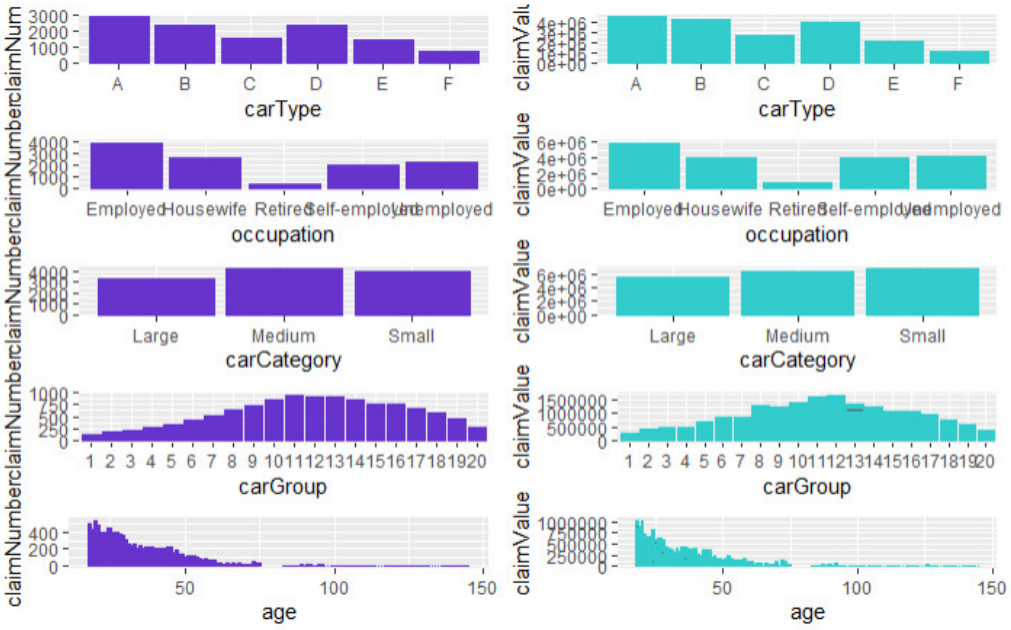

Figure 2: Écrémage de données

ganisés en plages spécifiques à l'utilisateur. On employe ce diagramme pour envisager de regrouper les valeurs de chaque attribut. Par exemple, concernant la variable "carType", la forme de l'occurrence entre ce type de données et le nombre de sinistres ou le montant total des sinistres a la même distribution. sinon, c'est l'un des critères à prendre en compte pour regrouper les valeurs de chaque attribut. (A, B, D comme un groupe, C et E comme un groupe et F).

1.2 Traitement des données

1.2.1 "Âge"

D'un coup d'œil sur les données, le problème de l'âge maximum de l'assuré est de 145 ans. On utilise des descriptions graphiques pour simuler la dispersion des observations (plot(datraw$age)), en combinant avec un graphique en boîte à moustaches pour identifier les valeurs aberrantes.

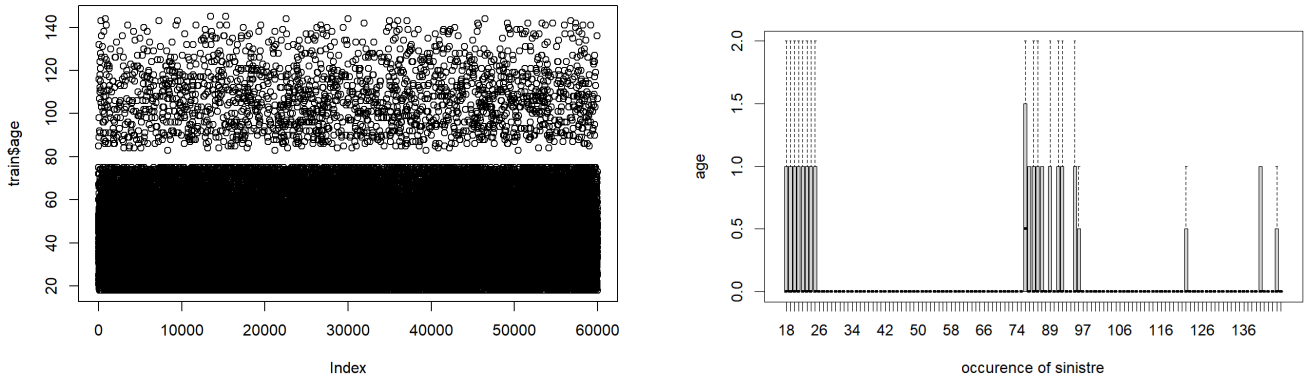

Table 1: Dispersion de l'âge des observation

Table 1 nous montre la dispersion de chaque âge et aussi la limite d'âge de 80 ans ou plus, ce qui nous donne une idée de la suppression de l'observation de plus de 80 ans.

1.2.2 "Gender"

En terme de "Gender", dans la diagramme première en barres, les hauteurs des barres proportionnelles au nombre de chaque valeur de sinistre, classé par sexe type. Généralement, la proportion de femmes (83,35%) qui ont eu le montant de sinistre 0 est plus élevée que celle des hommes (86,24%). Donc, lorsqu'on agrège avec le montant de sinistre et montre sur la diagramme circulaire, la valeur totale des sinistres du groupe par femmes est le tiers de celle des hommes. Mais cela peut tout à fait s'expliquer par le nombre d'observation qu'est la femme est un tiers de celui de l'homme. Et on va considérer l'existence de cette variable comme une variable explicative.

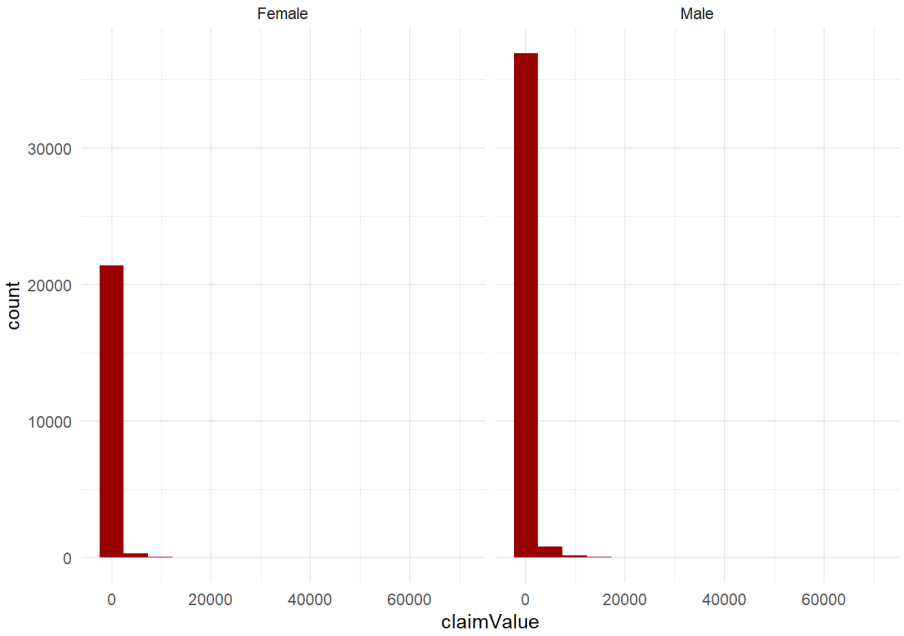

Figure 3: Diagramme en barres "Gender"

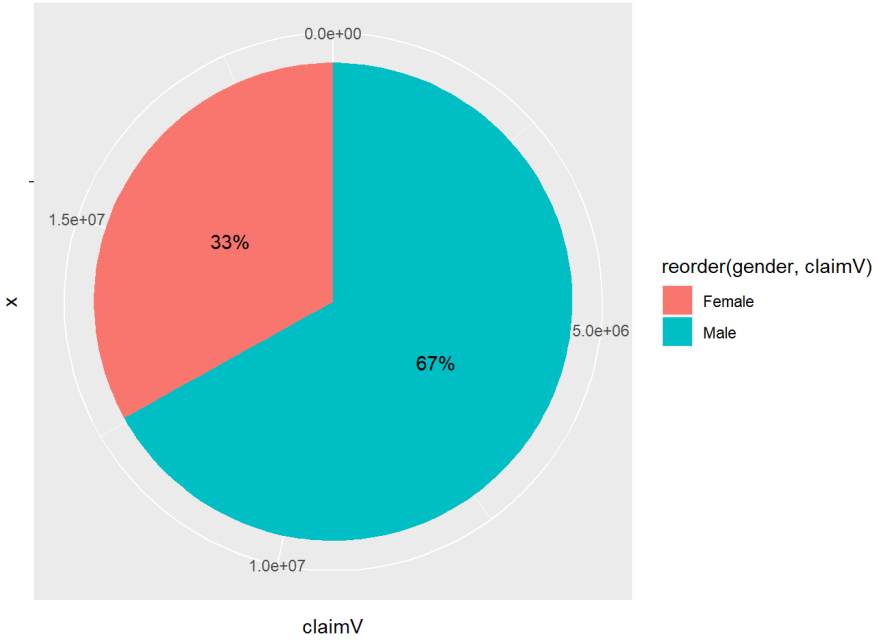

Figure 4: Diagramme circulaire de la varible "Gender" combiné avec le montant de sinistre

1.2.3 "Occupation"

Dans ce partie, on réfléchit l'impact de variable explicative "Occupation" sur la variable "ClaimValue". Basée sur nos données, le pourcentage de *employed* indépendants, de *self-employed* et de *housewife* représente la majeure partie du nombre d'observations, respectivement 31,3%, 20,3% et 20%. Dans le même temps, la proportion d'observateurs retraités et de *unemployed* est de 28,1%. À travers le diagramme, on peut réaliser que *retired* a le plus petit montant total de demande, soit 4%. En revanche, *unemployed* a environ 15%, mais le montant total des demandes atteint 22%.

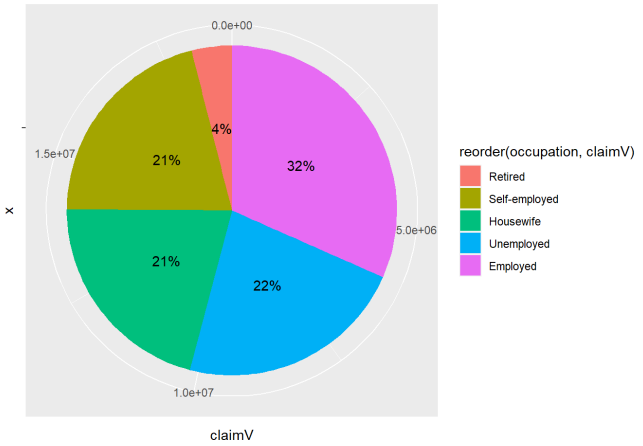

Figure 5: Diagramme circulaire de la varible "Occupation" combiné avec le montant de sinistre

1.2.4 "ClaimNumber" et "ClaimValue"

On a un grand nombre de montant de sinistre nul et quelques valeurs extremes. Le montant moyen de sinistres sur l'ensembles du portefeuille est de 312.61. Et pour les personnes sinistres le montant moyen est 2002 [summary(datraw[datraw$claimNumber!=0,]$claimValue]. On remarque donc que le grand nombre de personnes non sinistre permet de lisser la charge moyenne de sinistre. Donc, un modele de regression sur l'ensembles du portefeuilles causera un biais de sous estimation des montant de sinistres, donc de la prime pure.

Les sinistres nuls sont du en majorite aux personse non sinistres. On a 11 individus qui ont des sinistres declares mais pour qui on a payes aucunes indemnisations. C'est problablement des sinistres non responsables, ou des fausses declarations. Soit on les garde et considère que ce sont des sinistres qui ont été rembourses car les assures n'etaient pas responsables. Comme ca on ne perd pas l'information du nombre de sinistres.

Modelling

2.1 Théorie

On opte pour une stratégie dite de fréquence-sévérité des sinistres pour tarifer une observation (Frees, Derrig et Meyers, 2014 ; Parodi 2014 ; Denuit et al.,2007).
Notation:

- `La fréquence des sinistres F` est le nombre de sinistres N déposés par unité d'exposition au risque e.
- `La severité des sinistres S` fait référence au coût par sinistre et est définie par le montant moyen par sinistre déposé, c'est-à-dire le montant total des pertes L divisé par le nombre de sinistres N.
- `Le prix technique P` (ou prime pure) est alors le suivant:

$$\pi = E[\frac{L}{e}] = E[\frac{N}{e}] \times E[\frac{L}{N}|L > 0] = E[F] \times E[S] \tag{1}$$

On suppose l'existence d'indépendance entre la composante fréquence et la composante sévérité de la prime (*Klugman, Panjer et Willmot 2012*).

À prospos du modèle de Fréquence, toutes les données sont examinées avec *un poids d'exposition* au modèle comme les données de comptage du nombre de sinistres. Noté: Seuls les assurés ayant déposé au moins un sinistre (c'est-à-dire, L>0) contribuent à la calibration du modèle de Severité, et le nombre de sinistres N est utilisé comme un poids de cas dans la régression (Denuit et Lang 2004).

Dans notre projet, on va examiner les différentes approches telles que : CART, GLM, BOOSTING afin de choisir la meilleure technique pour tarifier la prime d'assurance automobile. Les étapes sont réprésentées comme suit:

Tout d'abord, on installe les bibliothèques nécessaires (`rpart` *Regression Tree*, `caret` *Classification and Regression Training*, `xgboost` comme *Extreme Gradient Boosting*, `gbm` *Generalized Boosted Regression Models*, `randomForest` *Breiman and Cutler's Random Forests for Classification and Regression* et `gridExtra` comme *Miscellaneous Functions for "Grid" Graphics*).
Ensuite, lance les données *datatrain*, *testraw* est la donnée qu'on va utiliser pour prédire la prime d'assurance, *datraw* est les données d'entraînement. On crée également une fonction qui peut transformer les données pour une meilleure utilisation de `xgboosting`.

2.2 Modélisation de la fréquence des sinistres

Comme la technique du boosting va être utilisée, l'optimisation des hyperparamètres est nécessaire.
Pour avoir une vue de référence, basée sur l'écart quadratique moyen (Root-mean-square deviation), on va utiliser le diagramme plot pour optimiser le nombre de cycles avec une valeur fixe du taux d'apprentissage = 0,1. On observe que le nombre optimal de répétitions est d'environ 65. `La cross validation` étant pour estimer la compétence des modèles d'apprentissage automatique. Ici, on choisit le nombre de folds est 5.

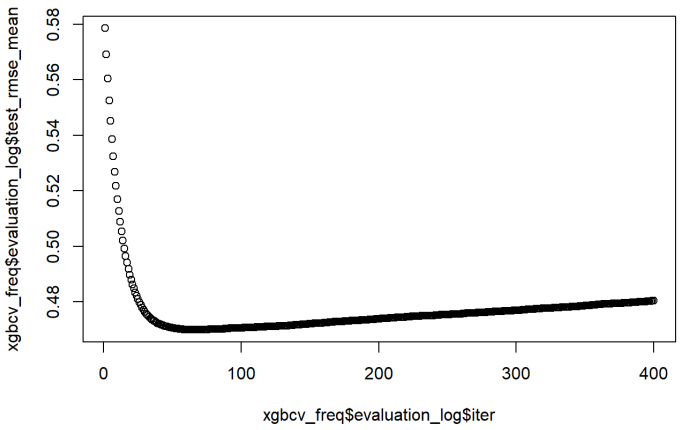

Figure 6: Relation entre le taux d'apprentissage et le nombre de cycles

Faire un zoom sur l'instant [50:200].

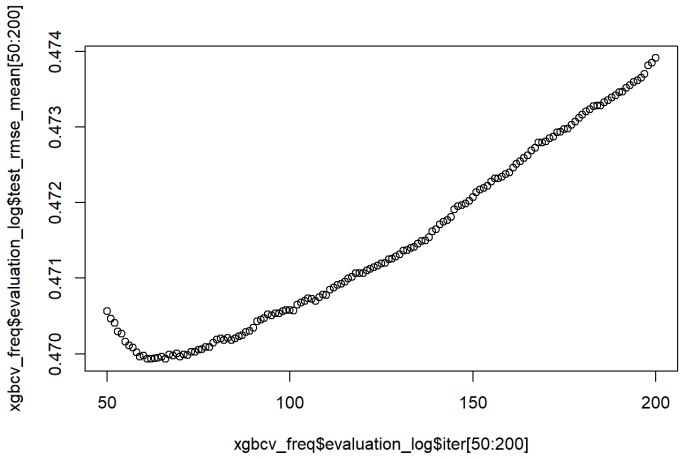

Figure 7: Zoom sur l'instant [50:200]

iters indique le nombre de fois que les paramètres de l'algorithme sont mis à jour. Encore une fois, si on applique un taux d'apprentissage de 0.1, il faut utiliser un nombre de cycles autour de 65.

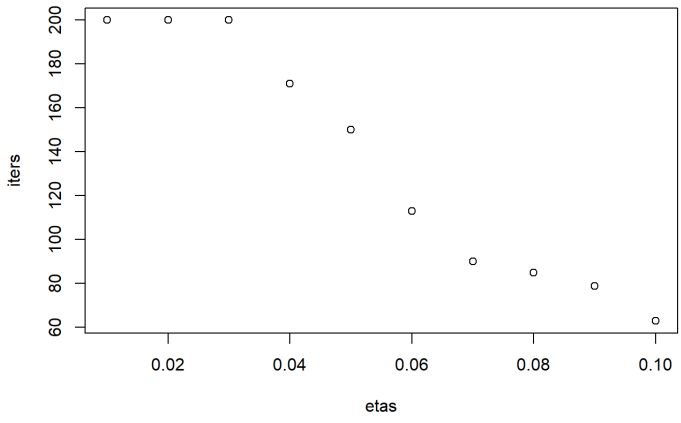

Figure 8: Relation entre taux d'apprentissage et le nombre de cycles

Tuning pour **max_depth** paramètre, on peut réaliser que `"Plus max_depth est élevé, plus le risque d'overfitting est important"`. Pour le modèle de Fréquence, on choisit la valeur

max_depth de 5 qui est expliqué dans l'image suivante.

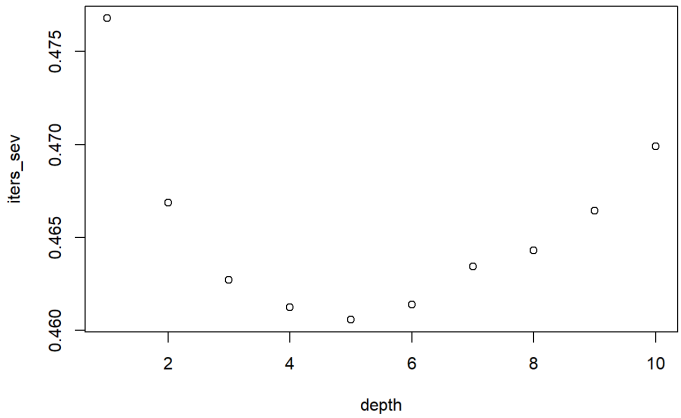

Figure 9: Max_depth

Ensuite, on va comparer les différentes approches pour modéliser la fréquence des sinistres, notamment : GLM, CART, BOOSTING. Le "tuning" des hyperparamètres du modèle CART consiste simplement à changer l'hyperparamètre cp - l'amélioration minimale qu'une division doit avoir. Ci-dessous, l'image symbolique utilisant un modèle de boîte à moustaches décrit la dispersion de l'écart quadratique moyen.

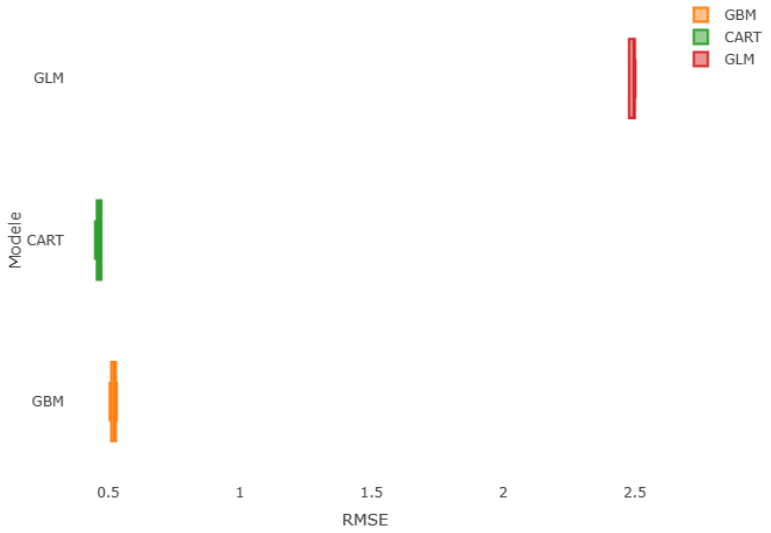

Figure 10: Erreur de prédiction

À travers `Firgure 10`,le niveau d'erreur moyen de prédiction des modèles GBM et CART est d'environ 0,4 et 0,5, tandis que le niveau d'erreur du modèle GLM peut atteindre 2,5.

2.3 Modélisation de la sévérité des sinistres

Premièrement, on sélectionne uniquement les données qui contiennent `claimValue > 0` pour trouver les sinistres attendu qu'une personne pourrait avoir lorsqu'elle dépose un sinistre. `ClaimValue < 50` est considéré comme trivial, `ClaimValue > 10000` est considéré comme trop grave, impossible à modéliser.

Visualiser les données pour mieux comprendre la distribution. On peut voir que cela apporte une forme similaire à la distribution gamma. Ainsi, la fonction objective de xgboosting sera définie comme "reg:gamma" et la même chose s'applique pour GLM. Cependant, le modèle CART ne supporte pas encore la méthode gamma, donc la technique de "ANOVA" est utilisée à cette place.

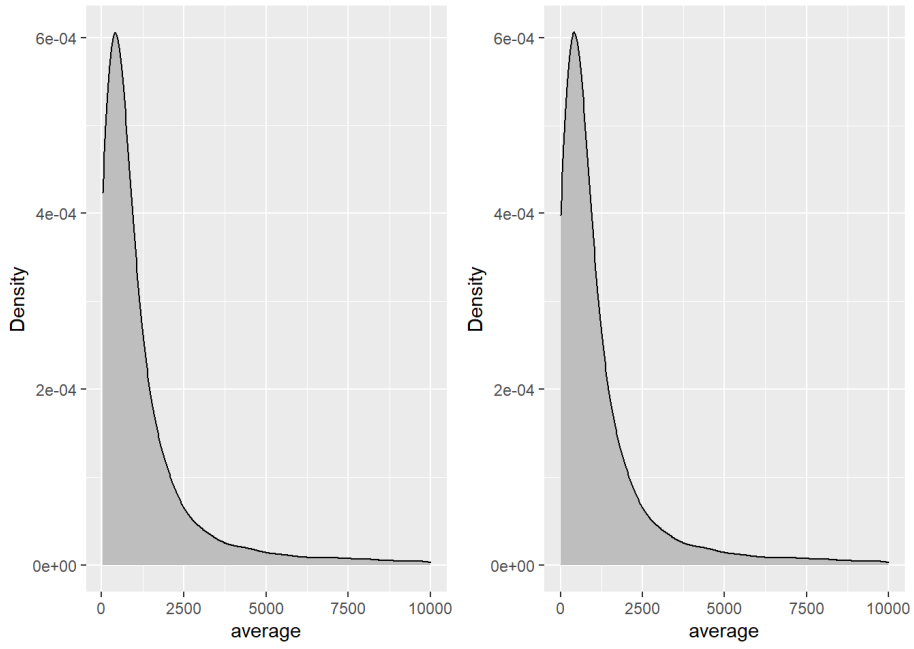

Figure 11: Distribution du montant des sinistres

La même stratégie de "tuning" pourrait être appliquée pour le modèle de prédiction de la sévérité

des sinistres. Tuning du nombre de cycles pour un taux d'apprentissage de valeur fixe. Pour le taux d'apprentissage de 0,1, la valeur optimale de nombre de cycles est d'environ 120.

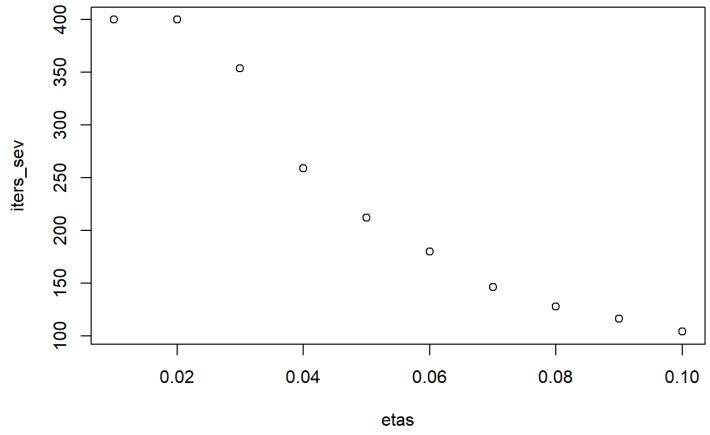

Figure 12: Max_depth pour le modèle de Sévérité

Tuning de la même max_depth que le modèle de fréquence. On choisit la valeur max_depth qui est la meilleure pour les 2 hyperparamètres déjà sélectionnés (nombre de cycles et le taux d'apprentissage eta) comme 2.

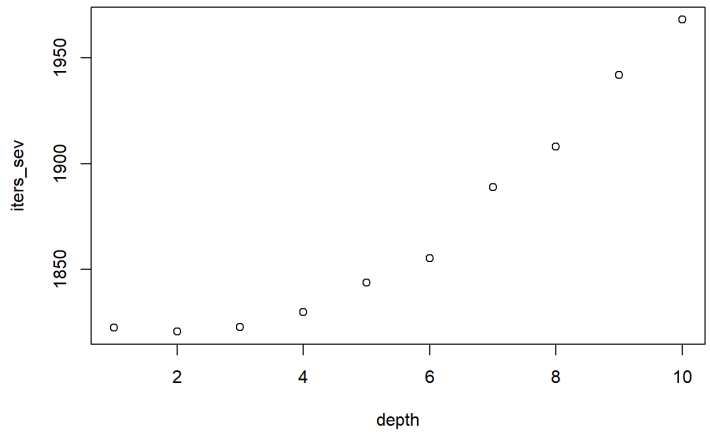

Figure 13: Le relation entre le nombre des cycles et le taux d'apprentissage

`Comparaison de différentes techniques de régression de la sévérité des sinistres.`
Malgré le fait que le mse n'est pas une bonne métrique pour estimer la performance d'un modèle lorsque les données sont asymétriques (on peut utiliser la déviance à cette place). Dans le but de synchroniser la métrique d'évaluation, la MSE (`Mean Squared Error`) est sélectionnée pour comparer ces modèles.

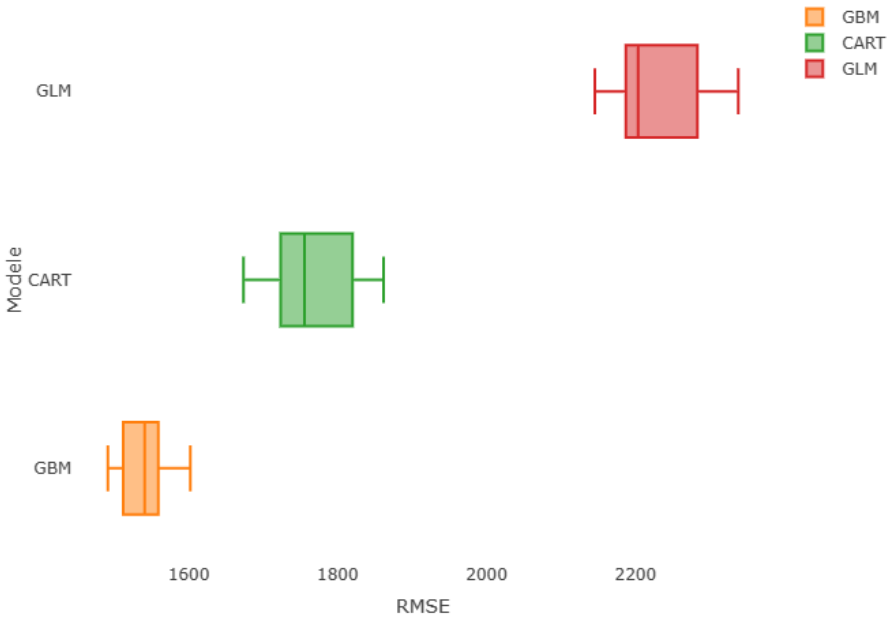

Figure 14: Erreur de prédiction

En conclusion, d'après les résultats des tests entre les différentes méthodes, la fréquence et la sévérité seront modélisées par la technique du boosting.

2.4 Interprétabilité

2.4.1 Fréquence

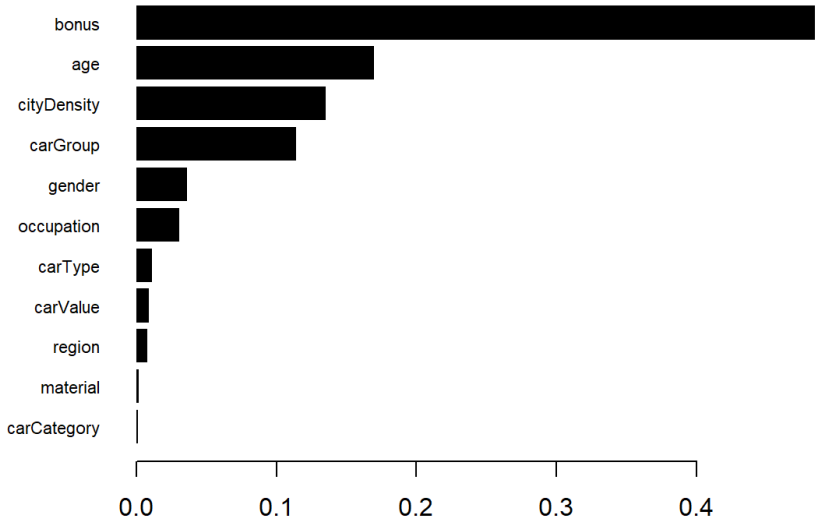

Figure 15: Erreur de prédiction

Utilise les diagrammes pour estimer et montrer l'importance dé variables explicatives avec la variable dependente. Les attributs `bonus` et `age` sont les variables qui ont le plus d'impact sur notre résult de prédiction.

Par figure 16, on peut réaliser que la tranche d'âge entre 20 et 40 ans présente la plus grande variance, et plus la tranche d'âge est ancienne, plus la stabilité est élevée.

Contrairement à la variable "Âge", plus le montant du bonus est élevé, plus la volatilité des résultats de prédiction est élevée. L'approche SHAP nous permet de creuser un peu plus dans la complexité des résultats du modèle prédictif, tout en nous permettant d'explorer les relations entre les variables pour le cas prédit. L'axe des y indique le nom de la variable, par ordre d'importance de haut en bas. L'axe des x indique l'importance de la variation en log-cotes. À

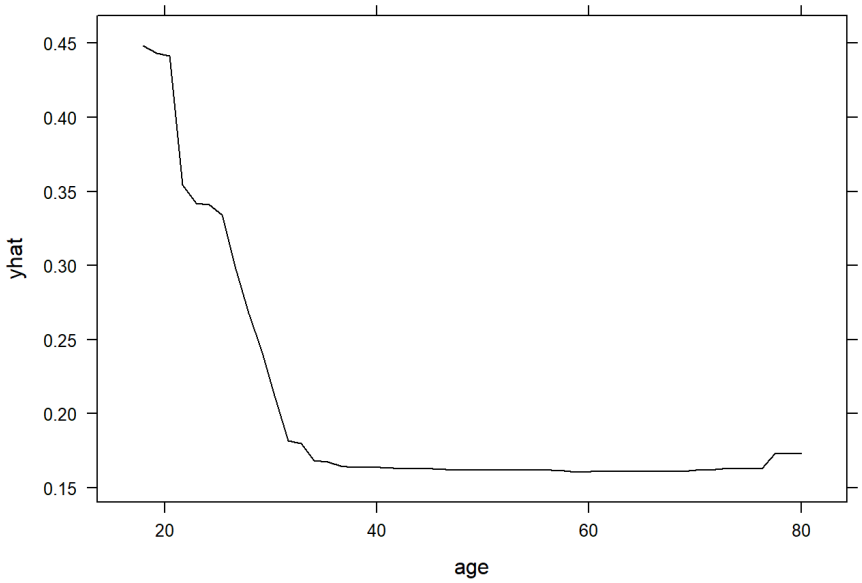

Figure 16: L'impact d'âge

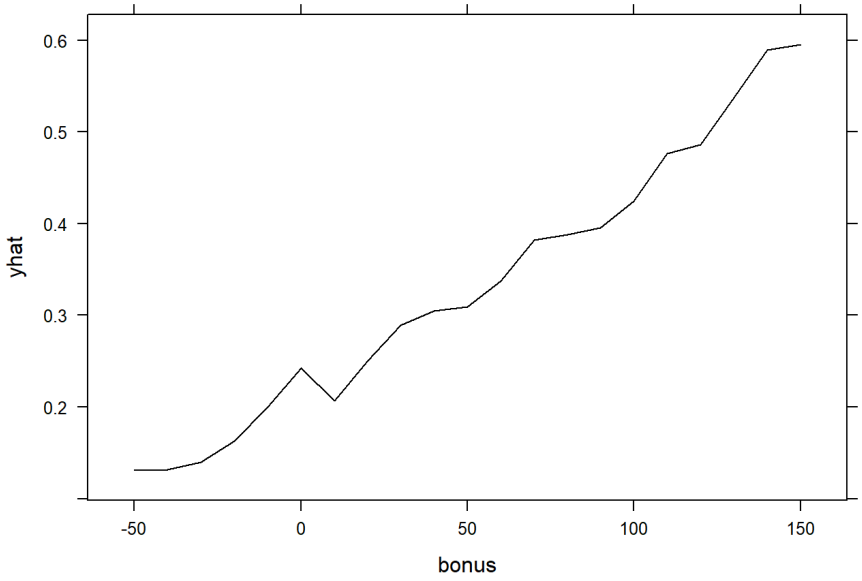

Figure 17: L'impact de bonus

partir de ce nombre, nous pouvons extraire la probabilité de succès.

Interpretation: On constate que le fait d'avoir un âge élevé est associé à des valeurs élevées et négatives sur la cible. Où le haut vient de la couleur et le négatif de la valeur x. Au contraire à variable "Âge", lorsque le niveau de bonus est élevé, la valeur du shap est élevée. Cette variable représente le niveau de risque élevé de l'observation, plus le bonus est élevé, plus le risque est élevé, ce qui conduit à une fréquence plus élevée de sinistre, puisque cela semble contre-intuitif.

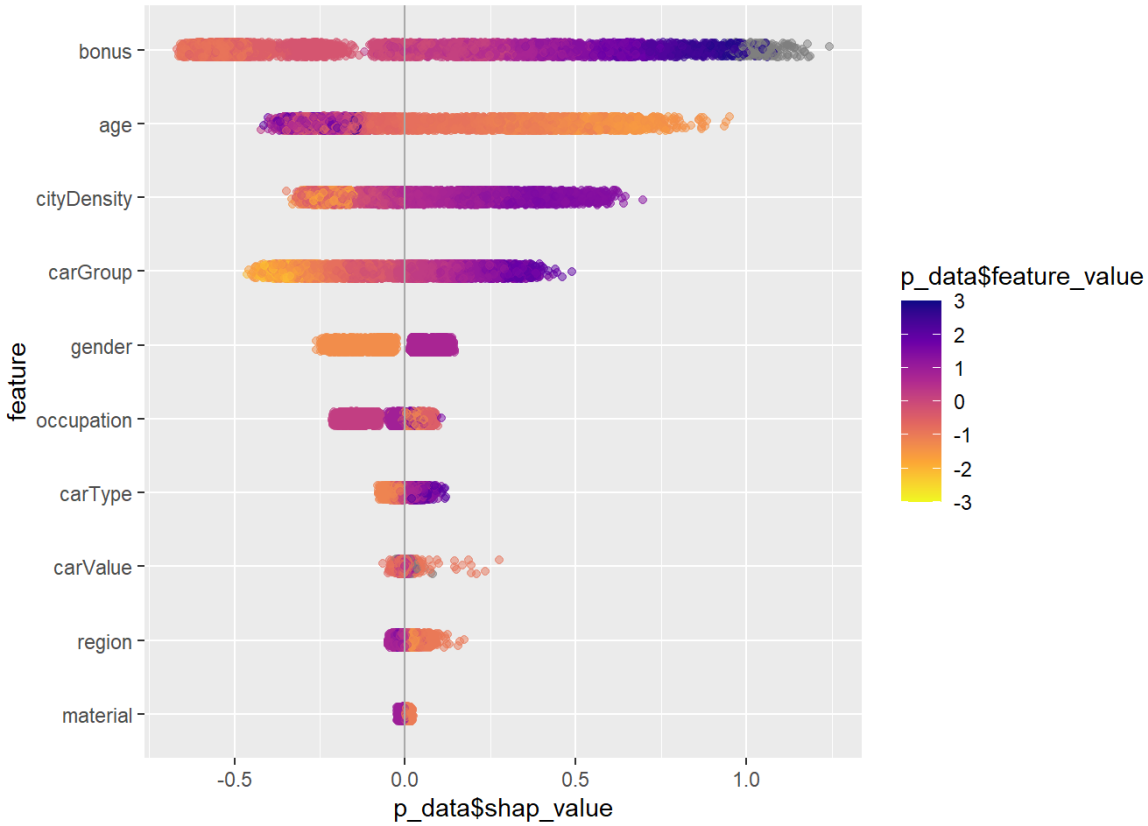

Figure 18: SHAP Value du modèle Fréquence

2.4.2 Sévérité

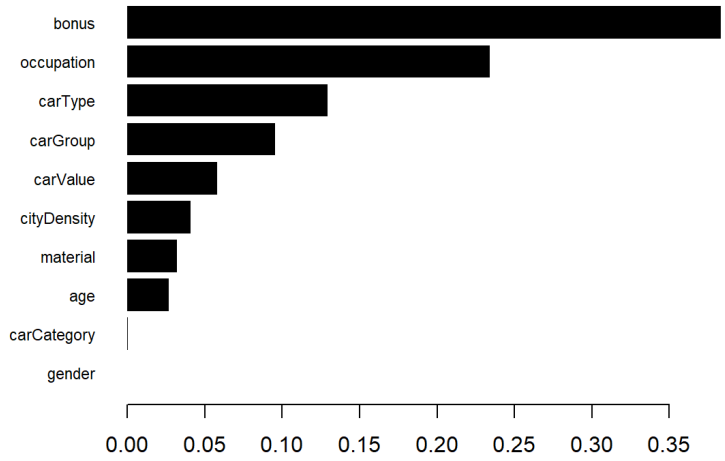

Figure 19: Variable importante du modèle Sévérité

Lors de l'analyse des extrêmes des observations, les attributs "Bonus" et "Occupation" sont ceux qui ont le plus d'impact sur le niveau de sévérité du sinistre, ou sur le montant total du sinistre.

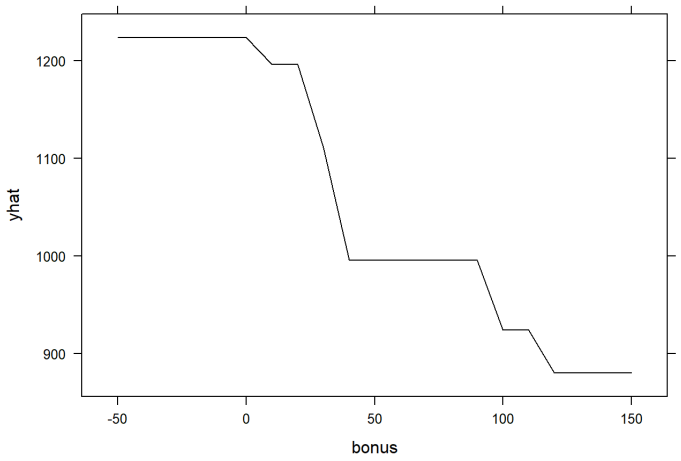

Figure 20: Impact de bonus

Sous le figure 20, on peut commenter l'influence et la volatilité de la variable Bonus sur la variable dépendante "Sévérité". Plus le niveau de bonus est élevé, par opposition à l'augmentation de la fréquence, plus la sévérité des accidents est faible.

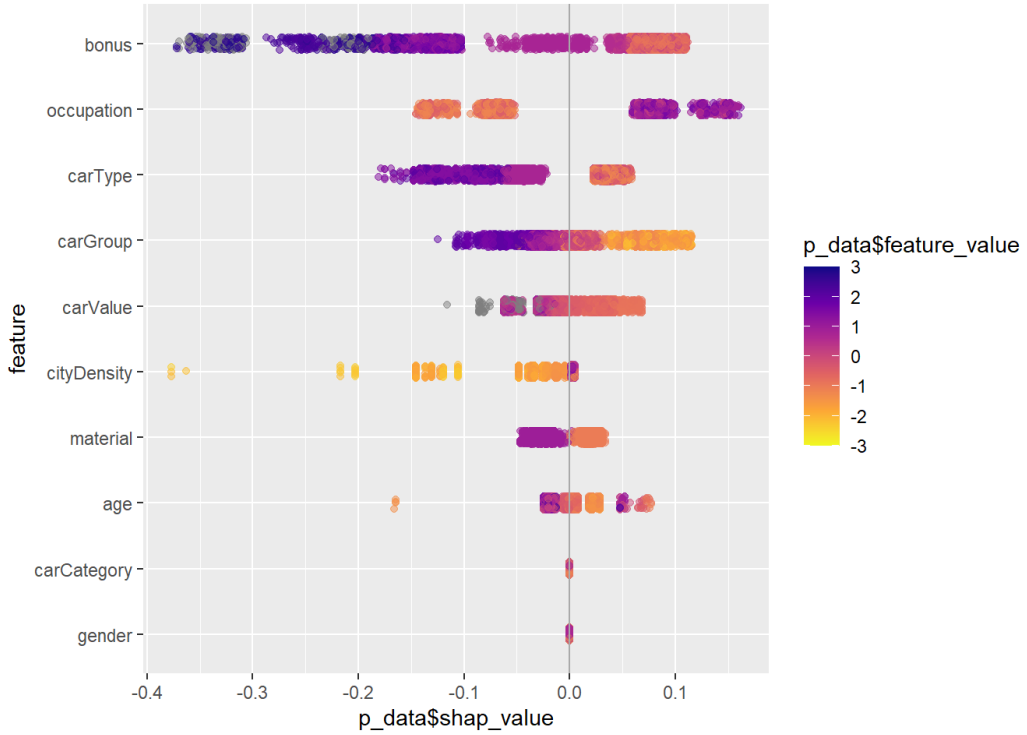

Figure 21: SHAP Value du modèle Sévérité

Le fonctionnement de Sharp Value a été expliqué. On peut observer que le fait d'avoir un bonus élevé est associé à des valeurs élevées et négatives sur les extrêmes des sinistres, ce qui est expliqué ci-dessus. Où élevé vient de la couleur et négatif du montant totale des sinistres.

Conclusion

Dans cette étude, nous avons adapté l'apprentissage automatique basé sur les arbres au problème de la tarification des assurances, sortant ainsi de la zone de confort de la tarification traditionnelle et de l'apprentissage automatique. Les GLM sont comparés aux arbres de régression et aux machines de renforcement de gradient. Ces techniques arborescentes peuvent être utilisées sur des données d'assurance, mais il faut faire attention aux hypothèses statistiques sous-jacentes sous la forme du choix de la fonction de perte. Tout d'abord, nous développons des plans tarifaires complets avec des techniques d'apprentissage automatique basées sur des arbres pour un portefeuille d'assurance vie réelle. Dans ce processus, même si nous utilisons la rmse, nous suggérons d'utiliser la déviance de poisson et gamma puisque la perte d'erreur quadratique classique n'est pas appropriée pour un problème de fréquence-sévérité. Deuxièmement, notre schéma de validation croisée élaboré donne une procédure de réglage bien pensée et minutieuse, nous permettant d'évaluer non seulement les performances de différentes méthodes, mais également la stabilité de nos résultats sur plusieurs plis de données. La machine d'amplification de gradient est systématiquement sélectionnée comme la meilleure approche de modélisation, à la fois par des mesures de performance hors échantillon et des critères d'évaluation de la portance du modèle. Cela implique qu'un assureur peut prévenir l'antisélection et générer des bénéfices en tenant compte de ce nouveau cadre de modélisation. Dans ce cas, une compagnie d'assurance peut toujours apprendre des informations précieuses sur la manière de constituer des portefeuilles rentables à partir d'un modèle technique interne et de les traduire en un produit commercial conforme à toutes ces exigences.

PREFERENCE

[1] - Henckaerts, Roel, et al. - **"Boosting insights in insurance tariff plans with tree-based machine learning methods"**
North American Actuarial Journal 25.2 (2021): 255-285

[2] - Frees, Edward W., Glenn G. Meyers and Richard A. Derrig - **"Predictive Modeling Applications in Actuarial Science."** 2004

[3] - M. Denuit, X. Marechal, S. Pitrebois, and JF. Walhin - **"Actuarial modeling of claim counts: Risk classification, credibility and bonus-malus systems".** John Wiley Sons Ltd, West Sussex, 2007

[4] - M. Denuit and S. Lang - **"Non-life rate-making with Bayesian GAMs.".**
Insurance: Mathematics and Economics, 35(3):627–647, 2004