OPTIMISATION DE LA TARIFICATION EN ASSURANCE GRÂCE À L'AUTOCALIBRATION : ÉTUDE COMPARATIVE DES MODÈLES GLM, GAM ET GBM

Noms et prénoms de l'étudiant : LEFFE GABARY John Kevin
Master en sciences actuarielle
Faculté des sciences
Année académique : 2022-2023
Promoteur de mémoire : TRUFIN Julien

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Je tiens à exprimer ma profonde gratitude envers toutes les personnes qui m'ont soutenu tout au long de la réalisation de ce mémoire.

Tout d'abord, je tiens à remercier sincèrement mon superviseur, le Prof. Julien Trufin, pour ses conseils éclairés, son soutien constant et sa disponibilité tout au long de ce projet. Ses connaissances approfondies dans le domaine de l'assurance non-vie ont été inestimables pour la réussite de ce travail.

Je souhaite également exprimer ma reconnaissance envers le Professeur Arthur Charpentier pour avoir mis à disposition son référentiel GitHub contenant des codes sources essentiels à la réalisation de ce mémoire. Ses travaux en collaboration avec le Prof. Michel Denuit et le Prof. Julien Trufin ont été une source majeure d'inspiration et ont grandement contribué à la qualité de mes analyses.

Je remercie également mes collègues et amis qui ont apporté leur soutien moral et leur encouragement tout au long de ce projet. Leurs discussions et idées ont enrichi mon travail de manière significative.

Enfin, je n'oublierai pas de remercier ma famille pour leur amour, leur soutien indéfectible et leurs encouragements tout au long de mes études.

Ce mémoire n'aurait pas été possible sans le concours de toutes ces personnes formidables. Merci infiniment pour votre contribution à la réussite de ce travail.

RÉSUMÉ

Ce mémoire propose une exploration de l'impact l'autocalibration dans le domaine de la tarification en assurance. L'étude se concentre spécifiquement sur trois catégories de modèles : les Modèles Linéaires Généralisés (GLM), les Modèles Additifs Généralisés (GAM) et les Machines à Gradient Boosting (GBM), en évaluant les effets de l'autocalibration sur les métriques clés de déviance et de lift. Les résultats de cette recherche mettent en évidence des améliorations substantielles après l'application de l'autocalibration. De manière particulièrement remarquable, cette technique affine l'ajustement des modèles en réduisant les valeurs de déviance. Les valeurs prédites par les modèles autocalibrés se superposent étroitement aux observations réelles, conférant ainsi une robustesse accrue à la précision des prédictions. L'analyse approfondie des courbes de lift simples pour chaque modèle démontre des améliorations significatives en matière de précision prédictive à la suite de l'autocalibration. À travers une variété de déciles de risque, les prédictions autocalibrées témoignent d'une proximité accrue avec les valeurs réelles. De surcroît, les courbes de lift doubles mettent en exergue la manière dont l'autocalibration réduit l'écart entre les valeurs prédites et les valeurs réelles, amplifiant la capacité des modèles à discriminer entre risques favorables et risques défavorables.

Mots-clés : Autocalibration, Tarification en assurance, Précision prédictive, Déviance, Courbes de lift, Gestion des risques.

ABSTRACT

This master's thesis offers a comprehensive exploration of the impact of autocalibration in the realm of insurance pricing. The study specifically focuses on three model categories : Generalized Linear Models (GLM), Generalized Additive Models (GAM), and Gradient Boosting Machines (GBM), evaluating the effects of autocalibration on key deviance and lift metrics. The research findings reveal substantial improvements resulting from the application of autocalibration. Notably, this technique enhances model fitting by reducing deviance values. Predicted values from autocalibrated models closely align with actual observations, thereby bolstering the robustness of predictive accuracy. A thorough analysis of simple lift curves for each model demonstrates significant enhancements in predictive precision post-autocalibration. Across various risk deciles, autocalibrated predictions exhibit closer proximity to actual values. Additionally, double lift curves illustrate how autocalibration diminishes the disparity between predicted and actual values, thereby magnifying the models' capability to differentiate between favorable and unfavorable risks.

Keywords : Autocalibration, Insurance pricing, Predictive Accuracy, Deviance, Lift Curves, Risk Management.

INTRODUCTION

Dans le domaine des sciences actuarielles, en particulier dans le contexte de l'assurance non-vie, la modélisation de la fréquence des sinistres revêt une importance capitale pour les compagnies d'assurance. L'estimation précise et équilibrée du nombre de sinistres susceptibles de se produire au sein d'une population d'assurés revêt une importance capitale pour évaluer les risques, établir les primes et prendre des décisions éclairées en matière de gestion des risques. Cette tâche de modélisation est complexe en raison de la nature aléatoire et imprévisible des événements de sinistres. Dans ce contexte, le développement de modèles statistiques robustes capables de saisir les caractéristiques de la fréquence des sinistres et de fournir des prédictions précises est essentiel.

Historiquement, les modèles linéaires généralisés (GLM) ont été largement utilisés pour estimer les primes en se basant sur des relations linéaires entre les variables explicatives et la variable de réponse. Cependant, avec les avancées technologiques et l'abondance de données, les modèles linéaires peuvent parfois se révéler insuffisants pour capturer la complexité des relations entre les variables. Face à cette limitation, des approches plus sophistiquées ont émergé, telles que les modèles additifs généralisés (GAM) et les méthodes d'ensemble comme les forêts aléatoires (RF) et les machines à gradient boosting (GBM). Ces techniques offrent une meilleure corrélation entre les primes prédites et les réclamations réelles. Cependant, elles peuvent parfois introduire un déséquilibre entre les primes estimées et les primes réellement observées, remettant en question la fiabilité des modèles de tarification.

Face à cette nécessité d'assurer que les primes calculées ne soient pas seulement précises, mais aussi impartiales et fidèles à la réalité des risques sous-jacents, Denuit, Charpentier, et Trufin (2021) ont développé une approche innovante basée sur le concept d'autocalibration. Cette méthode vise à restaurer l'équilibre global et local au sein des modèles de tarification.

Le présent mémoire a pour objectif de présenter cette nouvelle approche et d'évaluer son impact sur les mesures de performance à travers deux métriques d'évaluation de modèles : la déviance et les courbes de lift. Pour ce faire, nous débuterons par une exploration approfondie des bases théoriques des modèles GLM, GAM et GBM. Ensuite, nous étudierons en détail les principes fondamentaux de l'autocalibration, ainsi que sa mise en œuvre pratique dans le cadre de ces modèles. Enfin, nous évaluerons l'impact de l'autocalibration sur la qualité de la tarification en examinant sa capacité à rétablir l'équilibre entre les primes estimées et observées, en utilisant les courbes de lift et la déviance.

À travers cette analyse approfondie, nous aspirons à fournir une compréhension approfondie de l'importance de l'équilibre au sein des modèles de tarification et à offrir des éclairages pertinents sur l'efficacité de l'autocalibration pour restaurer l'équilibre dans le contexte de la tarification en assurance.

CHAPITRE 1

MODÈLES LINÉAIRES ET ADDITIFS GÉNÉRALISÉS

Sommaire

Dans le domaine de l'assurance non-vie, la modélisation de la fréquence de sinistres d'un portefeuille est d'une importance capitale pour évaluer les risques et calculer les primes d'assurance. Les actuaires cherchent à expliquer la valeur moyenne d'une variable cible, qui représente généralement des quantités telles que le nombre moyen de sinistre et ou le montant moyen de sinistre attendus sur une période donnée. Pour ce faire il font appel a diffèrents modèle de prédiction de la fréquence de sinistres et ou montant de sinistres. Parmis ces modèles on retrouve les modèles linéaires généralisés. Ils constituent un benchmark dans la tarification en assurance non vie. Dans ce chapitre nous presentons ces modèles ainsi qu'une de leur extension phare les modèles additifs généralisés.

1.1 Modèles Linéaires Généralisés (GLM)

Les Modèles Linéaires Généralisés sont une classe de modèles statistiques qui offre une approche flexible pour modéliser la relation entre une variable de réponse et des variables explicatives. Ils ont été introduits par John Nelder et Robert Wedderburn pour unifier différents types de modèles statistiques, tels que la régression linéaire, la régression logistique et la régression de Poisson. Les GLM permettent de modéliser la distribution de la variable de réponse en utilisant une fonction de lien pour relier la moyenne de la variable de réponse au prédicteur linéaire.

1.1.1 Structure des GLM

La modélisation d'une variable de réponse Y, en utilisant un GLM, implique trois composantes principales (Denuit et al., 2021, p. 99) : une composante aléatoire qui fait référence à la distribution de probabilité de la variable de réponse, une composante systématique qui renvoie au prédicteur linéaire, également appelé score, et la fonction de lien qui relie la moyenne de la variable de réponse au score.

Distribution de probabilité (Composante aléatoire)

La spécification de la distribution de probabilité constitue l'une des étapes fondamentales et incontournables des GLM. Elle revêt une importance capitale car elle permet de définir le comportement statistique de la variable de réponse. Pour ce faire, les GLM font appel à une famille de distributions connue sous le nom d'Exponentielle Dispersée (ED), qui englobe un ensemble de distributions caractérisées par des fonctions de densité de probabilité (PDF) ou de masse de probabilité (PMF) spécifiques. Cette famille de distributions ED est définie par un ensemble de paramètres cruciaux.

Pour chaque individu i, la distribution de probabilité dans un GLM de la famille ED est exprimée à travers la fonction de densité de probabilité (PDF) ou de masse de probabilité (PMF) suivante :

$$f_{Y_i}(y_i|\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - a(\theta_i)}{\frac{\phi}{\nu}}\right] \cdot c(y_i, \frac{\phi}{\nu}) \quad i = 1,\dots,n$$

Dans cette formulation :

- y_i [1] représente la valeur observée de la variable de réponse Y_i pour un individu donné.
- θ_i est un vecteur de paramètres de localisation à valeur réelle, également appelé paramètre canonique, qui est généralement lié à la moyenne de la distribution de Y_i.
- ϕ est le paramètre de dispersion positif qui contrôle la dispersion de la distribution en fonction de la valeur prédite par le modèle.
- ν est une constante positive connue sous le nom de poids. Les poids sont utilisés pour prendre en compte les cas où les observations sont regroupées ou pondérées différemment. Dans le cas où les données ne sont pas pondérées (comme dans notre cas), ν est égal à 1.
- $a(.)$ est une fonction convexe monotone de θ, connue sous le nom de fonction cumulante.Cette fonction joue un rôle crucial dans la détermination de la forme de la distribution de probabilité.
- $c(.)$ est une fonction de normalisation positive assurant que l'intégrale ou la somme sur toutes les valeurs possibles de y est égale à 1 (Denuit, Hainaut, & Trufin, 2019, p. 32).

L'évaluation de l'espérance et la variance de la variable de réponse Y_i pour chaque individu i se fait à l'aide des paramètres θ et ϕ :

1. Dans un modèle GLM, les réponses $Y_1, Y_2, \dots, Y_n$ mesurées sur les individus sont supposées indépendantes mais non nécessairement identiquement distribuées

$$\mu_i = E(Y_i) = a'(\theta_i)$$

$$\text{Var}(Y_i) = a''(\theta_i) \cdot \cdot \frac{\phi}{\nu}$$

Il convient de souligner que le choix de la distribution de probabilité dépend étroitement des caractéristiques propres à la variable de réponse. Dans le contexte spécifique de la modélisation de la fréquence de sinistres, il est courant d'opter pour la distribution de Poisson, laquelle se positionne comme un candidat naturel pour modéliser le nombre de sinistres signalés par les assurés (Denuit et al., 2019, p. 184).

Prédicteur linéaire ou Score (Composante systématique)

Le prédicteur linéaire, également appelé score, est une composante systématique des GLM qui permet de modéliser la relation entre la variable de réponse et les variables explicatives. Le score est construit en combinant linéairement les variables explicatives avec des coefficients correspondants. Pour une observation i donnée, le score η_i peut être calculé comme suit :

$$\eta_i = \mathbf{x}_i^T \cdot \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j \cdot x_{ij}$$

où :
- $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^T$ est le vecteur de variables explicatives pour l'observation i.
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ [2] est le vecteur de coefficients de régression à estimer.
- x_{ij} [3] est la valeur de la variable explicative j pour l'observation i.

Le prédicteur linéaire permet de capturer les relations linéaires entre les variables explicatives et la variable de réponse. Les coefficients $\beta_0, \beta_1, \ldots, \beta_p$ sont estimés à partir des données d'observation afin de trouver la meilleure combinaison linéaire des variables explicatives pour prédire la variable de réponse.

Fonction de lien

La fonction de lien est une composante essentielle des GLM qui relie la moyenne de la variable de réponse au prédicteur linéaire. Au lieu de modéliser directement la moyenne μ_i en fonction du score linéaire, une fonction de lien est utilisée pour transformer la moyenne de manière à ce qu'elle puisse varier dans l'ensemble des réels. Cette approche permet de prendre en compte les particularités de la variable réponse et d'obtenir des résultats interprétables.

La fonction de lien joue un rôle crucial dans la modélisation des effets des variables explicatives sur la variable de réponse. Elle permet de spécifier la forme de la relation entre le prédicteur linéaire et la moyenne de la variable de réponse. En utilisant la fonction de lien, on peut exprimer μ_i en fonction du score linéaire en inversant (grâce à la bijection de celle-ci) cette fonction et obtenir :

$$g(\mu_i) = \eta_i \iff \mu_i = g^{-1}(\eta_i)$$

où $g(\mu_i)$ est la fonction de lien appliquée à la moyenne μ_i et $g^{-1}(\text{score}_i)$ est l'inverse de la fonction de lien appliquée au score score_i.

Il est important de sélectionner une fonction de lien appropriée en fonction de la nature de la variable réponse, des objectifs de modélisation et des hypothèses sur la relation entre le score linéaire et la moyenne de la réponse. Dans la pratique, différentes fonctions de lien peuvent être utilisées, mais il est courant dans ce modèle, d'utiliser la fonction de lien canonique correspondant à la distribution de probabilité de la variable réponse, pour la propriété d'équilibre qu'elle confère.

1.1.2 GLM avec distribution de Poisson

La distribution de Poisson est couramment utilisée pour modéliser les variables discrètes, telles que les fréquences de sinistres, qui représentent le nombre de sinistres sur une période de temps donnée. La

2. La présence d'un intercept β_0 explique pourquoi le premier l'élément du vecteur $\mathbf{x}_i$ est un 1.
3. Les caractéristiques x_{ij} sont supposées être enregistrées sans erreurs ni valeurs manquantes.

distribution de Poisson est définie par un paramètre λ, qui représente la fréquence moyenne de sinistres attendus sur une unité de temps donnée.

La fonction de densité de probabilité de la distribution de Poisson est donnée par :

$$f(y|\lambda) = \frac{e^{-\lambda} \cdot \lambda^y}{y!}$$

où y est le nombre de sinistres observé et λ est le paramètre de localisation, représentant le taux de sinistres.

L'espérance et la variance de la distribution de Poisson peuvent être exprimées en fonction du paramètre λ :

$$\mathbb{E}(Y) = \lambda$$
$$\text{Var}(Y) = \lambda$$

1.1.3 Fonction de lien logarithmique

Pour la modélisation de la fréquence de sinistres avec une distribution de Poisson, la fonction de lien canonique est la fonction logarithmique. La fonction de lien logarithmique est définie comme suit :

$$g(\mu) = \ln(\mu)$$

où μ représente la moyenne de la variable de réponse.

La fonction de lien logarithmique permet d'interpréter les effets des variables explicatives sur la fréquence de sinistres de manière multiplicative plutôt qu'additive. Cela signifie que pour chaque unité de changement dans une variable explicative, l'effet sur la fréquence de sinistres est proportionnel à la valeur actuelle de la fréquence. En d'autres termes, la fonction de lien logarithmique permet de quantifier l'impact relatif des variables explicatives sur la fréquence de sinistres.

Par exemple, si le modèle estime un coefficient de régression de 0.2 pour une variable explicative donnée, cela signifie qu'une augmentation d'une unité dans cette variable est associée à une augmentation de 20% de la fréquence de sinistres, toutes choses étant égales par ailleurs.

1.1.4 Estimation par méthode du maximum de vraisemblance

Considérons un vecteur de réponses indépendantes Y de taille n appartenant à une famille de distributions exponentielles (ED) avec un paramètre canonique θ_i, qui dépend de μ_i via la relation $\mu_i = a'(\theta_i) = g^{-1}(\mathbf{x}_i^T\beta)$, et donc finalement de β. Considérons également une fonction de lien entre μ_i et les variables explicatives $x_1, x_2, \ldots, x_p$. Le but est d'estimer le vecteur de paramètres β en utilisant la méthode du maximum de vraisemblance (MV).

Fonction de vraisemblance et log-vraisemblance

La fonction de vraisemblance $L(y, \theta, \phi)$ est définie comme le produit des fonctions de densité de la distribution ED pour chaque observation i du vecteur de réponses Y. Elle s'exprime mathématiquement comme suit :

$$L(y, \theta, \phi) = \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - a(\theta_i)}{\frac{\phi}{\nu}}\right] \cdot c(y_i, \frac{\phi}{\nu})$$

Pour simplifier les calculs, il est courant d'utiliser la fonction log-vraisemblance $\ln(L(y, \theta, \phi))$, qui équivaut au logarithme naturel de $L(y, \theta, \phi)$. La log-vraisemblance est donnée par :

$$\ln(L(y, \theta, \phi)) = \sum_{i=1}^{n} \frac{y_i\theta_i - a(\theta_i)}{\phi} + \ln\left(c(y_i, \phi)\right)$$

Méthode du maximum de vraissemblance

Pour estimer les paramètres β optimaux, on recherche les valeurs de β qui maximisent la fonction log-vraisemblance. Cela équivaut à trouver les paramètres qui rendent les observations observées les plus probables, c'est-à-dire :

$$\hat{\beta} = \arg\max_{\beta} \ln(L(y, \theta, \phi)) = \arg\max_{\beta} \sum_{i=1}^{n} \left(\frac{y_i \theta_i - a(\theta_i)}{\phi} + \ln\left(c(y_i, \phi).\right) \right)$$

Ce qui conduit aux équations de dérivées partielles suivantes :

$$\frac{\partial \ln(L(y, \theta, \phi))}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \quad \forall j$$

avec $V(\mu_i) = a''(\theta_i)$ appelée fonction variance.

Ces équations de dérivées partielles permettant de trouver les paramètres β optimaux sont résolues à l'aide de méthodes itératives telles que l'algorithme de Fisher Scoring, l'algorithme de Newton-Raphson ou la méthode des moindres carrés itérativement pondérés (IRLS) (voir (Denuit et al., 2019, p. 112-129) ou (Wood, 2006, p. 63-67) pour plus de détails). Ces méthodes ajustent les valeurs des paramètres β à chaque itération jusqu'à ce qu'ils convergent vers les valeurs optimales qui maximisent la fonction log-vraisemblance.

Intervention des logiciels

Dans la pratique, l'estimation des paramètres dans les GLM est grandement facilitée par l'utilisation de logiciels usuels de modélisation statistique tels que R, SAS, Python. Ces logiciels fournissent des fonctions dédiées pour ajuster des GLM et effectuer l'estimation des paramètres.

Cas spécifique d'un modèle GLM avec distribution de Poisson

Dans le modèle GLM avec distribution de Poisson, avec fonction de lien canonique (fonction de lien logarithmique), la fonction de vraisemblance est donnée par :

$$L(y, \theta, \phi) = \prod_{i=1}^{n} \frac{\exp(-\lambda_i) \lambda_i^{N_i}}{N_i!}$$

où N_i est le nombre de sinistres observé pour l'individu ou police d'assurance i pendant une période d'observation v_i, et $\lambda_i = \exp(\mathbf{x}_i^T \beta)$ représente la fréquence attendue des sinistres.

La résolution de l'équation :

$$\hat{\beta} = \arg\max_{\beta} \ln(L(y, \theta, \phi)) = \arg\max_{\beta} \sum_{i=1}^{n} \left(-\lambda_i v_i + N_i \ln(\lambda_i) - \ln(N_i!)\right)$$

aboutit aux équations de dérivées partielles suivantes (comme nous l'avons vu précédemment de manière générale) :

$$\frac{\partial \ln(L(y, \theta, \phi))}{\partial \beta_j} = \sum_{i=1}^{n} (N_i - \lambda_i) \, x_{ij} = 0 \quad \forall j$$

Ces équations permettent d'obtenir les estimateurs $\hat{\beta}$ optimaux.

Propriété d'équilibre du modèle

Un résultat primordial dans l'estimation des paramètres par la méthode du maximum de vraisemblance, dans les modèles GLM avec fonction de lien canonique, est la propriété d'équilibre (issue du théorème 4.5 du livre de (Bühlmann & Gisler, 2005)). Dans le cas spécifique du modèle de Poisson avec fonction de lien logarithmique,

cette propriété est satisfaite. Elle implique que la somme des fréquences de sinistres prédites par le modèle pour toutes les observations doit être égale à la somme des fréquences réellement observées :

$$\sum_{i=1}^{n} v_i \cdot \lambda_i = \sum_{i=1}^{n} N_i$$

où v_i représente l'exposition au risque pour chaque police d'assurance i.

Cette propriété revêt une importance cruciale en tarification d'assurance, car elle garantit que le tarif calculé à partir du modèle est bien équilibré et cohérent avec les fréquences observées.

1.1.5 Mesure de performance

L'évaluation de la performance ou la qualité d'un modèle GLM est essentielle pour juger de son aptitude à s'ajuster aux données et à généraliser les résultats sur de nouvelles observations. Plusieurs techniques sont couramment utilisées pour évaluer la performance d'un modèle GLM, notamment la déviance dans la modélisation de la fréquence de sinistres.

Déviance

La déviance est une mesure utilisée pour évaluer l'ajustement d'un modèle GLM aux données observées. Sous sa forme normalisée, elle est calculée en comparant la log-vraisemblance du modèle ajusté $\ell(\hat{\mu}|Y)$ avec celle d'un modèle saturé $\ell(Y|Y)$ (un modèle parfaitement ajusté aux données) et en multipliant par un coefficient (-2). Mathématiquement, elle est exprimée comme suit :

$$D^*(\hat{\mu}|Y) = -2\left(\ell(Y|Y) - \ell(\hat{\mu}|Y)\right) = 2\sum_{i=1}^{n} \frac{\nu_i}{\phi}\left[y_i(\theta_i - \hat{\theta}_i) - a(\theta_i) + a(\hat{\theta}_i)\right]$$

La déviance est une mesure de l'ajustement global du modèle, où chaque observation contribue à sa valeur. Un D^* élevé indique que le modèle ajusté s'éloigne du modèle saturé pour plusieurs observations, tandis qu'une faible valeur de déviance indique un ajustement proche du modèle saturé pour plusieurs observations.

Plus formellement un test statistique d'adéquation du modèle est effectué en comparant sa valeur avec le quantile de niveau $1-\alpha$ de la distribution du chi-carré avec $n-p-1$ degrés de liberté. En effet, pour un niveau de confiance donné $1-\alpha$, si la déviance normalisée $D^*(\hat{\mu}|Y)$ dépasse le quantile de niveau $1-\alpha$ de la distribution du chi-carré avec $n-p-1$ degrés de liberté, on rejette l'hypothèse nulle selon laquelle le modèle est adéquat. Sinon, on accepte l'hypothèse nulle et considère le modèle comme approprié.

Pour le modèle de Poisson, la déviance non normalisée, notée $D^(\hat{\mu}|Y)$, est donnée par :

$$D^(\hat{\mu}|Y) = 2\sum_{i=1}^{n} \nu_i\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$$

Validation croisée

La validation croisée est une technique d'évaluation de la performance d'un modèle GLM qui permet d'estimer l'erreur de généralisation du modèle sur de nouvelles données non utilisées pour l'estimation des paramètres.

L'une des méthodes de validation croisée couramment utilisées est la validation croisée k-fold. Elle consiste à diviser l'ensemble de données en k sous-ensembles disjoints de taille similaire. Le modèle GLM est ensuite ajusté k fois en utilisant chaque sous-ensemble à tour de rôle comme ensemble de validation et les autres sous-ensembles comme ensemble d'entraînement. L'erreur de généralisation est ensuite estimée en moyennant les erreurs obtenues pour chaque itération. Cette erreur de généralisation peut être calculée à l'aide de la validation croisée k-fold en utilisant la déviance comme mesure.

Formellement, l'erreur de généralisation est donnée par :

$$PE(\hat{\mu}) = \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in D_j} D(y_i, \hat{\mu}_{-})$$

où :

— k est le nombre de partitions dans la validation croisée.
— D_j représente la j-ème partition (ou fold) des données.
— n est le nombre total d'observations dans l'ensemble d'apprentissage.
— $D(y_i, \hat{\mu}_{-})$ est la déviance calculée entre la variable de réponse réelle y_i et la valeur prédite $\hat{\mu}_{-}$.
— $\hat{\mu}_{-}$ désigne le modèle estimé entraîné sur toutes les données sauf la j-ème partition.

Une erreur de généralisation faible indique une meilleure capacité du modèle à prédire les nouvelles valeurs cibles, tandis qu'une erreur de généralisation plus élevée suggère un mauvais ajustement du modèle aux nouvelles données.

Sélection du modèle : Critères d'information

Pour comparer différents modèles GLM et sélectionner le modèle optimal, en plus de l'estimation de l'erreur de généralisation, des critères d'information tels que l'AIC (Critère d'Information d'Akaike) et le BIC (Critère d'Information Bayésien) sont couramment utilisés. Ces critères prennent en compte à la fois l'ajustement du modèle aux données et sa complexité.

L'AIC est calculé à l'aide de la formule suivante :

$$\text{AIC} = -2\log(L) + 2k$$

où L est la vraisemblance maximale du modèle et k est le nombre de paramètres estimés. Un modèle avec un AIC plus faible est préféré car il ajuste bien les données tout en ayant moins de paramètres.

Le BIC est calculé à l'aide de la formule suivante :

$$\text{BIC} = -2\log(L) + k\log(n)$$

où n est la taille de l'échantillon. Le BIC pénalise davantage les modèles plus complexes que l'AIC. Ainsi, le modèle avec un BIC plus faible est considéré comme préférable.

En utilisant ces critères d'information, il est possible de comparer différents modèles GLM et de sélectionner celui qui offre le meilleur compromis entre l'ajustement aux données et la complexité.

1.1.6 Avantages et limites des GLM

Les modèles GLM offrent de nombreux avantages :

— **Flexibilité :** Les GLM permettent la modélisation de différentes distributions de probabilité pour la variable réponse. Ils peuvent être appliqués à un large éventail de types de variables dépendantes, notamment les variables continues, binaires, catégorielles et de comptage (comme c'est le cas dans notre étude).
— **Interprétation aisée :** Les résultats des GLM sont faciles à interpréter. Les coefficients associés à chaque prédicteur fournissent des informations sur la direction et l'ampleur de l'effet de chaque variable explicative sur la variable réponse. Cela facilite la compréhension des relations entre les variables et permet de faire des inférences sur l'impact des prédicteurs sur le phénomène étudié.
— **Estimation robuste :** Les GLM utilisent la méthode du maximum de vraisemblance pour estimer les paramètres du modèle. Cette méthode fournit des estimateurs cohérents et asymptotiquement normaux, ce qui signifie que les estimations sont généralement fiables même avec des échantillons de taille limitée.
— **Gestion de la variance hétérogène :** Les GLM peuvent gérer les cas où la variance de la variable réponse n'est pas constante, c'est-à-dire la présence d'hétéroscédasticité, grâce à l'utilisation de la famille exponentielle dispersée. Cela améliore la précision des prédictions et permet de mieux modéliser la dispersion des données.

Cependant, les GLM présentent également quelques limitations :

— **Linéarité des relations :** Les GLM supposent une relation linéaire entre les prédicteurs et la variable réponse. Cependant, dans de nombreux cas réels, les relations peuvent être non linéaires (nous illustrerons cela dans l'une des section à venir). Lorsque cette hypothèse n'est pas vérifiée, les GLM peuvent fournir des ajustements insuffisamment précis et biaisés.
— **Gestion des interactions limitée :** Les GLM traditionnels ont des limitations dans la gestion des interactions entre les prédicteurs, surtout lorsque le nombre de prédicteurs est élevé. Inclure toutes les interactions potentielles est une tâches presque impossible, tandis que l'omission d'interactions importantes peut réduire la qualité des prédictions. La sélection appropriée des interactions est donc un défi important dans l'utilisation des GLM.
— **Sensibilité aux spécifications :** Les performances des GLM peuvent être sensibles aux spécifications, telles que le choix de la distribution et de la fonction de lien. Une spécification inappropriée peut entraîner des résultats incohérents et des estimations biaisées.
— **Distribution des résidus :** Les GLM supposent des distributions prédéfinies pour les résidus, telles que la distribution de Poisson ou la distribution binomiale négative. Cependant, la distribution réelle des résidus peut être plus complexe et ne correspondre que partiellement à ces distributions prédéfinies, ce qui peut affecter la précision des prédictions et la qualité de l'ajustement du modèle.
— **Colinéarité et identifiabilité :** La présence de colinéarité entre les prédicteurs peut rendre instable l'estimation des coefficients du modèle et rendre difficile l'identification des véritables facteurs de risque importants. La matrice de conception (*design matrix*) peut devenir singulière, ce qui invalide l'approche d'estimation utilisée dans les GLM (Denuit et al., 2019, p. 130-132).

Malgré ces limites, les Modèles Linéaires Généralisés restent un outil statistique puissant et largement utilisé pour la modélisation de données avec différentes distributions et types de variables dépendantes. Pour faire face à certaines de ces limites, des extensions des GLM, telles que les modèles additifs généralisés (GAM) ou les méthodes d'apprentissage automatique, peuvent être envisagées pour des analyses plus flexibles et précises.

1.2 Modèles Additifs Généralisés (GAM)

Les modèles additifs généralisés constituent une extension puissante des GLM permettant de modéliser les relations non linéaires et complexes entre une variable réponse y et une ou plusieurs variables prédictives $x_1, x_2, ..., x_p$. Ils ont été introduits par Hastie et Tibshirani (1986), et ont connu un développement significatif grâce aux contributions de Wood (2001). Cette section présente en détail les concepts fondamentaux des GAM, les techniques d'estimation et de sélection de modèles, ainsi que les considérations relatives à l'interprétation des résultats.

1.2.1 Structure générale des GAM

Tout comme dans les GLM, considérons des réponses $Y_1, Y_2, \ldots, Y_n$ mesurées sur n individus et un vecteur $x_i = (1, x_{i1}, ..., x_{ip})$ de dimension $p+1$ contenant les caractéristiques de l'individu i. Supposons de plus que ces caractéristiques sont subdivisées en p_{cat} variables catégorielles $x_{i1}, ..., x_{ip_{\text{cat}}}$ et $p_{\text{cont}} = p - p_{\text{cat}}$ variables continues $x_{i,\text{cat}+1}, ..., x_{ip}$. Dans un modèle GAM le score pour un individu i est donné par :

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j}^{p_{\text{cat}}} \beta_j x_{ij} + \sum_{j}^{p} f_j(x_{ij})$$

où :

— $g(.)$ est la fonction de lien qui lie la valeur attendue de la variable réponse Y à une combinaison linéaire des fonctions lisses des variables continues et des effets des variables catégorielles.
— μ_i est l'espérance de la variable réponse Y pour l'individu i.
— β_0 est l'intercept du modèle.
— p_{cat} est le nombre de variables catégorielles.
— p est le nombre total de prédicteurs (variables continues et catégorielles).
— x_{ij} est la variable prédicteur j (variable continue ou catégorielle) pour l'individu i.
— β_j sont les coefficients représentant les effets des variables catégorielles.
— $f_j(\cdot)$ sont les fonctions lisses des variables continues.

Les fonctions lisses $f_j(\cdot)$ sont des fonctions non paramétriques qui permettent de capturer les relations non linéaires entre les variables continues et la variable réponse. Elles peuvent être représentées à l'aide de bases de fonctions, telles que les bases de fonctions polynomiales ou les splines cubiques.

Les variables catégorielles sont représentées par des termes linéaires dans le modèle, où β_j est le coefficient correspondant à la variable catégorielle j. Ces termes linéaires permettent de capturer les effets des différentes catégories de la variable catégorielle sur la variable réponse.

1.2.2 Représentation des fonctions lisses

Dans cette section, nous abordons la représentation des fonctions lisses dans les Modèles Additifs Généralisés (GAM). Les fonctions lisses $f_j(x_j)$ sont des fonctions non paramétriques utilisées pour capturer les relations complexes entre les variables continues x_j et la variable réponse y. Les fonctions lisses sont modélisées en utilisant des bases de fonctions, telles que les bases de fonctions polynomiales ou les splines cubiques (James, Witten, Hastie, & Tibshirani, 2013, p. 275-282).

Bases de fonctions polynomiales

Les bases de fonctions polynomiales sont couramment utilisées pour représenter des fonctions lisses à faible complexité autour d'un point spécifié. Elles sont définies comme suit :

$$f_j(x_j) = \sum_{k=1}^{K} \beta_{jk} \cdot b_{jk}(x_j)$$

où :
- $f_j(x_j)$ est la fonction lisse pour la variable continue x_j.
- $b_{jk}(x_j)$ sont les fonctions de base polynomiales.
- β_{jk} sont les coefficients à estimer.

Les fonctions de base polynomiales $b_{jk}(x_j)$ peuvent être des polynômes d'ordre fixe, tels que les polynômes linéaires, quadratiques ou cubiques, ou des polynômes d'ordre variable.

Bases de splines cubiques

Les bases de splines cubiques offrent une représentation plus flexible des fonctions lisses sur l'ensemble du domaine des prédicteurs. Les splines cubiques sont construites à partir de sections de polynômes cubiques assemblées de manière à garantir la continuité en valeur et en dérivée aux points de jonction.

Les fonctions de base splines cubiques $b_{jk}(x_j)$ sont définies par des nœuds spécifiés x_i^* et sont définies comme suit :

$$f_j(x_j) = \sum_{i=1}^{q} \beta_{ji} \cdot R(x_j, x_i^*)$$

où :
- $f_j(x_j)$ est la fonction lisse pour la variable continue x_j.
- β_{ji} sont les coefficients à estimer.
- $R(x_j, x_i^*)$ est une fonction définie en morceaux qui prend la valeur $(x_j - x_i^*)^3$ pour $x_j \geq x_i^*$ et 0 sinon.
- x_i^* sont les nœuds de la spline cubique.

Les fonctions de base splines cubiques permettent de capturer des motifs complexes dans les données et offrent une flexibilité accrue pour modéliser des relations non linéaires.

En pratique, le choix entre les bases de fonctions polynomiales et les splines cubiques dépend de la complexité attendue des relations entre les variables continues et la variable réponse, ainsi que des contraintes de lissage souhaitées. Les splines cubiques sont souvent privilégiées lorsque des motifs non linéaires complexes sont anticipés, tandis que les bases de fonctions polynomiales peuvent être suffisantes pour modéliser des relations plus simples et locales. La sélection appropriée de la base de fonctions est essentielle pour obtenir un modèle GAM bien ajusté et informatif.

1.2.3 Calibration du modèle

Identifiabilité

Un problème important dans les GAM est que les fonctions lisses $f_j(x_{ji})$ ne sont identifiables qu'à une constante additive près. Cela signifie que l'ajout d'une constante à une fonction lisse ne modifie pas le modèle. Pour résoudre ce problème, des contraintes d'identifiabilité sont imposées sur les fonctions lisses en exigeant que la somme de toutes les valeurs prises par chaque fonction lisse soit égale à zéro :

$$\sum_i f_j(x_{ji}) = 0 \quad \text{pour tout } j = 1, \ldots, p$$

Ces contraintes peuvent être gérées en absorbant les constantes additives dans les fonctions de base, ce qui modifie les fonctions de base tout en éliminant le problème d'identifiabilité.

Contrôle du lissage

Pour contrôler le degré de lissage des fonctions f_j, des techniques de pénalisation sont utilisées dans l'estimation du modèle. La pénalité est définie comme la somme des carrés de la dérivée seconde des fonctions lisses :

$$\text{Pénalité} = \sum_j \int [f_j''(x)]^2 \, dx$$

Cette pénalité permet de contrôler la souplesse des fonctions lisses et d'éviter le surajustement du modèle aux données.

Estimation des paramètres et ajustement du modèle

L'estimation des paramètres du modèle GAM et l'ajustement des fonctions lisses aux données sont effectués en résolvant un problème d'optimisation. Nous cherchons à minimiser la déviance du modèle pénalisée par la pénalité de lissage :

$$\hat{\beta} = \arg\min_{\beta} \left\{ D(\beta) + \sum_j \lambda_j \int [f_j''(x)]^2 \, dx \right\}$$

où $D(\beta)$ est la déviance du modèle, λ_j sont les paramètres de lissage, β_j sont les coefficients des fonctions de base, et $\int [f_j''(x)]^2 \, dx$ est la pénalité pour la fonction lisse f_j (Wuthrich & Buser, 2023, p. 77-89).

Sélection des paramètres de lissage

La sélection des paramètres de lissage est une étape cruciale dans les Modèles Additifs Généralisés (GAM), car elle peut entraîner des ajustements très différents en faisant varier les paramètres de lissage λ_j. Les paramètres de lissage jouent un rôle essentiel dans la souplesse des fonctions lisses $f_j(x_j)$, qui capturent les relations non linéaires entre les variables prédictives et la variable réponse. Une souplesse insuffisante peut entraîner un sous-ajustement du modèle, tandis qu'une souplesse excessive peut conduire à un surajustement aux données.

Plusieurs méthodes peuvent être utilisées pour sélectionner les paramètres de lissage de manière optimale. Parmi ces méthodes, on trouve notamment :

- **La Validation Croisée (Cross-Validation, CV)** : Cette méthode consiste à diviser l'échantillon de données en plusieurs sous-échantillons. On ajuste ensuite le modèle en laissant un sous-échantillon de côté et en utilisant les autres pour l'estimation. On répète cette opération pour chaque sous-échantillon, et on calcule ensuite l'erreur de prédiction moyenne sur l'ensemble des sous-échantillons. Les valeurs des paramètres de lissage qui minimisent cette erreur de prédiction sont alors choisies.
- **La Validation Croisée Généralisée, en anglais Generalized Cross-Validation (GCV)** : La GCV est une méthode d'estimation du risque de prédiction qui permet de sélectionner les

paramètres de lissage sans avoir à réajuster le modèle plusieurs fois. Pour le cas général non gaussien, le GCV est défini par la formule suivante :

$$\text{GCV} = \frac{nD(\hat{\beta})}{(n - \text{trace}(F))^2}$$

où n est le nombre d'observations dans l'échantillon de données, $D(\hat{\beta})$ est la déviance pénalisée du modèle, $\hat{\beta}$ représente les coefficients estimés du modèle, et trace(F) est la somme des éléments diagonaux de la matrice de pénalité F associée aux fonctions lisses. Le GCV est utilisé pour sélectionner les valeurs optimales des paramètres de lissage en cherchant à minimiser l'erreur moyenne de prédiction tout en évitant le surajustement du modèle. Ce critère offre un avantage significatif par rapport à la validation croisée K-fold en termes d'efficacité computationnelle. Contrairement à la validation croisée K-fold, qui nécessite de réajuster le modèle K fois en divisant l'ensemble de données en K sous-ensembles, le critère GCV évalue la performance du modèle de manière analytique sans nécessiter de multiples ajustements du modèle. Cela le rend beaucoup plus rapide sur le plan computationnel, ce qui est essentiel dans les cas où le temps de calcul est un facteur critique (Wuthrich & Buser, 2023, p. 83). .

— **L'Estimation du Risque Non Biaisée (Unbiased Risk Estimation, UBRE)** : L'UBRE est une autre méthode d'estimation du risque généralisé qui permet également de sélectionner les paramètres de lissage sans réajuster le modèle à plusieurs reprises. Cependant, contrairement au GCV, l'UBRE estime l'erreur de prédiction en utilisant une approximation non biaisée du risque généralisé.

— **L'Estimation du Maximum de Vraisemblance Restreinte (Restricted Maximum Likelihood Estimation, RMLE)** : Le RMLE est une méthode d'estimation des paramètres dans les GAM qui utilise une version restreinte de la vraisemblance maximale. Contrairement à la vraisemblance maximale classique, le RMLE restreint certains paramètres, ce qui permet de régulariser le modèle et d'éviter le surajustement. Le RMLE estime les paramètres de lissage en cherchant à maximiser la vraisemblance restreinte du modèle.

En utilisant ces méthodes de sélection des paramètres de lissage, nous pouvons choisir les valeurs optimales des λ_j qui conduisent à un modèle GAM bien ajusté, informatif et adapté aux données observées (voir (Wood, 2006, p. 128-131) ou (Wuthrich & Buser, 2023, p. 82-90) pour plus de détails).

1.2.4 Interprétation dans les GAM

L'interprétation des fonctions lisses revêt une importance capitale pour comprendre le rôle des variables prédictives continues dans la prédiction de la variable réponse y. Les fonctions lisses sont des composantes non linéaires du modèle qui permettent de capturer les relations complexes entre les variables continues et la variable réponse. Pour mieux appréhender leur impact sur la variable réponse, les fonctions lisses peuvent être visualisées à l'aide de graphiques de lissage.

Les graphiques de lissage représentent graphiquement la forme des fonctions lisses en fonction des prédicteurs continus. Ils mettent en évidence des comportements non linéaires ou des interactions complexes entre les variables continues et la variable réponse. Par exemple, pour une variable continue x_j, le graphique de lissage affiche la courbe lisse $f_j(x_j)$ en fonction de x_j, montrant comment la variable x_j influence la variable réponse y de manière non linéaire.

1.2.5 Avantages et limites des GAM

Les Modèles Additifs Généralisés présentent plusieurs avantages significatifs par rapport aux modèles linéaires traditionnels :

— **Flexibilité :** Les GAM permettent de modéliser des relations non linéaires entre les variables prédictives et la variable réponse. Cela les rend adaptés à une large gamme de problèmes où les relations ne peuvent pas être bien approximées par des modèles linéaires.

— **Interprétation :** Les fonctions lisses utilisées dans les GAM permettent une interprétation plus aisée des relations entre les variables. Les graphiques de lissage peuvent être utilisés pour visualiser ces relations et aider à expliquer les résultats du modèle.

— **Capture des interactions :** Les GAM permettent de capturer facilement des interactions complexes entre les variables prédictives, ce qui est souvent difficile à réaliser avec les modèles linéaires traditionnels. Cela permet de modéliser des effets conjoints et synergiques entre les prédicteurs.

— **Robustesse aux valeurs aberrantes :** Les GAM sont généralement moins sensibles aux valeurs aberrantes que les modèles linéaires traditionnels, en particulier lorsque des techniques de lissage sont utilisées. Cela permet d'obtenir des estimations plus stables même en présence de données aberrantes.
— **Traitement des variables catégorielles :** Les GAM peuvent facilement intégrer des variables catégorielles dans le modèle en utilisant des termes linéaires pour représenter les effets des différentes catégories. Cela évite la nécessité de créer des variables indicatrices pour chaque catégorie, comme c'est souvent le cas avec les modèles linéaires traditionnels.

Malgré leurs avantages, les GAM présentent également certaines limites et considérations :

— **Choix de la fonction de lissage :** Le choix de la fonction de lissage peut être délicat, et il peut être difficile de déterminer quelle fonction est la plus appropriée pour représenter la relation entre une variable continue et la variable réponse. Une mauvaise spécification de la fonction de lissage peut entraîner un sous-ajustement ou un surajustement du modèle.
— **Choix des paramètres de lissage :** La sélection des paramètres de lissage est une étape cruciale dans les GAM, mais il n'existe pas de méthode unique pour choisir ces paramètres de manière optimale. Différentes méthodes de sélection peuvent conduire à des ajustements différents, et il est important de choisir judicieusement ces paramètres pour obtenir un modèle bien ajusté.
— **Interprétation des interactions :** Bien que les GAM permettent de capturer des interactions entre les variables prédictives, il peut être plus difficile d'interpréter ces interactions que dans les modèles linéaires traditionnels. Les interactions dans les GAM sont souvent plus complexes, ce qui peut rendre leur interprétation moins intuitive.
— **Complexité computationnelle :** Les GAM peuvent être plus complexes en termes de calculs que les modèles linéaires traditionnels, surtout lorsque de grandes quantités de données sont impliquées ou lorsque des fonctions de lissage complexes sont utilisées. Cela peut entraîner des temps de calcul plus longs et nécessiter des ressources informatiques supplémentaires.
— **Risque de surajustement :** Comme pour tout modèle statistique, il existe un risque de surajustement du modèle aux données d'entraînement, en particulier lorsque le degré de lissage est trop élevé. Une sélection appropriée des paramètres de lissage et l'utilisation de techniques de validation croisée peuvent aider à atténuer ce risque.

1.3 Implémentation des modèles GLM et GAM dans R

Avant de fournir les principaux résultats de notre analyse, nous faisons dans cette section, une brève présentation de notre base de données, une analyse descriptive sommaire de nos variable afin de détecter les différentes tendances existantes entres les variables explicatives et la réponse qu'est la fréquence de sinistre de notre portefeuille.

1.3.1 Description de la base de données

La base de données utilisée pour la construction des modèles présentés dans ce mémoire est la base de données "freMTPL2freq" du package *CASdatasets* dans R. Cette base de données est spécifiquement conçue pour l'analyse de la fréquence des sinistres dans le domaine de l'assurance automobile. Elle est basée sur des données réelles provenant d'entreprises d'assurance automobile en France, garantissant ainsi une représentation fidèle du domaine d'étude. Elle est composée de 678,013 observations et d'un ensemble de 12 variables pertinentes fournissant des informations clés sur les sinistres, les caractéristiques des véhicules assurés, les conducteurs et le contexte géographique permettant ainsi de capturer les facteurs qui peuvent influencer la fréquence des sinistres. Les variables explicatives dans notre base se présentent sous deux types principalement : les variables continues et les variables catégorielles.

Variables continues

— **Exposure** : Cette variable représente la période d'exposition du contrat d'assurance, mesurée en unités d'années. L'exposition est une variable essentielle pour ajuster la fréquence des sinistres, car un contrat exposé plus longtemps présente une probabilité plus élevée de réclamations. Par exemple, les contrats d'assurance avec une exposition plus élevée peuvent avoir une fréquence de sinistres plus élevée car ils sont actifs pendant une plus longue période, augmentant ainsi les chances d'accidents ou de dommages.

— **VehPower** : Il s'agit de la puissance du véhicule assuré. La puissance du véhicule peut être un indicateur du comportement de conduite, où des véhicules à plus haute puissance peuvent être associés à une conduite plus risquée et donc à une fréquence de sinistres plus élevée. Cette variable permettra de comprendre si les conducteurs de véhicules puissants sont plus susceptibles de déposer des réclamations d'assurance automobile.
— **VehAge** : Cette variable représente l'âge du véhicule assuré en années. L'âge du véhicule peut jouer un rôle important dans la fréquence des sinistres. Les véhicules plus anciens peuvent être plus susceptibles de présenter des problèmes mécaniques, ce qui peut entraîner un nombre accru de réclamations pour des réparations ou des accidents. En revanche, les véhicules plus récents peuvent être associés à des conducteurs plus prudents, ce qui pourrait entraîner une fréquence de sinistres plus faible.
— **DrivAge** : Il s'agit de l'âge du conducteur du véhicule assuré. L'âge du conducteur est un facteur bien établi dans l'analyse de l'assurance automobile, car les conducteurs plus jeunes et les conducteurs âgés ont tendance à présenter un risque plus élevé d'accidents. Les jeunes conducteurs, en raison de leur manque d'expérience, peuvent être plus enclins à prendre des risques au volant, tandis que les conducteurs plus âgés peuvent avoir une réactivité réduite, augmentant ainsi le risque d'accidents. Cette variable permettra d'évaluer l'impact de l'âge du conducteur sur la fréquence des sinistres.
— **BonusMalus** : Cette variable représente le coefficient de bonus-malus du conducteur, qui dépend de son historique de sinistres. Un coefficient de bonus élevé indique un historique de conduite responsable avec peu ou pas de sinistres, tandis qu'un coefficient de malus indique un historique comportant des sinistres. Cette variable sera importante pour comprendre comment l'historique de conduite du conducteur peut influencer la fréquence des sinistres.
— **Density** : Cette variable indique la densité de population de la région où le véhicule est assuré. Une densité de population plus élevée peut entraîner une circulation plus dense, augmentant ainsi les risques d'accidents et de réclamations. En revanche, dans les régions moins peuplées, le trafic pourrait être moins dense, ce qui pourrait entraîner une fréquence de sinistres plus faible. Cette variable sera essentielle pour comprendre comment l'environnement géographique peut jouer un rôle dans la fréquence des sinistres.

Variables catégorielles

En plus des variables continues, la base de données contient également des variables catégorielles :

— **VehBrand** : Cette variable représente la marque du véhicule assuré. La marque du véhicule peut être associée à des caractéristiques spécifiques du véhicule, de sa fiabilité à ses mesures de sécurité. Certaines marques peuvent être réputées pour leur sécurité et leur durabilité, ce qui pourrait se traduire par une fréquence de sinistres moins élevée, tandis que d'autres marques pourraient présenter un risque accru d'accidents. Cette variable permettra de comprendre comment la marque du véhicule peut influencer la fréquence des réclamations.
— **VehGas** : Il s'agit du type de carburant du véhicule assuré, soit "Regular" (essence ordinaire) ou "Diesel". Certains types de carburant pourraient être associés à une consommation de carburant plus élevée, ce qui pourrait augmenter la probabilité de déplacements plus fréquents et potentiellement de sinistres. Cette variable permettra d'évaluer l'impact du type de carburant sur la fréquence des sinistres.
— **Area** : Cette variable indique la zone géographique où le véhicule est assuré. Certaines zones géographiques peuvent présenter des caractéristiques spécifiques qui pourraient influencer la fréquence des sinistres. Par exemple, les zones urbaines avec une densité de population élevée pourraient avoir une fréquence de sinistres plus élevée en raison de la congestion du trafic, tandis que les zones rurales pourraient avoir une fréquence de sinistres moins élevée en raison de moins de trafic et d'accidents potentiels. Cette variable permettra de comprendre comment la zone géographique peut jouer un rôle dans la fréquence des sinistres.
— **Region** : Cette variable représente la région géographique où le contrat d'assurance est en vigueur. Les régions géographiques peuvent présenter des caractéristiques démographiques et géographiques uniques qui pourraient influencer la fréquence des sinistres. Par exemple, les régions avec des conditions météorologiques extrêmes pourraient connaître une fréquence de sinistres plus élevée. Cette variable permettra d'évaluer l'impact de la région géographique sur la fréquence des sinistres.

Ces variables fournissent des informations clés sur les caractéristiques des assurés, des véhicules et de leur environnement, ce qui permet de capturer les facteurs qui peuvent influencer la fréquence des

sinistres dans l'assurance automobile.

1.3.2 Statistique descriptive

Commençons par nous intéresser à notre variable de réponse qu'est la fréquence de sinistres. La Figure 1.1 illustre le caractère rare des sinistres automobiles dans l'échantillon. Sur la période d'observation, seuls 5,2% des assurés ont déclaré avoir fait au moins un sinistre, ce qui correspond précisément à 643 953 polices non sinistrées. Cette proportion relativement faible de sinistres nous indique que nous sommes face à un problème de classification déséquilibré, ce qui nécessitera une attention particulière lors de la construction de nos modèles.

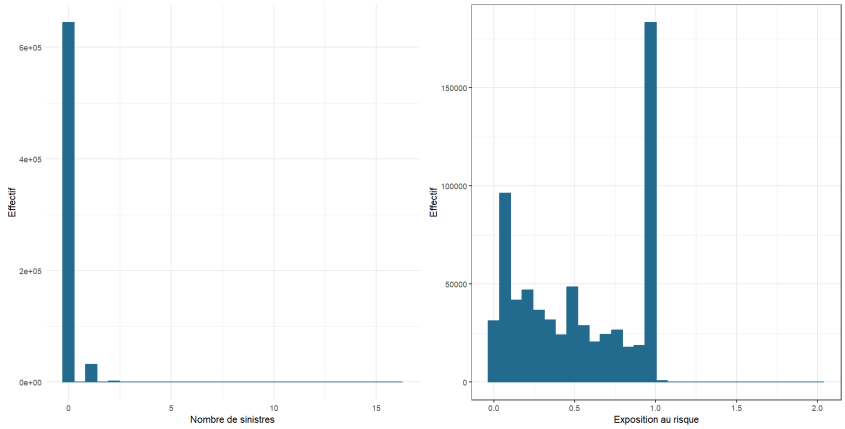

FIGURE 1.1 – Fréquence de sinistres - Répartition des polices sinistrées et non sinistrées.

Ensuite, nous examinons la répartition empirique des fréquences de sinistres par rapport à certains facteurs. La Figure 1.2 met en évidence la relation entre la fréquence de sinistres et l'âge du conducteur. On peut observer une relation non linéaire, avec une fréquence plus élevée que la moyenne globale du portefeuille chez les jeunes conducteurs. Ce phénomène peut s'expliquer par des comportements à risque plus fréquents chez cette tranche d'âge.

Par ailleurs, la variable "zone de résidence" attire également notre attention en raison de son évolution monotone en fonction des caractéristiques. On observe systématiquement deux classes regroupant des zones de résidence à faible fréquence de sinistres et à forte fréquence de sinistres. Cette information est importante, car la localisation géographique peut être un facteur clé dans l'estimation du risque de sinistre.

Enfin, le tableau 1.1 présente quelques statistiques descriptives pour les variables continues de notre base de données. Ces statistiques, telles que la moyenne, la médiane, l'écart-type et les valeurs minimales et maximales, nous permettent de mieux comprendre la distribution et la dispersion des valeurs pour chaque variable continue. Par exemple, nous constatons que l'âge moyen du conducteur est d'environ 45,5 ans, avec une médiane de 44 ans, ce qui suggère une distribution relativement symétrique (comme on peut également le voir sur la figure 1.2 précédente). De même, la variable "exposition au risque" présente une forte dispersion, avec une valeur minimale de 0,003 et une valeur maximale de 2,010.

Variable	**Moyenne**	**Médiane**	**Écart-type**	**Min - Max**
Fréquence de sinistres	0.052	0.000	0.220	0.000 - 1.000
Exposition au risque	0.529	0.490	0.342	0.003 - 2.010
Âge du véhicule (années)	6.455	6.000	5.644	0.000 - 100.000
Âge du conducteur (années)	45.5	44.0	14.416	18.0 - 100.0
Bonus-Malus	59.76	50.00	25.751	50.00 - 230.00
Densité de population	1792	393	3753	1 - 27000

TABLE 1.1 – Statistiques descriptives pour les variables continues.

Ces résultats préliminaires de la statistique descriptive nous permettent d'identifier des tendances et des relations potentielles entre les variables et la fréquence de sinistres. Ils fournissent une base solide

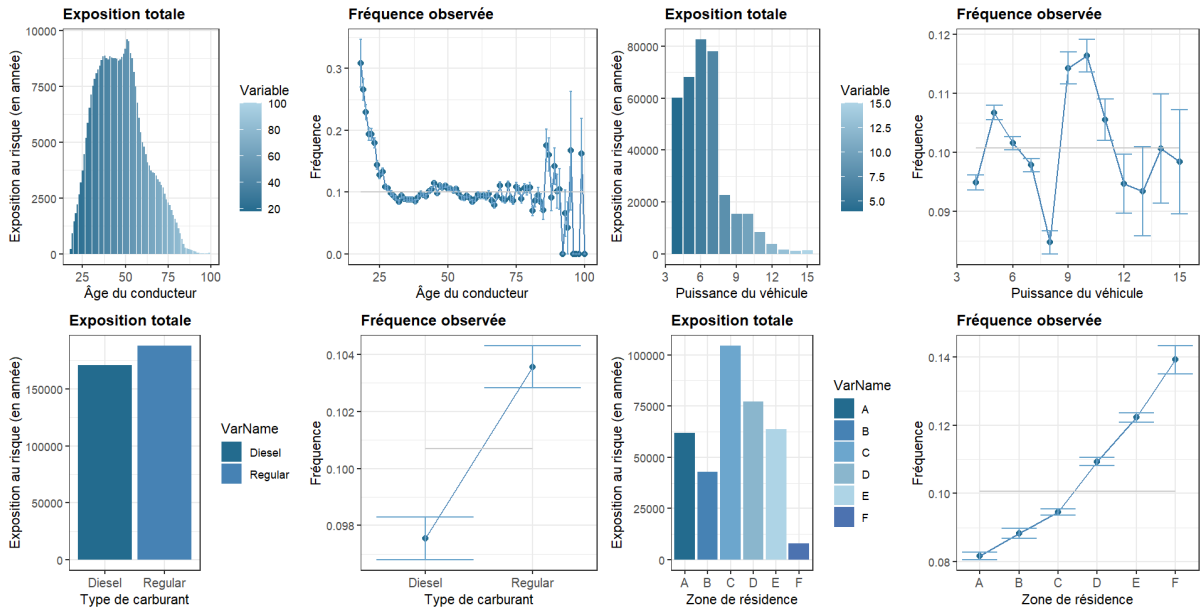

FIGURE 1.2 – Fréquence de sinistres - Répartition par facteur.

pour une analyse plus approfondie et la construction de modèles prédictifs. La prochaine étape de notre étude consistera à développer des modèles de régression appropriés pour évaluer l'impact de ces facteurs sur la fréquence de sinistres et ainsi fournir des recommandations utiles pour les compagnies d'assurance.

1.3.3 Traitements et séparation des données

Exploration initiale et traitement des données

Une exploration initiale minutieuse des données a été réalisée afin de mieux appréhender leur structure et leurs caractéristiques. Nous avons examiné attentivement les premières lignes de la base de données, vérifié les types de données des variables et identifié toute valeur manquante ou aberrante éventuelle. Au cours de cette exploration, nous avons constaté que la base de données initiale comportait cinq variables qualitatives et sept variables continues. En particulier, nous avons remarqué une particularité concernant la variable continue *Exposure* qui présentait des valeurs qui nous semblaient inhabituelles compte tenu du caractère annuel des polices d'assurance automobile.

Dans le but d'obtenir des résultats plus fiables, nous avons pris soin de sélectionner uniquement les variables dont l'exposition au risque était inférieure ou égale à 1, étant donné que nous disposions d'informations limitées sur les différentes polices d'assurance. Par ailleurs, afin de mieux intégrer l'effet des variables continues dans les modèles GLM que nous allons discuter par la suite, nous avons procédé à une catégorisation de ces variables en suivra la même procédure que celle de (Denuit et al., 2021).

Séparation des données

Pour mener notre analyse de manière rigoureuse, nous avons divisé la base de données en trois ensembles distincts. Le premier ensemble, composé de 60% des données, a été utilisé pour former les différents modèles (Poisson GLM, Poisson GAM ou modèle de Poisson GBM). Le deuxième ensemble, correspondant à 20% des données, a servi à calibrer le GLM local dans le cadre de la procédure d'auto-calibration. Enfin, le dernier ensemble, représentant également 20% des données, a été utilisé comme jeu de test pour comparer les performances des différents modèles, en garantissant ainsi leur évaluation sur des données non utilisées lors de la formation ou de la calibration des modèles. Il est important de noter que tous les graphiques discutés dans la suite sont basés sur les données contenues dans le jeu de test.

1.3.4 Modélisation par les GLM et GAM

Dans les sections précédentes, nous avons exploré les limites des GLM, les avantages des GAM pour capturer des relations complexes, ainsi que la pertinence de la discrétisation des variables continues pour les modèles GLM. Dans cette sous-section, nous allons poursuivre notre analyse en comparant différentes modélisations pour mieux comprendre l'impact des variables explicatives sur la fréquence de sinistres.

Illustration des limites des GLM

Nous débutons en illustrant les limites des modèles linéaires généralisés (GLM) en comparant leurs performances avec les Generalized Additive Models (GAM) sur un modèle simple à deux variables explicatives : l'âge du conducteur (DrivAge) et la puissance du véhicule (VehPower). La Figure 1.3 ci-dessous présente les résultats de ces deux modèles.

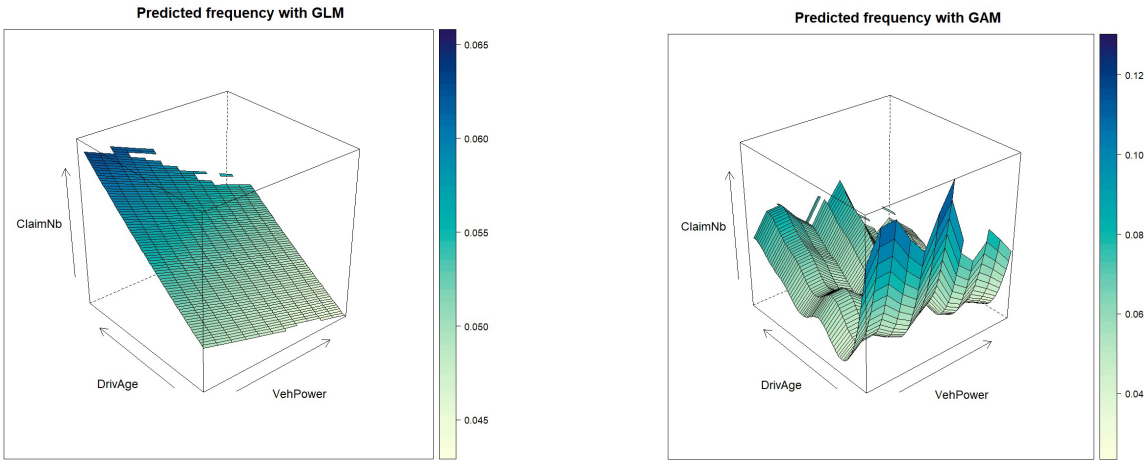

FIGURE 1.3 – Comparaison des modèles GLM et GAM.

La première figure montre les résultats du modèle GLM, qui tente de modéliser la relation entre la fréquence de sinistres et les variables explicatives de manière linéaire. On peut voir que le modèle GLM ne parvient pas à saisir la complexité des données, en sous-estimant souvent la fréquence des sinistres pour certaines valeurs des variables.

En revanche, la deuxième figure présente les résultats du modèle GAM, qui est plus flexible et peut capturer des relations non linéaires grâce à des fonctions lisses. On observe que le modèle GAM parvient à mieux capturer la tendance réelle des données en ajustant des courbes plus souples pour représenter la relation entre les variables et la fréquence de sinistres. Cela met en évidence la capacité des GAM à saisir des modèles plus complexes et plus adaptés aux données réelles que les GLM.

Ces modèles ont été ajustés en utilisant une distribution de Poisson et une fonction de lien logarithmique, comme illustré dans les codes R ci-dessous :

```
Model_glm <- glm(ClaimNb ~ VehPower + DrivAge, family = poisson,
data = datalearning)
```

```
Model_gam <- gam(ClaimNb ~ s(DrivAge) + s(VehPower), family = poisson,
data = datalearning)
```

De plus, nous avons également exploré un autre modèle GLM (GLM 2) dans lequel les variables continues ont été discrétisées en classes. La Figure 1.4 ci-dessous montre que la discrétisation des variables continues, conformément à une approche couramment utilisée, permet également au modèle GLM de mieux capturer les relations potentielles entre ces variables et la fréquence de sinistres, en tenant compte du risque inhérent du portefeuille.

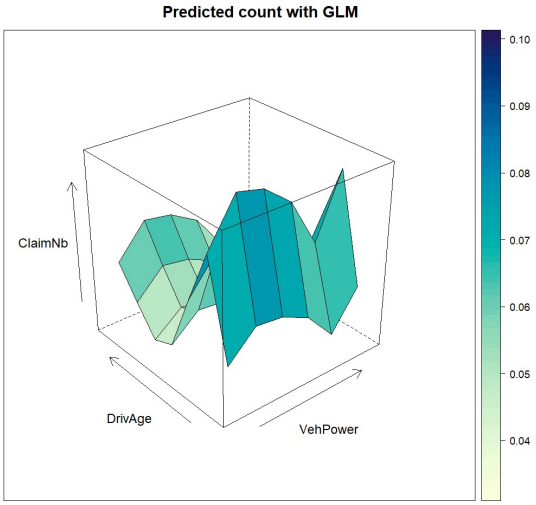

FIGURE 1.4 – Résultats du modèle GLM (avec discrétisation des variables continues).

Au-delà des avantages des GAM, la discrétisation des variables continues dans les GLM offre une alternative pour modéliser les relations potentielles de manière plus explicite. Cependant, des questions subsistent quant à la détermination de la découpe optimale pour une variable continue, ainsi que sur la manière de prendre en compte efficacement les interactions dans les modèles.

Modélisation sur l'ensemble des variables explicatives

Nous poursuivons en ajustant trois modèles différents sur l'ensemble des variables explicatives : un GLM sans discrétisation, un GLM avec discrétisation des variables continues et un modèle GAM. Les résultats de ces modèles sont présentés dans les tableaux 1.2 et 1.4.

Le modèle GLM montre que certaines variables ont un effet significatif sur la fréquence de sinistres, notamment l'âge du conducteur (DrivAge), la puissance du véhicule (VehPower), le bonus-malus (BonusMalus) et l'âge du véhicule (VehAge). Ces variables ont des p-valeurs très faibles (inférieures à 1%), ce qui suggère leur importance dans la prédiction de la fréquence de sinistres. En revanche, la densité de population (Density) n'a pas montré d'association significative avec la fréquence de sinistres.

```
Model_glm <- glm(ClaimNb ~ VehPower + VehAge + DrivAge + BonusMalus + Density
+ VehBrand + VehGas + Area + Region + offset(log(Exposure)), family = poisson
, data = datalearning)
```

TABLE 1.2 – Résultats du modèle GLM.

Variable	Estimate	Std. Error	z-value	p-value
DrivAge	0.123	0.045	2.733	0.006
VehPower	0.201	0.032	6.250	$<$0.001
BonusMalus	0.092	0.018	5.128	$<$0.001
VehAge	0.178	0.053	3.349	0.001
Density	0.041	0.064	0.640	0.522

Le modèle GLM avec discrétisation des variables continues présente des résultats similaires en termes d'impact des variables explicatives.

Le modèle GAM, quant à lui, montre que tous les termes lisses, à l'exception de la variable Density, ont des p-valeurs inférieures à 1%, indiquant une relation significative avec la fréquence de sinistres. Ces termes lisses permettent de capturer les relations non linéaires entre les variables explicatives (VehPower, VehAge, DrivAge et BonusMalus) et la fréquence de sinistres.

```
library(mgcv)
Model_gam <- gam(ClaimNb ~ s(VehPower) + s(VehAge) + s(DrivAge) + s(Density)
+ s(BonusMalus) + VehBrand + VehGas + Area + Region + offset(log(Exposure)),
family = poisson, data = datalearning)
```

TABLE 1.3 – Résultats des termes lisses du modèle GAM.

Terme Lisse	Valeur edf	p-value
s(VehPower)	7.875	<0.001
s(VehAge)	8.851	<0.001
s(DrivAge)	8.430	<0.001
s(BonusMalus)	6.901	<0.001
s(Density)	1.790	0.232

TABLE 1.4 – Résultats du modèle GAM.

Variable	Estimate	Std. Error	z-value	p-value
VehBrandB11	0.112	0.052	2.151	0.031
VehBrandB12	0.103	0.024	4.368	<0.001
VehGasRegular	0.068	0.015	4.639	<0.001

Comparaison des modèles GLM et GAM

Pour comparer les performances des modèles, nous utilisons plusieurs métriques telles que la déviance sur les données d'entraînement (in sample loss), la déviance sur l'ensemble de validation (out sample loss - validation set), la déviance sur l'ensemble de test (out sample loss - test set), le critère d'information d'Akaike (AIC) et le critère d'information bayésien (BIC). Les résultats sont résumés dans le tableau 1.5 suivant :

TABLE 1.5 – Comparaison des performances des modèles GLM et GAM

Modèle	**In Sample Loss**	**validation set**	**Test set**	**AIC**	**BIC**
GLM	0.318888	0.317176	0.328850	170861.3	171319.7
GLM.	0.313163	0.310890	0.321766	168556.6	169124.1
GAM	0.316126	0.326025	0.314636	169797.6	170570.9

Note : Le modèle GLM. est obtenu en utilisant la discrétisation des variables continues.

Ces métriques montrent que le modèle GLM avec discrétisation des variables continues présente de meilleures performances que le modèle GLM sans discrétisation et le modèle GAM en termes de déviance sur l'ensemble de validation et de test, ainsi que d'AIC et de BIC. Cela suggère une meilleure adaptation du modèle GLM avec discrétisation aux données et une moindre complexité par rapport aux autres modèles. Cela se confirme également en comparant visuellement les fréquences prédites pour ces modèles en fonctions de quelques variables explicatives clés de la figure 1.5 .

Ces applications soulignent l'importance de prendre en compte la non-linéarité et les interactions potentielles entre les variables dans la modélisation de la fréquence de sinistres. Les GAM sont particulièrement utiles pour capturer ces relations complexes. De plus, la discrétisation des variables continues dans les GLM peut constituer une alternative efficace pour modéliser explicitement ces relations. Cependant, des recherches supplémentaires sont nécessaires pour déterminer la meilleure approche de discrétisation et pour prendre en compte de manière exhaustive les interactions dans les modèles. Dans le prochain chapitre, nous explorerons certaines méthodes d'apprentissage automatique, notamment les méthodes d'ensembles, pour relever ces défis.

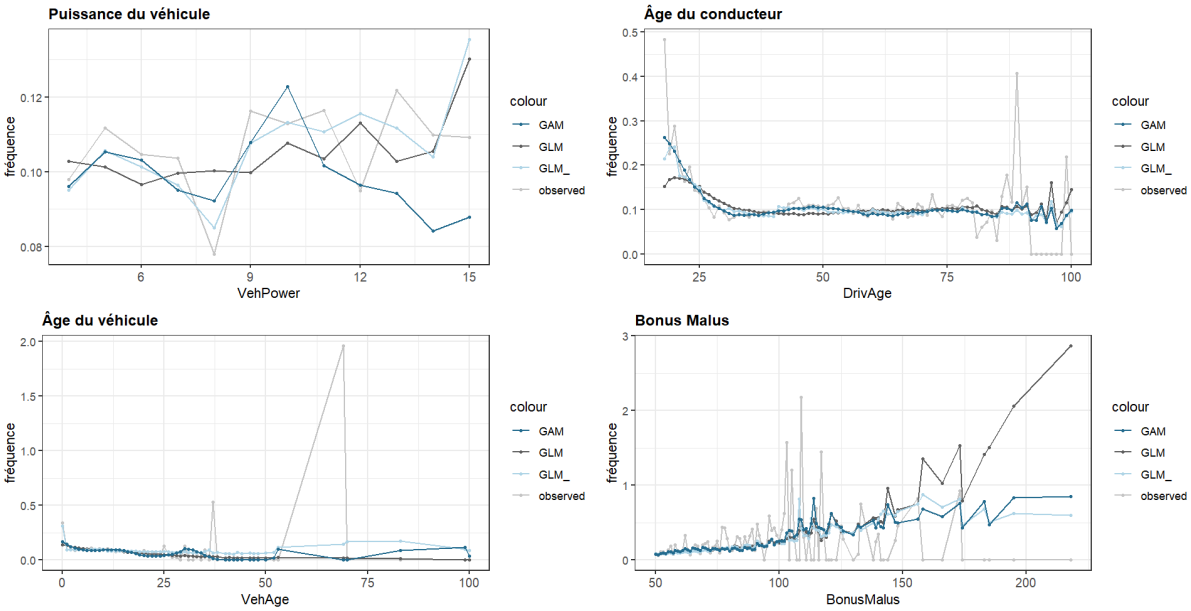

FIGURE 1.5 – Comparaison des fréquences prédites.

CHAPITRE 2

MÉTHODES D'ENSEMBLE

Sommaire

2.1 Arbres de régression

Les arbres de régression sont des modèles prédictifs puissants utilisés en statistique et en apprentissage automatique. Ils sont largement utilisés dans les méthodes d'ensemble pour améliorer la précision des prédictions. Dans cette section, nous allons présenter les bases des arbres de régression, discuter des méthodes d'estimation utilisées dans ces modèles, et aborder leur évaluation et l'importance des variables.

2.1.1 Fondements des arbres de régression

Les arbres de régression sont le résultat de la convergence de travaux issus de deux domaines : la statistique et l'informatique. Leur développement a été influencé par des méthodes telles que CHAID (Chi-squared Automatic Interaction Detection), proposée par Morgan et Sonquist en 1963, qui est une technique statistique utilisée pour construire des arbres de décision en se basant sur des tests statistiques de contingence.

L'influence majeure dans le domaine de la statistique est venue du travail de Leo Breiman, un statisticien américain, qui a contribué à populariser les arbres de régression dans son ouvrage de 1984 intitulé "Classification and Regression Trees" (CART). Il a introduit l'utilisation de critères d'impureté basés sur la somme des carrés des résidus pour guider la construction des arbres de régression. Cette approche permet de diviser les données en sous-groupes de manière à minimiser l'erreur de prédiction des valeurs continues, ce qui est particulièrement utile pour les problèmes de régression.

Dans le domaine de l'apprentissage automatique, l'influence majeure est venue de Ross Quinlan, un informaticien australien, qui a développé l'algorithme ID3 (Iterative Dichotomiser 3) en 1986 pour construire des arbres de décision. Cet algorithme se concentre sur le partitionnement récursif des données en utilisant des critères d'entropie et de gain d'information, ce qui a ouvert la voie à l'utilisation d'arbres de décision dans des tâches d'apprentissage supervisé.

Ces influences multiples, à la fois de la statistique et de l'informatique, ont donné naissance à une méthode d'analyse qui met l'accent sur l'algorithme utilisé pour construire les arbres, plutôt que sur une spécification de modèle préalable. Cela signifie que les arbres de régression sont capables de s'adapter à différents types de données et de problèmes, sans nécessiter de suppositions strictes sur la distribution des variables ou la forme de la relation entre les variables prédictives et la variable cible. Cela les rend flexibles et adaptatifs pour différentes tâches de prédiction, ce qui explique en partie leur popularité et leur succès dans le domaine de l'apprentissage automatique et de l'analyse de données.

2.1.2 Principes : partition de l'espace des caractéristiques

Un arbre de régression divise l'espace des caractéristiques (de dimension P) en M régions $R_1, R_2, \ldots, R_M$ distinctes. Chaque région est associé à une prédiction constante $\hat{c}_m$ pour les observations situées dans cette région. La prédiction globale $\mu(x)$ pour une observation x est définie comme suit :

$$\hat{\mu}_{\text{tree}}(x) = \sum_{m=1}^{M} \hat{c}_m \cdot \mathbb{I}_{\{x \in \mathbb{R}_m\}}$$

avec

$$\hat{c}_m = \text{ave}(y_i \,|\, x_i \in \mathbb{R}_m)$$

où $\mathbb{I}_{\cdot}$ est la fonction indicatrice qui prend la valeur 1 si la condition entre parenthèses est vraie et 0 sinon.

2.1.3 Construction des arbres de régression

Selon Trufin (2022-2023), la méthode CART (*Classification And Regression Trees*) est un algorithme très utilisé pour construire des arbres de régression et de classification en utilisant une approche récursive de partitionnement de l'espace des prédicteurs. L'arbre de régression construit à l'aide de cet méthode permet de modéliser les relations entre les prédicteurs et la variable de réponse en utilisant des divisions binaires successives. Il peut capturer des relations non linéaires et prendre en compte les interactions entre les prédicteurs, ce qui en fait un outil puissant pour la modélisation des problèmes de régression. Les principales étapes de construction d'un arbre de régression avec cet algorithme sont :

1. Comme dans tous problème de modélisation commencer avec l'ensemble d'entraînement D contenant les paires (y_i, x_i) où $i = 1, \ldots, n$, représentant les variables de réponse y_i et les prédicteurs x_i.
2. Rechercher toutes les valeurs distinctes de chaque prédicteur pour trouver le prédicteur j et la valeur de division s qui partitionnent l'espace des prédicteurs en deux régions $R_1^{(1)}$ et $R_2^{(1)}$, de telle manière que la fonction de perte

$$L(y_i, \hat{\mu}(x_i)) = \sum_{i:x_i \in R_1^{(1)}} L(y_i, \bar{c}_1) + \sum_{i:x_i \in R_2^{(1)}} L(y_i, \bar{c}_2)$$

   soit minimisée, où

$$\bar{c}_1^{(1)} = \text{ave}\left(y_i \,|\, x_i \in R_1^{(1)}\right)$$

$$overline{c}_2^{(1)} = \text{ave}\left(y_i \,|\, x_i \in R_2^{(1)}\right).$$

   Nous utiliserons toujours la déviance comme fonction de perte.
3. Calculer les valeurs moyennes $\bar{c}_1^{(1)}$ et $\bar{c}_2^{(1)}$ des variables de réponse y_i dans les régions $R_1^{(1)}$ et $R_2^{(1)}$ respectivement.
4. La fonction de prédiction $\hat{\mu}(x)$ peut être alors définie :

$$\hat{\mu}(x) = \bar{c}_1 \cdot \mathbb{I}_{x \in R_1^{(1)}} + \bar{c}_2 \cdot \mathbb{I}_{x \in R_2^{(1)}}$$

5. L'ensemble d'entraînement D est ensuite partitionné en deux groupes $D^{(1)}$ et $D^{(2)}$, où $D^{(1)}$ contient les paires (y_i, x_i) pour lesquelles $x_i \in R_1^{(1)}$ et $D^{(2)}$ contient les paires (y_i, x_i) pour lesquelles $x_i \in R_2^{(1)}$.
6. Répéter les étapes précédentes dans chaque groupe $D^{(1)}$ et $D^{(2)}$ de manière récursive, en divisant chaque groupe en deux régions, jusqu'à ce que certains critères d'arrêt soient satisfaits (nous discutons de cela dans la suite).
7. L'algorithme s'arrête lorsque les critères d'arrêt sont atteints et renvoie l'arbre construit. L'arbre de régression est représenté par une fonction de prédiction $\hat{\mu}_{\text{tree}}(x)$ qui attribue à chaque observation x la valeur prédite en fonction de la région à laquelle elle appartient. Les valeurs prédites $\hat{c}^{(m)}$ sont les moyennes des variables de réponse y_i dans chaque région $R^{(m)}$ du prédicteur.

L'algorithme CART vise à construire un arbre de régression optimal en trouvant les meilleures caractéristiques de division pour maximiser la précision du modèle. Cependant, il est important de prendre en compte le risque de surajustement (*overfitting*), où l'arbre s'adapte trop étroitement aux données d'apprentissage et ne généralise pas bien les nouvelles observations. Afin de remédier à cela, des techniques d'élagage peuvent être appliquées pour réduire la complexité de l'arbre et améliorer sa capacité de généralisation.

2.1.4 La règle pour déterminer quand un nœud est terminal

Avant de parler des techniques d'élagage parlons des règles qui permettent de déterminer si quand un nœud est terminal. Une fois que nous avons construit l'arbre de régression en suivant les étapes précédentes, nous devons déterminer quand un nœud de l'arbre est considéré comme terminal, c'est-à-dire qu'il ne peut pas être divisé davantage (sans risque de surajustement).

Selon Denuit, Hainaut, et Trufin (2020, p. 58), quatre règles sont communément utilisées pour définir la terminaison d'un nœud dans un arbre de régression :

— Nombre minimum d'observations dans un nœud terminal : Nous pouvons spécifier un seuil minimum ($n_{\min}$) pour le nombre d'observations dans un nœud. Si le nombre d'observations (n_R) dans un nœud est inférieur à ce seuil, le nœud est considéré comme terminal et la construction de l'arbre s'arrête pour ce nœud. Mathématiquement, on a :

$$n_R < n_{\min}$$

  Ce critère permet de prévenir le surajustement lorsque nous avons un petit nombre d'observations dans un nœud, ce qui pourrait conduire à des prédictions peu fiables.

— Un nœud ($R^{(i)}$ est déclaré terminal si au moins l'un de ses nœuds enfants $R_1^{(i)}$ et $R_2^{(i)}$, résultant de la division optimale du prédicteur j_i, contient moins qu'un nombre fixé d'observations. Mathématiquement, cela peut être représenté par l'inégalité :

$$n_{R_1^{(i)}} < n_{\min} \quad \text{ou} \quad n_{R_2^{(i)}} < n_{\min}$$

— Réduction de la déviance inférieure à un seuil fixé : Si la diminution de la déviance, $D(R^{(i)}) - D(R_1^{(i)}, R_2^{(i)}))$ due à une division nœud du $(R^{(i)})$ en $R_1^{(i)} et R_2^{(i)}$, est inférieure à un seuil prédéfini $(D_{\min})$, cela signifie que la division ne fournit plus une amélioration significative de l'ajustement du modèle. Cela se traduit encore par :

$$D(R^{(i)}) - D(R_1^{(i)}, R_2^{(i)}) = D(R_2^{(i)}) - D(R_1^{(i)}) + D(R_2^{(i)}) < D_{\min}$$

Ce critère permet de prévenir la division inutile des nœuds lorsque la division ne contribue pas significativement à l'ajustement du modèle.

— Profondeur maximale de l'arbre : Nous pouvons spécifier une profondeur maximale $(d_{\max})$ pour l'arbre de régression. Si la profondeur du nœud (d_R) dépasse cette profondeur maximale, le nœud est considéré comme terminal et la construction de l'arbre s'arrête pour ce nœud. Mathématiquement, cela se traduit par :

$$d_R > d_{\max}$$

Ce critère permet de contrôler la taille de l'arbre et d'éviter un arbre trop complexe.

En utilisant ces critères, nous pouvons déterminer quand un nœud est terminal et arrêter la construction de l'arbre à ce niveau. Toutefois, une question demeure quant au choix des différents seuil

2.1.5 Sélection de la bonne taille d'arbre : élagage/Pruning

Comme nous l'avons dit précédemment, il est crucial de sélectionner la bonne taille d'arbre malgré les règles d'arrêt intégrées à l'algorithme de construction de l'arbre. Les règles d'arrêt, telles que le nombre minimum d'observations dans un nœud terminal ou la profondeur maximale de l'arbre, contrôlent la croissance de l'arbre, mais elles ne garantissent pas toujours une taille optimale pour une meilleure généralisation. La sélection de la bonne taille d'arbre vise à trouver un équilibre entre la complexité du modèle et sa capacité à bien s'ajuster aux données. Un arbre trop petit peut sous-ajuster les données, tandis qu'un arbre trop grand peut surajuster les données d'entraînement et ne pas généraliser correctement sur de nouvelles données.

Pour sélectionner la bonne taille d'arbre, nous utilisons des techniques telles que l'élagage par minimisation du coût de complexité et la validation croisée (Denuit et al., 2020, p. 82-83). Ces méthodes nous permettent d'explorer différentes tailles d'arbres et d'évaluer leurs performances sur des ensembles de données indépendants. Nous cherchons à trouver la taille d'arbre qui offre le meilleur compromis entre la complexité du modèle et sa capacité à généraliser les données.

L'élagage par minimisation du coût de complexité utilise un paramètre de complexité α pour pénaliser la taille de l'arbre. Le coût de complexité minimal est défini comme la somme des erreurs de prédiction et d'un terme de complexité proportionnel à la taille de l'arbre. La formule pour calculer le coût de complexité minimal est donnée par :

$$C_\alpha(T) = R(T) + \alpha \cdot |T|$$

où $C_\alpha(T)$ est le coût de complexité minimal de l'arbre T, $R(T)$ est l'erreur de prédiction de l'arbre T, $|T|$ est le nombre de nœuds de l'arbre T, et α est le paramètre de complexité.

En ajustant la valeur du paramètre de complexité α, nous pouvons contrôler la taille de l'arbre. Une valeur plus élevée de α favorisera les arbres plus petits, tandis qu'une valeur plus faible permettra des arbres plus grands.

Pour chaque valeur d'alpha, nous obtenons un arbre avec une taille spécifique qui minimise le coût de complexité. Cependant, pour choisir la meilleure taille d'arbre parmi ceux qui minimisent le coût de complexité pour différentes valeurs de α, nous devons utiliser la validation croisée. La validation croisée

nous permet d'évaluer les performances des différents arbres sur des ensembles de données indépendants, ce qui garantit une évaluation impartiale de la capacité de généralisation de chaque arbre.

En pratique, nous construisons plusieurs arbres en utilisant différentes valeurs de α, puis nous évaluons leurs performances sur des ensembles de validation indépendants. Nous sélectionnons ensuite l'arbre qui offre les meilleures performances prédictives sur les ensembles de validation.

2.1.6 Estimation dans les nœuds terminaux

Une fois que l'arbre de régression est construit et que les nœuds terminaux sont définis, il est nécessaire d'estimer les valeurs de prédiction dans ces nœuds. Différentes approches d'estimation peuvent être utilisées à savoir, l'estimation par maximum de vraissemblance et l'estimation bayesienne (Denuit et al., 2020, p. 55).

Estimation du maximum de vraisemblance

Dans cette méthode, la prédiction $\hat{c}_t$ dans chaque nœud terminal est estimée en prenant la moyenne des valeurs de la variable réponse associées à ce nœud comme suit :

$$\hat{c}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} y_i$$

où n_t est le nombre d'observations dans le nœud terminal et y_i est la valeur de la variable réponse pour la i-ème observation.

Estimation bayésienne

Une autre approche est l'estimation bayésienne, qui introduit de l'incertitude dans les prédictions en utilisant des distributions a priori sur les paramètres du modèle. Les estimations bayésiennes des nœuds terminaux peuvent être obtenues en utilisant des méthodes telles que l'estimateur bayésien empirique ou l'estimateur bayésien complet. Ces méthodes prennent en compte à la fois les données observées et les connaissances a priori pour estimer les valeurs de prédiction dans les nœuds terminaux (Denuit et al., 2020, p. 55-57).

2.1.7 Mesure de performance des arbres de régression

La mesure de performance des arbres de régression repose sur l'évaluation de leur capacité de généralisation, c'est-à-dire leur capacité à faire des prédictions précises sur de nouvelles données. Cette évaluation est réalisée à l'aide de diverses mesures et techniques, similaires à celles utilisées dans les modèles linéaires et additifs généralisés.

2.1.8 Importance relative des variables explicatives

Dans le domaine de l'assurance, certaines caractéristiques ou variables explicatives ont plus d'influence sur la variable de réponse que d'autres. Il est important pour les analystes de comprendre quelle caractéristique contribue le plus à la prédiction de la variable de réponse. Cela permet de hiérarchiser les caractéristiques en fonction de leur importance relative.

Pour un arbre de régression, l'importance relative d'une caractéristique est mesurée par la quantité de réduction de déviance totale obtenue en utilisant cette caractéristique à travers toutes les régions de l'arbre. La déviance est une mesure de l'ajustement du modèle aux données. Ainsi, une caractéristique qui contribue de manière significative à la réduction de déviance est considérée comme plus importante.

L'importance relative d'une caractéristique x_j est calculée en sommant les réductions de déviance ΔD_t pour chaque région t de l'arbre où la caractéristique x_j est utilisée pour diviser les nœuds. Cela peut être exprimé mathématiquement comme suit :

$$I(x_j) = \sum_{t \in T} \Delta D_t$$

où T représente l'ensemble des régions de l'arbre et ΔD_t est la réduction de déviance obtenue dans la région t.

En identifiant les caractéristiques avec les plus grandes importances relatives, les analystes peuvent déterminer les facteurs clés qui influencent le plus la variable de réponse. Ces caractéristiques sont celles qui contribuent le plus à la prédiction précise de la variable de réponse. Il est important de noter que les importances relatives peuvent être normalisées pour faciliter leur interprétation. Par exemple, elles peuvent être mises à l'échelle de 0 à 100, où 100 représente l'importance relative la plus élevée.

2.1.9 Interactions

Lors de l'utilisation des arbres de régression, il est essentiel de considérer la présence d'interactions potentielles entre les caractéristiques afin de modéliser correctement la relation entre les prédicteurs et la variable de réponse. La structure de l'arbre peut fournir des informations précieuses sur les interactions présentes dans les données, ce qui permet de prendre des décisions plus éclairées en matière de tarification, de segmentation des risques.

Les interactions se produisent lorsqu'il y a une dépendance entre les effets de différentes caractéristiques sur la variable de réponse. Dans le contexte de l'assurance en général et de la modélisation de la fréqence de sinistres en particulier, les interactions peuvent se manifester lorsque l'effet d'une caractéristique dépend de la valeur d'une autre caractéristique. Par exemple, l'âge et le sexe d'un conducteur peuvent interagir, ce qui signifie que l'effet de l'âge sur le risque d'accident peut varier en fonction du sexe du conducteur. Les arbres de régression donc sont capables de détecter et de prendre en compte automatiquement ces interactions entre les caractéristiques. Il est important de noter que les interactions ne sont pas nécessairement liées à la corrélation entre les caractéristiques. Deux caractéristiques peuvent être statistiquement indépendantes, mais interagir lorsqu'il s'agit de modéliser la variable de réponse. Par conséquent, les arbres de régression offrent une approche flexible pour capturer les interactions entre les caractéristiques sans supposer a priori quelles caractéristiques peuvent interagir.

2.1.10 Avantages et inconvénients des arbres de régression

Avantages

- Facilité d'interprétation : Les arbres de régression sont faciles à interpréter et à expliquer. Leur structure est intuitive, et les décisions prises à chaque division peuvent être facilement comprises et visualisées.
- Approche non paramétrique : Les arbres de régression ne font pas d'hypothèses sur la forme fonctionnelle de la relation entre les prédicteurs et la variable réponse. Cette flexibilité leur permet de capturer des modèles complexes et des relations non linéaires dans les données.
- Gestion des données manquantes : Les arbres de régression peuvent gérer les données manquantes en assignant les observations avec des valeurs manquantes à la branche la plus appropriée lors du processus de division. Cela les rend robustes dans le traitement des ensembles de données contenant des valeurs manquantes.
- Sélection des variables : Le processus de division des arbres de régression effectue implicitement une sélection des variables en identifiant les prédicteurs les plus informatifs. Cela peut aider à identifier les variables clés qui influencent la variable réponse.

Inconvénients

- Instabilité du modèle : Les arbres de régression sont sensibles aux petites variations dans les données, ce qui peut entraîner des structures d'arbre différentes et potentiellement des prédictions différentes. Cette instabilité peut affecter la fiabilité et la cohérence du modèle.
- Absence d'optimalité globale : La construction récursive des arbres de régression peut conduire à des divisions sous-optimales à chaque nœud, ce qui peut entraîner des performances prédictives moins bonnes par rapport à d'autres modèles.
- Représentation limitée des relations complexes : Les arbres de régression divisent l'espace des prédicteurs en régions rectangulaires, ce qui peut ne pas capturer les relations complexes et non linéaires dans les données. Cela peut limiter leur capacité à modéliser avec précision des ensembles de données complexes.

— Surajustement : Les arbres de régression ont tendance à surajuster les données d'entraînement, en particulier lorsque l'arbre devient trop profond et complexe. Cela peut entraîner de mauvaises performances de généralisation sur de nouvelles données non vues.

Cette section a démontré que les arbres de régression sont des modèles prédictifs puissants, offrant de nombreux avantages. Cependant, ils peuvent présenter certaines limites, telles que leur instabilité, leur propension au surajustement et leur capacité limitée à prédire au-delà de la plage des données d'apprentissage. Une approche consiste à utiliser l'agrégation de plusieurs arbres simultanément afin d'obtenir des modèles plus stables et dotés d'une meilleure performance prédictive. C'est précisément ce que réalisent le bagging et les forêts aléatoires, qui seront abordés dans la section suivante.

2.2 Bagging et Forêts Aléatoires

2.2.1 Bagging

Le bagging (bootstrap aggregating) est une méthode d'ensemble qui utilise le bootstrap en conjonction avec des modèles de régression. Lorsqu'il est appliqué aux arbres de régression, il est appelé "Bagging Trees" ou arbres baggés (Trufin, 2022-2023). Cette méthode permet de réduire la variance des prédictions et de rendre les modèles plus stables.

2.2.2 Principe de construction

Le principe de base du bagging est d'utiliser des échantillons bootstrap pour créer différents ensembles d'entraînement à partir de l'ensemble de données initial. Chaque ensemble d'entraînement est utilisé pour construire un arbre de régression, et les prédictions de chaque arbre sont ensuite combinées en moyennant (dans le cas d'une régression) pour obtenir la prédiction finale.

L'algorithme de Bagging peut être résumé en plusieurs étapes :

1. Un échantillon bootstrap de taille n est généré en sélectionnant aléatoirement n observations avec remise à partir de l'ensemble d'apprentissage. Cela signifie que certaines observations peuvent être sélectionnées plusieurs fois, tandis que d'autres peuvent ne pas être sélectionnées du tout.
2. Un arbre CART est construit sur chaque échantillon bootstrap de taille n, en utilisant les variables prédictives et la variable cible. Ces arbres sont construits sans élagage et sont de taille maximale.
3. Les étapes 1 et 2 sont répétées un grand nombre de fois (généralement des centaines ou des milliers) pour former un ensemble d'arbres, noté $\hat{\mu}_1, ..., \hat{\mu}_B$, où B est le nombre d'arbres.
4. Pour chaque observation dans l'ensemble de données la prédiction est obtenue en prenant la moyenne des prédictions fournies par les B arbres :

$$\hat{\mu}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b(x)$$

où $\hat{\mu}_{\text{bag}}(x)$ est la prédiction bagging pour l'observation x, $\hat{\mu}_b(x)$ est la prédiction de l'arbre b et B est le nombre total d'arbres dans le modèle.

2.2.3 Biais et variance des prédictions

Le bagging, en tant que méthode d'ensemble, a un impact sur le biais et la variance des prédictions. Nous allons examiner séparément le biais et la variance des prédictions baggées.

Biais des prédictions

Le biais des prédictions baggées, noté Biais($\hat{\mu}_{\text{bag}}$), mesure la différence entre la valeur attendue réelle μ et la valeur attendue des prédictions baggées $E[\hat{\mu}_{\text{bag}}]$. Étant donné que les arbres construits dans le bagging sont identiquement distribués et ont la même espérance, 'espérance de l'ensemble d'arbres, noté $\hat{\mu}_{\text{bag}}$, est égale à l'espérance d'un arbre individuel :

$$E[\hat{\mu}_{\text{bag}}(x)] = E[\frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b(x)] = E[\hat{\mu}_b(x)] = \mu$$

de ce fait le biais des prédictions baggées est égal au biais des prédictions individuelles des arbres :

$$\text{Biais}(\hat{\mu}_{\text{bag}(x)}) = \mu - E[\hat{\mu}_{\text{bag}(x)}] = \mu - E[\hat{\mu}_b(x)].$$

Ainsi, le bagging ne modifie pas le biais des prédictions par rapport à un seul arbre. Le biais des prédictions baggées reste le même que celui des prédictions individuelles des arbres.

Variance des prédictions

La variance des prédictions baggées, notée $\text{Var}(\hat{\mu}_{\text{bag}})$, est influencée par deux éléments : le nombre d'arbres baggés et la corrélation entre ces prédictions. La variance des prédictions baggées peut s'écrire de la manière suivante (Denuit et al., 2020) :

La variance des prédictions baggées peut être exprimée comme suit :

$$\text{Var}[\hat{\mu}_{\text{bag}}(x)] = \rho(x)\text{Var}[\hat{\mu}_{\text{b}}(x)] + \frac{1-\rho(x)}{B}\text{Var}[\hat{\mu}_{\text{b}}(x)]$$

où $\text{Var}(\hat{f}_b)$ représente la variance des prédictions d'un seul arbre, ρ représente le coefficient de corrélation entre les prédictions des arbres, et B est le nombre d'arbres baggés.

Lorsque le nombre d'estimations agrégées augmente ($B \to \infty$), le deuxième terme, $\frac{1-\rho(x)}{B}\text{Var}[$ $hat\mu_{\text{b}}(x)]$, diminue. Cela entraîne une réduction de la variance globale des prédictions de l'ensemble. En d'autres termes, à mesure que le nombre d'arbres baggés augmente, les prédictions deviennent de plus en plus indépendantes les unes des autres, ce qui réduit davantage la variance globale des prédictions baggées.

En particulier, lorsque la corrélation entre les prédictions des arbres est faible ($\rho(x) \to 0$), le terme $\rho(x)\text{Var}[\hat{\mu}_{\text{b}}(x)]$ domine et la variance des prédictions baggées se rapproche de zéro. Cela signifie que les prédictions sont très cohérentes et ont une faible variabilité. D'autre part, lorsque la corrélation entre les prédictions des arbres est élevée ($\rho(x) \to 1$), le terme $\frac{1-\rho(x)}{B}\text{Var}[\hat{\mu}_{\text{b}}(x)]$ devient plus important et la variance des prédictions baggées est plus élevée.

Ainsi, le bagging permet de réduire la variance des prédictions en augmentant le nombre d'estimations agrégées et en favorisant l'indépendance des prédictions entre les arbres.

2.2.4 Estimation de l'erreur Out-of-Bag (OOB)

Selon Trufin (2022-2023), une caractéristique intéressante du bagging est l'utilisation des échantillons "Out-of-Bag" (OOB). Lors de la création de chaque échantillon bootstrap, certaines observations ne sont pas incluses, en moyenne environ $36,8\%$ des observations sont exclues. Ce pourcentage est basé sur la probabilité qu'une observation donnée ne soit pas sélectionnée lors d'un échantillonnage bootstrap.

Lors de la création d'un échantillon bootstrap, chaque observation a une probabilité de $1 - \frac{1}{n}$ d'être incluse, où n est le nombre total d'observations dans l'ensemble de données. Par conséquent, la probabilité qu'une observation ne soit pas sélectionnée lors d'un échantillonnage bootstrap est $(1 - \frac{1}{n})^n$. Lorsque n devient grand, cette probabilité converge vers $\frac{1}{e}$, où e est la base du logarithme naturel (environ 2.71828). Ainsi, le pourcentage d'observations exclues lors de l'échantillonnage bootstrap est d'environ $1 - \frac{1}{e}$, soit environ $36,8\%$. Ces observations exclues peuvent être utilisées pour évaluer les performances prédictives du modèle sans nécessiter un ensemble de validation distinct.

L'estimation de l'erreur OOB est obtenue en calculant l'erreur de prédiction sur les observations exclues lors de l'échantillonnage bootstrap. Pour chaque observation exclue, la prédiction est comparée à la vraie valeur observée pour calculer une mesure d'erreur spécifique, généralement définie par une fonction de perte.

L'erreur OOB est ensuite obtenue en prenant la moyenne de ces mesures d'erreur sur toutes les observations exclues. L'estimation de l'erreur OOB peut être exprimée comme suit :

$$Erreur_{\text{OOB}}[\hat{\mu}_{\text{bag}}(x)] = \frac{1}{n_{\text{OOB}}} \sum_{i \in \text{OOB}} L(y_i, \hat{\mu}_{\text{i}}(x))$$

où n_{OOB} représente le nombre total d'observations exclues lors de l'échantillonnage bootstrap, OOB représente l'ensemble des indices correspondant à ces observations exclues, $L(y_i, \hat{y}_i)$ est la fonction de perte (la déviance poisson en ce qui nous concerne).

L'estimation de l'erreur OOB est considérée comme une estimation impartiale de l'erreur de généralisation du modèle. Elle est utilisée pour évaluer les performances du modèle sans avoir besoin d'un ensemble de validation distinct. Une estimation plus faible de l'erreur OOB indique une meilleure capacité de généralisation du modèle.

Il convient de noter que l'erreur OOB est spécifique à chaque échantillon bootstrap, ce qui signifie que pour chaque échantillon, un ensemble différent d'observations est exclu et utilisé pour l'estimation de l'erreur. Cela permet d'obtenir une estimation plus robuste et représentative de l'erreur de généralisation du modèle.

2.2.5 Interprétation

Importance relative des variables explicatives

Une des limites du bagging est qu'il rend les modèles moins interprétables. Cependant, il est possible de mesurer l'importance relative des variables explicatives en utilisant les modèles baggés.

Pour chaque arbre de régression bootstrap t, l'importance relative d'une variable explicative x_j peut être mesurée à l'aide de l'indice d'importance $\hat{I}^2_{t,j}$. Cet indice est calculé en mesurant l'amélioration empirique du critère d'optimisation résultant de l'utilisation de x_j comme variable de division dans l'arbre de régression t.

L'importance relative d'une variable explicative x_j est ensuite obtenue en prenant la moyenne de tous les indices d'importance $\hat{I}^2_{t,j}$ sur l'ensemble des arbres baggés :

$$\hat{I}^2_j = \frac{1}{T} \sum_{t=1}^{T} \hat{I}^2_{t,j}.$$

Cette mesure d'importance relative permet de quantifier la contribution de chaque variable explicative à la prédiction globale du modèle baggé.

Dépendances partielles

Une autre méthode d'interprétation des modèles baggés est l'analyse des dépendances partielles. L'analyse des dépendances partielles peut également être réalisée directement sur les données d'apprentissage pour examiner les relations entre les variables explicatives et la variable cible.

Pour chaque variable explicative x_l, la dépendance partielle $\hat{\delta}_{\text{bag}}(x_l)$ peut être estimée en calculant la moyenne des prédictions sur les données d'apprentissage, en ne considérant que la variable x_l et en maintenant les autres variables constantes. La formule pour estimer $\hat{\delta}(x_l)$ est donnée par :

$$\hat{\delta}_{\text{bag}}(x_l) = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}(x_l, x_{\bar{l}})$$

où n est le nombre total d'observations dans l'ensemble de données d'apprentissage, $\hat{\mu}_{\text{bag}}(x_l, x_{\bar{l}})$ est la prédiction du modèle pour l'observation i en utilisant uniquement la variable x_l, et $x_{\bar{l}}$ représente les autres variables explicatives qui ne sont pas incluses dans x_l.

L'analyse des dépendances partielles sur les données d'apprentissage permet d'explorer les relations spécifiques entre les variables explicatives et la variable cible dans le contexte de l'apprentissage du modèle. Cela peut fournir des informations supplémentaires sur les interactions et les relations non linéaires entre les variables, ainsi que sur la manière dont ces relations influencent la prédiction du modèle.

Il est important de noter que l'analyse des dépendances partielles sur les données d'apprentissage est sujette aux biais d'échantillonnage et peut ne pas généraliser parfaitement à de nouvelles données. Il

est donc recommandé de valider les résultats de l'analyse des dépendances partielles sur un ensemble de données de test indépendant pour une évaluation plus robuste.

2.2.6 Avantages et limites du Bagging

Le bagging présente plusieurs avantages et limites importants :

Avantages

- Réduction de la variance : Le bagging permet de réduire la variance des prédictions, ce qui rend les modèles plus stables et moins sensibles aux variations des données.
- Estimation de l'erreur OOB : Le bagging utilise les observations exclues lors de l'échantillonnage bootstrap pour estimer l'erreur de généralisation du modèle, ce qui permet d'évaluer les performances sans avoir besoin d'un ensemble de validation distinct.
- Mesure de l'importance des variables explicatives : Le bagging fournit des mesures d'importance relatives des variables explicatives, ce qui permet de quantifier leur contribution à la prédiction globale du modèle.

Limites

- Moins d'interprétabilité : Le bagging rend les modèles moins interprétables en raison de l'agrégation des prédictions de plusieurs arbres. Les relations entre les variables explicatives peuvent être moins clairement définies.
- Temps de calcul : Le bagging nécessite la construction de plusieurs arbres, ce qui peut augmenter le temps de calcul nécessaire pour entraîner le modèle.
- Sensibilité aux données aberrantes : Le bagging peut être sensible aux données aberrantes présentes dans l'ensemble de données, car ces observations peuvent influencer négativement les prédictions agrégées.

2.2.7 Forêts aléatoires (en anglais Random Forest, RF)

Les forêts aléatoires sont une extension du bagging qui vise à améliorer davantage les performances des modèles en introduisant de l'aléa lors de la construction des arbres de décision. Les forêts aléatoires combinent le pouvoir de l'agrégation des prédictions du bagging avec la diversité des arbres aléatoires, ce qui en fait une méthode encore plus robuste et performante.

2.2.8 Principe de construction des forêts aléatoires

Le principe de construction des forêts aléatoires est similaire à celui du bagging, avec quelques différences clés qui ajoutent de l'aléa supplémentaire. Ainsi l'algorithme de constructuction des forêts aléatoires peut être résumé comme suit :

1. Pour chaque arbre de la forêt aléatoire :
   (a) Un échantillon bootstrap de taille n est généré en sélectionnant aléatoirement n observations avec remplacement à partir de l'ensemble d'apprentissage. Cela permet de créer un sous-ensemble d'observations pour l'arbre.
   (b) Une sélection aléatoire de m variables prédictives est effectuée parmi toutes les variables disponibles. Généralement, m est inférieur au nombre total de variables.
   (c) Un arbre de régression, tel qu'un arbre CART, est construit sur l'échantillon bootstrap en utilisant uniquement les variables sélectionnées.
2. Les prédictions de chaque arbre sont combinées en moyennant pour obtenir la prédiction finale $\hat{\mu}_{\mathrm{RF}}(x)$ pour chaque observation de la forêt aléatoire :

$$\hat{\mu}_{\mathrm{RF}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{\mu}_b(x)$$

où $\hat{\mu}_{\mathrm{RF}}(x)$ est la prédiction finale pour l'observation x, $\hat{\mu}_b(x)$ est la prédiction de l'arbre b, et B est le nombre total d'arbres dans la forêt aléatoire.

Remarquons au passage que tout comme celui du Bagging, l'algorithme de construction des forêts aléatoires est itératif, et chaque arbre est construit de manière indépendante des autres. La combinaison de l'échantillonnage bootstrap et de la sélection aléatoire des variables prédictives permet de créer un ensemble diversifié d'arbres dans la forêt.

2.2.9 Biais et variance des prédictions

Les forêts aléatoires partagent les mêmes propriétés de biais et de variance que le bagging.

En termes de biais, les prédictions des forêts aléatoires sont biaisées de la même manière que les prédictions individuelles des arbres constituant la forêt. Le biais des prédictions baggées reste le même que celui des prédictions individuelles des arbres.

En ce qui concerne la variance, les forêts aléatoires réduisent davantage la variance par rapport au bagging en introduisant de l'aléa lors de la construction des arbres. La sélection aléatoire des variables prédictives à chaque nœud permet de créer des arbres différents, ce qui favorise la diversité des prédictions. Cette diversité réduit la corrélation entre les arbres et conduit à une réduction de la variance globale des prédictions des forêts aléatoires.

2.2.10 Estimation de l'erreur Out-of-Bag (OOB)

L'estimation de l'erreur OOB dans les forêts aléatoires fonctionne de la même manière que dans le bagging. Les observations exclues lors de l'échantillonnage bootstrap sont utilisées pour évaluer les performances prédictives du modèle sans nécessiter un ensemble de validation distinct.

L'erreur OOB dans les forêts aléatoires est estimée en calculant l'erreur de prédiction sur les observations exclues lors de la construction de chaque arbre. Les prédictions OOB pour chaque observation sont obtenues en utilisant uniquement les arbres pour lesquels l'observation en question a été exclue lors de l'échantillonnage bootstrap.

L'estimation de l'erreur OOB est ensuite obtenue en prenant la moyenne des mesures d'erreur sur toutes les observations exclues. Cette estimation permet d'évaluer les performances de la forêt aléatoire sans avoir besoin d'un ensemble de validation distinct.

2.2.11 Interactions

Tout comme le bagging, les forêts aléatoires peuvent capturer les interactions entre les variables explicatives. Cependant, grâce à leur nature d'ensemble d'arbres, les forêts aléatoires offrent une meilleure capacité à détecter et à modéliser les interactions complexes.

Les arbres individuels de la forêt se concentrent sur différentes combinaisons de variables explicatives et créent des partitions qui représentent des sous-espaces avec des relations spécifiques. L'agrégation des prédictions de tous les arbres permet de prendre en compte les interactions entre les variables et contribue à la prédiction globale de la forêt.

Comparées au bagging, les forêts aléatoires ont deux avantages clés pour capturer les interactions :

1. **Variables aléatoires** : Les forêts aléatoires utilisent une sélection aléatoire des variables explicatives pour construire chaque arbre. Cela favorise l'inclusion de différentes variables dans les arbres, permettant d'explorer un espace plus large des combinaisons de variables et de détecter les interactions potentielles.

2. **Agrégation des prédictions** : L'agrégation des prédictions des arbres dans les forêts aléatoires prend en compte les interactions entre les variables. Les prédictions des arbres sont combinées de manière à minimiser l'erreur globale, renforçant ainsi les effets des interactions significatives et réduisant les effets des interactions non significatives ou indésirables.

Ces mécanismes confèrent aux forêts aléatoires une plus grande flexibilité pour modéliser les interactions entre les variables explicatives. Cependant, l'interprétation des interactions peut être plus complexe en raison de la combinaison des effets de plusieurs arbres. Des techniques spécifiques, telles que l'analyse des dépendances partielles, peuvent être utilisées pour comprendre et visualiser les interactions dans les forêts aléatoires.

Il est important de noter que la capacité des forêts aléatoires à capturer les interactions dépend de divers paramètres d'ajustement. Une exploration et une calibration appropriées de ces paramètres permettent de mieux exploiter le potentiel des forêts aléatoires pour modéliser les interactions entre les variables explicatives.

2.2.12 Calibration des paramètres

La calibration des paramètres dans les forêts aléatoires fait référence au processus d'ajustement des valeurs optimales des hyperparamètres du modèle. Elle peut être réalisée en utilisant différentes méthodes, telles que la recherche en grille (*grid search*) ou la recherche aléatoire (*random search*), associées à une validation croisée (*cross-validation*). Ces approches permettent d'évaluer les performances du modèle sur différentes combinaisons d'hyperparamètres et de sélectionner celles qui optimisent les performances du modèle.

Les forêts aléatoires possèdent plusieurs hyperparamètres qui peuvent être ajustés pour optimiser le modèle. Les plus couramment calibrés sont :

- **Le nombre d'arbres (`ntree`)** : Il s'agit du nombre d'arbres à inclure dans la forêt aléatoire. Un nombre plus élevé peut améliorer les performances du modèle, mais cela peut également entraîner un temps d'exécution plus long.
- **La taille minimale des nœuds (`nodesize`)** : Ce paramètre détermine le nombre minimum d'échantillons requis pour qu'un nœud puisse être scindé en deux sous-nœuds. Une valeur plus élevée peut limiter la complexité du modèle et le régulariser en évitant les divisions sur de petits groupes d'échantillons.
- **La fraction de variables à considérer pour chaque split (`mtry`)** : Ce paramètre contrôle le sous-ensemble de variables à prendre en compte lors de la recherche de la meilleure division d'un nœud. Une valeur plus faible favorise la diversité des arbres de la forêt en limitant la corrélation entre les arbres.

La calibration des hyperparamètres est une étape cruciale pour obtenir des performances optimales dans les forêts aléatoires. Il est recommandé d'expérimenter différentes combinaisons d'hyperparamètres en utilisant des méthodes de recherche systématique et en évaluant les performances du modèle à l'aide de la validation croisée. Une fois les hyperparamètres calibrés, le modèle peut être entraîné sur l'ensemble de données complet pour effectuer des prédictions précises.

Il est important de noter que les symboles entre parenthèses (`ntree`, `nodesize`, `mtry`) correspondent aux noms des paramètres utilisés dans le langage de programmation R que nous avons utilisé dans le cadre de ce mémoire.

2.2.13 Interprétation

Importance relative des variables explicatives

L'interprétation de l'influence relative des variables explicatives dans les forêts aléatoires est similaire à celle dans le bagging.

Dépendances partielles

L'analyse des dépendances partielles dans les forêts aléatoires fonctionne de la même manière que dans le bagging.

2.2.14 Interactions

De manière identique au bagging, les forêts aléatoires sont capables de capturer les interactions entre les variables explicatives. Toutefois, en utilisant le processus de sélection aléatoire des variables lors de la construction des arbres, les forêts aléatoires peuvent modéliser des interactions un peu plus complexes entre les variables.

2.2.15 Avantages et limites des Forêts Aléatoires

Les forêts aléatoires présentent plusieurs avantages et limites importants :

Avantages

- Réduction de la variance : Les forêts aléatoires réduisent la variance des prédictions par rapport au bagging en introduisant de l'aléa lors de la construction des arbres. Cela rend les modèles plus stables et moins sensibles aux variations des données.
- Estimation de l'erreur OOB : Les forêts aléatoires utilisent les observations exclues lors de l'échantillonnage bootstrap pour estimer l'erreur de généralisation du modèle, ce qui permet d'évaluer les performances sans avoir besoin d'un ensemble de validation distinct.
- Mesure de l'importance des variables explicatives : Les forêts aléatoires fournissent des mesures d'importance relatives des variables explicatives, ce qui permet de quantifier leur contribution à la prédiction globale du modèle.

Limites

- Moins d'interprétabilité : Les forêts aléatoires rendent les modèles moins interprétables en raison de l'agrégation des prédictions de plusieurs arbres et de la sélection aléatoire des variables prédictives. Les relations entre les variables explicatives peuvent être moins clairement définies.
- Temps de calcul : Les forêts aléatoires nécessitent la construction de plusieurs arbres, ce qui peut augmenter le temps de calcul nécessaire pour entraîner le modèle par rapport à un arbre de décision unique.
- Sensibilité aux données aberrantes : Les forêts aléatoires peuvent être sensibles aux données aberrantes présentes dans l'ensemble de données, car ces observations peuvent influencer négativement les prédictions agrégées.

2.3 Gradient Boosting

Le Gradient Boosting est une autre technique d'apprentissage automatique qui combine des modèles simples, appelés *weak learners*, pour former un modèle prédictif plus puissant. Il repose sur l'idée d'ajouter itérativement des modèles faibles à un modèle global en corrigeant les erreurs faites par les modèles précédents.

2.3.1 Principe de construction du Gradient Boosting

Le gradient boosting est une variante du boosting (Denuit et al., 2020). Son principe est de minimiser la fonction de perte en utilisant une approche d'optimisation par descente de gradient. Plus concrêtement, il s'agit de construire itérativement une séquence de modèles faibles (*weak learners*) en ajustant les résidus des prédictions précédentes. À chaque étape, un modèle faible est ajouté pour prédire les résidus restants.

Dans le contexte du gradient boosting :
- $\beta_{t,a_t} h(x; a_t)$ représente ce modèle faible.
- $f(x)$ représente l'apprentissage final résultant des modèles faibles individuels.

La procédure d'estimation des paramètres β_{t,a_t} et a_t du gradient boosting peut être résumée comme suit :

1. **Calcul des gradients (négatifs)**
   À chaque itération, les gradients négatifs, également appelés « pseudo-résidus », sont calculés. Le pseudo-résidu pour chaque observation i est défini comme :

$$r_i = - \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \bigg|_{f(x_i)=f_{t-1}(x_i)}, \quad i = 1, \ldots, n.$$

   où $L(y_i, f(x_i))$ est la fonction de perte mesurant l'écart entre la vraie valeur cible y_i et la prédiction actuelle $f(x_i)$.
2. **Estimation** a_t
   Dans cette étape, un modèle faible, est ajusté aux gradients négatifs $(r_{i=1,\ldots,n})$ obtenus à l'étape précédente afin d'estimation a_t
3. **Estimation de** β_{t,a_t}

À ce niveau le paramètre β_{t,a_t}, est estimé en minimisant la fonction de perte

$$\min_{\beta_t, a_t} \sum_{i=1}^{n} L\left(y_i, f_{t-1}(x_i) + \beta_t h(x_i; a_t)\right)$$

, où $f_{t-1}(x_i)$ représente les prédictions des itérations précédentes.

4. **Mise à jour du modèle**
   Les prédictions du modèle faible $\beta_{t,a_t} h(x; a_t)$, sont ajoutées aux prédictions précédentes $f_{t-1}(x)$, pour mettre à jour le modèle :

   $$f_t(x) = f_{t-1}(x) + \beta_t h(x; a_t).$$

   Les étapes précédentes sont répétées pour un nombre prédéterminé d'itérations, T, ou jusqu'à ce qu'un critère de convergence soit atteint. La prédiction finale est obtenue en sommant les prédictions de tous les apprenants faibles $f(x) = \sum_{t=1}^{T} \beta_t h(x; a_t)$.

2.3.2 Gradient Boosting avec les arbres

Le Gradient Boosting avec les arbres (plus généralement appelé Gradient Boosting Trees) est une variante du gradient boosting qui utilise des arbres de décision comme modèle faibles. Cette méthode combine les avantages du gradient boosting et des arbres de décision pour améliorer les performances prédictives du modèle.

Contrôle du surajustement

Pour prévenir le surajustement du modèle de gradient boosting, plusieurs stratégies peuvent être utilisées :

— **Contrôle du nombre d'itérations** : Le nombre d'itérations T (c'est-à-dire le nombre d'arbres) peut être contrôlé pour éviter un surapprentissage excessif. Une évaluation indépendante sur un échantillon de test ou une validation croisée peut être utilisée pour sélectionner la valeur optimale de T.

— **Réduction de la contribution de chaque arbre** : Une stratégie supplémentaire consiste à réduire la contribution de chaque arbre (*weak learner*) en utilisant un facteur d'échelle $\tau \in (0, 1]$. Cela revient à faire une mise à jour d'algorithme décrit précédemment de la manière suivante :

$$f_t(x) = f_{t-1}(x) + \tau \cdot \beta_{t,a_t} \cdot h(x; a_t)$$

. Le paramètre τ ralentit le taux d'apprentissage de la série d'arbres, ce qui conduit à une série plus longue mais avec une précision accrue. Empiriquement, de petits facteurs d'échelle τ (< 0.1) peuvent entraîner des améliorations significatives. Cependant, un τ faible nécessite plus d'itérations et donc plus de temps de calcul.

Réduction de la variance

Pour réduire la variance du modèle de gradient boosting, il est possible d'incorporer l'aléatoire comme partie intégrante de la procédure d'ajustement :

— **Échantillonnage aléatoire sans remise** : Dans l'étape d'estimation du paramètre , à chaque itération, un échantillon aléatoire simple est prélevé sans remise. Cet échantillon est généralement d'environ la moitié de la taille de l'ensemble de données d'entraînement complet. En conséquence, la variance des estimations individuelles des apprenants faibles à chaque itération augmente, mais il y a moins de corrélation entre ces estimations à différentes itérations. Cela se traduit par une réduction de la variance du modèle combiné.

— **Réduction du temps de calcul** : L'utilisation d'échantillons de taille réduite permet de réduire le temps de calcul nécessaire. En effet, le fait de prendre des demi-échantillons réduit le temps de calcul d'environ 50%.

Le Gradient Boosting Trees tire parti de ces modifications pour contrôler le surajustement, réduire la variance et améliorer les performances du modèle de gradient boosting. Ces ajustements permettent d'obtenir un modèle plus robuste et plus performant tout en conservant une certaine flexibilité dans la modélisation des relations entre les variables explicatives et la variable cible.

Fonctions de lien (Link functions)

Dans le contexte du Gradient Boosting, les fonctions de liaison sont utilisées de manière similaire aux modèles linéaires et additifs généralisés (GLM) abordés dans le chapitre 1.

2.3.3 Calibration des paramètres

La calibration des paramètres dans l'algorithme du Gradient Boosting est une étape essentielle pour obtenir de bonnes performances de prédiction. Les paramètres du Gradient Boosting peuvent être ajustés afin d'optimiser la performance du modèle et d'éviter le surajustement. Voici une liste des paramètres clés que l'on peut calibrer dans le Gradient Boosting :

- **Nombre d'arbres** (`n.trees`) : Il s'agit du nombre d'arbres à construire dans l'algorithme. Un nombre trop faible peut conduire à un sous-apprentissage, tandis qu'un nombre trop élevé peut entraîner un surapprentissage. Il est important de trouver un équilibre pour obtenir une performance optimale.
- **Profondeur des arbres** (`interaction.depth`) : Ce paramètre contrôle la complexité des arbres individuels dans l'algorithme. Une profondeur plus élevée permet aux arbres de capturer des relations plus complexes dans les données, mais peut également augmenter le risque de surapprentissage. Une valeur appropriée doit être choisie en fonction de la complexité du problème.
- **Taux d'apprentissage** (`shrinkage`) : Le taux d'apprentissage contrôle l'impact de chaque arbre sur la prédiction finale. Un taux d'apprentissage plus faible donne plus de poids aux itérations précédentes, ce qui peut améliorer la robustesse du modèle, mais nécessite généralement plus d'arbres pour atteindre une performance similaire. Il est recommandé d'expérimenter différentes valeurs pour trouver le bon compromis.
- **Fraction d'échantillonnage** (`bag.fraction`) : Ce paramètre détermine la fraction d'échantillonnage utilisée pour ajuster chaque arbre. En réduisant cette fraction, on peut réduire la variance du modèle, ce qui peut être bénéfique pour les ensembles de données de petite taille. Cependant, une valeur trop faible peut également entraîner une perte de performance si l'échantillon devient trop petit.
- **Nombre minimum d'observations dans les nœuds terminaux** (`n.minobsinnode`) : Ce paramètre impose un nombre minimum d'observations pour qu'un nœud soit considéré comme terminal dans un arbre. En augmentant cette valeur, on limite la complexité de l'arbre et on réduit le risque de surapprentissage. Cependant, une valeur trop élevée peut entraîner une sous-représentation de certains schémas dans les données.

Tout comme dans les forêts aléatoires, la calibration des hyperparamètres dans le Gradient Boosting utilise couramment deux approches : la recherche par grille (*grid search*) et la recherche aléatoire (*random search*), associées à la validation croisée (*cross-validation*).

2.3.4 Interprétation

Comme nous l'avons déjà discuté précédemment dans les sections sur les forêts aléatoires, l'interprétation des Gradient Boosting Machines (GBM) partage des concepts similaires. Cela comprend l'importance relative des variables, qui mesure l'influence de chaque variable sur les prédictions du modèle, ainsi que les dépendances partielles, qui permettent d'analyser l'effet spécifique d'une variable tout en maintenant les autres constantes. Ces approches sont utilisées pour comprendre la relation entre les variables explicatives et la variable cible dans les GBM, tout comme dans les forêts aléatoires.

En combinant l'importance relative des variables avec les dépendances partielles, nous obtenons une compréhension plus complète et approfondie du modèle. L'importance relative des variables nous indique quelles variables sont les plus influentes, tandis que les dépendances partielles nous permettent de comprendre comment ces variables interagissent avec la variable cible.

Ces outils d'interprétation peuvent être extrêmement précieux pour comprendre le fonctionnement du modèle, identifier les relations importantes entre les variables, détecter les interactions complexes et valider les hypothèses. Ils aident également à communiquer les résultats de manière claire et convaincante aux parties prenantes et à établir la confiance dans le modèle prédictif.

2.3.5 Interactions : statistique de Friedman

Les modèles basés sur les arbres de décision sont réputés pour leur capacité à prendre en compte les effets d'interaction entre les variables. Les interactions se produisent lorsque l'effet d'une variable sur la variable cible dépend de la valeur d'une autre variable. L'identification et la compréhension des interactions entre les variables peuvent fournir des informations précieuses pour interpréter le modèle.

La statistique H de Friedman, également connue sous le nom de H-statistique, est souvent utilisée pour mesurer les interactions entre les variables dans les modèles d'arbres de décision. Cette statistique permet de quantifier l'importance des interactions et d'identifier les paires de variables qui interagissent.

La statistique H de Friedman est calculée à l'aide de la formule suivante (Denuit et al., 2020) :

$$H^2_{j,k} = \frac{\sum_{i\in I}(\delta_{j,k}(x_{ji}, x_{ki}) - \delta_j(x_{ji}) - \delta_k(x_{ki}))^2}{\sum_{i\in I}\delta^2_{j,k}(x_{ji}, x_{ki})}$$

où : $\delta_{j,k}(x_{ji}, x_{ki})$ est la dépendance partielle de l'interaction entre les variables x_j et x_k évaluée aux valeurs observées x_{ji} et x_{ki}. $\delta_j(x_{ji})$ et $\delta_k(x_{ki})$ sont les dépendances partielles respectives des variables x_j et x_k évaluées aux valeurs observées x_{ji} et x_{ki} pour l'individu i. I est l'ensemble des indices des observations I données d'apprentissage.

La statistique $H^2_{j,k}$ mesure la variance expliquée par l'interaction entre les variables x_j et x_k par rapport à la variance totale. Une valeur de $H^2_{j,k}$ proche de 0 indique qu'il n'y a pas d'interaction significative entre les variables, tandis qu'une valeur proche de 1 indique une interaction forte.

De même, la statistique H^2_j peut être utilisée pour mesurer l'interaction d'une variable x_j avec toutes les autres variables du modèle. Elle est calculée comme suit :

$$H^2_j = \frac{\sum_{i\in I}(\delta(x_i) - \delta_j(x_{ji}) - \delta_{\bar{j}}(x_{\bar{j}i}))^2}{\sum_{i\in I}\delta^2(x_i)}$$

où $\delta(x_i)$ est la dépendance partielle globale du modèle évaluée à l'observation x_i, $\delta_j(x_{ji})$ est la dépendance partielle de la variable x_j évaluée à x_{ji}, et $\delta_{\bar{j}}(x_{\bar{j}i})$ est la dépendance partielle de toutes les autres variables sauf x_j évaluée à $x_{\bar{j}i}$.

En utilisant ces statistiques, il est possible d'identifier les interactions significatives entre les variables et de quantifier leur importance relative dans le modèle. Cela permet une meilleure compréhension des relations complexes entre les variables et la variable cible.

Il convient de noter que l'estimation précise des interactions peut demander des calculs intensifs, en particulier dans des modèles plus complexes ou avec un grand nombre de variables. Des techniques d'approximation et des méthodes d'échantillonnage peuvent être utilisées pour réduire la charge de calcul et estimer les interactions de manière plus efficace.

L'analyse des interactions entre les variables est essentielle pour une interprétation approfondie des modèles basés sur les arbres de décision. Elle permet de découvrir des relations complexes et de déterminer comment les variables interagissent pour influencer la prédiction du modèle.

2.3.6 Avantages et limites du gradient boosting

Le Gradient Boosting est une technique d'apprentissage automatique puissante qui présente plusieurs avantages, mais aussi certaines limites. Commençons par examiner les avantages du Gradient Boosting :

Avantages

1. **Haute précision prédictive :** Le Gradient Boosting est connu pour produire des modèles prédictifs de haute qualité. En combinant plusieurs modèles faibles, il est capable de capturer des relations complexes entre les variables explicatives et la variable cible, ce qui conduit à des prédictions précises.

2. **Gestion des données hétérogènes :** Le Gradient Boosting peut facilement gérer des ensembles de données hétérogènes, comprenant des variables de différents types (numériques, catégorielles, etc.). Il est capable de traiter des données structurées et non structurées sans nécessiter de prétraitement complexe.
3. **Interprétabilité :** Contrairement à certaines techniques d'apprentissage automatique "boîte noire", le Gradient Boosting offre une interprétabilité relativement élevée. Les dépendances partielles, les mesures d'importance des variables et d'autres techniques permettent de comprendre et d'expliquer les relations entre les variables et la prédiction du modèle.
4. **Robustesse aux données bruitées :** Le Gradient Boosting est robuste aux données bruitées ou aux valeurs aberrantes. En combinant plusieurs modèles, il est capable de réduire l'impact des valeurs atypiques et des erreurs de mesure, ce qui contribue à des prédictions plus fiables.
5. **Gestion de grandes dimensions :** Le Gradient Boosting peut gérer efficacement des ensembles de données avec un grand nombre de variables. Grâce à des techniques telles que la sélection automatique des caractéristiques et la régularisation, il est capable de gérer la dimensionnalité élevée tout en évitant le surajustement.

Limites

Malgré ses nombreux avantages, le Gradient Boosting présente également certaines limites :

1. **Sensibilité aux hyperparamètres :** Le Gradient Boosting nécessite un réglage précis des hyperparamètres pour obtenir de bonnes performances. Des hyperparamètres tels que le taux d'apprentissage, le nombre d'arbres, la profondeur des arbres, etc., doivent être soigneusement ajustés, ce qui peut être chronophage et nécessiter une expertise.
2. **Risque de surajustement :** Comme tout autre algorithme d'apprentissage automatique, le Gradient Boosting est susceptible de surajuster les données d'apprentissage si les hyperparamètres ne sont pas réglés correctement ou si le nombre d'arbres est trop élevé. Une régularisation appropriée et une validation croisée peuvent aider à atténuer ce risque.
3. **Temps d'entraînement plus long :** Le Gradient Boosting peut nécessiter plus de temps d'entraînement par rapport à certains autres algorithmes d'apprentissage automatique. En raison de la nature séquentielle de l'algorithme, chaque étape nécessite le calcul de gradients et la construction de nouveaux arbres, ce qui peut être coûteux en termes de temps de calcul.
4. **Sensibilité aux valeurs aberrantes :** Bien que le Gradient Boosting soit généralement robuste aux valeurs aberrantes, il peut être sensible à des valeurs aberrantes extrêmes. Les valeurs aberrantes extrêmes peuvent avoir un impact disproportionné sur les prédictions et affecter la qualité du modèle.
5. **Difficulté pour les données très déséquilibrées :** Le Gradient Boosting peut avoir des difficultés à gérer les ensembles de données très déséquilibrés, où les classes minoritaires sont sous-représentées. Des techniques telles que le sur-échantillonnage ou le sous-échantillonnage peuvent être nécessaires pour améliorer les performances sur de telles données.

2.4 Implémentation dans R du GBM

Dans cette section, nous mettons en pratique la construction et l'ajustement du modèle GBM en utilisant le langage de programmation R et le package "gbm3".

2.4.1 Construction du modèle GBM

La construction du modèle GBM est réalisée en utilisant la fonction *gbm()* du package *gbm3*. Cette fonction nécessite les données d'entraînement, une formule décrivant la relation entre les variables prédictives et la variable cible, ainsi que divers paramètres de configuration tels que le nombre d'itérations, le taux d'apprentissage et la profondeur des arbres.

Il est essentiel de calibrer correctement ces paramètres avant la construction du modèle. Pour cela, nous avons effectué une recherche par grille avec une validation croisée pour évaluer la performance de différents ensembles de modèles GBM. Nous avons testé un taux d'apprentissage de 1% (recommandé),

des profondeurs d'arbre de 1 à 10, et nous avons utilisé la mesure de performance pour choisir les meilleurs paramètres.

Les résultats de la calibration sont présentés dans le graphique 2.1, qui montre la performance en fonction des différentes combinaisons de paramètres testées.

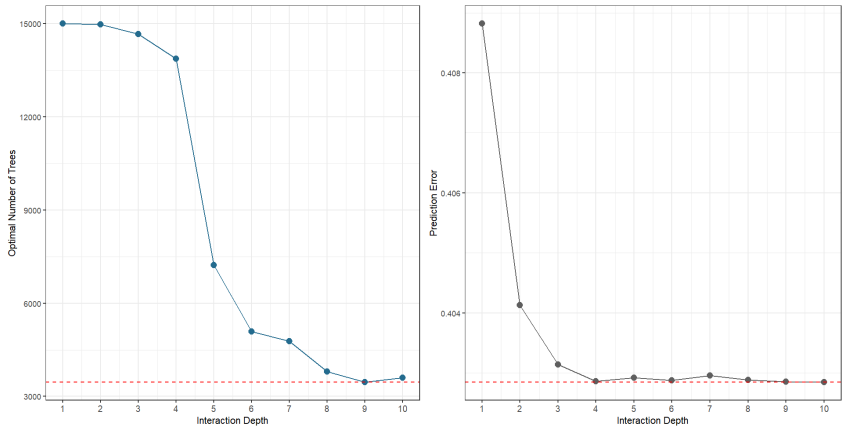

FIGURE 2.1 – Calibration des paramètres du GBM

Les paramètres optimaux obtenus sont les suivants :
— Taux d'apprentissage (shrinkage parameter) : 1%
— Profondeur des arbres (interaction depth) : 10
— Nombre minimum d'observations dans les nœuds (minobsinnode) : 1
— Fraction d'échantillonnage (bag.fraction) : 50%
— Nombre total d'arbres : 3601

2.4.2 Ajustement du modèle et interprétation

Une fois le modèle GBM construit avec les paramètres optimaux, nous l'avons ajusté sur les données d'entraînement. L'analyse des résultats révèle que les variables "BonusMalus", "VehAge", "VehBrand" et "DrivAge" ont une influence significative sur la fréquence des sinistres. La figure 2.2 présente l'importance relative de ces variables.

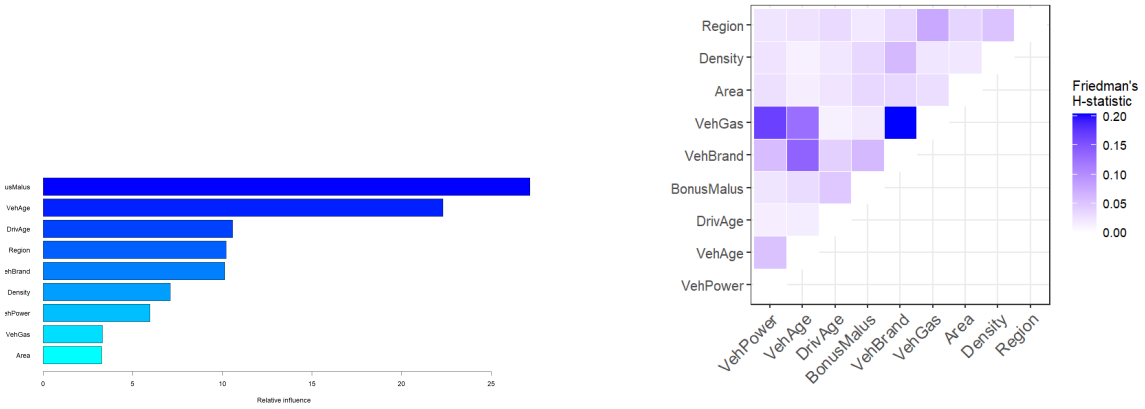

FIGURE 2.2 – Importance relative et intensité des interactions entre les variables explicatives selon la mesure de Friedman.

De plus, le graphe des dépendances partielles (figure 2.3) met en évidence les relations entre la fréquence des sinistres et chacune de ces variables. Cette visualisation permet de mieux comprendre le risque représenté par chaque assuré en fonction de ses caractéristiques.

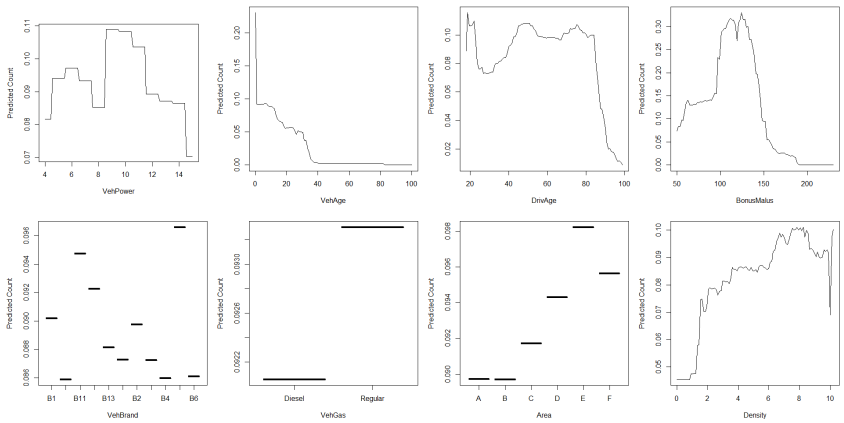

FIGURE 2.3 – Graphe des dépendances partielles

Comparaison avec les GLM et les GAM

Pour évaluer la performance du modèle GBM, nous l'avons comparé aux modèles GLM et GAM précédemment utilisés. Le tableau 2.1 présente les résultats de cette comparaison, notamment les pertes internes et externes (sur les ensembles de validation et de test), ainsi que les critères AIC et BIC.

TABLE 2.1 – Comparaison des performances des modèles

Modèle	**In Sample Loss**	**validation set**	**Test set**	**AIC**	**BIC**
GLM	0.318888	0.317176	0.328850	170861.3	171319.7
GLM.	0.313163	0.310890	0.321766	168556.6	169124.1
GAM	0.316126	0.326025	0.314636	169797.6	170570.9
GBM	0.298511	0.313775	0.303663	NaN	NaN

Note : Le modèle GLM. est obtenu en utilisant la discrétisation des variables continues.

Cette comparaison montre que le modèle GBM présente de meilleures performances en termes de pertes internes et externes, ainsi que des critères AIC et BIC plus favorables. De plus, la figure 2.4 illustre la corrélation entre le GBM et la fréquence des sinistres, renforçant ainsi sa fiabilité.

À travers cette application nous venons avons monté comment modèles d'apprentissage automatique sont capables de produire des scores qui ont une meilleure corrélation avec la réponse par rapport au modèles traditionnels de modélisation de la fréquence de sinistres. Cependant pour y parvenir, celà nécessite un temps énorme pour la calibration des hyperparamètres et nous remarquons également que bien fortement correlés un transfert de prime (que nous voyons comme la fréquence modélisée en considérant les montants de sinstres unitaires) est partagé entre les assurés. Celà se voit grâce à la sous-estimation des primes issues du modèle pour les catégories très risqués. Le troisieme et dernier chapitre de notre mémoire est consacré à la présentation détaillée de la nouvelle technique basée sur le concept "l'autocalibration" proposé par Denuit et al. (2021) pour palier à ce déséquilibre.

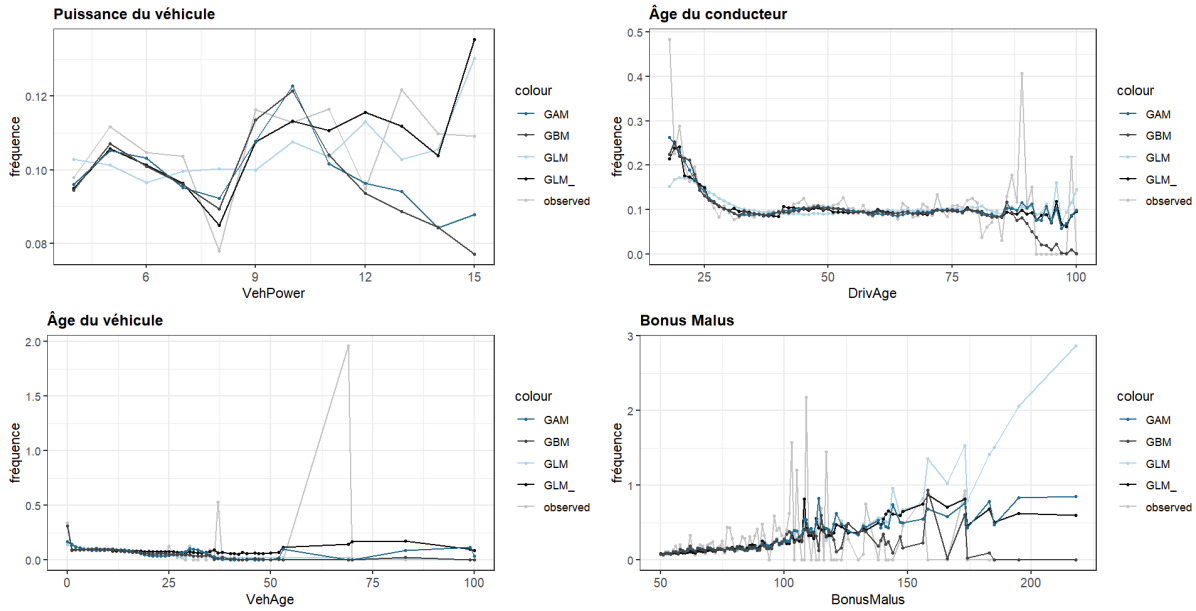

FIGURE 2.4 – Comparaison des modèles GLM, GAM et GBM

CHAPITRE 3

TARIFICATION EN ASSURANCE ET AUTOCALIBRATION

Sommaire

3.1 Problème de la tarification en assurance

La tarification en assurance est un processus essentiel qui vise à évaluer les risques associés à une police d'assurance et à déterminer les primes appropriées pour couvrir ces risques. Le but de celle-ci est de trouver un équilibre entre les primes payées par les assurés et les sinistres indemnisés par l'assureur. Cependant, ce processus est complexe en raison de plusieurs défis spécifiques au domaine de l'assurance.

Tout d'abord, la tarification doit prendre en compte la variabilité des risques. Chaque assuré présente des caractéristiques individuelles qui influencent la probabilité de survenance d'un sinistre et le montant potentiel du sinistre. Par conséquent, les primes doivent être différenciées en fonction de ces caractéristiques pour refléter de manière adéquate le niveau de risque de chaque assuré. De plus, elle doit tenir compte des contraintes réglementaires et légales incluant des limites sur les écarts de primes entre les différents groupes d'assurés.

Autrement dit, supposons que nous ayons une variable aléatoire Y représentant la réponse liée aux sinistres (fréquence, gravité ou pertes totales), et un vecteur $\mathbf{X} = (X_1, X_2, ..., X_p)$ des caractéristiques des assurés. L'objectif de la tarification en assurance est d'estimer la prime pure $\mu(\mathbf{X})$, qui représente l'espérance conditionnelle de la variable aléatoire Y étant donné les caractéristiques $\mathbf{X}$:

$$\mu(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}].$$

Pour ce faire, les acteurs de l'assurance utilisent des modèles statistiques tels que les modèles linéaires généralisés (GLM), les modèles additif généralisés (GAM) ou d'autres méthodes avancées comme le gradient boosting machine (GBM) ou les réseaux de neurones pour approximer cette prime pure par une estimation $\hat{\mu}(\mathbf{X})$, adaptée aux données historiques sur les sinistres et les caractéristiques des assurés.

Cependant, l'utilisation de modèles statistiques avancés basés sur des données historiques peut introduire des biais dans les prédictions, compromettant ainsi l'équilibre global et local entre les primes et les sinistres. Les modèles peuvent surestimer ou sous-estimer les primes prévues par rapport aux sinistres réels observés, entraînant ainsi des problèmes d'adéquation tarifaire.

Pour remédier à ces défis, Denuit et al. (2021) ont proposé une approche novatrice basée sur le concept d'autocalibration. Approche qui consiste à ajuster les modèles de tarification pour que les primes prévues correspondent de manière plus précise aux sinistres réels attendus et ainsi améliorer la rentabilité pour les compagnies d'assurance.

Les avancées récentes dans les méthodes d'apprentissage automatique basées sur des données historiques fiables offrent de nouvelles possibilités pour améliorer la précision des modèles de tarification en fournissant des prédictions très corrélées avec la prime. Cependant, l'utilisation de ces méthodes peut introduire des biais dans les prédictions, compromettant ainsi l'équilibre global et local entre les primes et les sinistres.

Dans les sections suivantes, nous plongerons plus en profondeur dans le concept d'autocalibration, explorant ses aspects théoriques et pratiques, tout en évaluant son impact sur l'amélioration de la déviance du modèle et sur la courbe de lift, tout comme l'ont fait Ciatto, Verelst, Trufin, et Denuit (2022). Notre objectif est de démontrer comment l'autocalibration peut être une approche prometteuse pour améliorer la tarification en assurance en rétablissant un équilibre adéquat entre les primes et les sinistres, tout en surmontant les défis spécifiques du domaine de l'assurance.

3.2 Définition de l'autocalibration

L'autocalibration est un concept fondamental en tarification actuarielle et en modélisation d'assurance. Il vise à assurer un équilibre entre les primes d'assurance prévues et les pertes réelles, en ajustant le prédicteur initial pour qu'il soit en accord avec les caractéristiques du portefeuille d'assurance. L'objectif est d'améliorer la précision des primes tarifaires tout en maintenant la cohérence entre les prévisions et les résultats observés.

3.2.1 Définition formelle

Formellement, un prédicteur $\pi(\mathbf{X})$ est considéré comme autocalibré si la prime prévue $\pi(\mathbf{X})$ est égale à l'espérance conditionnelle des sinistres réels Y donnée la prime prévue $\pi(\mathbf{X})$ pour chaque profil de risque $\mathbf{X}$:

$$\pi(X) = E[Y|\pi(X)] \quad \text{presque sûrement.}$$

En d'autres termes, un prédicteur autocalibré fournit des primes qui sont équilibrées avec les sinistres réels attendus, garantissant ainsi une tarification juste et adéquate. Cette définition est basée sur la présentation générale du concept donnée par Kruger et Ziegel (2020). Un aspect essentiel de l'autocalibration qu'il faut souligner est qu'elle garantit également que le prédicteur est moins variable que la variable réponse au sens de l'ordre convexe (Denuit et al., 2021).

3.2.2 Objectifs de l'autocalibration

L'autocalibration poursuit deux objectifs principaux :

1. **Équilibre local :** en imposant l'équation $\pi(\mathbf{X}) = \mathbb{E}[Y|\pi(\mathbf{X})]$, l'autocalibration assure que les primes prévues correspondent étroitement aux sinistres attendus dans des sous-ensembles de profils de risque similaires. Cela permet d'obtenir un équilibre local, où les primes reflètent les caractéristiques spécifiques des assurés dans chaque groupe homogène.
2. **Équilibre global :** il vise également à maintenir l'équilibre global entre les primes prévues et les pertes totales attendues dans l'ensemble du portefeuille d'assurance (cf. propriété d'équilibre vu précédemment). En d'autres termes, la somme des primes prévues doit être égale à la somme des sinistres attendus, garantissant ainsi une tarification globalement équilibrée.

3.2.3 L'autocalibration d'un prédicteur donné

Les auteurs ont supposé que la fonction $\hat{\pi}(X) \mapsto E[Y|\hat{\pi}(X)]$ est continuellement croissante, ce qui est une exigence raisonnable pour toute prime candidate et peut être testée à l'aide des outils développés après Bowman et al. (1998). Ensuite, ils ont montré qu'une manière simple de rétablir l'équilibre global consiste à passer de $\hat{\pi}$ à sa version corrigée d'équilibre $\hat{\pi}_{BC}$ définie comme suit :

$$\hat{\pi}_{BC}(X) = E[Y|\hat{\pi}(X)]$$

Cette version corrigée s'approche de l'espérance $E[Y]$, comme le montre leur Propriété 5.1. Il est à noter que la croissance de la fonction de régression $\hat{\pi}(X) \mapsto E[Y|\hat{\pi}(X)]$ n'est en réalité pas nécessaire dans la Propriété 5.1 de Denuit et al. (2021), car nous avons directement :

$$E[Y|\hat{\pi}_{BC}(X)] = E[E[Y|\hat{\pi}(X); \hat{\pi}_{BC}(X)]|\hat{\pi}_{BC}(X)] = E[E[Y|\hat{\pi}(X)]|\hat{\pi}_{BC}(X)] = \hat{\pi}_{BC}(X)$$

Pour mettre en œuvre l'autocalibration, une approche de régression polynomiale locale peut être utilisée. Cela permet de passer de $\hat{\pi}$ à $\hat{\pi}_{BC}$. Veuillez consulter Loader (1999) pour plus de détails sur l'approche de régression polynomiale locale avec le package "Locfit" de R. Dans cette approche, seule la prime $\hat{\pi}$ nécessitant une autocalibration est prise en compte. Un modèle linéaire généralisé local (GLM) est ajusté aux données $(Y_i, e_i, \hat{\pi}(x_i))$ comprises dans l'ensemble de données de validation, pour une exposition pertinente e_i. Prenons en compte un profil de risque spécifique x. Des poids uniformes sont attribués à chaque élément $(Y_i, e_i, \hat{\pi}(x_i))$ dans le voisinage $V(x)$ de x. Ainsi, une fonction de poids rectangulaire (ou uniforme) est spécifiée. Le voisinage $V(x)$ regroupe la fraction δ des données ayant les primes les plus proches $\hat{\pi}(x_i)$ de $\hat{\pi}(x)$. Ici, δ agit comme un paramètre de lissage et contrôle la taille des sous-portefeuilles où l'équilibre local est imposé. La valeur optimale du paramètre δ est sélectionnée par validation croisée de vraisemblance.

L'équation de vraisemblance du GLM local est la suivante :

$$\sum_{i \in V(x)} y_i = \sum_{i \in V(x)} e_i \hat{\pi}_{BC}(x)$$

Cette équation permet de transférer une partie de l'expérience aux valeurs voisines de $\hat{\pi}$ pour obtenir $\hat{\pi}_{BC}$. Ainsi, un modèle linéaire généralisé local constant, ou intercept-only GLM, implémente l'équilibre local dans les sous-portefeuilles regroupant les assurés ayant des valeurs prédites similaires.

3.3 Méthodologie de l'autocalibration dans R

Les étapes principales de la mise en œuvre de l'autocalibration proposée par Denuit et al. (2021) sont les suivantes :

1. **Tarification initiale :** Appliquez une tarification classique en utilisant le prédicteur initial $\pi^{(0)}(\mathbf{X})$, qui peut être basé sur des modèles statistiques ou d'autres méthodes traditionnelles de tarification. Cependant, ce prédicteur ne satisfait pas nécessairement l'autocalibration.
2. **Calibration du GLM local constant :** Dans cette étape, nous calibrons un modèle de Régression Linéaire Généralisée (GLM) local constant pour chaque profil de risque $\mathbf{x}$. Le GLM local constant est un modèle simple qui ne comprend qu'un terme intercept, c'est-à-dire qu'il est constant pour chaque voisinage de profils de risque.
   (a) **Définition du voisinage $V(\mathbf{x})$:** Pour chaque profil de risque $\mathbf{x}$, nous définissons un voisinage $V(\mathbf{x})$ qui rassemble un pourcentage α des données ayant les primes $\pi^{(0)}(\mathbf{x}_i)$ les plus proches de $\pi^{(0)}(\mathbf{x})$. Cela permet de créer des sous-ensembles de profils de risque similaires, où l'on imposera ensuite l'équilibre local.
   (b) **Attribution des poids ϕ_i :** À l'intérieur de chaque voisinage $V(\mathbf{x})$, nous attribuons des poids ϕ_i à chaque observation en fonction de la distance entre les primes $\pi^{(0)}(\mathbf{x}_i)$ et $\pi^{(0)}(\mathbf{x})$. Les poids $\phi_i(\pi^{(0)}(\mathbf{x}))$ sont plus grands pour les assurés dont les primes sont proches de $\pi^{(0)}(\mathbf{x})$ et vice versa. Le paramètre α contrôle la taille des sous-ensembles où l'on impose l'équilibre local.
   (c) **Sélection du paramètre α optimal :** Déterminez la valeur optimale du paramètre α en utilisant la validation croisée par vraisemblance. Cette étape permet de trouver le bon équilibre entre la calibration locale et globale. Un α plus élevé donnera des sous-ensembles plus importants avec un meilleur équilibre local, tandis qu'un α plus faible donnera des sous-ensembles plus petits avec une meilleure calibration globale.
3. **Calcul du prédicteur Balanced-Corrected ($\pi^{(BC)}(\mathbf{X})$) :** Pour chaque profil de risque $\mathbf{x}$, on calcule le prédicteur "Balanced-Corrected" $\pi^{(BC)}(\mathbf{X})$ en utilisant l'équation de vraisemblance du GLM local constant ajusté avec les poids $\phi_i(\pi^{(0)}(\mathbf{x}))$ pour ce profil spécifique. Ce prédicteur autocalibré sera plus équilibré et corrigera les déséquilibres globaux et locaux du prédicteur initial.
4. **Utilisation du prédicteur autocalibré :** Enfin, on utilise le prédicteur autocalibré $\pi^{(BC)}(\mathbf{X})$ pour ajuster les tarifs d'assurance. Ce prédicteur restaurera les équilibres globaux et locaux dans les sous-portefeuilles d'assurance, ce qui devrait améliorer la calibration des primes par rapport aux sinistres réels et garantir une tarification plus précise.

3.3.1 Sélection du paramètre de lissage

Dans le processus d'autocalibration, la détermination du paramètre de lissage optimal α est une étape cruciale pour obtenir un modèle de lissage bien équilibré entre la flexibilité locale et la régularisation globale. Pour cette sélection, nous utilisons la méthode de la validation croisée par vraisemblance (LCV) en nous appuyant sur les fonctionnalités fournies par le package `locfit`, qui met en œuvre la régression polynomiale locale.

La procédure de sélection du paramètre de lissage α consiste à générer un graphique de validation croisée (LCV plot) en fonction des degrés de liberté (DF) du modèle, pour différentes valeurs de α. Le LCV plot affiche la statistique de validation croisée par vraisemblance en fonction de la souplesse du modèle, représentée par les DF. Ce graphique nous permet d'identifier la valeur de α qui optimise la vraisemblance et, par conséquent, qui offre le meilleur ajustement du modèle.

Le LCV plot est réalisé à l'aide de la fonction `lcvplot` qui, en interne, appelle la fonction `lcv` pour calculer les statistiques de validation croisée. Nous obtenons ainsi des LCV plots pour le modèle Poisson GLM, le modèle Poisson GAM et le modèle Poisson boosté.

Suite à l'analyse des LCV plots de la figure 3.1 suivante, nous avons sélectionné les valeurs optimales pour α. Dans notre cas, une valeur commune 5% a été choisie pour les différents modèles. Cette valeur permet de garantir à nos modèles une flexibilité pour capturer les variations locales tout en évitant le surajustement sur de nouvelles données non vues.

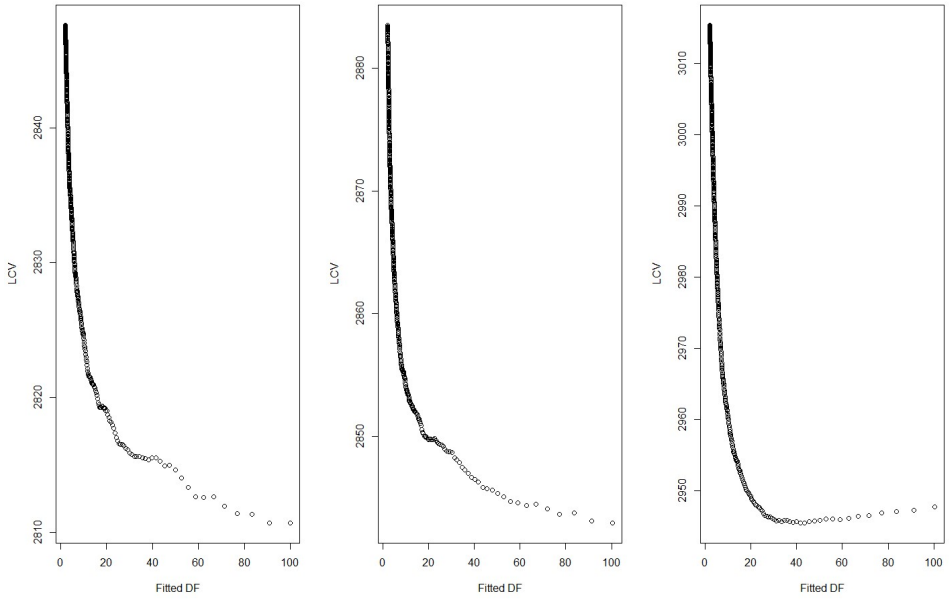

FIGURE 3.1 – LCV plots pour le GLM (à gauche), le GAM (au milieu) et le GBM (à droite)

Ces graphiques nous permettent de visualiser les performances du modèle pour différentes valeurs de α et d'identifier clairement les valeurs optimales qui conduisent à un ajustement optimal du modèle.

3.4 Mesures de lift pour l'évaluation des modèles

Comme déjà mentionné dans les pécédents chapitres, l'évaluation des modèles est une étape cruciale de la modélisation. Cette étape permet d'analyser la capacité prédictive des modèles en d'autres termes évaluer leurs facultés à différencier avec précision les meilleurs risques des pires risques. Parmi les méthodes d'évaluation, les méthodes de lift offrent une approche visuelle puissante pour évaluer la performance des modèles de régression en assurance.

3.4.1 Définition du lift

Le Lift est une mesure qui évalue la valeur économique d'un modèle de régression en termes de sa capacité à améliorer la différenciation entre les meilleurs et les pires risques. Contrairement à l'ajustement du modèle sur les données d'entraînement, le Lift se concentre spécifiquement sur la capacité du modèle à distinguer les risques favorables (risques souhaités) des risques défavorables (risques indésirables). En assurance, cette capacité est cruciale pour établir une tarification équitable et éviter la sélection adverse.

3.4.2 Mesure de lift

Tevet et Khare (2013) distinguent plusieurs mesures notamment :

— **Graphique de quantiles simples ou Courbe de Lift Simple ou :** Les graphiques de quantiles simples sont un outil visuel puissant pour évaluer la capacité d'un modèle de régression à différencier les meilleurs risques des pires risques. La création d'une courbe de Lift simple implique les étapes suivantes :

1. Tri des données en fonction des prédictions du modèle (par exemple, les primes ou les taux de sinistres prédits) de la plus petite à la plus grande valeur.

2. Répartition des données en quantiles de taille égale (quintiles, déciles, etc.) regroupant les individus en groupes de risques similaires.
3. Calcul des moyennes des prédictions du modèle et des valeurs observées réelles dans chaque quantile.
4. Tracé de la courbe de Lift simple pour montrer la capacité du modèle à différencier les risques favorables des risques défavorables.

— **Courbe de Lift Double :** Une courbes de Lift double permet de comparer directement deux modèles de régression (par exemple, Modèle A et Modèle B, qui produisent des estimations de prime attendu pour chaque assuré). Le processus de création d'une courbe de Lift double est similaire à celui d'une courbe de Lift simple, mais avec l'ajout d'une étape pour comparer les résultats des deux modèles :

1. Calcul du Ratio de Tri pour chaque observation du jeu de test, en comparant les prédictions des Modèles A et B.
2. Tri des données en fonction du rapport de tri, du plus petit au plus grand.
3. Répartition des données en quantiles (quintiles, déciles, etc.) pour regrouper les individus en groupes de risques similaires basés sur le rapport de tri.
4. Calcul des moyennes pour chaque modèle dans chaque quantile.
5. Tracé de la courbe de Lift double pour permettre une comparaison directe des performances des deux modèles.

— **Loss Ratio Charts :** Les Loss Ratio Charts sont une variante des courbes de Lift qui utilisent le rapport des coûts réels et prédits (par exemple, le rapport des coûts de sinistres réels sur les coûts de sinistres prédits) pour évaluer la performance du modèle. Les étapes pour créer un Loss Ratio Chart sont similaires à celles d'une courbe de Lift simple ou double, mais les coûts réels et prédits sont représentés plutôt que les prédictions sous forme de primes ou de taux de sinistres.

— **Indice de Gini :** L'indice de Gini, nommé en l'honneur du statisticien et sociologue Corrado Gini, est couramment utilisé en économie pour quantifier l'inégalité des revenus nationaux. Dans le contexte de l'assurance, l'indice de Gini peut être utilisé pour évaluer la capacité d'un modèle de régression à différencier les risques favorables des risques défavorables. Un indice de Gini plus élevé indique une meilleure capacité de discrimination du modèle.

Pour une description détaillée de ces méthodes voir Goldburd, Khare, Tevet, et Guller (2019).

3.4.3 Avantages des mesures de Lift

Les mesures de Lift offrent plusieurs avantages importants dans la validation et la sélection de modèles :

— **Facilité d'interprétation** : Les courbes de Lift offrent une visualisation claire de la performance du modèle, ce qui les rend faciles à interpréter même pour les non-experts en modélisation.
— **Comparaison directe** : Les courbes de Lift double permettent de comparer directement deux modèles, ce qui facilite la sélection du meilleur modèle pour un problème donné.
— **Focus sur les risques importants** : Les méthodes de Lift se concentrent sur la capacité du modèle à différencier les meilleurs risques des pires risques, ce qui est particulièrement utile dans les problèmes où les événements rares sont d'une grande importance.

Dans le cadre de notre étude, nous utiliserons les méthodes de courbes de Lift simple et de courbes de Lift double avec 10 déciles pour évaluer la performance et comparer nos modèles GLM, GAM et GBM avant et après leur autocalibration.. En utilisant une courbe de lift simple, pour determiner quel modèle offre un meilleur lift (prime pure différentiée), nous prenons en compte trois critères :

— **Précision prédictive :** Il s'agit de la capacité de chaque modèle à prédire avec précision les primes pures réelles dans chaque quantile.
— **Monotonie :** Nous nous assurons que les primes pures prédites et les primes pures réelles augmentent de manière monotone à mesure que le quantile augmente, bien que de légères inversions soient acceptables.
— **Distance verticale entre les premiers et derniers quantiles :** Le premier quantile contient les risques pour lesquels le modèle prévoit une expérience favorable, tandis que le dernier quantile contient les risques pour lesquels le modèle prévoit une expérience défavorable. Une différence significative (ou "lift") entre les primes pures réelles dans les quantiles avec les coûts de sinistre

prédits les plus faibles et les plus élevés indique que le modèle peut efficacement distinguer les meilleurs et les pires risques.

Dans la courbe de lift double, le modèle meilleur modèle dans notre analyse est celui qui correspond le plus étroitement à la prime pure réelle dans chaque quantile.

3.5 Résultats de l'autocalibration

3.5.1 Impact de l'autocalibration sur la déviance

En examinant les différences de déviance avant et après l'autocalibration, nous observons que pour les trois modèles (GLM, GAM et GBM), les valeurs de déviance sont plus faibles après l'autocalibration comme le montre la figure 3.2.

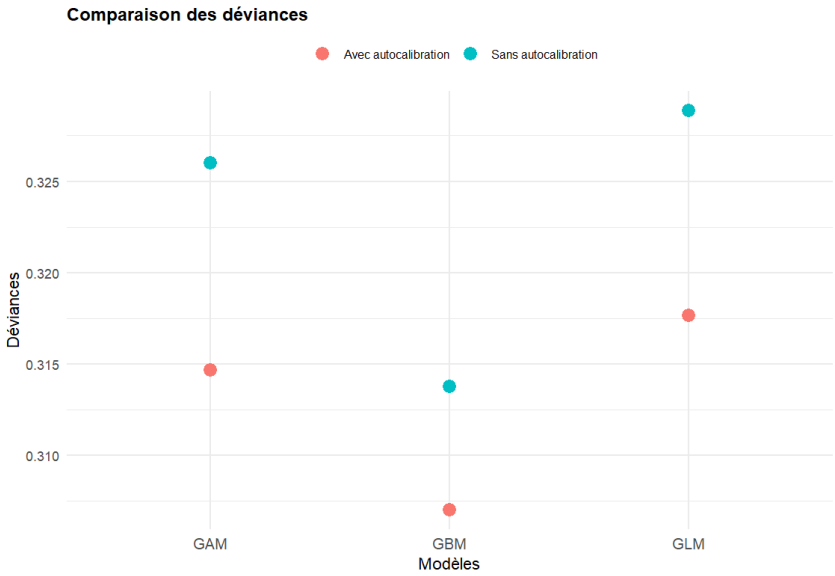

FIGURE 3.2 – Déviances avant et après autocalibration pour les modèles GLM, GAM et GBM.

En effet, ces résultats indiquent que l'autocalibration a permis d'améliorer l'ajustement des modèles, car les déviances diminuent après l'autocalibration. Par exemple, pour le modèle GAM, la déviance passe de 0.326025 sans autocalibration à 0.314664 avec autocalibration. Pour le modèle GBM, la déviance diminue de 0.313775 à 0.307047 après l'autocalibration. Une déviance plus faible indique que les prédictions des modèles se rapprochent davantage des valeurs réelles, ce qui est un indicateur positif de la précision prédictive du modèle.

En outre, les différences de déviance sont toutes positives, ce qui suggère que l'autocalibration a contribué à réduire les écarts entre les prédictions des modèles et les valeurs réelles. Par exemple, pour le modèle GAM, la différence de déviance avant et après l'autocalibration est de 0.011058, ce qui montre une nette amélioration de l'ajustement après l'autocalibration.

Ces résultats confirment les travaux de Denuit et al. (2021) qui ont montré que l'autocalibration permet de restaurer l'équilibre global et local des modèles en ajustant les primes prédites pour qu'elles correspondent plus étroitement aux totaux observés. En rétablissant l'équilibre, l'autocalibration contribue à améliorer la pertinence et la fiabilité des prévisions du modèle.

3.5.2 Impact de l'autocalibration sur le lift

Courbe de lift simple

L'analyse des courbes de lift simple pour les modèles GLM, GAM et GBM avant et après autocalibration nous permet d'évaluer l'impact de cette technique sur la précision prédictive de chaque modèle, en particulier dans différents déciles de risque.

Tout d'abord, observons le graphique 3.3 pour le modèle GLM :

Le modèle GLM présente une amélioration significative de la précision prédictive après autocalibration, comme le montre la courbe de lift simple. Dans chaque décile, les valeurs prédites avec autocalibration (GLM_BC) sont plus proches des valeurs observées que les valeurs prédites sans autocalibration (GLM). Cela indique que l'autocalibration a permis de mieux ajuster le modèle aux données réelles et d'améliorer sa capacité à prédire les primes pures dans chaque quantile de risque. L'impact positif de l'autocalibration est visible à la fois pour les risques favorables et défavorables, démontrant ainsi son efficacité globale dans l'amélioration de la précision prédictive du modèle GLM.

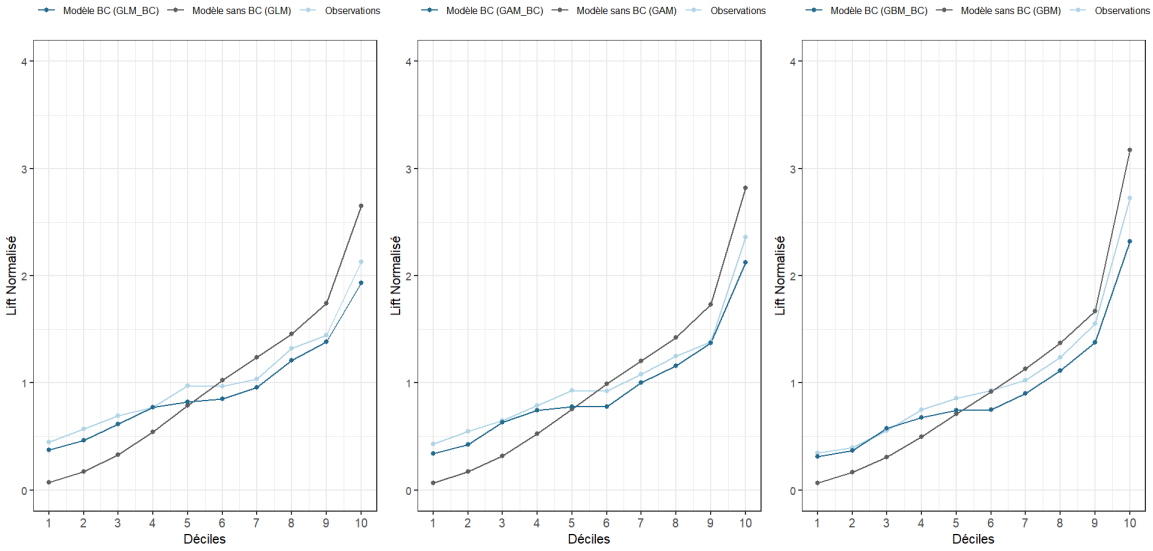

FIGURE 3.3 – Courbe de lift simple pour les modèles GLM, GAM et GBM.

TABLE 3.1 – Valeur de Lift avant et après autocalibartion

Modèle	Lift Avant	Lift Après
GLM	2.77809	1.51474
GAM	2.87585	1.81707
GBM	3.06783	1.01145

Ensuite, examinons les résultats pour le modèle GAM :

Le modèle GAM présente également une amélioration significative de la précision prédictive après autocalibration, comme le montre la courbe de lift simple. Les valeurs prédites avec autocalibration (GAM_BC) se rapprochent davantage des valeurs observées dans chaque décile par rapport aux valeurs prédites sans autocalibration (GAM). Cela confirme que l'autocalibration a permis d'ajuster le modèle aux données réelles et d'améliorer sa capacité à prédire les primes pures dans différentes classes de risque. L'impact positif de l'autocalibration est observé à la fois pour les risques favorables et défavorables, soulignant ainsi son efficacité globale dans l'amélioration de la précision prédictive du modèle GAM.

Enfin, analysons les résultats pour le modèle GBM :

Le modèle GBM présente une amélioration significative de la précision prédictive après autocalibration, comme le montre la courbe de lift simple. Dans chaque décile, les valeurs prédites avec autocalibration (GBM_BC) se rapprochent davantage des valeurs observées que les valeurs prédites sans autocalibration (GBM). Cela démontre que l'autocalibration a permis de mieux ajuster le modèle aux données réelles et d'améliorer sa capacité à prédire les primes pures dans différents niveaux de risque. Comme pour les autres modèles, l'impact positif de l'autocalibration est évident pour les risques favorables et défavorables, mettant en évidence son efficacité globale dans l'amélioration de la précision prédictive du modèle GBM.

L'analyse globale des courbes de lift simple pour les modèles GLM, GAM et GBM confirme que l'autocalibration a un impact significatif et positif sur la précision prédictive de chaque modèle. L'auto-

calibration a permis d'ajuster les modèles aux données réelles, améliorant ainsi leur capacité à prédire les primes pures dans différentes classes de risque. Ces résultats renforcent les travaux antérieurs de Denuit et al., montrant que l'autocalibration permet de restaurer l'équilibre local et global des modèles d'assurance, améliorant ainsi leurs performances dans la tarification et la gestion des risques. Ces conclusions ont des implications importantes pour le développement et l'utilisation de modèles actuariels plus précis et fiables, contribuant ainsi à une meilleure prise de décision et à une gestion plus efficace des risques dans l'industrie de l'assurance.

Courbe de lift double

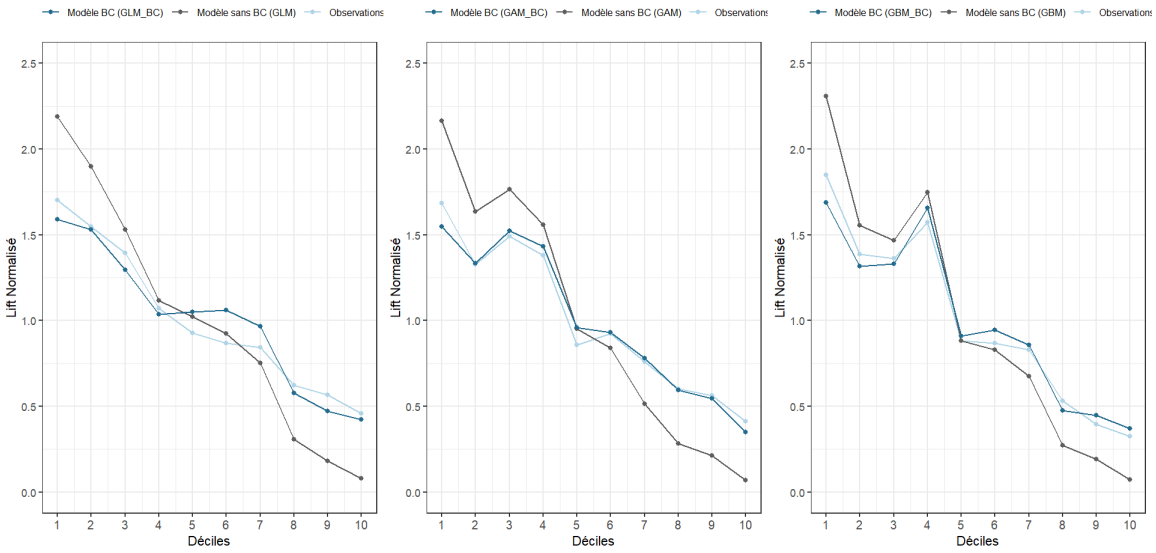

FIGURE 3.4 – Courbe de lift courbe pour les modèles GLM, GAM et GBM.

Les courbes de lift double de la figure 3.4 montrent un rapprochement significatif des valeurs prédites après l'autocalibration vers les valeurs réelles. Ce phénomène est particulièrement visible dans les déciles supérieurs, où les valeurs observées sont nettement un peu plus proches des valeurs prédites après l'autocalibration.

Le rapprochement des valeurs autocalibrées vers les valeurs réelles indique que l'autocalibration a permis d'améliorer la précision prédictive des modèles actuariels. Avant l'autocalibration, les modèles pouvaient présenter des écarts importants entre les valeurs prédites et les valeurs réelles, ce qui peut affecter la capacité du modèle à distinguer les risques favorables des risques défavorables. Cependant, grâce à l'autocalibration, ces écarts ont été réduits, ce qui indique que les modèles sont désormais plus proches de la réalité et mieux adaptés aux caractéristiques des données réelles. Cette amélioration du rapprochement entre les valeurs prédites et réelles est un indicateur positif de l'efficacité de l'autocalibration pour restaurer l'équilibre local et global dans les modèles actuariels.

Les courbes de lift double montrent clairement que l'autocalibration a permis d'améliorer la performance prédictive des modèles en les rapprochant davantage des valeurs réelles. Cette amélioration renforce la capacité des modèles à prendre des décisions plus précises en matière de tarification et de gestion des risques, ce qui peut avoir un impact significatif sur les résultats financiers et opérationnels des compagnies d'assurance. Ces résultats confirment l'importance de l'autocalibration en tant qu'outil d'amélioration des performances des modèles actuariels et soulignent son utilité dans le domaine de l'assurance. Ces résultats de cette analyse confirment l'efficacité de l'autocalibration pour restaurer l'équilibre des modèles d'assurance et améliorer leurs performances. Par exemple, l'autocalibration a permis d'améliorer la précision prédictive du modèle GAM avec une différence de déviance positive de 0.011361 pour les données de test. Ces résultats sont en accord avec les travaux de Denuit et al. (2021) et ceux de Ciatto et al. (2022). L'utilisation de l'autocalibration peut donc être bénéfique pour améliorer la précision des prévisions et renforcer l'utilisation de modèles plus flexibles et sophistiqués dans le domaine de l'assurance.

CONCLUSION

En conclusion, cette étude a mis en évidence l'importance cruciale de l'autocalibration dans l'amélioration des performances prédictives des modèles de tarification. En se concentrant sur les Modèles Linéaires Généralisés, les Modèles Additifs Généralisés et les Machines à Gradient Boosting, nous avons démontré de manière probante que l'autocalibration peut considérablement améliorer la performance de ces modèles.

L'analyse des résultats montre que l'autocalibration conduit à une meilleure adéquation des modèles aux données réelles en réduisant les valeurs de déviance. Cela se traduit par des prédictions plus précises et une meilleure évaluation des risques. De plus, les courbes de lift simples et doubles démontrent que l'autocalibration renforce la capacité des modèles à distinguer entre les différentes classes de risques, ce qui est essentiel pour une tarification précise et une gestion optimale des portefeuilles d'assurance.

Toutefois, il convient de noter que l'efficacité de l'autocalibration dépend en grande partie de la sélection appropriée des paramètres de lissage. Une approche méthodique est nécessaire pour en tirer pleinement parti. De plus, bien que les résultats soient prometteurs, il est crucial de poursuivre les recherches pour évaluer l'applicabilité de l'autocalibration dans des contextes plus diversifiés et pour explorer ses limites potentielles.

En définitive, tout comme l'on montré Denuit et al. (2021), l'autocalibration se révèle être un outil puissant pour affiner les prédictions et renforcer la prise de décision. En intégrant cette approche dans les processus actuariels, les entreprises d'assurance peuvent espérer obtenir des évaluations plus précises des risques, ce qui se traduira par des décisions plus informées et une meilleure gestion des portefeuilles.

RÉFÉRENCES

Bühlmann, H., & Gisler, A. (2005). *A course in credibility theory and its applications*. Springer.

Ciatto, N., Verelst, H., Trufin, J., & Denuit, M. (2022). *Does autocalibration improve goodness of lift ?*

Denuit, M., Charpentier, A., & Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance : Mathematics and Economics*, *101*, 485–497.

Denuit, M., Hainaut, D., & Trufin, J. (2019). *Effective statistical learning methods for actuaries i : Glms and extensions*. Springer.

Denuit, M., Hainaut, D., & Trufin, J. (2020). *Effective statistical learning methods for actuaries ii : Tree-based methods and extensions*. Springer.

Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2019). *Generalized linear models for insurance rating*. Arlington, VA : Casualty Actuarial Society.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297-318.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : with applications in r* (Illustrated éd., Vol. 103). Springer New York.

Kruger, F., & Ziegel, J. F. (2020). Generic conditions for forecast dominance. *Journal of Business Economic Statistic*. (In press)

Loader, C. (1999). *Local regression and likelihood*. New York : Springer.

Tevet, D., & Khare, A. (2013). *Introduction to predictive modeling using glms a practitioner's viewpoint*.

Trufin, J. (2022-2023). *Actuf506 : Assurance non-vie 2 - partie 1 : Tarification a priori*. Course. (Université Libre de Bruxelles, ULB)

Wood, S. (2001, June). Mgcv : GAMs and generalized ridge regression for R. *R News*, *1*(2). Consulté sur https://journal.r-project.org/articles/RN-2001-015/

Wood, S. (2006). *Generalized additive models : An introduction with r* (1st éd.). Chapman & Hall/CRC.

Wuthrich, M. V., & Buser, C. (2023, June). *Data analytics for non-life insurance pricing*. Consulté sur https://ssrn.com/abstract=2870308 (Swiss Finance Institute Research Paper No. 16-68) doi: 10.2139/ssrn.2870308

CODE SOURCE

Le code source de mon mémoire est disponible sur ma page GitHub à l'adresse suivante : https://github.com/JohnGAB7/Master_Thesis_2023.

J'aimerai encore rappeler ici qu'une grande partie du code, notamment la partie portant sur l'autocalibration, a été construite en utilisant le référentiel GitHub du Professeur Arthur Charpentier disponible à l'adresse : https://github.com/freakonometrics/autocalibration/.