Mémoire présenté devant l'Université de Paris-Dauphine
pour l'obtention du Certificat d'Actuaire de Paris-Dauphine
et l'admission à l'Institut des Actuaires
le 2 décembre 2022

Par : Markéta KRÚPOVÁ
Titre : Construction d'un modèle de Machine Learning interprétable pour la tarification en assurance non-vie

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

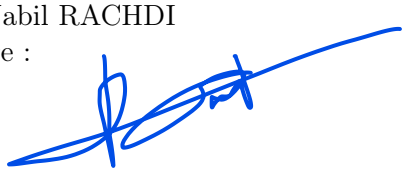
Les signataires s'engagent à respecter la confidentialité ci-dessus

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : ADDACTIS France
Signature :

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

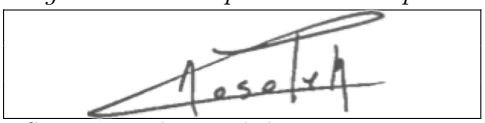
Directeur de Mémoire en entreprise :
Nom : Nabil RACHDI
Signature :



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

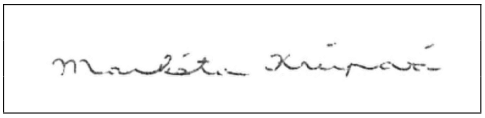
Secrétariat :

Signature du responsable entreprise



Bibliothèque :

Signature du candidat



Résumé

Dans un contexte d'accroissement perpétuel de la concurrence et d'une pression réglementaire très forte, la justesse et la précision actuarielle du tarif sont des enjeux clefs pour le secteur de l'assurance non-vie. Les modèles linéaires généralisés traditionnellement utilisés donnent lieu à un tarif multiplicatif favorisant l'interprétabilité et l'applicabilité opérationnelle. D'autre part, les méthodes et algorithmes issus de l'Apprentissage Statistique, ou *Machine Learning*, se focalisent sur un objectif prédictif et permettent de construire des modèles ayant une précision accrue et une forte capacité à capturer des non-linéarités inhérentes à la donnée. La performance de ces modèles se heurte cependant à leur complexité et leur faible interprétabilité. Au-delà d'une analyse séparée des modèles linéaires généralisés et des modèles de *Machine Learning*, l'objectif de ce mémoire est de construire une méthode exploitant conjointement les avantages de ces deux approches, un modèle atteignant un fort pouvoir prédictif tout en conservant une interprétabilité intrinsèque. Partant des données d'un assureur, nous appliquons alors le modèle additif généralisé dont nous exploitons la structure additive combinant habituellement des fonctions splines pour introduire des fonctions composantes incorporant des modèles de *Machine Learning*. Au lieu d'interpréter un modèle de *Machine Learning*, nous construisons ainsi un modèle de *Machine Learning* intrinsèquement interprétable, un modèle *Glass Box*. Nous montrons l'intérêt de ce type de modélisation dans la tarification en assurance non-vie.

Mots-clés : *Assurance non-vie*, *Modèle linéaire généralisé*, *Modèle additif généralisé*, *Machine Learning*, *Interprétabilité*, *Modèle Glass Box*.

Abstract

In the face of the ever-increasing competition and mounting regulatory pressure, actuarial precision and accuracy shape the art of setting the price in the non-life insurance sector. Generalized linear models are the standard method in pricing of non-life insurance products, leading to a multiplicative tariff that is immediately interpretable and operationally efficient. In recent years, the advent of Machine Learning has been termed the next frontier of innovation and productivity, focusing on predictive power and capturing the inherent non-linearity of the data. However, the high accuracy of complex models comes at the cost of interpretability. Beyond the separate analysis of the generalized linear models and Machine Learning methods, the aim of this master thesis is to bring these two classes of models with their complementing properties together. In other words, the objective is to propose a model that achieves high predictive accuracy while retaining much of the intelligibility of linear models. Working with insurer data, we apply the generalized additive model which explicitly decomposes a complex function into one-dimensional building-blocks, usually the spline functions, and we extend the framework to Machine Learning components. Rather than interpreting Machine Learning models, we construct models that are intrinsically interpretable, the Glass Box models. Our experiments show the potential of this approach in the field of non-life insurance pricing.

Keywords : *Non-life insurance*, *Generalized linear model*, *Generalized additive model*, *Machine Learning*, *Interpretability*, *Glass Box model*

Note de Synthèse

Introduction

Le marché de l'assurance non-vie est caractérisé par des exigences réglementaires et techniques fortes. L'optimisation tarifaire est alors au coeur des travaux de tarification structurant l'intégralité de la chaîne de valeur. Avec une structure paramétrique adaptée à la nature des données assurantielles, les modèles linéaires généralisés constituent aujourd'hui la référence pour la tarification. Si les modèles de *Machine Learning* offrent un meilleur éclairage des risques sous-jacents, leur manque d'interprétabilité exclut une utilisation plus systématique. Dans une vision plus inclusive et collaborative entre les deux paradigmes de modélisation, nous proposons alors un modèle de *Machine Learning* qui intègre la notion d'interprétabilité de manière intrinsèque (Figure 1). Pour cela nous nous appuyons sur la modélisation additive généralisée, combinant traditionnellement des fonctions splines, et nous introduisons des composantes univariées des modèles de *Machine Learning*.

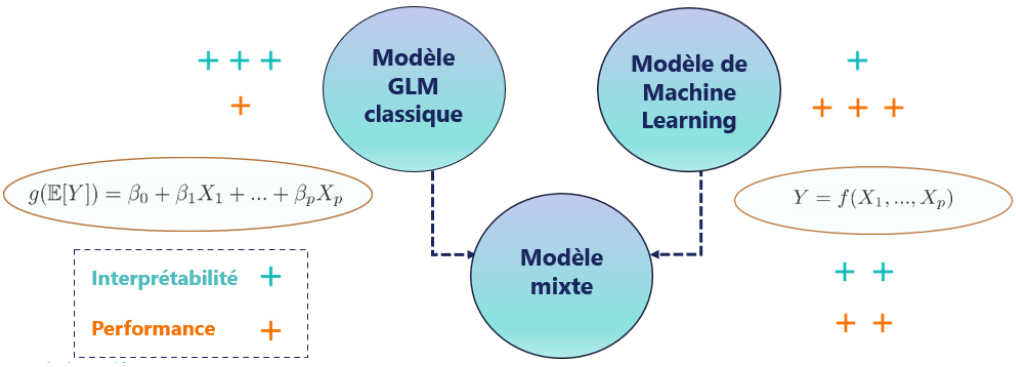


FIGURE 1 – Positionnement du problème : construire un modèle à la fois performant et interprétable.

Dans le cadre des travaux de ce mémoire, nous nous appuyons sur les données d'un assureur et nous nous focalisons sur la modélisation de la garantie DTA (dommage tout accident) en assurance automobile.

Modèle linéaire généralisé

Dans un modèle linéaire généralisé, l'espérance de la variable réponse est modélisée en fonction d'une combinaison linéaire des variables explicatives, quantifiant indépendamment l'effet de chaque prédicteur

$$g(\mathbb{E}[Y_i]) = \underbrace{\beta_0}_{\text{effet moyen}} + \underbrace{\beta_1 X_{i,1}}_{\text{effet } X_1} + \underbrace{\beta_2 X_{i,2}}_{\text{effet } X_2} + \ldots + \underbrace{\beta_p X_{i,p}}_{\text{effet } X_p}, \; i = 1, 2, \ldots, n.$$

La loi Gamma est particulièrement adaptée aux distributions asymétriques à droite comme des coûts de sinistres. Pour la fréquence, la loi de Poisson permet de modéliser un phénomène de comptage.

L'utilisation de la fonction de lien logarithmique dans le cas de la distribution Gamma et Poisson conduit à un modèle multiplicatif

$$\mathbb{E}[Y_i] = \exp(\beta_0) \times \exp(\beta_1 X_{i,1}) \times \exp(\beta_2 X_{i,2}) \times \dots \exp(\beta_p X_{i,p}),\ i = 1, 2, \dots, n.$$

L'estimation des paramètres $\beta_0, \dots, \beta_p$ s'effectue par la méthode du maximum de vraisemblance.

Modèles de Machine Learning

Les modèles de *Machine Learning* n'imposent pas une structure linéaire entre la variable à expliquer et les variables explicatives. Basées sur une procédure d'apprentissage plutôt que sur une procédure d'estimation, ils permettent une intégration efficace des non-linéarités inhérentes à la donnée tout comme une prise en compte des interactions entre les variables. Pour atteindre des performances de prédiction accrues, les modèles de *Machine Learning* s'appuient souvent sur des architectures sophistiquées, notamment des structures arborescentes et neuronales. Parmi les algorithmes célèbres, nous pouvons citer les arbres de décision, la forêt aléatoire (*Random Forest*), l'algorithme *Extreme Gradient Boosting* ou le perceptron multicouche. Un principe central en apprentissage statistique est l'agrégation des modèles, avec notamment :

- la procédure de *Bagging*, un apprentissage indépendant puis une moyennisation des résultats ;
- la procédure de *Boosting*, un apprentissage adaptatif et récursif.

Si les modèles de *Machine Learning* permettent de modéliser la sinistralité dans sa complexité en proposant une segmentation du risque très précise, leur principal défaut reste leur manque d'interprétabilité. Ce statut de boîte noire les rend inadaptés aux contraintes opérationnelles de la tarification en assurance non-vie.

Modèle additif généralisé et ses déclinaisons

Les modèles additifs généralisés sont une extension des modèles linéaires généralisés. L'espérance de la variable réponse est modélisée en fonction d'une combinaison linéaire des variables explicatives, mais la contribution individuelle de chaque prédicteur se trouve complexifiée. En effet, au lieu des coefficients constants $\beta_j,\ j = 1, \dots, p$, les variables explicatives participent à la prédiction au moyen des fonctions composantes non-paramétriques $f_j,\ j = 1, \dots, p$

$$g(\mathbb{E}[Y_i]) = \underbrace{\beta_0}_{\text{effet moyen}} + \underbrace{f_1(X_{i,1})}_{\text{effet } X_1} + \underbrace{f_2(X_{i,2})}_{\text{effet } X_2} + \dots + \underbrace{f_p(X_{i,p})}_{\text{effet } X_p},\ i = 1, 2, \dots, n.$$

Le maintien de la structure additive intrinsèquement interprétable permet un ajout parcimonieux de la non-linéarité dans le modèle. Traditionnellement, ce sont des fonctions de lissage, notamment des fonctions splines, qui sont utilisées comme fonctions composantes dans un modèle additif généralisé. Réécrites dans une base adaptée, les fonctions composantes se mettent alors sous la forme

$$f_j(x_j) = \sum_{q=1}^{Q_j} \beta_{jq} b_{jq}(x_j),\ j = 1, \dots, p.$$

Le modèle additif généralisé combinant des fonctions splines peut ainsi s'interpréter comme un modèle linéaire généralisé hyperparamétré. Par la suite, l'estimation des coefficients est basée sur la méthode du maximum de vraisemblance, avec une pénalisation supplémentaire pour promouvoir l'aspect lisse des fonctions composantes. Cette estimation s'effectuant de manière simultanée, la capacité du modèle

à extraire de manière fine le comportement de la donnée s'en trouve limitée.

Dans ce mémoire, nous souhaitons nous appuyer sur une procédure d'apprentissage plutôt que sur une procédure d'estimation, apprendre l'information au lieu de la lisser. À ce titre, nous introduisons des composantes univariées des modèles de *Machine Learning* dans un modèle additif généralisé (Figure 2).

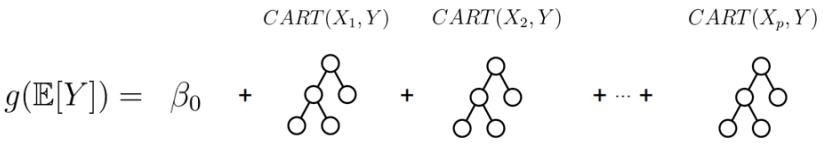


FIGURE 2 – Exemple d'un modèle additif généralisé combinant des arbres de décision

Les travaux de ce mémoire portent essentiellement sur les modèles additifs généralisés ayant une structure arborescente. Les arbres de décision CART sont alors utilisés comme briques de base, puis une complexification de l'architecture s'opère par les procédures de *Bagging* et de *Boosting*. Pour garantir l'interprétabilité intrinsèque du modèle de *Machine Learning* ainsi développé, une procédure d'apprentissage spécifique est mise en oeuvre (Figure 3). L'aspect essentiel de la modélisation réside dans l'apprentissage cyclique des fonctions composantes. Plus précisément :

- dans une approche *Boosting*, chaque fonction composante est apprise sur les résidus de la précédente sous la forme d'un arbre de décision avec une logique d'amélioration récursive (lecture horizontale du modèle) ;
- dans une approche *Bagging* locale, les arbres appris au cours des différentes itérations sont agrégés indépendamment pour chaque variable explicative avec une logique d'amélioration itérative (lecture verticale du modèle) ;
- dans une approche *Bagging* globale, ce processus d'apprentissage est réitéré plusieurs fois pour plus de robustesse (lecture générale du modèle).

Pour l'implémentation, nous nous appuyons sur les travaux de l'équipe de recherche de Microsoft INTERPRETML TEAM (2021) avec l'algorithme *Explainable Boosting Machine* que nous adaptons pour permettre une meilleure adéquation à la structure des données assurantielles.

Application à la modélisation de la garantie DTA en assurance automobile

Selon le principe général de la tarification en assurance non-vie, la fréquence et la sévérité des sinistres sont estimés indépendamment. Si plusieurs modélisations sont proposées pour les deux volets, afin de répondre à l'objectif du mémoire, il s'agit de comparer essentiellement :

- le modèle linéaire généralisé (**GLM**), le modèle de référence ;
- un modèle de *Machine Learning*, ici le modèle *Extreme Gradient Boosting* (**XGB**), le modèle de performance ;
- le nouveau modèle développé dans ce mémoire, le modèle additif généralisé avec une structure arborescente (**GAM arbres**).

Pour le modèle linéaire généralisé, la mise en place d'un modèle multiplicatif permet de construire une grille tarifaire immédiatement interprétable et explicable. Pour les modèles additifs généralisés,

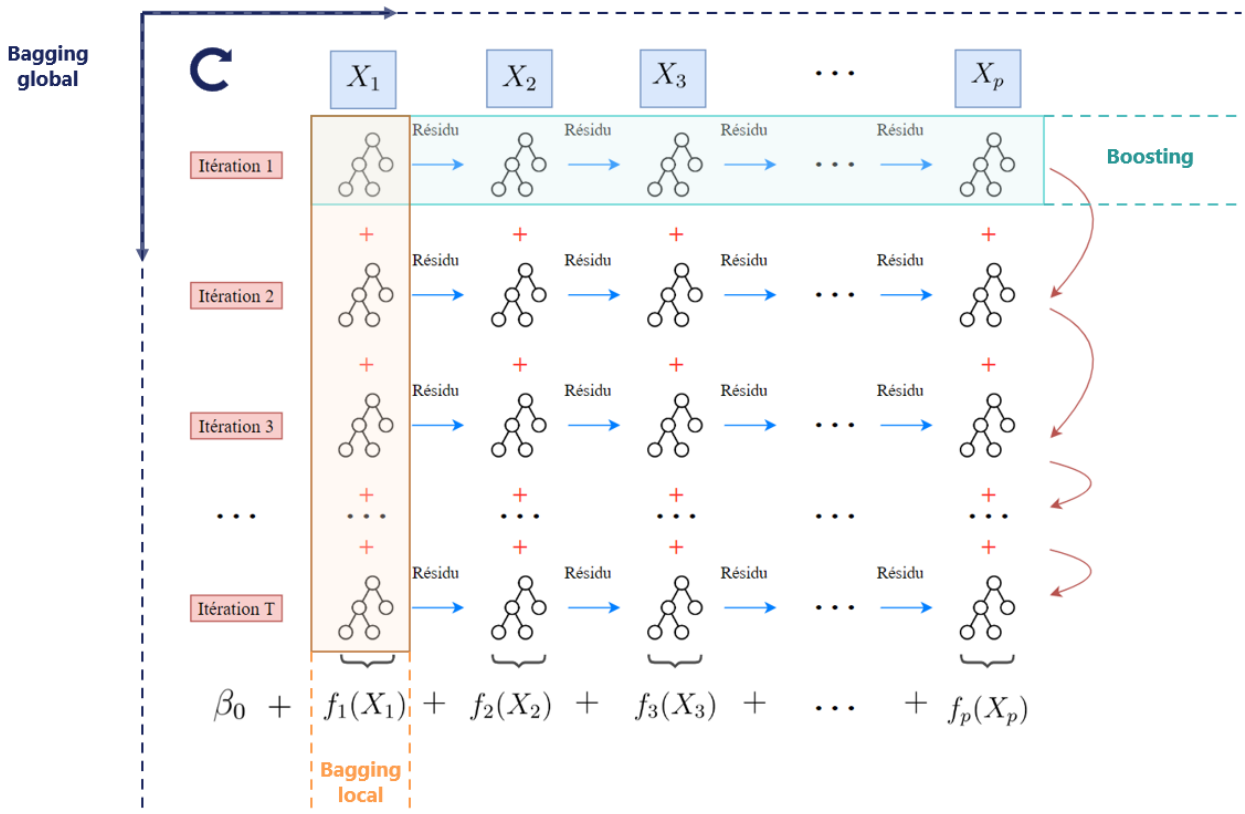


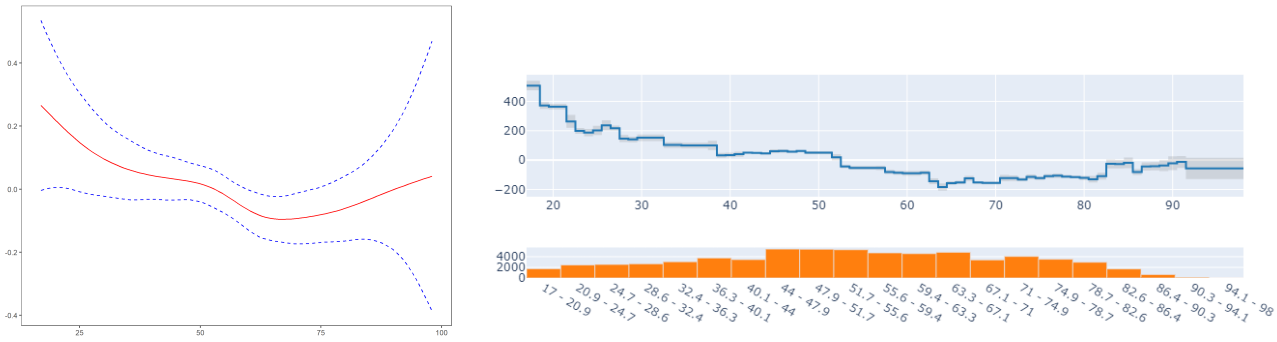
FIGURE 3 – Procédure d'apprentissage dans un modèle additif généralisé ayant une structure arborescente

la visualisation des fonctions composantes fournit une description exacte de comment est réalisée la prédiction au niveau de chaque variable explicative (Figure 4). Nous remarquons alors que l'utilisation des fonctions splines comme fonctions composantes conduit à une allure lisse et moyennisée. Les fonctions composantes obtenues à partir d'une structure arborescente permettent d'identifier avec précision le comportement de chaque profil de risque. Cependant, les variations observées ne sont pas toujours intuitives et il n'est pas aisé de leur associer systématiquement un sens actuariel. Un arbitrage est alors à faire entre précision et pertinence actuarielle des modèles.

Les métriques de performance globales telles que l'erreur quadratique moyenne, la RMSE, et l'erreur absolue moyenne, la MAE, opèrent un rapprochement entre des structures algorithmiques différentes et permettent donc une comparaison des modèles. Nous établissons ainsi que le modèle additif généralisé se positionne entre le modèle linéaire et le modèle de *Machine Learning* (Table 1).

Prédiction	**Coût moyen**		**Fréquence**	
Métrique	**RMSE**	**MAE**	**RMSE**	**MAE**
GLM	2027	1308	0,1382	0,0364
GAM arbres	2017	1304	0,1380	0,0364
XGB	2016	1293	0,1380	0,0364

TABLE 1 – Résultats de comparaison des modèles pour la modélisation du coût moyen et de la fréquence sur un échantillon de test : GLM, GAM arbres avec une structure arborescente et *Explainable Boosting Machine*.

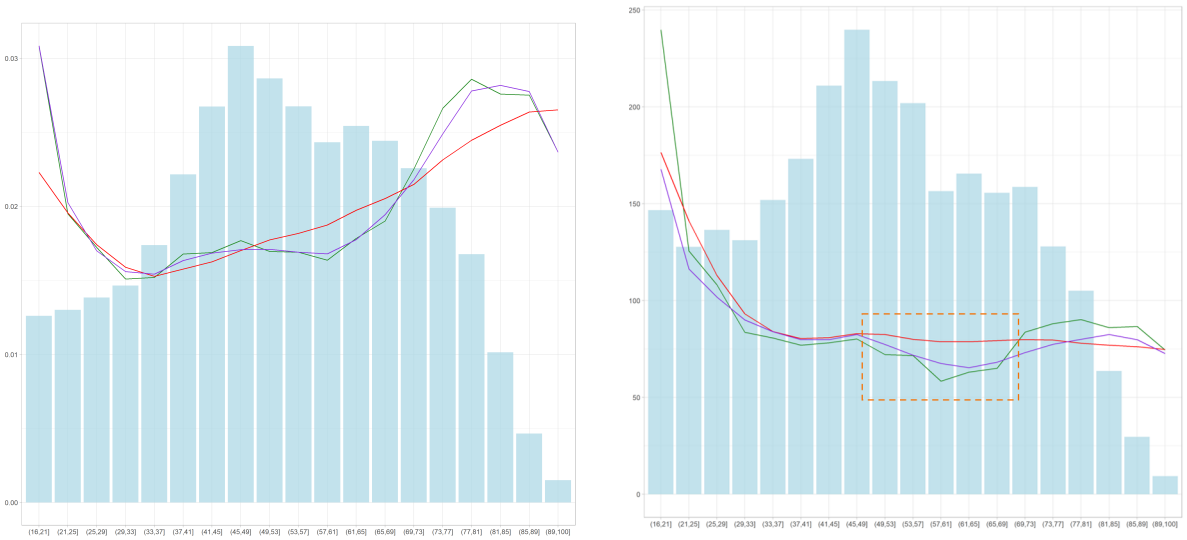


(a) spline (b) structure arborescente

FIGURE 4 – Fonction composante associée à la variable âge conducteur dans le modèle additif généralisé

Pour avoir une vision locale des performances, nous nous intéressons aux graphes des valeurs observées et prédites variable par variable (Figure 5). Nous observons alors que le modèle linéaire généralisé a tendance à moyenniser l'information, jusqu'à omettre des détails importants dans le comportement du risque. Le modèle additif généralisé avec une structure arborescente s'ajuste mieux à la courbe des valeurs observées en captant plus efficacement la volatilité des données (Figure 5a).

La finalité de la modélisation séparée du coût moyen et de la fréquence est la modélisation de la prime pure. Nous comparons alors le risque observé, représenté par le *Burning Cost*, au risque prédit par le modèle linéaire généralisé d'une part, et le modèle additif généralisé avec une structure arborescente d'autre part (Figure 5b). Nous remarquons de nouveau une tendance moyennisante pour le modèle linéaire généralisé et un meilleur ajustement à la volatilité du risque observé pour le modèle additif.



(a) fréquence (b) *Burning Cost*

FIGURE 5 – Risque observé et prédit pour la variable âge conducteur sur un échantillon de test avec les valeurs observées en vert, le modèle GLM en rouge et le modèle GAM arbres en violet.

Dans une perspective d'optimisation tarifaire, nous analysons la variable δ définie par

$$\delta = \frac{\text{PP EBM}}{\text{PP GLM}}.$$

Ainsi, nous quantifions et qualifions les écarts entre la prime pure prédite par le modèle linéaire généralisé (PP GLM) et celle prédite par le modèle additif généralisé ayant une structure arborescente (PP EBM). Dans une approche descriptive puis explicative, nous identifions une surestimation du risque par rapport au modèle de référence sur les profils des jeunes conducteurs peu expérimentés. En revanche, sur les profils des conducteurs confirmés de 30 à 60 ans environ, nous observons une sous-estimation du risque par rapport au modèle linéaire généralisé.

Conclusion

Dans ce mémoire, nous avons proposé un modèle à la frontière des modèles linéaires généralisés et des modèles de *Machine Learning*, un modèle qui intègre la notion de l'apprentissage tout en restant interprétable. Les contributions principales de ce travail de modélisation sont alors les suivantes :

- la formalisation mathématique de la mécanique d'apprentissage garantissant l'interprétabilité intrinsèque d'un modèle de *Machine Learning*, l'illustration du processus avec de nombreux schémas et l'établissement de liens avec les autres méthodes de *Machine Learning* ;
- l'application du modèle additif généralisé avec une structure arborescente à la tarification en assurance non-vie, plus précisément à la modélisation de la garantie DTA en assurance automobile ;
- la déclinaison de l'algorithmie *Explainable Boosting Machine* existante pour répondre aux contraintes de la tarification imposant une structure multiplicative en mettant en oeuvre deux approches correctives :
  - une approche *ex ante* consistant à exploiter le méta-algorithme *Cyclic Boosting Machine* (Wick et al. (2019)) pour l'apprentissage des fonctions composantes ;
  - une approche *ex post* basée sur la modification de la variable à prédire et une correction du biais de prédiction.
- l'analyse critique des résultats et l'identification des profils intéressants pour l'assureur en vue d'une optimisation tarifaire.

Limites de l'étude et ouvertures

Parmi les principales limites et perspectives de cette étude, nous comptons :

- l'algorithmie *Explainable Boosting Machine* existante ne permettant pas l'utilisation de bonnes fonctions de lien et distributions dans le modèle additif généralisé, et donc la nécessité de mettre en place des approches correctives ;
- l'optimisation de l'ordre dans lequel les variables explicatives sont apprises dans la modèle additif généralisé combinant des composantes univariées des modèles de *Machine Learning* (Figure 3), cet ordre étant potentiellement significatif pour une meilleure prise en compte des liens de causalité entre les variables ;
- l'investigation plus détaillée de l'approche par réseaux de neurones dans la modélisation additive généralisée ;
- la vision univariée proposée dans ce mémoire : une intégration parcimonieuse des interactions du second ordre pourrait enrichir le modèle, sans en dégrader l'interprétabilité ;
- la définition des métriques de performance plus adaptées à la nature des données assurantielles ainsi que des indicateurs permettant une quantification et une qualification efficace des écarts de comportement entre les modèles ;
- l'extension de l'étude à d'autres garanties, notamment des garanties plus volatiles comme la garantie vol ou la garantie responsabilité civile.

Synthesis note

Introduction

The non-life insurance sector is characterized by strong regulatory and technical requirements. Tariff optimization is therefore at the heart of pricing issues and structures the entire value chain. With a parametric structure adapted to the nature of the insurance data, generalized linear models are the standart reference for pricing. If Machine Learning models offer a better insight into the underlying risks, their lack of interpretability precludes their more systematic use. In a more inclusive and collaborative vision between the two modeling paradigms, we propose a Machine Learning model that intrinsically integrates the notion of interpretability. To do so, we rely on generalized additive modeling, traditionally combining spline functions, and we introduce univariate components of Machine Learning models.

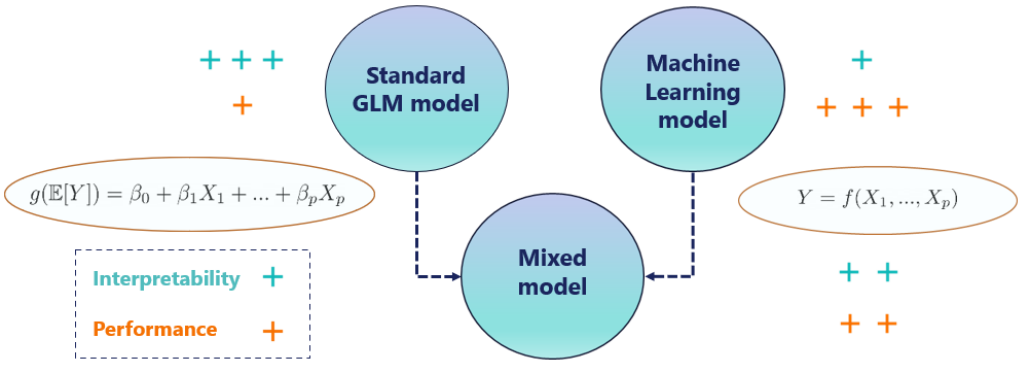


Figure 6 – Problem positioning : construct a model both performant interpretable model.

In this masetr thesis, we use data from an insurer and focus on the modeling of the own damage coverage in automobile insurance.

Generalized linear models

In a generalized linear model, the expectation of the target variable is modelled by a linear function of explanatory variables, independently quantifying the effect of each predictor

$$g(\mathbb{E}[Y_i]) = \underbrace{\beta_0}_{\text{global average}} + \underbrace{\beta_1 X_{i,1}}_{X_1 \text{ effect}} + \underbrace{\beta_2 X_{i,2}}_{X_2 \text{ effect}} + \ldots + \underbrace{\beta_p X_{i,p}}_{X_p \text{ effect}}, \; i = 1, 2, \ldots, n.$$

The Gamma distribution is particularly suitable for right-skewed distributions such as claim costs. For frequency, the Poisson distribution allows to model a counting phenomenon. The use of the logarithmic link function in the case of the Gamma and Poisson distributions leads to a multiplicative model

$$\mathbb{E}[Y_i] = \exp(\beta_0) \times \exp(\beta_1 X_{i,1}) \times \exp(\beta_2 X_{i,2}) \times \ldots \exp(\beta_p X_{i,p}),\ i = 1, 2, \ldots, n.$$

The parameters $\beta_0, \ldots, \beta_p$ are estimated by the maximum likelihood method.

Machine Learning models

Machine Learning models do not impose a linear dependency between the target and the explanatory variables. Based on a learning strategy rather than an estimation procedure, they allow an efficient integration of both the non-linearities inherent to the data and the interactions between the variables. In order to achieve high predictive accuracy, Machine Learning models often rely on sophisticated architectures, such as tree and neural ensembles. Decision tree, Random Forest, Extreme Gradient Boosting and multilayer perceptron rank among the most popular algorithms. Model aggregation is a central principle in statistical learning, especially with:

- the *Bagging* approach, an independent learning followed by an averaging procedure ;
- the *Boosting* approach, an adaptative and recursive learning.

Despite their ability to model the claims in all their complexity by proposing a very precise segmentation of the risk, Machine Learning models give rise to a significant drawback : the lack of interpretability. This black box status renders them unsuitable for the operational constraints of non-life insurance pricing.

GAM modelling and its declinations

Generalized additive models are an extension of generalized linear models. The expectation of the response variable is modelled by a linear function of explanatory variables, but the individual contribution of each predictor is complexified. Instead of the constant coefficients $\beta_j,\ j = 1, \ldots, p$, the explanatory variables participate in the prediction as non-parametric shape functions $f_j,\ j = 1, \ldots, p$

$$g(\mathbb{E}[Y_i]) = \underbrace{\beta_0}_{\text{global average}} + \underbrace{f_1(X_{i,1})}_{X_1 \text{ effect}} + \underbrace{f_2(X_{i,2})}_{X_2 \text{ effect}} + \ldots + \underbrace{f_p(X_{i,p})}_{X_p \text{ effect}},\ i = 1, 2, \ldots, n.$$

Maintaining the inherently interpretable additive structure allows for a parsimonious addition of nonlinearity into the model. Smoothing parameters such as spline functions traditionally form the shape functions of a generalized additive model. Rewritten in an adapted basis, the shape functions can be viewed as

$$f_j(x_j) = \sum_{q=1}^{Q_j} \beta_{jq} b_{jq}(x_j),\ j = 1, \ldots, p.$$

The generalized additive model combining spline shape functions can thus be interpreted as a hyper-parameterized generalized linear model. Subsequently, the estimation of the coefficients is based on the maximum likelihood method, with an additional penalty to promote the smoothness of the shape functions. Since this estimation is performed simultaneously, the ability of the model to extract the behavior of the data in a fine-grained manner is limited.

In this study, we aim to promote a learning strategy rather than an estimation procedure, learn the information instead of smoothing it. As such, we introduce univariate components of Machine Learning algorithms in a generalized additive model.

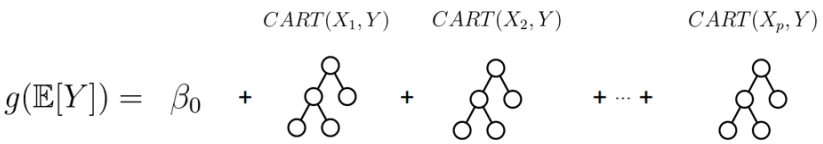


Figure 7 – Exemple of a generalized additive model combining decision trees

In this master thesis, we focus on generalized additive models with a tree structure. The CART decision trees are used as building blocks, then a complexification of the architecture is carried out by the Bagging and Boosting procedures. To maintain the intrinsic interpretability of the Machine Learning model thus developed, a specific learning procedure is implemented (Figure 8). This modelling strategy essentially revolves around the cyclic learning of the shape functions. More precisely:

- within a *Boosting* approach, each shape function is learned from the residuals of the previous one in the form of a decision tree with a recursive improvement logic (horizontal reading of the model) ;
- within a local *Bagging* approach, the trees learned during the different iterations are aggregated independently for each explanatory variable with an iterative improvement logic (vertical reading of the model) ;
- within a global *Bagging* approach, this learning process is repeated several times for more robustness (general reading of the model).

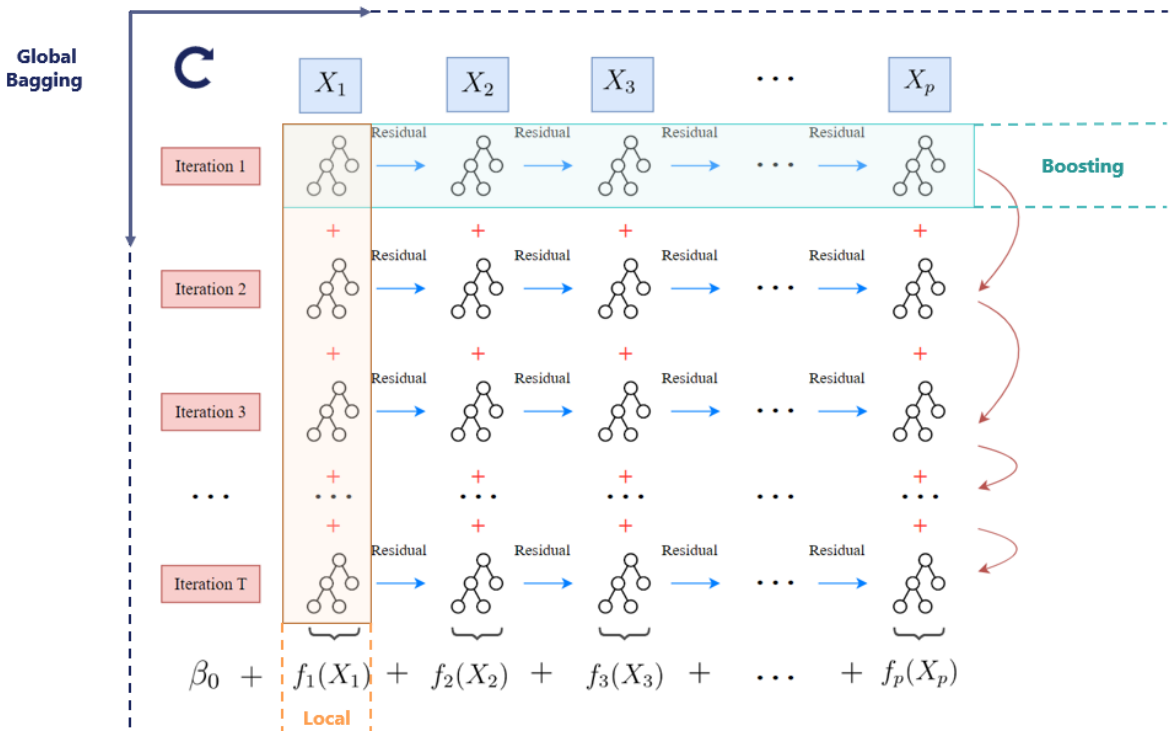


Figure 8 – Learning procedure in a generalized additive model with a tree structure

For the implementation, we rely on the work of the Microsoft Research team InterpretML Team (2021)

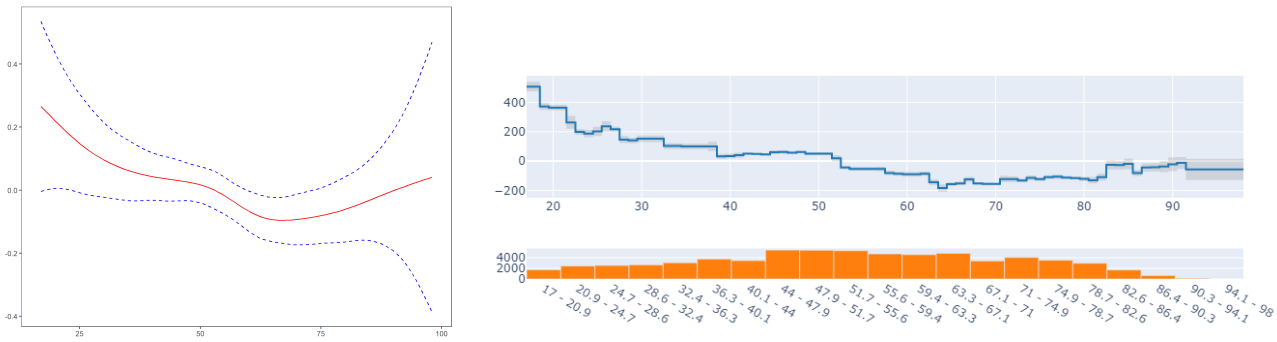
with the algorithm *Explainable Boosting Machine* that we adapt to allow a better fit to the specific structure of the insurance data.

Application to the pricing of the own damage coverage

According to the general principle of non-life insurance pricing, the frequency and severity of claims are estimated independently. While several models are proposed for the two components, in order to meet the goal of this master thesis, we mainly focus on comparing:

- the generalized linear model (**GLM**), the standard model ;
- a Machine Learning model, in this case the *Extreme Gradient Boosting* algorithm (**XGB**), the performance model ;
- the new model developped during this master thesis, the generalized additive model with a tree structure (**GAM tree**).

For the generalized linear model, the use of a multiplicative structure allows us to construct an immediately interpretable and explainable tariff grid. For the generalized additive models, the visualization of the shape functions provides an exact description of how the prediction is made at the level of each explanatory variable (Figure 9). A smooth and regularized output can be obtained with spline functions as shape functions. With tree shape functions, we can precisely identify the behavior of each risk profile. However, the variations observed are not always intuitive and it is not straightforward to systematically embed them with an actuarial interpretation. A trade-off must therefore be made between precision and actuarial relevance of the models.



(a) spline

(b) tree structure

Figure 9 – Shape function of the driver age variable in a generalized linear model

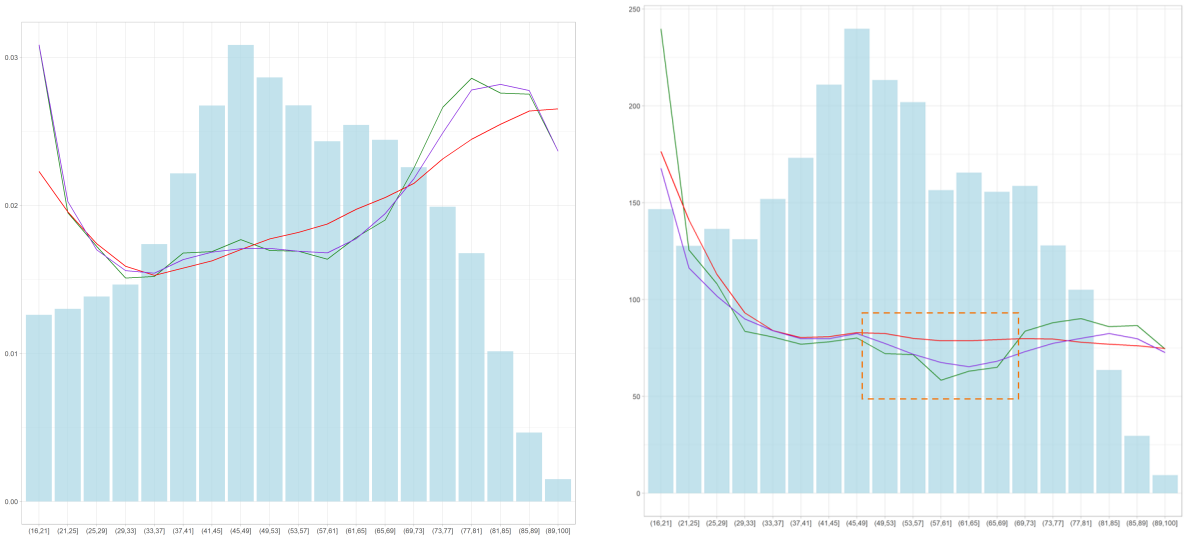
Global performance metrics such as root mean square error, RMSE, and mean absolute error, MAE, reconcile different algorithmic structures and thus allow a comparison of the models. We establish that the generalized additive model positions itself between the linear model and the *Machine Learning* model (Table 2).

Prediction	**Severity**		**Frequency**	
Metric	**RMSE**	**MAE**	**RMSE**	**MAE**
GLM	2027	1308	0,1382	0,0364
GAM tree	2017	1304	0,1380	0,0364
XGB	2016	1293	0,1380	0,0364

Table 2 – Comparison of results for the modelling of frequency and severity on a test sample : GLM, GAM with a tree structure and *Explainable Boosting Machine*.

In order to have a more local vision of the algorithm performances, we construct graphs of the observed and predicted values variable by variable (Figure 10). We observe that the generalized linear model tends to average the information, to the point of omitting important details in the behavior of the risk. The generalized additive model with a tree structure fits the observed value curve better by capturing the volatility of the data more effectively (Figure 10a).

The purpose of considering the frequency and severity of claims separately is to model the pure premium. Subsequently, we compare the observed risk, represented by the Burning Cost, to the risk predicted by the generalized linear model on the one hand, and the generalized additive model with a tree structure on the other hand (Figure 10b). We notice an averaging trend for the generalized linear model and a better fit to the volatility of the observed risk for the generalized additive model.



(a) fréquence (b) *Burning Cost*

Figure 10 – Observed and predicted risk for the driver age variable on a test sample with the observed values in green, the GLM model in red and the GAM tree model in purple.

With a tariff optimization objective, we analyze the variable δ defined by

$$\delta = \frac{\text{PP EBM}}{\text{PP GLM}}.$$

We quantify and qualify the differences between the pure premium predicted by the generalized linear model (PP GLM) and the one predicted by the generalized additive model with a tree structure (PP EBM). In a descriptive, then explanatory approach, we identify an overestimation of the risk compared to the reference model on the profiles of young drivers with little seniority. Conversely, on the profiles of experienced drivers in the age span 30 to 60 years, we observe an underestimation of the risk compared to the generalized linear model.

Conclusion

In this master thesis, we proposed a model at the frontier of generalized linear models and Machine Learning models, a model that integrates the notion of learning while remaining interpretable. The main contributions of this modeling work are thus the following:

- the mathematical formalization of the learning mechanism ensuring the intrinsic interpretability of a Machine Learning model, the illustration of the procedure with numerous diagrams and the establishment of links with other Machine Learning methods ;
- the application of the generalized additive model with a tree structure to non-life insurance pricing, more precisely to the modeling of the own damage coverage in automobile insurance ;
- the adaptation of the existing Explainable Boosting Machine algorithm to meet the insurance pricing constraints imposing a multiplicative structure, the implementation two corrective approaches:
  — an *ex ante* approach consisting of using the meta-algorithm Cyclic Boosting Machine (Wick et al. (2019)) for the learning of shape functions ;
  — an *ex post* approach based on the modification of the target variable and the subsequent correction of the prediction bias.
- critical analysis of the results and identification of profiles of interest to the insurer with a tariff optimization goal.

Study limitations and research perspectives

Key limitations and perspectives of this study include:

- the current version of the *Explainable Boosting Machine* algorithm not allowing the use of suitable link functions and distributions in the generalized additive modelling, and thus the need to consider corrective approaches ;
- the optimization of the order in which the explanatory variables are learned in the generalized additive model combining univariate components of the Machine Learning models (Figure 8), this order being potentially significant for a better integration of causal links between the variables ;
- a more detailed investigation of the neural network approach in the generalized additive modeling ;
- the univariate view proposed in this study: a parsimonious integration of second-order interactions could enrich the model, without degrading its interpretability ;
- the definition of performance metrics better adapted to the nature of insurance data, as well as of indicators allowing an efficient quantification and qualification of the behavior differences of the models ;
- the extension of the study to other coverages in automobile insurance, in particular more volatile coverages such as theft or civil liability.

Remerciements

Je tiens à adresser mes premiers remerciements à l'ensemble de l'équipe Pricing & Analytics P&C d'ADDACTIS France pour m'avoir accueilli, intégré et accompagné tout au long de mes travaux de mémoire.

Je remercie particulièrement Guillaume ROSOLEK pour l'opportunité qu'il m'a offerte en intégrant son équipe et pour ses conseils, actuariels et professionnels.

Je remercie également mon directeur de mémoire, Nabil RACHDI, pour son enthousiasme scientifique, son expertise en *Data Science*, son encadrement et la confiance qu'il m'a accordée.

Mes remerciements sont aussi destinés à Bilal SADOU et Pierre CHATELAIN pour leur précieuse aide technique, leur expertise actuarielle et leur disponibilité.

J'adresse mes sincères remerciements à l'ensemble des professeurs qui m'ont accompagné et guidé pendant mon parcours académique. Dans le cadre de ce mémoire, je souhaite remercier en particulier mon tuteur académique, Quentin GUIBERT, pour son suivi attentif et la pertinence de ses remarques.

Je remercie également l'équipe des stagiaires, des alternants et des consultants pour l'entraide et les bons moments partagés. Je remercie particulièrement Margaux REGNAULT et Mathilde ROCHELLE pour les relectures de mon mémoire.

Je remercie Samuel BOISADAM pour son aide et son soutien constants.

V neposlední řadě děkuji své rodině. Za všechno.

Table des matières

Introduction

Dans un marché fortement réglementé et concurrentiel tel que celui de l'assurance non-vie, les problématiques de tarification sont au coeur des travaux actuariels et structurent l'intégralité de la chaîne de valeur. Si les modèles linéaires généralisés (GLM) prédisent simplement et efficacement la fréquence et la sévérité des sinistres, donnant lieu à une équation tarifaire multiplicative, les méthodes issues de l'Apprentissage Statistique, ou *Machine Learning*, offrent un éclairage nouveau sur les risques sous-jacents ainsi qu'un meilleur pouvoir prédictif. Toujours plus précis et performants, ces algorithmes permettent de modéliser la sinistralité dans sa complexité en intégrant des interactions, des effets non-linéaires, voire des relations de causalité qu'un modèle linéaire moyennise ou passe sous silence.

Le processus de tarification en assurance non-vie impose de produire un tarif maintenable, pilotable et explicable à tout moment aux tiers, contraintes opérationnelles qu'un modèle de *Machine Learning* opaque et volatil ne saurait satisfaire. Pour le Conseil de Stabilité Financière (CSF), « le manque d'interprétabilité et d'auditabilité des méthodes d'Intelligence Artificielle et du *Machine Learning* pourrait devenir un risque macroéconomique[*] ». Autrement dit, la difficulté à expliquer de manière précise et consistante la décision automatique d'un algorithme constitue un frein à l'intelligibilité et l'explicabilité, insécables au secteur de la finance et de l'assurance. L'excédant de complexité déployé pour accéder à une compréhension plus fine du risque devient un risque en lui-même.

La nécessité d'un algorithme complexe est cependant inhérente à la complexité même du phénomène que l'on modélise. Dans la littérature actuarielle, la construction d'un modèle de *Machine Learning* s'établit souvent en parallèle d'un modèle GLM classique dans un objectif de maximisation de performance ou d'enrichissement *ex post* du modèle GLM par des interactions bien choisies. Ce mémoire ambitionne de proposer une vision inclusive et collaborative des modèles linéaires généralisés et des modèles de *Machine Learning*, une approche combinant la structure additive interprétable de la première configuration avec la complexité et la précision de la seconde. Au lieu d'interpréter un modèle de *Machine Learning*, nous tâcherons ainsi de construire un modèle de *Machine Learning* intrinsèquement interprétable. Pour cela, nous adopterons le cadre du modèle additif généralisé (GAM) qui ajoute de la non-linéarité dans une équation tarifaire multiplicative par l'introduction des fonctions splines. Partant de cette structure, nous intégrerons des modèles de *Machine Learning* comme fonctions composantes dans le modèle GAM, affinant ainsi la contribution individuelle de chaque prédicteur, mais limitant la complexité du modèle.

Après une première partie introductive sur les notions de l'assurance non-vie et la présentation des données (Chapitre 1), nous décrirons le modèle GLM comme le modèle référent du marché puis les modèles de *Machine Learning* comme les modèles de performance, exposant ainsi les principes et les limites des deux modélisations (Chapitre 2). Nous insisterons notamment sur les notions d'interprétabilité et d'explicabilité. Par la suite, nous détaillerons le cadre du modèle GAM, avec des fonctions splines comme fonctions composantes dans un premier temps et les modèles de *Machine Learning* dans

∗. traduction libre de l'anglais : "The lack of interpretability and auditability of Artificial Intelligence and Machine Learning methods could become a macro-level risk" (FSB (2017))

un second (Chapitre 2). Nous comparerons enfin quantitativement et qualitativement les différentes configurations proposées et mettrons en évidence l'intérêt du modèle GAM combinant des modèles de *Machine Learning* pour la modélisation de la prime pure (Chapitre 4).

Chapitre 1

Notions d'assurance IARD et présentation des données

L'objectif de ce premier chapitre est de présenter succinctement les principes fondamentaux d'une opération d'assurance pour exposer par la suite les spécificités de l'assurance IARD (Incendie, Accidents et Risques Divers), également appelée assurance dommages ou assurance non-vie. A travers la définition du cadre juridique, des principaux acteurs et des dynamiques afférentes au secteur, il s'agit de comprendre les enjeux particuliers de l'assurance automobile dans le but d'apporter une réponse pertinente et ciblée à la problématique de ce mémoire. La deuxième partie est consacrée à la présentation de la base de données, l'analyse exploratoire et les principaux retraitements. L'exposé théorique se base essentiellement sur les ouvrages Mathématiques de l'Assurance Non-Vie (Denuit et Charpentier (2005)) et Assurance IARD (Suru (2020)) et est complété par les chiffres actuels de l'Autorité de Contrôle Prudentiel et de Résolution (ACPR) et de France Assureurs.

1.1 La tarification en IARD

1.1.1 Principe de l'assurance dommage

Définitions juridiques et présentation générale

A défaut d'une définition juridique précise dans les Codes Civil, du Commerce ou même des Assurances, la vision qui est proposée est celle communément admise définissant l'assurance comme une opération par laquelle un organisme, l'assureur, s'engage à exécuter une prestation au profit d'une autre personne, le bénéficiaire, en cas de réalisation d'un événement aléatoire, le risque, en contrepartie du versement d'une somme d'argent par l'assuré, la prime. Les relations entre les parties sont définies par un contrat d'assurance.

Plusieurs acteurs interviennent ainsi dans une opération d'assurance :

- l'assureur qui couvre le risque et s'engage à payer la prestation en cas de réalisation du sinistre ;
- l'assuré qui court le risque et s'engage à payer la prime ;
- le bénéficiaire qui peut être l'assuré, une personne désignée ou la victime et qui reçoit la prestation ;

- éventuellement d'autres preneurs de risques à qui l'assureur cède le risque (réassureurs, marchés financiers, ...)

Contrairement à d'autres industries, l'assureur perçoit les primes avant de verser les prestations. Les coûts réalisés de production ne sont connus qu'a posteriori : on parle de cycle inversé de production. Pour évaluer le montant de la prime demandée à l'assuré, l'assureur doit donc utiliser le passé pour prédire le futur. Pour cela, il met en place des outils mathématiques sophistiqués, essentiellement statistiques et probabilistes, et s'appuie sur les données historiques pour évaluer la variabilité des risques futurs. Dans ce contexte, la pertinence, l'exhaustivité et l'exactitude des données apparaissent comme des critères essentiels pour la construction des modèles et une attention particulière doit être accordée à leur traitement.

La recherche de modèles probabilistes permettant une représentation adéquate des risques doit s'appuyer sur deux principes : la mutualisation et la segmentation. La mutualisation repose sur l'idée que l'agrégation d'un grand nombre d'événements incertains et indépendants aboutit à un risque dont l'incertitude est faible. Cependant, les risques ne sont pas homogènes en général et il y a besoin de les regrouper en groupes afin de pouvoir appliquer une prime différenciée à chacun d'entre eux : c'est le principe de la segmentation.

En France, les organismes d'assurance peuvent être des sociétés d'assurance, des mutuelles ou des instituts de prévoyance. Les sociétés d'assurance sont des sociétés anonymes ou à forme mutuelle. Elles sont régies par le Code des Assurances et sont adhérentes à la Fédération Française des Sociétés d'Assurance (FFSA). Les mutuelles et les instituts de prévoyance sont des organismes à but non lucratif relevant respectivement du Code de la Mutualité et du Code de la Sécurité Sociale. Ces trois types d'organismes sont soumis au contrôle unique de l'Autorité de Contrôle Prudentiel et de Résolution (ACPR), autorité administrative indépendante chargée de veiller à ce que les entités soumises à son contrôle soient en mesure de tenir à tout moment les engagements contractés envers les assurés ou les adhérents. L'article L310-1 du Code des Assurances distingue :

- les entreprises d'assurance vie qui « contractent des engagements dont l'exécution dépend de la durée de la vie humaine, s'engagent à verser un capital en cas de mariage ou naissance d'enfants, ou font appel à l'épargne en vue de la capitalisation et contractent à cet effet des engagements déterminés » ;
- les entreprises de prévoyance qui « sous forme d'assurance directe couvrent les risques de dommages corporels liés aux accidents et à la maladie » ;
- les entreprises d'assurance IARD qui « sous forme d'assurance directe couvrent d'autres risques y compris ceux liés à une activité d'assistance » ;

L'assurance IARD offre donc un moyen de protection contre les aléas ne relevant pas de la vie humaine. Elle a pour objet la protection du patrimoine de l'assuré en compensant les pertes financières liées à la dégradation de ses biens ou des biens des tiers. Notons que l'article L310-1 du Code des Assurances prévoit qu'il n'est pas possible d'assurer dans une même entreprise à la fois des activités vie et des activités IARD tandis que l'activité de prévoyance peut être assurée par l'une ou l'autre des entités.

Assurance IARD

Le principe indemnitaire de l'assurance IARD a des spécificités par rapport à la nature forfaitaire des prestations d'assurance vie :

- l'aléa moral qui désigne l'éventualité qu'un assuré augmente sa prise de risque par rapport à la situation où il supporterait entièrement les conséquences négatives des risques auxquels il

s'expose. L'aléa moral conduit donc à des dysfonctionnements du marché d'assurance non-vie car l'assuré n'est pas toujours incité à prévenir les sinistres, ce qui est moins le cas en assurance vie ;
- l'antisélection qui naît du fait que les individus connaissent mieux leur propre risque que la compagnie d'assurance. Ils peuvent donc utiliser ce supplément d'information lorsqu'ils souscrivent un contrat d'assurance, ce qui a pour conséquence qu'une prime reflétant le risque moyen dans la population attire les individus à haut risque et, *a contrario*, écarte les individus à bas risque. Ce phénomène existe également en assurance vie mais dans une moindre mesure ;
- l'horizon de placement des primes qui est typiquement de quelques années voire de quelques mois pour les contrats IARD tandis qu'il représente en moyenne une dizaine d'années pour les contrats vie.

Aujourd'hui, une part importante de l'assurance vie correspond en réalité à de l'activité d'épargne, l'aléa repose sur le moment de remboursement des sommes et la marge est issue du volume de chiffre d'affaires. Pour les contrats IARD, la rentabilité technique provient de la justesse et de la précision de l'évaluation actuarielle, à savoir la tarification et le provisionnement. L'activité est très régulée, pour les assureurs comme pour les assurés, et le marché est particulièrement concurrentiel.

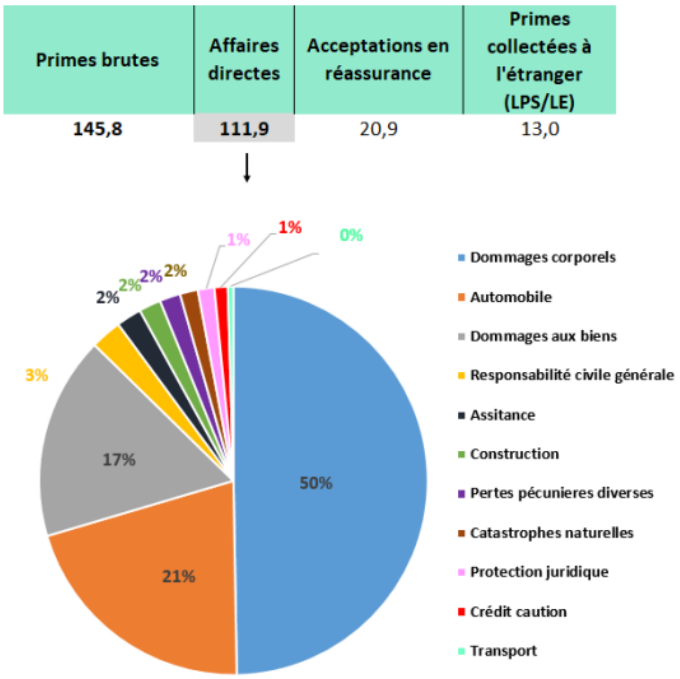


FIGURE 1.1 – Primes acquises en assurance non-vie, en milliards d'euros. Source : ACPR (2021).

En dépit du niveau très bas des taux d'intérêt et d'une crise sanitaire puis économique, le secteur de l'assurance fait preuve d'une forte capacité de résistance et demeure le premier marché national au sein de l'Union Européenne (ACPR (2021)). Cependant, à la différence de l'évolution constatée en assurance vie, les primes collectées en assurance non-vie n'ont progressé que légèrement en 2020, passant de 144,1 milliards d'euros en 2019 à 145,2 milliards d'euros en 2020, les affaires directes atteignant 111,9 milliards d'euros (Figure 1.1). Si les effets de la crise sanitaire ont eu peu d'impact sur le niveau des primes, les équilibres techniques se sont dégradés. Le résultat technique en assurance non-vie s'établit à 3,2 milliards d'euros en 2020, en diminution de 41% par rapport à 2019. Cette dégradation est due à une hausse des charges, notamment de la sinistralité. Cette dernière se distribue inégalement sur les

différentes lignes d'activité : le nombre de sinistres est en baisse sur le marché de l'assurance automobile et l'assurance santé mais une sur-sinistralité est observée sur la ligne d'activité des pertes pécuniaires diverses et de la prévoyance collective. L'accroissement des primes collectées ne permet pas de compenser la hausse des prestations et le ratio combiné s'établit à 97% en 2020 contre 96% en 2019.

Assurance automobile

Ce mémoire se focalise sur de l'assurance automobile en France et s'appuie sur des données d'un assureur. L'assurance automobile représente une part importante du marché de l'assurance IARD, pesant environ 21% en 2020 des primes acquises en affaires directes (Figure 1.1). En France, elle est obligatoire depuis 1958 au titre de la garantie responsabilité civile, à savoir les dommages causés à autrui par un véhicule terrestre à moteur. La durée d'un contrat type est d'un an et la garantie obligatoire de responsabilité civile est souvent complétée par des garanties en cas de bris de glace, de vol ou de dommage collision. D'autres garanties existent comme l'assistance juridique, le contenu privé ou la protection contre les événements climatiques. Ces dernières sont généralement limitées par des franchises, des plafonds et des exclusions qui permettent à l'assureur de lutter contre l'aléa moral et de restreindre son périmètre de prise en charge. S'ajoutent à cela des mesures d'incitation visant à encourager les assurés à adopter un comportement réduisant leur probabilité de survenance de sinistre dont la plus connue est le Coefficient de Réduction-Majoration (CRM). Ce système réglementaire, dont la grille d'évaluation est définie dans le Code des Assurances, récompense l'absence d'un accident responsable en appliquant un coefficient qui minore la prime d'assurance de référence.

Sur les dix dernières années, le coût moyen des sinistres progresse à un rythme annuel de 5,7% pour les accidents corporels et de 3,5% pour les accidents matériels, une inflation que ne parvient pas à compenser la diminution de la fréquences des sinistres. En 2020, quelle que soit la garantie, la fréquence de sinistres recule sensiblement sous l'effet de la baisse du trafic routier durant la crise sanitaire mais est en partie contrebalancée par l'inflation, notamment par une hausse de 8,1% du prix des pièces détachées. Du fait de la sous-sinistralité concomitante à la quasi-stagnation des primes collectées, le ratio combiné s'améliore en assurance automobile et s'établit à 94,7%, un niveau inférieur à 100 pour la première fois depuis 2005 (France Assureurs (2021)).

L'assurance automobile s'inscrit dans un contexte de changements réglementaires et technologiques qui créent de nouvelles dynamiques. En France, le marché est très mature et compétitif, avec près de 100 acteurs recensés par France Assureurs. De plus en plus de bancassureurs profitent d'une clientèle très large sur laquelle ils possèdent déjà nombreuses données pertinentes. Depuis le 1er janvier 2015, la loi Hamon, dite « loi sur la consommation », permet aux assurés ayant souscrit une assurance auto/moto/habitation de résilier leur contrat après un an d'adhésion, sans être exposés à des pénalités. Cette mesure a apporté beaucoup de liquidité et d'activité au marché de l'assurance, intensifiant la concurrence entre les différents acteurs qui font maintenant face à des assurés moins captifs et plus sujets à des sollicitations commerciales. Le développement de l'assurance directe, les comparateurs en ligne et plus récemment l'intégration de la télématique transforment le processus de souscription tandis que les avancées technologiques, guidées par l'avènement du *Big Data* et des méthodes de *Machine Learning*, ouvrent la voie vers une tarification encore plus fine (Santenar et al. (2022)).

1.1.2 La tarification

Traditionnellement, la construction d'un produit d'assurance est structurée en cinq étapes :

1. définir l'assuré (B to B ou B to C) ;
2. lister les garanties incluses dans le contrat ;
3. lister les exclusions ou les règles de prévention ;
4. définir les biens assurées ;
5. calculer la prime.

Le produit d'assurance est ensuite affecté à l'une des branches pour lesquelles l'organisme possède l'agrément, c'est-à-dire l'autorisation d'exercer l'activité d'assurance.

L'enjeu principal de la tarification non-vie est la séparation entre la prime pure et la prime commerciale. La prime pure (**PP**) correspond à la tarification du risque, c'est-à-dire à l'espérance mathématique des sinistres. Elle couvre le montant probable du sinistre dépendant de la garantie souscrite. La prime commerciale (Figure 1.2) est la prime réellement versée par l'assuré. Elle est obtenue en ajoutant à la prime pure les chargements d'acquisition et de gestion (**C1**), qui couvrent les coûts de l'entreprise, et les chargements de sécurité (**C2**), qui sont souvent proportionnels à la variance du sinistre ou à la prime pure et couvrent donc le risque de mauvaise tarification.

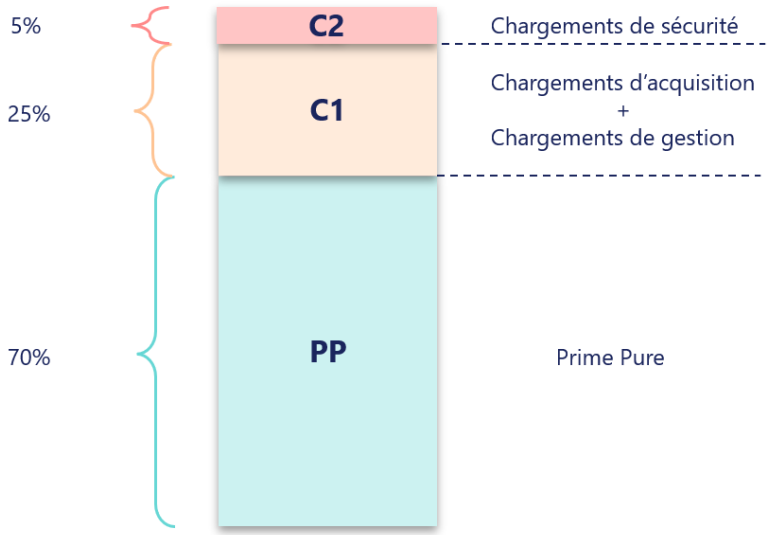


FIGURE 1.2 – Décomposition de la prime commerciale en assurance non-vie

Selon le modèle général fréquence/sévérité, on définit la variable aléatoire S comme la somme

$$S = \sum_{k=1}^{N} B_k, \tag{1.1}$$

où N est le nombre de sinistres pour une période donnée et B_k, $k = 1, ..., N$ représente le montant des sinistres. On suppose que N est une variable discrète tandis que les B_k sont des variables continues positives. Les hypothèses standard en tarification IARD sont les suivantes :

(H_1) : $B_k \overset{\text{iid}}{\sim} B$, *ie* les montants des sinistres ont le même comportement aléatoire ;
(H_2) : $\forall\, k = 1, ..., N$, $B_k \perp N$, *ie* la fréquence est indépendante de la sévérité.

Les hypothèses (H_1) et (H_2) permettent d'obtenir une décomposition coût-fréquence de la sinistralité

$$\mathbb{E}[S] = \mathbb{E}[N]\,\mathbb{E}[B]. \tag{1.2}$$

Par la suite, la fréquence et le montant des sinistres seront estimés indépendamment (la démonstration complète est disponible en Annexe (A.1)). En assurance automobile, le fonctionnement est pyramidal : l'assuré part généralement d'un socle, composé de la garantie obligatoire de responsabilité civile et d'une ou plusieurs garanties principales auxquelles il peut ajouter des modules ou options afin de construire la mesure de protection qui correspond à son utilité. Lors de la souscription, ces garanties peuvent être choisies indépendamment les unes des autres et nécessitent donc chacune une tarification spécifique. Les garanties communément incluses sont définies de la façon suivante :

- bris de glace : dommages causés au pare-brise, vitres latérales et lunette arrière ;
- dommage tout accident : ensemble des dommages matériels subis par le véhicule assuré, que le conducteur soit responsable ou non de l'accident ;
- vol : dommages subis par le véhicule lors d'un vol total, partiel ou d'une tentative ;
- incendie : dommages matériels subis par le véhicule assuré lors d'un incendie, de la chute de la foudre ou d'une explosion ;
- protection du conducteur : aide psychologique en cas de traumatisme suite à un accident ;
- protection juridique : prise en charge des frais engagés au tribunal en cas de litige avec un tiers ;
- tempête, grêle, neige qui : dommages matériels subis par le véhicule assuré lors des aléas naturels mentionnés plus haut.

1.2 Présentation de la base de données et analyse exploratoire

1.2.1 Description générale de la base de données

Ce mémoire s'appuie sur les données d'un assureur, régulièrement utilisées pour des travaux de tarification et de recherche et développement au sein du cabinet. La base de données étant déjà fiabilisée, l'objectif de l'analyse exploratoire ici proposée est d'identifier un nombre volontairement restreint de variables explicatives les plus représentatives du risque. Ce choix est d'une part motivé par un souci d'interprétabilité au moment de l'analyse des modèles. En effet, nous ne souhaitons conserver que les variables explicatives immédiatement parlantes d'un point de vue métier pour vérifier la cohérence de notre modélisation. D'autre part, l'algorithmie que l'on souhaite mettre en place nécessite de s'appuyer sur un petit nombre de prédicteurs pour permettre un paramétrage correct avec des temps computationnels raisonnables. Ces aspects seront détaillés par la suite dans le Chapitre 3. La sélection des variables explicatives s'appuie donc sur des méthodes statistiques couplées à l'expertise métier.

Dans tout ce qui suit, les applications seront réalisées avec `R` (R Core Team (2022)) et `Python` (Van Rossum et Drake (2009)). Les packages spécifiques seront cités au fur et à mesure.

Pour des raisons de stockage et de manipulation, deux bases de données sont utilisées, une pour la modélisation du coût moyen, et une pour la modélisation de la fréquence. Dans la base de données fréquence, une ligne correspond à un contrat tandis que dans la base de données coût moyen, une ligne représente un sinistre. Les informations relatives à un même contrat sont réparties sur plusieurs lignes dans le but d'isoler les différentes périodes de vie d'un contrat qui présentent une homogénéité des caractéristiques et de situation. Chaque ligne correspond ainsi à une image, la création d'une nouvelle image étant provoqué par la reconduction annuelle du contrat ou un avenant.

La garantie ici modélisée est la garantie dommage tout accident (DTA) sur une période allant de 2010 à 2021. Le risque est supposé homogène sur cette période. Cependant, une variable temporelle sera incluse dans l'étude pour capter une tendance dans l'évolution du coût moyen ou de la fréquence qui ne serait pas due à un phénomène structurel. La volumétrie des deux bases de données est précisée dans la Table 1.1 et la typologie des variables explicatives est résumée dans la Figure 1.3.

	Observations	**Variables**
Coût moyen	83 099	100
Fréquence	4 947 201	139

TABLE 1.1 – Taille de la base de données pour la modélisation du coût moyen et de la fréquence

FIGURE 1.3 – Résumé du jeu de données

La première étape consiste à faire une pré-sélection des variables basée sur un savoir faire métier. Une cinquantaine de variables est ainsi supprimée dont :

- les variables relatives à la modélisation des autres garanties ;
- les variables qui apportent une information redondante avec d'autres variables ;
- les variables révélées non-significatives dans les études précédentes ;
- les variables dont l'intégration dans le modèle est difficile, notamment les variables qualitatives possédant un trop grand nombre de modalités que l'on n'arrive pas à regrouper en groupes homogènes et stables.

1.2.2 Statistiques descriptives et sélection non-supervisée de variables

Toute étude d'un jeu de données doit être précédée d'une étape descriptive à l'aide d'outils, certes rudimentaires, mais robustes, en privilégiant les représentations graphiques. L'objectif est de se familiariser avec les données, identifier les premiers éléments explicatifs et effectuer une sélection non-supervisée des variables.

Analyse des corrélations

Dans cette partie, on s'intéresse à l'étude simultanée de deux variables X et Y observées sur un même échantillon. Notons n la taille de l'échantillon et $(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$ les observations. L'analyse des corrélations est une méthode statistique permettant de mesurer la force de la relation linéaire entre deux variables. Un fort lien de corrélation indique que l'essentiel de l'information portée

par l'une des variables est portée par l'autre, donc l'une des deux peut être supprimée. L'analyse des corrélations permet ainsi d'éviter la redondance de l'information, étape qui aura son importance lors de la modélisation. Disposant de données quantitatives et qualitatives, différentes mesures de dépendance $\rho_{x,y}$ sont calculées selon le type des variables :

- la corrélation de Pearson si X et Y sont quantitatives (**cas 1**) ;
- le V de Cramér si X et Y sont qualitatives (**cas 2**) ;
- la corrélation multiple si X et quantitative et Y qualitative, ou l'inverse (**cas 3**).

Dans le **cas 1**, on a $\rho_{x,y} = \frac{Cov(x,y)}{\sqrt{Var(x)\,Var(y)}}$, où on considère :

- la covariance entre X et Y : $Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}_n \overline{y}_n$, avec $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$;
- la variance de X : $Var(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x}_n)^2$;
- la variance de Y : $Var(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\overline{y}_n)^2$.

Dans le cas **cas 2**, on suppose que X et Y sont deux variables qualitatives de modalités $m_1, \ldots, m_J$ et $l_1, \ldots, l_K$ et on définit :

- les effectifs conjoints, comptant le nombre de fois que les modalités m_j et l_k sont observées sur un même individu : $n_{jk} = \sum_{i=1}^{n} \mathbb{1}_{\{x_i = m_j\}} \mathbb{1}_{\{y_i = l_k\}}$;
- les effectifs marginaux comptant le nombre de fois que la variable X (respectivement Y) prend la modalité m_j (respectivement l_k) : $n_{j\cdot} = \sum_{k=1}^{K} n_{jk}$ et $n_{\cdot k} = \sum_{j=1}^{J} n_{jk}$;
- la statistique du khi-deux :

$$\chi^2 = \sum_{j,k} \frac{(n_{jk} - \frac{n_{j\cdot} n_{\cdot k}}{n})^2}{\frac{n_{j\cdot} n_{\cdot k}}{n}} \tag{1.3}$$

On définit alors le V de Cramér comme $\rho_{x,y} = \sqrt{\frac{\frac{\chi^2}{n}}{min(J-1, K-1)}}$.

Dans le **cas 3**, on suppose que la variable X est quantitative et que la variable Y est qualitative de deux modalités 0 et 1 (ce résultat pourra ensuite être généralisé). On définit :

- $\overline{x}_n^0$ la moyenne de la variable X pour tous les points où la variable Y prend la modalité 0 ;
- $\overline{x}_n^1$ la moyenne de la variable X pour tous les points où la variable Y prend la modalité 1 ;
- la fréquence marginale de la modalité 0 : $f_0 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i = 0}$;
- la fréquence marginale de la modalité 1 : $f_1 = 1 - f_0$.

On définit alors la corrélation multiple par $\rho_{x,y} = \frac{\overline{x}_n^1 - \overline{x}_n^0}{Var(x)} \sqrt{f_0 f_1}$.

La fonction `assoc` du package `greybox` (Svetunkov, 2022) permet de résumer les résultats dans une matrice de corrélation (Figure A.1). Nous observons alors que les variables du véhiculier SRA sont positivement corrélées, tout comme les règles de comptage de sinistres. La variable donnant l'année civile (`annee`) est très fortement positivement corrélée à la variable déterminant l'exercice comptable afférent au phénomène observé (`exercice`). Par la suite, l'idée sera de conserver une variable par groupe de variables corrélées. A l'aide de la théorie des graphes, la corrélation peut également être représentée de manière graphique. Pour cela, il est possible de combiner la fonction `assoc` à la fonction `network_plot` du package `corrr` (Kuhn et al., 2020). La Figure 1.4 montre les résultats pour le variables du véhiculier SRA. Nous remarquons ainsi que la variable `prix_SRA` permet de reconstituer les liens avec les autres variables explicatives. La Figure 1.5 fournit le graphe des corrélations complet. Il est ainsi possible d'identifier les grandes typologies de variables explicatives résumées dans la Figure 1.3. Il est notamment intéressant d'observer les relations qui existent entre les caractéristiques du conducteur

et du contrat. Mentionnons par exemple la variable `lien_COND` qui donne le lien avec le conducteur principal et joue sur la majoration appliquée aux conducteurs novices (`majo`), ou l'âge du conducteur (`age_COND`) négativement corrélé au coefficient de réduction-majoration (`cof_CRM`). Notons également que les variables décrivant l'équipement de la voiture (`equip1`, `equip2` et `equip3`) sont corrélées.

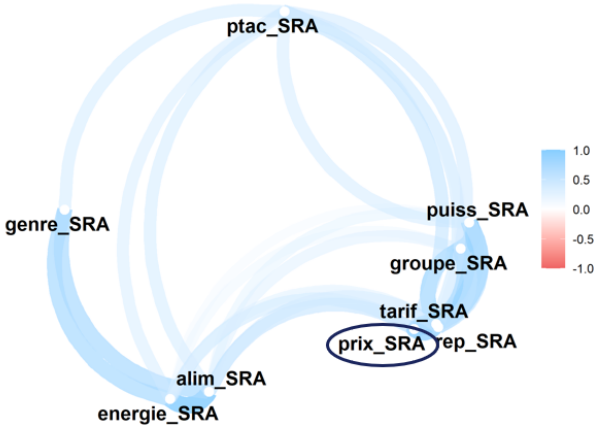


FIGURE 1.4 – Graphe des corrélations pour les variables du véhiculier

Méthodes factorielles de représentation et discrimination : ACP et ACM

L'analyse en composantes principales (ACP) et l'analyse des correspondances multiples (ACM) sont des méthodes descriptives permettant de synthétiser l'information contenue dans un jeu de données multivarié (BESSE (1992)). Au-delà de la visualisation et de la description des données, l'analyse factorielle fournit un système de représentation des variables latentes formées à partir de combinaisons linéaires des variables initiales, et comporte donc un aspect explicatif servant d'outil à la décision. Pour l'implémentation, nous allons nous appuyer sur le package `FactoMineR` (LÊ et al. (2008)).

L'analyse en composantes principales vise à produire une description synthétique d'un tableau de données relatif à p_1 variables quantitatives et n individus. Certaines informations contenues dans les différentes variables sont redondantes, l'objectif de l'ACP est alors d'exhiber un petit nombre de directions permettant de rendre compte de la dispersion des données d'intérêt. L'ACP se décline autour de la recherche des axes et des composantes principales. D'une certaine manière, les axes principaux définissent un nouveau repère dans $\mathbb{R}^{p_1}$, en lieu de celui engendré par la base canonique. De même, les projections du nuage de points sur ces nouveaux axes, i.e. les composantes principales, définissent ce qu'il conviendrait d'appeler des méta-variables. Ce sont des combinaisons linéaires des variables initiales décorrélées entre elles auxquels il faut donner une signification pour la description des données. Le premier plan factoriel, c'est-à-dire le plan formé par les deux premières composantes principales, fournit un résumé du jeu de données tandis que les autres plan factoriels permettent d'accéder à des informations de détail.

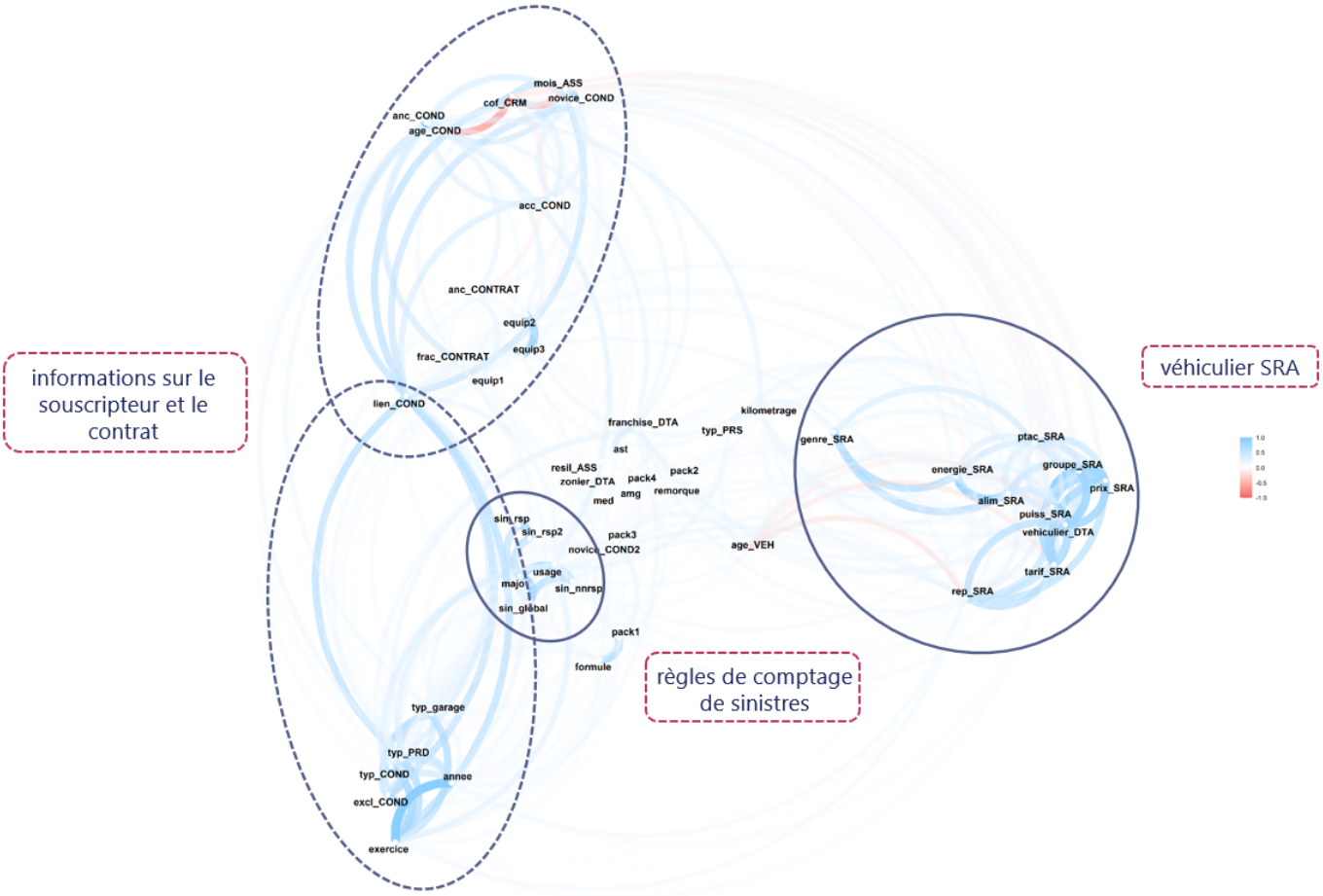


FIGURE 1.5 – Graphe des corrélations complet

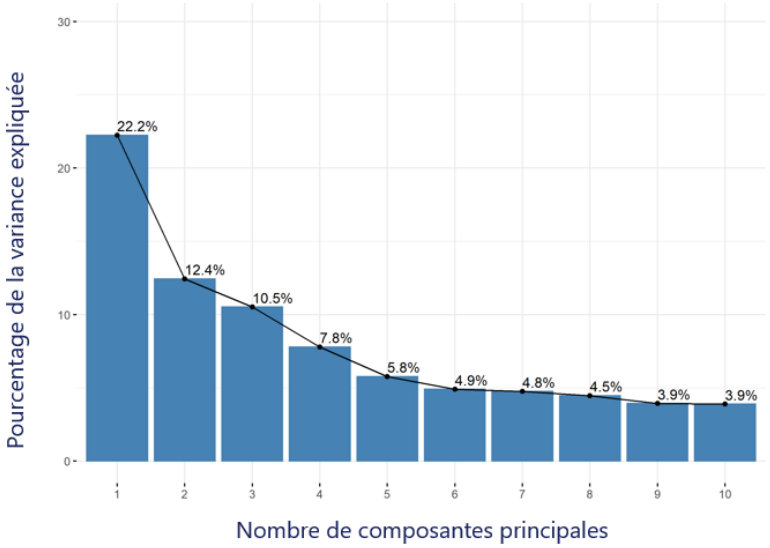


FIGURE 1.6 – Éboulis des valeurs propres pour la modélisation du coût moyen

Pour la modélisation du coût, les 10 premières composantes principales expliquent plus de 80% de l'inertie du nuage des points (Figure 1.6). Dans le premier plan factoriel, on capture environ 35% de l'inertie totale (Figure 1.7). La Figure 1.7 montre également que la première méta-variable peut être associée aux caractéristiques du véhicule avec les variables du véhiculier, tandis que la seconde porte les informations sur le contrat et le souscripteur. Les variables sont coloriées en fonction de leur contribution à l'inertie du nuage, ce qui permet de détecter les variables les plus influentes (en bleu) par groupes de variables corrélées : `prix_SRA`, `equip3` et `sin_global`. Finalement, on projette la variable coût moyen (`CM`) dans le premier plan factoriel et on remarque l'association avec la première méta-variable, un lien avec les caractéristiques du véhicule donc. La zone est indiquée schématiquement par des tirets violets.

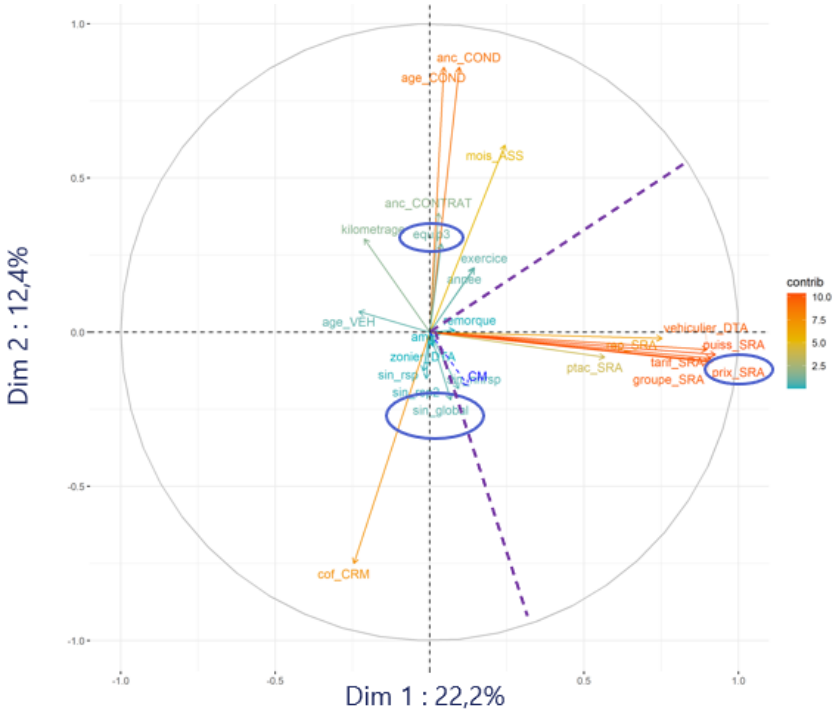


FIGURE 1.7 – Graphe des variables dans le premier plan factoriel pour la modélisation du coût moyen

Pour la modélisation de la fréquence, les 10 premières composantes principales expliquent environ 76% du nuage des points (Figure 1.8) et le premier plan factoriel capture presque 32% de l'inertie totale (Figure 1.9). On peut associer la même signification aux méta-variables dans le premier plan factoriel, à savoir un lien avec les caractéristiques du véhicule pour le premier axe et un lien avec caractéristiques du contrat et du conducteur pour le second. On retrouve également les mêmes variables influentes en termes de contribution. Cependant, la variable `freq` semble être plus portée par la seconde méta-variable. Autrement dit, la fréquence des sinistres est davantage expliquée par les informations sur le contrat et le conducteur dans le premier plan factoriel.

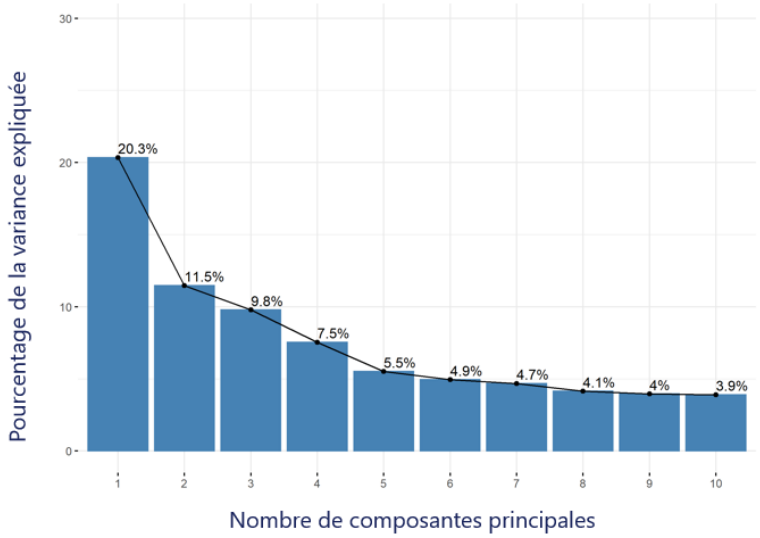


FIGURE 1.8 – Éboulis des valeurs propres pour la modélisation de la fréquence

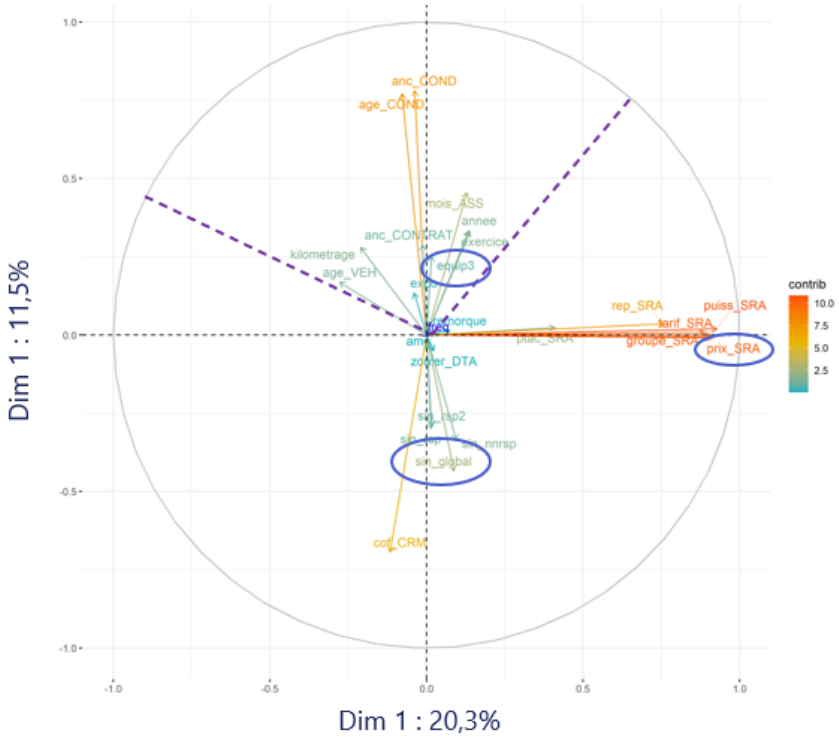


FIGURE 1.9 – Graphe des variables dans le premier plan factoriel pour la modélisation de la fréquence

Analyse des correspondances multiples

L'analyse des correspondances multiples est une extension de l'analyse factorielle des correspondances permettant de décrire les relations entre p_2 variables qualitatives simultanément observées sur n individus. Il s'agit d'une généralisation de l'analyse en composantes principales aux variables qualitatives.

L'analyse factorielle des correspondances se base sur la construction d'une table de contingence associée aux observations de la variable X de modalités $m_1, \ldots, m_J$ et de la variable Y de modalités $l_1, \ldots, l_K$. Ses éléments génériques sont l'effectif conjoint $n_{jk} = \sum_{i=1}^{n} \mathbb{1}_{\{x_i=m_j\}} \mathbb{1}_{\{y_i=l_k\}}$ et les effectifs marginaux $n_{j\cdot} = \sum_{k=1}^{K} n_{jk}$ et $n_{\cdot k} = \sum_{j=1}^{J} n_{jk}$. On en déduit les profils-lignes $\left\{ \frac{n_{j1}}{n_{j\cdot}}, \ldots, \frac{n_{jk}}{n_{j\cdot}}, \ldots, \frac{n_{jK}}{n_{j\cdot}} \right\}$ et les profils-colonnes $\left\{ \frac{n_{1k}}{n_{\cdot k}}, \ldots, \frac{n_{jk}}{n_{\cdot k}}, \ldots, \frac{n_{Jk}}{n_{\cdot k}} \right\}$ qui permettent de caractériser chaque modalité. Ces espaces sont munis des métriques du χ^2 (Équation 1.3) exprimant les correspondances entre modalités en termes de distances. L'analyse factorielle des correspondances est alors le résultat d'une double analyse en composantes principales : l'ACP des profils-lignes et l'ACP des profils-colonnes. L'analyse des correspondances multiples se base sur la construction d'un tableau de Burt qui comporte toutes les tables de contingence des variables prises deux à deux. La démarche suivie est ainsi celle de l'analyse factorielle des correspondances. Les résultats de l'ACM, réalisée sur la base de données coût moyen, sont résumés dans la Figure 1.11.

La Figure 1.10 montre qu'avec 10 composantes principales, on ne parvient à expliquer que 37% de l'inertie du nuage de points. Dans le premier plan factoriel, on capture environ 15% de l'inertie totale (Figure 1.11). Les variables les plus contributives (en bleu) sont le type du conducteur (`typ_COND`) et le lien avec le conducteur principal (`lien_COND`). En revanche, la variable donnant le lieu de stationnement du véhicule (`typ_garage`) est peu contributive et proche au sens de la distance du χ^2 d'autres variables contributives. Elles sera donc supprimée de l'étude.

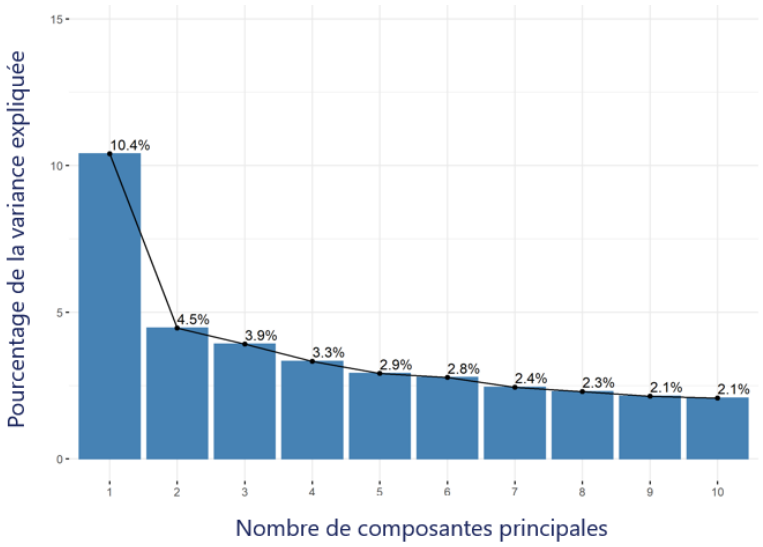


FIGURE 1.10 – Éboulis des valeurs propres pour l'analyse des correspondances multiples

A l'issue de l'analyse exploratoire, nous supprimons 16 variables explicatives de l'étude. Pour porter l'information du véhiculier SRA, c'est la variable `prix_SRA` qui sera retenue, pour le comptage de sinistres la variable `sin_global`, pour l'équipement de la voiture la variable `equip3` et pour la tendance temporelle la variable `annee`.

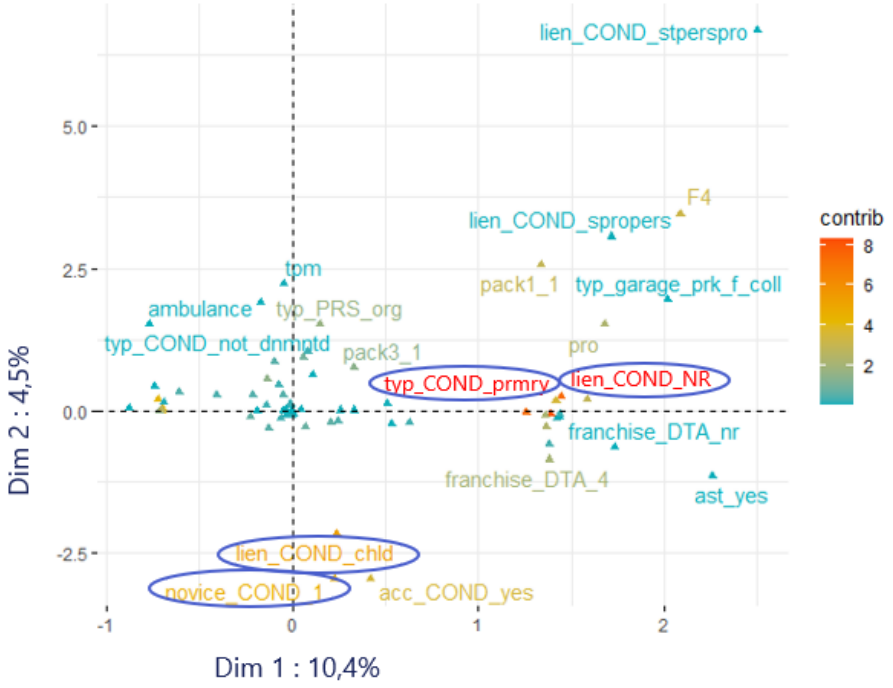


FIGURE 1.11 – Graphe des variables dans le premier plan factoriel pour l'analyse des correspondances multiples

1.2.3 Sélection supervisée de variables

La sélection non-supervisée permettant de supprimer qu'un nombre limité de variables, une procédure supervisée est mise en oeuvre pour réduire significativement la volumétrie verticale du jeu de données et faciliter la modélisation subséquente. Nous nous appuyons d'une part sur le sens actuariel que nous souhaitons valoriser dans la modélisation par un choix judicieux de variables explicatives. D'autre part nous croisons les résultats donnés par :

- la significativité des coefficients dans le modèle linéaire généralisé (pour les détails techniques voir Section 2.2.2) ;
- l'importance des variables dans l'algorithme *Random Forest* (Section 2.3.2) ;
- la sélection des variables par la pénalisation Lasso (Section 2.2.3).

Les Figures 1.12 et 1.13 (respectivement 1.14 et 1.15) montrent les 20 variables les plus importantes données par la méthode *Random Forest* et la méthode Lasso pour la modélisation du coût moyen (respectivement la fréquence). On remarque alors que l'on retrouve globalement les mêmes variables discriminantes, mais pas dans le même ordre. S'appuyant sur la procédure *Random Forest*, c'est la variable zonier (`zonier_DTA`) qui ressort en premier pour le coût moyen (Figure 1.12) tandis que pour la fréquence, il s'agit de l'âge du conducteur (`age_COND`) (Figure 1.14). Après la variable donnant la présence d'un conducteur secondaire novice (`novice_COND2`), la procédure Lasso met en relief la variable véhiculier (`prix_SRA`) pour le coût moyen (Figure 1.13) et de nouveau l'âge du conducteur pour la fréquence (Figure 1.15).

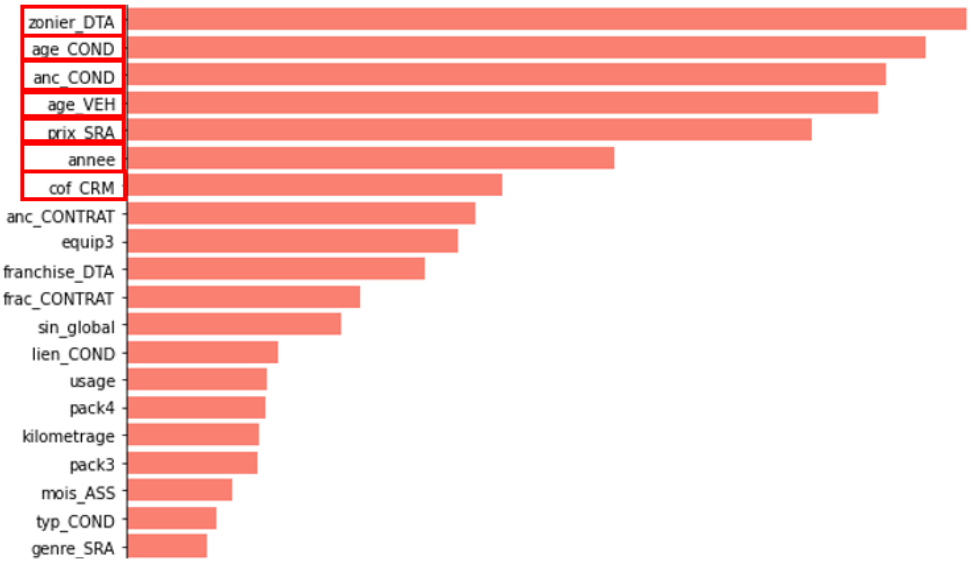


FIGURE 1.12 – Importance des variables par la méthode *Random Forest* pour la modélisation du coût moyen

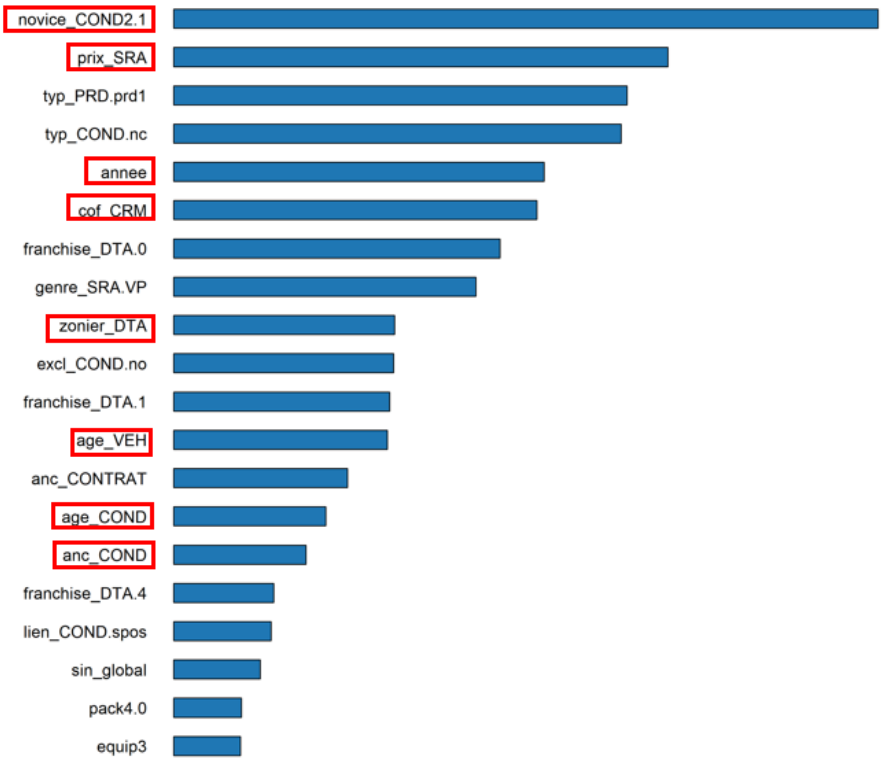


FIGURE 1.13 – Importance des variables par la méthode Lasso pour la modélisation du coût moyen

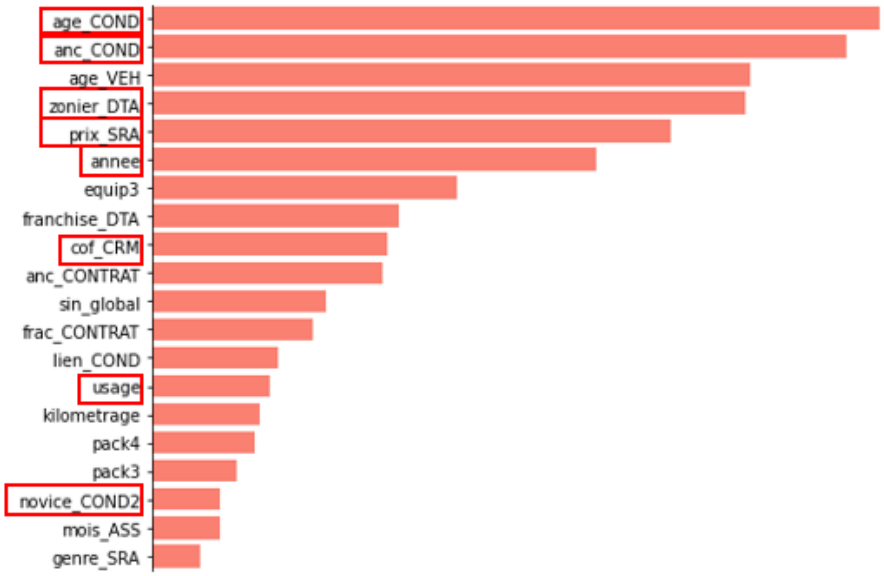


FIGURE 1.14 – Importance des variables par la méthode *Random Forest* pour la modélisation de la fréquence

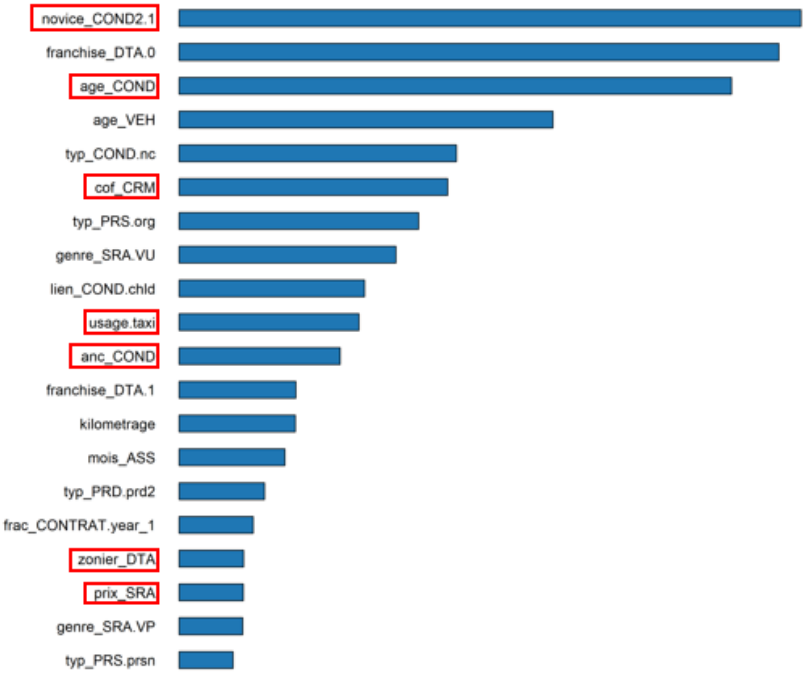


FIGURE 1.15 – Importance des variables par la méthode Lasso pour la modélisation de la fréquence

A l'issue de la pré-sélection, la sélection non-supervisée et la sélection supervisée, la Table 1.2 donne le détail de 8 variables qui peuvent être retenues pour la modélisation. Elles sont indiquées en rouge sur les Figures 1.14, 1.13, 1.14 et 1.15.

Nom de la variable	Description de la variable	coût moyen	fréquence
novice_COND2	présence d'un conducteur secondaire novice	×	×
prix_SRA	variable véhiculier donnant la classe de prix du véhicule	×	×
annee	variable donnant l'année civile	×	×
anc_COND	ancienneté du permis du conducteur principal en années	×	×
age_COND	âge du conducteur principal en années	×	×
age_VEH	âge du véhicule en années	×	
cof_CRM	coefficient de réduction-majoration	×	×
zonier_DTA	variable zonier relative à la garantie modélisée	×	×
usage	variable décrivant l'usage du véhicule		×

TABLE 1.2 – Description de 8 variables qui peuvent être retenues respectivement pour la modélisation du coût moyen et de la fréquence à l'issue de la pré-sélection, la sélection non-supervisée et la sélection supervisée des variables

1.2.4 Détermination du seuil des sinistres graves

Les sinistres graves, ou exceptionnels, sont caractérisés par l'association d'un coût important et d'une fréquence faible. La présence des sinistres graves, en particulier pour des garanties telles que la responsabilité civile, peut nuire à la qualité des modèles élaborés. Il est donc important de déterminer un seuil de sinistralité extrême en se basant sur des critères d'homogénéité des risques, de la suffisance des données et de la cohérence. De nombreuses méthodes statistiques existent pour étudier le comportement asymptotique des valeurs prises par une variable aléatoire. Ces méthodes, relevant essentiellement de la théorie des valeurs extrêmes (DE HAAN et al. (2006)), ne seront pas détaillées dans ce mémoire. Néanmoins, une revue rapide des principes fondamentaux sera proposée (LAUREILLE et al. (2019)) pour avoir un aperçu du comportement des sinistres graves pour la garantie DTA.

D'après le théorème de Fisher-Tippett-Gnedenko, la famille de distribution des valeurs extrêmes a pour fonction de répartition

$$H_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)_+^{-\frac{1}{\gamma}}\right) & \text{si } \gamma \neq 0 \\ \exp(-\exp(x)) & \text{si } \gamma = 0. \end{cases}$$

Selon le signe du paramètre γ, appelé indice des valeurs extrêmes ou indice de queue, on distingue trois domaines d'attraction :

- si $\gamma < 0$, la famille de distribution appartient au domaine d'attraction de Weibull. Il contient des lois dont la fonction de survie n'a pas de queue de distribution ;
- si $\gamma = 0$, la famille de distribution appartient au domaine d'attraction de Gumbel. Il contient les lois dont la fonction de survie est à décroissance exponentielle, c'est-à-dire les lois à queue légère comme la loi normale ou la loi Gamma.
- si $\gamma > 0$, la famille de distribution appartient au domaine d'attraction de Fréchet. Il contient les lois dont la fonction de survie est à décroissance polynomiale, c'est-à-dire les lois à queue lourde, comme la loi de Pareto ou la loi de Student.

Soit X une variable aléatoire et $X_u = [X - u | X > u]$ la variable aléatoire des excès résiduels au-delà du seuil u. D'après le théorème de Pickands-Balkema-de Hann, la distribution conditionnelle des excès résiduels, définie formellement par

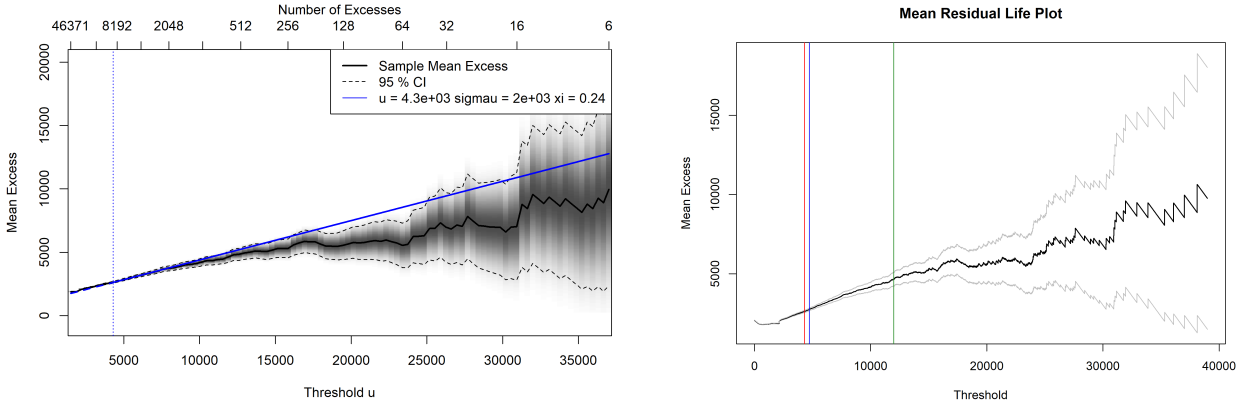
$$F_u(x) = \mathbb{P}(X - u \leq x | X > u) = 1 - \frac{S(x+u)}{S(u)},$$

où S désigne la fonction de survie, converge vers une loi de Pareto généralisée (**GPD**) lorsque la taille de l'échantillon, et donc le seuil u, tendent vers l'infini. Par conséquent, en choisissant un seuil convenable, une loi de Pareto généralisée pourra être ajustée aux montants des sinistres dépassant ce seuil.

La fonction moyenne des excès (ou *Mean Excess plot*) est couramment employée pour déterminer le seuil des sinistres extrêmes. En effet, si X_u suit une loi de Pareto généralisée $\mathbf{GPD}(\xi, \sigma_u)$, alors $\mathbb{E}[X_u] = \frac{\sigma_u}{1-\xi}$. La propriété de stabilité de la loi de Pareto généralisée permet d'écrire que pour tout seuil $v > u$ tel que $\xi < 1$, $X_v = [X - v | X > v] \sim \mathbf{GPD}(\xi, \sigma_u + \xi(v-u))$, et donc

$$\mathbb{E}[X_v] = \frac{\sigma_u + \xi(v-u)}{1-\xi},$$

qui est une fonction affine en v de coefficient directeur $\frac{\xi}{1-\xi}$ et d'ordonnée à l'origine $\frac{\sigma_u - \xi u}{1-\xi}$. Il s'agira alors de rechercher le seuil à partir duquel la fonction moyenne des excès empirique est affine. La Figure 1.16 fournit les résultats de cette approche en s'appuyant sur les packages `evmix` (Hu et Scarrott (2018)) et `POT` (Ribatet et Dutang (2022)). Le package `evmix` donne directement une estimation du seuil des graves $u = 4300$ ainsi qu'une estimation des paramètres de la loi de Pareto généralisée par maximum de vraisemblance $\xi = 0,24$ et $\sigma_u = 2000$. La fonction moyenne des excès obtenue avec le package `POT` permet de détecter une rupture dans le comportement affine autour de $u = 12000$. Ce seuil est indiqué en vert sur le graphe et correspond par ailleurs à l'estimateur de Hill défini plus bas (Figure 1.17a). La ligne bleue correspond au seuil donné par la méthode Gerstengarbe (Figure 1.17b) et la ligne rouge au seuil obtenu avec le package `evmix`.



(a) package `evmix` (b) package `POT`

FIGURE 1.16 – Détermination du seuil des graves avec la fonction moyenne des excès

Soient maintenant $X_1, \ldots, X_n$ n variables aléatoires indépendantes et de même fonction de répartition F appartenant au domaine d'attraction de Fréchet. L'estimateur de Hill de γ est défini par

$$\hat{\gamma}_n^H(k_n) = \frac{1}{k_n} \sum_{i=1}^{k_n} \ln(X_{n-i+1,n}) - \ln(X_{n-k_n,n}).$$

Le *Hill plot* consiste à tracer le graphe des points $\left(k_n, \hat{\gamma}_n^H(k_n),\ 1 < k_n < n\right)$. Cette méthode repose sur la propriété de stabilité de la loi de Pareto généralisée. La Figure 1.17a montre une stabilité de l'estimateur à partir de 660 excès, correspondant à un seuil de 12 000.

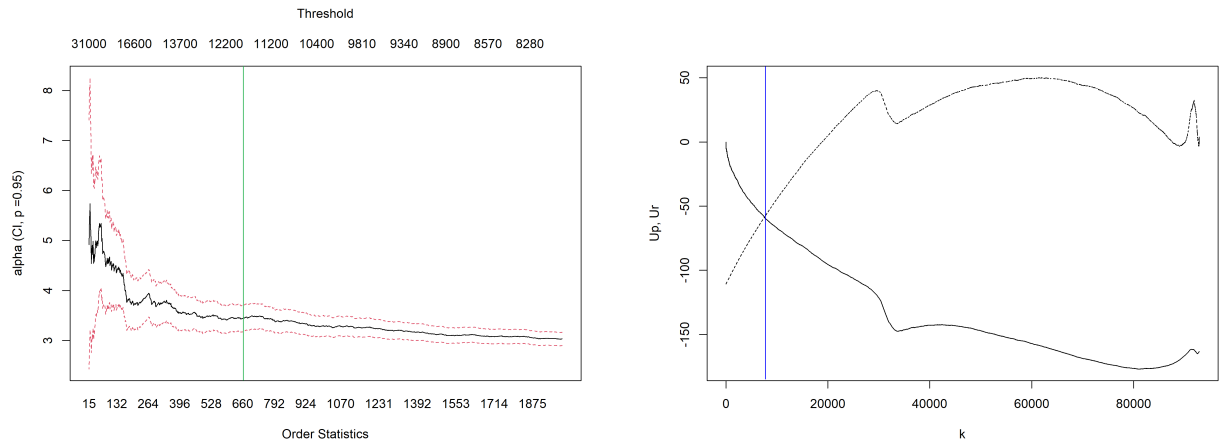
Dans la méthode de Gerstengarbe, il s'agit d'observer le comportement de la série des écarts consécutifs puis d'identifier le point de changement de comportement. Pour identifier ce dernier, un test séquentiel est effectué deux fois, pour les différences du début à la fin du jeu de données et inversement. Soit $X(1) < X(2) < \dots X(n)$ un échantillon ordonné de taille n. On définit la série des écarts du début jusqu'à la fin par

$$\begin{cases} \Delta_i &= X(i) - X(i-1),\ 2 < i < n \\ \Delta_1 &= 0 \\ U_i &= \frac{U_i^* - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i-1)(i+5)}{72}}} \\ U_i^* &= \sum_{k=2}^{i} |\Delta_1, \dots, \Delta_{k-1} \leq \Delta_k|. \end{cases}$$

Puis la série des écarts de la fin jusqu'au début par

$$\begin{cases} \Delta_j &= X(j) - X(j+1),\ 1 < j < n-1 \\ \Delta_n &= 0 \\ U_j &= \frac{U_j^* - \frac{(n-j)(n-j+1)}{4}}{\sqrt{\frac{(n-j)(n-j+1)(n-j+6)}{72}}} \\ U_j^* &= \sum_{k=j}^{n-1} |\Delta_{k+1}, \dots, \Delta_n \leq \Delta_k|. \end{cases}$$

L'intersection du tracé des deux courbes $\{(U_i, i),\ 1 < i < n\}$ et $\{(U_j, j),\ 1 < j < n\}$ sur le même graphe permet de déterminer le seuil optimal. La Figure 1.17b donne ce seuil à $u = 4720$ et se rapproche donc du résultat obtenu avec la fonction moyenne des excès (Figure 1.16).



(a) Estimateur de Hill (b) Méthode de Gerstengarbe

FIGURE 1.17 – Détermination du seuil des graves avec l'estimateur de Hill et la méthode de Gerstengarbe

Au-delà des indicateurs statistiques, il est possible d'étudier la typologie des sinistres extrêmes. Nous pouvons notamment vérifier si $X_u \sim \mathbf{GPD}(\xi, \sigma_u)$ en s'appuyant sur les diagrammes quantile-quantile,

ou QQ-plot, permettant d'évaluer la pertinence de l'ajustement d'une distribution donnée à une loi en comparant la position de certains quantiles dans la population observée avec leur position dans la population théorique. En utilisant l'estimation du seuil des graves $u = 4300$ et des paramètres $\xi = 0,24$ et $\sigma_u = 2000$ de la loi de Pareto généralisée obtenue avec le package `evmix` (Figure 1.16a), nous observons alors une bonne adéquation (Figure 1.18). Ainsi, l'ajustement d'une loi de Pareto aux montants des sinistres dépassant le seuil $u = 4300$ semble pertinent et un comportement extrême de la donnée confirmé.

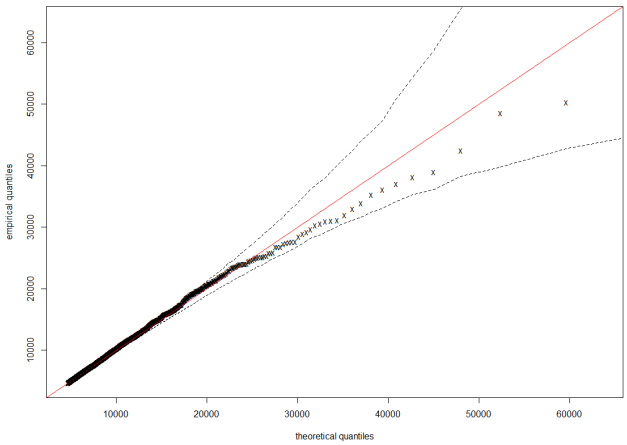


FIGURE 1.18 – QQ-plot de la distribution du coût moyen au-delà du seuil $u = 4300$ obtenu avec le package `tea` (OSSBERGER (2020))

Néanmoins, le principal défaut de ces méthodes statistiques demeure leur interprétabilité et leur applicabilité. En effet, les différents seuils déterminés ne sont pas forcément cohérents entre eux et il n'est pas aisé de leur associer un sens actuariel. Avec un seuil inférieur à 5 000, la sinistralité extrême représenterait 7% de la base, ce qui semble trop important pour la garantie DTA. Le seuil de 12 000 est plus cohérent et pourrait indiquer la destruction complète de la voiture d'un prix moyen. Par la suite, nous choisissons donc de ne pas considérer la séparation en sinistres graves et attritionnels pour la modélisation de la garantie DTA, estimant que la destruction complète de la voiture est un signal qui doit pouvoir être capté par les variables issues du véhiculier SRA. Nous reviendrons cependant sur le seuil de sinistralité extrême dans le Chapitre 3 (Section 3.3.1).

Dans ce chapitre, nous avons délimité le cadre actuariel dans lequel s'inscrit notre étude. Nous avons décrit les spécificités de l'assurance non-vie en nous focalisant sur l'assurance automobile et en nous appuyant sur des chiffres récents. Nous avons également rappelé les principes de la tarification et défini la garantie DTA (dommage tout accident) qui sera modélisée par la suite. La base de données assureur servant de base à la modélisation a été explorée sous l'angle de l'analyse des corrélations et des méthodes factorielles de discrimination. Une sélection des variables explicatives a ainsi été réalisée. Une analyse de la sinistralité extrême a également été proposée.

Chapitre 2

Présentation des modèles actuels et de leurs limites

2.1 Introduction

2.1.1 Contexte général

Depuis les années 1990, les modèles linéaires généralisés constituent l'outil de référence employé par les actuaires pour la tarification des produits d'assurance non-vie. De conception simple et transparente, ces modèles permettent de produire une segmentation tarifaire cohérente et commercialement applicable. Quantifiant l'impact de chaque variable explicative sur la variable réponse, ils répondent directement au besoin d'interprétabilité de la profession avec une structure paramétrique adaptée à la nature des données. Par ailleurs, ce souci d'intelligibilité permet de conserver un sens économique dans la modélisation mathématique.

Ces dernières années, l'avènement du *Big Data*, couplé à la démocratisation des techniques de *Machine Learning*, ouvre de nouvelles perspectives dans le processus de prise de décision. Que ce soit dans le diagnostic médical, dans le domaine du web et du multimédia ou dans le secteur de la finance et de l'assurance, la possibilité de transformer les données collectées en information permet d'accéder à une compréhension plus fine du phénomène que l'on étudie, en expliquer les causes principales, anticiper certaines de ses conséquences, gérer les risques consubstantiels. Ces évolutions récentes s'effectuent néanmoins en parallèle d'un renforcement de la réglementation et de la transparence. Le Règlement Général de la Protection des Données (RGPD) (Parlement Européen et Conseil de l'Union Européenne (2016)) stipule que « le recours à un processus de décision automatique doit se justifier par la capacité à fournir des renseignements exacts et utiles sur le fonctionnement interne de l'algorithme, et la pertinence de l'évaluation des conséquences d'une telle décision sur l'individu concerné[†] ». Suite à la publication du livre blanc pour une approche de l'Intelligence Artificielle (IA) basée sur l'excellence et la confiance (Commission Européenne (2020)), de nombreuses propositions de textes réglementaires ont été publiées dont un établissant des règles harmonisées sur l'IA (Ebers et al. (2021)). Modéliser avec finesse les phénomènes observés, puis expliquer les prédictions du modèle de manière précise et consistante, est une question pratique, éthique et juridique (Besse et al. (2018)). Plus particulièrement, dans le secteur de l'assurance, la nécessité de justifier le tarif auprès des clients et des autorités de

†. traduction libre de l'anglais : "the existence of automated decision-making should carry meaningful information about the logic involved, as well as the signifcance and the envisaged consequences of such processing for the data subject"

contrôle se heurte à l'excédent de complexité et au manque d'interprétabilité des modèles de Machine Learning, freinant ainsi leur applicabilité opérationnelle. Le besoin de confiance et de transparence est très présent et très prisé.

Dans ce contexte, l'Intelligence Artificielle Explicable (*eXplainable Artificial Intelligence* (XAI)) apparaît comme un concept central pour concilier les exigences de performance et de transparence. Les mots clés qu'il convient alors de distinguer sont l'interprétabilité et l'explicabilité (Gilpin et al. (2018)). Contrairement à un modèle boîte noire (*Black Box*) dont les données d'entrée et de sortie sont connues mais dont le fonctionnement reste opaque, un modèle interprétable est un modèle dont la représentation interne des données est compréhensible et pilotable par un expert de Machine Learning. Le concepteur d'un tel modèle est en mesure de justifier le bon ou le mauvais fonctionnement de l'algorithme et de quantifier la confiance dans ses résultats. L'explicabilité, quant à elle, consiste à fournir une information sous une forme compréhensible par un utilisateur non expert de la donnée. Un modèle explicable est donc interprétable mais l'inverse n'est pas nécessairement vrai.

L'objectif de ce chapitre est triple. Dans un premier temps, il s'agit de rappeler les fondements théoriques du modèle linéaire pour définir le modèle linéaire généralisé, ses hypothèses et ses limites. Dans un deuxième temps, les principaux modèles de *Machine Learning* sont présentés, avec un accent particulier sur l'agrégation des modèles. La troisième partie est consacrée aux outils d'interprétabilité des modèles de *Machine Learning*.

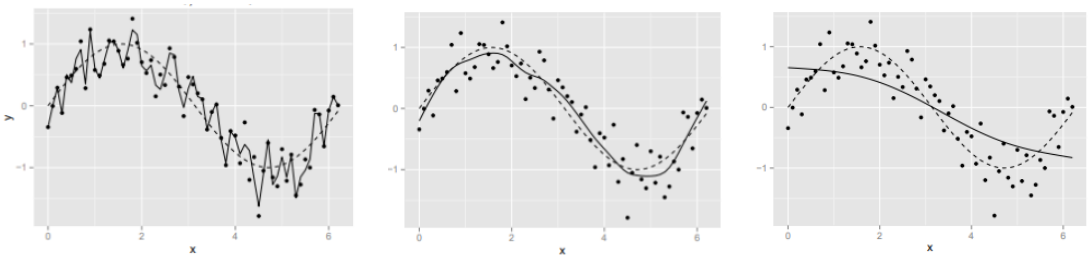
2.1.2 Introduction des notations

Dans tout ce qui suit, soit Y une variable à expliquer (ou encore réponse ou variable cible) et $X_1, ..., X_p$ les p variables explicatives (ou encore prédicteurs ou variables de contrôle). On se place dans un contexte d'apprentissage supervisé et on observe un échantillon statistique $\mathcal{D}_n = \{(X_i, Y_i)\,;\, i = 1, \dots, n\}$ d'une loi conjointe inconnue $\mathcal{P}$ sur $\mathcal{X} \times \mathcal{Y}$, où $\mathcal{X} \subset \mathbb{R}^p$ et $\mathcal{Y} \subset \mathbb{R}_+$. Le cadre choisi est donc celui de la régression : pour le coût moyen, il s'agira de prédire des valeurs réelles positives tandis que pour la fréquence, ce seront des entiers naturels. On note $\mathbf{x} = (x_1, \dots, x_p)$ une observation de la variable $\mathbf{X} = (X_1, \dots, X_p)$ et y une observation de la variable Y. Une règle de prévision, est une fonction mesurable $f : \mathcal{X} \to \mathcal{Y}$ qui associe la sortie $f(\mathbf{x})$ à l'entrée $\mathbf{x} \in \mathcal{X}$. Pour mesurer la qualité de prévision, on introduit une fonction de perte $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ telle que $l(y, y) = 0$ et $l(y, y') > 0$ pour $y \neq y'$. Étant donnée une fonction de perte l et une règle de prévision f, le risque ou l'erreur de généralisation est défini par $R_{\mathcal{P}}(f) = \mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{P}}[l(Y, f(\mathbf{X}))]$ et décrit le comportement moyen de la fonction de perte. En l'absence de toute information ou hypothèse sur la loi $\mathcal{P}$, c'est-à-dire dans un cadre non-paramétrique, on remplace $\mathcal{P}$ par $\mathcal{P}_n$, une mesure empirique associée à l'échantillon $\mathcal{D}_n$, et on s'intéresse alors au risque empirique $\hat{R}_{\mathcal{P}_n}(f, \mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n} l(Y_i, f(\mathbf{X}_i))$ qui exprime la qualité d'ajustement du modèle sur l'échantillon observé. La minimisation du risque empirique est une extension de la procédure d'estimation d'un modèle. Un prédicteur f^* est dit optimal, ou oracle, si $R_{\mathcal{P}}(f^*) = inf_{f\in\mathcal{F}} R_{\mathcal{P}}(f)$, où $\mathcal{F}$ désigne la classe de tous les prédicteurs. L'objectif est alors de déterminer un modèle $F \in \mathcal{F}$ pour lequel le risque de l'estimateur $\hat{f}_F(\mathcal{D}_n)$ est proche de celui de l'oracle. En particulier, on souhaite minimiser l'excès du risque qui se décompose en deux parties

$$\hat{R}_{\mathcal{P}_n}(\hat{f}_F(\mathcal{D}_n)) - R_{\mathcal{P}}(f^*) = \underbrace{\left\{\hat{R}_{\mathcal{P}_n}(\hat{f}_F(\mathcal{D}_n)) - inf_{f\in F} R_{\mathcal{P}}(f)\right\}}_{\text{erreur d'estimation : variance}} + \underbrace{\left\{inf_{f\in F} R_{\mathcal{P}}(f) - R_{\mathcal{P}}(f^*)\right\}}_{\text{erreur d'approximation : biais}}. \quad (2.1)$$

L'erreur d'estimation, ou encore l'erreur stochastique ou la variance, est l'erreur causée par le choix de la règle de prévision. L'erreur d'approximation, ou encore l'erreur systématique ou le biais, est, quant à elle,

liée à la complexité de l'espace ou de la classe dans laquelle la solution est recherchée. Très généralement, plus un modèle est complexe, plus il est flexible et peut s'ajuster aux données, et donc plus le biais diminue. En revanche, la complexité du modèle augmentant avec le nombre de paramètres à estimer, cette diminution du biais s'accompagne souvent d'une augmentation de la variance. Un tel modèle peut s'avérer défaillant lorsqu'il s'agira de généraliser, c'est-à-dire s'appliquer aux données qui n'ont pas participé à son estimation. L'objectif est alors de construire un modèle qui capture suffisamment la complexité des données mais qui ne surapprend pas, un modèle qui réalise l'équilibre entre le biais et la variance. La Figure 2.1 montre un exemple d'ajustement aux données bruitées supposées suivre un signal sinusoïdal avec des fonctions splines pénalisées (pour plus de détails voir la Section 3.2.2). La première fonction spline s'ajuste bien aux données, la troisième pas assez. Toutes les deux répliquent mal le phénomène sinusoïdal étudié, l'une en apprenant du bruit, l'autre en sous-apprenant la tendance. Le meilleur modèle est celui proposé par la deuxième fonction spline : un compromis entre le biais et la variance. Ce paradigme va guider les réflexions de ce chapitre.



(a) Surajustement (b) Ajustement optimal (c) Sous-ajustement

FIGURE 2.1 – Exemple d'ajustement aux données bruitées (points) suivant un signal sinusoïdal (trait en pointillé) avec des fonctions splines pénalisées (trait plein). Source : LARSEN (2015).

2.2 Du modèle linéaire au modèle linéaire généralisé

Le modèle linéaire généralisé est une extension du modèle linéaire introduite dans les années 1970 (J. A. NELDER et WEDDERBURN (1972)). Transformant la structure de dépendance linéaire entre la variable réponse et les prédicteurs par des fonctions adaptées, il constitue aujourd'hui l'approche paramétrique standard pour la tarification en assurance non-vie. Les notions d'estimation des paramètres, de tests de significativité et de qualité d'ajustement ne seront pas détaillées pour le modèle linéaire. En revanche, elles seront explicitées pour le modèle linéaire généralisé.

2.2.1 Modèle linéaire

L'écriture du modèle linéaire (*Linear Model*, **LM**) consiste à supposer que les p variables explicatives vérifient

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p} + \epsilon_i,\ i = 1, 2, \ldots, n \tag{2.2}$$

avec les hypothèses suivantes :

1. Les termes $X_1, ..., X_p$ sont déterministes ;
2. Les coefficients réels $\beta_0, ..., \beta_p$ sont constants ;

3. Les ϵ_i sont des termes d'erreur indépendants identiquement distribués de loi normale $\mathcal{N}(0, \sigma^2)$.

Matriciellement, $Y = X\beta + \epsilon$. X est appelée la matrice modèle (ou matrice de *design*) et les paramètres $\beta_0, \ldots, \beta_p$ sont estimés par minimisation du critère des moindres carrés (**MC**)

$$\hat{\beta}_{MC} = \arg\min_{\beta} ||Y - X\beta||^2 \tag{2.3}$$

On montre alors que

$$\hat{\beta}_{MC} = (X'X)^{-1}X'Y.$$

La qualité d'ajustement peut être évaluée à l'aide du coefficient de détermination R^2. Ce dernier donne la part de variation de Y expliquée par le modèle linéaire et est défini par

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \in [0, 1], \tag{2.4}$$

où $\hat{y}_i$ est la valeur prédite correspondante à y_i et $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Plus le coefficient de détermination est proche de 1, meilleur est l'ajustement du modèle. Si le R^2 constitue un indice de qualité, il a la propriété d'être croissant avec le nombre de variables. Il ne peut donc servir qu'à comparer deux modèles de même niveau, c'est-à-dire ayant le même nombre de variables explicatives.

2.2.2 Modèle linéaire généralisé

Les modèles linéaires généralisés (*Generalized Linear Model*, **GLM**) sont caractérisés par trois éléments : une composante aléatoire, une composante déterministe et une relation fonctionnelle entre la composante aléatoire et déterministe (McCullagh et J. A. Nelder (2019)).

Composante aléatoire

La composante aléatoire identifie la distribution de probabilité de la variable à expliquer Y. On observe n variables aléatoires $\underline{Y} = (Y_1, \ldots, Y_n)$ indépendantes et on suppose qu'elles admettent des distributions issues d'une structure exponentielle, c'est-à-dire que la loi de Y est dominée par une mesure de référence et que la vraisemblance de Y calculée en y par rapport à cette mesure s'écrit

$$f_Y(y, \omega, \phi) = \exp\left\{\frac{y\omega - b(\omega)}{\gamma(\phi)} + c(y, \phi)\right\}, \tag{2.5}$$

où $b(\cdot)$, $c(\cdot)$ et $\gamma(\cdot)$ sont des fonctions et ω et ϕ sont des paramètres. Cette formulation inclut la plupart des lois usuelles, notamment la loi gaussienne, binomiale, Poisson, Gamma et Gamma inverse. L'écriture de l'équation 2.5 fait apparaître une propriété fondamentale de la famille exponentielle : $\mathbb{E}[Y] = b'(\omega)$ et $Var(Y) = b''(\omega)\,\gamma(\phi)$. Le paramètre ω est appelé paramètre naturel de la famille exponentielle. Si la fonction γ est de la forme $\gamma(\phi) = \phi$, ce qui est le cas pour la loi gaussienne et la loi Gamma inverse, le paramètre ϕ est appelé paramètre de dispersion. L'expression de la structure exponentielle (Équation 2.5) se met alors sous la forme canonique en posant

$$f_Y(y,\omega) = a(\omega)\, d(y)\, \exp(yQ(w)),$$

avec $a(\omega) = \exp(-\frac{b(\omega)}{\phi})$, $d(y) = \exp(c(y,\phi))$ et $Q(\omega) = \frac{\omega}{\phi}$.

Example 2.1. La loi Gamma peut être utilisée pour modéliser des variables unimodales à valeurs positives comme des durées de vie ou des coûts de sinistres. Elle est particulièrement adaptée pour des distributions fortement asymétriques avec une décroissance rapide en épaisseur de queue. Une distribution Gamma $\mathcal{G}(\alpha,\beta)$ est caractérisée par deux paramètres α et β qui affectent respectivement la forme et l'échelle de sa représentation graphique. Sa densité est donnée par

$$f_Y(y) = \frac{\beta^\alpha y^{\alpha-1}}{\Gamma(\alpha)} \exp(-\beta y) = \exp\Big(-\alpha\big(\frac{\beta}{\alpha}y - \ln(\frac{\beta}{\alpha})\big) + \alpha\ln(\alpha) + (\alpha-1)\ln(y) - \ln(\Gamma(\alpha))\Big),$$

où $y \in \mathbb{R}_+^*$ et $\alpha,\beta > 0$. Nous retrouvons donc l'écriture de l'équation 2.5 avec

$$\begin{cases} \omega & = \frac{\beta}{\alpha} \\ b(\omega) & = \ln(\frac{\beta}{\alpha}) \Rightarrow b'(\omega) = \frac{\alpha}{\beta} = \mathbb{E}[Y] \\ \gamma(\phi) & = -\frac{1}{\alpha} \Rightarrow b''(\omega)\gamma(\phi) = \frac{\alpha}{\beta^2} = Var(Y) \\ c(y,\phi) & = \alpha(\ln(\alpha)) + (\alpha-1)\ln(y) - \ln(\Gamma(\alpha)). \end{cases}$$

Example 2.2. Pour modéliser un phénomène de fréquence, nous pouvons utiliser la loi de Poisson qui est une loi de probabilité discrète décrivant le comportement du nombre d'occurrences d'un événement dans un intervalle de temps fixé, indépendamment du temps écoulé depuis l'événement précédent. La densité de la loi de Poisson $\mathcal{P}(\lambda)$ est donnée par

$$f_Y(y) = \exp(-\lambda)\,\frac{\lambda^y}{y!} = \exp(-\lambda + y\ln(\lambda) - \ln(y!)),$$

où $y \in \mathbb{N}$ et $\lambda > 0$. Nous retrouvons donc l'écriture de l'équation 2.5 avec

$$\begin{cases} \omega & = \ln(\lambda) \\ b(\omega) & = \lambda = \exp(\omega) \Rightarrow b'(\omega) = b''(\omega) = \lambda = \mathbb{E}[Y] = Var(Y) \\ \gamma(\phi) & = 1 \\ c(y,\phi) & = -\ln(y!). \end{cases}$$

Composante déterministe

La composante déterministe du modèle est le prédicteur linéaire, soit le vecteur à n composantes $\eta = X\beta$.

Fonction de lien

La fonction de lien g établit la relation entre le prédicteur linéaire $\eta = X\beta$ et la moyenne μ de la distribution de Y. Elle est supposée monotone et différentiable. L'écriture du modèle linéaire généralisé consiste alors à supposer que les p variables explicatives vérifient

$$g(\mathbb{E}[Y_i]) = \underbrace{\beta_0}_{\text{effet moyen}} + \underbrace{\beta_1 X_{i,1}}_{\text{effet } X_1} + \underbrace{\beta_2 X_{i,2}}_{\text{effet } X_2} + \ldots + \underbrace{\beta_p X_{i,p}}_{\text{effet } X_p}, \; i = 1, 2, \ldots, n \tag{2.6}$$

La fonction de lien qui associe la moyenne μ au paramètre naturel de la famille exponentielle ω est appelée fonction de lien canonique. Dans ce cas, $g(\mu) = \omega = \eta = X\beta$. Pour la loi Gamma, la fonction de lien canonique est donnée par la fonction inverse. Pour la loi de Poisson, il s'agit de la fonction logarithme. Dans le cas de la distribution normale, la fonction de lien identité permet de retrouver le modèle linéaire (Équation 2.2). La moyenne de la variable réponse est alors exprimée directement en fonction du prédicteur linéaire et donne lieu à un modèle additif. En d'autres mots, les effets de chaque variable explicative sont sommés. L'utilisation de la fonction de lien logarithme dans le cas de la distribution Gamma et Poisson conduit à un modèle multiplicatif. En effet, $\mathbb{E}[Y] = g^{-1}(\eta) = g^{-1}(X\beta)$ et l'inverse de la fonction logarithme étant la fonction exponentielle, on a

$$\mathbb{E}[Y_i] = \exp(\beta_0) \times \exp(\beta_1 X_{i,1}) \times \exp(\beta_2 X_{i,2}) \times \ldots \exp(\beta_p X_{i,p}), \; i = 1, 2, \ldots, n. \tag{2.7}$$

Le lien logarithmique n'est pas le lien canonique associé à la loi Gamma. Néanmoins, son utilisation permet d'obtenir une grille tarifaire multiplicative immédiatement interprétable et explicable. Le choix du lien logarithmique est ainsi devenu très classique dans la modélisation en assurance non-vie.

Estimation des paramètres

Pour estimer les paramètres $\beta_0, \ldots, \beta_p$, on utilise la méthode du maximum de vraisemblance (**EMV**). Par indépendance des observations, la vraisemblance du n-échantillon des variables indépendantes $\underline{Y} = (Y_1, \ldots, Y_n)$ est donnée par

$$\beta \to L(\underline{Y}, \beta) = \prod_{i=1}^{n} f_{Y_i}(y_i, \omega_i).$$

La log-vraisemblance vaut

$$\beta \to l(\underline{Y}, \beta) = \sum_{i=1}^{n} \ln(f_{Y_i}(y_i, \omega_i)),$$

où β, η, μ et ω sont liés par le modèle. L'estimateur du maximum de vraisemblance associé vérifie donc $\hat{\beta}_{EMV} \in \arg\max_\beta L(\underline{Y}, \beta) = \arg\max_\beta l(\underline{Y}, \beta)$. En particulier, si g est la fonction de lien canonique, on a $\omega_i = \mathbf{x}_i\beta$ et donc

$$l(\underline{Y}, \beta) = \sum_{i=1}^{n} \frac{y_i \mathbf{x}_i \beta - b(\mathbf{x}_i\beta)}{\gamma(\phi)} + c(y, \phi).$$

On s'intéresse alors au score

$$S(\underline{Y}, \beta) = \Big(\frac{\partial}{\partial \beta_0} l(\underline{y}, \beta), \ldots, \frac{\partial}{\partial \beta_p} l(\underline{y}, \beta) \Big)' \tag{2.8}$$

L'estimateur du maximum de vraisemblance vérifie donc $S(\underline{Y}, \hat{\beta}_{EMV}) = \mathbf{0_{p+1}}$. Si g est la fonction de lien canonique

$$\forall\, j = 0, \dots, p,\ \frac{\partial}{\partial \beta_j} l(\underline{Y}, \beta) = \sum_{i=1}^{n} \frac{x_{i,j}}{\gamma(\phi)} (y_i - b'(\mathbf{x}_i \beta)) = 0.$$

Le système n'est linéaire en β que si $b'(x) = x$, c'est-à-dire dans le cas du modèle linéaire (Équation 2.2). Pour tous les autres modèles linéaires généralisés (Équation 2.6), le système est non-linéaire et il n'existe pas toujours de formule analytique pour cet estimateur. Cependant, la détermination de $\hat{\beta}_{EMV}$ est associée à un problème d'optimisation convexe qui être traité par un algorithme de type Newton-Raphson, un algorithme itératif basé sur le développement de Taylor à l'ordre 1 du score (Équation 2.8). Pour plus de détails, il est possible de consulter l'algorithme IRLS (*Iteratively Reweighted Least Squares*, McCullagh et J. Nelder (1989)). L'évaluation de la qualité d'un modèle ainsi construit s'articule autour de trois axes :

- la significativité des variables explicatives ;
- l'ajustement du modèle ;
- la robustesse du modèle.

Test de significativité

Dans le cadre du modèle linéaire généralisé, la loi de l'estimateur du maximum de vraisemblance n'est connue qu'asymptotiquement. Aussi, les procédures de test sont menées dans un cadre asymptotique. On distingue généralement le test des modèles emboîtés et le test d'un paramètre β_j.

Le test des modèles emboîtés permet de déterminer si un sous-ensemble de variables explicatives est suffisant pour expliquer la cible Y. Soient deux modèles emboîtés M_1 et M_2 définis respectivement par $g(\mu) = X_1\beta_1$ et $g(\mu) = X_2\beta_2$, avec M_1 un sous-modèle de M_2. On veut tester

$$(H_0)\ :\ M_1 \qquad \textit{contre} \qquad (H_1)\ :\ M_2.$$

On considère alors le test du rapport vraisemblance dont la statistique de test est donnée par

$$T = -2 \ln \left[\frac{L(\underline{Y}, \hat{\beta}_1)}{L(\underline{Y}, \hat{\beta}_2} \right] = -2\big(l(\underline{Y}, \hat{\beta}_1) - l(\underline{Y}, \hat{\beta}_2)\big), \tag{2.9}$$

où $\hat{\beta}_1$ et $\hat{\beta}_2$ sont les estimateurs du maximum de vraisemblance dans le modèle M_1 et M_2 respectivement. Sous certaines conditions de régularité de la densité de probabilité,

$$T \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(k_2 - k_1),$$

avec k_1 et k_2 les dimensions des sous-espaces engendrés par les colonnes de X_1 et X_2 respectivement. L'hypothèse (H_0) est donc rejetée si

$$T > v_{1-\alpha,\, k_2 - k_1},$$

où $v_{1-\alpha,k_2-k_1}$ désigne le $(1-\alpha)$-quantile de la loi $\chi^2(k_2-k_1)$. Ce test peut également être vu sous l'angle de la déviance, définie comme l'écart entre la log-vraisemblance du modèle d'intérêt M et celle du modèle saturé M_{sat}. Le modèle saturé est le modèle le plus complet possible, c'est-à dire comportant autant de paramètres que d'observations. La déviance du modèle M est donnée par

$$\mathcal{D}(M) = -2\big(l(\underline{Y},\hat{\beta}) - l(\underline{Y},\hat{\beta}_{sat})\big). \tag{2.10}$$

La statistique de test T (Équation 2.9) se réécrit alors comme $T = \mathcal{D}(M_1) - \mathcal{D}(M_2)$.

Pour évaluer la significativité d'un paramètre β_j dans le modèle, on met en oeuvre un Z-test. On souhaite tester

$$(H_0) \ : \ \beta_j = 0 \qquad \textit{contre} \qquad (H_1) \ : \ \beta_j \neq 0.$$

Sous (H_0), on a l'approximation suivante

$$\big(\hat{\beta}_{EMV}\big)_j \overset{\mathcal{L}}{\simeq} \mathcal{N}\big(0, [\mathcal{I}_n(\hat{\beta}_{EMV})^{-1}]_{jj}\big),$$

où $\mathcal{I}_n(\beta) = X'WX$ désigne la matrice d'information de Fisher avec W la matrice diagonale de pondération définie par $[W]_{ii} = \frac{1}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$. L'hypothèse (H_0) est donc rejetée si

$$T_j = \frac{|(\hat{\beta}_{EMV})_j|}{\sqrt{[\mathcal{I}_n(\hat{\beta}_{EMV})^{-1}]_{jj}}} > z_{1-\frac{\alpha}{2}},$$

où $z_{1-\frac{\alpha}{2}}$ est le $(1-\frac{\alpha}{2})$-quantile de la loi $\mathcal{N}(0,1)$.

Qualité d'ajustement

Les coefficients du modèle linéaire généralisé étant estimés par maximum de vraisemblance, et non par minimisation du critère des moindres carrés, le coefficient de détermination R^2 (Équation 2.4) ne constitue pas un indicateur adapté pour mesurer la qualité d'ajustement d'un modèle M. On introduit alors le pseudo R^2 défini en fonction de la déviance du modèle nul M_0, c'est-à-dire d'un modèle composé uniquement de l'intercept

$$R^2_{pseudo} = \frac{\mathcal{D}(M_0) - \mathcal{D}(M)}{\mathcal{D}(M_0)} = 1 - \frac{\mathcal{D}(M)}{\mathcal{D}(M_0)} \in [0,1]. \tag{2.11}$$

Comme pour le coefficient de détermination, plus le pseudo R^2 est proche de 1, meilleur est l'ajustement du modèle.

Robustesse du modèle

La robustesse du modèle peut être évaluée à l'aide de l'étude graphique des résidus. Les résidus bruts sont définis par $\hat{\epsilon}_i = y_i - \hat{y}_i$. Ces résidus n'ayant pas toujours la même variance, il est difficile de les comparer à un comportement type attendu. En normalisant les résidus bruts par l'écart-type s_i de $\hat{y}_i$,

on obtient les résidus standardisés de Pearson $r_P = \frac{y_i - \hat{y}_i}{s_i}$. D'autres types de résidus peuvent être définis, notamment en incluant la notion de déviance. L'étude réalisée en Annexe C.1 permet de graphiquement les résidus bruts (Figure C.1a) aux résidus de Pearson et de déviance (Figure C.1b).

Offset

Pour la modélisation de la fréquence, il est important de tenir compte de l'exposition de chaque observation. En effet, les individus constituant la base d'apprentissage ne sont pas tous observés sur une durée d'un an. L'objectif étant de prédire le nombre moyen de sinistres pour l'année suivante, un ajustement au prorata temporis doit être appliqué. L'espérance λ de la loi de Poisson $\mathcal{P}(\lambda)$ devient λe, où $e \in [0, 1]$. Appliqué à l'équation 2.7, cet ajustement devient

$$\begin{aligned}\mathbb{E}[Y_i] &= e_i \times \big(\exp(\beta_0) \times \exp(\beta_1 X_{i,1}) \times \exp(\beta_2 X_{i,2}) \times \dots \exp(\beta_p X_{i,p})\big) \\ &= \exp\big(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots \beta_p X_{i,p} + \ln(e_i)\big),\ i = 1, 2, \dots, n.\end{aligned}$$

Une variable explicative, appelée *offset*, est ainsi ajoutée dans le modèle. Son coefficient est fixé à 1 et ne doit donc pas être estimé.

2.2.3 Sélection de modèle

La modélisation statistique vise en général trois objectifs : une finalité descriptive, explicative et prédictive. Si ces objectifs sont complémentaires, il convient néanmoins de distinguer les spécificités et les limites de chacun, notamment lors d'une sélection des variables. Une finalité descriptive se résume à rechercher de façon exploratoire les liens entre Y et d'autres variables, souvent nombreuses, afin d'en sélectionner un sous-ensemble pertinent. Cette recherche peut être alimentée par une analyse des corrélations ou des analyses factorielles (Section 1.2.2). Un objectif explicatif est, quant à lui, sous-tendu par une connaissance *a priori* du domaine concerné. Il s'agit de confirmer ou d'infirmer une hypothèse en se basant sur des tests statistiques conduisant à la prise de décision désirée. Cependant, la significativité de la présence d'une variable basée sur le test de nullité de son coefficient (Section 2.2.2) n'est pas un indicateur de l'importance de cette variable pour la qualité de la prévision. C'est la finalité prédictive qui met l'accent sur la qualité de la prévision et conduit à rechercher des modèles parcimonieux afin de réduire la variance.

AIC et BIC

Le critère d'information d'Akaïke (AIC) (Akaike (1974)) et le critère d'information de Bayes (BIC) (Schwarz (1978)) sont des indicateurs basés sur la qualité de prévision permettant de comparer deux modèles non-emboîtés. L'AIC s'applique à tout modèle estimé par maximisation de vraisemblance L et se définit par

$$AIC = -2\ln(L) + 2\frac{d}{n},$$

où $d = p + 1$ désigne le nombre de paramètres du modèle. Il s'agit donc d'une pénalisation de la déviance du modèle par deux fois le nombre de paramètres libres. Le BIC cherche asymptotiquement le modèle associé à la plus grande probabilité *a posteriori* et se met sous la forme

$$BIC = -2\ln(L) + \ln(n)\frac{d}{n}.$$

Il est possible de montrer que dès que $n > e^2$, le BIC tend à pénaliser plus lourdement les modèles complexes, mais lorsque n tend vers l'infini la probabilité pour le BIC de choisir le bon modèle tend vers 1. Quelque soit le critère retenu, il suffira de choisir le modèle présentant le plus faible AIC ou BIC.

Algorithmes de sélection de variables

En pratique, lorsque p est grand, il n'est pas réaliste de tester les 2^p modèles au sens du critère AIC ou BIC. Par capacité croissante d'optimalité, les stratégies algorithmiques suivantes sont alors proposées :

- méthode ascendante (*forward selection*) : le modèle initial est vide et, à chaque pas, on ajoute la variable qui permet de réduire le plus le critère AIC du modèle obtenu. La procédure s'arrête lorsque toutes les variables ont été introduites ou lorsque l'AIC ne décroît plus ;
- méthode descendante (*backward elimination*) : le modèle initial est complet et, à chaque pas, on retire la variable qui conduit à l'AIC le plus faible. La procédure s'arrête lorsque l'AIC ne décroît plus ;
- méthode *stepwise* (*bidirectional elimination*) : une combinaison des deux méthodes précédentes qui introduit une étape d'élimination après chaque étape de sélection pour retirer du modèle les variables qui seraient devenues moins significatives de part l'introduction de nouvelles variables.

Régression régularisée ou pénalisée

Les méthodes de régularisation correspondent à l'ajout d'une pénalité envers la complexité du modèle afin de favoriser la parcimonie. Dans un souci de robustesse, il s'agit d'augmenter le biais en faveur d'une réduction de variance. Il y a principalement deux façons de pénaliser le modèle linéaire :

- contraindre les paramètres du modèle en les rétrecissant par une régression qui opère une régularisation par pénalisation l_2 ;
- simplifier le modèle par une régression Lasso (*Least Absolute Shrinkage and Selection Operator*) (Tibshirani (1996)) qui réalise une sélection des variables par pénalisation l_1.

La régression Ridge conserve toutes les variables dans le modèle mais elle les empêche de prendre de trop grandes valeurs en appliquant une pénalité de type l_2 aux paramètres. Si l'on considère le modèle linéaire (Équation 2.2), l'estimateur Ridge de $\beta = (\beta_0, \ldots, \beta_p)'$ est défini par

$$\hat{\beta}_{Ridge} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

où λ est le terme de pénalisation qui détermine le degré d'ajustement aux données. Plus la pénalité augmente, plus la solution obtenue est régulière, ou encore plus le biais augmente et la variance diminue. Par conséquent, le calibrage du paramètre λ est crucial et s'effectue par validation croisée (Section 2.3.4).

La régression Lasso est une solution qui favorise l'interprétabilité. Avec la pénalité l_1, elle privilégie naturellement les coefficients nuls et réalise ainsi une sélection des variables. L'estimateur Lasso de β est défini par

$$\hat{\beta}_{Lasso} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Si la régression Ridge réduit de façon simultanée la valeur des coefficients de variables corrélées, la régression Lasso est plutôt indifférente aux phénomènes de corrélation et a tendance à retenir arbitrairement l'une des variables parmi les variables corrélées. Graphiquement, nous pouvons l'observer avec la Figure 2.2 qui représente les courbes d'équi-erreur $||y_i - \hat{y}_i||$ de la régression non-pénalisée avec des ellipses ainsi que les zones de contrainte imposées par les pénalités l_1 et l_2. L'intersection entre les deux surfaces correspond à la solution du problème de régularisation.

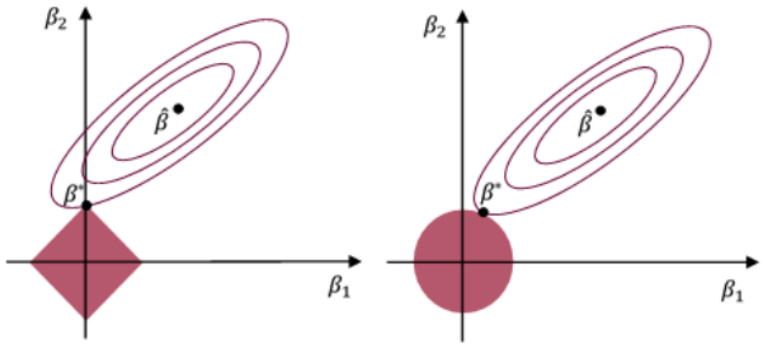


FIGURE 2.2 – Régression Lasso (à gauche) et Ridge (à droite). Source : GUILLOT (2015).

Il est possible de combiner les régressions Ridge et Lasso par l'introduction simultanée des pénalités l_1 et l_2 avec la méthode Elastic-Net. Deux paramètres sont alors à calibrer par validation croisée : λ et α. Le paramètre λ correspond au terme de pénalisation global, la valeur $\alpha = 0$ à la régression Ridge et la valeur $\alpha = 1$ à la régression Lasso. L'estimateur Elastic-Net est ainsi défini par

$$\hat{\beta}_{Elastic-Net} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} X_{i,j}\beta_j \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \right\}.$$

Pour implémenter les différentes méthodes de régression régularisée, nous utilisons le package `glmnet` (J. FRIEDMAN et al. (2010)). En pratique, la solution Lasso est souvent retenue par les assureurs pour sa robustesse et son efficience en termes de temps de calcul, atout indispensable permettant un pilotage régulier du portefeuille. Des alternatives intéressantes combinant la régression Lasso avec le lissage des fonctions splines peuvent être proposées. A ce titre, il est possible de consulter le mémoire d'actuariat SAVOYE (2021).

Dans la section 2.2, nous avons établi le cadre théorique du modèle linéaire généralisé. Nous avons en particulier détaillé l'estimation des paramètres et la pénalisation du modèle. L'objectif de la section suivante est de sortir de ce cadre paramétrique pour envisager des méthodes plus flexibles et adaptatives avec des modèles de *Machine Learning*. Nous tâchons ainsi d'évaluer les avantages et les inconvénients du cadre non-paramétrique pour la tarification en assurance non-vie.

2.3 Modèles de Machine Learning

Dans la tradition statistique, la notion de modèle est centrale, surtout avec une finalité explicative. Il s'agit d'approcher la réalité en se basant sur une théorie physique, économique biologique ou autre. Cette dernière guidera les hypothèses probabilistes, les critères d'ajustement et les tests de validité. L'interprétation du rôle de chaque variable explicative est prépondérante dans cette démarche. En revanche, si l'objectif est essentiellement prédictif, la volumétrie des données importante et les relations entre les variables complexes, il apparaît que la notion d'apprentissage est d'une pertinence et d'une efficacité accrue (Vapnik (1999)). L'estimation devient un apprentissage, la prévision remplace l'explication. Plus flexibles et progressives, les méthodes de *Machine Learning* ouvrent un nouveau champ de possibilités dans la tarification en assurance non-vie. Les modèles ne se basent plus sur l'estimation par maximum de vraisemblance mais étendent leur construction à plusieurs fonctions de perte. De plus, il n'imposent pas une structure linéaire entre la variable à expliquer et les variables explicatives et intègrent implicitement les interactions entre les variables, qui deviennent alors une expression de la complexité du modèle.

L'objectif de ce mémoire est de proposer une nouvelle manière d'apprendre afin de construire des modèles de *Machine Learning* intrinsèquement interprétables. Les avantages et les inconvénients du cadre non-paramétrique ici présenté sont alors mesurés à l'aune de l'interprétabilité et dans l'anticipation de la méthodologie présentée dans le Chapitre 3. Nous établissons ainsi le cadre théorique des modèles de *Machine Learning*, leurs spécificités et leurs limites. Nous accordons une importance particulière à la définition mathématique des notions pour être en mesure de prendre du recul lors de la manipulation des concepts du Chapitre 3. De nombreux modèles d'apprentissage existent. Nous nous focalisons ici sur les arbres de décision dans un contexte de régression, puis sur les techniques d'agrégation des modèles ayant cette structure arborescente. Une présentation rapide des réseaux de neurones est également proposée. Pour implémenter les différents algorithmes de *Machine Learning*, nous nous appuyons essentiellement sur les packages `caret` (Kuhn (2022)), `mlr` (Bischl et al. (2016)) et `scikit-learn` (Pedregosa et al. (2011)).

2.3.1 Arbres de décision

Un arbre binaire de décision (*Classification And Regression Tree*, CART) réalise un partitionnement dyadique de l'espace grâce à une hiérarchie de questions discriminantes (Breiman et al. (2017)). L'ensemble des individus est regroupé à la racine de l'arbre puis, à chaque division, s'opère un partitionnement en deux groupes, noeuds, plus homogènes au sens du prédicteur. Soit v un noeud de l'arbre, X_j un prédicteur utilisé pour partitionner v et s_j le seuil qui divise v en deux noeuds fils, le noeud gauche $v_G := v_G(s_j)$ et le noeud droit $v_D := v_D(s_j)$. On a $v_G = \{X_{i,j} \leq s_j\}$ et $v_D = \{X_{i,j} > s_j\}$. Le critère de division repose sur la définition d'une fonction d'hétérogénéité et la longueur de chaque branche est alors proportionnelle à la décroissance de l'hétérogénéité, induite par la division. Dans le cas de la régression, l'hétérogénéité D_v du noeud v est définie par la variance

$$D_v = \frac{1}{|v|} \sum_{i \in v} (y_i - \overline{y}_v)^2,$$

où $|v|$ désigne l'effectif du noeud v. Pour chaque noeud, l'objectif est de rechercher la division qui contribuera à la plus forte décroissance de l'hétérogénéité des noeuds fils gauche et droit. Autrement dit, il s'agit de minimiser la variance intraclasse ou encore

$$\frac{|v_G|}{n}\sum_{i\in v_G}(y_i-\overline{y}_{v_G})^2+\frac{|v_D|}{n}\sum_{i\in v_D}(y_i-\overline{y}_{v_D})^2, \tag{2.12}$$

où $|v_G|$ et $|v_D|$ désigne l'effectif du noeud gauche et droit respectivement. Lorsque le partitionnement est interrompu, chaque noeud terminal de l'arbre devient une feuille à laquelle est affectée une valeur qui correspond à la moyenne des observations de cette feuille. Les sous-groupes ainsi créés respectent un critère d'homogénéité que l'ensemble des données ne respecte pas. Cependant, cette démarche fournit un arbre qui peut être excessivement raffiné et dépendant des données qui ont participé à son estimation. Pour favoriser la robustesse, un arbre optimal est à rechercher entre l'arbre trivial réduit à une seule feuille et l'arbre maximal homogène. C'est l'objectif de la procédure d'élagage (*pruning*) de l'arbre qui correspond à une étape de sélection de modèle.

Les arbres de décision ne requièrent pas d'hypothèses sur la distribution et intègrent implicitement une procédure de sélection des variables ainsi qu'une prise en compte des interactions. Si la profondeur de l'arbre est limitée, à savoir si le nombre de niveaux entre la racine et les feuilles est faible, le modèle reste intelligible. De plus, les résultats se présentent sous une forme graphique aisément interprétable, ce qui permet d'envisager une visualisation de la structure tarifante sous-jacente. Une telle représentation peut être obtenue avec le package `rpart` (Therneau et Atkinson (2022)). La Figure 2.3 montre les résultats pour la modélisation du coût moyen. Le dégradé des couleurs (du bleu clair au violet) permet d'identifier les zones de faible et de fort coût. Les interactions sont détectées par les partitionnements successifs de l'espace. On remarque alors que les variables `prix_SRA`, `age_COND` et `anc_COND` sont particulièrement discriminantes pour la segmentation du coût des sinistres. Pour la modélisation de la fréquence (Figure 2.4), le dégradé du jaune clair à l'orange montre les zones de faible et de forte fréquence. Hormis l'exposition, il met en relief les variables `age_COND` et `cof_CRM`.

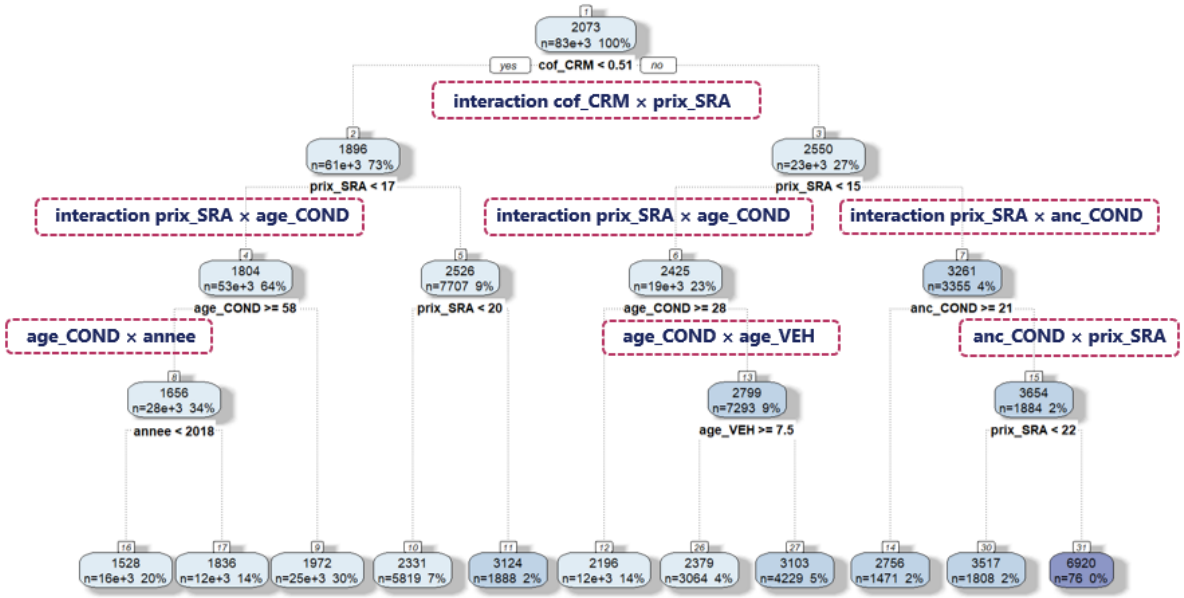


Figure 2.3 – Arbre de décision pour la modélisation du coût moyen

Les arbres de décision souffrent cependant d'une grande instabilité due à la structure hiérarchique

de l'algorithme. En effet, une erreur de division en début de l'arbre est propagée tout au long de la construction. Par essence, le modèle ainsi construit manque de régularité, surtout si le phénomène à modéliser est lui-même régulier. Plus qu'un algorithme de prédiction robuste, un arbre de décision doit donc être vu comme une méthode de partitionnement de l'espace et de segmentation du risque. Si l'algorithme CART est majoritairement utilisé dans les arbres de décision, d'autres alternatives existent. En effet, lors du partitionnement, l'algorithme CART peut souffrir d'un biais dans la sélection des variables optimales (Dutang et Guibert (2022)). De nombreux autres algorithmes ont alors été proposés, notamment la méthode *Model Based trees* (MOB) reposant sur une sélection des variables de partitionnement selon un test de fluctuation et une construction d'un sous-modèle local par noeud (Zeileis et al. (2008)).

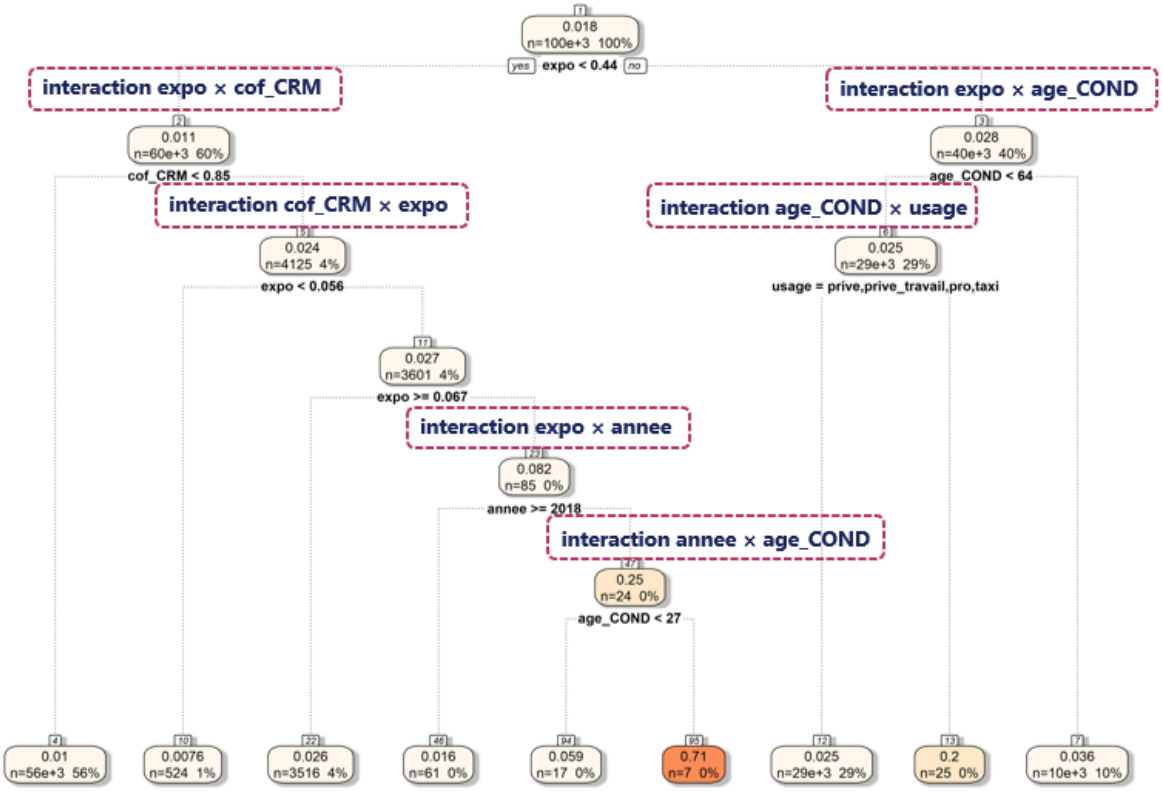


Figure 2.4 – Arbre de décision pour la modélisation de la fréquence

2.3.2 Agrégation des modèles

Les méthodes d'agrégation en *Machine Learning* sont basées sur l'hypothèse que la combinaison de plusieurs modèles sous-optimaux permet d'aboutir à un modèle collectif performant. Les défauts des modèles faibles, tels qu'une forte instabilité ou une complexité insuffisante, deviennent alors un attribut nécessaire à la réduction de la variance et du biais par une agrégation astucieuse. Les principes d'agrégation peuvent s'appliquer à toute méthode de modélisation mais n'apportent un réel intérêt que dans le cas des modèles peu robustes. Ainsi, l'utilisation de ces algorithmes n'a guère de sens pour la régression linéaire mais s'avère particulièrement pertinente pour les arbres de décision (Section 2.3.1) dont la structure arborescente est suffisamment flexible tout en étant perfectible. Deux grands paradigmes de modélisation sont à distinguer :

- le *Bagging* qui repose sur une construction itérative d'une famille de modèles avec une stratégie aléatoire ;
- le *Boosting* qui repose sur une construction récurrente d'une famille de modèles avec une stratégie adaptative.

Les notions de *Bagging* et *Boosting* seront présentées de manière détaillée car elles constituent l'un des piliers de la méthode présentée dans le Chapitre 3. De nombreuses illustrations seront alors proposées pour accompagner les formules mathématiques. Un résumé des enjeux principaux soulevés les méthodes de *Bagging* et de *Boosting* est donné par la Figure 2.5.

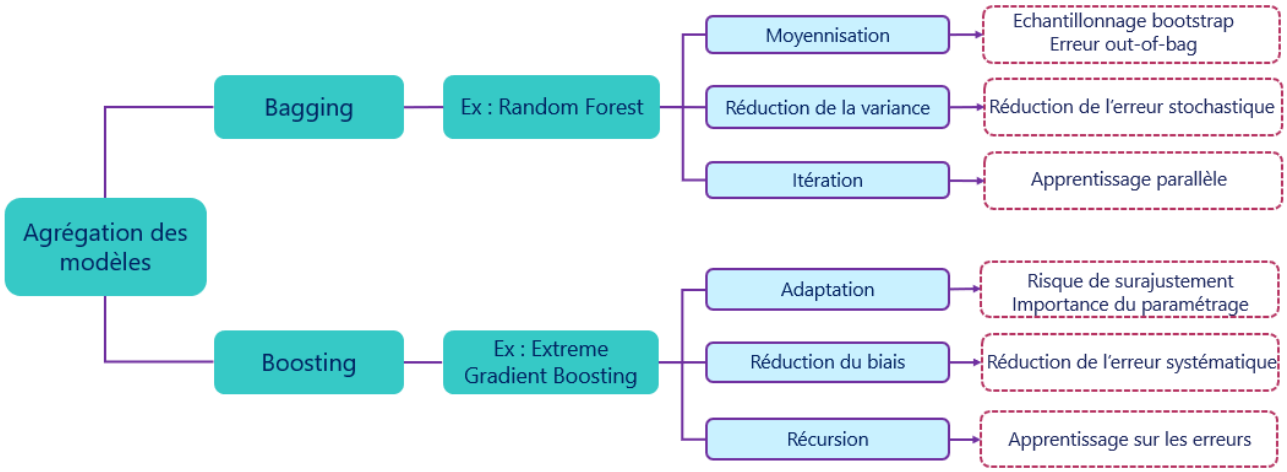


FIGURE 2.5 – Résumé des techniques d'agrégation des modèles en *Machine Learning*.

Bagging et l'algorithme Random Forest

Le principe du *Bagging* (contraction de *Bootstrap Aggregating*) consiste à moyenniser les prévisions de plusieurs modèles indépendants afin de réduire la variance. En reprenant les notations de la Section 2.1, soit $\mathbf{z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ un échantillon de loi $\mathcal{P}$ et $f(\mathbf{x})$ un modèle en fonction de $\mathbf{x} = (x_1, \ldots, x_p)$. En considérant B échantillons indépendants notés $\{z_b\}_{b=1,\ldots,B}$, une estimation par la stratégie de *Bagging* est donnée par

$$\hat{f}_B(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{z_b}(\mathbf{x}).$$

Il s'agit donc d'une moyenne des résultats obtenus pour les modèles associés à chaque échantillon. Il est naturel d'accompagner ce calcul par une erreur *out-of-bag* (OOB). Pour la i-ème observation $(\mathbf{x}_i, y_i)$, on calcule la moyenne $\hat{y}^{(i)}(\mathbf{x}_i)$ des $\hat{f}_{z_b}(\mathbf{x}_i)$ pour les échantillons z_b ne contenant pas la i-ème observation

$$\hat{y}^{(i)}(\mathbf{x}_i) = \frac{1}{B_i} \sum_{b : i \notin z_b} \hat{f}_{z_b}(\mathbf{x}_i).$$

L'erreur *out-of-bag* due à la i-ème observation est alors mesurée par $(y_i - \hat{y}^{(i)})^2$ et l'erreur *out-of-bag*

globale vaut

$$err_{OOB} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}^{(i)})^2. \tag{2.13}$$

Cependant, il n'est pas réaliste de considérer B échantillons indépendants. Ces échantillons sont donc remplacés par B réplications d'échantillons *bootstrap* obtenus chacun par n tirages avec remise selon la mesure empirique $\mathcal{P}_n$. Les distributions des échantillons *bootstrap* sont identiques et on montre que $\forall\ b = 1, \ldots, B$, $\mathbb{E}[\hat{f}_B(\mathbf{x})] = \mathbb{E}[\hat{f}_{z_b}(\mathbf{x})]$. Autrement dit, le biais de l'estimateur *baggé* est le même que celui des estimateurs agrégés. En notant $\sigma^2(\mathbf{x}) = Var(\hat{f}_{z_b}(\mathbf{x}))$, $b = 1, \ldots, B$ et $\rho(\mathbf{x}) = Corr(\hat{f}_{z_b}(\mathbf{x}), \hat{f}_{z'_b}(\mathbf{x}))$, $b \neq b'$, il vient que $Var(\hat{f}_B(\mathbf{x})) = \rho(\mathbf{x})\sigma^2(\mathbf{x}) + \frac{1-\rho(\mathbf{x})}{B}\sigma^2(\mathbf{x})$. Donc pour B suffisamment grand, la variance de l'estimateur *baggé* est en général inférieure à celle des estimateurs agrégés car on a souvent $\rho(\mathbf{x}) < 1$. Ainsi, le *bagging* apporte un gain sur la variance lorsque les estimateurs agrégés sont différents, c'est-à-dire décorrélés. Les arbres de décision étant très sensibles aux variations d'échantillonnage, ils constituent des candidats idéaux pour l'agrégation des estimateurs.

Dans le cas spécifique des arbres de décision, Breiman (1996) propose une amélioration du *Bagging* par l'ajout d'une composante aléatoire. L'algorithme, connu sous le nom de *Random Forest* (**RF**), ou forêt aléatoire, consiste en effet à appliquer la stratégie de *Bagging* aux arbres de décision en ajoutant un tirage aléatoire de m variables explicatives parmi les p (Figure 2.6). Ainsi, à chaque étape de construction d'un arbre, d'autres variables sont nécessairement mises en avant. La sélection aléatoire d'un nombre $m \ll p$ de prédicteurs potentiels accroît donc significativement la variabilité. Chaque modèle de base est moins performant, sous-optimal, mais l'agrégation conduit finalement à des bons résultats. Comme pour le *Bagging*, l'évaluation itérative de l'erreur *out-of-bag* (Équation 2.13) permet de contrôler le nombre B d'arbres dans la forêt. Le nombre m de variables tirées aléatoirement peut être un paramètre sensible, avec le choix par défaut $m = \frac{p}{3}$ dans un problème de régression pas toujours optimal. Dans tous les cas, il s'agit de trouver un bon compromis entre volume de calculs et qualité de prévision (Section 2.3.4).

Si l'algorithme *Random Forest* a le potentiel d'augmenter significativement la qualité de prévision d'un arbre de décision seul, il génère une perte d'interprétabilité. Néanmoins des informations pertinentes sur le fonctionnement de l'algorithme peuvent être obtenues par le calcul et la représentation graphique des indices proportionnels à l'importance des variables dans le modèle agrégé. Deux critères sont ainsi proposés pour évaluer l'importance d'une variable $X_j \in \mathcal{X}$ dans le modèle :

- la *Mean Decrease Accuracy* qui repose sur une permutation aléatoire des valeurs de la variable X_j. Plus la qualité de prévision, estimée par l'erreur *out-of-bag* (Équation 2.13), est dégradée par la permutation des valeurs de X_j, plus cette variable est importante. Il s'agit d'une mesure globale de l'importance des variables ;
- la *Mean Decrease Gini* qui est basée sur la décroissance de l'hétérogénéité (Équation 2.12). L'importance de la variable X_j correspond alors à la somme pondérée des décroissances d'hétérogénéité induites lorsqu'elle est utilisée pour définir la division associée à un noeud. Il s'agit d'une mesure locale de l'importance des variables.

Boosting et l'algorithme Extreme Gradient Boosting

Le *Boosting* a pour but d'améliorer un modèle prédictif faible, c'est-à-dire prédisant légèrement mieux que le hasard, par combinaison avec d'autres modèles faibles afin de produire un modèle collectif fort. L'idée originale de Schapire de 1990 a été affinée dans l'algorithme *Adaboost* (contraction de *Adaptative*

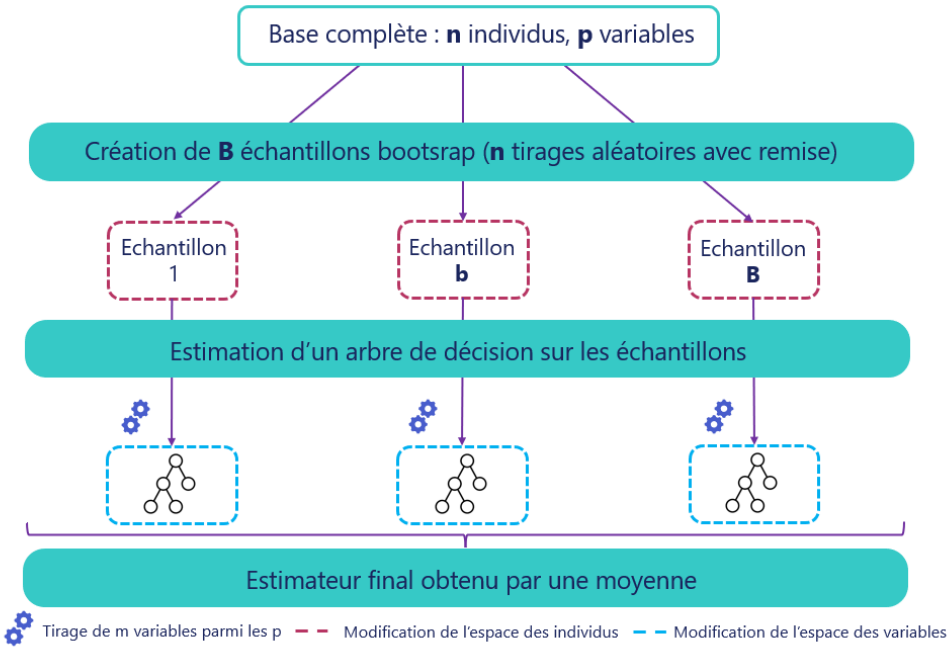


FIGURE 2.6 – Fonctionnement de l'algorithme *Random Forest*

Boosting) pour la prévision d'une variable binaire (FREUND, SCHAPIRE et al. (1996)) et adaptée depuis au cadre de la régression. Tout comme dans la procédure de *Bagging*, il s'agit de construire une famille de modèles qui seront ensuite agrégés par une moyenne pondérée. Cependant, la famille des modèles est construite de manière récurrente, chaque modèle est une version adaptative du précédent. En effet, il s'agit de donner plus de poids, lors de l'estimation suivante, aux observations mal ajustées lors de l'estimation précédente. L'algorithme concentre donc ses efforts sur les observations les plus difficiles à ajuster, réduisant ainsi le biais, tandis que l'agrégation de l'ensemble des modèles réduit la variance.

Dans l'algorithme *Adaboost*, les arbres de décision sont utilisés comme modèles de base (Figure 2.7). Dans un problème de classification, les poids de chaque observation sont initialisés à $\frac{1}{n}$, puis évoluent pour l'estimation de chaque nouveau modèle. Dans un problème de régression, plutôt que de jouer sur les pondérations, un nouvel échantillon avec remise est tiré à chaque itération selon une loi p initialisée par la distribution uniforme. Pour un échantillon $\mathbf{z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ et un modèle $f(\mathbf{x})$, notons z_m un échantillon tiré dans $\mathbf{z}$ selon $\mathbf{p}$ et $\hat{f}_{z_m}$ une estimation sur z_m. A chaque itération $m = 1, \ldots, M$ et pour chaque individu $i = 1, \ldots, n$, il s'agit alors de calculer

$$\begin{cases} l_m(i) & = l(y_i, \hat{f}_{z_m}(\mathbf{x}_i)), & \text{la fonction de perte ;} \\ \hat{\epsilon}_m & = \sum_{i=1}^{n} p_i l_m(i), & \text{le taux d'apprentissage ;} \\ \omega_i & = g(l_m(i)) p_i, & \text{les poids avec } g \text{ une fonction continue croissante ;} \\ p_i & = \frac{\omega_i}{\sum_{i=1}^{n} \omega_i}, & \text{les nouvelles probabilités.} \end{cases}$$

Pour le rééchantillonnage, la procédure *bootstrap* est ainsi répliquée, mais avec des probabilités inversement proportionnelles à la qualité d'ajustement de l'itération précédente. L'estimateur final est donnée par

$$\hat{f}_M(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} \ln\left(\frac{1}{\hat{\epsilon}_m}\right)\hat{f}_{z_m}(\mathbf{x}).$$

Le taux d'apprentissage $\hat{\epsilon}_m$ rend la procédure de *Boosting* plus ou moins conservatrice. En effet, si $\hat{\epsilon}_m < 1$, moins de corrections sont appliquées au prochain modèle ajouté à la famille, ce qui a pour conséquence de ralentir le processus d'apprentissage, et donc d'éviter le risque de surajustement.

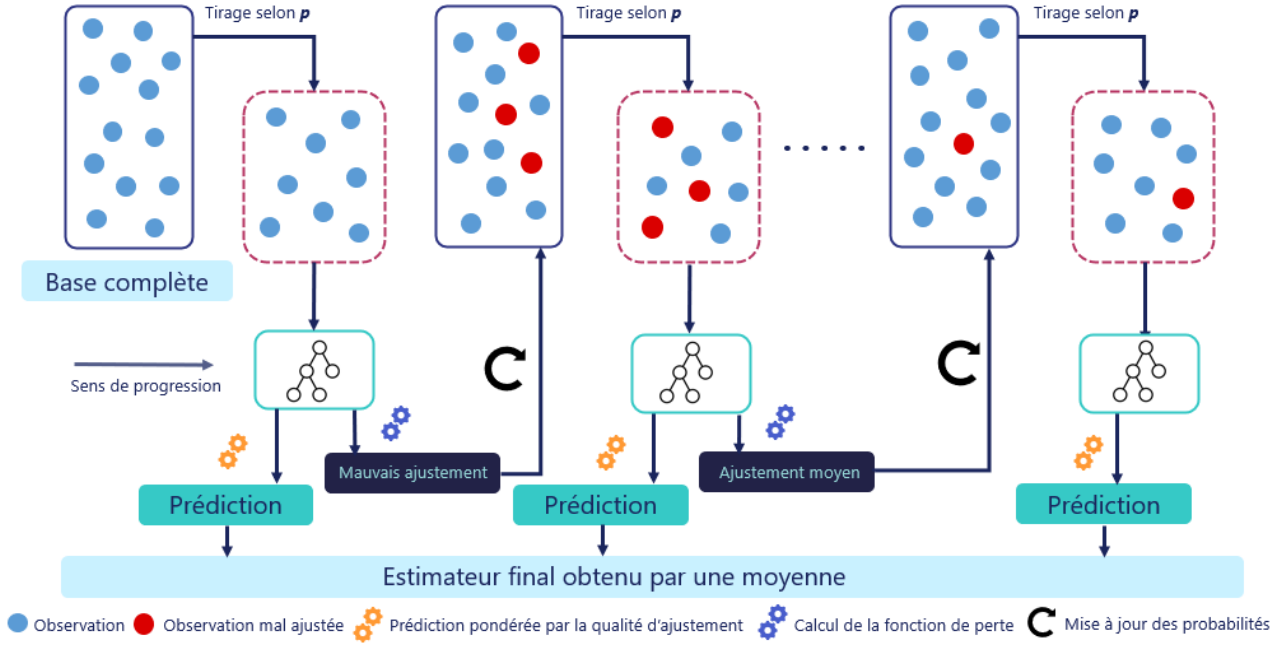


FIGURE 2.7 – Fonctionnement de l'algorithme *Adaboost*

Le *Gradient Boosting* (**GB**) est une adaptation de l'algorithme *Adaboost* basée sur une fonction de perte l convexe et différentiable (J. H. FRIEDMAN (2002)). Les modèles sont agrégés de telle sorte que l'ajout d'un nouveau modèle à la combinaison marque un pas vers une meilleure solution. Pour améliorer les propriétés de convergence, ce pas est franchi dans la direction du gradient de la fonction de perte. A l'étape m, l'estimation devient donc

$$\hat{f}_{z_m}(\mathbf{x}) = \hat{f}_{z_{m-1}}(\mathbf{x}) - \gamma \sum_{i=1}^{n} \nabla_{\hat{f}_{z_{m-1}}} l(y_i, \hat{f}_{z_{m-1}}(\mathbf{x_i})),$$

où γ désigne le meilleur pas de descente déterminé par

$$min_\gamma \sum_{i=1}^{n} \left[l(y_i, \hat{f}_{z_{m-1}}(\mathbf{x_i})) - \gamma \frac{\partial l(y_i, \hat{f}_{z_{m-1}}(\mathbf{x_i}))}{\partial \hat{f}_{z_{m-1}}(\mathbf{x_i})} \right].$$

Le *Gradient Boosting* constitue ainsi une méthode de descente dans laquelle le modèle est corrigé à mesure que les itérations se succèdent, et devient particulièrement adapté aux données. Une amélioration du *Gradient Boosting* a été proposée récemment avec l'algorithme *Extreme Gradient Boosting* (**XGB**) (CHEN et GUESTRIN (2016)). Trois axes nouveaux sont ainsi envisagés dans la procédure de *Boosting* :

- la parallélisation des calculs ;
- l'ajout d'un terme de régularisation ;
- la gestion des valeurs manquantes.

L'ajout d'un terme de régularisation est particulièrement intéressant pour le contrôle du surapprentissage. En effet, une nouvelle fonction objectif $\mathcal{L}$ est alors considérée en complétant la fonction de perte l par un terme de pénalisation. Elle est définie par

$$\mathcal{L}(f) = \sum_{i=1}^{n} l(y_i, f_M(\mathbf{x}_i)) + \sum_{m=1}^{M} \Omega(f_{z_m}),$$

où $\Omega(f) = \alpha|f| + \frac{1}{2}\beta||\mathbf{w}||^2$ avec $|f|$ le nombre de feuilles de l'arbre de décision et $\mathbf{w}$ le vecteur des valeurs attribuées à chacune de ses feuilles. Le terme Ω s'interprète alors comme une combinaison de régularisation Ridge et de pénalisation Lasso (Section 2.2.3). Si l'algorithme *Extreme Gradient Boosting* est flexible est performant, la complexification est très sensible, notamment avec le nombre de paramètres à calibrer (Section 2.3.4).

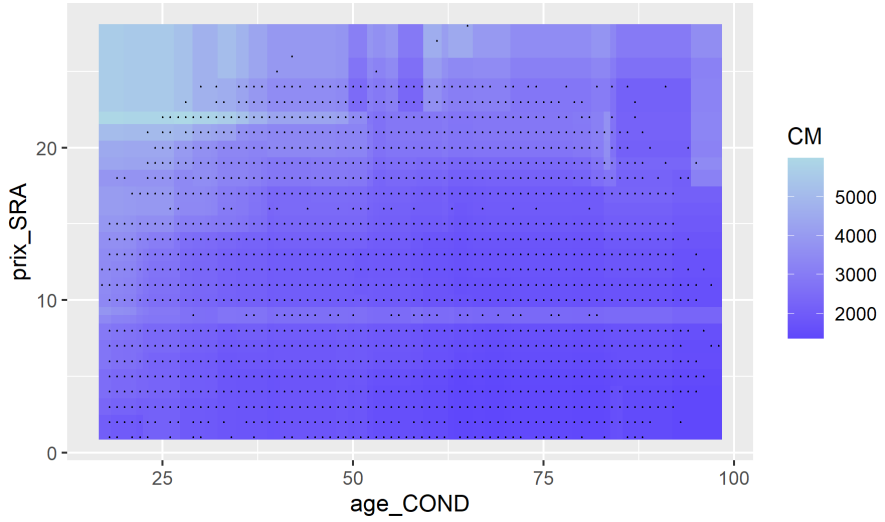


FIGURE 2.8 – Exemple de prédiction dans le plan `age_COND` et `prix_SRA` obtenu à partir de l'algorithme *Extreme Gradient Boosting* pour la modélisation du coût moyen

La Figure 2.8 montre un exemple de prédiction dans le plan âge conducteur (`age_COND`) et classe de prix du véhicule (`prix_SRA`) avec l'algorithme *Extreme Gradient Boosting* pour la modélisation du coût moyen. Nous observons ainsi une segmentation du risque très précise avec plusieurs zones de coût élevé, notamment pour les jeunes conducteurs ayant une voiture chère.

Dans la section 2.3.2, nous avons introduit l'agrégation des modèles qui est un concept central en *Machine Learning*. Nous avons en particulier détaillé les techniques de *Bagging* et de *Boosting* en se basant sur les arbres de décision comme briques de base. Nous avons vu le gain de performance que ces méthodes pouvaient engendrer et constaté la perte d'interprétabilité en découlant. Une autre famille de modèles très particulière est formée par les réseaux de neurones. Dans le cadre de nos travaux de développement d'une nouvelle mécanique d'apprentissage, nous avons privilégié la structure arborescente pour des raisons de temps et de facilité d'implémentation. L'approche par

réseaux de neurones a néanmoins été investiguée, c'est pour cela que ses fondements théoriques sont succinctement présentés dans la section suivante et que plusieurs de ses déclinaisons sont explorés dans le Chapitre 3, notamment dans la Section 3.3.2.

2.3.3 Réseaux de neurones

Un réseau de neurones est un système dont la conception est à l'origine inspirée des comportements du cerveau humain, et qui par la suite s'est rapproché des méthodes statistiques. C'est en 1943 que les premières notions formelles d'un réseau de neurones apparaissent, mais le succès de la méthode est plus récent, avec des résultats spectaculaires obtenus notamment en reconnaissance d'image et en apprentissage profond (Mishra et Gupta (2017)). Un réseau de neurones est défini comme l'association en un graphe plus ou moins complexe d'objets élémentaires, les neurones. Les réseaux de neurones se distinguent par :

- leur architecture : l'organisation du graphe, des connexions, la présence ou non des boucles de rétroaction ;
- leur complexité : le nombre de couches et le nombre de neurones par couche ;
- le type des neurones : leurs fonctions de transition ou d'activation ;
- l'objectif visé : l'apprentissage supervisé ou non, l'optimisation, les systèmes dynamiques...

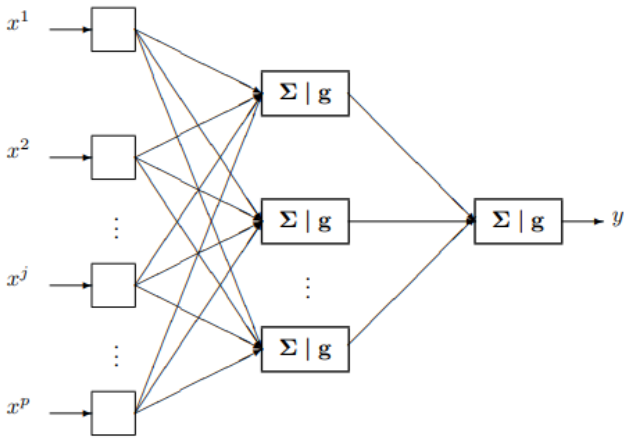


Figure 2.9 – Exemple de perceptron multicouche élémentaire avec une couche cachée à trois neurones. Source : Laurent-Bonneau (2020)

Un exemple de réseau d'architecture simple est le perceptron multicouche. Ce dernier est composé d'une couche d'entrée, d'une ou de plusieurs couches cachées et d'une couche de sortie qui fournit la réponse du système. Chaque couche cachée ajoute un niveau de non-linéarité dans le modèle. L'optimisation des poids, qui représentent la mémoire ou la connaissance répartie du réseau, s'effectue par l'algorithme de rétropropagation du gradient. Un théorème dit d'approximation universelle montre que cette structure élémentaire est suffisante pour prendre en compte les problèmes classiques de modélisation en apprentissage statistique (Hornik (1991)).

2.3.4 Optimisation des paramètres

Dans les algorithmes de *Machine Learning*, le calibrage des paramètres est souvent crucial pour garantir de bonnes performances des modèles de prédiction. Ces paramètres résument la complexité des méthodes et demandent donc une attention particulière.

Validation croisée

La validation croisée en V segments (ou *V-fold cross validation*) consiste à partager aléatoirement la base de données en V segments puis à faire jouer à chacun des segments le rôle d'échantillon de test tandis que les $V-1$ segments restants constituent l'échantillon d'apprentissage, et ce de manière itérative. La validation croisée est utilisée pour estimer le risque de prévision ou les paramètres qui minimisent ce risque. Plus précisément, soit $\tau : \{1, \dots, n\} \to \{1, \dots, V\}$ la fonction d'indexation qui attribue à chaque observation son segment. L'estimation du risque de prévision par validation croisée est donnée par

$$\hat{R}_{CV}(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, \hat{f}^{(-\tau(i))}(\mathbf{x_i})),$$

où $\hat{f}^{(-v)}$ désigne l'estimation de $\hat{f}$ sans prendre en compte la v-ème partie de l'échantillon. La Figure 2.10 détaille le fonctionnement pour $V = 5$.

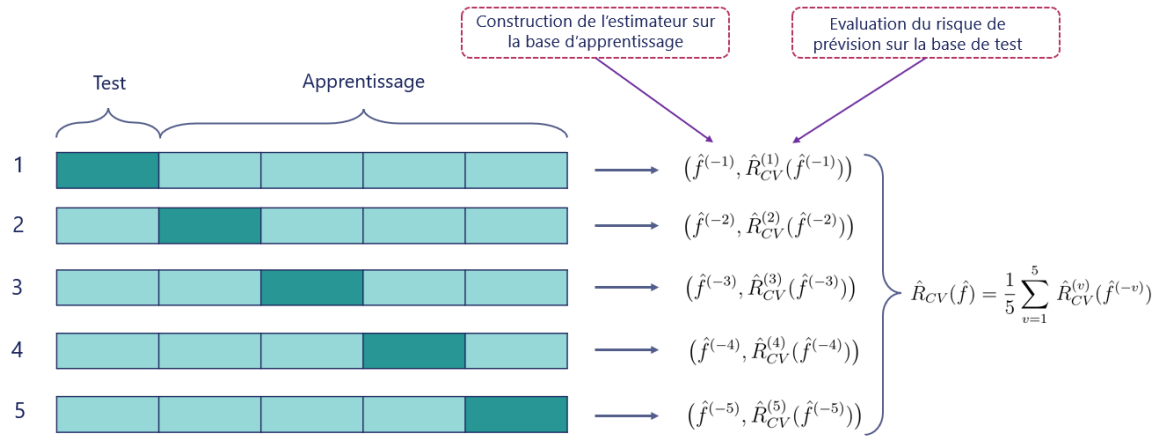


FIGURE 2.10 – Mécanisme de validation croisée pour $V = 5$

Minimiser le risque estimé par validation croisée permet ainsi de choisir le paramètre optimal $\hat{\theta}$ au sein d'une famille paramétrée

$$\hat{\theta} = \arg\min_{\theta} \hat{R}_{CV}(\theta).$$

Grid Search et Random Search

Les algorithmes de *Machine Learning* requièrent souvent une optimisation simultanée de plusieurs paramètres. La recherche sur une grille, ou *Grid Search* est une méthode d'optimisation permettant de croiser chacune des hypothèses et de construire un modèle pour chaque combinaison des paramètres. Par exemple, dans une forêt aléatoire, on peut vouloir calibrer le nombre B d'arbres et le nombre m de variables tirées aléatoirement selon les hypothèses suivantes :

- $B \in \{500, 750, 1000\}$;
- $m \in \{5, 10, 15\}$.

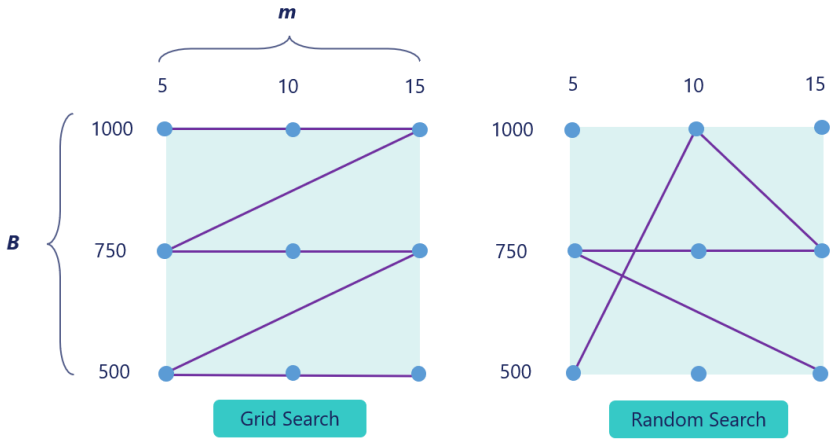


FIGURE 2.11 – Exemple d'optimisation des paramètres dans l'algorithme de la forêt aléatoire par *Grid Search* et *Random Search*

Nous aurons ainsi 9 modèles à construire et à évaluer par validation croisée. Cependant, tester toutes les combinaisons possibles peut s'avérer coûteux en temps, notamment lorsque le nombre de paramètres à optimiser est important, comme dans l'algorithme *Extreme Gradient Boosting*, ou lorsque la base de données est volumineuse. Le mécanisme de recherche aléatoire, ou *Random Search*, constitue alors une alternative à cette approche et permet de sélectionner les combinaisons des paramètres à tester. L'exhaustivité de la recherche peut être pilotée par l'utilisateur. La Figure 2.11 résume le fonctionnement de deux méthodes d'optimisation sur l'exemple de la forêt aléatoire.

2.3.5 Métriques de performance

Pour juger de la qualité d'un modèle de *Machine Learning* et comparer plusieurs modèles entre eux, on s'intéresse souvent aux métriques de performance. Dans le cadre de la régression, les métriques habituelles sont :

- le carré moyen des erreurs (*Mean Squared Error*, **MSE**) qui est la moyenne arithmétique des carrés des écarts entre les observations et les prévisions du modèle : $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$;
- l'erreur quadratique moyenne (*Root Mean Square Error*, **RMSE**) qui correspond à la racine carrée du **MSE** : $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$;
- l'erreur absolue moyenne (*Mean Absolute Error*, **MAE**) qui est la moyenne arithmétique des valeurs absolues des écarts : $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$.

Dans l'erreur quadratique moyenne, les écarts sont mis au carré avant d'être moyennisés, ce qui a pour effet de pénaliser les erreurs importantes. Dans l'erreur absolue moyenne, les écarts individuels ont le même poids. Quelque soit le critère retenu, il s'agira de retenir le modèle qui minimise les erreurs.

Pour les données assurantielles, ces métriques classiques ne sont pas toujours adaptées. Rappelons qu'un bon prédicteur est un prédicteur qui minimise la fonction de perte l. Dans le cadre d'un modèle linéaire de fréquence/sévérité, on s'intéresse alors aux métriques utilisant la déviance poissonienne ou la déviance Gamma, par exemple le pseudo R^2 (Équation 2.11). En effet, les métriques basées sur la déviance, et donc sur la vraisemblance (Équation 2.10), sont adaptées à la distribution sous-jacente de la donnée, notamment aux processus de comptage pour la modélisation de la fréquence ou aux distributions asymétriques à droite pour la modélisation du coût moyen. Dans ce mémoire, nous souhaitons comparer plusieurs modèles où la notion de vraisemblance n'existe pas naturellement. C'est pour cela que nous allons nous appuyer essentiellement sur les métriques de performance classiques, en prenant en compte leurs limites. Pour la modélisation de la fréquence, nous envisagerons également de transformer le problème de régression en un problème de classification multi-classe et nous nous intéresserons alors à l'*Accuracy* des prédictions définie par

$$Accuracy = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}}.$$

Dans la Section 2.3, nous avons détaillé le fonctionnement d'un arbre de décision, puis envisagé des structures algorithmiques plus complexes avec l'agrégation des modèles et les réseaux de neurones. Nous avons également présenté les procédures d'optimisation des paramètres dans un modèle de *Machine Learning* et défini quelques métriques pour évaluer leur performance. La section suivante aura pour objectif d'introduire les méthodes qui peuvent être envisagées pour palier au défaut principal des modèles de *Machine Learning*, leur interprétabilité.

2.4 Interprétabilité des modèles de Machine Learning

2.4.1 Introduction

Les modèles de *Machine Learning* prennent en compte efficacement les non-linéarités inhérentes à la donnée et les phénomènes d'interaction entre les variables. Cependant, avec des architectures complexes, la hausse de performance s'accompagne souvent d'une perte d'interprétabilité. Pour avoir confiance dans la modélisation proposée et justifier son utilité, la question d'interprétabilité est capitale. Elle se décline souvent en deux volets, avec d'une part l'évaluation du rôle des variables dans le modèle et d'autre part l'explication du comportement de ce dernier. Autrement dit, il s'agit d'apporter une réponse aux questions de qui et comment lors du développement du modèle pour sa validation interne. De plus, dans le secteur de la finance et de l'assurance, ce souci d'interprétabilité est couplé à un besoin d'explicabilité. Pour se conformer aux exigences réglementaires et répondre aux attentes du client, les décisions du modèle doivent être intelligibles, même pour un utilisateur non expert de la donnée.

À l'heure actuelle, l'applicabilité opérationnelle des modèles de *Machine Learning* reste encore limitée. Afin de permettre leur intégration dans le processus de décision, la communauté scientifique montre un intérêt croissant pour les questions d'interprétabilité. Cependant, la définition des critères d'un modèle interprétable ne rencontre pas un consensus général. L'interprétabilité est communément

définie comme « le degré à partir duquel un humain peut régulièrement prédire le résultat d'un modèle »(Molnar (2020)). Le degré d'interprétabilité d'un modèle s'évalue alors selon les critères suivants (Delcaillau et al. (2020)) :

- la confiance : la possibilité de s'appuyer sur le modèle en toute sérénité lors de la prise de décision ;
- la causalité : l'adéquation des résultats du modèle avec les phénomènes du monde réel ;
- la transférabilité : l'aptitude du modèle à s'adapter à des situations légèrement différentes, sa capacité de généralisation ;
- l'informativité : la capacité du modèle à fournir une information utile pour la prise de décision au-delà de la simple optimisation mathématique ;
- une prise de décision juste et éthique, l'absence du biais.

Au-delà du respect des directives du RGPD, la notion du biais requiert une attention particulière dans le secteur assurantiel. En effet, l'hyper-individualisation du tarif, favorisée par le développement de l'Intelligence Artificielle et des modèles de *Machine Learning*, remet en cause le principe même de l'assurance, à savoir la mutualisation des risques entre tous les assurés. Si la mutualisation recule au détriment de la segmentation, des biais importants doivent être évités dans la construction des algorithmes pour préserver un principe d'équité (Barry et Charpentier (2022)). Par-delà, l'interprétabilité des modèles demeure une préoccupation majeure.

Selon Molnar (2020), les différentes méthodes d'interprétabilité d'un modèle peuvent être classées selon trois axes :

- intrinsèque (*a priori*) ou *a posteriori* ;
- spécifique ou agnostique ;
- locale ou globale.

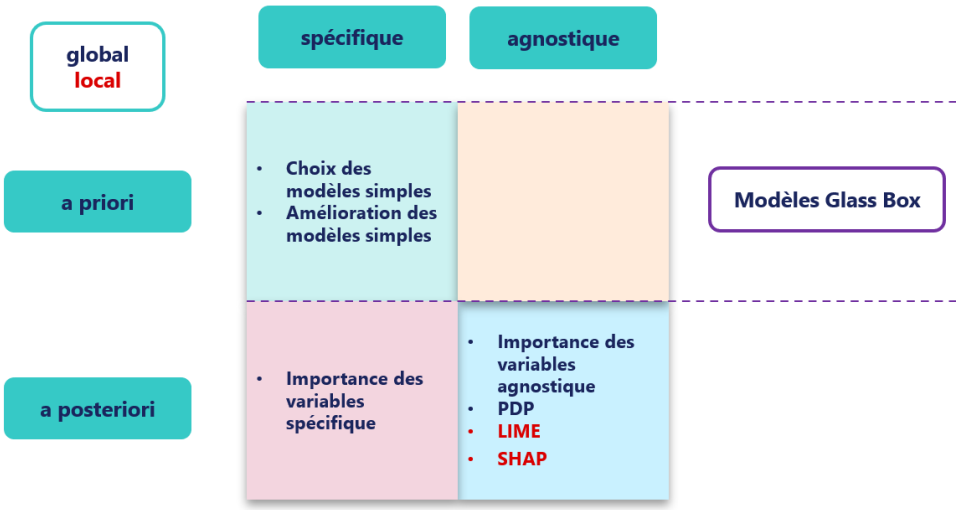


Figure 2.12 – Résumé des méthodes d'interprétation des modèles de *Machine Learning*

Dans un modèle interprétable *a priori*, on pense l'interprétabilité dès la construction du modèle. Les modèles considérés comme intrinsèquement interprétables sont alors souvent des modèles simples, comme les modèles linéaires généralisés (Section 2.2.2) ou les arbres de décision de faible profondeur (Section 2.3.1). Dans une approche *a posteriori*, il s'agit de gagner de la visibilité sur le fonctionnement

des modèles complexes, et ce une fois qu'ils ont été ajustés. En effet, les modèles obtenus par agrégation (Section 2.3.2) ou par un réseau de neurones (Section 2.3.3) s'apparentent à des boîtes noires, des *Black Box*, ce n'est que l'analyse postérieure qui peut les rendre intelligibles. Par la suite, une interprétation spécifique est entièrement liée au modèle ou à la classe de modèles choisie tandis qu'une interprétation agnostique est indépendante de la modélisation utilisée et donc applicable à tout algorithme. Finalement, une méthode d'interprétation locale se limite à l'explication d'une prédiction individuelle tandis qu'une méthode globale aborde le comportement du modèle dans son intégralité.

Les méthodes d'interprétabilité sont très nombreuses et puisent dans différentes branches mathématiques. Dans cette section, nous nous focalisons sur l'importance des variables, les graphes de dépendance partielle et les fondements théoriques de la méthode LIME. À titre indicatif, une revue rapide de la méthode SHAP est également proposée. La Figure 2.12 fournit un résumé des méthodes d'interprétabilité abordées. Elle permet également d'anticiper la position de la modélisation proposée dans le Chapitre 3, à savoir les modèles *Glass Box*, par rapport aux autres méthodes : une interprétabilité intrinsèque située à la frontière du spécifique et de l'agnostique.

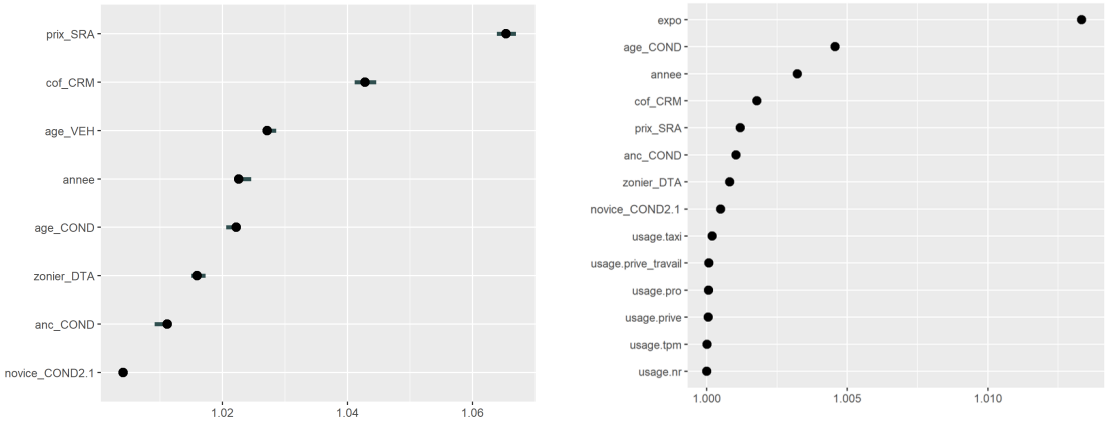
2.4.2 Comprendre le rôle des variables explicatives

Importance des variables

L'importance d'une variable explicative dans le modèle est définie par la hausse du risque de prévision suite à la permutation des valeurs de cette variable (Molnar (2020)). En effet, la permutation des valeurs d'une variable modifie la relation entre le prédicteur et la cible. Si une permutation des valeurs d'une variable affecte significativement les performances du modèle, c'est que cette variable est importante dans la prédiction. Il s'agit d'une mesure agnostique de l'importance des variables. Elle peut donc être utilisée pour comparer les algorithmes ayant une structure arborescente ou neuronale. Formellement, pour une variable à prédire Y, un ensemble de prédicteurs $\mathbf{X} = (X_1, \ldots, X_p)$, un modèle f et une fonction de perte l, la procédure d'estimation de l'importance des variables est la suivante :

1. Estimer le risque de prévision du modèle $\hat{R}(\hat{f}) = l(y, \hat{f}(\mathbf{x}))$;
2. Pour chaque variable explicative X_j, $j = 1, \ldots, p$:
   - Générer une matrice $\mathbf{X}_{perm,j}$ en permutant les valeurs de la j-ème variable dans X ;
   - Estimer le risque de prévision $\hat{R}_{perm,j}(\hat{f}) = l(y, \hat{f}(\mathbf{x}_{perm,j}))$;
   - Calculer l'importance de la j-ème variable par $imp_j = \hat{R}_{perm,j} - \hat{R}(\hat{f})$ ou $imp_j = \frac{\hat{R}_{perm,j}}{\hat{R}(\hat{f})}$;
3. Ordonner les variables explicatives par ordre décroissant de l'importance des variables imp_j, $j = 1, \ldots, p$.

Dans l'algorithme *Random Forest*, l'importance des variables agnostique est donnée par la *Mean Decrease Accuracy* (Section 2.3.2). L'autre mesure d'importance, la *Mean Decrease Gini*, est basée sur la décroissance de l'hétérogénéité qui est une notion propre à un arbre de décision. Il s'agit donc d'une mesure spécifique de l'importance des variables dont un exemple a été étudié dans le Chapitre 1 (Figures 1.12 et 1.14). La Figure 2.13, obtenue avec le package `iml` (Molnar et al. (2018)), montre l'importance des variables agnostique obtenue à partir de l'algorithme *XGBoost*. Les trois variables jugées les plus importantes pour la modélisation du coût moyen (Figure 2.13a) sont alors la classe de prix du véhicule (`prix_SRA`), le coefficient de réduction majoration (`cof_CRM`) et l'âge du véhicule (`age_VEH`). Pour la modélisation de la fréquence, les trois variables les plus importantes, hormis l'exposition, sont l'âge du conducteur (`age_COND`), l'année civile (`annee`) et `cof_CRM` (Figure 2.13b).



(a) coût moyen (b) fréquence

FIGURE 2.13 – Importance des variables agnostique obtenue à partir de l'algorithme *XGBoost*

Analyse PDP

L'analyse PDP (*Partial Dependance Plot*) est une méthode d'interprétation globale qui a pour objectif de montrer l'effet marginal moyen d'une ou plusieurs variables explicatives sur la prédiction du modèle (J. H. FRIEDMAN (2001)). Notons $\mathbf{X}_S$ l'ensemble des variables explicatives pour lesquelles on souhaite connaître l'effet sur la prédiction et $\mathbf{X}_C$ les variables restantes. On a ainsi $\mathbf{X} = \mathbf{X}_S \cup \mathbf{X}_C$. On émet l'hypothèse forte de la non-corrélation entre les variables explicatives de l'ensemble S et celles de l'ensemble C. Pour un modèle f, la fonction de dépendance partielle est définie par

$$\hat{f}_S(\mathbf{x}_S) = \mathbb{E}_{\mathbf{X}_C}[\hat{f}(\mathbf{x}_S, \mathbf{X}_C)] = \int \hat{f}(\mathbf{x}_S, \mathbf{x}_C)\, d\mathbb{P}_{\mathbf{X_C}}(\mathbf{x}_C).$$

Prenons l'exemple simple de $S = \{1\}$ et $C = \{2, \ldots, p\}$. La base des prédicteurs $(X_{i,j})_{1\leq i\leq n, 1\leq j\leq p}$ se réécrit ainsi comme $(X_{i,1}, X_{i,C})_{1\leq i\leq n}$ et la procédure d'estimation de la fonction de dépendance partielle $\hat{f}_S$ est la suivante :

1. Pour $i = 1, \ldots, n$:
   - Copier la base des prédicteurs en remplaçant la valeur de la variable x_1 par la valeur constante $x_{i,1}$: $(x_{i,1}, x_{k,C})_{1\leq k\leq n}$;
   - Calculer la prédiction $\hat{f}$ sur cette base de prédicteurs : $\hat{f}(x_{i,1}, x_{k,C})$ pour $k = 1, \ldots, n$;
   - Estimer $\hat{f}_1(x_{i,1})$ par $\frac{1}{n}\sum_{k=1}^{n} \hat{f}(x_{i,1}, x_{k,C})$;
2. Tracer le graphique des points $(x_{i,1}, \hat{f}_1(x_{i,1}))$ pour $i = 1, \ldots, n$, le graphe de dépendance partielle.

Les Figures 2.14 et 2.15 montrent le graphe de dépendance partielle pour les variables âge conducteur (`age_COND`) et classe de prix du véhicule (`prix_SRA`) dans la modélisation du coût moyen. Elles sont obtenues à partir de l'algorithme de *Random Forest* avec le package `InterpretML`[*]. Les courbes en couleur en train fin correspondent aux graphes de dépendance partielle obtenus sur des échantillons aléatoires de la base de données, la courbe bleue en trait épais donne l'allure moyenne. Pour la variable âge conducteur, l'effet marginal moyen (fonction `Average Response` sur la Figure 2.14) a une tendance décroissante jusqu'à 35 ans environ. Un pic de croissance s'observe sur l'intervalle de 40 à 50 ans, puis

∗. Le package `InterpretML` est présenté en détail dans le Chapitre 3

de nouveau une tendance décroissante jusqu'à 75 ans. Sur la dernière tranche d'âge, l'effet marginal moyen augmente modérément. Cette allure générale est cohérente avec les évolutions constatées en assurance automobile. Le coût moyen de sinistres est plus élevé pour les jeunes conducteurs et pour les conducteurs âgés. L'augmentation observée sur l'intervalle de 40 à 50 ans peut être expliquée par la présence d'un conducteur novice, typiquement due au fait que les parents prêtent leur voiture aux enfants.

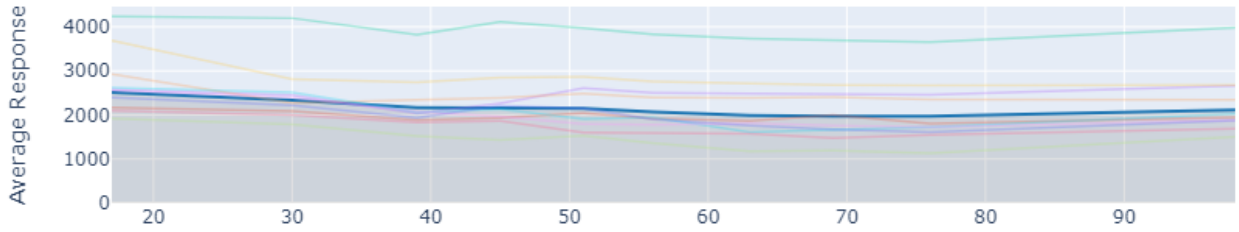


FIGURE 2.14 – Graphe de dépendance partielle pour la variable `age_COND` obtenu à partir de l'algorithme *Random Forest* pour la modélisation du coût moyen

Pour la variable donnant la classe de prix de la voiture, la Figure 2.15 montre un effet marginal moyen globalement croissant, avec une accélération de la croissance à partir de la classe 15. Ce résultat a également une cohérence actuarielle. Plus le prix de la voiture est élevé, plus le coût moyen de sinistre sera important pour la garantie DTA.

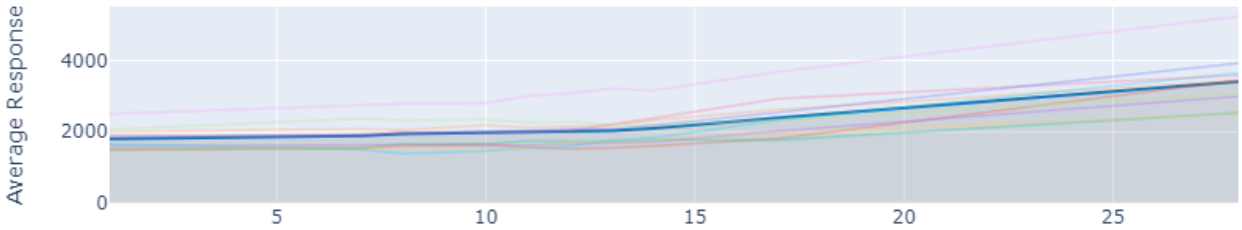


FIGURE 2.15 – Graphe de dépendance partielle pour la variable `prix_SRA` obtenu à partir de l'algorithme *Random Forest* pour la modélisation du coût moyen

2.4.3 Comprendre les prédictions individuelles

LIME

La méthode LIME (*Locally Interpretable Model-agnostic Explanations*, RIBEIRO et al. (2016)) est l'une des premières approches locales apparues dans le domaine du *Machine Learning* interprétable. Elle consiste à expliquer la prédiction pour un individu par un modèle simple, et ce uniquement avec les individus de son voisinage. Ainsi, si un modèle est globalement complexe, il peut être localement interprétable. Soit M_1 le modèle que l'on cherche à expliquer, par exemple un modèle *XGBoost* (Section 2.3.2), de règle de prévision f_1 et M_2 le modèle d'interprétation, par exemple un modèle linéaire (Section 2.2.1). La règle de prévision f_2 associée au modèle M_2 est obtenue en suivant la procédure suivante :

1. Créer un nouvel échantillon $\tilde{\mathbf{X}}$ en perturbant les données initiales $\mathbf{X}$;
2. Appliquer le modèle M_1 sur cet échantillon : $\hat{y}_1 = \hat{f}_1(\tilde{\mathbf{x}})$;
3. Pondérer chaque observation de l'échantillon $\tilde{\mathbf{X}}$ en fonction de sa proximité avec les données initiales ;
4. Construire le modèle d'interprétation M_2 sur les données pondérées en résolvant le problème d'optimisation
$$\hat{f}_2 = \arg\min_{f_2} l_{\pi_x}(f_1, f_2) + \Omega(f_2),$$
où l est la fonction de perte, π_x une mesure de proximité définissant la taille du voisinage autour de $\mathbf{x}$ et Ω une fonction traduisant la complexité du modèle, par exemple le nombre de paramètres dans le cas d'un modèle linéaire.

La Figure 2.16 résume le principe de fonctionnement de la méthode LIME dans un problème de classification binaire avec deux variables explicatives. La zone bleue (respectivement rouge) répertorie les points associés à la classe 1 (respectivement 0) par le modèle M_1. Par la suite, les croix et les points représentent les données simulées pour l'apprentissage du modèle d'interprétation M_2. La taille du motif est proportionnelle au poids du point considéré en fonction de la distance de ce dernier à l'observation d'intérêt, représentée par la grande croix rouge. Finalement, la droite grise en pointillés est la règle de décision obtenue par le modèle M_2, en l'occurrence un modèle linéaire.

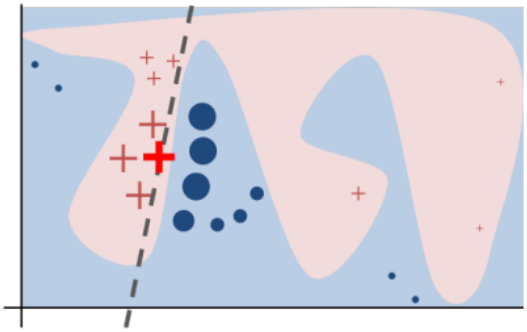


FIGURE 2.16 – Illustration du principe de fonctionnement de la méthode LIME. Source : DELCAILLAU et al. (2020).

Si la méthode LIME permet d'expliquer localement la prédiction d'un modèle complexe, elle est très sensible à l'échantillonnage des données et le choix du noyau utilisé pour mesurer la proximité des observations.

SHAP

La méthode SHAP (*Shapley Additive Explanations*, LUNDBERG et LEE (2017)) constitue une autre approche d'interprétation locale. Elle emprunte à la théorie des jeux en se basant sur la valeur de Shapley pour définir une mesure du poids des variables dans les prédictions du modèle. Dans un jeu coopératif, la valeur de Shapley donne une répartition équitable des gains parmi les joueurs d'une coalition (WINTER (2002)). Si l'on transpose ces notions à l'interprétabilité des modèles, on établit les associations suivantes :

- le jeu représente la tâche de prédiction pour une variable $X_j \in \mathcal{X}$;

- le gain correspond à la prédiction actuelle de cette variable moins la prédiction moyenne de toutes les autres variables du jeu de données ;
- les joueurs représentent les valeurs des variables $X_j,\ j = 1, \ldots, p$ qui collaborent pour recevoir le gain.

Finalement, on peut définir la valeur de Shapley comme la contribution marginale moyenne d'une variable explicative sur toutes les coalitions possibles. Les fondements mathématiques de cette approche ainsi que le détail des calculs peuvent être consultés dans Winter (2002) et Lundberg et Lee (2017). Par ailleurs, une application de la méthode SHAP à la tarification en assurance non-vie peut trouvée dans le mémoire d'actuariat Bucci (2021).

Dans ce chapitre, nous avons envisagé d'une part le cadre paramétrique, avec les modèles linéaires généralisés, et d'autre part le cadre non-paramétrique avec les modèles de *Machine Learning*. Nous avons construit l'exposé théorique autour des notions d'estimation, d'apprentissage, de pénalisation, d'interprétabilité et d'explicabilité. Nous avons ainsi défini les concepts mathématiques principaux qui sont revisités et déclinés dans le Chapitre 3. Une importance particulière a été accordée à l'interprétabilité des modèles et a permis de définir la classe des modèles intrinsèquement interprétables par opposition aux méthodes interprétables *a posteriori*.

Chapitre 3

Présentation du modèle additif généralisé

3.1 Introduction

Le Chapitre 2 a permis d'exposer les enjeux spécifiques liés à la modélisation statistique dans la tarification en assurance non-vie. Si la structure paramétrique des modèles linéaires généralisés assure l'interprétabilité et la facilité d'utilisation, et la structure non-paramétrique des modèles de *Machine Learning* favorise la performance et la finesse de modélisation, il apparaît pertinent de combiner les avantages de ces deux méthodes dans une structure semi-paramétrique. Il s'agit en effet de se positionner entre les deux paradigmes de modélisation et proposer un modèle mixte qui puisse être à la fois performant et interprétable (Figure 3.1). Un tel cadre semi-paramétrique est donné notamment par les modèles additifs généralisés (Hastie et Tibshirani (1987)).

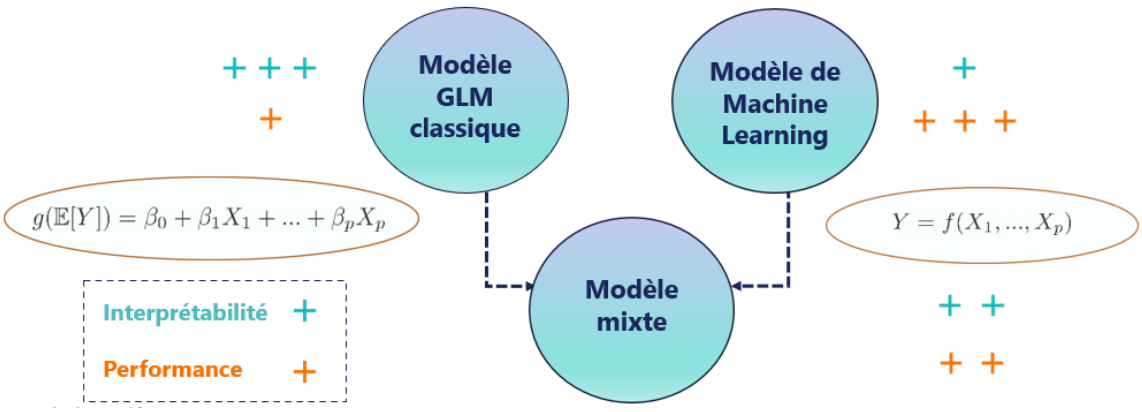


Figure 3.1 – Positionnement du problème : construire un modèle à la fois performant et interprétable.

Les modèles additifs généralisés sont une extension des modèles linéaires généralisés (Équation 2.6) où l'espérance de la variable réponse Y est expliquée en fonction d'une combinaison linéaire des prédicteurs, mais la contribution individuelle de chaque prédicteur se trouve complexifiée. En effet, la dépendance entre la variable à prédire et les variables explicatives n'est plus nécessairement modélisée linéairement. Au lieu des coefficients constants $\beta_j,\ j = 1, \ldots, p$, les variables explicatives participent à la prédiction au travers de fonctions composantes non-paramétriques $f_j,\ j = 1, \ldots, p$. Le modèle s'écrit alors

$$g(\mathbb{E}[Y_i]) = \underbrace{\beta_0}_{\text{effet moyen}} + \underbrace{f_1(X_{i,1})}_{\text{effet } X_1} + \underbrace{f_2(X_{i,2})}_{\text{effet } X_2} + \ldots + \underbrace{f_p(X_{i,p})}_{\text{effet } X_p}, \; i = 1, 2, \ldots, n \qquad (3.1)$$

L'intérêt essentiel d'un tel type de modélisation réside dans l'ajout parcimonieux de la non-linéarité dans la structure additive d'un modèle linéaire généralisé. La contribution de chaque prédicteur peut être arbitrairement complexe, mais la modélisation individuelle de cette contribution rend le modèle final intrinsèquement intérprétable. Par la suite, la visualisation de cette fonction composante donne une description exacte de comment est réalisée la prédiction au niveau de la variable. Un bon modèle réalisant l'équilibre entre le biais et la variance, la possibilité de contrôler notre *a priori* sur la régularité de la relation prédictive au travers de la régularité des fonctions composantes apparaît d'autant plus intéressante.

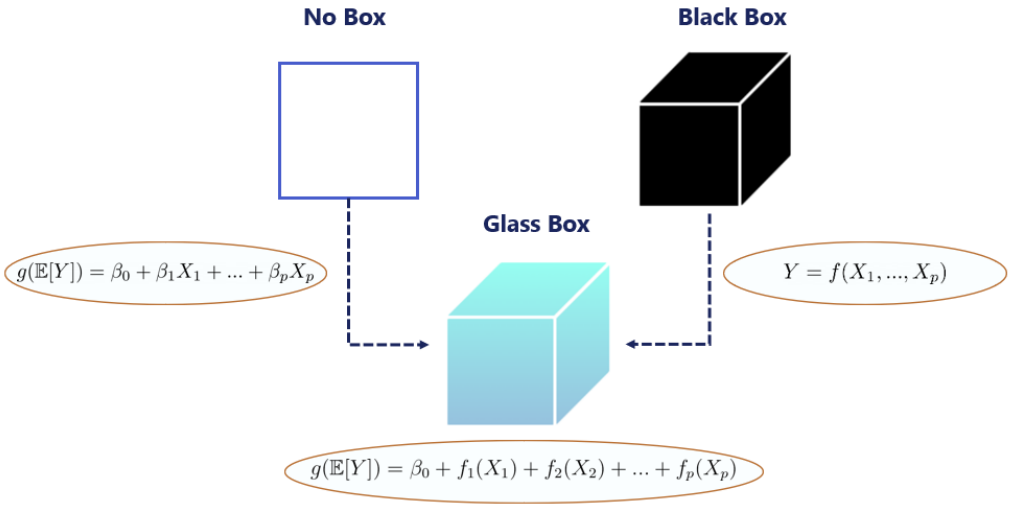


FIGURE 3.2 – Reformulation du problème : construire un modèle de *Machine Learning* intrinsèquement interprétable.

Traditionnellement, les fonctions composantes dans un modèle additif généralisé sont des fonctions de lissage, notamment des fonctions splines. Si elles adressent efficacement le problème de contrôle de régularité, elles peinent souvent à capturer la richesse des données et sous-apprennent le phénomène sous-jacent (LARSEN (2015)). Adopter ce type de modélisation mais s'approprier davantage la notion d'apprentissage au travers d'une compréhension plus fine de la donnée semble aller dans le sens de l'évolution amorcée par les modèles de *Machine Learning*. En effet, la mise en place d'une démarche d'innovation pilotée par les données est au coeur des préoccupations actuelles (SANTENAR et al. (2022)). Dans le cadre spécifique de la tarification, une vision plus inclusive et collaborative entre les modèles linéaires généralisés et les modèles de *Machine Learning* est de plus en plus demandée. A ce titre, et en se basant sur un article de l'équipe de recherche de Microsoft (NORI et al. (2019)), nous introduisons des fonctions composantes incorporant des modèles de *Machine Learning* dans le modèle additif généralisé. Dit autrement, nous considérons d'une part le modèle linéaire généralisé, le modèle de référence que l'on appellera également *No Box* car non fondé sur une notion d'apprentissage. D'autre part, nous travaillons avec un modèle de *Machine Learning*, un modèle boîte noire ou *Black Box* car très peu interprétable. Entre les deux, nous positionnons un modèle boîte transparente ou *Glass Box*[*], un modèle qui intègre

∗. Le terme *Glass Box* est emprunté à l'équipe de recherche de Microsoft. L'objectif est de rendre compte de l'interprétabilité intrinsèque de la méthode de *Machine Learning* développée par opposition aux modèles *Black Box*.

la notion d'apprentissage tout en restant interprétable (Figure 3.2). Reprenant le schéma de la Figure 2.12, il s'agit de proposer une méthode qui intègre la notion d'interprétabilité *a priori* de manière globale, une méthode que l'on peut positionner à la frontière de l'interprétabilité spécifique et agnostique.

Tout comme dans le modèle linéaire généralisé, l'ajout parcimonieux des interactions permet un enrichissement du modèle (J. A. Nelder et Wedderburn (1972)). Cette approche est naturellement extensible au cadre de la modélisation additive généralisée combinant des fonctions de *Machine Learning*. Dans ce mémoire, nous nous focalisons sur des composantes univariées des modèles de *Machine Learning*. L'ajout des interactions du second ordre sera introduit succinctement dans la Section 3.3.1 et envisagé plus généralement en perspectives.

L'objectif de ce chapitre est d'asseoir le cadre théorique de la modélisation additive généralisée combinant des composantes univariées des modèles de *Machine Learning*. Pour cela, les fondements du modèle additif généralisé utilisant des fonctions splines sont rappelés. L'introduction des modèles de *Machine Learning* comme fonctions composantes vient concurrencer les fonctions splines et s'appuit abondamment sur les notions du Chapitre 2. Nous tâchons en particulier de distinguer le lissage de l'information opérée par les fonctions splines (Section 3.2) et l'apprentissage de l'information induit par les modèles de *Machine Learning* (Section 3.3). Dans tout ce qui suit, l'exemple de la variable âge conducteur (`age_COND`) pour la visualisation des fonctions composantes est proposé.

3.2 GAM Lissage

Dans les modèles additifs généralisés traditionnels, la composante déterministe $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$ des modèles linéaires généralisés (Équation 2.6) est remplacée par une composante de lissage non-paramétrique $\eta = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$. Compte tenu de la structure sous-jacente de la donnée, les fonctions composantes f_j, $j = 1, \ldots, p$ peuvent être non-linéaires. La variable réponse Y appartient toujours à la famille exponentielle mais sa relation de dépendance avec les variables explicatives est ainsi rendue plus flexible (Figure 3.3). Cette flexibilité et proximité de la donnée vient en contrepartie de deux nouveaux problèmes théoriques : la représentation des fonctions de lissage et leur degré de régularité.

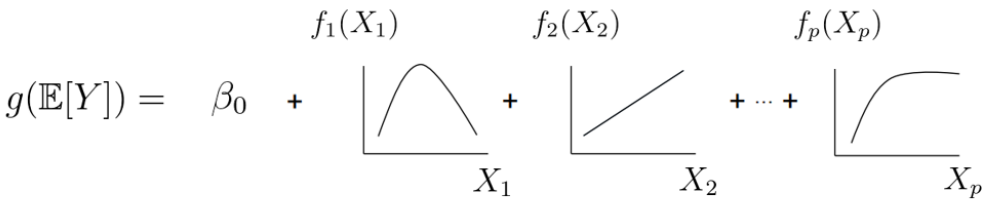


Figure 3.3 – Structure d'un modèle additif généralisé traditionnel

L'objectif de cette section est d'introduire les fonctions de lissage comme fonctions composantes dans le modèle additif généralisé. Plus précisément, il s'agit de partir d'un exemple avec une base simple et d'établir les liens qui existent avec le modèle linéaire généralisé. La procédure d'estimation des fonctions

composantes est ensuite détaillée pour permettre de pointer les différences conceptuelles avec l'approche par modèles de *Machine Learning* présentée plus loin dans la Section 3.3.1.

3.2.1 Base polynomiale

Pour introduire les fonctions de lissage, il convient de repartir du cadre du modèle linéaire. Pour simplifier, nous considérons un modèle avec une seule variable explicative

$$Y_i = f(X_i) + \epsilon_i,\ i = 1, \ldots, n, \tag{3.2}$$

où f désigne la fonction de lissage et ϵ_i les termes d'erreur indépendants identiquement distribués de loi normale $\mathcal{N}(0, \sigma^2)$. Supposons de plus que $X_i \in [0, 1]$. Pour estimer f avec les outils théoriques afférents au modèle linéaire (Section 2.2), l'équation 3.2 doit pouvoir se mettre sous la forme de l'équation 2.2. Ceci passe par le choix d'une base où la fonction f est représentée sous la forme

$$f(x) = \sum_{q=1}^{Q} \beta_q b_q(x), \tag{3.3}$$

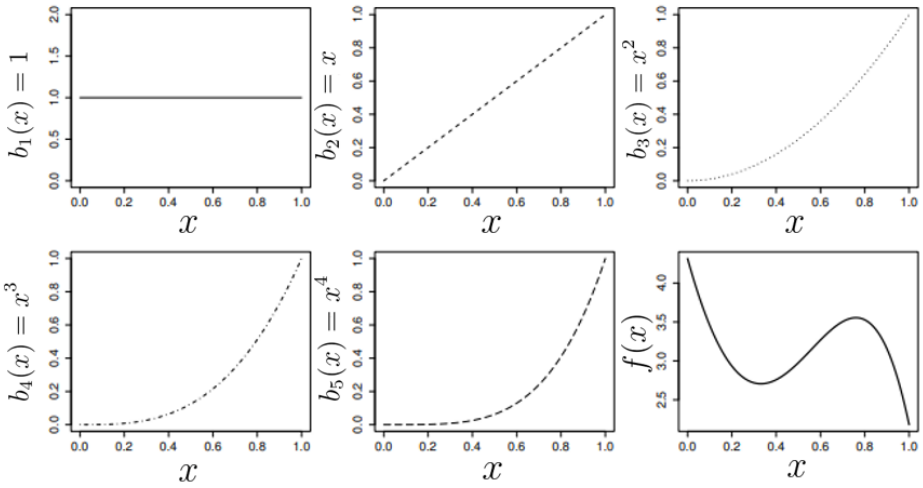


FIGURE 3.4 – Illustration du principe de décomposition d'une fonction dans une base polynomiale. Source : WOOD (2006), modifié.

où $\beta_q,\ q = 1, \ldots, Q$ sont les paramètres à estimer et $b_q,\ q = 1, \ldots, Q$ désignent les fonctions de base. Un premier cadre simple est celui d'une base polynomiale. Supposons que f est un polynôme de degré 4. Une base de cette espace est donnée par

$$\begin{cases} b_1(x) &= 1 \\ b_2(x) &= x \\ b_3(x) &= x^2 \\ b_4(x) &= x^3 \\ b_5(x) &= x^4. \end{cases}$$

La Figure 3.4 montre les fonctions de base $b_q,\ q = 1, \ldots, 5$, sur les cinq premiers panels, puis la fonction finale f sur le dernier. Elle est obtenue par une addition des fonctions de base après multiplication par un paramètre $\beta_q,\ q = 1, \ldots, 5$. L'équation 3.3 se réécrit alors comme

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + \beta_5 x^4.$$

Le modèle de l'équation 3.2 devient ainsi

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + \beta_5 X_i^4 + \epsilon_i,$$

où les paramètres $\beta_q,\ q = 1, \ldots, 5$ sont estimés par moindres carrés (Équation 2.3). Cependant, la base polynomiale est peu utilisée en pratique. En effet, les polynômes sont des fonctions trop rigides, avec une propension à osciller et une difficulté à représenter des parties plates (Wood (2006)). Les fonctions splines possèdent de meilleures propriétés théoriques.

3.2.2 Base de l'espace des splines

Une spline cubique naturelle est une fonction constituée d'une succession de polynômes de degré 3 qui se raccordent correctement, c'est-à-dire tels que la fonction soit deux fois continûment dérivable sur $\mathbb{R}$. Les points de raccordement sont appelés les noeuds. On peut montrer que la spline cubique naturelle est la fonction qui oscille le moins possible parmi toutes les fonctions qui interpolent les données. Cependant, dans une optique prédictive, il est plus judicieux de lisser les données plutôt que de les interpoler, suivre l'allure générale sans pour autant passer exactement par tous les points. On s'intéresse alors aux splines de régression et on choisit l'emplacement des noeuds, typiquement des noeuds équidistribués sur l'ensemble des valeurs observées de x ou une distribution par quantiles. Soit $\left\{x_q^*,\ q = 1, \ldots, Q-2\right\}$ l'ensemble des locations des noeuds. Il est possible de montrer (Wahba (1990)) que la base de l'espace des splines cubiques naturelles s'écrit alors

$$\begin{cases} b_1(x) & = 1 \\ b_2(x) & = x \\ b_{q+2}(x) & = R(x, x_q^*),\ q = 1, \ldots, Q-2, \end{cases}$$

où

$$R(x,z) = \frac{\left((z-\frac{1}{2})^2 - \frac{1}{12}\right)\left((x-\frac{1}{2})^2 - \frac{1}{12}\right)}{4} - \frac{\left((|x-z|-\frac{1}{2})^4 - \frac{1}{2}(|x-z|-\frac{1}{2})^2 + \frac{7}{240}\right)}{24}.$$

La Figure 3.5 illustre le même principe de fonctionnement que la Figure 3.4, mais avec des splines cubiques naturelles d'ordre $Q = 5$ où les noeuds ont été choisis tels que $x_1^* = \frac{1}{6}$, $x_2^* = \frac{3}{6}$ et $x_3^* = \frac{5}{6}$.

Le modèle 3.2 s'écrit alors

$$Y = X\beta + \epsilon,$$

où le i-ème rang de la matrice modèle est $\mathbf{X}_i = [1, X_i, R(X_i, X_1^*), R(X_i, X_2^*), \ldots, R(X_i, X_{Q-2}^*)]$. Les paramètres $\beta_q,\ q = 1, \ldots, Q$ s'estiment alors par moindres carrés (Équation 2.3). Une très bonne base de l'espace des splines est formée par les fonctions B-Splines. Il s'agit des splines cubiques naturelles

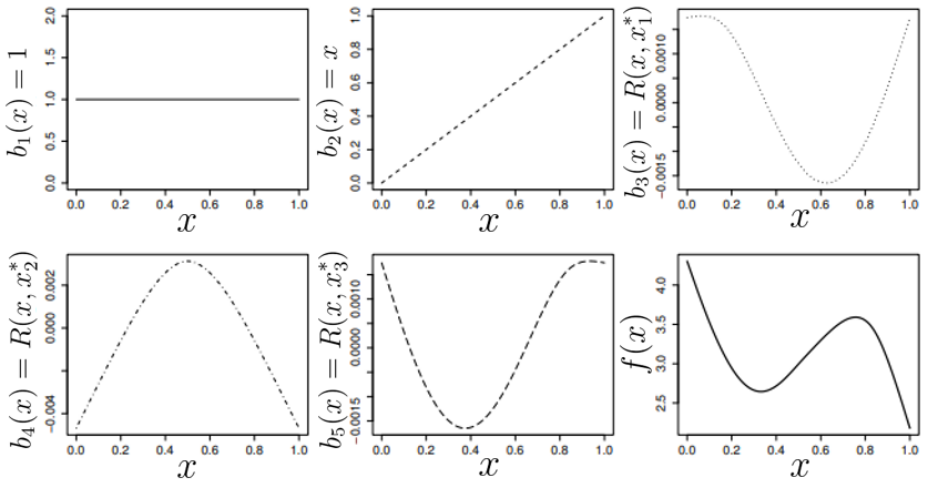


FIGURE 3.5 – Illustration du principe de décomposition d'une fonction dans une base de l'espace des splines naturelles. Source : WOOD (2006), modifié.

à support borné, c'est-à-dire nulles en dehors d'un intervalle borné. Elles possèdent des propriétés théoriques intéressantes, notamment la propriété de la partition de l'unité aux noeuds (WOOD (2006)).

3.2.3 Pénalisation des splines

L'unique contrôle de régularité des splines cubiques passe par le nombre de noeuds : moins il y a de noeuds, plus la fonction est régulière. L'idée est alors d'appliquer une pénalité supplémentaire de type $\int_0^1 (f''(x))^2 dx$ lors de l'estimation des coefficients β_q, $q = 1, \ldots, Q$. La fonction f dépendant linéairement des paramètres β, la pénalité peut se réécrire comme

$$\int_0^1 \left(f''(x)\right)^2 dx = \beta' S \beta,$$

où S est une matrice de coefficients connus. L'objectif de minimisation devient alors

$$||Y - X\beta||^2 + \lambda \beta' S \beta. \tag{3.4}$$

Tout comme dans la pénalisation Ridge ou Lasso (Section 2.2.3), c'est le paramètre λ qui détermine le degré d'ajustement aux données, et donc la régularité de la solution finale. Il est calibré par validation croisée (Section 2.3.4). Les splines cubiques ainsi pénalisées sont communément désignées sous le nom de P-Splines.

3.2.4 Estimation des paramètres dans un modèle GAM

L'écriture du modèle additif généralisé nécessite l'estimation simultanée de plusieurs fonctions composantes. Reprenant les notations de la Section 3.2.1, on pose

$$f_j(x_j) = \sum_{q=1}^{Q_j} \beta_{jq} b_{jq}(x_j),\ j = 1, \dots, p.$$

L'idée consiste alors à exhiber une matrice modèle $\tilde{X}_j \in \mathcal{M}_{n,Q_j}(\mathbb{R})$ pour chaque fonction composante f_j. Si $\mathbf{f}_j$ est le vecteur tel que $f_{j_q} = f_j(x_{jq})$ et $\tilde{\beta}_j = [\beta_{j1}, \beta_{j2}, \dots, \beta_{jQ_j}]'$ alors $\mathbf{f}_j = \tilde{X}_j \tilde{\beta}_j$, où $(\tilde{X}_j)_{qt} = b_{jt}(x_{jq})$. L'équation 3.1 se réécrit alors comme

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

où $\mathbf{X} = [\mathbf{1}_n, \tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p]$ est la matrice modèle complète avec n lignes et $1 + \sum_{j=1}^{p} Q_j$ colonnes et $\boldsymbol{\beta} = [\beta_0, \tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_p]'$. Le modèle additif généralisé s'interprète ainsi comme un modèle linéaire généralisé hyperparamétré. Par conséquent, log-vraisemblance $L(\underline{Y}, \boldsymbol{\beta})$ du modèle peut être calculée et les paramètres $\beta_0, \tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_p$ peuvent être estimés par la méthode du maximum de vraisemblance (Section 2.2.2). Reprenant l'équation 3.4, une pénalité $\tilde{\beta}_j' \tilde{S}_j \tilde{\beta}_j$ est mise en place pour chaque fonction composante f_j. On s'intéresse alors à la vraisemblance pénalisée

$$L_p(\underline{Y}, \boldsymbol{\beta}) = L(\underline{Y}, \boldsymbol{\beta}) - \frac{1}{2} \sum_{j=1}^{p} \lambda_j \tilde{\beta}_j' \tilde{S}_j \tilde{\beta}_j,$$

où les paramètres λ_j contrôlent l'équilibre entre le biais et la variance pour chaque fonction composante f_j. La résolution s'effectue par l'algorithme PIRLS (*Penalized Iteratively Reweighted Least Squares*, Marx et Eilers (1998)), qui est une extension de l'algorithme IRLS utilisé pour l'estimation des paramètres dans un modèle linéaire généralisé (Section 2.2.2). Cette méthode est implémentée dans le package `mgcv` (Wood (2011)) de manière à sélectionner automatiquement les paramètres optimaux. En résumé, il s'agit donc de suivre la procédure suivante :

1. Transformer le modèle additif généralisé en un modèle linéaire généralisé hyperparamétré ;
2. Calculer la log-vraisemblance ;
3. Pénaliser le modèle et calculer la log-vraisemblance pénalisée ;
4. Estimer les paramètres par la méthode du maximum de vraisemblance.

D'autres procédures d'estimation et de pénalisation des fonctions composantes dans un modèle additif généralisé existent mais nous ne focalisons ici sur la vision adaptative du modèle linéaire généralisé. Pour plus de détails, il est possible de consulter Wood (2006). Les Figures 3.6 et 3.7 donnent un exemple de fonctions composantes obtenues avec des splines cubiques naturelles à l'aide du package `mgcv`. Le trait en rouge montre l'allure de la fonction composante tandis que les tirets bleus représentent l'intervalle de confiance. Utilisant une fonction de lien logarithmique pour la modélisation du coût et de la fréquence, l'axe des ordonnées est en échelle logarithmique. Il faut donc appliquer la fonction exponentielle pour quantifier la variation du coût ou de la fréquence. Pour la modélisation du coût moyen, la Figure 3.6 indique deux zones de coût élevé, une pour les jeunes conducteurs et l'autre pour les conducteurs plus âgés. Pour la modélisation de la fréquence, l'augmentation pour les conducteurs âgés est beaucoup plus marquée que pour les jeunes conducteurs (Figure 3.7).

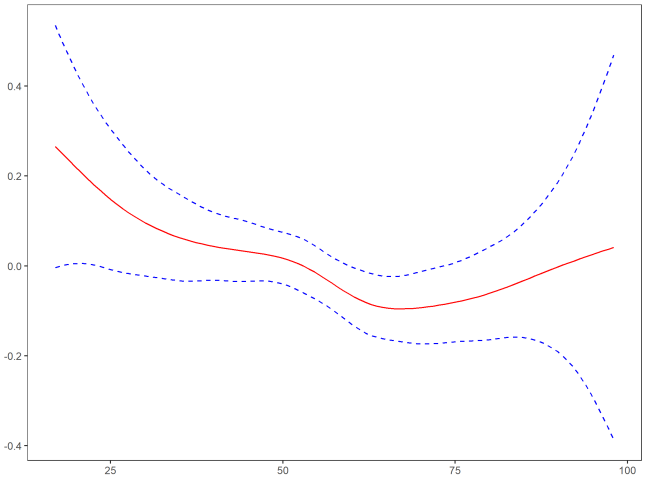


FIGURE 3.6 – Fonction composante associée à la variable `age_COND` dans le modèle GAM avec des splines cubiques naturelles pour la modélisation du coût moyen

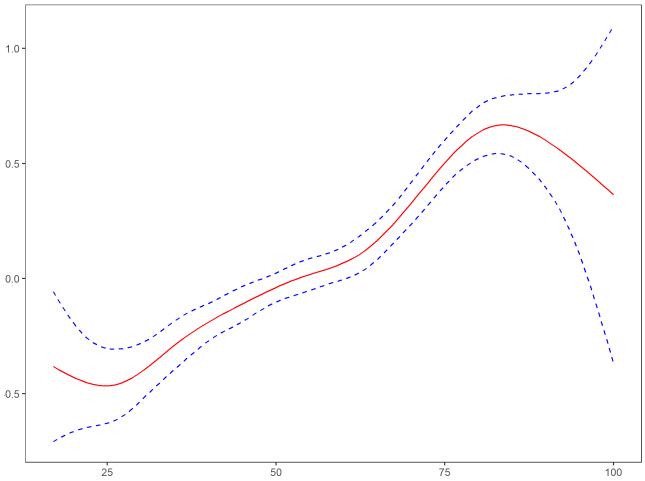


FIGURE 3.7 – Fonction composante associée à la variable `age_COND` dans le modèle GAM avec des splines cubiques naturelles pour la modélisation de la fréquence

Dans la Section 3.2, nous avons exploré l'ajout de la non-linéarité dans un modèle additif par des fonctions splines. Nous avons ainsi établi qu'un modèle additif généralisé combinant des fonctions splines pouvait se réécrire comme un modèle linéaire généralisé hyperparamétré. Nous avons également vu les procédures de lissage et de pénalisation. Dans la section suivante, nous nous intéressons à l'apprentissage de l'information plutôt qu'à son lissage. Nous introduisons ainsi la nouvelle mécanique d'apprentissage garantissant l'interprétabilité en incorporant des composantes univariées des modèles de *Machine Learning* dans un modèle additif généralisé.

3.3 GAM Machine Learning

La structure du modèle additif généralisé autorise chaque variable explicative à contribuer de manière spécifique à la prédiction finale. Conjuguer cette structure avec des fonctions polynomiales ou des fonctions splines permet de poursuivre la discussion sur les notions d'explicabilité, d'estimation et de lissage. En effet, un tel modèle additif généralisé pouvant s'écrire comme un modèle linéaire généralisé hyperparamétré (Section 3.2.4), les fonctions composantes et les paramètres de lissage sont estimés simultanément. La capacité à extraire de manière fine le comportement de la donnée s'en trouve assez limitée. Par conséquent, la notion d'apprentissage doit pouvoir être exploitée davantage. Cependant, dans un milieu fortement réglementé comme celui de l'assurance, ce qui est appris est tout aussi important que la précision de l'apprentissage. Les modèles innovants ambitionnent de capturer de la non-linéarité inhérente à la donnée tout en restant intrinsèquement interprétables. Cette interprétabilité inclut nécessairement la quantification et la visualisation de l'apport de chaque variable explicative dans le modèle et une compréhension détaillée du processus prédiction.

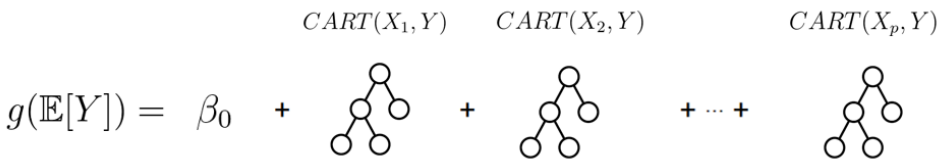


FIGURE 3.8 – Structure d'un modèle additif généralisé combinant des arbres de décision

L'objectif de cette section est de montrer comment les modèles additifs généralisés combinant des composantes univariées des modèles de *Machine Learning* peuvent contribuer à cette finalité. Un exemple d'une telle structure est donné par la Figure 3.8. Nous détaillons le fonctionnement d'un modèle additif généralisé avec une structure arborescente qui constitue alors le coeur de ce mémoire (Section 3.3.1). Une revue rapide de l'alternative avec une structure neuronale est également proposée pour prendre du recul par rapport à la nouvelle méthodologie introduite (Section 3.3.2).

3.3.1 GAM avec une structure arborescente

Dans cette sous-section, nous allons présenter le fonctionnement d'un modèle additif généralisé combinant des composantes univariées des modèles de *Machine Learning*. Si la structure arborescente sera utilisée au cours de la démonstration, c'est essentiellement la nouvelle manière d'apprentissage que nous souhaitons expliciter ici. En effet, c'est cette dernière qui définit ce que l'on appelle un modèle de *Machine Learning* intrinsèquement interprétable, un modèle *Glassbox* se situant à mi-chemin entre un modèle paramétrique et un modèle boîte noire. Dans ce qui suit, nous allons nous appuyer sur les articles de recherche LOU, CARUANA et GEHRKE (2012), CARUANA et al. (2015), WICK et al. (2019) et INTERPRETML TEAM (2021). La méthodologie étant récente et ses applications assez peu documentées, notre objectif premier sera d'expliciter et d'illustrer le formalisme mathématique sous-jacent. Dans un second temps, nous nous pencherons sur l'implémentation. Nous tâcherons en particulier de combler les lacunes de l'algorithmie existante pour l'adapter à notre problème de prédiction et modéliser séparément le coût moyen et la fréquence.

Bagging, Boosting, Binning et apprentissage des fonctions composantes

Les premiers modèles additifs généralisés combinant des fonctions de *Machine Learning* sont ceux ayant une structure arborescente (Lou, Caruana et Gehrke (2012)). Reprenant les concepts de la Section 2.3.2, plusieurs niveaux de complexité sont alors à explorer :

- les arbres de décision seuls ;
- un apprentissage indépendant de plusieurs arbres de décision, puis une moyennisation des résultats avec une procédure de type *Bagging* ;
- un apprentissage adaptatif de plusieurs arbres de décision, puis une agrégation des résultats avec une procédure de type *Boosting* ;
- une combinaison des procédures de *Bagging* et *Boosting* sur les arbres de décision.

Les articles de référence sur le sujet (Lou, Caruana et Gehrke (2012) et Caruana et al. (2015)) établissent que la combinaison des procédures de *Boosting* et *Bagging* est d'une pertinence accrue pour la modélisation additive généralisée. Dit autrement, une amélioration récursive, puis itérative du modèle additif généralisé couplée à une structure arborescente permet de surpasser les performances du modèle traditionnel combinant des fonctions splines. C'est ce cadre d'apprentissage qui sera détaillé en s'appuyant plus précisément sur l'algorithme *Explainable Boosting Machine* (**EBM**) développé par Nori et al. (2019).

L'algorithme *Explainable Boosting Machine* est un modèle *Glass Box*, un modèle de *Machine Learning* intrinsèquement interprétable utilisant des arbres CART peu profonds comme briques de base. Au niveau des fonctions composantes, cette architecture du modèle d'apprentissage se traduit par $f(\cdot) = CART(\cdot, \cdot)$. Pour satisfaire la contrainte d'interprétabilité, l'apprentissage des fonctions composantes $f_j,\ j = 1, \ldots, p$ dans le modèle additif généralisé (Équation 3.1) s'effectue de manière indépendante en suivant un raisonnement cyclique. Contrairement par exemple à l'algorithme *Gradient Boosting*, où les variables explicatives sont traitées simultanément, un apprentissage cyclique réserve un traitement spécifique, parallélisable, à chaque variable explicative $X_j,\ j = 1, \ldots, p$. S'ajoute à cela la procédure de *Binning*, ou groupement des données par classe, avec un principe d'équidistribution ou de distribution par quantiles pour chaque prédicteur X_j. Les variables explicatives, séparées en *bins* $b_j^k,\ k = 1, \ldots, K$, sont ainsi considérées de manière indépendante en un cycle à la suite duquel une modification adéquate à la prédiction $\hat{Y}$ de Y est appliquée. Le nombre K de *bins* est identique pour chaque prédicteur. Plus précisément, soit f_j^k la fonction composante relative au j-ème prédicteur et au k-ème *bin*. Notons également $t = 1, \ldots, T$ l'itération courante et $\tau = t - 1$ l'itération précédente dans le cycle à mesure que l'apprentissage de l'algorithme progresse. Les fonctions composantes f_j^k sont estimées selon le principe général suivant :

1. Calcul de la moyenne globale β_0 de y à travers tous les *bins* de toutes les variables explicatives :

$$\beta_0 = \frac{1}{n} \sum_{i=1}^{n} y_i \,;$$

2. Initialisation des fonctions composantes $f_j,\ j = 1, \ldots, p$ dans tous les *bins* k :

$$\forall\, k = 1, \ldots K,\ f_1^k, \ldots, f_p^k = 0 \,;$$

3. Itération cyclique à travers tous les prédicteurs $j = 1, \ldots, p$, calcul des facteurs partiels g et des

facteurs agrégés f à chaque cycle et pour chaque *bin* k :

$$\begin{cases} \hat{y}_{k,\tau} & = \beta_0 + \sum_{j=1}^{p} f_{j,\tau}^{k}, & \text{estimation de } y \text{ à l'itération précédente ;} \\ g_{j,t}^{k} & = \sum_{x_{i,j} \in b_j^k} y_i - \hat{y}_{k,\tau}, & \text{résidu pour la variable } j \text{ dans le } \textit{bin } k \text{ à l'itération courante ;} \\ f_{j,t}^{k} & = \sum_{s=1}^{t} g_{j,s}^{k}, & \text{somme des facteurs partiels disponibles à l'itération courante.} \end{cases}$$

4. Arrêt si le nombre maximal d'itérations est atteint ou s'il n'y a plus d'amélioration au niveau de la fonction de perte l.

Dans l'algorithme *Explainable Boosting Machine*, la fonction de perte l correspond l'opposé de la log-vraisemblance associée à la fonction de densité de la distribution de probabilité utilisée dans le modèle additif généralisé. Reprenant la Section 2.2.2, la fonction de perte à minimiser dans le cas d'une loi de Poisson $\mathcal{P}(\lambda)$ est

$$l(y, \hat{y}) = \sum_{i=1}^{n} y_i \ln(\hat{y}_i) - \hat{y}_i.$$

Pour la loi Gamma $\mathcal{G}(\alpha, \beta)$, la fonction de perte est donnée par

$$l(y, \hat{y}) = \sum_{i=1}^{n} \hat{y}_i y_i - \alpha \ln(\frac{\hat{y}_i}{y_i}) - (\alpha - 1) \ln(y_i),$$

où α est un paramètre du modèle.

Le détail des calculs pour toutes les variables explicatives est disponible dans l'Algorithme 1. La progression de la première itération (en rouge) pour l'apprentissage de la première fonction composante relative au premier prédicteur (en bleu) est indiquée pour faciliter la compréhension de l'aspect cyclique de la procédure. Afin d'expliciter le fonctionnement de la procédure de *Binning*, considérons une seule itération de l'Algorithme 1 avec $n = 6$, $p = 4$ et $K = 3$ et un principe d'équidistribution pour le groupement des données. La Figure 3.9 permet de visualiser une telle matrice des prédicteurs.

	X_1	X_2	X_3	X_4	
1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$k = 1$
2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$	
3	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	$X_{3,4}$	$k = 2$
4	$X_{4,1}$	$X_{4,2}$	$X_{4,3}$	$X_{4,4}$	
5	$X_{5,1}$	$X_{5,2}$	$X_{5,3}$	$X_{5,4}$	$k = 3$
6	$X_{6,1}$	$X_{6,2}$	$X_{6,3}$	$X_{6,4}$	

FIGURE 3.9 – Exemple d'un *Binning* avec un principe d'équidistribution au niveau d'une matrice de prédicteurs simple

Pour la variable X_1, la composition des *bins* $b_1^k,\ k \in \{1,2,3\}$ est alors la suivante :

$$\begin{cases} X_{1,1}, X_{2,1} \in b_1^1 \\ X_{3,1}, X_{4,1} \in b_1^2 \\ X_{5,1}, X_{6,1} \in b_1^3 \end{cases}$$

La prédiction $\hat{y}_1$ au niveau du *bin* $k = 1$ est alors obtenue par $\hat{y}_1 = \beta_0 + f_1^1 + f_2^1 + f_3^1 + f_4^1$, où

$$\begin{cases} f_1^1 &= CART(X_{1,1}, X_{2,1}, Y) \\ f_2^1 &= CART(X_{1,2}, X_{2,2}, Y) \\ f_3^1 &= CART(X_{1,3}, X_{2,3}, Y) \\ f_4^1 &= CART(X_{1,4}, X_{2,4}, Y). \end{cases}$$

Pareillement, pour le *bin* $k = 2$, on a $\hat{y}_2 = \beta_0 + f_1^2 + f_2^2 + f_3^2 + f_4^2$, où

$$\begin{cases} f_1^2 &= CART(X_{3,1}, X_{4,1}, Y) \\ f_2^2 &= CART(X_{3,2}, X_{4,2}, Y) \\ f_3^2 &= CART(X_{3,3}, X_{4,3}, Y) \\ f_4^2 &= CART(X_{3,4}, X_{4,4}, Y). \end{cases}$$

Finalement, pour le *bin* $k = 3$, on obtient $\hat{y}_3 = \beta_0 + f_1^3 + f_2^3 + f_3^3 + f_4^3$, où

$$\begin{cases} f_1^3 &= CART(X_{5,1}, X_{6,1}, Y) \\ f_2^3 &= CART(X_{5,2}, X_{6,2}, Y) \\ f_3^3 &= CART(X_{5,3}, X_{6,3}, Y) \\ f_4^3 &= CART(X_{5,4}, X_{5,4}, Y). \end{cases}$$

Dans l'algorithme *Explainable Boosting Machine*, l'aspect *Boosting* vient de l'apprentissage sur les résidus lors de l'estimation des fonctions composantes. En effet, la première fonction composante f_1 est apprise sur X_1 dans le but de prédire Y. La seconde fonction composante f_2 est apprise sur X_2 dans le but de prédire $y - f_1(x_1)$, la troisième f_3 dans le but de prédire $y - f_1(x_1) - f_2(x_2)$, et ainsi de suite :

$$\begin{cases} f_1(X_1) &= CART(X_1, Y) \\ r_j &= Y - f_j(X_j),\ \forall\, j = 1, \ldots, p-1 \\ f_j(X_j) &= CART(X_j, r_{j-1}),\ \forall\, j = 2, \ldots, p \\ f_p(X_p) &= CART(X_p, r_{p-1}) \end{cases}$$

Cette procédure est répétée T fois de manière indépendante. La fonction composante finale étant une moyenne des fonctions composantes apprises à chaque itération, l'aspect *Bagging*, que l'on peut qualifier de local, vient de l'agrégation indépendante des résultats pour chaque prédicteur. Dans la version développée par Nori et al. (2019), le processus décrit dans l'Algorithme 1 est répété plusieurs fois, puis les résultats sont moyennisés pour produire la prédiction finale. Il y a donc un aspect *Bagging* global, tel décrit dans la Section 2.3.2, qui complète l'algorithme. Ces considérations sont résumées dans la Figure 3.10.

Algorithm 1 Un cycle complet de l'algorithme **EBM**

Initialisation
$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$ ▷ Estimation de la moyenne globale sur l'échantillon d'apprentissage
for $k = 1, \ldots, K$ **do**
  $f_1^k, \ldots, f_p^k = 0$ ▷ Première estimation des fonctions composantes
end for
Itération 1 : $t = 1$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 1
  $\hat{y}_{k,\tau} \leftarrow \beta_0 + \underbrace{0 + \ldots + 0}_{p \text{ fois}}$ ▷ Estimation courante de y_i
  $g_{1,1}^k \leftarrow \sum_{x_{i,1} \in b_1^k} y_i - \hat{y}_{k,\tau}$
  $f_{1,1}^k \leftarrow g_{1,1}^k$ ▷ Première mise à jour de la première fonction composante
end for
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 2
  $\hat{y}_{k,\tau} \leftarrow \beta_0 + f_{1,1}^k + \underbrace{0 + \ldots + 0}_{p-1 \text{ fois}}$
  $g_{2,1}^k \leftarrow \sum_{x_{i,2} \in b_2^k} y_i - \hat{y}_{k,\tau}$
  $f_{2,1}^k \leftarrow g_{2,1}^k$
end for
... ▷ Calcul des fonctions composantes pour les prédicteurs 3 à $p-1$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur p
  $\hat{y}_{k,\tau} \leftarrow \beta_0 + f_{1,1}^k + f_{2,1}^k + \ldots + f_{p-1,1}^k + 0$
  $g_{p,1}^k \leftarrow \sum_{x_{i,p} \in b_p^k} y_i - \hat{y}_{k,\tau}$
  $f_{p,1}^k \leftarrow g_{p,1}^k$
end for ▷ Fin de l'itération 1
Itération 2 : $t = 2$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 1
  $\hat{y}_{k,\tau} \leftarrow \beta_0 + f_{1,1}^k + \ldots + f_{p,1}^k$
  $g_{1,2}^k \leftarrow \sum_{x_{i,1} \in b_1^k} y_i - \hat{y}_{k,\tau}$
  $f_{1,2}^k \leftarrow g_{1,1}^k + g_{1,2}^k$ ▷ Deuxième mise à jour de la première fonction composante
end for
for $i = k, \ldots, K$ **do** ▷ Prédicteur 2
  $\hat{y}_{k,\tau} \leftarrow \beta_0 + f_{1,2}^k + f_{2,1}^k + \ldots + f_{p,1}^k$
  $g_{2,2}^k \leftarrow \sum_{x_{i,2} \in b_2^k} y_i - \hat{y}_{k,\tau}$
  $f_{2,2}^k \leftarrow g_{2,1}^k + g_{2,2}^k$
end for
... ▷ Calcul des fonctions composantes pour les prédicteurs 3 à $p-1$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur p
  $\hat{y}_{k,\tau} \leftarrow \beta_0 + f_{1,2}^k + \ldots + f_{p-1,2}^k + f_{p,1}^k$
  $g_{p,2}^k \leftarrow \sum_{x_{i,p} \in b_p^k} y_i - \hat{y}_{k,\tau}$
  $f_{p,2}^k \leftarrow g_{p,1}^k + g_{p,2}^k$
end for ▷ Fin de l'itération 2
... ▷ **Itérations 3 à** $p-1$
Itération T **:** $t = T$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 1
  $\hat{y}_{k,\tau}\beta_0 + f_{1,T-1}^k + \ldots + f_{p,T-1}^k$
  $g_{1,T}^k \leftarrow \sum_{x_{i,1} \in b_1^k} y_i - \hat{y}_{k,\tau}$
  $f_{1,T}^k \leftarrow g_{1,1}^k + \ldots + g_{1,T}^k$ ▷ Tème mise à jour de la première fonction composante
end for
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 2
  $\hat{y}_{k,\tau} \leftarrow \beta_0 + f_{1,T}^k + f_{2,T-1}^k + \ldots + f_{p,T-1}^k$
  $g_{2,T}^k \leftarrow \sum_{x_{i,2} \in b_2^k} y_i - \hat{y}_{k,\tau}$
  $f_{2,T}^k \leftarrow g_{2,1}^k + \ldots + g_{2,T}^k$
end for
... ▷ Calcul des fonctions composantes pour les prédicteurs 3 à $p-1$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur p
  $\hat{y}_{i,\tau} \leftarrow \beta_0 + f_{1,T}^k + \ldots + f_{p-1,T}^k + f_{p,T-1}^k$
  $g_{p,T}^k \leftarrow \sum_{x_{i,p} \in b_p^k} y_i - \hat{y}_{k,\tau}$
  $f_{p,T}^k \leftarrow g_{p,1}^k + \ldots + g_{p,T}^k$
end for ▷ Fin de l'itération T

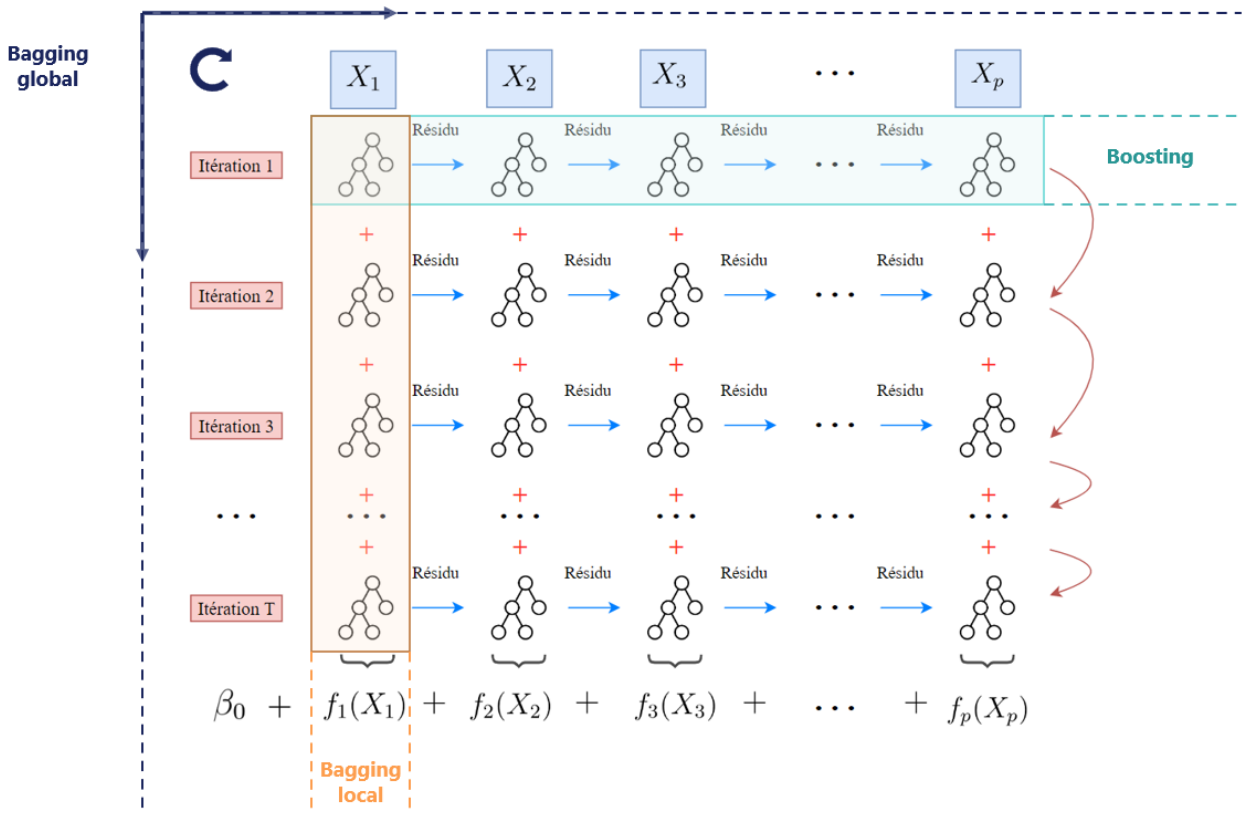


FIGURE 3.10 – Fonctionnement de l'algorithme *Explainable Boosting Machine*

L'ordre dans lequel les variables explicatives X_j, $j = 1, \ldots, p$ sont apprises peut potentiellement être significatif. En effet, l'apprentissage cyclique permet d'exploiter naturellement les relations de causalité entre les prédicteurs, le second apprenant sur les résidus du premier et ainsi de suite. Ainsi, si la première variable explicative permet de capter l'effet général d'un phénomène, les autres variables peuvent être utilisées pour apporter des informations de détail. Dans un contexte assurantiel, nous pouvons imaginer par exemple que l'effet général lié à l'âge est porté par la variable âge du conducteur principal. L'ancienneté de permis va venir préciser cette information en apportant des renseignements sur l'expérience de conduite. Avec la variable nombre de mois d'assurance, l'expérience d'assurance du conducteur peut être agrégée à l'ensemble. Les trois variables citées étant fortement corrélées, un apprentissage cyclique dans un ordre astucieusement choisi peut amener un supplément d'information lors de l'apprentissage supervisé, et donc une meilleure capacité de généralisation du modèle. La Figure 3.11 donne une représentation de la chaîne d'apprentissage évoquée.

Finalement, la procédure de *Binning* facilite l'apprentissage des événements rares si la séparation en groupes est convenablement choisie. L'optimisation s'effectuant localement dans chaque *bin* b_j^k, les observations atypiques sont traitées à part pour chaque prédicteur X_j, ce qui différencie une fois de plus l'algorithme *Explainable Boosting Machine* des autres méthodes de *Machine Learning*. Notons au passage une similitude avec le choix des noeuds dans la base de l'espace des splines (Section 3.2.2). En résumé, les aspects principaux de la modélisation sont les suivants :

- apprentissage cyclique des fonctions composantes ;
- amélioration récursive de l'algorithme grâce à l'aspect *Boosting* ;
- amélioration itérative de l'algorithme grâce à l'aspect *Bagging*, local et global ;

FIGURE 3.11 – Exemple d'exploitation des relations de causalité entre les variables explicatives dans l'algorithme *Explainable Boosting Machine* pour l'apprentissage de l'effet de l'âge

- prise en compte des relations de causalité entre les variables explicatives ;
- apprentissage des événements rares grâce à l'optimisation locale par *bins*.

Algorithme EBM dans le package `InterpretML`

Une première version de l'algorithme *Explainable Boosting Machine* accessible à l'utilisateur est disponible dans le package `InterpretML` (INTERPRETML TEAM (2021)). L'utilisateur peut piloter l'apprentissage de l'algorithme et la sophistication du modèle développé en plaçant le curseur sur :

- la construction de la structure arborescente en contrôlant la profondeur et la complexité des arbres de décision utilisés comme briques de base ;
- la procédure de *Boosting* en calibrant notamment le taux d'apprentissage (Section 2.3.2) ;
- la procédure de *Bagging*, locale et globale, en optimisant le nombre d'itérations et l'ajout d'une composante aléatoire ;
- la procédure de *Binning* en sélectionnant le type et la finesse du groupement de données.

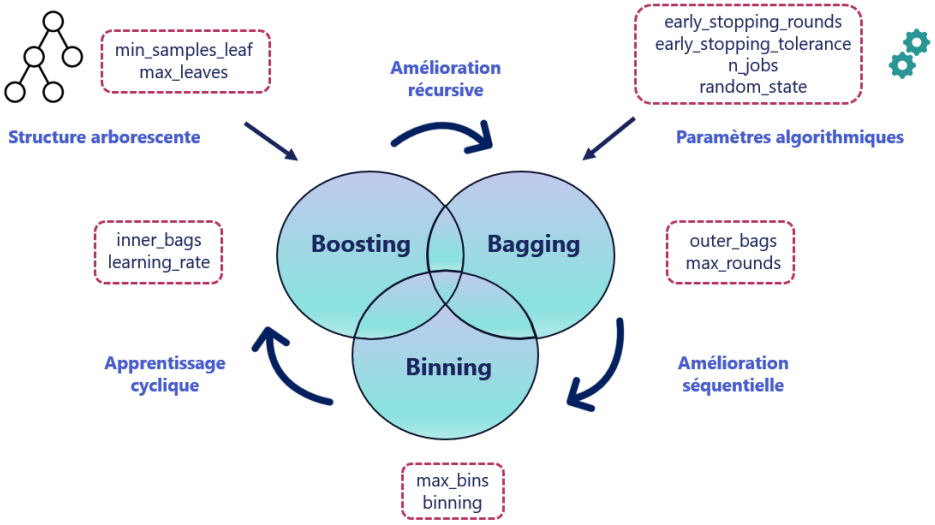


FIGURE 3.12 – Aspects essentiels du paramétrage de l'algorithme *Explainable Boosting Machine* dans le package `InterpretML`

L'ensemble des paramètres est optimisé en s'appuyant sur les méthodes de *Grid Search* ou *Random Search* (Section 2.3.4). Les significations mathématiques et les valeurs par défaut de tous les paramètres disponibles sont explicitées en Annexes (Section B.1.1). Les pistes à explorer lors du surajustement ou du sous-ajustement du modèle sont également indiquées. La Figure 3.12 fournit un résumé de la procédure de paramétrage.

Le taux d'apprentissage dans la procédure de *Boosting* mérite une attention particulière. En effet, il permet de piloter la force du lien de causalité entre les variables explicatives apparu grâce à l'apprentissage cyclique des fonctions composantes. Plus le taux d'apprentissage est important, plus l'ordre des prédicteurs sera implicitement pris en compte, et plus les relations de causalité seront amplifiées. Inversement, si le taux d'apprentissage est très faible, les potentielles relations de causalité entre les variables seront ignorées. L'allure des fonctions composantes est étroitement liée à la finesse de la procédure de *Binning* et à la complexité des arbres de décision utilisés comme briques de base dans le modèle additif généralisé. Ces aspects ayant un impact direct sur l'interprétabilité du modèle final, une étude de sensibilité aux paramètres est réalisée en Annexes (Section B.1.1).

Le package `InterpretML` propose une interface graphique qui permet une visualisation aisée des fonctions composantes dans le modèle additif généralisé. Les axes sont directement dans l'unité de la variable à prédire (fonction `Score`). Une représentation de la procédure de *Binning* est disponible en bas de la figure sous forme d'un histogramme qui montre le nombre d'observations dans chaque *bin* (fonction `Density`). Les intervalles de confiance sont obtenus grâce au paramètre `outer_bags` avec la procédure de *Bagging* globale décrite plus haut. Sur la Figure 3.13, on peut observer que le fait d'être un jeune conducteur augmente de plus de 400 le coût moyen. Ce dernier diminue par paliers jusqu'à environ 65 ans, avec une légère augmentation autour des 40-50 ans. Cette dernière pourrait être expliqué par la présence d'un second conducteur novice, typiquement liée au fait que les parents prêtent leur voiture aux enfants. A partir de 70 ans, le coût a une tendance croissante mais ne dépasse pas la barre du coût nul.

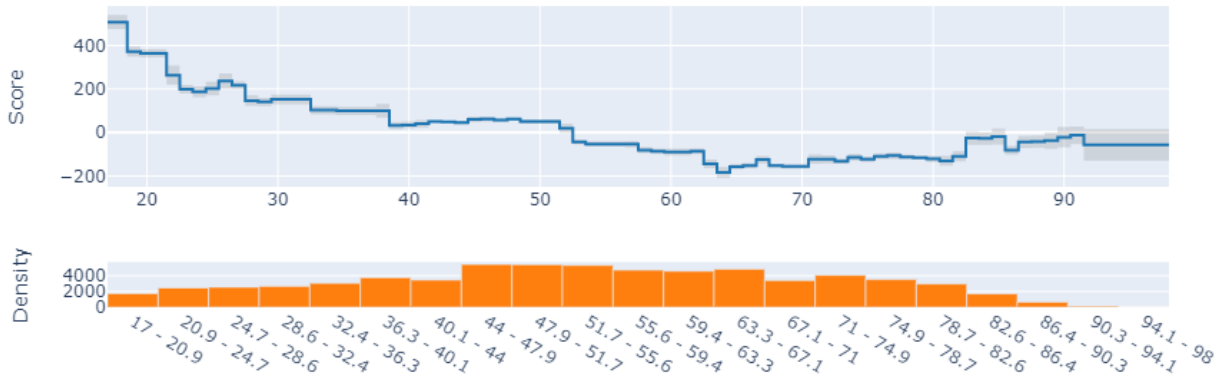


FIGURE 3.13 – Fonction composante associée à la variable `age_COND` dans le modèle GAM avec une structure arborescente pour la modélisation du coût moyen

Pour la modélisation de la fréquence, la Figure 3.14 montre deux pics distincts. Le premier, très marqué, s'observe pour les jeunes conducteurs jusqu'à 20 ans environ, avec une fréquence de sinistres atteignant 0,015. La deuxième augmentation intervient pour les conducteurs plus âgés, à partir de 70 ans environ, où la croissance est plus graduelle, puis stagnante à partir de 80 ans pour diminuer à nouveau à partir de 90 ans. Comparé aux Figures 3.6 et 3.7, les courbes obtenues avec l'algorithme *Explainable Boosting*

Machine présentent un aspect beaucoup moins lisse. Plutôt qu'une tendance générale, elles permettent d'identifier avec précision le comportement de chaque tranche d'âge, et ce avec un intervalle de confiance reflétant l'exposition. Cependant, les variations observées ne sont pas toujours intuitives et il est difficile de leur associer systématiquement un sens actuariel. Notons enfin que l'augmentation de la fréquence de sinistre pour les jeunes conducteurs est plus importante sur la Figure 3.14 que sur la Figure 3.7.

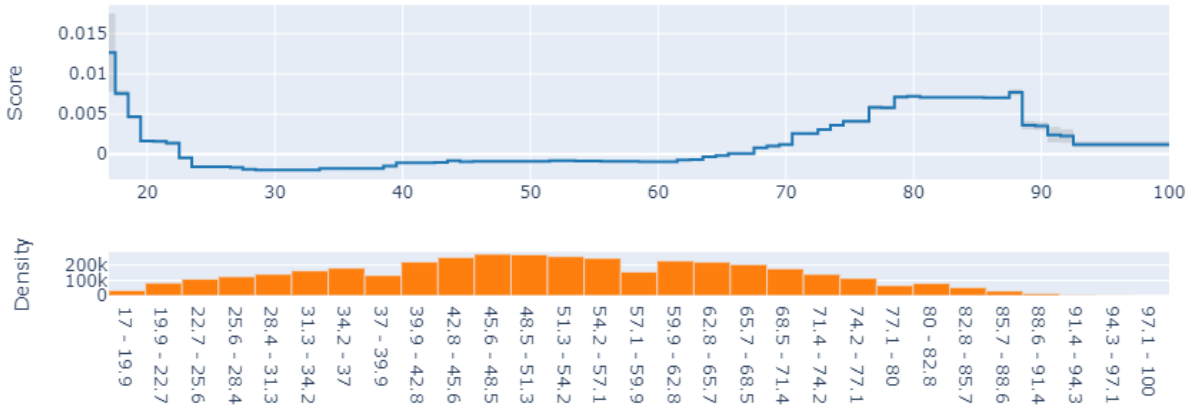


FIGURE 3.14 – Fonction composante associée à la variable `age_COND` dans le modèle GAM avec une structure arborescente pour la modélisation de la fréquence

Dans l'algorithme *Explainable Boosting Machine*, les variables qualitatives sont traitées avec la procédure de *label encoding*. Cette dernière permet d'associer un entier unique à chaque modalité en suivant l'ordre alphabétique. La Figure 3.15 montre la contribution de la variable qualitative `usage` dans la modélisation de la fréquence. Nous remarquons que les modalités `ambulance`, `taxi` et `tpm` (transport de marchandises) sont celles qui font augmenter le plus la fréquence de sinistres. Au contraire, la modalité `nr` (non-renseigné) fait diminuer la fréquence. La modalité `prive` fait plus augmenter la fréquence que la modalité `prive_travail` alors que cette dernière correspond à un usage plus large. En s'appuyant sur une analyse plus actuarielle, il serait judicieux de regrouper les modalités `ambulance`, `taxi` et `tpm` avec la modalité `pro` (usage professionnel) car elles définissent des risques similaires et sont assez peu représentées dans la base. De même, la modalité `prive` pourrait être regroupée avec la modalité `prive_travail` pour former une classe correspondant à un usage personnel. Pour la variable `nr`, en s'appuyant sur un principe de prudence, il serait envisageable de l'exclure de l'étude ou de la regrouper avec la modalité `prive_travail` qui est dominante.

Le package `InterpretML` permet également de considérer les interactions du second ordre. Pour cela, il s'appuie sur l'algorithme FAST (*Fast Interaction Detection*) (LOU, CARUANA, GEHRKE et HOOKER (2013)) pour déterminer le nombre optimal d'interactions à inclure dans le modèle additif généralisé, puis apprend une fonction composante par couple de variables explicatives en suivant la procédure décrite dans l'Algorithme 1. La Figure 3.16 montre la contribution du couple âge conducteur et classe de prix du véhicule (`age_COND` et `prix_SRA`) pour la modélisation du coût moyen. Nous observons ainsi deux zones à risque : l'association d'une voiture chère avec les jeunes conducteurs d'une part et les conducteurs âgés d'autre part. Ce résultat pouvait également s'observer dans le plan de prédiction obtenu avec l'algorithme *Extreme Gradient Boosting* (Figure 2.8).

FIGURE 3.15 – Contribution de la variable qualitative `usage` dans l'algorithme EBM

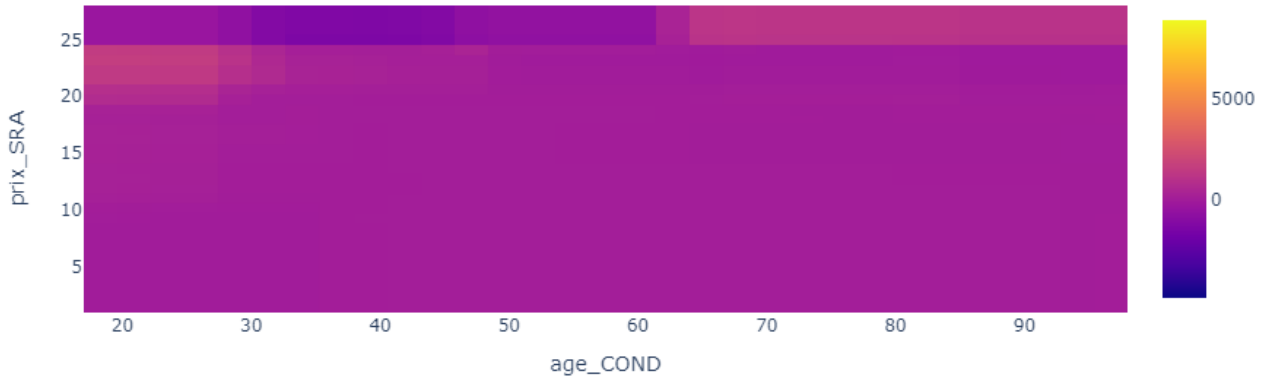


FIGURE 3.16 – Contribution du couple `age_COND` et `prix_SRA` pour la modélisation du coût moyen

Modèle additif et modèle multiplicatif

Dans le modèle linéaire généralisé tout comme dans le modèle additif généralisé, la fonction de lien occupe une place centrale car elle permet de transformer la structure de dépendance linéaire entre la variable réponse et les prédicteurs, et ce d'une manière adaptée à la distribution sous-jacente de la donnée. Le choix de la loi de distribution est déterminant, l'estimation des coefficients s'appuyant sur la maximisation de vraisemblance (Section 2.2.2 et 3.2.4). Dans l'algorithme *Explainable Boosting Machine*, l'optimisation interne est basée sur la minimisation d'une fonction de perte qui correspond à l'opposé de la log-vraisemblance (INTERPRETML TEAM (2021)). Cependant, dans la version actuelle du package `InterpretML`, seule la fonction de lien identité est implémentée pour un problème de régression, avec donc une hypothèse de distribution gaussienne pour la variable réponse. Pour un problème de classification, la fonction de lien logistique peut être utilisée avec une distribution binomiale pour la variable réponse. Si le cadre de la classification n'est pas celui exploré dans ce mémoire, nous pourrions par exemple envisager l'application de l'algorithme *Explainable Boosting Machine* avec la fonction de lien logistique à la prédiction de la probabilité qu'un sinistre soit grave ou pas dans un modèle de propension. Dans tous les cas, avec seulement deux fonctions de lien et distributions, l'algorithmie existante disponible dans le package `InterpretML` n'est pas suffisamment mature pour permettre à l'utilisateur de mettre en place une modélisation additive généralisée pour des problèmes de prédiction

plus complexes. Si la flexibilité de l'algorithme *Explainable Boosting Machine*, avec un paramétrage fin, permet de remédier en partie à ce défaut, le modèle ainsi construit possède une structure additive, tandis que les contraintes de la tarification imposent une structure multiplicative avec une fonction de lien logarithme. Pour répondre à ce problème, deux pistes sont envisagées dans ce mémoire[*] :

- une approche *ex ante* avec l'exploitation du méta-algorithme pour l'apprentissage des fonctions composantes dans l'algorithme *Explainable Boosting Machine* ;
- une approche *ex post* avec la prédiction du logarithme de la variable réponse Y et une correction du biais de prédiction.

L'analyse du code interne de l'algorithme *Explainable Boosting Machine* (InterpretML Team (2022)) montre que l'apprentissage des fonctions composantes dans le modèle additif généralisé s'appuie sur un méta-algorithme nommé *Cyclic Boosting Machine* (**CBM**, Wick et al. (2019)). Une version expérimentale non-optimisée de l'algorithme est disponible dans le dépôt GitHub Microsoft (2022). Cette dernière peut être utilisée de manière multiplicative, avec l'hypothèse d'une distribution Gamma pour la variable réponse Y. Tout comme dans l'algorithme *Explainable Boosting Machine*, le fonctionnement est cyclique, mais la notion de résidu est appréhendée différemment. Dans l'Algorithme 1, elle est basée sur une différence entre les prédictions tandis que dans l'algorithme *Cyclic Boosting Machine*, elle repose sur un quotient. La mise à jour des facteurs partiels g et des facteurs agrégés f adapte naturellement l'opération de somme en une opération de produit. Le détail des calculs est disponible dans l'Algorithme 2. Cependant, à l'heure actuelle, l'algorithme *Cyclic Boosting Machine* n'est pas suffisamment robuste pour mettre en oeuvre un paramétrage correct par *Grid Search* ou *Random Search* (Section 2.3.4). De plus, certains paramètres permettant de piloter l'apprentissage et la complexité du modèle développé ne sont pas encapsulés dans le méta-algorithme. En particulier, il n'est pas possible de contrôler directement la complexité des arbres de décision agrégés avec des paramètres tels que `min_samples_leaf` ou `max_leaves`. Un résumé de ces aspects est fourni par la Figure 3.17.

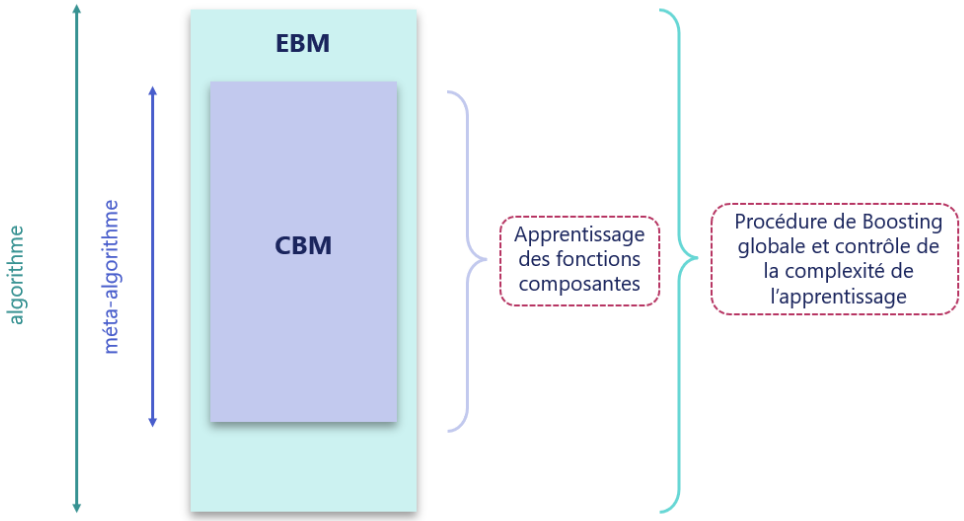
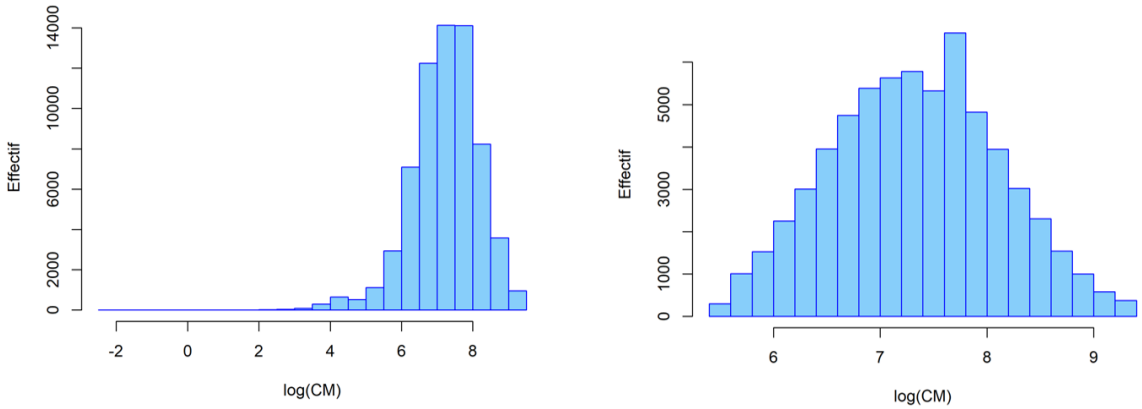


Figure 3.17 – Encapsulation de l'algorithme *Cyclic Boosting Machine* dans l'algorithme *Explainable Boosting Machine*

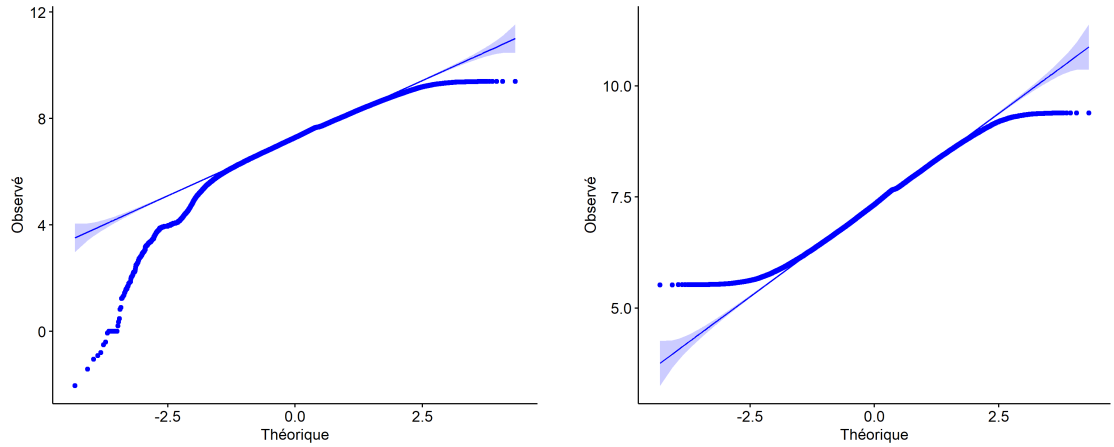
∗. Pour l'application, nous nous limiterons à la modélisation du coût moyen. En effet, la prédominance des zéros dans la distribution de la fréquence, à savoir l'absence de sinistre, soulève des problèmes de modélisation difficilement compatibles avec la méthodologie que l'on souhaite mettre en oeuvre.

Dans la deuxième approche corrective, on considère une nouvelle variable réponse $\tilde{Y} = \log(Y)$. Si on fait l'hypothèse que $\tilde{Y}$ suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, alors Y suit une loi log-normale $\mathcal{LN}(\mu, \sigma^2)$. En particulier, $\mathbb{E}[Y] = \exp(\mu + \frac{\sigma^2}{2})$. Cela suggère de corriger le biais dans la prédiction pour de prendre en compte que $\mathbb{E}[\exp(Y)] \neq \exp(\mathbb{E}[Y])$. Concrètement, une modification de la distribution de la variable $\tilde{Y}$ doit d'abord être appliquée afin de permettre à cette dernière d'avoir une allure normale. Pour cela, on utilise le seuil de sinistralité extrême fixé à $u_1 = 12000$ (Section 1.2.4) et on détermine un seuil u_2 permettant de corriger le début de la distribution. On s'appuie d'une part sur les histogrammes (Figure 3.18) et d'autre part sur les diagrammes quantile-quantile, ou QQ-plot, permettant d'évaluer la pertinence de l'ajustement d'une distribution donnée à une loi normale (Figure 3.19). Un seuil $u_2 = 250$ peut ainsi être retenu.



(a) pas de seuil (b) $u_2 = 250$

FIGURE 3.18 – Histogramme de la distribution du logarithme du coût moyen



(a) pas de seuil (b) $u_2 = 250$

FIGURE 3.19 – QQ-plot de la distribution du logarithme du coût moyen obtenu avec le package `ggpubr` (KASSAMBARA (2020))

Notons également que des tests statistiques de normalité tels que le test de Shapiro-Wilk ou le test de Kolmogorov-Smirnov auraient pu être utilisés. Cependant, nous nous limitons ici à une recherche de pseudo-normalité dans une optique de correction de distribution afin de faciliter la modélisation subséquente. Une analyse visuelle est ainsi estimée suffisante. La Figure 3.20 montre la fonction composante associée à la variable `age_COND` pour la prédiction du logarithme du coût moyen dans le

modèle additif généralisé avec une structure arborescente. L'échelle de l'axe des ordonnées est ainsi directement comparable à celle de la Figure 3.6. Nous remarquons de nouveau des coûts plus élevés pour des jeunes conducteurs jusqu'à 20 ans environ, puis pour des conducteurs plus âgés avec des larges intervalles de confiance à partir de 90 ans.

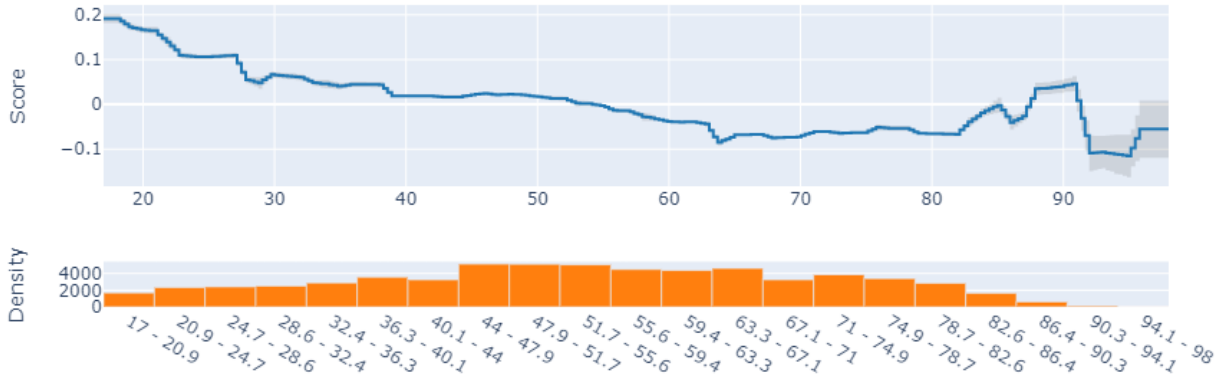


FIGURE 3.20 – Fonction composante associée à la variable `age_COND` dans le modèle GAM avec une structure arborescente pour la modélisation du logarithme du coût moyen

Dans la Section 3.3.1, nous avons présenté le fonctionnement d'un modèle additif généralisé avec une structure arborescente. Nous avons détaillé l'apprentissage cyclique des fonctions composantes tout comme les procédures de *Boosting* et *Bagging* permettant une amélioration récursive, puis itérative de l'algorithme. Nous avons vu que cette dernière agissait à un niveau local et global. Pour répondre aux contraintes d'une structure tarifaire multiplicative, nous avons tâché de décliner l'algorithmie existante pour proposer un modèle adapté aux distributions et aux fonctions de lien souhaitées. L'analyse graphique des fonctions composantes a permis de pointer les premières différences avec l'approche par splines. L'objectif de la sous-section suivante est d'introduire les composantes univariées des modèles de *Machine Learning* ayant une structure neuronale.

3.3.2 GAM avec une structure neuronale

Récemment, des modèles additifs généralisés combinant des réseaux de neurones ont été introduits (AGARWAL et al. (2021)). Tout comme avec les arbres de décision, il s'agit d'apporter une compréhension plus fine de la donnée en exploitant davantage la notion de l'apprentissage, tout en maintenant une structure additive intrinsèquement interprétable (Figure 3.21).

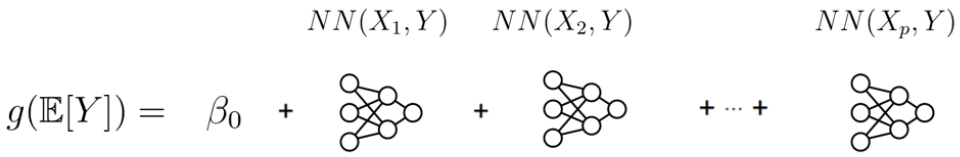


FIGURE 3.21 – Structure d'un modèle additif généralisé combinant des réseaux de neurones

L'utilisation des réseaux de neurones comme briques élémentaires dans le modèle apporte plus de flexibilité et facilite l'extension à d'autres contextes de prédiction, notamment la classification. Les fonctions composantes dans l'équation 3.1 sont estimées par l'algorithme de rétropropagation du gradient (Section 2.3.3). Une description plus détaillée du fonctionnement de l'algorithme peut être trouvée dans le dépôt GitHub développé par GOOGLE RESEARCH TEAM (2022). Par la suite, la visualisation des fonctions composantes fournit une représentation exacte de comment est réalisée la prédiction. Un exemple est donné par la Figure 3.22.

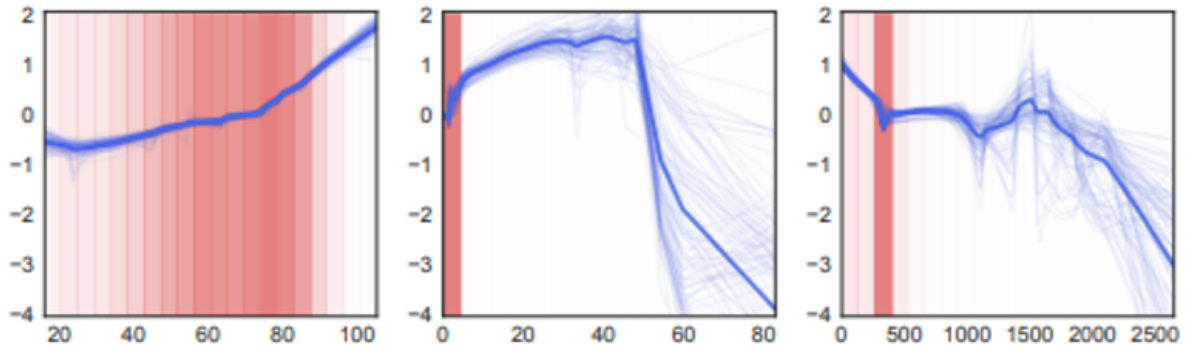


FIGURE 3.22 – Exemple de 3 fonctions composantes dans un modèle additif généralisé combinant des réseaux de neurones. Source : AGARWAL et al. (2021).

Un avantage de la structure de réseaux de neurones par rapport à la structure arborescente est la facilité de combinaison avec d'autres méthodes, que ce soit dans une optique de maximisation de performance avec l'apprentissage profond, ou avec un objectif de gain d'intelligibilité. Un exemple intéressant est donné par l'algorithme *Combined Actuarial Neural Network* (CANN, SCHELLDORFER et WUTHRICH (2019)) qui intègre le modèle linéaire généralisé dans une structure neuronale et combine les résultats sur la couche de sortie du réseau. Les avantages respectifs des deux méthodes sont ainsi combinés, à savoir l'interprétabilité de l'un avec la finesse de modélisation de l'autre.

Dans ce chapitre, nous avons établi le cadre théorique des modèles additifs généralisés. Nous avons distingué l'approche par splines, avec un lissage inhérent de l'information, et l'approche par composantes univariées des modèles de *Machine Learning* avec un apprentissage guidé de l'information. Nous avons ainsi défini une nouvelle mécanique d'apprendre garantissant l'interprétabilité intrinsèque d'un modèle de *Machine Learning*. Dans la démonstration, nous nous sommes

essentiellement appuyés sur une structure arborescente en combinant les procédures de *Boosting* et de *Bagging*. Nous avons également envisagé une structure neuronale et évoqué la possibilité de l'ajout des interactions dans le modèle.

Chapitre 4

Comparaison des modèles

4.1 Introduction

Dans les trois premiers chapitres, nous avons progressivement présenté le contexte assurantiel, la base de données, les modèles classiques en tarification non-vie et les modèles de *Machine Learning*. La notion d'interprétabilité a permis de guider l'exposé théorique du Chapitre 2 et d'amener la nouvelle mécanique d'apprentissage présentée dans le Chapitre 3. Entre un modèle linéaire généralisé et un modèle de *Machine Learning*, nous avons ainsi positionné un modèle *Glass Box*, un modèle de *Machine Learning* intrinsèquement interprétable (Figure 3.2).

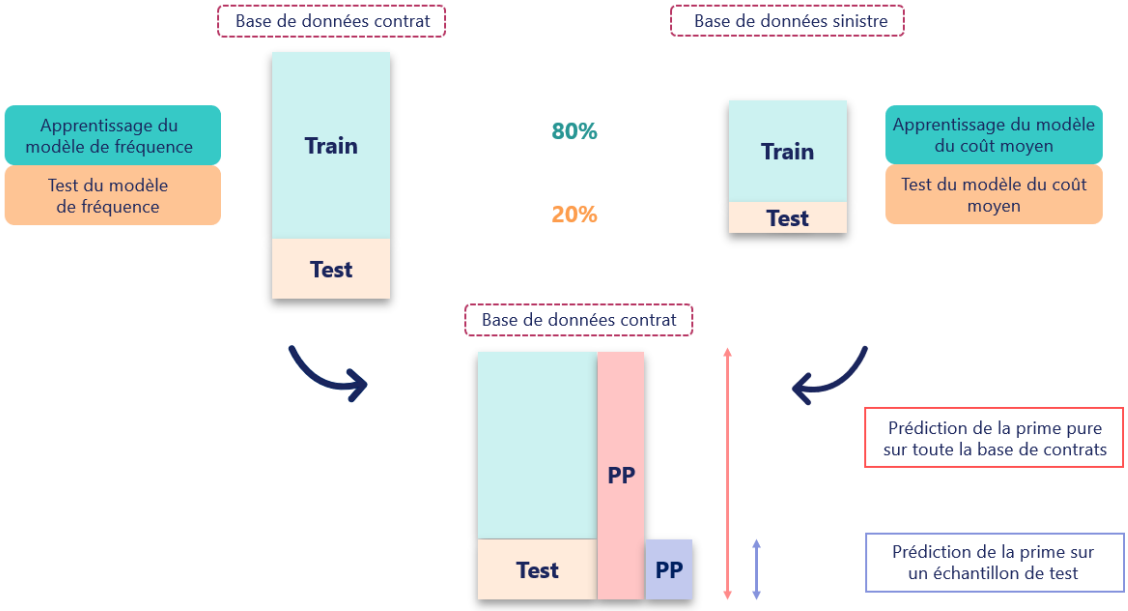


FIGURE 4.1 – Bases d'apprentissage et de test pour la modélisation de la garantie DTA

Les hypothèses standard en tarification non-vie permettent d'obtenir une décomposition coût-fréquence de la sinistralité (Équation 1.2). Pour la modélisation de la garantie dommage tout accident (DTA) en assurance automobile, le coût moyen et la fréquence des sinistres ont ainsi été étudiés séparément en s'appuyant sur deux bases de données d'assureur, la base de sinistres et la base de contrats (Section 1.2.1). Pour garantir l'objectivité et la comparabilité des résultats, nous avons systématiquement

distingué la base d'apprentissage et la base de test dans nos modélisations. Pour entraîner les modèles, nous avons utilisé 80% de la base. Le test de leur capacité de généralisation a ainsi été effectué sur les 20% restants. Nous avons également veillé à construire des bases d'apprentissage et de test ayant des caractéristiques similaires, à savoir approximativement les mêmes distributions pour la variable à prédire et les prédicteurs quantitatifs et la même répartition des modalités pour les prédicteurs qualitatifs. Finalement, pour la modélisation de la prime pure, nous avons combiné les modèles de coût moyen et de la fréquence dans une approche multiplicative. Ces considérations sont résumées sur la Figure 4.1.

Nous avons vu que les modélisations du coût moyen et de la fréquence soulèvent des enjeux et des limites différents. Néanmoins, pour les deux composantes, nous avons proposé un modèle linéaire généralisé (modèle de référence, Section 2.2), un ou plusieurs modèles de *Machine Learning* (modèles de performance, Section 2.3) et une ou plusieurs déclinaisons de l'algorithmie explicitées dans le Chapitre 3 (nouveaux modèles). Nous nous sommes appuyés sur les variables explicatives décrites dans la Table 1.2 (Section 1.2.2).

L'objectif de ce chapitre est d'analyser l'ensemble des résultats obtenus durant ce mémoire. Nous proposons ainsi des indicateurs de performance quantitatifs et qualitatifs pour examiner les différents modèles de fréquence et de sévérité, et ce au niveau local et global. Au-delà d'une simple comparaison des modèles, nous accordons une importance particulière à l'analyse actuarielle. Après une interprétation des résultats (Section 4.2), nous nous intéressons à leur applicabilité opérationnelle (Section 4.3), puis aux mouvements tarifaires induits par les différentes modélisations (Section 4.4).

4.2 Interprétation des résultats

Dans cette section, nous nous penchons sur l'interprétation des résultats et l'analyse qualitative des modèles du coût moyen et de la fréquence. Rappelons que les modèles de *Machine Learning* ne sont interprétables qu'à *posteriori*. Qu'elle soit spécifique ou agnostique (voir les Sections 2.4.2 et 2.4.3), leur interprétation nécessite des analyses supplémentaires. Face à cela, les modèles linéaires généralisés et les modèles additifs généralisés offrent une interprétabilité intrinsèque, ils sont immédiatement intelligibles. Rappelons également que les modèles *Glass Box* se situent à la frontière de l'agnostique et du spécifique des méthodes interprétables *a priori*, ce qui leur confère une certaine spécificité (Figure 2.12). L'interprétabilité des modèles de *Machine Learning* étant évoquée dans la Section 2.4, nous nous focalisons ici sur modèles linéaires généralisés (Section 2.2) et les modèles additifs généralisés (Sections 3.2 et 3.3).

4.2.1 Analyse des coefficients du modèle GLM

Les modèles linéaires généralisés sont traditionnellement utilisés pour la tarification en assurance non-vie. Avec l'application de la fonction de lien logarithmique, ils permettent d'obtenir une grille tarifaire multiplicative immédiatement interprétable et explicable, constituant ainsi le modèle de référence des assureurs. Pour la prédiction du coût moyen, la loi Gamma, permettant de modéliser des distributions à valeurs positives asymétriques à droite, est adoptée. La loi de Poisson, décrivant le nombre d'occurrences d'un phénomène discret, est utilisée pour la modélisation de la fréquence.

Modélisation du coût moyen

La Table 4.1 montre l'exemple d'une grille tarifaire obtenue pour la modélisation du coût moyen avec le modèle linéaire généralisé de loi Gamma en utilisant les 8 variables explicatives de la Table 1.2. Les coefficients de ces variables sont statistiquement significatifs au sens du Z-test défini dans la Section 2.2.2. Nous nous baserons sur cette grille tarifaire standard pour comparer les autres modèles. Rappelons que le lien logarithmique n'est pas le lien canonique associé à la loi Gamma (Section 2.2.2). Néanmoins, son utilisation permet d'aboutir à un modèle multiplicatif, sans en altérer la structure et les performances (Équation 2.7). Les coefficients du modèle sont estimés par maximum de vraisemblance. En leur appliquant la fonction inverse du logarithme, la fonction exponentielle, ils deviennent alors ce que l'on appelle des *multipliers*.

Variable	β_j	$\exp(\beta_j)$
`novice_COND2.1`	0,211	1,236
`prix_SRA`	0,032	1,032
`annee`	0,030	1,031
`anc_COND`	-0,002	0,998
`age_COND`	-0,003	0,997
`cof_CRM`	0,801	2,227
`age_VEH`	-0,013	0,988
`zonier_DTA`	0,008	1,008

TABLE 4.1 – Coefficients du modèle GLM et *multipliers* pour la modélisation du coût moyen

L'analyse des coefficients de la Table 4.1 fournit alors les renseignements suivants :

- la présence d'un conducteur secondaire novice (`novice_COND2.1`) augmente de plus de 23% le coût moyen des sinistres pour la garantie DTA par rapport à la modalité de référence qui renseigne sur l'absence d'un second conducteur novice ;
- le passage d'une classe de prix du véhicule (`prix_SRA`) à la classe de prix qui lui est immédiatement supérieure augmente de 3% le coût moyen ;
- chaque année, le coût moyen augmente de 3% ;
- l'augmentation d'un an de l'ancienneté de permis du conducteur principal (`anc_COND`) diminue de 0,2% le coût moyen ;
- l'augmentation d'un an de l'âge du conducteur principal (`anc_COND`) diminue de 0,3% le coût moyen ;
- l'augmentation d'un point du coefficient de réduction-majoration (`cof_CRM`) fait multiplier par 2 le coût moyen ;
- l'augmentation d'un an de l'ancienneté du véhicule diminue de 1,2% le coût moyen ;
- le passage d'une classe de risque géographique dans le `zonier_DTA` à une classe de risque immédiatement supérieure augmente de 0,8% le coût moyen.

D'un point de vue actuariel, ces valeurs sont globalement cohérentes. Plus le conducteur est âgé et expérimenté, moins le coût de sinistres sera important. Inversement, plus la voiture est chère et plus la zone géographique est risquée, plus le coût augmente. L'augmentation du coût moyen avec le temps peut s'expliquer par l'inflation, notamment une hausse du prix des pièces détachées évoquée dans la Section 1.1.1. Le coefficient de réduction-majoration est une image de la sinistralité passée. Pour son interprétation, nous nous référons au système bonus-malus français, plus précisément aux articles A121-1 à A121-2 du Code des Assurances. Une valeur de 0,5 correspond à 13 années sans accident, une valeur de 1,5 à 2 accidents responsables par un. D'après notre modèle, pour passer d'un état à

l'autre, le coût moyen est alors multiplié par deux. Cependant, d'un point de vue commercial, cette augmentation semble être trop importante.

Pour analyser plus en détail la qualité d'ajustement du présent modèle linéaire généralisé utilisant les 8 variables explicatives de la Table 1.2, une analyse des résidus tout comme une comparaison avec un modèle utilisant les 34 variables explicatives retenues à l'issue de l'analyse exploratoire (Section 1.2.2) est effectuée. Cette étude peut être consultée en Annexe C.1.1.

Modélisation de la fréquence

De manière similaire à la modélisation du coût moyen, l'utilisation de la fonction de lien logarithme avec la loi de Poisson pour la modélisation de la fréquence conduit à un modèle multiplicatif.

Variable	β_j	$\exp(\beta_j)$
`novice_COND2.1`	0,483	1,622
`prix_SRA`	0,022	1,022
`annee`	-0,039	0,961
`anc_COND`	-0,008	0,992
`age_COND`	0,017	1,018
`cof_CRM`	1,902	6,699
`zonier_DTA`	0,008	1,008
`usage.prive`	0,036	1,086
`usage.pro`	0,085	1,215
`usage.taxi`	0,204	1,598
`usage.tpm`	0,198	1,578
`usage.ambulance`	0,275	1,883
`usage.nr`	-0,261	0,548

TABLE 4.2 – Coefficients du modèle GLM et *multipliers* pour la modélisation de la fréquence

La grille tarifaire de la Table 4.2 est obtenue à partir des variables explicatives de la Table 1.2 dont les coefficients sont tous statistiquement significatifs au sens du Z-test. Elle donne les renseignements suivants :

- la présence d'un conducteur secondaire novice augmente de plus de 60% la fréquence de sinistres pour la garantie DTA par rapport à la modalité de référence qui renseigne sur l'absence d'un second conducteur novice ;
- le passage d'une classe de prix du véhicule à une classe de prix immédiatement supérieure augmente de 2% la fréquence ;
- chaque année, la fréquence diminue de 4% ;
- l'augmentation d'un an de l'ancienneté de permis du conducteur diminue de 0,8% la fréquence ;
- l'augmentation d'un an de l'âge du conducteur augmente de 1,8% la fréquence ;
- l'augmentation d'un point du coefficient de réduction-majoration fait multiplier par plus de 6 la fréquence ;
- le passage d'une classe de risque géographique dans le `zonier_DTA` à une classe de risque immédiatement supérieure augmente de 0,8% la fréquence ;
- les modalités de la variable `usage` s'interprètent de la manière suivante :
  - si la voiture est utilisée à des fins personnels (`prive`), la fréquence de sinistres augmente de 8,6% par rapport à la modalité de référence, qui décrit un usage privé couplé à des

déplacements sur le lieu de travail (`prive_travail`) ;
— si la voiture est utilisée à des fins professionnels non précisés par ailleurs (`pro`), la fréquence augmente de 21% ;
— si la voiture est utilisée pour des services de taxi (`taxi`), la fréquence augmente de 60% ;
— si la voiture est utilisée pour le transport des marchandises (`tpm`), la fréquence augmente de 58% ;
— si la voiture est utilisée pour des services d'ambulance (`ambulance`), la fréquence augmente de 88% ;
— si l'usage de la voiture n'est pas renseigné (`nr`), la fréquence diminue de 45%.

Contrairement à la modélisation du coût moyen, quelques incohérences sont à relever dans les résultats de la modélisation de la fréquence. L'augmentation de l'âge du conducteur devrait, en moyenne, s'accompagner d'une baisse de la fréquence de sinistres. Pour la variable `usage`, l'augmentation de la fréquence pour un usage professionnel par rapport à la modalité de référence décrivant un usage personnel est cohérente. En effet, une sursinistralité est souvent afférente à un usage professionnel, notamment celui de taxi et d'ambulance. Cependant, la modalité `prive` correspond à un usage plus restreint que la modalité `prive_travail`, donc la fréquence devrait être en baisse. Comme nous l'avons déjà évoqué en commentant la contribution de la variable `usage` dans le modèle additif généralisé ayant une structure arborescente (Figure 3.15), il aurait été judicieux de regrouper les modalités correspondant à un usage professionnel (`pro`, `ambulance`, `taxi` et `tpm`) et les modalités correspondant à un usage personnel (`prive` et `prive_travail`). Pour la modalité non-renseigné (`nr`), il aurait fallu envisager soit la suppression, soit le regroupement avec la modalité majoritaire, à savoir la modalité `prive_travail`. La multiplication par 6 lors de l'augmentation d'un point dans le coefficient de réduction-majoration semble également beaucoup trop importante. Les variations observées sur les autres variables sont cohérentes. La diminution de la fréquence des sinsitres avec le temps peut s'expliquer notamment par la mise en place des différentes mesures de prévention ces dernières années (Dubessy (2021)).

Tout comme pour la modélisation du coût moyen, une analyse plus détaillé du présent modèle linéaire généralisé est disponible en Annexe C.1.2.

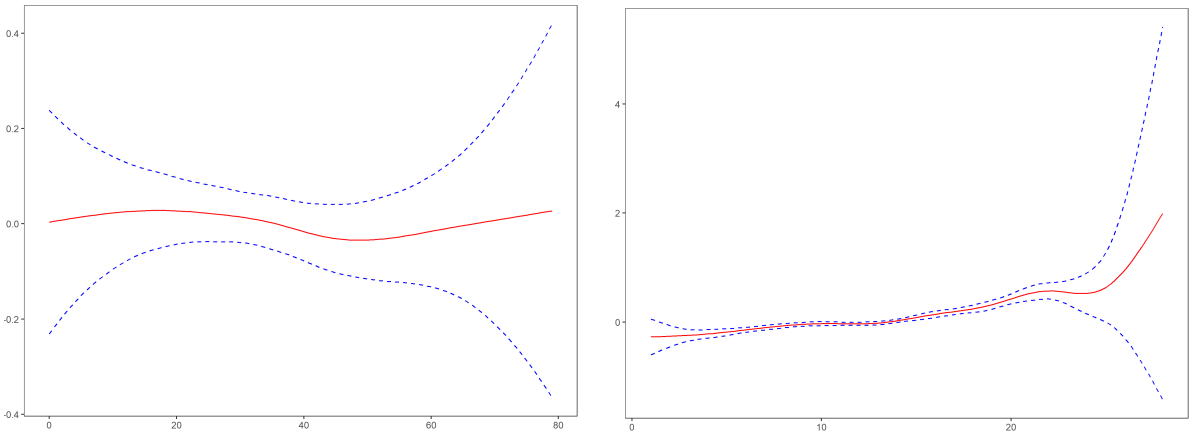
4.2.2 Analyse des fonctions composantes du modèle GAM

Dans un modèle additif généralisé, la dépendance entre la variable à prédire et les variables explicatives n'est plus nécessairement modélisée linéairement. Au lieu des coefficients constants, les variables explicatives participent à la prédiction au travers des fonctions composantes non-paramétriques (Équation 3.1). L'interprétation des modèles additifs généralisés revient essentiellement à l'analyse graphique de ces fonctions composantes, qui donnent une description exacte de comment est réalisée la prédiction au niveau de chaque variable explicative. Autrement dit, il s'agit de passer d'une analyse par lecture des coefficients à une analyse par lecture des fonctions. Dans le Chapitre 3, nous avons déjà commenté les fonctions composantes relatives à la variable âge conducteur (`age_COND`) pour la modélisation du coût moyen (Figures 3.6 et 3.13) et de la fréquence (Figures 3.7 et 3.14). Dans l'algorithme *Explainable Boosting Machine* (EBM), c'est-à-dire dans le modèle additif généralisé ayant une structure arborescente (Section 3.3.1), les fonctions composantes permettent d'identifier avec précision le comportement de chaque profil de risque. Cependant, les variations observées ne sont pas toujours intuitives et il n'est pas aisé de leur associer systématiquement un sens actuariel. Dans l'Annexe B.1.1, nous avons notamment discuté de l'influence de certains paramètres sur l'interprétabilité des fonctions composantes. Inversement, les fonctions composantes dans un modèle additif généralisé combinant des fonctions splines proposent une allure lisse, moyennisée et interprétable au risque d'omettre parfois des détails importants dans le comportement étudié. Finalement, dans un modèle linéaire généralisé, l'information donnée par

un coefficient est souvent insuffisante pour décrire la complexité du phénomène sous-jacent. Il s'agit alors de faire un arbitrage entre précision et pertinence actuarielle des modèles. L'objectif de cette sous-section est de poursuivre la discussion sur d'autres variables significatives dans la modélisation du coût moyen et de la fréquence.

Modélisation du coût moyen

La fonction composante associée à la variable ancienneté de permis du conducteur principal (`anc_COND`) présente une allure générale similaire dans le modèle additif généralisé ayant une structure arborescente (Figure 4.3) et dans celui combinant des fonctions splines (Figure 4.2a). La tendance est globalement décroissante jusqu'à environ 50 ans, puis croissante sur les âges plus élevés avec des intervalles de confiance plus larges reflétant le manque d'exposition. Cependant, sur la Figure 4.3, on observe deux pics distincts, l'un correspondant à 8 ans d'ancienneté environ et l'autre, plus marqué, à 30 ans. Si on fait l'hypothèse que l'on acquiert le permis de conduire à 18 ans, le premier pic n'a pas vraiment de signification actuarielle. En revanche, le second pic correspond à l'âge de 50 ans, c'est-à-dire à l'âge où les parents ont l'habitude de prêter leur voiture aux enfants. L'augmentation soudaine du coût moyen peut alors être expliquée par la présence d'un second conducteur novice. Sur la Figure 4.2a cette information n'apparaît pas ponctuellement, elle est lissée sur l'intervalle allant de 10 à 30 ans. Rappelons que dans la Figure 4.3, les axes sont directement dans l'unité de la variable à prédire (fonction `Score`) et une représentation de la procédure de *Binning* est disponible en bas de la figure sous forme d'un histogramme qui montre le nombre d'observations dans chaque *bin* (fonction `Density`). Ainsi, un niveau à 200 pour la tranche d'ancienneté de 0 à 4 ans environ traduit directement une augmentation de 200 par rapport à la barre du coût nul.



(a) `anc_COND` (b) `prix_SRA`

FIGURE 4.2 – Fonctions composantes dans le modèle GAM avec des splines cubiques pour la modélisation du coût moyen

Pour la variable donnant la classe de prix du véhicule (`prix_SRA`), les deux modèles additifs généralisés montrent une tendance croissante, avec une accélération importante à partir de la classe 23 environ associée à des larges intervalles de confiance. Sur la fonction composante du modèle avec une structure arborescente (Figure 4.4), on observe un pic associé aux classes 8 et 9 qui n'apparaît pas sur la fonction spline (Figure 4.2b). Cette augmentation peut être expliquée notamment par le manque d'exposition sur cette classe.

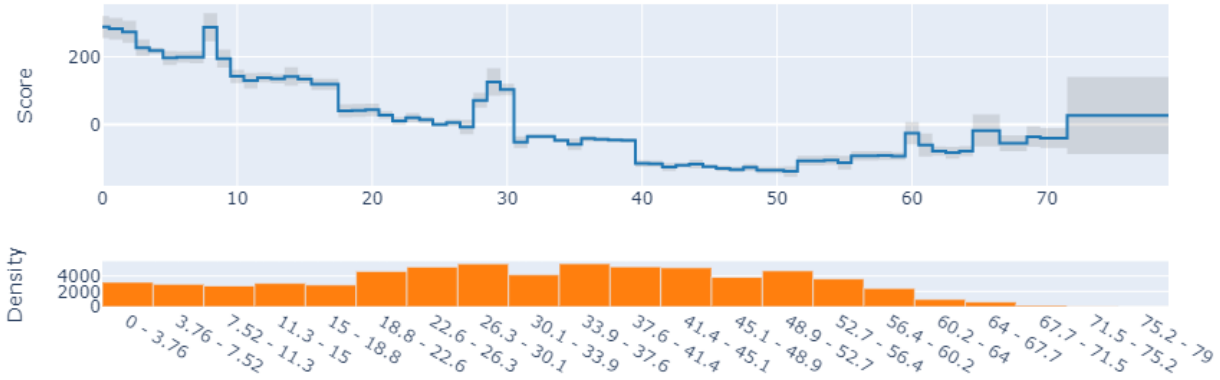


FIGURE 4.3 – Fonction composante associée à la variable `anc_COND` dans le modèle GAM avec une structure arborescente pour la modélisation du coût moyen

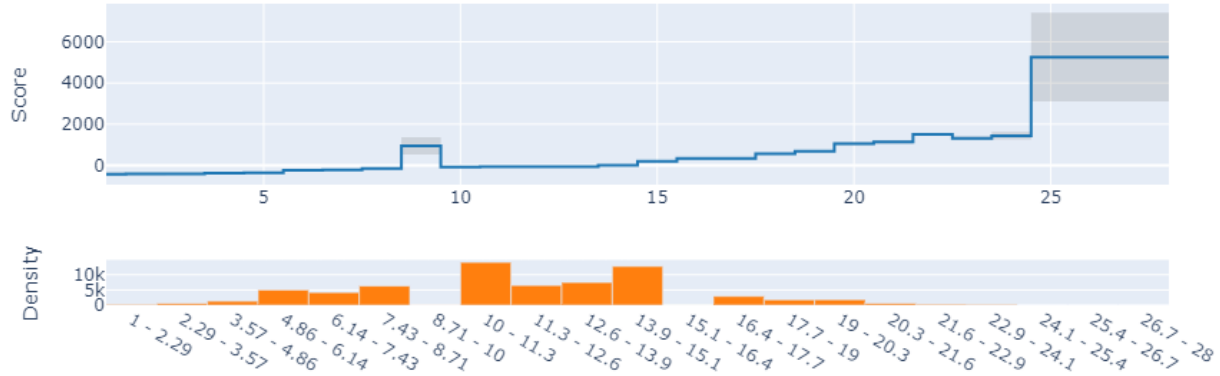


FIGURE 4.4 – Fonction composante associée à la variable `prix_SRA` dans le modèle GAM avec une structure arborescente pour la modélisation du coût moyen

Modélisation de la fréquence

Pour la modélisation de la fréquence, l'analyse des fonctions composantes relatives à la variable ancienneté de permis donne des résultats très similaires à ceux obtenus pour la variable âge conducteur (Figures 3.7 et 3.14). On observe en effet une augmentation de la fréquence de sinistres pour les conducteurs novices, puis pour les conducteurs expérimentés. L'augmentation de la fréquence pour les conducteurs expérimentés est cependant plus lente et graduelle. De nouveau, la tendance est plus lisse pour la fonction spline (Figure 4.5) et plus segmentée pour la fonction issue d'une structure arborescente (Figure 4.6).

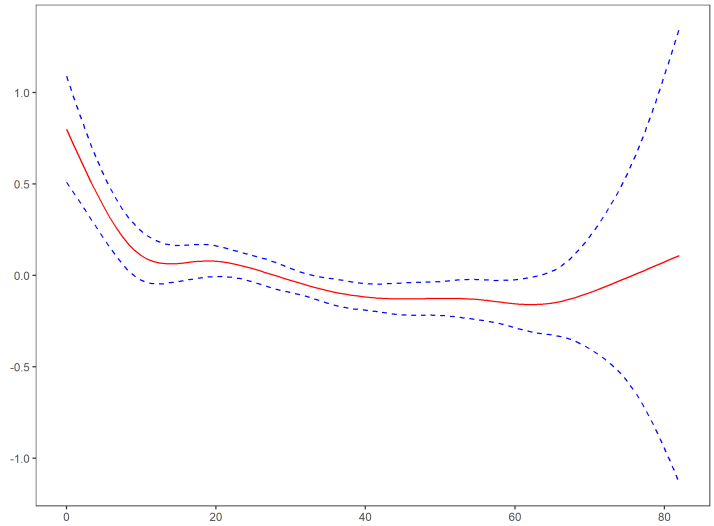


FIGURE 4.5 – Fonction composante associée à la variable `anc_COND` dans le modèle GAM avec des splines cubiques pour la modélisation de la fréquence

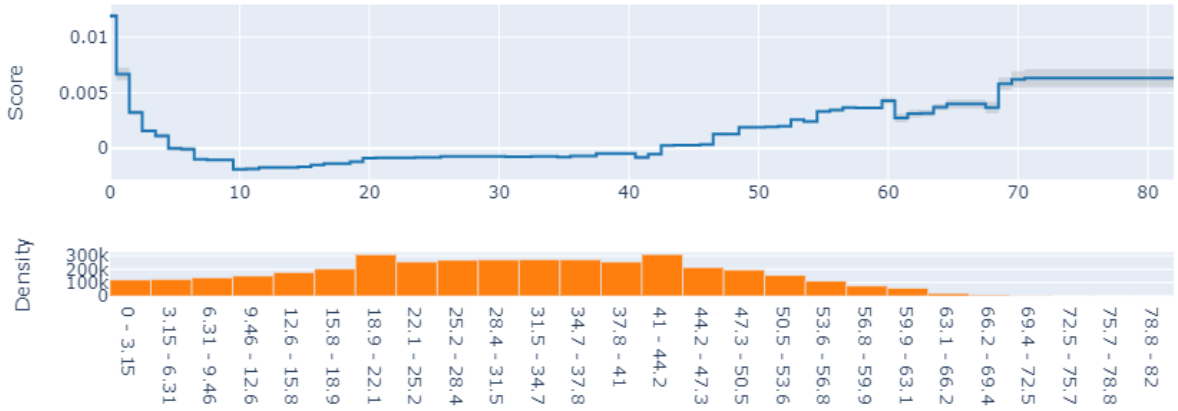


FIGURE 4.6 – Fonction composante associée à la variable `anc_COND` dans le modèle GAM avec une structure arborescente pour la modélisation de la fréquence

4.2.3 Analyse univariée des valeurs observées et prédites

Pour évaluer la qualité des modèles au niveau local, il convient de s'intéresser aux graphes des valeurs observées et prédites variable par variable sur un échantillon de test (Figure 4.1). Sous l'hypothèse de non-corrélation entre les variables explicatives, ces graphes peuvent renseigner sur l'effet marginal de chaque variable. Cependant, des corrélations existent dans notre jeu de données, notamment entre les variables âge du conducteur et ancienneté de permis (Figure 1.5). Il convient donc de nuancer les interprétations en prenant en compte cet aspect. Rappelons que les modèles ici comparées sont :

- le modèle linéaire généralisé (**GLM**, en rouge) ;
- le modèle additif généralisé combinant des splines cubiques (**Spline cubique**, en orange) ;
- le modèle issu de l'algorithme *Explainable Boosting Machine* (**EBM**, en violet), à savoir le

modèle additif généralisé ayant une structure arborescente.

Nous allons considérer séparément les modèles de coût moyen et de la fréquence et relever les différences de comportement induites par les trois modélisations.

Modélisation du coût moyen

Le graphe des valeurs observées et prédites pour la variable âge conducteur (Figure 4.7) montre un bon ajustement pour l'ensemble des modèles, avec des limites à soulever pour le modèle linéaire généralisé. Les courbes des deux modèles additifs sont très proches. Elles épousent mieux la courbe des valeurs observées, notamment sur l'intervalle des âges allant de 50 à 60 où le modèle linéaire généralisé a une tendance à surestimer. Globalement, les modèles additifs généralisés captent plus en détail la volatilité des données tandis que le modèle linéaire moyennise l'information.

Pour la variable donnant la classe de prix du véhicule (Figure 4.8), ce comportement se confirme. Notons également que sur les classes peu exposées, notamment celles correspondant à un prix élevé (classes 18 à 28 environ), l'ajustement des modèles additifs généralisés est sensiblement meilleur.

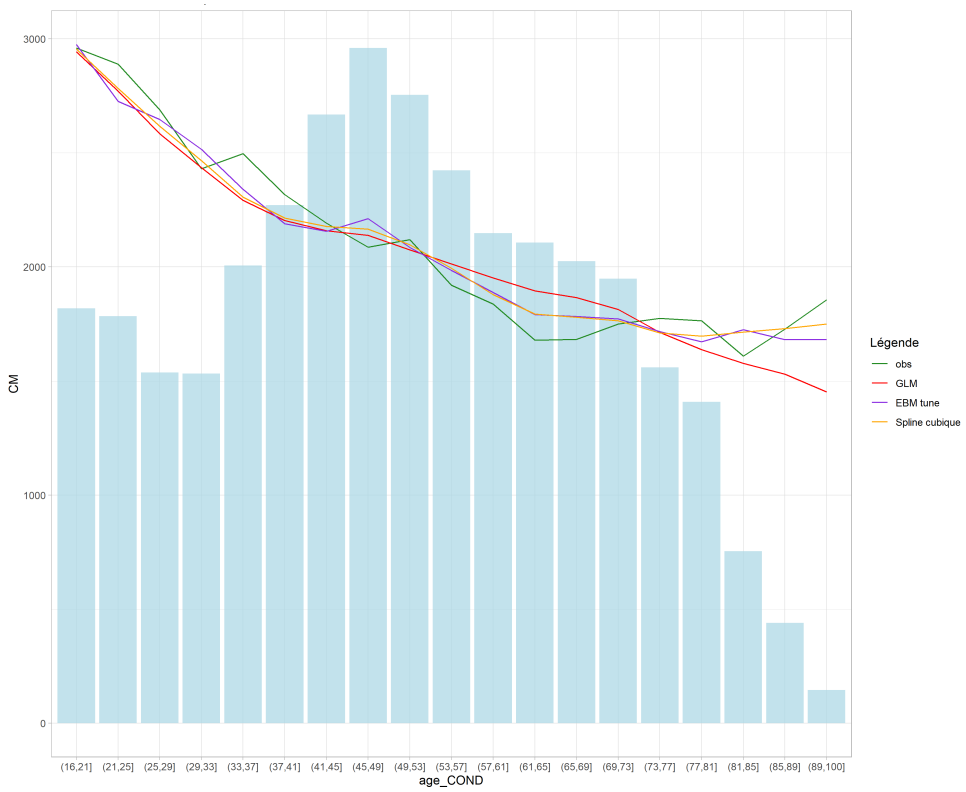


FIGURE 4.7 – Valeurs observées et prédites pour la modélisation du coût moyen : comparaison des modèles GLM, GAM splines et EBM sur la variable `age_COND`

Modélisation de la fréquence

Les tendances observées pour la modélisation du coût moyen se confirment et s'accentuent pour la modélisation de la fréquence. Le graphe obtenu pour la variable âge conducteur (Figure 4.9) montre un

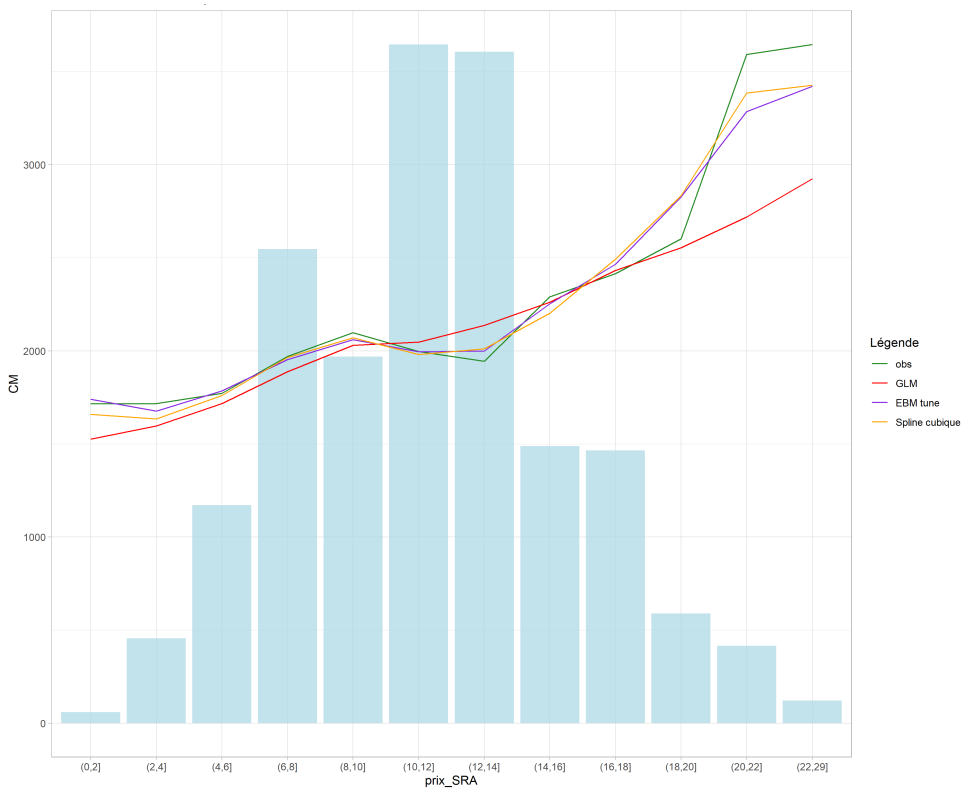


FIGURE 4.8 – Valeurs observées et prédites pour la modélisation du coût moyen : comparaison des modèles GLM, GAM splines et EBM sur la variable `prix_SRA`

modèle linéaire généralisé qui moyennise l'information au point d'omettre des détails importants dans le comportement du risque. Sur l'intervalle allant de 50 à 70 ans notamment, il surestime la fréquence et sur l'intervalle allant de 70 à 100 ans, il la sous-estime. Les modèles additifs généralisés répliquent bien la courbe des valeurs observées. Pour les très jeunes conducteurs (17 à 23 ans) ainsi que pour les conducteurs âgés (77 à 100 ans), le modèle additif généralisé ayant une structure arborescente semble être plus précis que le modèle combinant les fonctions splines.

4.3 Applicabilité opérationnelle

Après une première analyse qualitative, nous allons maintenant nous intéresser aux métriques de performance globales permettant de comparer les modèles. Nous évoquerons également les temps computationnels et les spécificités techniques des différentes modélisations. Dans cette section, nous allons analyser trois types de modèles :

- les modèles linéaires généralisés (**GLM**) ;
- les modèles additifs généralisés (**Spline cubique**, **EBM**, **EBM log** et **CBM**) ;
- les modèles de *Machine Learning* : les arbres de décision (**CART**, Section 2.3.1) et l'algorithme *Extreme Gradient Boosting* (**XGB**, Section 2.3.2).

Nous chercherons ainsi à positionner le nouveau modèle décrit dans le Chapitre 3 entre le modèle de référence et les modèles dits de performance. Nous tâcherons également d'évaluer de manière critique l'applicabilité opérationnelle et la pertinence actuarielle des différents modèles développés.

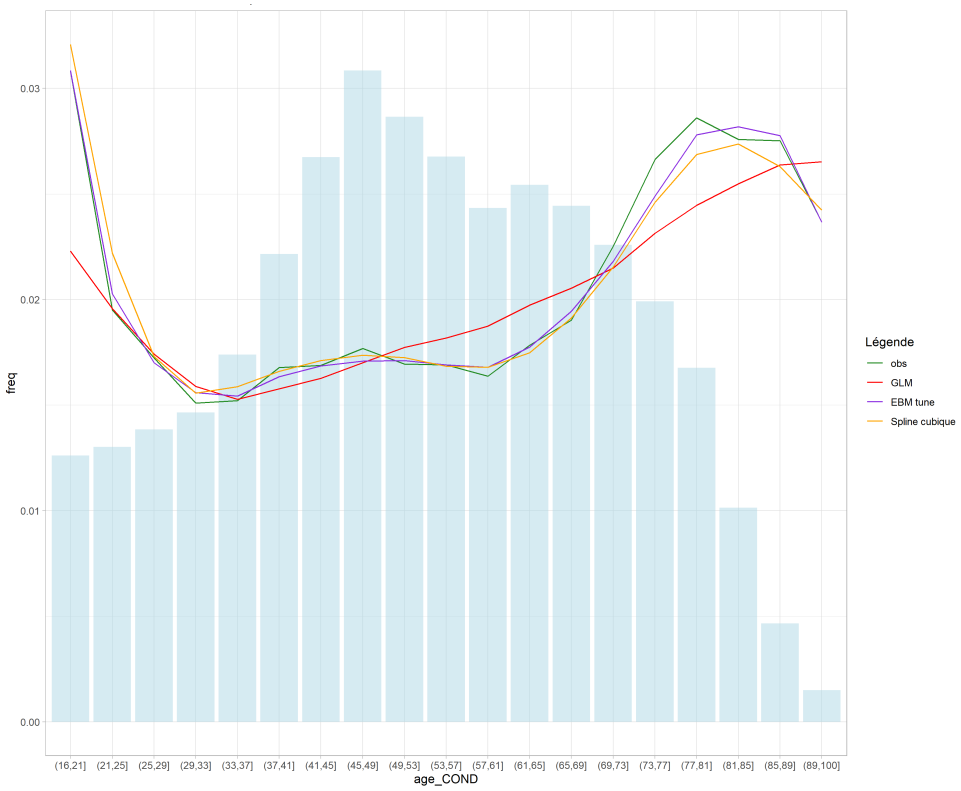


FIGURE 4.9 – Valeurs observées et prédites pour la modélisation de la fréquence : comparaison des modèles GLM, GAM splines et EBM sur la variable `age_COND`

4.3.1 Performance et temps computationnel

Les métriques de performance donnent une image du comportement global des modèles et opèrent ainsi un rapprochement entre des structures algorithmiques très différentes. Pour la modélisation du coût moyen et de la fréquence, l'utilisation des métriques classiques permet d'évaluer si le modèle est bon en moyenne et en médiane avec respectivement l'erreur quadratique moyenne $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ et l'erreur absolue moyenne $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. Cependant, comme nous l'avons évoqué dans la Section 2.3.5, l'application de ces métriques de performance à la modélisation du coût moyen ou de la fréquence fait apparaître des limites. En effet, pour prendre en compte des distributions asymétriques à droite ou un phénomène de comptage qui est discret par principe, des métriques basées sur la déviance Gamma ou la déviance poissonienne sont plus pertinentes, mais s'adaptent difficilement au cadre des modèles de *Machine Learning*. Dans tous les cas, la comparaison des métriques de performance sur les échantillons d'apprentissage et de test donne une quantification de la qualité d'apprentissage des modèles. En effet, si une dégradation importante des métriques s'observe entre l'échantillon d'apprentissage et de test, c'est que le modèle surapprend et possède une mauvaise capacité de généralisation.

Pour la modélisation du coût moyen, la Table 4.3 montre que les meilleures performances sont enregistrées par l'algorithme *Extreme Gradient Boosting*. Les deux modèles additifs généralisés donnent des résultats très similaires, avec une légère surperformance pour le modèle avec des fonctions splines. L'arbre de décision CART enregistre les pires performances de prédiction, montrant ainsi que l'agrégation des modèles de *Machine Learning* (Section 2.3.2) est souvent nécessaire pour garantir de bons résultats. Nous remarquons également que l'écart de performance entre un modèle de *Machine Learning*,

dont le nombre d'interactions n'est pas limité *a priori*, et un modèle additif généralisé combinant des composantes univariées des modèles de *Machine Learning* n'est pas significatif. Pour compléter les résultats de performances en valeur absolue, nous proposons également une vision en valeurs relatives par rapport au modèle de référence, à savoir le modèle linéaire généralisé.

Métrique	**RMSE**			**MAE**			**R2**	
Algorithme	Train	Test	%	Train	Test	%	Train	Test
GLM	2109	2027	100	1329	1308	100	0,062	0,061
Spline	2098	2016	99,5	1324	1301	99,4	0,072	0,072
EBM	2099	2017	99,5	1326	1304	99,7	0,071	0,071
CART	2133	2051	101,2	1341	1317	100,7	0,040	0,039
XGB	2072	**2016**	99,5	1302	**1293**	98,9	0,095	**0,072**

TABLE 4.3 – Résultats de comparaison des modèles pour la modélisation du coût moyen : GLM, Spline cubique, EBM, CART et XGB

Pour la modélisation de la fréquence, les écarts de performance entre les modélisations sont très peu significatifs, la différence se fait souvent au niveau de la quatrième décimale ou n'est pas détectée (Table 4.4). De nouveau, c'est l'algorithme *Extreme Gradient Boosting* qui enregistre les meilleures performances et l'arbre de décision CART les pires. Nous remarquons que le modèle additif généralisé ayant une structure arborescente se positionne entre le modèle linéaire généralisé et le modèle de *Machine Learning*, en dépassant les performances du modèle utilisant des fonctions splines. Si les différences entre les 5 modélisations ne sont pas notables au regard des métriques étudiées, les écarts apparaissent clairement sur les graphes des valeurs observées et prédites (Figures 4.9 et C.3). Autrement dit, les métriques de performance globales ont tendance à passer sous silence les différences de comportement qui s'observent au niveau local. Concernant l'*Accuracy*, les cinq modèles comparés ont les mêmes performances. Ce résultat vient du déséquilibre dans la distribution des données de comptage, avec une prédominance des zéros, et donc des prédictions concentrées essentiellement autour de cette valeur. Une vision en valeurs relatives par rapport au modèle de référence n'est pas proposée car les écarts de performance sont peu significatifs.

Métrique	**RMSE**		**MAE**		**Accuracy**	
Algorithme	Train	Test	Train	Test	Train	Test
GLM	0,1380	0,1382	0,0364	0,0364	0,9818	0,9817
Spline	0,1379	0,1381	0,0364	0,0364	0,9818	0,9817
EBM	0,1378	0,1380	0,0364	0,0364	0,9818	0,9817
CART	0,1381	0,1382	0,0365	0,0365	0,9818	0,9817
XGB	0,1378	**0,1380**	0,0363	**0,0364**	0,9818	**0,9817**

TABLE 4.4 – Résultats de comparaison des modèles pour la modélisation de la fréquence : GLM, Spline cubique, EBM, CART et XGB

Dans le Chapitre 3, nous avons vu les différences conceptuelles entre les modèles additifs généralisés combinant des composantes univariées des modèles de *Machine Learning* et des fonctions splines. Nous avons notamment souligné l'apprentissage et le lissage de l'information (Section 3.3). Au niveau des métriques de performance, il n'y a pas d'écart significatif entre les deux modélisations (Tables 4.3 et 4.4). Cependant, si on s'intéresse au coût computationnel, on soulève une différence considérable. En effet, la Table 4.5 montre que pour la modélisation de la fréquence, l'algorithme *Explainable Boosting Machine* avec les paramètres optimisés est 7 fois plus rapide que la configuration avec les fonctions splines. Dans le secteur assurantiel, les portefeuilles doivent être régulièrement pilotés. D'un point de

vue technique et opérationnel, le temps computationnel est donc un paramètre important à prendre en compte.

Pred / Algo	**Coût moyen**	**Fréquence**
Spline	2,60	204
EBM défaut	0,06	1,19
EBM param	2,0	29,4

TABLE 4.5 – Temps computationnel en minutes des modèles additifs généralisés

4.3.2 Modèle additif et modèle multiplicatif

Les contraintes de la tarification en assurance non-vie imposent de produire des modèles avec une structure multiplicative, cette dernière permettant l'obtention d'une grille tarifaire immédiatement interprétable et facilement pilotable. Dans la Section 3.3.1, nous avons évoqué les lacunes de l'implémentation actuelle de l'algorithme *Explainable Boosting Machine* et proposé plusieurs pistes pour y remédier. En effet, dans le package `InterpretML` (INTERPRETML TEAM (2021)), la possibilité d'utiliser uniquement la fonction de lien identité avec une distribution gaussienne donne lieu à un modèle additif. Pour retrouver une structure multiplicative, nous avons ainsi tâché de décliner le méta-algorithme *Cyclic Boosting Machine* utilisé pour l'apprentissage des fonctions composantes dans le modèle additif généralisé. Nous nous sommes également intéressés à la prédiction du logarithme de la variable cible. Les résultats de cette approche sont résumés dans la Table 4.6 et illustrés sur la Figure 4.10. Nous remarquons alors que les deux méthodologies proposées adressent efficacement le problème de prédiction.

Métrique	**RMSE**			**MAE**			**R2**	
Algorithme	Train	Test	%	Train	Test	%	Train	Test
GLM	1695	1674	100	1203	1192	100	0,067	0,069
Spline	1688	1664	99,4	1197	1183	99,3	0,075	0,080
EBM	1686	1665	99,5	1197	1185	99,5	0,077	0,079
EBM log	1693	1672	99,8	1192	**1179**	98,9	0,069	0,071
CBM	1690	**1664**	99,4	1197	1183	99,3	0,073	**0,080**

TABLE 4.6 – Résultats de comparaison des modèles pour la modélisation du coût moyen sur un échantillon réduit : GLM, EBM, EBM log et CBM

La prédiction du logarithme du coût moyen, avec une correction subséquente du biais de prédiction (Section 3.3.1), enregistre les meilleures performances au niveau de l'erreur absolue moyenne. Pour l'erreur quadratique moyenne, c'est le méta-algorithme *Cyclic Boosting Machine* qui donne les meilleurs résultats*. Cette métrique a une tendance à pénaliser les erreurs importantes (Section 2.3.5). Rappelons que le modèle *Cyclic Boosting Machine* émule intrinsèquement la distribution sous-jacente des données, avec une procédure adéquate de l'apprentissage des fonctions composantes (Algorithme 2). Graphiquement, nous remarquons qu'il a un comportement légèrement différent que les déclinaisons l'algorithme *Explainable Boosting Machine* (Figure 4.10). Dans les deux cas, les approches correctives envisagées dépassent les performances de l'algorithme *Explainable Boosting Machine* dans sa version additive, et par-delà les performances du modèle linéaire généralisé.

∗. La différence avec le modèle additif généralisé combinant des fonctions splines se fait au niveau de la première décimale

Rappelons qu'à l'heure actuelle, une étude similaire ne peut pas être envisagée pour la modélisation de la fréquence, l'algorithme *Cyclic Boosting Machine* étant peu robuste et les données de comptage trop déséquilibrées. L'applicabilité opérationnelle d'une telle approche s'en trouve limitée. Les premiers résultats semblent toutefois encourageants. L'analyse graphique et les métriques de performance montrent qu'avec les bonnes fonctions de lien et distributions, l'algorithme *Explainable Boosting Machine* a le potentiel de s'approcher encore plus des performances des modèles de *Machine Learning*, tout en restant intrinsèquement interprétable.

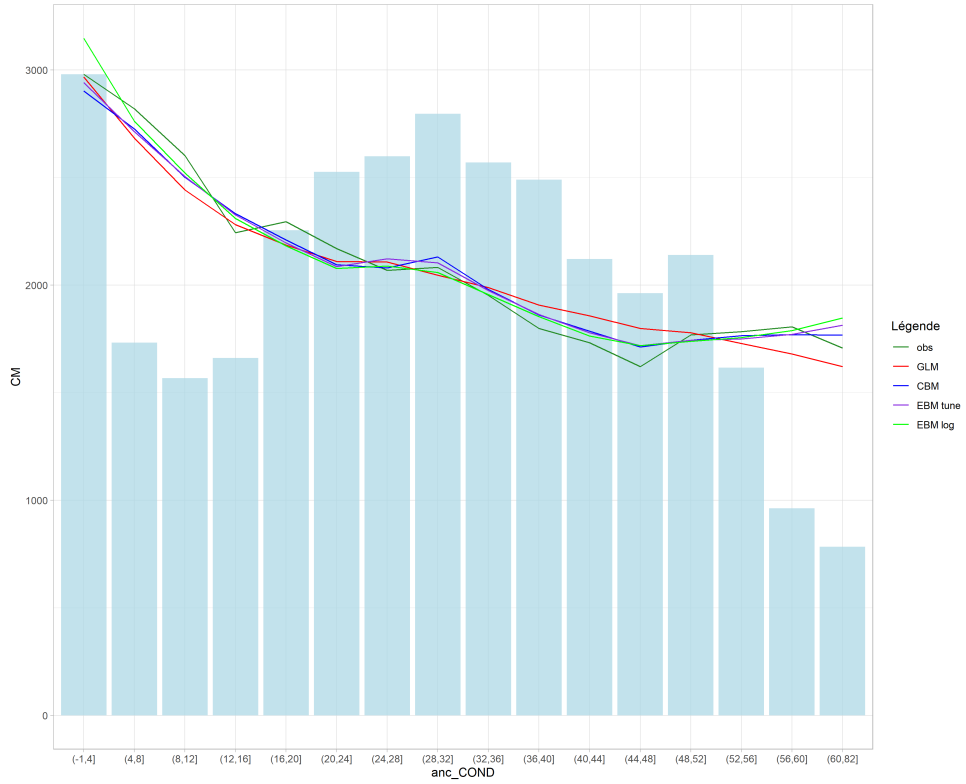


FIGURE 4.10 – Valeurs observées et prédites pour la modélisation du coût moyen : comparaison des modèles GLM, CBM et EBM sur la variable `anc_COND` ou son logarithme

4.3.3 Offset et Machine Learning

Pour la modélisation de la fréquence, il est important de tenir compte de l'exposition de chaque observation en appliquant un ajustement au *prorata temporis* dans la prédiction du nombre moyen de sinsitres (Section 2.2.2). Si l'intégration d'une variable *offset* est aisée dans les modèles linéaires et dans les modèles additifs généralisés, l'adaptation au cadre non-paramétrique des modèles de *Machine Learning* n'est pas évidente. Dans l'algorithme *Explainable Boosting Machine*, il n'est pas possible, à l'heure actuelle, de considérer une variable *offset* dans la modélisation. Nous avons donc opté pour la conservation de la variable exposition parmi les variables explicatives en s'assurant que sa contribution est croissante (Figure C.4). Une autre approche, non-équivalente, consiste à pondérer les observations par l'exposition associée. De même, cette option n'est pas disponible dans l'implémentation actuelle de l'algorithme *Explainable Boosting Machine*. Finalement, une approche alternative consisterait à prédire directement la fréquence des sinistres au lieu de prédire le nombre et de considérer l'exposition en *offset*.

4.4 Prime pure et mouvements tarifaires

Dans ce mémoire, nous avons cherché à construire un modèle de *Machine Learning* intrinsèquement interprétable. Les Sections 4.2 et 2.4 ont permis d'analyser quantitativement et qualitativement les modèles du coût moyen et de la fréquence. La modélisation et l'analyse de la prime pure constituent l'aboutissement naturel de ce travail. Dans cette section, nous comparerons les primes pures prédites par le modèle linéaire généralisé et le modèle additif généralisé ayant une structure arborescente, à savoir l'algorithme *Explainable Boosting Machine*. Nous proposerons tout d'abord une analyse univariée des primes prédites et du *Burning Cost* (Section 4.4.1). Dans un deuxième temps, nous chercherons à identifier les profils où le comportement des deux modèles de prime pure diffère (Section 4.4.2). Nous proposerons ainsi une segmentation des écarts de prédiction en vue d'une optimisation tarifaire.

4.4.1 Analyse du Burning Cost et des primes prédites

Dans notre approche, le *Burning Cost* est défini comme le quotient de la somme des coûts observés par la somme des expositions. L'analyse du *Burning Cost* et des primes prédites sur un échantillon de test permet alors de comparer le risque observé au risque estimé par les modèles. La Figure 4.11 montre le résultat pour la variable âge conducteur. Comme constaté sur les graphes des valeurs observées et prédites pour la modélisation du coût moyen et de la fréquence, le modèle linéaire généralisé moyennise le comportement étudié. Le modèle additif généralisé avec une structure arborescente parvient mieux à capter la volatilité des données. Nous observons en particulier un meilleur ajustement sur l'intervalle allant de 50 à 70 ans, où le modèle linéaire surestime le risque. Ce résultat se confirme sur la variable ancienneté du permis (Figure C.5). Sur l'intervalle allant de 30 à 45 ans, le modèle additif généralisé arrive à reproduire la baisse du risque observé tandis que le modèle linéaire conserve le niveau de risque correspondant à 25 ans d'ancienneté. Nous observons également un écart de prédiction pour les conducteurs peu expérimentés avec une ancienneté de permis variant entre 4 et 16 ans. Notons également que sur les profils peu exposés, notamment sur les grands âges, le modèle additif est plus fidèle à la réalité que le modèle linéaire.

4.4.2 Surtarification et sous-tarification

Les graphes des valeurs observées et prédites (Figures 4.11 et C.5) montrent des différences de comportement entre le modèle linéaire généralisé et le modèle additif ayant une structure arborescente. Pour visualiser et quantifier les écarts entre la prime pure prédite par le modèle linéaire généralisé (PP GLM) et celle prédite par le modèle additif (PP EBM), nous nous intéressons alors à la variable δ définie par

$$\delta = \frac{\text{PP EBM}}{\text{PP GLM}}.$$

La variable δ s'interprète de la manière suivante :

- si $\delta > 1$; le modèle EBM surtarifie par rapport au modèle GLM ;
- si $\delta < 1$; le modèle EBM sous-tarifie par rapport au modèle GLM ;
- si $\delta = 1$; les modèles EBM et GLM proposent le même niveau de tarification.

Dans un premier temps, nous nous intéressons à l'allure moyenne de la variable δ en fonction de quelques variables significatives. La Figure 4.12a (respectivement 4.12b) montre les résultats pour la variable âge conducteur (respectivement ancienneté du permis). La ligne rouge permet d'identifier le même

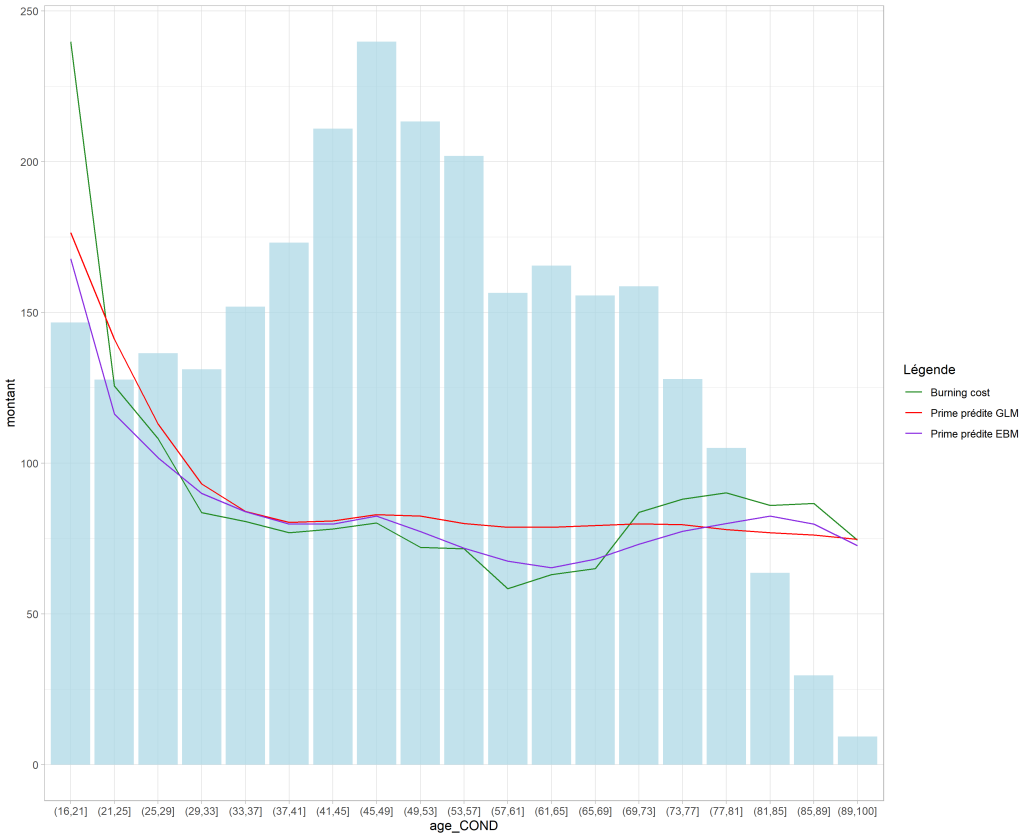
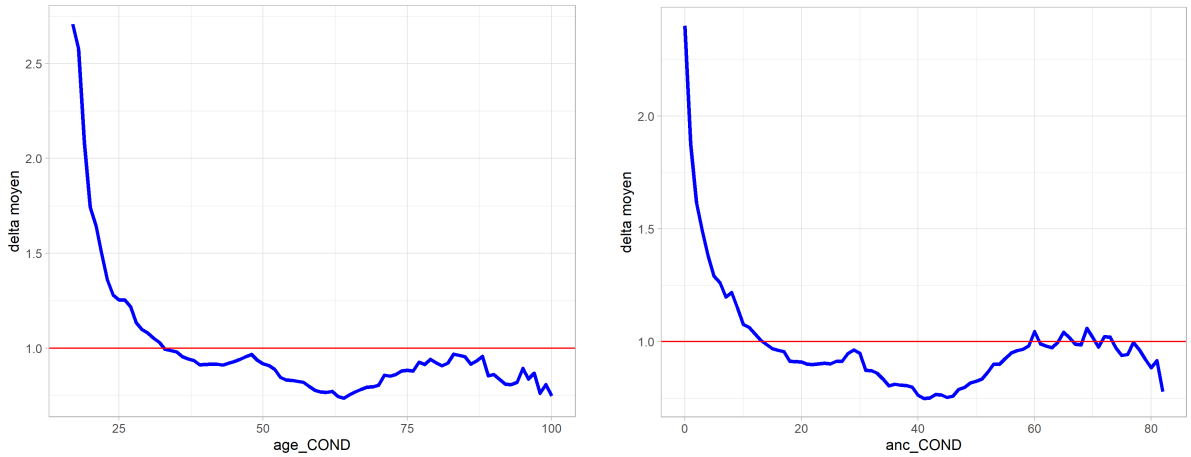


FIGURE 4.11 – *Burning Cost* et les primes prédites par les modèles GLM et EBM pour la variable `age_COND`

niveau de tarification par les deux modèles, à savoir $\delta = 1$. Pour les deux variables explicatives, nous observons alors que le modèle EBM surtarifie par rapport au modèle GLM pour les jeunes conducteurs (jusqu'à 30 ans environ) et les conducteurs inexpérimentés (jusqu'à 12 ans d'ancienneté environ). En revanche, pour les conducteurs confirmés (35 à 60 ans environ, 15 à 40 ans d'ancienneté), le modèle EBM sous-tarifie par rapport au modèle GLM. Cette tendance pouvait déjà s'observer sur le graphe des valeurs observées et prédites (Figures 4.11 et C.5). L'évolution est de nouveau croissante pour les conducteurs âgés, mais reste souvent sous la barre de $\delta = 1$. Le pic qui s'observe vers 45 ans ou vers 30 ans d'ancienneté peut être rapproché des analyses faites sur les fonctions composantes du modèle additif généralisé associées à la variable âge conducteur (Figure 3.13) et ancienneté du permis (Figure 4.3). En effet, la présence d'un conducteur secondaire novice, associée souvent au moment ou les parents prêtent leur voiture aux enfants, est ponctuellement traduite par le modèle EBM par une augmentation du risque, une tendance que le modèle GLM moyennise. Les résultats pour la variable donnant la classe de prix du véhicule sont disponible en Annexes (Figure C.6). Nous remarquons alors une surtarification importante du modèle EBM par rapport au modèle GLM pour les classes de prix très élevées.

Dans un deuxième temps, nous visualisons la distribution des différents profils en fonction de la variable δ. La Figure 4.13 montre l'exemple pour la variable ancienneté du permis. Nous retrouvons alors les résultats observés sur les Figures C.5 et 4.12b. Le modèle EBM sous-tarifie par rapport au modèle GLM sur l'intervalle allant de 30 à 60 ans environ. En revanche, pour des anciennetés de permis faibles, il propose des tarifs plus élevés. Pour la variable donnant la classe de prix du véhicule, une surtarification s'observe sur les classes de prix comprises dans l'intervalle de 20 à 30 (Figure C.7).



(a) `age_COND` (b) `anc_COND`

FIGURE 4.12 – Évolution de δ en fonction de variables significatives

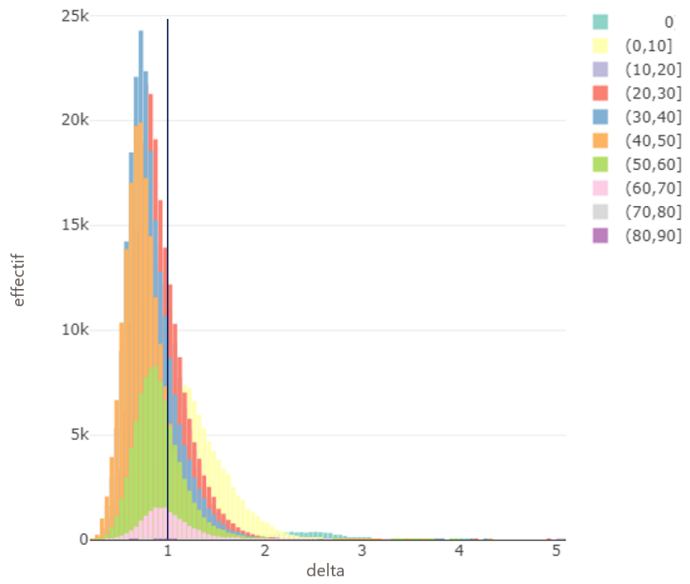


FIGURE 4.13 – Distribution de la variable `anc_COND` en fonction de δ

Finalement, dans une approche explicative plutôt que descriptive, nous pouvons segmenter les écarts de prévision à l'aide d'un arbre de décision CART réalisant un partitionnement de l'espace (Section 2.3.1). Les variables donnant les plus grands écarts supérieurs à 1 sont alors le coefficient de réduction-majoration, la classe de prix de véhicule et l'ancienneté du permis. Plus précisément, la structure arborescente indique que les conducteurs avec un coefficient de réduction-majoration supérieur à la valeur de référence 0,5 ($\texttt{cof\_CRM} \geq 0,61$) et une ancienneté de permis faible ($\texttt{anc\_COND} \geq 1,5$ mais $\texttt{anc\_COND} < 4,5$) auront une prime pure plus élevée avec le modèle EBM. Ce résultat corrobore les analyses précédentes : le modèle additif généralisé a une compréhension du risque différente pour les jeunes conducteurs en particulier.

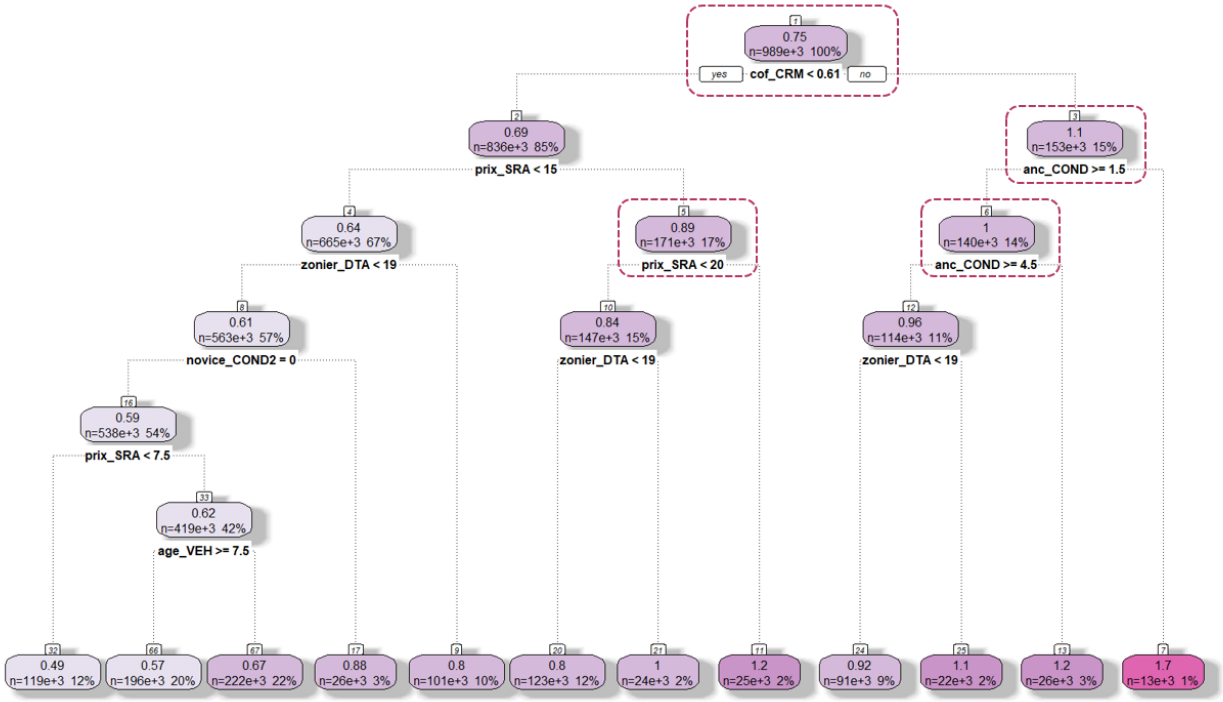


FIGURE 4.14 – Arbre de décision pour segmenter la variable δ

Si nous avons commencé à analyser des profils intéressants en vue d'une optimisation tarifaire à l'aide de l'indicateur δ, une limite importante est à soulever dans cette approche. En effet, la variable δ permet d'identifier les zones où le modèle additif généralisé avec une structure arborescente surestime ou sous-estime le risque par rapport au modèle de référence, le modèle linéaire généralisé. Cependant, il est important de savoir si cette surtarification ou sous-tarification s'effectue à raison. Plus précisément, il faudrait envisager des indicateurs qui prennent en compte le risque observé, notamment

$$\delta_{GLM} = \frac{\text{Burning Cost}}{\text{PP GLM}}$$

et

$$\delta_{EBM} = \frac{\text{Burning Cost}}{\text{PP EBM}}.$$

Dans une approche de type frontière de décision avec les machines à vecteur de support (*Support-vector machine*, UKIL (2007)), il serait également intéressant de créer un indicateur qui combine le *Burning Cost*, la prime pure prédite par le modèle linéaire généralisé et celle par le modèle additif généralisé. Par ailleurs, le modèle additif généralisé combinant des fonctions splines et celui ayant une structure arborescente étant comparables en termes de performances, il serait intéressant d'appliquer cette approche pour analyser leur comportement sur des tranches d'âge ou sur des zones particulières.

Conclusion

Dans ce mémoire, nous avons cherché à construire des modèles de *Machine Learning* intrinsèquement interprétables pour la tarification en assurance non-vie. Afin de proposer une vision plus inclusive et collaborative entre les modèles linéaires généralisés et les modèles de *Machine Learning*, nous nous sommes alors intéressés aux modèles additifs généralisés. En conservant la structure additive intrinsèquement interprétable, nous avons opéré un ajout parcimonieux de la non-linéarité dans le modèle au travers des fonctions composantes. Dépassant le cadre traditionnel donné par les fonctions splines, nous avons combiné des composantes univariées des modèles de *Machine Learning* dans un modèle additif généralisé.

Si l'approche par fonctions splines possède déjà de nombreuses applications dans la tarification en assurance non-vie, l'utilisation des modèles de *Machine Learning* comme fonctions composantes dans un modèle additif généralisé constitue un élément innovant que l'on souhaite transmettre dans ce mémoire. En effet, au-delà de l'introduction d'une nouvelle méthode de modélisation, les modèles *Glass Box* déterminent une nouvelle mécanique d'apprentissage, potentiellement applicable à d'autres architectures et dans d'autres contextes. La concession au niveau de la structure pour un apprentissage cyclique des fonctions composantes, couplée éventuellement à une procédure de *Boosting* et de *Bagging*, définit une nouvelle classe des modèles, des modèles de *Machine Learning* intrinsèquement interprétables.

Appliqués à la tarification de la garantie DTA (dommage tout accident) en assurance automobile, les modèles additifs généralisés ayant une structure arborescente confirment leur position d'intermédiaire entre les modèles linéaires généralisés et les modèles de *Machine Learning*. En termes de performance, ils sont comparables aux modèles additifs généralisés combinant des fonctions splines, mais pour un temps computationnel considérablement inférieur et un potentiel afférent à la structure semi-paramétrique restant à explorer. En effet, la limite principale de la modélisation additive généralisée avec une structure arborescente réside dans la non robustesse de l'algorithmie existante, ne permettant pas l'utilisation des fonctions de lien et des distributions adaptées à la structure des données assurantielles et, *in fine*, la construction d'une grille tarifaire multiplicative. Néanmoins, les solutions envisagées dans ce mémoire comblent en partie ces lacunes et libèrent le potentiel du modèle. Dans de futurs travaux, il serait intéressant de se pencher sur l'ajout des interactions dans le modèle additif généralisé. L'approche par composantes univariées des réseaux de neurones constitue une autre piste à investiguer plus en détail par la suite, tout comme l'optimisation de l'ordre dans lequel les variables explicatives sont apprises dans un modèle additif généralisé de type *Machine Learning*. Les travaux de recherche et développement en cours portent également sur la définition des métriques de performance plus adaptées ainsi que des indicateurs ayant le potentiel de qualifier et quantifier efficacement les différences de comportement entre les modèles. Enfin, à l'aune de ces éléments, l'extension de l'étude à d'autres garanties plus volatiles, notamment les garanties vol et responsabilité civile, semble pertinente.

En définitive, la démarche par composantes univariées des modèles de *Machine Learning* ne vient pas concurrencer l'approche traditionnelle de tarification. Elle offre cependant une alternative et un nouvel

éclairage sur le comportement sous-jacent de la sinistralité qu'un modèle linéaire généralisé moyennise et un modèle combinant des fonctions splines lisse.

Bibliographie

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R. et Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems* 34, p. 4699-4711.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control* 19.6, p. 716-723.

Autorité de Contrôle Prudentiel et de Résolution (2021). Chiffres du marché français de la banque et de l'assurance 2020. Rapport technique.

Barry, L. et Charpentier, A. (2022). L'équité de l'apprentissage machine en assurance.

Besse, P. (1992). PCA stability and choice of dimensionality. *Statistics & Probability Letters* 13.5, p. 405-410.

Besse, P., Castets-Renard, C., Garivier, A. et Loubes, J.-M. (oct. 2018). L'IA du Quotidien peut elle être Éthique ? working paper. URL : https://hal.archives-ouvertes.fr/hal-01886699.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. et Jones, Z. M. (2016). mlr: Machine Learning in R. *Journal of Machine Learning Research* 17.170, p. 1-5.

Breiman, L. (1996). Bagging predictors. *Machine learning* 24.2, p. 123-140.

Breiman, L., Friedman, J. H., Olshen, R. A. et Stone, C. J. (2017). Classification and regression trees. Routledge.

Bucci, S. (2021). Étude et implémentation de techniques d'analyse de sensibilité dans les modèles de tarification non-vie. Application à la tarification à l'adresse. Mémoire d'actuariat. Paris : ENSAE.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. et Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, p. 1721-1730.

Chen, T. et Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, p. 785-794.

Commission Européenne (2020). Livre blanc sur l'intelligence artificielle : une approche européenne d'excellence et de confiance. Rapport technique.

De Haan, L., Ferreira, A. et Ferreira, A. (2006). Extreme value theory: an introduction. T. 21. Springer.

Delcaillau, D., Ly, A., Vermet, F. et Papp, A. (2020). Interprétabilité des modèles : état des lieux des méthodes et application à l'assurance. *arXiv preprint arXiv:2007.12919*.

Denuit, M. et Charpentier, A. (2005). Mathematiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement.

Dubessy, F. (2021). Assurance automobile : miser sur la prévention. *L'Argus de l'assurance* 7750, p. 38-40.

Dutang, C. et Guibert, Q. (2022). GLM et tarification en assurance non-vie : nouvelles approches à partir d'arbres et de forêts GLM. *L'Actuariel* 44, p. 60-62.

Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H. et Steinrötter, B. (2021). The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). *J* 4.4, p. 589-603.

Financial Stability Board (2017). Artificial intelligence and machine learning in financial services - market developments and financial stability implication. Rapport technique.

France Assureurs (2021). Rapport annuel 2020. Rapport technique.

Freund, Y., Schapire, R. E. et al. (1996). Experiments with a new boosting algorithm. *icml*. T. 96. Citeseer, p. 148-156.

Friedman, J., Hastie, T. et Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33.1, p. 1-22.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, p. 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis* 38.4, p. 367-378.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. et Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, p. 80-89.

Google Research Team (2022). Neural Aditive Models: Interpretable Machine Learning with Neural Nets. GitHub Repository. url : https://neural-additive-models.github.io/ (visité le 02/09/2022).

Guillot, A. (2015). Apprentissage statistique en tarification non-vie : quel avantage opérationnel ? Mémoire d'actuariat. Paris : ENSAE.

Hastie, T. et Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association* 82.398, p. 371-386.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks* 4.2, p. 251-257.

Hu, Y. et Scarrott, C. (2018). evmix: An R package for Extreme Value Mixture Modeling, Threshold Estimation and Boundary Corrected Kernel Density Estimation. *Journal of Statistical Software* 84.5, p. 1-27.

InterpretML Team (2021). InterpretML. Jupyter Book. url : https://interpret.ml/docs/intro.html (visité le 25/08/2022).

InterpretML Team (2022). InterpretML GitHub. GitHub Repository. url : https://github.com/interpretml (visité le 01/09/2022).

Kassambara, A. (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. url : https://CRAN.R-project.org/package=ggpubr.

Kuhn, M. (2022). caret: Classification and Regression Training. R package version 6.0-91. url : https://CRAN.R-project.org/package=caret.

Kuhn, M., Jackson, S. et Cimentada, J. (2020). corrr: Correlations in R. R package version 0.4.3. url : https://CRAN.R-project.org/package=corrr.

Larsen, K. (2015). GAM: the predictive modeling silver bullet. *Multithreaded. Stitch Fix* 30, p. 1-27.

Laureille, X., Quennelle, P. et Ittah, M. (2019). Détermination des seuils de sinsitres graves et optimisation de la réassurance. *Conférence 100% actuaires, 100% data science*.

Laurent-Bonneau, B. (2020). Réseaux de neurones. Support de cours. INSA Toulouse.

Lê, S., Josse, J. et Husson, F. (2008). FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software* 25.1, p. 1-18.

Lou, Y., Caruana, R. et Gehrke, J. (2012). Intelligible models for classification and regression. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 150-158.

Lou, Y., Caruana, R., Gehrke, J. et Hooker, G. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 623-631.

Lundberg, S. M. et Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.

Marx, B. D. et Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis* 28.2, p. 193-209.

McCullagh, P. et Nelder, J. (1989). Generalized linear models (Vol. 37) CRC press.

McCullagh, P. et Nelder, J. A. (2019). Generalized linear models. Routledge.

Microsoft (2022). Cyclic Boosting Machines. GitHub Repository. url : https://github.com/microsoft/CBM (visité le 01/09/2022).

Mishra, C. et Gupta, D. (2017). Deep machine learning and neural networks: An overview. *IAES International Journal of Artificial Intelligence* 6.2, p. 66.

Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Molnar, C., Bischl, B. et Casalicchio, G. (2018). iml: An R package for Interpretable Machine Learning. *JOSS* 3.26, p. 786.

Nelder, J. A. et Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* 135.3, p. 370-384.

Nori, H., Jenkins, S., Koch, P. et Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Ossberger, J. (2020). tea: Threshold Estimation Approaches. R package version 1.1. url : https://CRAN.R-project.org/package=tea.

Parlement Européen et Conseil de l'Union Européenne (2016). Règlement (UE) 2016/679 du Parlement Européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données). OJ L. 119/1.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12.Oct, p. 2825-2830.

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. url : https://www.R-project.org/.

Ribatet, M. et Dutang, C. (2022). POT: Generalized Pareto Distribution and Peaks Over Threshold. R package version 1.1-10. url : https://CRAN.R-project.org/package=POT.

Ribeiro, M. T., Singh, S. et Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, p. 1135-1144.

Santenar, I., Majkowski, E., Manchester, P. et Bong, A. S.-Y. (2022). 2022 Global Insurance Outlook: growing with people, purpose and tech. Rapport technique. EY.

Savoye, S. (2021). Développement et mise en application d'un algorithme de tarification sur un modèle coût/fréquence en confrontation avec le GLM. Mémoire d'actuariat. Paris : Université Paris-Dauphine.

Schelldorfer, J. et Wuthrich, M. V. (2019). Nesting classical actuarial models into neural networks. *Available at SSRN 3320525*.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, p. 461-464.

Suru, A. (2020). Assurance IARD. Economica.

Svetunkov, I. (2022). greybox: Toolbox for Model Building and Forecasting. R package version 1.0.5. url : https://CRAN.R-project.org/package=greybox.

Therneau, T. et Atkinson, B. (2022). rpart: Recursive Partitioning and Regression Trees. R package version 4.1.16. url : https://CRAN.R-project.org/package=rpart.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, p. 267-288.

Ukil, A. (2007). Support vector machine. *Intelligent Systems and Signal Processing in Power Engineering*. Springer, p. 161-226.

Van Rossum, G. et Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA : CreateSpace.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks* 10.5, p. 988-999.

Wahba, G. (1990). Spline models for observational data. Society for Industrial et Applied Mathematics.

Wick, F., Kerzel, U. et Feindt, M. (2019). Cyclic boosting-an explainable supervised machine learning algorithm. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, p. 358-363.

Winter, E. (2002). The shapley value. *Handbook of game theory with economic applications* 3, p. 2025-2054.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73.1, p. 3-36.

Wood, S. N. (2006). Generalized additive models: an introduction with R. chapman et hall/CRC.

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

Annexe A

Notions d'assurance IARD et présentation des données

A.1 La tarification en IARD

Les hypothèses standard en tarification IARD sont les suivantes :

(H_1) : $B_k \overset{\text{iid}}{\sim} B$, *ie* les montants des sinistres ont le même comportement aléatoire ;
(H_2) : $\forall\, k = 1, ..., N$, $B_k \perp N$, *ie* la fréquence est indépendante de la sévérité.

Sous (H_1) et (H_2), à l'aide de l'espérance conditionnelle, une décomposition coût-fréquence de la sinistralité peut ainsi être obtenue :

$$\begin{aligned}\mathbb{E}[S] &= \mathbb{E}\left[\sum_{k=1}^{N} B_k\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{N} B_k | N\right]\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{N} \mathbb{E}[B_k|N]\right] = \mathbb{E}\left[\sum_{k=1}^{N} \mathbb{E}[B_k]\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{N} \mathbb{E}[B]\right] = \mathbb{E}[N]\,\mathbb{E}[B].\end{aligned} \tag{A.1}$$

A.2 Présentation de la base de données et analyse exploratoire

La Figure A.1 donne la matrice des corrélations complète des variables explicatives du modèle. Elle permet en particulier de visualiser une zone de forte corrélation entre les variables du véhiculier SRA, puis une forte corrélation entre les variables définissant les règles de comptage de sinsitres.

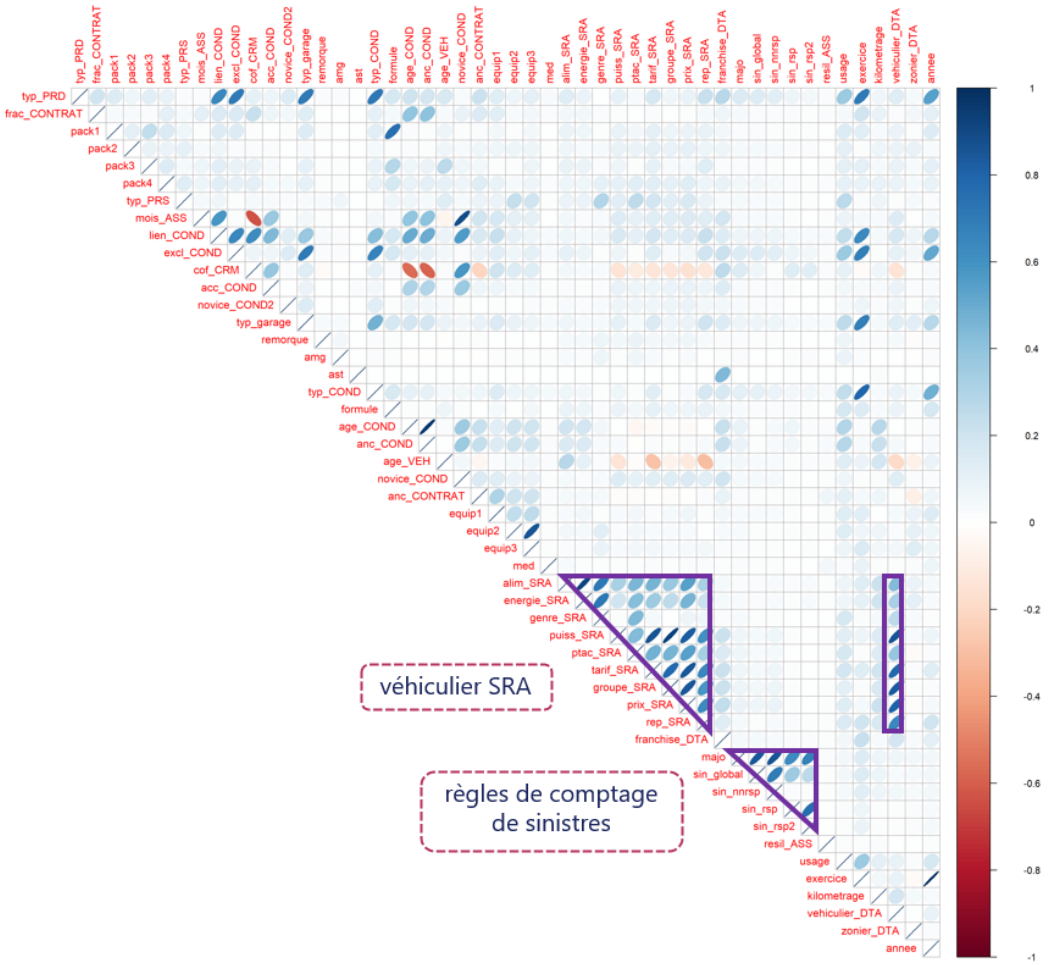


FIGURE A.1 – Matrice des corrélations complète

Annexe B

Présentation du modèle additif généralisé

B.1 GAM Machine Learning

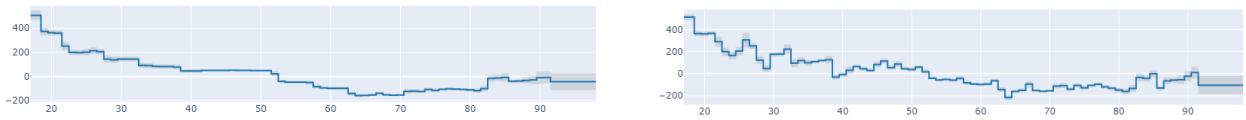
B.1.1 Algorithme EBM dans le package `InterpretML`

Dans le package `InterpretML`, l'implémentation de l'algorithme *Explainable Boosting Machine* s'appuie sur les paramètres suivants :

- `max_bins` = 256 : paramètre de *Binning*, donne le nombre maximal de *bins* pour chaque variable explicative X_j ;
- `binning` = `quantile` : paramètre de *Binning*, décrit le type de groupement des données ;
- `outer_bags` = 8 : paramètre de *Bagging* global, détermine la taille de l'échantillon tiré aléatoirement sur lequel la procédure décrite dans l'Algorithme 1 sera répétée. Cet accroissement de la variabilité donne une indication de la sensibilité de l'algorithme à une légère perturbation dans l'échantillon d'apprentissage et permet de produire des intervalles de confiance en s'appuyant sur le calcul de l'écart-type. La valeur par défaut 8 signifie que 80% de l'échantillon d'apprentissage est pris en compte ;
- `inner_bags` = 0 : paramètre de *Boosting*, permet d'ajouter de la variabilité dans la procédure en tirant des échantillons aléatoires sur lesquels les fonctions composantes seront apprises lors d'une itération. Cette étape permet d'accroître les performances prédictives mais est très coûteuse algorithmiquement. La valeur par défaut 0 signifie que cette option est désactivée ;
- `learning_rate` = 0,01 : paramètre de *Boosting*, définit le taux d'apprentissage (Section 2.3.2) ;
- `early_stopping_rounds` = 50 : paramètre qui permet d'enclencher l'arrêt précoce de l'algorithme dès que la fonction de perte commence à augmenter. Plus précisément, il donne le nombre maximal de cycles où une dégradation peut être observée avant que l'algorithme ne s'arrête définitivement ;
- `early_stopping_tolerance` = 0,0001 : deuxième paramètre qui permet d'enclencher l'arrêt précoce de l'algorithme, détermine le seuil au-delà duquel on considère que l'apprentissage de l'algorithme s'améliore ;
- `max_rounds` = 5 000 : paramètre de *Bagging* local, fixe le nombre maximal d'itérations T ;
- `min_samples_leaf` = 2 : paramètre relatif à la construction des arbres de décision (Section 2.3.1), donne le nombre minimal d'observations par feuille, permet de contrôler la complexité du modèle ;
- `max_leaves` = 3 : deuxième paramètre relatif à la construction des arbres de décision, donne le nombre maximal de feuilles dans un arbre, permet de contrôler la complexité du modèle.
- `n_jobs` : paramètre algorithmique, fixe nombre de tâches à exécuter en parallèle ;

- `random_state` : paramètre algorithmique, initialise la graine.

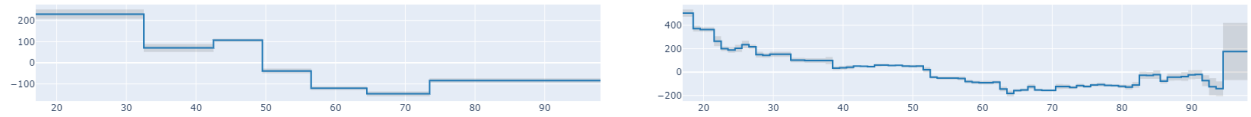
Si un surajustement s'observe, une diminution des paramètres `max_bins` et `early_stopping_rounds` en parallèle d'une augmentation du paramètre `early_stopping_tolerance` est conseillée. Si le modèle sous-ajuste, il s'agira de mettre en place la procédure inverse. La complexité de l'arbre de décision peut également être pilotée. Une augmentation du paramètre `max_leaves`, et donc mécaniquement de la profondeur de l'arbre, a pour effet d'amplifier les variations observées dans l'allure des fonctions composantes, donnant lieu à un modèle plus volatil. La Figure B.1 montre qu'en multipliant par cinq le paramètre `max_leaves`, toutes choses égales par ailleurs, les variations dans chaque *bin* au niveau des fonctions composantes sont beaucoup plus importantes et les intervalles de confiance plus larges (Figure B.1b).



(a) `max_leaves` = 3 (b) `max_leaves` = 15

FIGURE B.1 – Variations observées dans la fonction composante associée à la variable `age_cond` dans le modèle GAM avec une structure arborescente pour la modélisation du coût moyen suite au changement du paramètre `max_leaves`

L'aspect en escalier des fonctions composantes est fortement lié à la finesse de la procédure de *Binning*. Pour `max_bins` = 8 (Figure B.2a), nous obtenons un partitionnement en huit escaliers de la fonction composante associée à la variable âge conducteur. Pour la prédiction du coût moyen, le résultat est alors aisément interprétable : le coût moyen est élevé sur la tranche de 18 à 32 ans, il diminue sur la tranche de 33 à 41 ans, puis raugmente sur la tranche de 42 à 50 ans. Il a une tendance décroissante jusqu'à 74 ans environ et augmente à nouveau sur les âges plus élevés. Pour `max_bins` = 1 024 (Figure B.2b), nous obtenons un partitionnement extrêmement précis pour chaque tranche d'âge. Un arbitrage est alors à faire entre la précision du modèle et son applicabilité actuarielle. Pour favoriser la robustesse et l'interprétabilité, les paramètres `max_leaves` et `max_bins` doivent être limités. En revanche, si l'on souhaite se focaliser sur un segment en particulier et étudier une dérive de la sinistralité, une revue à la hausse de ces paramètres peut être envisagée.



(a) `max_bins` = 8 (b) `max_bins` = 1 024

FIGURE B.2 – Variations observées dans la fonction composante associée à la variable `age_cond` dans le modèle GAM avec une structure arborescente pour la modélisation du coût moyen suite au changement du paramètre `max_bins`

B.1.2 Modèle additif et modèle multiplicatif

Algorithm 2 Un cycle complet de l'algorithme **CBM**

Initialisation
$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$ ▷ Estimation de la moyenne globale sur l'échantillon d'apprentissage
for $k = 1, \ldots, K$ **do**
  $f_1^k, \ldots, f_p^k = 1$ ▷ Première estimation des fonctions composantes
end for
Itération 1 : $t = 1$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 1
  $\hat{y}_{k,\tau} \leftarrow \beta_0 \times \underbrace{1 \times \ldots \times 1}_{p \text{ fois}}$ ▷ Estimation courante de y_i
  $g_{1,1}^k \leftarrow \frac{\sum_{x_{i,1} \in b_1^k} y_i}{\sum_{x_{i,1} \in b_1^k} \hat{y}_{k,\tau}}$
  $f_{1,1}^k \leftarrow g_{1,1}^k$ ▷ Première mise à jour de la première fonction composante
end for
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 2
  $\hat{y}_{k,\tau} \leftarrow \beta_0 \times f_{1,1}^k \times \underbrace{1 \ldots \times 1}_{p-1 \text{ fois}}$
  $g_{2,1}^k \leftarrow \frac{\sum_{x_{i,2} \in b_2^k} y_i}{\sum_{x_{i,2} \in b_2^k} \hat{y}_{k,\tau}}$
  $f_{2,1}^k \leftarrow g_{2,1}^k$
end for
... ▷ Calcul des fonctions composantes pour les prédicteurs 3 à $p-1$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur p
  $\hat{y}_{k,\tau} \leftarrow \beta_0 \times f_{1,1}^k \times f_{2,1}^k \times \ldots \times f_{p-1,1}^k \times 1$
  $g_{p,1}^k \leftarrow \frac{\sum_{x_{i,p} \in b_p^k} y_i}{\sum_{x_{i,p} \in b_p^k} \hat{y}_{k,\tau}}$
  $f_{p,1}^k \leftarrow g_{p,1}^k$
end for ▷ Fin de l'itération 1
Itération 2 : $t = 2$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 1
  $\hat{y}_{k,\tau} \leftarrow \beta_0 \times f_{1,1}^k \times \ldots \times f_{p,1}^k$
  $g_{1,2}^k \leftarrow \frac{\sum_{x_{i,1} \in b_1^k} y_i}{\sum_{x_{i,1} \in b_1^k} \hat{y}_{k,\tau}}$
  $f_{1,2}^k \leftarrow g_{1,1}^k \times g_{1,2}^k$ ▷ Deuxième mise à jour de la première fonction composante
end for
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 2
  $\hat{y}_{k,\tau} \leftarrow \beta_0 \times f_{1,2}^k \times f_{2,1}^k \times \ldots \times f_{p,1}^k$
  $g_{2,2}^k \leftarrow \frac{\sum_{x_{i,2} \in b_2^k} y_i}{\sum_{x_{i,2} \in b_2^k} \hat{y}_{k,\tau}}$
  $f_{2,2}^k \leftarrow g_{2,1}^k \times g_{2,2}^k$
end for
... ▷ Calcul des fonctions composantes pour les prédicteurs 3 à $p-1$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur p
  $\hat{y}_{k,\tau} \leftarrow \beta_0 \times f_{1,2}^k \times \ldots \times f_{p-1,2}^k \times f_{p,1}^k$
  $g_{p,2}^k \leftarrow \frac{\sum_{x_{i,p} \in b_p^k} y_i}{\sum_{x_{i,p} \in b_p^k} \hat{y}_{k,\tau}}$
  $f_{p,2}^k \leftarrow g_{p,1}^k \times g_{p,2}^k$
end for ▷ Fin de l'itération 2
... ▷ **Itérations 3 à** $p-1$
Itération T **:** $t = T$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 1
  $\hat{y}_{k,\tau} \leftarrow \beta_0 \times f_{1,T-1}^k \times \ldots \times f_{p,T-1}^k$
  $g_{1,T}^k \leftarrow \frac{\sum_{x_{i,1} \in b_1^k} y_i}{\sum_{x_{i,1} \in b_1^k} \hat{y}_{k,\tau}}$
  $f_{1,T}^k \leftarrow g_{1,1}^k \times \ldots \times g_{1,T}^k$ ▷ Tème mise à jour de la première fonction composante
end for
for $k = 1, \ldots, K$ **do** ▷ Prédicteur 2
  $\hat{y}_{k,\tau} \leftarrow \beta_0 \times f_{1,T}^k \times f_{2,T-1}^k \times \ldots \times f_{p,T-1}^k$
  $g_{2,T}^k \leftarrow \frac{\sum_{x_{i,2} \in b_2^k} y_i}{\sum_{x_{i,2} \in b_2^k} \hat{y}_{k,\tau}}$
  $f_{2,T}^k \leftarrow g_{2,1}^k \times \ldots \times g_{2,T}^k$
end for
... ▷ Calcul des fonctions composantes pour les prédicteurs 3 à $p-1$
for $k = 1, \ldots, K$ **do** ▷ Prédicteur p
  $\hat{y}_{k,\tau} \leftarrow \beta_0 \times f_{1,T}^k \times \ldots \times f_{p-1,T}^k \times f_{p,T-1}^k$
  $g_{p,T}^k \leftarrow \frac{\sum_{x_{i,p} \in b_p^k} y_i}{\sum_{x_{i,p} \in b_p^k} \hat{y}_{k,\tau}}$
  $f_{p,T}^k \leftarrow g_{p,1}^k \times \ldots \times g_{p,T}^k$
end for ▷ Fin de l'itération T

Annexe C

Comparaison des modèles

C.1 Interprétation des résultats

C.1.1 Modélisation du coût moyen

La Table permet de comparer la qualité d'ajustement des modèles de coût moyen au sens des métriques adaptées à l'étude des modèles linéaires (pour plus de détail, voir les Sections 2.2.2 et 2.2.3). Les modèles analysés sont :

- le modèle GLM de distribution Gamma avec la fonction de lien logarithmique utilisant les 34 variables explicatives sélectionnées à l'issue de la pré-sélection et de la sélection non-supervisée (Section 1.2.2) ;
- le modèle GLM de distribution Gamma avec la fonction de lien logarithmique utilisant les 8 variables explicatives sélectionnées à l'issue de la pré-sélection et des sélections supervisées et non-supervisées (Table 1.2).

Métrique	**Déviance**	**pseudo R^2**	**AIC**
GLM 34	50 019	0,082	1 139 105
GLM 8	50 464	0,074	1 139 679

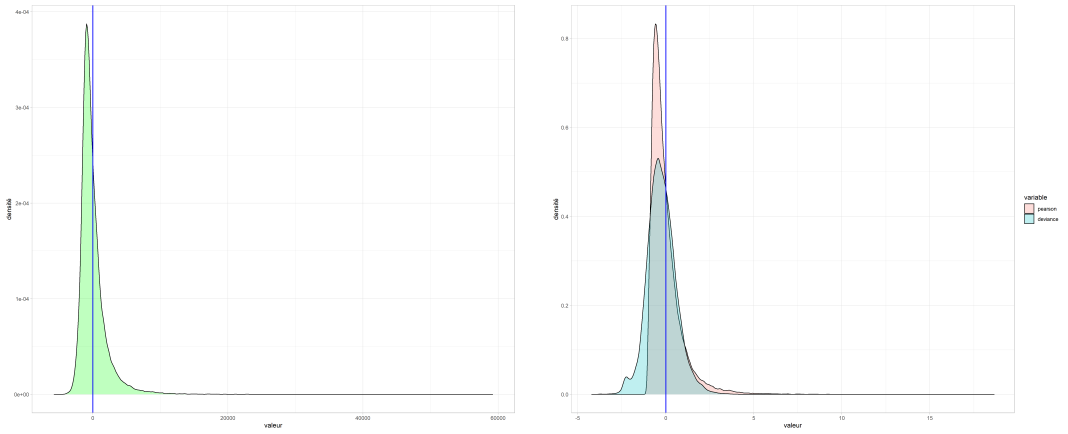
TABLE C.1 – Comparaison de la qualité d'ajustement des modèles de coût moyen

La Table C.1 montre alors que même le modèle linéaire généralisé avec 34 variables explicatives possède des meilleures propriétés d'ajustement. Néanmoins, la modélisation avec seulement 8 variables explicatives n'engendre qu'une très faible dégradation. Par la suite, l'analyse graphique de la distribution des résidus du modèle linéaire généralisé utilisant 8 variables explicatives permet de constater un bon comportement. Les Figures C.1a et C.1b montrent que les valeurs se concentrent essentiellement autour de 0, avec une distribution assez symétrique pour les résidus de Pearson et les résidus basés sur la notion de déviance, formalisme adapté à l'étude des modèles linéaires généralisés.

C.1.2 Modélisation de la fréquence

Pareillement à la modélisation du coût moyen, nous comparons :

- le modèle GLM de distribution Poisson avec la fonction de lien logarithmique utilisant les 34



(a) Résidus bruts (b) Résidus de Pearson et résidus de déviance

FIGURE C.1 – Distribution des résidus pour la modélisation du coût moyen avec un modèle linéaire généralisé utilisant 8 variables explicatives

variables explicatives sélectionnées à l'issue de la pré-sélection et de la sélection non-supervisée (Section 1.2.2) ;

- le modèle GLM de distribution Poisson avec la fonction de lien logarithmique utilisant les 8 variables explicatives sélectionnées à l'issue de la pré-sélection et des sélections supervisées et non-supervisées (Table 1.2).

Métrique	**Déviance**	**pseudo R^2**	**AIC**
GLM 34	572 823	0,0179	718 159
GLM 8	578 173	0,0088	723 416

TABLE C.2 – Comparaison de la qualité d'ajustement des modèles de fréquence

De nouveau, la Table C.2 montre que la modélisation avec 8 variables explicatives est comparable à celle avec 34. Une analyse graphique des résidus permet de constater que l'essentiel des valeurs se concentre autour de 0.

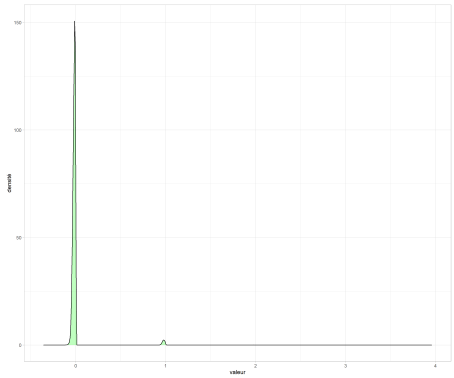


FIGURE C.2 – Distribution des résidus pour la modélisation de la fréquence avec un modèle linéaire généralisé utilisant 8 variables explicatives

C.1.3 Applicabilité opérationnelle

C.1.4 Performance et temps computationnel

La Figure C.3 montre le graphe des valeurs observées et prédites pour la modélisation de la fréquence sur la variable `age_COND`. On remarque que l'arbre de décision CART moyennise la tendance, sans capter les variations significatives de la donnée. Le modèle *Extreme Gradient Boosting* s'ajuste très bien aux valeurs observées, avec une allure similaire à la courbe obtenue par l'algorithme *Explainable Boosting Machine* (Figure 4.9).

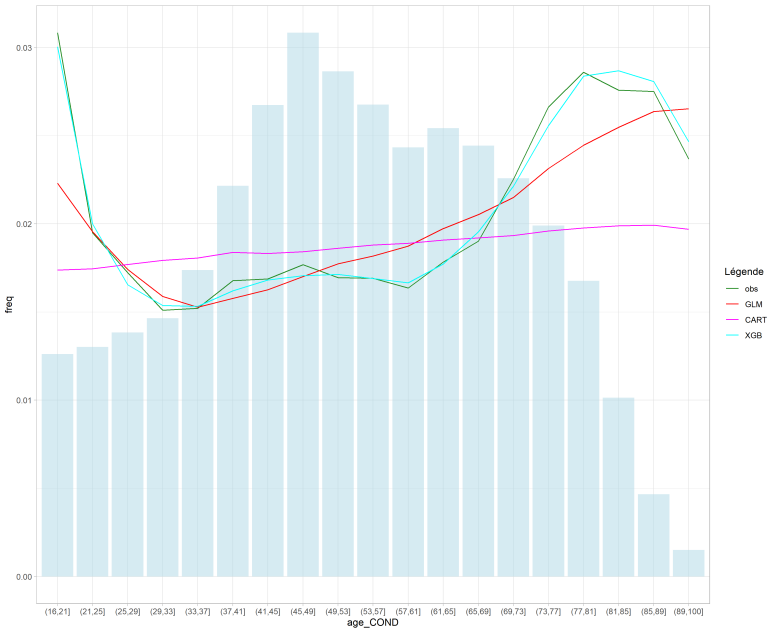


FIGURE C.3 – Valeurs observées et prédites pour la modélisation de la fréquence : comparaison des modèles GLM, CART et XGB sur la variable `age_COND`

C.1.5 Offset et Machine Learning

La Figure C.4 permet de visualiser la fonction composante associée à la variable exposition dans l'algorithme *Explainable Boosting Machine*. Nous remarquons alors une croissance quasi-linéaire avec le nombre de sinistres.

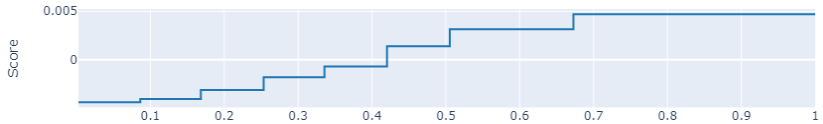


FIGURE C.4 – Fonction composante associée à la variable exposition dans le modèle GAM avec une structure arborescente

C.2 Prime pure et mouvements tarifaires

C.2.1 Analyse du Burning Cost et des primes prédites

La Figure C.5 permet de comparer le risque observé, représenté par le *Burning Cost*, au risque prédit par le modèle linéaire généralisé et le modèle additif généralisé avec une structure arborescente. L'analyse est réalisée sur la variable ancienneté du permis et montre en particulier un meilleur aujustement aux données observées par le modèle additif sur l'intervalle des âges allant de 30 à 50 ans.

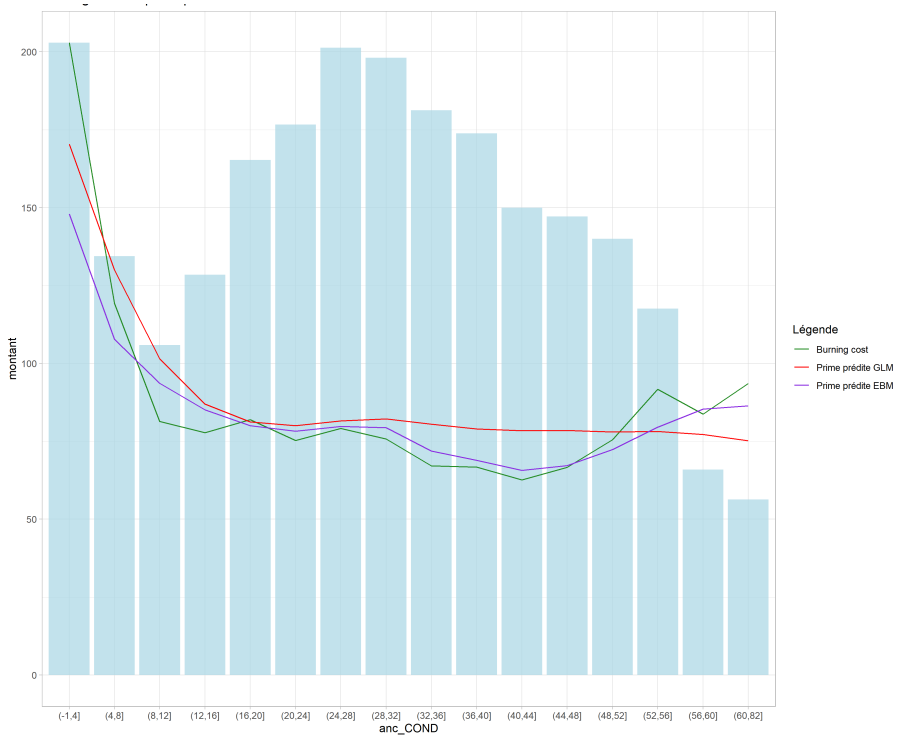


FIGURE C.5 – *Burning Cost* et les primes prédites par les modèles GLM et EBM pour la variable `anc_COND`

C.2.2 Surtarification et sous-tarification

Dans la Figure C.6, nous pouvons observer l'évolution de la variable δ en fonction de la classe de prix du véhicule. Pour les classes de prix faibles, le modèle EBM sous-tarifie par rapport au modèle GLM, puis la tendance s'inverse à partir de la classe de prix 16 pour augmenter significativement à partir de la classe de prix 25. Nous notons un pic sur la classe 9 qui fait écho à celui observé sur la Figure 4.4. La Figure C.7 montre la distribution des classes de prix en fonction de δ. En cohérence avec la Figure C.6, nous notons une légère sous-tarification pour les classes de prix faibles et une surtarification plus importante pour les classe de prix comprises entre 20 et 30.

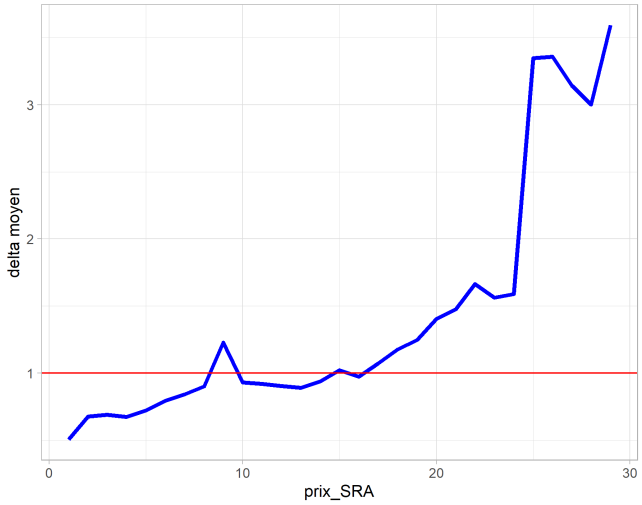


FIGURE C.6 – Évolution de δ en fonction de la variable `prix_SRA`

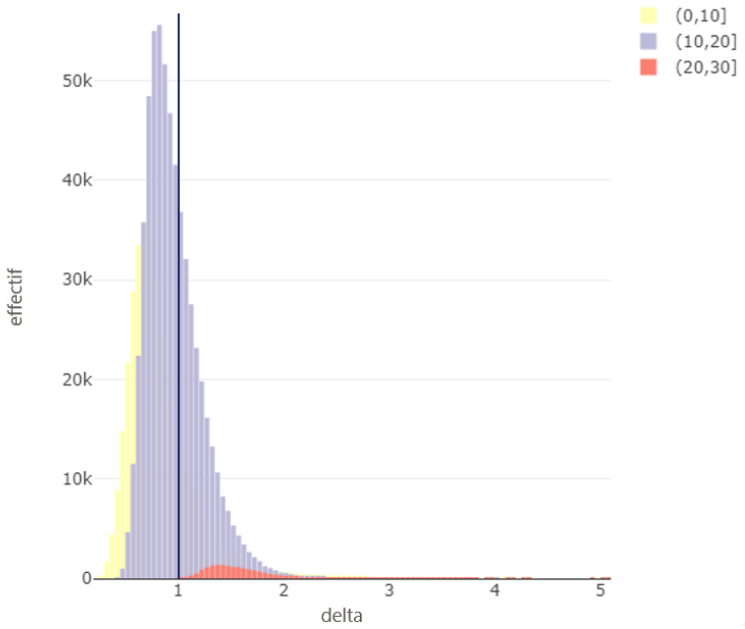


FIGURE C.7 – Distribution de la variable `prix_SRA` en fonction de δ