risks

MDPI

Article

Modelling Motor Insurance Claim Frequency and Severity Using Gradient Boosting †

Carina Clemente [1,‡], Gracinda R. Guerreiro [2,3,*,‡] and Jorge M. Bravo [4,5,6,7,‡]

1 NOVA IMS—Information Management School, Universidade Nova de Lisboa, 1070-312 Lisbon, Portugal; m20200314@novaims.unl.pt
2 FCT NOVA, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
3 CMA-FCT-UNL, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
4 NOVA IMS—Information Management School, Universidade Nova de Lisboa, MagIC, 1070-312 Lisbon, Portugal; jbravo@novaims.unl.pt
5 Department of Economics, University Paris-Dauphine PSL, 75016 Paris, France
6 CEFAGE-UE, 7000-809 Évora, Portugal
7 BRU-ISCTE-IUL, 1649-026 Lisbon, Portugal
* Correspondence: grg@fct.unl.pt
† This paper is a substantially extended and revised version of a conference paper to be presented at CAPSI 2023.
‡ These authors contributed equally to this work.

Abstract: Modelling claim frequency and claim severity are topics of great interest in property-casualty insurance for supporting underwriting, ratemaking, and reserving actuarial decisions. Standard Generalized Linear Models (GLM) frequency–severity models assume a linear relationship between a function of the response variable and the predictors, independence between the claim frequency and severity, and assign full credibility to the data. To overcome some of these restrictions, this paper investigates the predictive performance of Gradient Boosting with decision trees as base learners to model the claim frequency and the claim severity distributions of an auto insurance big dataset and compare it with that obtained using a standard GLM model. The out-of-sample performance measure results show that the predictive performance of the Gradient Boosting Model (GBM) is superior to the standard GLM model in the Poisson claim frequency model. Differently, in the claim severity model, the classical GLM outperformed the Gradient Boosting Model. The findings suggest that gradient boost models can capture the non-linear relation between the response variable and feature variables and their complex interactions and thus are a valuable tool for the insurer in feature engineering and the development of a data-driven approach to risk management and insurance.

Keywords: gradient boosting; non-life insurance pricing; expert systems; predictive modelling; risk management; actuarial science

Citation: Clemente, Carina, Gracinda R. Guerreiro, and Jorge M. Bravo. 2023. Modelling Motor Insurance Claim Frequency and Severity Using Gradient Boosting. *Risks* 11: 163. https://doi.org/10.3390/risks11090163

Academic Editor: Shengkun Xie

Received: 31 July 2023
Revised: 30 August 2023
Accepted: 6 September 2023
Published: 12 September 2023

1. Introduction

Modelling claim frequency and claim severity are topics of great interest in property-casualty insurance (e.g., Third Party Liability (TPL) Motor Insurance) and a crucial step for making appropriate underwriting, ratemaking, and reserving actuarial decisions. To this end, insurers tend to model separately the claim frequency and average claim severity using GLMs, in which the response variable—claim counts and claim amounts—is expressed through specific link transforms as linear combinations of feature (rating) variables such as the driver's age, car brand, education, or distance driven (Garrido et al. 2016; Renshaw 1994).

The standard frequency–severity model has, however, some limitations. First, the model assumes a linear relationship between the response variable and the predictors, with empirical studies documenting nonlinear effects between, e.g., claim severity and

the insured's age (Cunha and Bravo 2022; Frees and Valdez 2009). Alternative approaches using Generalized Additive Models (GAM) can overcome the linear predictor constraint of GLMs but have difficulty capturing the complex interactions among feature variables (Verbelen et al. 2018).

Second, the standard model assumes claim frequency and claim severity are independent. In practice, empirical studies show that claim counts and amounts are often dependent and negatively correlated in auto and health insurance (see, e.g., Frees et al. 2011; Garrido et al. 2016; Gschlößl and Czado 2007; Shi et al. 2015). Several authors have proposed mixed copula-based models linking the discrete distribution of claim counts and the continuous distribution of average claim size to deal with the dependencies (see, e.g., Czado et al. 2012; Frees and Wang 2006; Gao and Li 2023; Krämer et al. 2013; Shi 2016; Shi and Zhao 2020; Shi et al. 2015). An alternative modelling strategy is a conditional approach in which the number of claims is used as a feature variable in the GLM for the average claim size (see, e.g., Frees et al. 2011; Garrido et al. 2016; Jeong and Valdez 2020). Another approach is the multivariate conditional auto-regressive model and the multivariate Tobit model, which incorporates the correlation between the claim frequency and severity by modelling claim frequencies or accident rates at different severity levels (see, e.g., Aguero-Valverde 2013; Zeng et al. 2017).

Third, GLMs assign full credibility to the data, i.e., they assume the dataset contains enough observations for the parameter estimates to be considered fully credible. In practice, in segmented property-casualty insurance portfolios such as vehicle insurance, the issue of credibility must be addressed by considering, e.g., Generalized Linear Mixed Models (GLMMs) or Elastic Net GLMs (Katrien and Valdez 2011; Qian et al. 2016). Fourth, GLMs belittle conceptual uncertainty in empirical modelling, with recent literature highlighting the advantages of model ensembles in risk management (see, e.g., Bravo 2021).

The failure to flexibly capture the nonlinear relation between the claim frequency (severity) and often overlapping risk factors in GLMs and GAMs and the availability of larger datasets, including non-conventional data, shifted the attention towards the use of machine learning and deep learning methods in motor insurance modelling. Paefgen et al. (2013) and Baecke and Bocca (2017) used, respectively, Decision Trees (DT), Artificial Neural Networks (ANN), and Random Forests (RF) to predict claim counts in Usage-based Insurance (UBI) products such as pay-as-you-drive and pay-how-you-drive. Quan and Valdez (2018) compared the usage of univariate and multivariate response variables when predicting frequency in several non-auto coverages using the Classification and Regression Trees (CART), RF and Gradient Boosting (GB) models.

Pesantez-Narvaez et al. (2019) and Meng et al. (2022) examined the use of boosting machines in UBI claim probability prediction. The former concluded that the performance of boosting is less robust than classical logistic regression but attributed this to the small number of covariates considered in the study and the absence of hyperparameter tuning. Fauzan and Murfi (2018) analyse the accuracy of XGBoost in auto-insurance claim prediction and conclude that XGBoost shows increased accuracy in terms of normalised Gini when compared to the alternative methods AdaBoost, Stochastic GB, RF, and ANN. Su and Bai (2020) investigated the use of a stochastic gradient boosting algorithm and a profile likelihood approach to estimate parameters for both the claim frequency and average claim severity distributions in a French auto insurance dataset and concluded that the approach outperforms standard models.

To develop a full tariff plan for a Belgian TPL motor cover, Henckaerts et al. (2021) investigated the performance of simple regression trees, random forest, and boosted trees using the GLM as a benchmark and concluded that boosted trees outperformed GLMs. Similarly, Noll et al. (2020) predicted the claim frequency in a French motor TPL dataset using regression trees, GB, ANN and GLMs and concluded that GB and ANN outperformed the GLM, but also stated that the development of the benchmark model could have been improved. Su and Bai (2020) predicted the frequency and severity of the TPL motor cover, combining the stochastic gradient boost and a profile likelihood approach to estimate the

parameters of the distributions. This work adds to previous literature by introducing the dependence between claim frequency and claim average cost using the claim frequency as a predictor in the regression model for the severity. The authors concluded that abandoning the independence assumption contributes to increasing the model performance when compared to state-of-the-art models.

Some studies focus on other covers with great exposure, such as Collision. Staudt and Wagner (2021) developed frequency prediction on a Swiss motor portfolio, using GLM and GAM as reference models and two random forest models, one for claim severity and the other for log-transformed claim severity. The usage of the log-normal transformation of severity did not lead to any performance gains when the random forest was applied; however, it was still the favourite choice for explaining the right-skewed claims. Globally, GAM has a better performance.

Against this background, following and summarising the obtained results in Clemente (2023), this paper investigates the performance of Gradient Boosting with Decision Trees as base learners to separately model the claim frequency and the claim severity distributions of an international insurer auto insurance big dataset and compare it with that obtained using a standard GLM model. Boosting is one of the most popular ensemble learning methods, in some cases complemented with a model selection from a larger model space before aggregation. The method consecutively combines a large number of base weak learners in an additive form to tackle conceptual uncertainty in empirical research, capturing the nonlinear relation between the claim counts and amounts and feature variables and their complex interactions. First, our work contributes to the recent literature (see, e.g., Noll et al. 2020; Su and Bai 2020) by empirically investigating the performance of GBM and other machine learning methods to model claim frequency and severity of the TPL motor cover. In the risk management and insurance literature, Yang et al. (2018) developed a gradient-boosting Tweedie compound Poisson model and concluded that the model makes a more accurate premium prediction than GLM and GAM Tweedie compound Poisson models. Zhou et al. (2022) propose a boosting-assisted zero-inflated Tweedie model, called EMTboost, to cope with extremely unbalanced zero-inflated data.

Contrary to other machine learning methods with similar predictive accuracy, GB provides interpretable results, which makes it particularly attractive for modelling motor insurance losses. In GB models, complex interactions are simply modelled and may be included in the pricing structure. The feature selection is performed as an integral part of the application of the model, and this allows for a flexible approach when using GB models for insurance pricing. Actuaries may choose between different ways of using the potential of GB models: (a) adopt the GB model as the modelling tool to produce a new pricing structure or (b) identify statistically significant variables and interactions from the GB approach and include them on a GLM model, to improve the accuracy and prediction power of the model.

Second, our work uses a proprietary large, rich auto insurance database consisting of 0.8 million TPL vehicle insurance policies in force between 1 January 2016 and 31 December 2019, covering individuals against property damage, corresponding to 2.46 million observation duration (exposure to risk). Besides the response variables, the dataset includes 36 feature variables characterising the policyholder, the insurance policy, and the insured vehicle. This differentiates from most previous research, often using small (and selected) publicly available datasets made available for illustration purposes and research (e.g., the French Motor TPL Insurance Claims Data available in the Kaggle data repository). Third, we have implemented an extensive data pre-processing framework, including data cleaning, feature selection and engineering, outlier treatment and dimensionality reduction, as well as a detailed hyperparameter tuning approach using a nested k-fold cross-validation resampling procedure. In TPL motor insurance, the raw data often contain missing values or inconsistent data values resulting from human or computer error or the combination of multiple databases, outliers, irrelevant or redundant information, or inconsistent formats that can negatively impact the precision and reliability of data-driven models. Previous

research in multiple scientific domains concluded that data pre-processing and hyperparameter tuning positively contribute to improving the performance of machine learning and deep learning models (see, e.g., Chollet 2021).

A key requirement in the insurance (and financial) industry is the need for transparent and interpretable pricing models which are easily explainable to all stakeholders (e.g., managers, customers, shareholders, regulators, auditors), see, e.g., Kuo and Lupton (2023). Insurance ratemaking models are highly regulated, and they must meet specific requirements (see, e.g., the regime "algorithmic accountability" of decision-making machine algorithms imposed by the European Union's General Data Protection Regulation (European Parliament 2016, effective 25 May 2018, in which insurers are held accountable for their pricing models in terms of transparency, fairness, and solidarity) before being deployed in practice. To respond to that requirement, in this paper, we estimate two important tools to interpret the GB model, namely, variable importance measures and partial dependence plots.

The remainder of the paper is structured as follows. Section 2 summarises the GBM model used in the paper. Section 3 details the empirical strategy adopted, including the dataset information, the data pre-processing framework, and the hyperparameter tuning approach. Section 4 presents and discusses the main results. Section 5 concludes and sets out the agenda for further research.

2. Gradient Boosting Machines

A common task in the application of statistical learning, machine learning and deep learning methods in finance, insurance, and risk management is to develop a parametric or non-parametric classification, regression, or ranking model from the data. Empirical work in these domain-specific areas is subject to significant uncertainty about model specification. This may be the consequence of the lack of a universally accepted theory that has been empirically verified as a (near) perfect explanation of reality (theory uncertainty), the multiple ways in which theories can be empirically tested (specification uncertainty), heterogeneity uncertainty and variable independence (Steel 2020).

One way to circumvent model uncertainty is to pursue a data-driven approach, learning the model directly from the data. The customary approach to data-driven modelling is to neglect model risk and pursue a "winner-takes-all" perspective by which, for each dataset, a unique believed to be the "best" model is selected from a set of candidate (weak) learners using some method or statistical criteria (goodness-of-fit, predictive), see Bravo and Ayuso (2021). The statistical inference then proceeds conditionally upon the assumption that the selected model happens to be a good approximation to the true data generating process.

To tackle conceptual uncertainty and overcome the shortcomings of individual learners, an alternative approach is model combination, i.e., building an ensemble of (homogeneous or diverse) classifiers (e.g., artificial neural networks, support vector regressions, GLMs, recurrent neural networks), often complemented with a model selection from a larger model space before aggregation Jose and Winkler (2008). Ensemble methods aim at finding a static or dynamic composite model that better approximates the actual data generation process and its multiple sources of uncertainty. Empirical studies show that they can provide superior predictive accuracy relative to single learners in several domain-specific areas (Ashofteh et al. 2022; Ayuso et al. 2021; Bravo 2021; Kim and Baek 2022). Examples of successful applications of machine-learning ensemble techniques in different domains include random forests (Breiman 2001), artificial neural network ensembles (Hansen and Salamon 1990; Shu and Burn 2004), Bayesian model ensembles (Bravo and Ayuso 2021; Raftery et al. 1997), bootstrap aggregating (bagging), boosting and meta-learning strategies for expert combination such as stacking (Ashofteh and Bravo 2021; Wolpert 1992), arbitrating (Ortega et al. 2001), dynamic combiners (Sergio et al. 2016) or mixture of experts (Jacobs et al. 1991).

In gradient boosting machines, the learning process proceeds by consecutively building an ensemble of shallow and weak base-learners (e.g., linear models, smooth models, or decision trees), with each step learning and improving on the previous one to form a committee that is capable of accurate estimating the response variable. The algorithm is constructed such that the new base learners are maximally correlated with the negative gradient of the loss function (e.g., squared-error loss, Adaboost) of the whole ensemble (Friedman 2001). The approach is quite flexible and can be customised to any data-driven task and has proven considerable achievement in real-world applications (Hanafy and Ming 2021; Henckaerts et al. 2021).

Formally, let y denote a random response variable and $\mathbf{x}$ a set of input or predictor variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Using a training sample $(y, \mathbf{x})_{i=1}^{N}$ of known $(y, \mathbf{x})$-values, the goal is to obtain an estimate of the approximation $\hat{F}(\mathbf{x})$ of the function $F^*(\mathbf{x})$ mapping the unknown functional dependence $\mathbf{x} \xrightarrow{F} y$, that minimises the expected value of some specified loss function $\mathcal{L}(y, F(\mathbf{x}), \mathbf{w})$ over the joint distribution of all $(y, \mathbf{x})$-values,

$$F^* = \underset{F}{arg\,min}\, E_x\big[E_y(\mathcal{L}(y, F(\mathbf{x}), \mathbf{w})\big], \tag{1}$$

possibly complemented by a weights function $\mathbf{w} = (w_1, w_2, \ldots, w_n)$. For instance, for claim frequency modelling, the weight is the exposure to risk, typically the number of policy years, and the response variable is the number of claims divided by $\mathbf{w}$. In claim severity modelling exercises, the weight is the number of claims, with the average claim size (claims cost divided by the number of claims) as the response variable.

To make the estimation problem tractable, a common procedure is to restrict the function search space to a member of a parametric family of functions $F(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ is a finite set of parameters whose joint values identify the individual learners. Following Friedman (2001), in this paper, we focus on a class of additive expansions of the form

$$F\Big(\mathbf{x}, \{\beta_m, \mathbf{a}_m\}_{i=1}^{M}\Big) = \sum_{m=1}^{M} \beta_m\, h(x, \mathbf{a}_m), \tag{2}$$

where $h(\mathbf{x}; \mathbf{a})$ is a base or weak learner function of the input variables with parameters $\mathbf{a} = \{a_1, a_2, \ldots\}$. Choosing a parametric model transforms the function optimisation problem into a parameter optimisation problem:

$$\{\beta_m, \mathbf{a}_m\}_{i=1}^{M} = \underset{\{\beta_m, \mathbf{a}_m\}}{arg\,min} \sum_{i=1}^{N} \mathcal{L}\left(y_i, \sum_{m=1}^{M} \beta_m\, h(\mathbf{x}, \mathbf{a}_m), w_i\right). \tag{3}$$

Given M iteration steps, the parameter estimates can be written in the incremental form. For $m = 1, 2, \ldots, M$, we can write

$$(\beta_m, \mathbf{a}_m) = \underset{\beta, a}{arg\,min} \sum_{i=1}^{N} \mathcal{L}(y_i, F_{m-1}(\mathbf{x}_i) + \beta\, h(\mathbf{x}_i, \mathbf{a}), w_i), \tag{4}$$

with incremental steps or "boosts" defined by the optimisation method

$$F_m(\mathbf{x}_i) = F_{m-1}(\mathbf{x}_i) + \beta_m\, h(\mathbf{x}_i, \mathbf{a}_m). \tag{5}$$

The numerical optimisation is resolved by GBM through a two-step process using the steepest-descent algorithm, which is based on consecutive improvements along the direction of the gradient of the loss function in which, for each interaction, the pseudo-residuals are used to assess the regions of the predictor space for which the model does not have a good performance, and therefore improve the fit in the direction of better overall performance. In this paper, we consider decision trees as base learners $h(\cdot, \cdot)$. This means

parameters a_m are the splitting variables and splitting points that define the tree, and the base learner is of the following form:

$$h(x_i, \{R_{lm}\}_1^L) = \sum_{l=1}^{L} \bar{y}_{lm}\, \mathbb{I}(\mathbf{x} \in R_{lm}), \tag{6}$$

where $\bar{y}_{lm}$ is the mean of the pseudo-residuals $\tilde{y}_{im}$ for observation i in iteration m over the region R_{lm}, forming a subdivision of the space R where the vector $\mathbf{x}$ takes its values. Decision trees are commonly selected as base learners in gradient boosting because: (i) of their ability to capture complex interactions and nonlinear relationships in the data, creating splits and branches to signify intricate decision boundaries, which are crucial when dealing with datasets containing nontrivial patterns, (ii) they inherently provide feature importances, helping the algorithm to identify and focus on the features which contribute the most to reducing the prediction error, (iii) of their flexibility and adaptability, since they can fit both micro and macro trends in the data and error correction mechanisms, (iv) of their interpretability, since single decision trees are relatively easy to read and visualize, helping in understanding the model's decision-making process, a crucial element in insurance pricing and risk management, (v) of their robustness to outliers and capacity to handle missing data, (vi) of is capacity to generate ensemble diversity through feature selection, alternative splitting criteria and different depth levels, (vii) of their computational tractability when compared to some other machine learning algorithms, making it feasible to work with large datasets such as in this study. Since the value of the base learners $h(\cdot, \cdot)$ is constant for each region of the tree, $\beta h(x_i, \mathbf{a}_m)$ can be simplified to γ and Equation (4) re-written as:

$$\gamma_{lm} = \underset{\gamma}{arg\,min} \sum_{i=1}^{N} \mathcal{L}(y_i, F_{m-1}(\mathbf{x}_i) + \gamma), \tag{7}$$

with incremental boosts for each region R_{lm} updated using γ_{lm}

$$\hat{F}_m(\mathbf{x}_i) = \hat{F}_{m-1}(\mathbf{x}_i) + \lambda\, \gamma_{lm}\, \mathbb{I}(\mathbf{x} \in R_{lm}), \tag{8}$$

with $\lambda (0 < \lambda \leq 1)$ the learning rate (also known as the shrinking parameter) determining the learning pace of the algorithm by shrinking updates for $\mathbf{x} \in R_{lm}$. A lower value of λ outputs a better performance, reducing overfitting but also increasing the computational power required because more trees are necessary for the algorithm to minimise the pseudo-residuals and to converge. Usually, λ is fixed at the lowest value possible within the computational restraints (Henckaerts et al. 2021). The performance of the GBM model investigated in this paper is tested against the results provided by the benchmark GLM approach. The model fitting, forecasting, simulation procedures, and additional computations have been implemented using an R (version 4.2.0) software routine.

3. Empirical Strategy

3.1. Dataset Information and Treatment

The automobile insurance database used in this study was supplied, for the work of Clemente (2023), by a European insurer operating in European and non-European insurance markets. The proprietary dataset used in this study is private and is not available for public use. The dataset consists of 799,587 Third Party Liability motor insurance policies covering individuals against property damage, in force between 1 January 2016 and 31 December 2019, corresponding to 2,464,181 observation duration or exposure-to-risk (fraction of the year when the policy was in force). Of these, a total of 78,264 insurance claims were recorded during the four-year period with a total incurred cost of 97.9 million euros. In addition to response variables, the dataset includes 36 characteristic variables that characterise the policyholder (e.g., age, education, job, marital status, seniority of driver's license), the insurance policy (e.g., coverage, payment method), and the insured vehicle (e.g., age of vehicle, car brand, driving km per year, fuel type, number of vehicle seats).

We have implemented an extensive data pre-processing framework including data cleaning, data pre-processing, feature selection and engineering, outlier treatment, and dimensionality reduction that allows for variable evaluation, model implementation and accuracy improvement. For instance, the correlation analysis identified a strong correlation between the location-related variables, such as Municipality, District, Delegation, and Driving Zone, as well as between Driving Zone and distribution method or between NBexe (new business or renewed policies) and the driver's age. As a result of this pre-processing stage, the final dataset used for model calibration consists of 2,464,181 observation durations and 21 feature variables, summarised in Table 1.

Table 1. Final feature variables.

Variable	Levels	Description
UEN	RIF, ZRT	Type of client (RIF—Individual, ZRT—direct channel)
Client Time on Book (years)	1 to 21 (individually), 21+, 999	The seniority of the policyholder in the company
Payment Instalments	1/year, 2/year, 4/year, 12/year	N. payments per year
Agent Delegation	22 different levels, from PD1 to PD22	Policy distribution channel
Direct Debit Payment	Non-DB, DB	If the policy payments come from direct-charge or not
Policy Time on Book (years)	1 to 21 (individually), 21+	The policy's seniority, time since contract initiation
Vehicle Brand	708 different levels from OM1 to OM708, unknown	Vehicle Brand
Vehicle Seats	2, 3, 4, 5, 6, 7, 8, 9, 11+, 999	N. of seats in the vehicle
Engine Capacity	32 levels (1–50, . . . , 1000–1100, . . . , 5000+)	Engine capacity of the vehicle
Horse Power	0–50, 50–100, 100–150, 150–200, 200+	Vehicle power, measured in horsepower
Vehicle Weight (kg)	32 levels (<50, . . . , 1700–1800, . . . , 3500+)	Vehicle Weight
Vehicle Value as New (Euro)	14 levels (<7000, . . . , 25,000–30,000, . . . 500,000+)	Initial price of the vehicle, as if it was new
Fuel	8 distinct levels, from OF1 to OF8, without fuel, other, unknown	Type of fuel
District	22 different levels, from ODC1 to ODC22, unknown	The policyholder's (usual driver) District of residence
Bonus–Malus	20 levels (−5, −4, . . . , 0, 1,. . . , 13, 14)	Bonus–Malus System (BMS)
Years of Driving	1 to 21 (individually), 21+, 999	Seniority of the driver's license
Vehicle Age (years)	1 to 30 (individually), 30+, 999	Age of the vehicle
Driver Age	0–17, 18 to 85 (individually), 85+, unknown	Age of the usual driver
Cover Capital (Euro)	CapMin, CapMax	CapMax if the policy has the optional 59M TPL capital, or CapMin otherwise
NBexe (New Business)	RN, NB, FNB RN (renewal), NB (New Business), FNB (fake new business)	
Own Damage Cover	Yes, No	Yes (the policy includes own damage coverage) No (otherwise)

The dataset also includes a discrete quantitative variable representing the number of claims reported per policy, per year, and the corresponding claim amount. The average annual claim rate was 4.834%, with a variance of 0.09289, with only 3.07% of the policies

reporting claims during the four-year study period. The corresponding average cost per claim was EUR 1251, with a standard deviation of EUR 1972.37. The average cost per insurance policy during this period was EUR 122.45, with a standard deviation of EUR 749.75.

In the pre-processing stage, performed in Clemente (2023), we have analysed the relationship between the response variables and the feature variables. For instance, Figure 1 represents the relationship between the driver's age and the claim frequency (on top) and the claim severity (on bottom). The average driver's age in the insurance portfolio is 51 years old, with ages ranging between 18 and 93 years old. Figure 1 shows that the frequency and the severity are significantly different between age groups, with the peak at 21 years of age, then declining with age up to 70 years old. After this age, an increase in both claim frequency and severity may be explained by a natural reduction of driving skills and less exposure to risk.

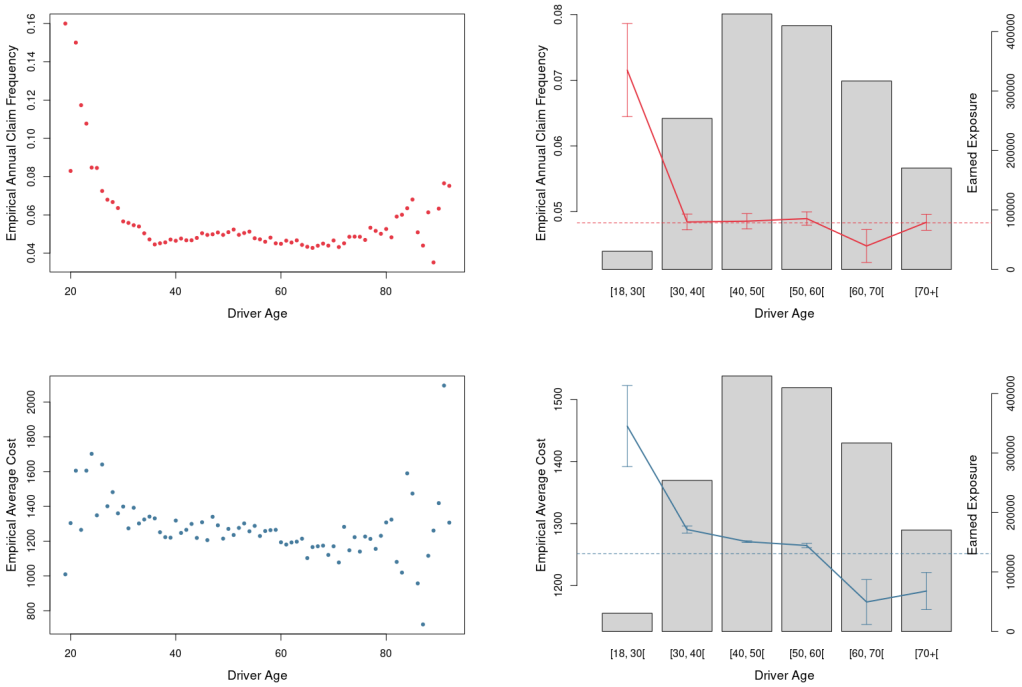

Figure 1. Claim frequency (**top** row) and claim severity (**bottom** row) vs. age of the driver-individual (**left** panel) and by group ages (**right** panel).

3.2. Tuning Approach

Machine learning methods usually rely on training data to construct a model, validation data to tune the parameters to be applied, and test data to evaluate the out-of-sample model performance. A fundamental part of successfully training a tree-based model is to control model complexity, taking into account the bias-variance trade-off. A large tree has low bias and high variance, whereas a small tree has a high bias but low variance. For validating machine learning and deep learning models, it is customary to apply the so-called nested cross-validation since it can handle both the selection of the best set of hyperparameters and error estimation (Henckaerts et al. 2021). The method starts by partitioning a dataset $\mathcal{D}$ into K disjoint, equally sized and stratified sets $\mathcal{D} = \cup_{k=1}^{K} \mathcal{D}_k$ ordered on claim frequency or claim severity. By imposing stratification based on claim frequency/severity, the goal is to obtain a similar distribution of the claim frequency/severity in each of the K subsets. In this paper, we use a nested K-fold cross-validation resampling procedure (with $K \in [1, 2, \ldots, 6]$) for evaluating and comparing the learning algorithms and tuning machine learning hyperparameters. Next, the iterative procedure consists of a double loop of cross-validation, with the inner loop serving for hyperparameter tuning and the independent outer loop serving for assessing model performance. Specifically, in

the inner loop, cross-validation is performed for every hyperparameter combination, and the cross-validation error is computed by applying the loss function and averaging on the validation data sets. For each model, the optimal hyperparameters are those that minimise the cross-validation error. The inner loop comprises K_1 folds and the outer loop of K_2 folds, with the total number of trained models equal to K_1K_2. The training data is divided into a learning set (80%) and a validation set (20%) considering random folds of nearly the same size, mutually exclusive, and stratified (Hastie et al. 2009). The fold partition used in this paper represents a compromise between the objective of reducing the generalisation error and the computational burden (Boehmke and Greenwel 2020).

For both the claim frequency and claim severity models, a grid search procedure was used to calibrate the boosting and the decision tree-specific hyperparameters (Su and Bai 2020). The procedure systematically works through a range of hyperparameter combinations in order to find the ones that result in the best predictive performance for the model. Specifically, we calibrate the GBM algorithm for the number of decision trees (N) accounting for (i) overfitting with $N \in \{100, 250, 400, 500, 750, 1000\}$ in the claim frequency model and $N \in \{100, 150, 200, 250, 300, 400, 500\}$ in the claim severity model; (ii) controlling for different values of the learning rate (shrinkage factor) assuming $\lambda \in \{0.1, 0.05, 0.01\}$ in both the claim frequency and claim severity models. For both GLM and GBM, the weights were defined as unitary for the claim severity model (each claim is an observation), whereas for the claim frequency model, the response variable was redefined as the number of claims divided by exposure to risk, see, e.g., Ohlsson and Johansson (2010) and Wüthrich and Merz (2023) for details on standard non-life insurance modelling. Regarding the decision tree-specific hyperparameters, for both models, we have systematically investigated multiple combinations controlling for the tree depth (with integer values ranging between 1 and 5) for the minimum number of observations in the terminal nodes, which determine the complexity of each tree (we assumed a 1% rate) and for the bag (subsampling) fraction, i.e., the proportion of the training set observations randomly selected to propose the next tree in the expansion (bag fraction $\in \{0.7, 0.8, 0.9, 0.95\}$). Overall, to find the GBM optimal parameters, the grid search procedure investigated 360 combinations in the claim frequency model and 420 parameter specifications in the claim severity model.

For applying machine learning algorithms such as decision trees or GBMs, we need to specify which loss function to minimise when training the model. For example, the typical loss function for regression problems is the squared error loss, which is only appropriate when the data are normally distributed. However, in practice, claim frequency and claim severity are not normally distributed. Claim frequency typically is assumed to be Poisson distributed, whereas claim severity usually is assumed to be Gamma distributed. Because of this, authors such as Wüthrich and Buser (2023) and Wüthrich and Merz (2023) suggest using Poisson deviance and Gamma deviance as an appropriate loss function in GBM models, a methodological option we pursued in this paper. Furthermore, in GLMs, estimating the model parameters by finding the values that maximise the log-likelihood function is equivalent to minimising the unscaled deviance loss.

Table 2 summarises the optimal set of parameters for each of the six folds tested in the claim frequency (Part A) and claim severity (Part B) models, respectively. They represent the hyperparameter combination that generated the smallest cross-validation iteration error (the out-of-sample Poisson (Gamma) deviance in the claim frequency (severity) model) in that fold. In the claim frequency model, we can observe that the maximum number of optimal trees is achieved for smaller values of shrinkage factor, a well-known behaviour identified in similar studies using GBMs. The average optimal number of trees is significantly smaller in the claim severity model compared to the claim frequency model. In a significant number of cases, the claim severity model uses trees with only one split as weak learners, making the models additive and without interactions.

Table 2. Optimal tuning parameters per fold.

A. Claim frequency model					
# Fold (*K*)	**# Trees**	**Shrinkage Factor**	**Interaction Depth**	**Bag Fraction**	**Poisson Deviance [1]**
1	37	0.10	2	0.95	0.2802844
2	64	0.10	2	0.95	0.2802088
3	642	0.01	2	0.80	0.2802078
4	116	0.10	1	0.95	0.2793700
5	239	0.05	2	0.95	0.2791459
6	47	0.10	4	0.95	0.2796919
Average	190	0.077	2	0.925	–
B. Claim severity model					
# Fold (*K*)	**# Trees**	**Shrinkage Factor**	**Interaction Depth**	**Bag Fraction**	**OOS Gamma Deviance [2]**
1	133	0.05	1	0.95	15.76648
2	125	0.05	1	0.95	15.76562
3	56	0.05	2	0.70	15.76658
4	33	0.1	1	0.80	15.76578
5	59	0.05	2	0.70	15.76709
6	75	0.1	1	0.95	15.76697
Average	80	0.067	1	0.925	–

[1] Optimal tuning parameters and out-of-sample Poisson deviance (models with smaller deviance are better) estimated considering random samples of 50,000 observations extracted from a 1.97 million observations training set used for calibrating the frequency model. [2] Optimal tuning parameters and out-of-sample Gamma deviance estimated considering the whole set of 18,801 observations.

4. Results

4.1. Model Performance

To estimate both the optimal GBM model and the benchmark standard GLM model, we use 80% of the observations as training data and the remaining 20% as testing data. Table 3 reports the in-sample and out-of-sample loss for the claim frequency and claim severity. In the claim frequency model, the empirical results show, for both the training and test samples, that the gradient boosting model exhibits lower deviance compared to the classical GLM model. Differently, in the claim severity model, the standard GLM model significantly outperformed the optimal GBM model, exhibiting lower gamma deviance. The poorer performance of the GBM model in the claim severity model may be explained, first, by the significantly smaller sample size (circa 18,000 observations) compared to the 1.97 million available in the claim frequency model. It is well known that machine learning models tend to perform better the larger the available dataset. Second, the higher volatility of the claim amount distribution compared to the claim count distribution makes it harder to model using the GBM model.

Table 3. Total Poisson and Gamma deviance for frequency and severity models, for both sub-samples and all data.

Sample	**Total Poisson Deviance**		**Total Gamma Deviance**	
	GLM	**GBM**	**GLM**	**GBM**
Training (80%)	432,456	428,621	8787	10,545
Testing (20%)	107,914	106,773	2162	2624
All (100%)	540,406	535,685	10,954	13,209

4.2. Model Interpretation

Contrary to single decision trees, GBM combining multiple decision trees is hard to interpret since the model cannot be represented and visualised with a two-dimensional graphic. To overcome the black-box nature of GBM, (Breiman 2001; Friedman and Popescu 2008; Hastie et al. 2009) introduced two important tools to interpret the GBM model: (i) variable importance measures and (ii) partial dependence plots. Variable importance is a measure of how important the feature variables are in predicting the response. Formally, for a specific explanatory variable $x_l, l \in \{1, 2, \ldots, n\}$, in the m-th decision tree, the variable importance is measured by summing the improvements in the loss function over all the internal nodes $J-1$ for which the variable x_l was used as the splitting variable:

$$\mathcal{J}_l(m) = \sum_{j=1}^{J-1} \mathbb{I}(v(j) = l)(\Delta\mathcal{L})_j \tag{9}$$

In ensemble techniques such as GBMs, feature importance is measured by averaging the importance of variable x_l over the different trees included in the ensemble, i.e.,

$$\mathcal{J}_l = \frac{1}{M} \sum_{m=1}^{M} \mathcal{J}_l(m) \tag{10}$$

with influences averaged over all trees and normalised so that they add up to 100%.

4.2.1. Variable Importance Measure

Figure 2 shows the variable importance scores for the optimal GBM claim frequency and claim severity models taking, for each fold, the average over all trees and discarding features with $\mathcal{J}_l < 0.1\%$. The results show, for each cross-validation number of folds, that the policyholder's (usual driver's) district of residence, the Bonus–Malus level, the vehicle brand, the frequency of premium payments, and the policy's seniority are the five most important variables in predicting the claim frequency. Other important variables for predicting auto insurance claim frequency are the driver's age, the vehicle's age, the client's seniority, and the vehicle's horsepower. The finding also shows that the variable importance scores can fluctuate according to the cross-validation number of folds used in tuning the GBM model.

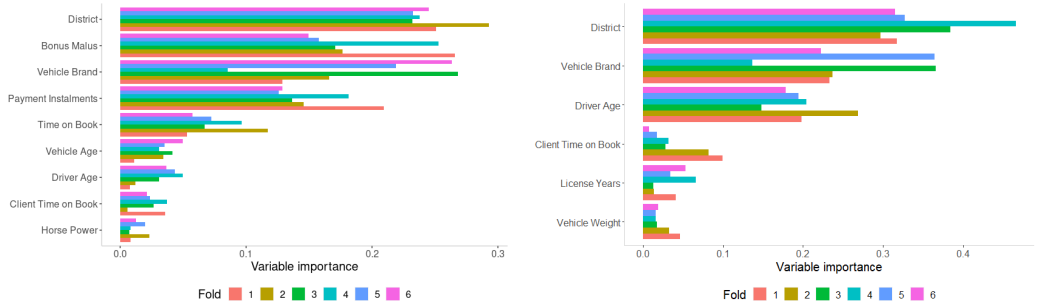

Figure 2. Variable importance in the optimal GBM per data fold, claim frequency (**left**) and claim severity (**right**) models.

Similarly, the right plot of Figure 2 represents the variable importance scores for the optimal GBM claim severity model, discarding again features with $\mathcal{J}_l < 0.1\%$. The findings show, for each cross-validation number of folds, that the usual driver district of residence, the vehicle brand, the driver's age, the years of driving experience, and policy seniority are the five most important variables in predicting the auto insurance claim severity. The finding suggests that the variable importance is not homogeneous over the number of K-folds used in tuning the GBM claim severity model.

4.2.2. Partial Dependence Plots

In GLM models, the additive monotonic form of the linear predictor and the low degree of interacting variables augment model interpretability. Differently, in gradient boosting, the influences measured by variable importance scores do not provide any explanations about how a given feature actually affects the response. However, in decision-tree GBMs, visualisation tools such as Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots can be used to visualise the effect of the predictor on the modelled response (claim frequency) after marginalising the remaining explanatory variables. Partial dependence plots exhibit the average effect of a feature on the predictions obtained from a model, see Hastie et al. (2009), whereas ICE plots disaggregate the averaged data, providing a method of inspecting how the instance's prediction changes when a feature varies. In PDPs, we calculate the predictions for a specific variable x_l while averaging over the values of the remaining features x_G, i.e.,

$$f_l(x_l) = \frac{1}{N}\sum_{h=1}^{N} f_{model}(x_l, x_{h,G}), \tag{11}$$

where $x_{h,G}$ denotes the values of the remainder feature variables for observation h, n is the number of observations in the training subset and G is the complement set to variable l.

Figure 3 depicts the graphical representation of the PDP effect of the policyholder's (usual driver's) district of residence on the claim frequency per data fold, considering a sample of 1000 observations. The results suggest that Districts 4, 8, 14, 17, and 19 exhibit a higher similar risk of reporting a claim across all folds.

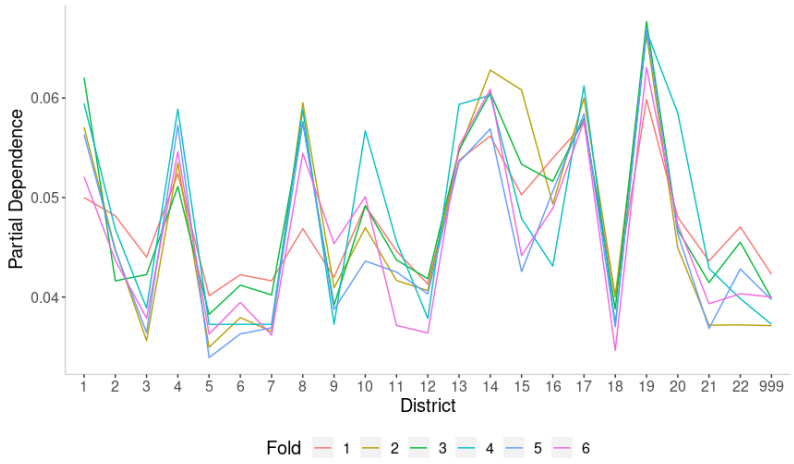

Figure 3. PDP of the policyholder's district of residence on claim frequency, per data fold, using a sample of 1000 observations.

Similarly, Figure 4 depicts the graphical representation of the PDP effect of the driver's age on claim severity, per data fold, considering again a sample of 1000 observations. The findings clearly show the inverse relationship between the driver's age (and driving experience) and claim severity, with younger, inexperienced drivers more likely to report higher average claim amounts per claim count.

In Figure 5, we randomly select 1000 policies from fold 5 to produce the ICE plot for the feature vehicle vehicle brand. Each line of the plot represents how the response changes when the vehicle brand changes, keeping all other variables constant. The blue line represents the average of these lines, i.e., the partial dependence. ICE plots allow us to capture heterogeneity in the relationship between the feature variable and the response created by variable interaction, see Goldstein et al. (2015).

The patterns in Figures 5 and 6 suggest that the ICE lines tend to follow the same trend as the average. However, in Figure 5, the overlapping crossing lines observed for

vehicle brands number 15, 46, 47, 75, and 102, deviating from the average, indicate a possible interaction between the vehicle brand and another feature. On the other hand, in Figure 6, the concentration around the average, especially for districts 4, 5, 7, 8, 11, 15 and 16, shows that keeping all the other risks constant, the severity of a claim is less sensitive in these districts.

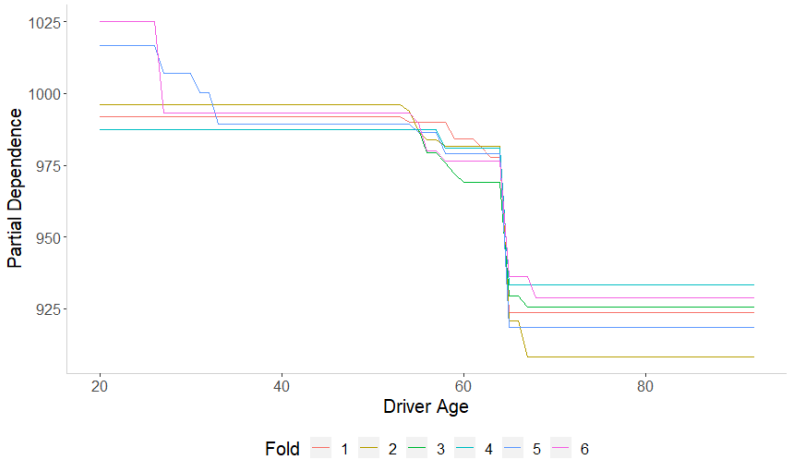

Figure 4. PDP representing the effect of the Driver Age on severity, per data fold, using a sample of 1,000 observations on the training dataset.

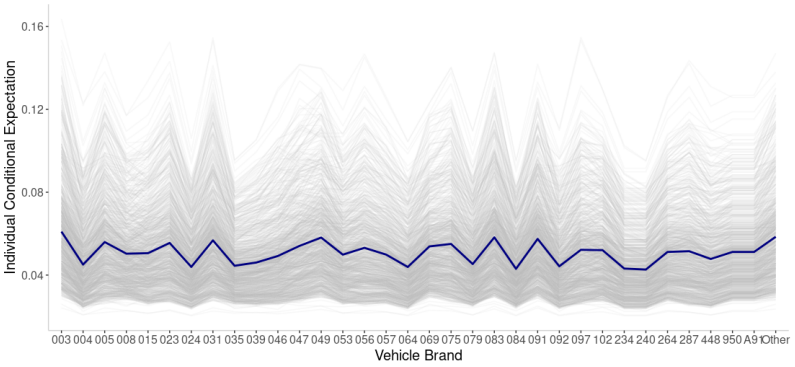

Figure 5. Effect of the Vehicle Brand on the frequency model as Partial Dependence (in dark blue) and Individual Conditional Expectation (in grey), considering 5-fold cross-validation.

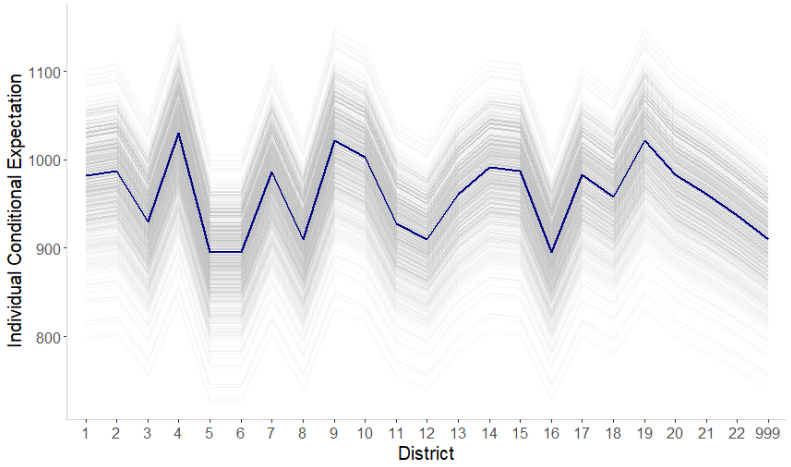

Figure 6. Effect of the District on the severity model as Partial Dependence (in dark blue) and Individual Conditional Expectation (in grey), considering data fold 2.

4.2.3. GLM Model Results

Table 4 shows the regression output of the GLM models trained for the claim frequency (left panel) and the claim severity (right panel) models. We used a forward stepwise procedure to select the variables to include in the regression model, testing for the individual significance of each feature variable complemented by the Likelihood Ratio test to assess whether a given block of variables significantly contributes to improving the overall model performance.

The findings show that the type of fuel, the vehicle's brand, age, and horsepower, the number of premium payments per year, the driver's district of residence, age and years of driving, and the policy seniority significantly contribute to predicting claim frequency. Going deeper, the results show, first, that the probability of reporting a claim in an auto accident is lower for automobiles using fuel type G2 (gasoline) compared to vehicles using diesel. Second, the results also show that the higher the frequency of premium payments per year (G2 = monthly and semi-annual payments; G3 = quarterly payments), the higher the likelihood of a reported auto insurance claim against benchmark policies with a single payment per year. Third, drivers living in the countryside (level G2) and small cities (level G3) tend to have a lower probability of reporting an auto accident. Regarding the driver's age, the findings show that, compared to the benchmark age group comprising individuals aged between 36 and 44 years old (level G2), younger drivers (G1: age $\in [18, 35]$) exhibit a lower probability of a claim, whereas older drivers (G3: age $\in [45, 65]$, G4: age $\in [66, 70]$; G5: age $\in [71, 76]$; G6: age ≥ 77) exhibit higher claim risk. Older and more powerful vehicles evidence a lower likelihood of reporting a claim. Regarding the driver's experience, the findings show that compared to the benchmark age group comprising individuals between 18 and 32 years of driving (Level G6), less experienced drivers with from one to 17 years of driving experience (Levels G1 to G5) exhibit a significantly higher probability of reporting a claim, whereas more experienced drivers tend to have low claim risk (levels G7 and G8). Finally, the more senior the insurance policy, the lower tends to be the claim risk (levels G2 to G6) compared to one-year policies (level G1).

Moving now to the claim severity model (Table 4, right panel), the findings reveal a much simpler model with the vehicle's brand, the driver's age, and the policyholder's (usual driver) district of residence as the only statistically significant variables. The claim severity of auto accidents in small villages in the countryside is lower compared to large urban areas (level G2) but higher in smaller cities across the country (level G3). The claim severity of accidents involving younger drivers is significantly higher compared to adult drivers and lower for older (and more experienced) drivers.

Table 5 summarises the list of features selected by both the GBM model and the benchmark GLM claim frequency and claim severity model. The results show that, out of the nine main variables identified as important by the GBM claim frequency model (with a minimum of 0.1% of variable importance score per fold), only the Bonus–Malus level and Client Seniority (client time on book) were not selected in the GLM model. Bonus–Malus Systems (BMS), rewarding claim-free years by discounts and penalising at-fault accidents with premium surcharges are a powerful incentive for safe driving. However, it is also well-known that BMS can encourage the non-reporting of claims to avoid premium penalties. Because of this, some forms of BMS introduce varying (escalating) deductibles that prevent malus evasion. The results also show that, out of the six key variables identified as important by the GBM claim severity model, only the years of driving experience and the vehicle weight were not selected in the corresponding GLM model. Although the findings show that most data features are present in both prediction models considered in this study, the results suggest that the GBM approach has a slightly higher capability of selecting the feature variables that best differentiate claim frequency and claim severity risks in TPL auto insurance.

Table 4. GLM outputs for Claim Frequency and Claim Severity.

Variable	Level	Claim Frequency			Claim Severity		
		β	Std. Error	p-Value	β	Std. Error	p-Value
Intercept	–	−2.8334	0.0176	≈0	6.8804	0.01083	≈0
Fuel	G2	−0.1034	0.00958	≈0	–	–	–
Vehicle Brand	G2	0.0421	0.01034	4.7×10^{-5}	−0.0327	0.01268	9.9×10^{-3}
Payment Instalments	G2	0.2138	0.00989	≈0	–	–	–
	G3	0.4288	0.01447	≈0	–	–	–
District	G2	−0.3905	0.01142	≈0	−0.0863	0.02079	3.3×10^{-5}
	G3	−0.2743	0.01039	≈0	0.0909	0.02463	2.2×10^{-4}
Driver Age	G1	−0.1597	0.02357	≈0	0.0736	0.03396	3×10^{-2}
	G3	0.02032	0.02122	≈0	–	–	–
	G4	0.5415	0.03041	≈0	−0.0938	0.01644	1.14×10^{-8}
	G5	0.1544	0.01219	≈0	−0.0938	0.01644	1.14×10^{-8}
	G6	0.2874	0.02380	≈0	−0.0938	0.01644	1.14×10^{-8}
Years Driving	G1	0.9770	0.06352	≈0	–	–	–
	G2	0.6739	0.05099	≈0	–	–	–
	G3	0.4385	0.03676	≈0	–	–	–
	G4	0.2549	0.02842	≈0	–	–	–
	G5	0.1419	0.01926	≈0	–	–	–
	G7	−0.0608	0.01373	9.5×10^{-6}	–	–	–
	G8	−0.1289	0.01644	≈0	–	–	–
Vehicle Age	G2	−0.1607	0.01454	≈0	–	–	–
Horse Power	G2	−0.1625	0.04331	≈0	–	–	–
Time on Book	G2	−0.1032	0.01533	≈0	–	–	–
	G3	−0.1806	0.01498	≈0	–	–	–
	G4	−0.2969	0.01517	≈0	–	–	–
	G5	−0.3910	0.01671	≈0	–	–	–
	G6	−0.5036	0.03379	≈0	–	–	–

Table 5. Variables included in the frequency and severity models, according to both approaches: GLM and GBM (only those with over 0.1% variable importance).

Frequency		Severity	
GLM	**GBM**	**GLM**	**GBM**
Age of the Driver	Age of the Driver	Age of the Driver	Age of the Driver
Years of Driving	–	–	Years of Driving
District	District	District	District
–	Client Time on Book	–	–
Policy Time on Book	Policy Time on Book	–	–
Payment Instalments	Payment Instalments	Payment Instalments	Payment Instalments
–	Bonus–Malus	–	–
Brand	Brand	Brand	Brand
Vehicle Age	Vehicle Age	–	–
Horse Power	Horse Power	–	–
Fuel	–	–	–
–	–	–	Vehicle Weight

4.2.4. Searching for Interactions

Friedman's H-statistic (Friedman and Popescu 2008), evaluates the strength of interactions between feature variables by measuring how much of the prediction variance stems from the interaction effect. In this paper, we focus on two-way interactions, but the statistic can be applied to any number of variables.

Let $f_l(x_l)$ and $f_u(x_u)$ be the one-way partial dependence of the feature variables x_l and x_u, and $f_{lu}(x_l, x_u)$ the two-way partial dependence defined by Equation (11). The H-statistic is defined as

$$H_{lu}^2 = \frac{\sum_{h=1}^{N}\left[f_{lu}\left(x_l^{(h)}, x_u^{(h)}\right) - f_l\left(x_l^{(h)}\right) - f_u\left(x_u^{(h)}\right)\right]^2}{\sum_{h=1}^{N}\left[f_{lu}^2\left(x_l^{(h)}, x_u^{(h)}\right)\right]}, \tag{12}$$

where $x_l^{(h)}$ denotes observed value x_l for observation h. By definition, by squaring and scaling, we obtain H-statistc values in the range between 0 and 1, with 0 representing the absence of interaction between two variables and 1 signalling that the effect of a feature on the response variable is attributable to the interaction only.

To further check for interactions between feature variables, Friedman's H-statistic was estimated for all possible combinations. Table 6 reports the obtained results for the 10 strongest two-way interactions between all feature variables in the GBM frequency model, considering data fold 5. The H-statistic results suggest that the features of vehicle brand and frequency of premium payment may interact in explaining claim frequency (H-statistic = 0.2255). Important but weaker interaction effects are also suggested between the policyholder's district of residence and policy seniority (H-statistic = 0.2004) and between client seniority and the vehicle's age (H-statistic = 0.1560). As a result of this analysis, the interaction between the policyholder's district of residence and policy seniority was included in the GLM model. However, as the model performance has not improved significantly, the interaction was removed from the final GLM model.

Table 6. Friedman's H-statistic for the 10 strongest two-way interactions between all feature variables in the GBM claim frequency model, considering 5-fold cross-validation.

Variables	H-Statistic
(Payment Instalments, Vehicle Brand)	0.2255
(District, Policy seniority)	0.2004
(Client seniority, Vehicle Age)	0.1560
(Bonus–Malus, Payment Instalments)	0.1424
(Payment Instalments, Policy seniority)	0.1355
(District, Vehicle Brand)	0.1147
(Bonus–Malus, District)	0.1038
(District, Payment Instalments)	0.0868
(Bonus–Malus, Vehicle Brand)	0.0867
(District, Vehicle Age)	0.0695

Considering data fold 5, Figure 7 shows the effect of the feature vehicle brand on claim frequency as partial dependence, grouped by Payment Instalments. The plot suggests that for brands 24, 49, 64, 92, and 448, the claim frequency risk associated with insurance policies with quarterly premium payment is superior to that of other payment frequencies. For other car brands, such as brands 3, 53, 57, and 97, the different premium payment frequencies do not seem to affect the claim frequency predictions.

Figure 7. Grouped partial dependence plot for the frequency GBM model, considering data fold 5.

5. Conclusions

Modelling claim frequency and claim severity is a critical task in ratemaking in property-casualty insurance. The type of data available for this exercise typically includes risk factors (policy details, policyholder information, insured vehicle characteristics, driving records) and numerical response variables, making statistical learning and supervised machine learning methods particularly suitable for this task. Generalised Linear Models are the industry benchmark for developing analytic insurance pricing models. The standard GLM and GAM frequency–severity models assume a linear or additive relationship between the response variable and the feature variables, that the claim risk and claim severity are independent, and tend to assign full credibility to the data.

To overcome these restrictions, this paper investigates the predictive performance of Gradient Boosting with decision trees as base learners to model the claim frequency and the claim severity distributions of an auto insurance big dataset and compare it with that obtained using a standard GLM model. The gradient boosting algorithm is a machine learning method for optimising prediction accuracy carrying out variable selection, particularly useful in the context of high-dimensional data. The model results in prediction rules which share some interpretability characteristics as the standard statistical model fits. The gradient boosting algorithm combines learners with "poor" performance (high prediction error) as regression or classification trees to produce a highly accurate prediction rule with easily interpretable results.

The use of gradient boosting models with decision trees as base learners in auto insurance ratemaking permits the segmentation of a portfolio of policyholders into groups of homogeneous risk profiles based on some feature variables, inducing transparency and intra-group risk pooling under common asymptotic (group size) conditions. The use of ensemble (model combination) techniques combining multiple decision trees instead of selecting single learners addresses conceptual uncertainty concerns in responsible insurance pricing and provides a sounder basis for statistical inference. Ensemble models aim to find a composite model that better approximates the actual data generation process and its multiple sources of uncertainty. They have a long tradition in the statistical and forecasting literature, yet they are relatively underexplored in insurance pricing (Bravo 2021). A fundamental part of successfully training machine learning models is to control model complexity and overfitting, considering the bias-variance trade-off. Instead of relying on built-in tuning strategies, we performed an extensive grid search procedure using nested cross-validation among a predefined tuning grid for evaluating and comparing the performance of the learning algorithms and tuning machine learning hyperparameters and for analysing the stability of the results across multiple data folds.

The results of the out-of-sample performance measure show that the predictive performance of the Gradient Boosting model is superior to that of the standard GLM model in the Poisson claim frequency model. On the contrary, in the claim severity model, the classical

GLM significantly outperformed the GBM. The findings for the claim frequency model suggest gradient boosting models can capture the nonlinear relation between the response variable and risk variables and their complex interactions and are thus a valuable tool for feature engineering and the development of a data-driven approach to auto insurance ratemaking and risk management. The poorer performance of the GBM model in the claim severity model may be explained by the small sample size available for training the model and the significant volatility of the claim amount distribution.

Regarding model interpretation, the variable importance measures allowed us to identify the most relevant variables in the frequency and severity models. The finding also shows that the variable importance scores can fluctuate according to the cross-validation number of folds used in tuning the GBM model. An interesting result for both the claim frequency and the claim severity model is that the most important risk factors in the gradient boosting machines are those selected in the corresponding GLMs. A similar conclusion was obtained by Henckaerts et al. (2021) using a portfolio of motor Third Party Liability from a Belgian insurer in 1997. The results suggest, however, that the GBM approach has a slightly higher capability of selecting the feature variables that best differentiate claim frequency and claim severity risks in TPL auto insurance. The fact that both models include a similar selection of risk factors means gradient boosting models can assist in the selection of the candidate feature variables (and their complex interactions) to consider in the tuning of the GLM to be used in pricing and risk management.

The partial dependence plots and individual conditional expectation plots provide additional insight into a selection of noteworthy effects for the claim risk model. An important and well-known effect in auto insurance pricing detected by PDP and ICE plots is the interaction between driver age and driving experience and claim frequency and severity. The results highlight a clear inverse relationship between driver age and claim severity, with younger drivers more likely to have a serious accident. Further research should investigate the performance of GBM against other supervised machine learning methods (e.g., Random Forest, Classification and Regression Tree, K-Nearest Neighbours, and Artificial Neural Networks-based models).

Author Contributions: C.C. participated in the Conceptualization, Methodology, Investigation and Writing—review and editing. C.C. was also responsible for the software implementation, Formal Analysis and Data Curation. G.R.G. participated in Conceptualization, Methodology, Investigation, writing—original draft preparation and was responsible for Validation. J.M.B. participated in the Investigation, writing—original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by national funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 and UIDP/00297/2020—Center for Mathematics and Applications—(G.R. Guerreiro) and grants UIDB/04152/2020—Centro de Investigação em Gestão de Informação (MagIC) and UIDB/00315/2020—BRU-ISCTE-IUL—(J.M. Bravo).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: 3rd Party Data. Restrictions apply to the availability of these data.

Acknowledgments: The authors express their gratitude to the insurer that allowed them to perform this study, by making available (to the authors) the dataset used in this study. The authors are grateful to the anonymous referees for the comments, questions and suggestions made to an earlier version of this paper.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Aguero-Valverde, Jonathan. 2013. Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. *Accident Analysis & Prevention* 59: 365–73.

Ashofteh, Afshin, and Jorge M. Bravo. 2021. A Conservative Approach for Online Credit Scoring. *Expert Systems With Applications* 176: 114835. [CrossRef]

Ashofteh, Afshin, Jorge M. Bravo, and Mercedes Ayuso. 2022. A New Ensemble Learning Strategy for Panel Time-Series Forecasting with Applications to Tracking Respiratory Disease Excess Mortality during the COVID-19 pandemic. *Applied Soft Computing* 128: 109422. [CrossRef]

Ayuso, Mercedes, Jorge M. Bravo, Robert Holzmann, and Eduard Palmer. 2021. Automatic indexation of pension age to life expectancy: When policy design matters. *Risks* 9: 96. [CrossRef]

Baecke, Philippe, and Lorenzo Bocca. 2017. The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems* 98: 69–79. [CrossRef]

Boehmke, Bradley, and Brandon Greenwel. 2020. *Hands-On Machine Learning with R*, 1st ed. Boca Raton: CRC Press, Taylor & Francis.

Bravo, Jorge M. 2021. Pricing Participating Longevity-Linked Life Annuities: A Bayesian Model Ensemble approach. *European Actuarial Journal* 12: 125–59. [CrossRef]

Bravo, Jorge M., and Mercedes Ayuso. 2021. Linking Pensions to Life Expectancy: Tackling Conceptual Uncertainty through Bayesian Model Averaging. *Mathematics* 9: 3307. [CrossRef]

Breiman, Leo. 2001. Random Forests. *Machine Learning* 45: 5–32. [CrossRef]

Chollet, François. 2021. *Deep Learning with Python*, 2nd ed. New York: Manning.

Clemente, Carina. 2023. A Refreshed Vision of Non-Life Insurance Pricing—A Generalized Linear Model and Machine Learning Approach. Master's thesis, NOVA IMS, Lisbon, Portugal.

Cunha, Lourenço, and Jorge M. Bravo. 2022. Automobile Usage-Based-Insurance: Improving Risk Management using Telematics Data. Paper presented at 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), Madrid, Spain, June 22–25, pp. 1–6.

Czado, Claudia, Rainer Kastenmeier, Eike C. Brechmann, and Aleksey Min. 2012. A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal* 4: 278–305. [CrossRef]

European Parliament. 2016. *General Data Protection Regulation*. Regulation (EU) 2016/679. Strasbourg: European Parliament.

Fauzan, Muhammad A., and Hendri Murfi. 2018. The Accuracy of XGBoost for Insurance Claim Prediction. *International Journal of Advances in Soft Computing and Its Applications* 10: 159–71.

Frees, Edward W., and Emiliano A. Valdez. 2009. Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin: The Journal of the IAA* 39: 165–97. [CrossRef]

Frees, Edward W., and Ping Wang. 2006. Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics* 38: 360–73. [CrossRef]

Frees, Edward W., Jie Gao, and Marjorie A. Rosenberg. 2011. Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal* 15: 377–92. [CrossRef]

Friedman, Jerome H. 2001. Greedy boosting approximation: A gradient boosting machine. *Annals of Statistics* 29: 1189–232. [CrossRef]

Friedman, Jerome H., and Bogdan E. Popescu. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics* 2: 916–54. [CrossRef]

Gao, Guangyuan, and Jiahong Li. 2023. Dependence modeling of frequency-severity of insurance claims using waiting time Author links open overlay panel. *Insurance: Mathematics and Economics* 109: 29–51.

Garrido, Jose, Christian Genest, and Juliana Schulz. 2016. Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics* 70: 205–15. [CrossRef]

Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* 24: 44–65. [CrossRef]

Gschlößl, Susanne, and Claudia Czado. 2007. Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal* 3: 202–25. [CrossRef]

Hanafy, Mohamed, and Ruixing Ming. 2021. Machine learning approaches for auto insurance big data. *Risks* 9: 42. [CrossRef]

Hansen, Lars K., and Peter Salamon. 1990. Neural networks Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12: 993–1001. [CrossRef]

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning—Data Mining, Inference, and Prediction*, 2nd ed. Springer Series in Statistics. Berlin: Springer.

Henckaerts, Roel, Marie-Pier Coté, Katrien Antonio, and Roel Verbelen. 2021. Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods. *North American Actuarial Journal* 25: 255–85. [CrossRef]

Jacobs, Robert A., Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation* 3: 79–87. [CrossRef] [PubMed]

Jeong, Himchan, and Emiliano A. Valdez. 2020. Predictive compound risk models with dependence. *Insurance. Mathematics and Economics* 94: 182–85. [CrossRef]

Jose, Victor R., and Robert L. Winkler. 2008. Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting* 24: 163–69. [CrossRef]

Katrien, Antonio, and Emiliano A. Valdez. 2011. Statistical Concepts of a Priori and a Posteriori Risk Classification in Insurance. *Advances in Statistical Analysis* 96: 187–224.
Kim, Donghwan, and Jun-Geol Baek. 2022. Bagging ensemble-based novel data generation method for univariate time series forecasting. *Expert Systems with Applications* 203: 117366. [CrossRef]
Krämer, Nicole, Eike C. Brechmann, Daniel Silvestrini, and Claudia Czado. 2013. Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics* 53: 829–39. [CrossRef]
Kuo, Kuo, and Daniel Lupton. 2023. Towards Explainability of Machine Learning Models in Insurance Pricing. *Variance 16*. Available online: https://variancejournal.org/article/68374-towards-explainability-of-machine-learning-models-in-insurance-pricing (accessed on 5 September 2023).
Meng, Shengwang, Yaqian Gao, and Yifan Huang. 2022. Actuarial intelligence in auto insurance: Claim frequency modeling with driving behavior features and improved boosted trees. *Insurance: Mathematics and Economics* 106: 115–27. [CrossRef]
Noll, Alexander, Robert Salzmann, and Mario V. Wüthrich. 2020. Case Study: French Motor Third-Party Liability Claims. *SSRN Eletronic Journal*, 1–41. [CrossRef]
Ohlsson, Esbjörn, and Björn Johansson. 2010. *Non-Life Insurance Pricing with Generalized Linear Models*, 2nd ed. Berlin: Springer.
Ortega, Julio, Moshe Koppel, and Shlomo Argamon. 2001. Arbitrating among competing classifiers using learned referees. *Knowledge and Information Systems* 3: 470–90. [CrossRef]
Paefgen, Johannes, Thorsten Staake, and Frédéric Thiesse. 2013. Evaluation and aggregation of pay-as-you-drive insurance rate factors: a classification analysis approach. *Decision Support Systems* 56: 192–201. [CrossRef]
Pesantez-Narvaez, Jessica, Monserrat Guillen, and Manuela Alcañiz. 2019. Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks* 7: 70. [CrossRef]
Qian, Wei, Yi Yang, and Hui Zou. 2016. Tweedie's Compound Poisson Model with Grouped Elastic Net. *Journal of Computational and Graphical Statistics* 25: 606–25. [CrossRef]
Quan, Zhiyu, and Emiliano A. Valdez. 2018. Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling* 6: 377–407. [CrossRef]
Raftery, Adrian, David Madigan, and Jennifer Hoeting. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92: 179–91. [CrossRef]
Renshaw, Arthur E. 1994. Modelling the claims process in the presence of covariates. *ASTIN Bulletin* 24: 265–85. [CrossRef]
Sergio, Anderson, Tiago P. F. de Lima, and Teresa B. Ludermir. 2016. Dynamic selection of forecast combiners. *Neurocomputing* 218: 37–50. [CrossRef]
Shi, Peng. 2016. Insurance ratemaking using a copula-based multivariate Tweedie model. *Scandinavian Actuarial Journal* 2016: 198–215. [CrossRef]
Shi, Peng, and Zifeng Zhao. 2020. Regression for copula-linked compound distributions with application in modelling aggregate insurance claims. *The Annals of Applied Statistics* 14: 357–80. [CrossRef]
Shi, Peng, Xiaoping Feng, and Anastasia Ivantsova. 2015. Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics* 64: 417–28. [CrossRef]
Shu, Chang, and Donald H. Burn. 2004. Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resources Research* 40: 1–10. [CrossRef]
Staudt, Yves, and Joel Wagner. 2021. Assessing the performance of random forests for modeling claim severity in collision car insurance. *Risks* 9: 53. [CrossRef]
Steel, Mark F. J. 2020. Model Averaging and Its Use in Economics. *Journal of Economic Literature* 58: 644–719. [CrossRef]
Su, Xiaoshan, and Manying Bai. 2020. Stochastic gradient boosting frequency-severity model of insurance claims. *PLoS ONE* 15: e0238000. [CrossRef] [PubMed]
Verbelen, Roel, Katrien Antonio, and Gerda Claeskens. 2018. Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society. Series C. Applied Statistics* 67: 1275–304. [CrossRef]
Wolpert, David H. 1992. Stacked generalization. *Neural Networks* 5: 241–59. [CrossRef]
Wüthrich, Mario V., and Christoph Buser. 2023. *Data Analytics for Non-Life Insurance Pricing*. Swiss Finance Institute Research Paper No. 16-68. Zürich: ETH Zurich. [CrossRef]
Wüthrich, Mario V., and Michael Merz. 2023. *Statistical Foundations of Actuarial Learning and Applications*. Cham: Springer.
Yang, Yi, Wei Qian, and Hui Zou. 2018. Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. *Journal of Business & Economic Statistics* 36: 456–70.
Zeng, Qiang, Huiying Wen, Helai Huang, Xin Pei, and Sze Wong. 2017. A multivariate random-parameters Tobit model for analyzing highway crash rates by injury severity. *Accident Analysis & Prevention* 99: 184–91.
Zhou, He, Wei Qian, and Yang Yang. 2022. Tweedie Gradient Boosting for Extremely Unbalanced Zero-inflated Data. *Communications in Statistics—Simulation and Computation* 51: 5507–29. [CrossRef]