ISCTE IUL
REPOSITÓRIO
INSTITUTO UNIVERSITÁRIO DE LISBOA

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2026-07-28

Deposited version:
Publisher Version

Peer-review status of attached file:
Peer-reviewed

Citation for published item:
Clemente, C., Guerreiro, G. R. & Bravo, J. M. (2023). Gradient boosting in motor insurance claim frequency modelling. In CAPSI 2023 Proceedings. (pp. 53-69).: Associacao Portuguesa de Sistemas de Informacao.

Further information on publisher's website:
10.18803/capsi.v23.53-69

Publisher's copyright statement:
This is the peer reviewed version of the following article: Clemente, C., Guerreiro, G. R. & Bravo, J. M. (2023). Gradient boosting in motor insurance claim frequency modelling. In CAPSI 2023 Proceedings. (pp. 53-69).: Associacao Portuguesa de Sistemas de Informacao., which has been published in final form at https://dx.doi.org/10.18803/capsi.v23.53-69. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Serviços de Informação e Documentação, Instituto Universitário de Lisboa (ISCTE-IUL)
Av. das Forças Armadas, Edifício II, 1649-026 Lisboa Portugal
Phone: +(351) 217 903 024 | e-mail: administrador.repositorio@iscte-iul.pt
https://repositorio.iscte-iul.pt

Association for Information Systems
AIS Electronic Library (AISeL)

CAPSI 2023 Proceedings | Portugal (CAPSI)

10-21-2023

Gradient Boosting in Motor Insurance Claim Frequency Modelling

Carina Clemente
NOVA IMS, m20200314@novaims.unl.pt

Gracinda R. Guerreiro
NOVA FCT & NOVA MATH, Universidade Nova de Lisbo, grg@fct.unl.pt

Jorge M. Bravo
NOVA IMS & MagIC, Universidade Nova de Lisboa; BRU-ISCTE-IUL; CEFAGE-UE, jbravo@novaims.unl.pt

Follow this and additional works at: https://aisel.aisnet.org/capsi2023

Recommended Citation

Clemente, Carina; Guerreiro, Gracinda R.; and Bravo, Jorge M., "Gradient Boosting in Motor Insurance Claim Frequency Modelling" (2023). *CAPSI 2023 Proceedings*. 5.
https://aisel.aisnet.org/capsi2023/5

This material is brought to you by the Portugal (CAPSI) at AIS Electronic Library (AISeL). It has been accepted for inclusion in CAPSI 2023 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Gradient Boosting in Motor Insurance Claim Frequency Modelling

Carina Clemente, NOVA IMS, Portugal, m20200314@novaims.unl.pt

Gracinda R. Guerreiro, NOVA FCT & NOVA MATH, Universidade Nova de Lisboa, Portugal, grg@fct.unl.pt

Jorge M. Bravo, NOVA IMS & MagIC, Universidade Nova de Lisboa; BRU-ISCTE-IUL; CEFAGE-UE, Portugal, jbravo@novaims.unl.pt

Abstract

Modelling claim frequency and claim severity are topics of great interest in property-casualty insurance for supporting underwriting, ratemaking, and reserving actuarial decisions. This paper investigates the predictive performance of Gradient Boosting with Decision Trees as base learners to model the claim frequency in motor insurance using a private cross-country large insurance dataset. The Gradient Boosting algorithm combines many weak base learners to tackle conceptual uncertainty in empirical research. The findings show that the Gradient Boosting model is superior to the standard Generalised Linear Model in the sense that it provides closer predictions in the claim frequency model. The finding also shows that Gradient Boosting can capture the nonlinear relation between the claim counts and feature variables and their complex interactions being, thus, a valuable tool for feature engineering and the development of a data-driven approach to risk management.

Keywords: Gradient Boosting; Non-life Insurance Pricing; Expert systems; Predictive modelling; Risk Management

1. INTRODUCTION

Modelling claim frequency and claim severity are topics of great interest in property-casualty insurance (e.g., Third Party 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 Generalised linear models (GLMs), in which the response variable – claim counts and claim amounts – is expressed, through specific canonical link transforms, as linear combinations of feature (rating) variables such as the age of the driver, the brand of the car, the driver's education level or the distance driven (Renshaw 1994; Garrido et al. 2016).

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 non-linear effects between, e.g., claim severity and the insured's age (Frees et al. 2009; Cunha and Bravo, 2022). Alternative approaches using Generalised Additive Models (GAM) can

ISSN 2183-489X
http://dx.doi.org/10.18803/capsi.v23.53-69

overcome the linear predictor constraint of GLMs, but they have difficulty capturing the complex interactions among feature variables (Verbelen et al., 2018). Second, the standard model assumes that claim frequency and claim severity are independent. In practice, empirical studies show that claim counts and amount are often dependent and negatively correlated in auto and health insurance (see, e.g., Gschlößl and Czado 2007; Frees et al. 2011; Shi et al. 2015; Garrido et al. 2016). Several authors have proposed copula-based models to deal with the dependencies (see, e.g., Czado et al. 2012; Shi 2016). Third, GLMs assign full credibility to the data, i.e. they assume the dataset has 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., generalised linear mixed model (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 et al., 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 utilizing the 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 alternative methods AdaBoost, Stochastic GB, RF, and ANN. Su and Bai (2020) investigate 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 conclude that the approach outperforms standard models.

To develop a full tariff plan for a Belgian TPL motor cover, Henckaerts et al. (2020) 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) predict the claim frequency in a French motor TPL dataset using regression trees, GB, ANN, and GLMs and conclude 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 the 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 severity. The authors conclude that abandoning the independence assumption contributes to increasing the model performance compared to state-of-the-art models.

Some studies focus on other covers with great exposure, such as collisions. 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, this paper investigates the performance of Gradient Boosting with decision trees as base learners to model the claim frequency distribution of an international insurer auto insurance big dataset and compare it with that obtained using a standard GLM model. Then, we estimate variable importance measures and partial dependence plots to interpret the GB model. Boosting is one of the most popular ensemble learning methods, in some cases complemented by a model selection from a larger model space prior to 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. We have implemented an extensive data preprocessing framework and hyperparameter tuning approach using a nested k-fold cross-validation resampling procedure.

The rich auto insurance database used in this study consists of 0.8 million Third Party Liability (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. In addition to the response variables, the dataset includes 36 feature variables characterizing the policyholder, the insurance policy, and the insured vehicle.

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.

The findings show that the GB model is superior to the standard GLM model in the sense that it provides closer predictions and lower deviance in the Poisson claim frequency model. The remainder of the paper is structured as follows. Section 2 summarises the key materials and methods used in the paper. Section 3 details the empirical strategy adopted, including the dataset information, the data preprocessing 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. The empirical work in these domain-specific areas is subject to significant uncertainty regarding 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 bypass 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-take-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 (goodness-of-fit, predictive) criteria (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 to find 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; Kim and Baek, 2022; Bravo, 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 (Raftery et al., 1997; Bravo et al., 2021, 2023), bootstrap aggregating (bagging), boosting and meta-learning strategies for expert combination such as stacking (Wolpert, 1992; Ashofteh and Bravo, 2021), arbitrating (Ortega et al., 2001), dynamic combiners (Sergio et al., 2016) or mixture of experts (Jacobs et al., 1991).

In gradient boosting machines (GBMs), 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 (Henckaerts et al., 2020; Hanafy & Ming, 2021).

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

$$F^* = \arg\min_{F} E_x \left[E_y \left(\mathcal{L}\big(y, F(\boldsymbol{x})\big) \right) | \boldsymbol{x} \right] \tag{1}$$

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 $F(\boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots\}$ 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

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

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

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

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

$$(\beta_m, \boldsymbol{a}_m) = \arg\min_{\beta,a} \sum_{i=1}^{N} \mathcal{L}\big(y_i, F_{m-1}(\boldsymbol{x}_i) + \beta h(\boldsymbol{x}_i; \boldsymbol{a})\big) \tag{4}$$

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

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

The numerical optimization 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 a direction of better overall performance. In this paper, we consider decision trees as base learners $h(.,.)$. 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:

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

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}. Since the value of the base learners $h(.,.)$ is constant for each region of the tree, $\beta\, h(\boldsymbol{x_i}, \boldsymbol{a}_m)$ can be simplified to γ and equation (4) re-written as:

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

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

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

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 $\boldsymbol{x} \in R_{lm}$. A lower value of λ outputs a better performance, reducing overfitting, but also increases the computational power required, because more trees are necessary to converge to a good solution. 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 generalized linear model (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 is a private dataset supplied by a European insurer operating in multiple markets. The dataset is not available for public use and the insurer prefers to remain anonymous. It consists of 799 587 Third Party Liability (TPL) motor insurance policies, in force between 1 January 2016 and 31 December 2019, covering individuals against property damage, corresponding to 2 464 181 observation duration or exposure-to-risk (fraction of the year when the policy was in force). For these, a total of 78 264 insurance claims were registered during the four years with a total incurred cost of 97.9 million euros. Besides the response variables, the dataset includes 36 feature variables characterizing 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 the vehicle, car brand, driving km per year, type of fuel, number of vehicle seats).

We have implemented a data preprocessing framework including data cleaning, data pre-processing, feature selection and engineering, outlier treatment, and dimensionality reduction. 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. The final dataset used for model calibration consists of 2 464 181 observation duration and 21 feature variables (Table 1).

The response variable of the frequency model is *Claim Count*, a discrete quantitative variable representing the number of claims reported per policy, per year, considering the exposure-to-risk in each year (fraction of the year in which the policy was in force) (Ohlsson and Johansson, 2010). Both the response variable and the exposure-to-risk were directly collected from the dataset. During the study period, the average annual claim rate was 4.8%. and only 3.07% of the contracts reported claims. For each year of observation, all policies in force in that year were included, with all feature variables properly updated.

During the pre-processing stage, we analysed the relationship between the response variable and the feature variables. For instance, Figure 2 represents the claim frequency according to the driver's age. The average driver's age in the insurance portfolio is 51 years, with ages ranging between 18 and 93 years. Figure 1 shows that the claim frequency is significantly distinct between age groups, with the peak at 21 years of age, then declining with age.

Variable Name	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 x year, 2 x year, 4 x year, 12 x year	Premium frequency, i.e., number of payments per year
Agent Delegation	22 different levels, from P_D1 to P_D22	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 O_M1 to O_M708, unknown	Vehicle Brand
Vehicle Seats	2, 3, 4, 5, 6, 7, 8, 9, 11+, 999	# of seats in the vehicle
Engine Capacity (cc)	32 levels (1-50, …, 1000-1100, …, 5000+)	Engine capacity of the vehicle
Horse Power (hp)	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 (€)	14 levels (<7000,…,25000-30000,… 500000+)	Initial value of the vehicle, as if it was new.
Fuel	8 distinct levels, from O_F1 to O_F2, without fuel, other, unknown	Type of fuel
District	22 different levels, from O_DC1 to O_DC22, unknown	The policyholder's (usual driver) District of residence
Bonus-Malus Level	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 (years)	0-17, 18 to 85 (individually), 85+, unknown	Age of the usual driver
Cover Capital (€)	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).

Table 1 – Final feature variables list for each policy of the dataset.

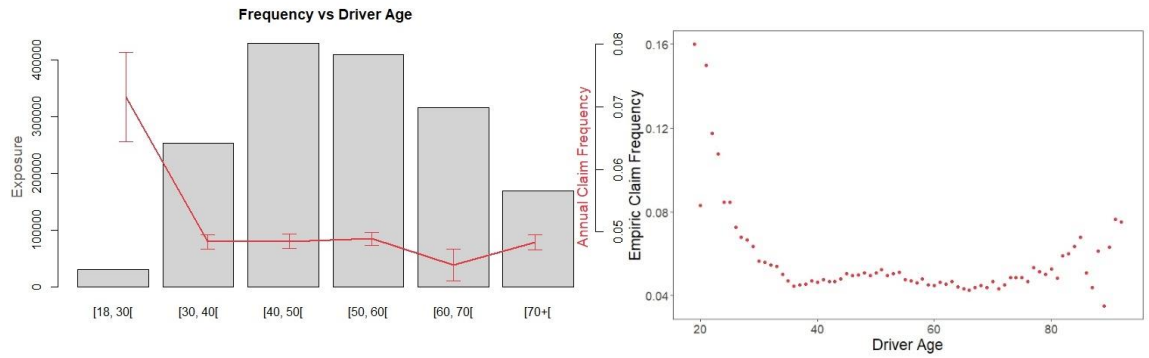


Figure 1 - Claim frequency vs Driver Age - individual (left panel) and by group ages (right panel)

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 considering 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. The analysis was performed using R Studio (v4.0.2), applying the *gbm* package. We use a nested k-fold cross-validation resampling procedure for evaluating and comparing the learning algorithms, checking for overfitting, and tunning the machine learning hyperparameters considering values in the range $k \in [1,2,\dots,6]$. This process consists of a double loop of cross-validation: the inner loop serves for tuning the hyperparameters, whereas the independent outer loop serves for assessing the quality of the model. The inner loop is composed of k_1 folds and the outer loop of k_2 folds, with the total number of trained models equal to $k_1 k_2$. 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., 2008). The fold partition used in this paper represents a compromise between the objective of reducing the generalization error and the computational burden (Boehmke & Greenwel, 2020).

To calibrate the boosting and the decision tree-specific hyperparameters, we have adopted a grid search procedure (Su & Bai, 2020). This technique develops a model for each possible combination of all hyperparameters, searching for the architecture that generates the best results. Specifically, we calibrate the GBM algorithm for the number of decision trees accounting for overfitting with $N \in \{100, 250, 400, 500, 750, 1000\}$, controlling for different values of the learning rate (shrinkage factor) $\lambda \in \{0.1, 0.05, 0.01\}$. Regarding the tree-specific hyperparameters, we have investigated multiple combinations controlling for the tree depth, with 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, the grid search procedure investigated 360 different models to find the

GBM optimal parameters. Table 2 summarises the optimal set of parameters for each of the six folds tested in the claim frequency model. They represent the combination that generated the smallest cross-validation iteration error in that fold (the out-of-sample Poisson deviance). We can observe that the maximum number of optimal trees is achieved for smaller values of shrinkage factor, a well-known behaviour between these parameters identified in similar studies using GBMs.

# Fold	# Trees	Shrinkage factor	Interaction Depth	Bag Fraction	Poisson Deviance
1	37	0.1	2	0.95	0.2802844
2	64	0.1	2	0.95	0.2802088
3	642	0.01	2	0.80	0.2802078
4	116	0.1	1	0.95	0.2793700
5	**239**	**0.05**	**2**	**0.95**	**0.2791459**
6	47	0.1	4	0.95	0.2796919
Average	**190**	**0.077**	**2**	**0.925**	**0.2798181**

Table 2 - Optimal tunning parameters per fold.

Note: Optimal tunning parameters and out-of-sample Poisson deviance (Models with smaller deviance are better) estimated considering random samples of 50000 observations extracted from a 1.97 million observations training set used for calibrating the frequency model.

4. RESULTS

4.1. Model Performance

We use 80% of the 1.97 million observations as training data and the remaining 20% as testing data to estimate both the optimal GBM model and the benchmark standard GLM model. The latter was developed using the EMBLEM software (by Willis Towers Watson) following the standard forward stepwise procedure of variable selection, to optimize ratemaking to each risk profile. Table 3 reports the in-sample and the out-of-sample loss for the competing models. The results for both samples show that the gradient boosting model exhibits lower deviance when compared to the classical GLM model, thus contributing to increasing the predictive accuracy in auto insurance claim frequency modelling.

Sample	Poisson Deviance	
	GLM	GBM
Training (80%)	432 449	428 621
Testing (20%)	107 896	106 773
All (100%)	540 362	535 685

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

4.2. Model Interpretation

In this subsection, we use two important tools, variable importance measures, and partial dependence plots (Friedman, 2008) to interpret the GBM model. Figure 2 shows the variable importance scores for the optimal GBM model, taking, for each fold, the average over all trees and discarding features with variable importance lower than 0.1%. This measure is based on the number of times a variable is selected for splitting in a decision tree, with influences averaged over all trees and standardised so that they add up to 100%,

The results show, for each cross-validation number of folds, that the policyholder's (usual driver's) district of residence, the bonus-malus system (BMS), 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 to predict 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.

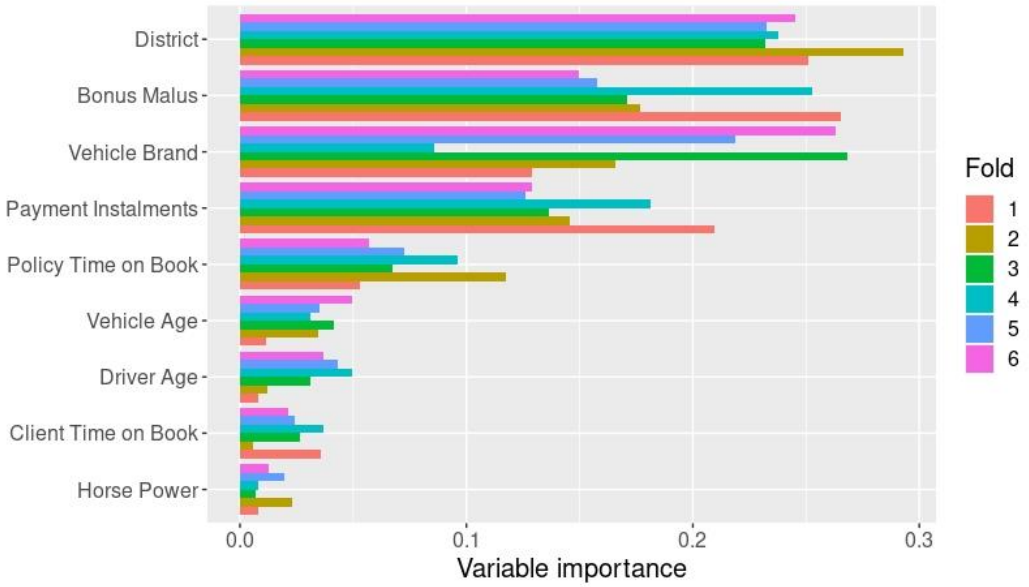


Figure 2 - Variable importance in the optimal GBM per data fold, claim frequency model

Table 4 summarises the list of features selected by both the GBM model and the benchmark GLM model. The results show that, out of these nine main variables identified as important by the GBM model (with a minimum of 0.1% of variable importance score per fold), only the Bonus-Malus System 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 penalizing at-fault accidents with premium surcharges are a powerful incentive for safe driving. However, it is also well-known

that BMS systems can also encourage non-reporting of claims in order to avoid the malus. Because of this, some forms of BMS introduce varying (escalating) deductibles that prevent malus evasion.

GLM	GBM
District	District
	Bonus-Malus System
Vehicle Brand	Vehicle Brand
Payment Instalments	Payment Instalments
Policy Time on Book	Policy Time on Book
Vehicle Age	Vehicle Age
Driver's Age	Driver's Age
	Client Time on Book
Horse Power	Horse Power
Fuel	
Years of Driving Licence	

Table 4 - Variables included in the frequency models: GLM and GBM

Although the findings show that most data features are present in both prediction models considered in this study, the results suggest that the Gradient Boosting approach has a slightly higher ability to select the feature variables that best differentiate the frequency of claims in TPL auto insurance.

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 affects the response. However, in decision-tree GBMs, visualization tools such as partial dependence plots and individual conditional expectation (ICE) plots can be used to visualize the effect of the predictor on the modeled response (claim frequency) after marginalizing out the remaining explanatory variables. Partial dependence plots exhibit the average effect of a feature, whereas ICE plots disaggregate the averaged data, providing a method of inspecting how the instance's prediction changes when a feature varies.

In Figure 3, we depict the graphical representation of the partial dependence effect of the policyholder's (usual driver's) district of residence on 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.

In Figure 4 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 (Goldstein et al. 2015). Figure 4 shows that the ICE lines tend to follow the same trend as the average. However, the

overlapping crossing lines observed for vehicle brands number 15, 46, 47, 75, and 102 deviating from the average suggest a possible interaction between the vehicle brand and another feature.

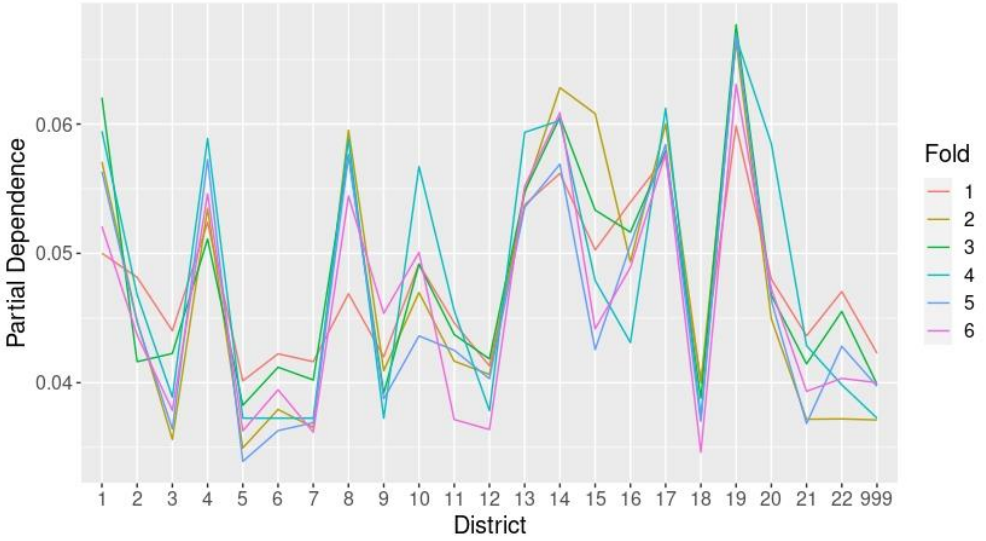


Figure 3 - Partial dependence plot of the policyholder's district of residence on claim frequency, per data fold, using a sample of 1 000 observations

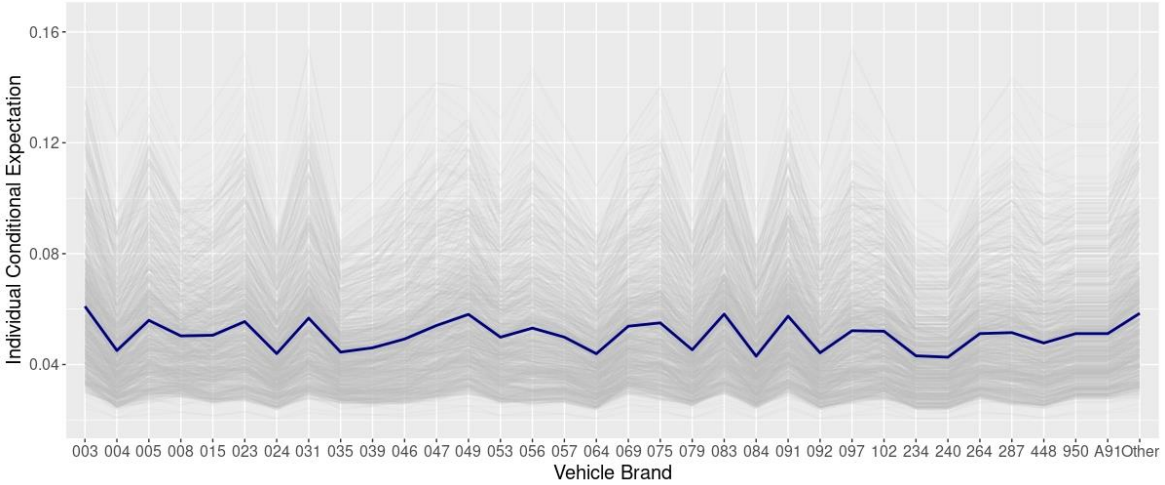


Figure 4 - Effect of the Vehicle Brand on the frequency model as Partial Dependence (in dark blue) and Individual Conditional Expectation (in grey), considering 5-fold cross-validation

To further check for interactions between variables, Friedman's H-statistic (Friedman and Popescu, 2008) was estimated for all possible combinations. By definition, the statistic ranges 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. Table 5 reports Friedman's H-Statistic 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 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 vehicle age (H-statistic = 0.1560). As a result of this analysis, the interaction between the policyholder's district of residence and the policy seniority was included in the GLM model.

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

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

Figure 5 shows the effect of the feature vehicle brand on claim frequency as partial dependence (in dark blue) and the individual conditional expectation (in grey), considering data fold 5, grouped by Payment Instalments. The plot suggests that for brands 24, 49, 64, 92, and 448 the claim frequency risk associated to 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.

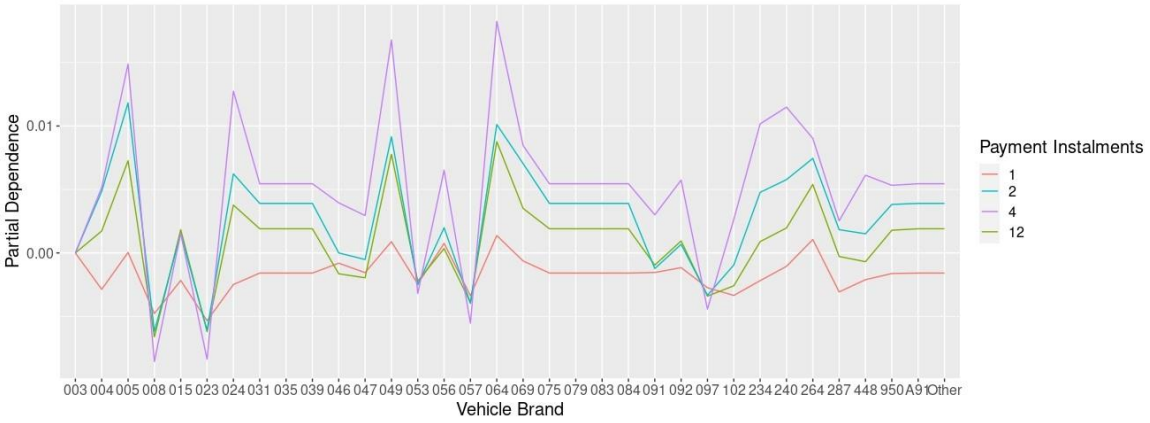


Figure 5 - Effect of the Vehicle Brand on the frequency model as Partial Dependence (in dark blue) and Individual Conditional Expectation (in grey), considering data fold 5

5. DISCUSSION AND CONCLUSIONS

The insurance field has registered significant growth over the last decades in several countries, and with that comes the rise of competitiveness. It is in every insurance company's best interest to make sure that their prices correctly reflect the risks they are underwriting, which is partially done through the development and implementation of fair pricing models. Modelling claim frequency is a critical task in ratemaking in property-casualty insurance.

The gradient boost algorithm combines models with 'poor' performance (high prediction error), such as regression or classification trees, to produce a highly accurate prediction rule with easily interpretable results. The adoption of gradient boosting models with decision trees as base learners in auto insurance ratemaking facilitates the breakdown of a portfolio of policyholders into homogeneous risk profiles based on some feature variables and promotes transparency and intra-group risk pooling under common asymptotic (group size) conditions.

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. The findings show that GBM may offer higher accuracy (i.e., lower deviance) when compared to classical GLM.

An interesting result of the claim frequency model is that the most important risk factors in the gradient boosting machines are those selected in the corresponding GLMs. These results are consistent with that obtained by Henckaerts et al. (2021) using a portfolio of motor Third Party Liability from a Belgian insurer in 1997. This means that Gradient Boosting results could also be used to aid the selection of the candidate variables (and their complex interactions) to consider in fitting the GLM, by setting a starting point that most likely includes the most significant variables. The analysis of partial dependence plots and individual conditional expectation plots offers additional insight into a selection of noteworthy effects for the claim frequency model. Further research should investigate the performance of GBM against GLM for claim frequency considering overdispersion distributions (e.g., Negative Binomial) as well as for claim severity. Comparing 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) is also left for further research.

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).

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 suggestions made to an earlier version of this paper.

REFERENCES

Ashofteh, A., & Bravo, J. M. (2021). A Conservative Approach for Online Credit Scoring. *Expert Systems With Applications*, 176, 114835.

Ashofteh, A., Bravo, J. M., & Ayuso, M. (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.

Ayuso, M., Bravo, J. M., Holzmann, R. & Palmer, E. (2021). Automatic indexation of pension age to life expectancy: When policy design matters. *Risks*, 9(5), 96.

Baecke, P. & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, 98, 69-79.

Boehmke, B. & Greenwel, B. (2020). Hands-on Machine Learning with R. Taylor and Francis.

Bravo, J. M. (2022). Pricing Participating Longevity-Linked Life Annuities: A Bayesian Model Ensemble approach. *European Actuarial Journal*, 12, 125–159.

Bravo, J. M., & Ayuso, M. (2021). Linking Pensions to Life Expectancy: Tackling Conceptual Uncertainty through Bayesian Model Averaging. *Mathematics,* 9(24), 3307.

Bravo, J. M., Ayuso, M., Holzmann, R. & Palmer, E. (2021). Addressing the Life Expectancy Gap in Pension Policy. *Insurance: Mathematics and Economics*, 99, 200-221.

Bravo, J. M., Ayuso, M., Holzmann, R., & Palmer, E. (2023). Intergenerational Actuarial Fairness when Longevity Increases: Amending the Retirement Age. Insurance: Mathematics and Economics. (available at https://doi.org/10.1016/j.insmatheco.2023.08.007).

Breiman, L. (2001) Random Forests. *Machine Learning*, 45, 5–32.

Cunha, L. & Bravo, J. M. (2022). Automobile Usage-Based-Insurance: Improving Risk Management using Telematics Data. *CISTI'2022 - 17th Iberian Conference on Information Systems and Technologies*. (available at https://doi.org/10.23919/CISTI54924.2022.9820146).

Czado, C., Kastenmeier, R., Brechmann, E. C., & Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 2012(4), 278-305.

Fauzan, M. A., & Murfi, H. (2018), The Accuracy of XGBoost for Insurance Claim Prediction, *International Journal of Advances in Soft Computing and Its Applications*, *10*(2), 159–171.

Frees, E. W., Gao, J., & Rosenberg, M. A. (2011). Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal*, 15(3), 377–392.

Frees, E. W., Shi, P., & Valdez, E. A. (2009). Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin: The Journal of the IAA*, 39(1), 165–197.

Friedman, J. (2001). Greedy boosting approximation: a gradient boosting machine. *Annals of Statistics,* 29, 1189–1232.

Friedman, J. H, & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2 (3), 916-954.

Garrido, J., Genest, C. & Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70, 205-215.

Goldstein, A., Kapelner, A. Bleich, J. & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics,* 24 (1), 44-65.

Gschlößl, S., & Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 3, 202–225.

Hanafy, M., & Ming, R. (2021). Machine learning approaches for auto insurance big data. *Risks*, 9(2), 1–23.

Henckaerts, R., Côté, M. P., Antonio, K., & Verbelen, R. (2021). Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods. *North American Actuarial Journal,* 25, 255–85.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79–87.

Jose, V. R. R., &Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, 24(1), 163–169.

Katrien, A. Valdez, E. (2011). Statistical Concepts of a Priori and a Posteriori Risk Classification in Insurance. *Advances in Statistical Analysi*s, 96 (2), 187–224.

Kim, D. & Baek, Jun-Geol (2022). Bagging ensemble-based novel data generation method for univariate time series forecasting. *Expert Systems with Applications*, 203, 117366.

Meng, S., Gao, Y. & Huang, Y. (2022). Actuarial intelligence in auto insurance: Claim frequency modeling with driving behavior features and improved boosted trees. *Insurance: Mathematics and Economics*, 106,115-127.

Noll, A., Salzmann, R., & Wüthrich, M. v. (2020). Case Study: French Motor Third-Party Liability Claims. *SSRN Eletronic Journal*, 1–41.

Ohlsson, E.; Johansson, B. (2010), Non-Life Insurance Pricing with Generalized Linear Models, 2nd ed. Berlin: Springer.

Ortega, J., Koppel, M., & Argamon, S. (2001). Arbitrating among competing classifiers using learned referees. *Knowledge and Information Systems*, 3(4), 470–490.

Paefgen, J., Staake, T. & Thiesse, F. (2013). Evaluation and aggregation of pay-as-you-drive insurance rate factors: a classification analysis approach. *Decision Support Systems*, 56, 192-201.

Pesantez-Narvaez, J., Guillen, M. & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks*, 7 (2), p. 70.

Qian, W., Yang, Y., & Zou, H. (2016). Tweedie's Compound Poisson Model with Grouped Elastic Net. *Journal of Computational and Graphical Statistics*, 25(2), 606–625.

Quan, Z. and Valdez, E. A. (2018) Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling*, *6*(1), 377–407.

Raftery, A., Madigan, D. & Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association,* 92, 179–91.

Renshaw, A.E. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin*, 24(2), 265-285.

Sergio, A., Lima, T. & Ludermir, T. (2016). Dynamic selection of forecast combiners. *Neurocomputing*, 218, 37-50.

Shi, P. (2016) Insurance ratemaking usibng a copula-based multivariate Tweedie model. *Scandinavian Actuarial Journal*, 2016(3), 198-215, doi: 10.1080/03461238.2014.921639

Shi, P., Feng, X., Ivantsova, A. (2015). Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64, 417–428.

Shu, C., & Burn, D. H. (2004). Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resources Research* 40, 1–10.

Staudt, Y. and Wagner, J. (2021) Assessing the performance of random forests for modeling claim severity in collision car insurance. *Risks*, *9*(3), 53.

Su, X., Bai, M. (2020) Stochastic gradient boosting frequency-severity model of insurance claims. PLoS ONE 15(8): e0238000. (available at https://doi.org/10.1371/journal.pone.0238000).

Verbelen, R., Antonio, K. & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. Journal of the Royal Statistical Society. Series C. *Applied Statistics,* 67(5), 1275-1304.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2), 241–259.