PLOS ONE

RESEARCH ARTICLE

Stochastic gradient boosting frequency-severity model of insurance claims

Xiaoshan Su*, Manying Bai

Department of Finance, Beihang University, Beijing, China

* suxiaoshan@buaa.edu.cn

Abstract

The standard GLM and GAM frequency-severity models assume independence between the claim frequency and severity. To overcome restrictions of linear or additive forms and to relax the independence assumption, we develop a data-driven dependent frequency-severity model, where we combine a stochastic gradient boosting algorithm and a profile likelihood approach to estimate parameters for both of the claim frequency and average claim severity distributions, and where we introduce the dependence between the claim frequency and severity by treating the claim frequency as a predictor in the regression model for the average claim severity. The model can flexibly capture the nonlinear relation between the claim frequency (severity) and predictors and complex interactions among predictors and can fully capture the nonlinear dependence between the claim frequency and severity. A simulation study shows excellent prediction performance of our model. Then, we demonstrate the application of our model with a French auto insurance claim data. The results show that our model is superior to other state-of-the-art models.

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

Editor: Feng Chen, Tongii University, CHINA

Received: June 12, 2020

Accepted: August 6, 2020

Published: August 31, 2020

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The research was supported by National Natural Science Foundation of China (Grant: 71333014, 71571007).

Competing interests: The authors have declared that no competing interests exist.

Introduction

Insurance claims modeling is a topic of great concern in non-life insurance. The model helps an insurer accurately estimate potential loss and make appropriate actuarial decisions. Specifically, the model enables an insurer to set a fair premium for each individual policy. It is important to charge the policyholder with a fair premium. For instance, Dionne, Gouriéroux, and Vanasse [1] point out that in auto insurance, if an insurer charges too little for young drivers and too much for old drivers, young drivers will be attracted while old drivers will switch to competitors. Then, the insurer loses profitable and gain underpriced policies, both resulting in economic losses. Further, the model helps the insurer determine a suitable level of risk capital. The underestimation of loss can make the insurer not hold enough risk capital and hence raise bankruptcy risk. In contrary, the overestimation can reduce liquid capital of the insurer and then hamper business expansion. Thus, an accurate model of insurance claims is significant to competency and profits of an insurer.

The frequency-severity model is a standard model of insurance claims, which separately models the claim frequency and average claim severity. The claim frequency examines the number of claims and the average claim severity takes account of the average amount of claims

conditional on occurence. The claim frequency and severity depend on characteristics of an individual policy. For instance, in auto insurance, the characteristics include the age, gender and motor vehicle record points of the policyholder, per capital income and population density of the policyholder's residential area, age and model of the vehicle, etc. Thus, there is a need of predictive models. The traditional frequency-severity model uses generalized linear models (GLM) for modeling the claim frequency and severity. The frequency part often uses Poisson or negative binomial regressions and the severity part uses gamma or inverse Gaussian regressions. There is a large literature extending the model to capture different features of the data. For instance, multivariate models can give a joint analysis of the frequency or severity at different levels of classification. Anastasopoulos, Shankar, Haddock, and Mannering [2] introduce a multivariate Tobit model to study accident rates categorized by severities. The conditional autoregressive model can be used for accommodating spatial correlation. Huang, Song, Xu, Zeng, Lee, and Abdel-Aty [3] develop a macro-level Bayesian spatial model with conditional autoregressive prior and a micro-level Bayesian spatial joint model for predicting the claim frequency. Zeng, Wen, Wong, Huang, Guo, and Pei [4] use a bivariate conditional autoregressive model to simultaneously analyze daytime and nightime claim frequencies. Aguero-Valverde [5] introduces a multivariate conditional autoregressive model to estimate excess claim frequencies at different severity levels. Generalized linear mixed models and the other random parameters models can be used to capture unobserved heterogeneity across observations. Barua, El-Basyouny, and Islam [6] develop a multivariate random parameters conditional autoregressive model to predict claim frequencies. Zeng, Wen, Huang, Pei, and Wong [7] propose a multivariate random parameters Tobit model to analyze accident rates by severity. Zeng, Guo, Wong, Wen, Huang, and Pei [8] introduce a multivariate random parameters spatio-temporal Tobit model to accommodate spatio-temporal correlation and interaction. Dong, Ma, Chen, and Chen [9] use a mixed logit model to examine the difference of single-vehicle and multivehicle accident probabilities. Chen, Chen, and Ma [10] adopt a mixed logit model to analyze the hourly accident probability for highway segments. Chen, Song, and Ma [11] develop a random parameters bivariate ordered probit model to investigate the injury severity of the two drivers involved in the same crash.

However, there are two major limitations in the frequency-severity model. First, the model has a linear predictor form. In practice, there are nonlinear effects from predictors. For instance, in auto insurance, the nonlinear relation between the claim severity and the insured's age is well documented (Frees, Shi, and Valdez [12]). Generalized additive models (GAM) developed in Hastie and Tibshirani [13] and popularized by Wood [14] overcome the restrictive linear form by modeling continuous variables with smooth functions estimated from data. However, the additive form of GAM models can't automatically capture complex interactions among predictors. Though interaction terms can be manually added to the structure of the model, identifying interactions terms can be tedious when many predictors are involved. Missing important interactions can reduce prediction accuracy. Second, the standard frequency-severity model assumes an independent relation between the claim frequency and severity. However, in practice, the claim frequency and severity are often dependent. For instance, in auto insurance, the claim frequency and severity are often negatively correlated (Gschlößl and Czado [15]). Home insurance claims due to natural hazard such as earthquake or flood are both large and frequent in affected areas. Frees, Gao, and Rosenberg [16] also point out that the claim frequency has a significant effect on the claim severity for outpatient expenditures. Gschlößl and Czado [15], Frees, Gao, and Rosenberg [16], Erhardt and Czado [17], Shi, Feng, and Ivantsova [18] and Garrido, Genest, and Schulz [19] capture the dependence between the claim frequency and severity by treating the claim frequency as a predictor variable in the regression model for the average claim severity. Shi, Feng, and Ivantsova [18] show that the

predictor method applied to the GLM frequency-severity model can only measure a linear relation between the claim frequency and severity. Czado, Kastenmeier, Brechmann, and Min [20], Krämer, Brechmann, Silvestrini, and Czado [21] and Shi, Feng, and Ivantsova [18] model the joint distribution of the claim frequency and average claim severity through copulas. However, popular copulas, such as elliptical and Archimedean copulas, can only capture the symmetric or limited dependence structures. The multivariate conditional autoregressive model (Aguero-Valverde [5]) with its random parameters version (Barua, El-Basyouny, and Islam [6]) and the multivariate Tobit model (Anastasopoulos, Shankar, Haddock, and Mannering [2]) with its random parameters version (Zeng, Wen, Huang, Pei, and Wong [7]) and its random parameters spatio-temporal version (Zeng, Guo, Wong, Wen, Huang, and Pei [8]) accommodate the correlation between the claim frequency and severity by modeling claim frequencies or accident rates at different severity levels. But the usage of finitely many severity levels only partially captures the dependence between the claim frequency and severity. Thus, there is a need to develop a data-driven dependent frequency-severity model, which can learn the optimal model structure from the data and can flexibly capture the nonlinear dependence between the claim frequency and severity.

Boosting is one of the most successful ensemble learning methods, which combines a large number of weak prediction models (weak learners) in an additive form to enhance prediction performance. The seminal work is Freund and Schapire [22], which introduce a boosting algorithm named AdaBoost for classification. Breiman et al. [23] and Breiman [24] observe an intrinsic connection between the AdaBoost algorithm and the functional gradient descent algorithm. Friedman, Hastie, Tibshirani, et al. [25] reveal another important fact that the AdaBoost and other boosting algorithms are additive models, i.e., an additive combination of weak learners. Then, they propose a general boosting algorithm named gradient boosting for both of classification and regression. The algorithm can be viewed as an estimation method for an additive model that combines weak learners. From this new perspective, many boosting regression models are developed. They are different in forms when different loss functions, weak learners or optimization schemes are used. Friedman, Hastie, and Tibshirani [26] and Friedman [27, 28] develop boosting regression models with the least-squares, least absolute deviation and Huber loss functions. Ridgeway [29, 30] propose the boosting Poisson regression and boosting proportional hazards regression models. Kriegler and Berk [31] introduce the boosting quantile regression model. In actuarial literature, Noll, Salzmann, and Wuthrich [32] show that the boosting Poisson regression model performs better than the GLM model in predicting the claim frequency. Yang, Qian, and Zou [33] develop a gradient boosting Tweedie compound Poisson model, where they use a profile likelihood approach to estimate the index and dispersion parameters. They show that the model makes more accurate premium prediction than GLM and GAM Tweedie compound Poisson models. In order to cope with extremely unbalanced zero-inflated data, Zhou, Yang, and Qian [34] introduce a gradient boosting zero-inflated Tweedie compound Poisson model by using a similar method. In fact, the method that combines the gradient boosting algorithm and profile likelihood approach can be used to develop any gradient boosting exponential family regression models. Sigrist and Hirnschall [35] apply the method to develop a gradient boosting Tobit model for predicting defaults on loans made to Swiss small and medium-sized companies. They show that the model outperforms other state-of-the-art models in predictive performance.

In this paper, we apply the method to develop a gradient boosting frequency-severity model (D-FSBoost). We illustrate the model with a Poisson distribution for modeling the claim frequency and with a gamma distribution for modeling the claim severity. We use the profile

likelihood approach to estimate the dispersion parameter in the gamma distribution. The gradient boosting frequency-severity model with other exponential family distributions for modeling the claim frequency and severity can be developed in the same manner. Following Gschlößl and Czado [15], Frees, Gao, and Rosenberg [16], Erhardt and Czado [17], Shi, Feng, and Ivantsova [18] and Garrido, Genest, and Schulz [19], we capture the dependence between the claim frequency and severity by treating the claim frequency as a predictor in the regression model for the average claim severity. Since the gradient boosting gamma regression model can learn the optimal model structure from the data, the D-FSBoost model can fully capture the nonlinear dependence between the claim frequency and severity. The D-FSBoost model inherits all advantages of boosting models, such as the data-driven model structure, high prediction accuracy, automatic feature selection and high capacities of avoiding overfitting problems, etc. In a simulation study, we demonstrate that the D-FSBoost model can flexibly capture the nonlinear relation between the claim frequency (severity) and predictors and complex and higher order interactions among predictors and can fully capture the nonlinear dependence between the claim frequency and severity. We compare the D-FSBoost model with GLM and GAM frequency-severity models and show that the D-FSBoost model can make more accurate prediction in claim frequency and severity distributions. We apply the D-FSBoost model to analyze a French auto insurance claim data. We provide further evidence on the dependence between the claim frequency and severity and indicate that the frequency-severity model can be significantly improved by taking the claim frequency as a predictor in the regression model for the average claim severity. We also show that the D-FSBoost model is superior to other state-of-the-art models in prediction of pure premium.

The rest of this paper is organized as follows. In section 2, we review the gradient boosting algorithm and introduce the D-FSBoost model. In section 3, we show high prediction accuracy of the model in a simulation study. Finally, in section 4, we apply the model to analyze a French auto insurance claim data.

Stochastic gradient boosting frequency-severity model

In this section, we introduce the stochastic gradient boosting algorithm. Then, we show the implementation of the D-FSBoost model.

Stochastic gradient boosting

In this subsection, we briefly review the stochastic gradient boosting algorithm in Friedman [28]. Denote by $\mathbf{x} = (x_1, \ldots, x_p)$ the set of predictors and y the response variable. Given a training sample $\{y_i, \mathbf{x}_i\}_{i=1}^{d}$ and a loss function $\Psi(y, f(\mathbf{x}))$, the algorithm estimates the optimal prediction function $\hat{f}(\mathbf{x})$ by minimizing loss over the training sample,

$$\hat{f}(\mathbf{x}) = \arg\min_{f(\mathbf{x})} \sum_{i=1}^{d} \Psi(y_i, f(\mathbf{x}_i)), \tag{1}$$

where $f(\mathbf{x})$ is constrained to a form of a sum of weak learners as

$$f(\mathbf{x}) = h(\mathbf{x}; \mathbf{a}_0) + \sum_{m=1}^{M} \beta_m h(\mathbf{x}; \mathbf{a}_m), \tag{2}$$

where $h(\mathbf{x}; \mathbf{a}_m)$ is a weak learner with a parameter vector $\mathbf{a}_m$, $\beta_m \in \mathbb{R}$ is an expansion coefficient, M is the number of weak learners.

The algorithm estimates the function $\hat{f}(\mathbf{x})$ in a forward stagewise manner. Let the constant $f_0(\mathbf{x})$ be an initial estimate of $\hat{f}(\mathbf{x})$ as

$$f_0(\mathbf{x}) = h(\mathbf{x}; \mathbf{a}_0) = \arg\min_{\rho} \sum_{i=1}^{d} \Psi(y_i, \rho). \tag{3}$$

Denote by $f_{m-1}(\boldsymbol{x})$ the estimate of $\hat{f}(\mathbf{x})$ at the $(m-1)^{\text{th}}$ step. Then, at the m^{th} step, the algorithm randomly selects a subsample of size $\tilde{d} < d$, $\{\tilde{y}_i, \tilde{\mathbf{x}}_i\}_1^{\tilde{d}}$, computes the negative gradient

$$\tilde{z}_i = -\frac{\partial \Psi(\tilde{y}_i, f(\tilde{\mathbf{x}}_i))}{\partial f(\tilde{\mathbf{x}}_i)}\bigg|_{f(\tilde{\mathbf{x}}_i)=f_{m-1}(\tilde{\mathbf{x}}_i)}, \tag{4}$$

and then fits the weak learner $h(\mathbf{x}; \mathbf{a}_m)$ by minimizing the following least square sum

$$\mathbf{a}_m = \arg\min_{\mathbf{a}} \sum_{i=1}^{\tilde{d}} (\tilde{z}_i - h(\tilde{\mathbf{x}}_i; \mathbf{a}))^2. \tag{5}$$

The optimal value of β_m is determined by

$$\beta_m = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{\tilde{d}} \Psi(\tilde{y}_i, f_{m-1}(\tilde{\mathbf{x}}_i) + \boldsymbol{\beta} h(\tilde{\mathbf{x}}_i; \mathbf{a}_m)). \tag{6}$$

Then, the current estimate of $\hat{f}(\mathbf{x})$ is updated as

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + v\boldsymbol{\beta}_m h(\mathbf{x}; \mathbf{a}_m), \tag{7}$$

where $0 < v \leq 1$ is the shrinkage factor that controls the learning rate. Friedman [27] points out that small v reduces overfitting and enhances predictive performance.

The algorithm reduces to a standard gradient boosting algorithm when the full sample is used at each iteration in place of the randomly selected subsample. Friedman [28] shows that the stochastic gradient boosting algorithm has a faster computation speed and higher prediction accuracy.

The D-FSBoost model

In this subsection, we introduce the dependent frequency-severity model. Then, we estimate mean parameters by using the stochastic gradient boosting algorithm.

In the frequency-severity model, we model the claim frequency N with a Poisson distribution with the parameter $\lambda > 0$,

$$f_N(n|\lambda) = \frac{\lambda^n}{n!} e^{-\lambda} \quad \text{for} \quad n = 0, 1, 2, \ldots. \tag{8}$$

For $N > 0$, denote by

$$\tilde{Y} = \frac{Y_1 + \ldots + Y_N}{N} \tag{9}$$

the average claim severity, where Y_j is the j^{th} claim amount. We model the average claim

severity $\tilde{Y}$ conditional on N via a gamma distribution with parameters $\mu_N > 0$ and $\delta > 0$

$$f_{\tilde{Y}|N}(s|\mu_N, \delta) = \frac{1}{s\Gamma\left(\frac{1}{\delta}\right)}\left(\frac{s}{\mu_N\delta}\right)^{\frac{1}{\delta}} e^{-\frac{s}{\mu_N\delta}} \quad \text{for} \quad s > 0, \tag{10}$$

where we model the dependence between the claim frequency and severity by making the mean parameter μ_N depend on N.

Denote by $\boldsymbol{x}$ the vector of predictors representing characteristics of an individual policy. We assume that the parameters λ and μ_N are determined by the following two regression models:

$$\log(\lambda) = F_N(\boldsymbol{x}; \boldsymbol{\alpha}) \qquad \text{and} \qquad \log(\mu_N) = F_{\tilde{Y}|N}(\boldsymbol{x}, N; \boldsymbol{\beta}), \tag{11}$$

where log link functions are used, $F_N : \mathbb{R}^p \to \mathbb{R}$ and $F_{\tilde{Y}|N} : \mathbb{R}^p \times \mathbb{N} \to \mathbb{R}$ are two regression functions, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ denote the vector of parameters for F_N and $F_{\tilde{Y}|N}$, respectively. The functions F_N and $F_{\tilde{Y}|N}$ are restricted to linear and additive forms in GLM and GAM models, respectively. In our model, F_N and $F_{\tilde{Y}|N}$ are ensembles of weak learners.

For the time being, we assume that the dispersion parameter δ is given. We will estimate δ later. Then, we apply the stochastic gradient boosting algorithm to estimate the functions F_N and $F_{\tilde{Y}|N}$.

Denote by $\{n_i, s_i, \boldsymbol{x}_i\}$ the claim frequency, the average claim severity and the vector of predictors for the i^{th} policy, respectively. We consider θ independent insurance policies. Then, we have the log-likelihood function as follows:

$$\begin{aligned}\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta|\{n_i, s_i, \boldsymbol{x}_i\}_{i=1}^{\theta}) = & \underbrace{\sum_{i=1}^{\theta} \log \frac{e^{n_i F_N(\boldsymbol{x}_i;\boldsymbol{\alpha})}}{n_i!} e^{-e^{F_N(\boldsymbol{x}_i;\boldsymbol{\alpha})}}}_{l_1(\boldsymbol{\alpha})} \\ & + \underbrace{\sum_{i=1}^{\theta} \log \frac{1}{s_i\Gamma\left(\frac{1}{\delta}\right)}\left(\frac{s_i}{\delta e^{F_{\tilde{Y}|N}(\boldsymbol{x}_i, n_i;\boldsymbol{\beta})}}\right)^{\frac{1}{\delta}} e^{-\frac{s_i}{\delta e^{F_{\tilde{Y}|N}(\boldsymbol{x}_i, n_i;\boldsymbol{\beta})}}}}_{l_2(\boldsymbol{\beta}, \delta)}.\end{aligned} \tag{12}$$

Since maximizing the above log-likelihood function is equivalent to maximizing $l_1(\boldsymbol{\alpha})$ and $l_2(\boldsymbol{\beta}, \delta)$, respectively, we use negative log-likelihood functions $-l_1(\boldsymbol{\alpha})$ and $-l_2(\boldsymbol{\beta}, \delta)$ as loss functions and estimate the functions $F_N(\boldsymbol{x};\boldsymbol{\alpha})$ and $F_{\tilde{Y}|N}(\boldsymbol{x}, N; \boldsymbol{\beta})$ by minimizing loss over the sample $\{n_i, s_i, \boldsymbol{x}_i\}_{i=1}^{\theta}$,

$$\hat{f}(\mathbf{x}) = \arg\min_{f(\mathbf{x})} \sum_{i=1}^{\theta} \Psi_1(n_i, f(\mathbf{x}_i)) \qquad \text{and} \qquad \hat{g}(\mathbf{x}, N) = \arg\min_{g(\mathbf{x},N)} \sum_{i=1}^{\theta} \Psi_2(s_i, g(\mathbf{x}_i, n_i)), \tag{13}$$

where

$$\begin{cases} \Psi_1(n_i, f(\mathbf{x}_i)) & = -\log \dfrac{e^{n_i f(\mathbf{x}_i)}}{n_i!} e^{-e^{f(\mathbf{x}_i)}} \\ \Psi_2(s_i, g(\mathbf{x}_i, n_i)) & = -\log \dfrac{1}{s_i \Gamma\left(\dfrac{1}{\delta}\right)} \left(\dfrac{s_i}{\delta e^{g(\mathbf{x}_i, n_i)}}\right)^{\frac{1}{\delta}} e^{-\dfrac{s_i}{\delta e^{g(\mathbf{x}_i, n_i)}}} \end{cases}, \tag{14}$$

and the functions $f(\mathbf{x})$ and $g(\mathbf{x}, N)$ are confined to the form of a sum of weak learners as (2).

Then, the gradient boosting algorithm estimate $\hat{f}(\mathbf{x})$ and $\hat{g}(\mathbf{x}, N)$ in a forward stagewise manner. The initial estimates are computed as

$$\begin{cases} f_0(\mathbf{x}) = \arg\min_{\rho} \sum_{i=1}^{\theta} \Psi_1(n_i, \rho) \\ g_0(\mathbf{x}, N) = \arg\min_{\rho} \sum_{i=1}^{\theta} \Psi_2(s_i, \rho) \end{cases}. \tag{15}$$

Denote by $f_{m-1}(\mathbf{x})$ and $g_{m-1}(\mathbf{x}, N)$ the estimates of $\hat{f}(\mathbf{x})$ and $\hat{g}(\mathbf{x}, N)$ at the $(m-1)^{\text{th}}$ step, respectively. At the m^{th} step, the algorithm randomly selects a subsample of size $\tilde{\theta} < \theta$, $\{\tilde{n}_i, \tilde{s}_i, \tilde{\mathbf{x}}_i\}_{i=1}^{\tilde{\theta}}$, and computes the negative gradient

$$\begin{cases} \tilde{z}_i^f & = \tilde{n}_i - e^{f_{m-1}(\tilde{\mathbf{x}}_i)} \\ \tilde{z}_i^g & = \dfrac{\tilde{s}_i e^{-g_{m-1}(\tilde{\mathbf{x}}_i, \tilde{n}_i)} - 1}{\delta} \end{cases}. \tag{16}$$

Then, the algorithm fits weak learners $h^f(\mathbf{x}; \mathbf{a}_m^f)$ and $h^g(\mathbf{x}, N; \mathbf{a}_m^g)$ by minimizing the following least square sums,

$$\begin{cases} \mathbf{a}_m^f & = \arg\min_{\mathbf{a}} \sum_{i=1}^{\tilde{\theta}} (\tilde{z}_i^f - h^f(\tilde{\mathbf{x}}_i; \mathbf{a}))^2 \\ \mathbf{a}_m^g & = \arg\min_{\mathbf{a}} \sum_{i=1}^{\tilde{\theta}} (\tilde{z}_i^g - h^g(\tilde{\mathbf{x}}_i, \tilde{n}_i; \mathbf{a}))^2 \end{cases}. \tag{17}$$

We use K-terminal node regression trees as weak learners, i.e.,

$$\begin{cases} h^f(\mathbf{x}; \mathbf{a}_m^f) = \sum_{k=1}^{K} \hat{n}_k \mathbf{1}_{\{\mathbf{x} \in U_{k,m}\}} \\ h^g(\mathbf{x}, N; \mathbf{a}_m^g) = \sum_{k=1}^{K} \hat{s}_k \mathbf{1}_{\{(\mathbf{x}, N) \in V_{k,m}\}} \end{cases}, \tag{18}$$

where

$$\begin{cases} \hat{n}_k &= \dfrac{\sum_{i=1}^{\tilde{\theta}} \tilde{n}_i \mathbf{1}_{\{\tilde{\mathbf{x}}_i \in U_{k,m}\}}}{\sum_{i=1}^{\tilde{\theta}} \mathbf{1}_{\{\tilde{\mathbf{x}}_i \in U_{k,m}\}}} \\ \hat{s}_k &= \dfrac{\sum_{i=1}^{\tilde{\theta}} \tilde{s}_i \mathbf{1}_{\{(\tilde{\mathbf{x}}_i, \tilde{n}_i) \in V_{k,m}\}}}{\sum_{i=1}^{\tilde{\theta}} \mathbf{1}_{\{(\tilde{\mathbf{x}}_i, \tilde{n}_i) \in V_{k,m}\}}} \end{cases}, \tag{19}$$

and $\{U_{k,m}\}_{k=1}^{K}$ and $\{V_{k,m}\}_{k=1}^{K}$ are disjoint regions of $\mathbf{x}$ and $(\mathbf{x}, N)$ spaces, respectively, which represent terminal nodes of regression trees. In this case, the parameters $\mathbf{a}_m^f$ and $\mathbf{a}_m^g$ are splitting variables and split points of regression trees, which determine the regions $\{U_{k,m}\}_{k=1}^{K}$ and $\{V_{k,m}\}_{k=1}^{K}$. The optimization problem (17) is solved by a greedy algorithm with a least squared splitting criterion (Friedman [27]).

Once the weak learners $h^f(\mathbf{x}; \mathbf{a}_m^f)$ and $h^g(\mathbf{x}, N; \mathbf{a}_m^g)$ are obtained, the optimal expansion coefficients β_m^f and β_m^g are solved by

$$\begin{cases} \beta_m^f &= \arg\min_{\beta} \sum_{i=1}^{\tilde{\theta}} \Psi_1(\tilde{n}_i, f_{m-1}(\tilde{\mathbf{x}}_i) + \beta \sum_{k=1}^{K} \hat{n}_k \mathbf{1}_{\{\tilde{\mathbf{x}}_i \in U_{k,m}\}}) \\ \beta_m^g &= \arg\min_{\beta} \sum_{i=1}^{\tilde{\theta}} \Psi_2(\tilde{s}_i, g_{m-1}(\tilde{\mathbf{x}}_i, \tilde{n}_i) + \beta \sum_{k=1}^{K} \hat{s}_k \mathbf{1}_{\{(\tilde{\mathbf{x}}_i, \tilde{n}_i) \in V_{k,m}\}}) \end{cases}. \tag{20}$$

We can obtain the better estimation of $\hat{f}(\mathbf{x})$ and $\hat{g}(\mathbf{x}, N)$ by replacing a single expansion coefficient β_m^f (β_m^g) with the optimal coefficient $\gamma_{k,m}^f$ ($\gamma_{k,m}^g$), $k = 1, \ldots, K$ for each region $U_{k,m}$ ($V_{k,m}$), $k = 1, \ldots, K$. The optimal coefficients $\gamma_{k,m}^f$ ($\gamma_{k,m}^g$), $k = 1, \ldots, K$ are solved by

$$\begin{cases} \gamma_{k,m}^f &= \arg\min_{\gamma} \sum_{\tilde{\mathbf{x}}_i \in U_{k,m}} \Psi_1(\tilde{n}_i, f_{m-1}(\tilde{\mathbf{x}}_i) + \gamma) \\ \gamma_{k,m}^g &= \arg\min_{\gamma} \sum_{(\tilde{\mathbf{x}}_i, \tilde{n}_i) \in V_{k,m}} \Psi_2(\tilde{s}_i, g_{m-1}(\tilde{\mathbf{x}}_i, \tilde{n}_i) + \gamma) \end{cases}. \tag{21}$$

We have explicit solutions as follows:

$$\begin{cases} \gamma_{k,m}^f &= \log\left(\dfrac{\sum_{i=1}^{\tilde{\theta}} \tilde{n}_i \mathbf{1}_{\{\tilde{\mathbf{x}}_i \in U_{k,m}\}}}{\sum_{i=1}^{\tilde{\theta}} e^{f_{m-1}(\tilde{\mathbf{x}}_i)} \mathbf{1}_{\{\tilde{\mathbf{x}}_i \in U_{k,m}\}}} \right) \\ \gamma_{k,m}^g &= \log\left(\dfrac{\sum_{i=1}^{\tilde{\theta}} \tilde{s}_i e^{-g_{m-1}(\tilde{\mathbf{x}}_i, \tilde{n}_i)} \mathbf{1}_{\{(\tilde{\mathbf{x}}_i, \tilde{n}_i) \in V_{k,m}\}}}{\sum_{i=1}^{\tilde{\theta}} \mathbf{1}_{\{(\tilde{\mathbf{x}}_i, \tilde{n}_i) \in V_{k,m}\}}} \right) \end{cases}. \tag{22}$$

Then, the estimates of $\hat{f}(\mathbf{x})$ and $\hat{g}(\mathbf{x}, N)$ are updated as

$$\begin{cases} f_m(\mathbf{x}) &= f_{m-1}(\mathbf{x}) + \nu \sum_{k=1}^{K} \gamma_{k,m}^{f} \mathbf{1}_{\mathbf{x} \in U_{k,m}} \\ g_m(\mathbf{x}, N) &= g_{m-1}(\mathbf{x}, N) + \nu \sum_{k=1}^{K} \gamma_{k,m}^{g} \mathbf{1}_{(\mathbf{x},N) \in V_{k,m}} \end{cases}, \tag{23}$$

where we set $\nu = 0.03$ in our implementation.

The procedures are repeated M times. Then, we obtain $f_M(\mathbf{x})$ and $g_M(\mathbf{x}, N)$ as the final estimates.

The D-FSBoost algorithm is summarized as follows:

The D-FSBoost Algorithm

1. Initialize $f_0(\mathbf{x})$ and $g_0(\mathbf{x}, N)$,

$$\begin{cases} f_0(\mathbf{x}) = \arg\min_{\rho} \sum_{i=1}^{\theta} \Psi_1(n_i, \rho) \\ g_0(\mathbf{x}, N) = \arg\min_{\rho} \sum_{i=1}^{\theta} \Psi_2(s_i, \rho) \end{cases}. \tag{24}$$

2. For $m = 1$ to M do

1. Generate a random subsample $\{\tilde{n}_i, \tilde{s}_i, \tilde{\mathbf{x}}_i\}_{i=1}^{\tilde{\theta}}$.
2. Compute the negative gradient $(\tilde{z}_1^f, \ldots, \tilde{z}_{\tilde{\theta}}^f)$ and $(\tilde{z}_1^g, \ldots, \tilde{z}_{\tilde{\theta}}^g)$,

$$\begin{cases} \tilde{z}_i^f &= \tilde{n}_i - e^{f_{m-1}(\tilde{\mathbf{x}}_i)} \\ \tilde{z}_i^g &= \dfrac{\tilde{s}_i e^{-g_{m-1}(\tilde{\mathbf{x}}_i, \tilde{n}_i)} - 1}{\delta} \end{cases}, \ i = 1, \ldots, \tilde{\theta}. \tag{25}$$

3. K-terminal node regression trees fit two datasets $\{\tilde{z}_i^f, \tilde{\mathbf{x}}_i\}_{i=1}^{\tilde{\theta}}$ and $\{\tilde{z}_i^g, (\tilde{\mathbf{x}}_i, \tilde{n}_i)\}_{i=1}^{\tilde{\theta}}$ with a least squared splitting criterion and obtain the regions $\{U_{k,m}\}_{k=1}^{K}$ and $\{V_{k,m}\}_{k=1}^{K}$.
4. Compute the optimal coefficient for each region $U_{k,m}$ ($V_{k,m}$), $k = 1, \ldots, K$,

$$\begin{cases} \gamma_{k,m}^{f} &= \log\left(\dfrac{\sum_{i=1}^{\tilde{\theta}} \tilde{n}_i \mathbf{1}_{\{\tilde{\mathbf{x}}_i \in U_{k,m}\}}}{\sum_{i=1}^{\tilde{\theta}} e^{f_{m-1}(\tilde{\mathbf{x}}_i)} \mathbf{1}_{\{\tilde{\mathbf{x}}_i \in U_{k,m}\}}} \right) \\ \gamma_{k,m}^{g} &= \log\left(\dfrac{\sum_{i=1}^{\tilde{\theta}} \tilde{s}_i e^{-g_{m-1}(\tilde{\mathbf{x}}_i, \tilde{n}_i)} \mathbf{1}_{\{(\tilde{\mathbf{x}}_i, \tilde{n}_i) \in V_{k,m}\}}}{\sum_{i=1}^{\tilde{\theta}} \mathbf{1}_{\{(\tilde{\mathbf{x}}_i, \tilde{n}_i) \in V_{k,m}\}}} \right) \end{cases}, \ k = 1, \ldots, K. \tag{26}$$

5. Update the estimates of $\hat{f}(\mathbf{x})$ and $\hat{g}(\mathbf{x}, N)$ as

$$\begin{cases} f_m(\mathbf{x}) &= f_{m-1}(\mathbf{x}) + \nu \sum_{k=1}^{K} \gamma_{k,m}^{f} \mathbf{1}_{\mathbf{x} \in U_{k,m}} \\ g_m(\mathbf{x}, N) &= g_{m-1}(\mathbf{x}, N) + \nu \sum_{k=1}^{K} \gamma_{k,m}^{g} \mathbf{1}_{(\mathbf{x},N) \in V_{k,m}} \end{cases}. \tag{27}$$

end

3. Return $f_M(\mathbf{x})$ and $g_M(\mathbf{x}, N)$.

Estimating δ and choice of tuning parameters

We estimate the dispersion parameter δ using the profile likelihood approach. The D-FSBoost algorithm determines the value of $\boldsymbol{\beta}$ for each fixed δ. Denote by $\boldsymbol{\beta}_\delta$ the estimated value of $\boldsymbol{\beta}$. Then, the profile log-likelihood function for δ is given by

$$\tilde{l}(\delta) = l_2(\boldsymbol{\beta}_\delta, \delta). \tag{28}$$

The value of the dispersion parameter δ is obtained by maximizing the profile log-likelihood function $\tilde{l}(\delta)$:

$$\hat{\delta} = \arg\max_{\delta} \tilde{l}(\delta). \tag{29}$$

To reduce computations, we calculate $\hat{\delta}$ by doing a simple grid search over S grid points $\{\delta_1, \ldots, \delta_S\}$, i.e.,

$$\hat{\delta} = \arg\max_{\delta \in \{\delta_1, \ldots, \delta_S\}} \tilde{l}(\delta). \tag{30}$$

In the implementation, we need select tuning parameters, including the size of trees K and the number of trees M. The value of K controls the degree of the interaction among the predictors $\boldsymbol{x}$ or $(N, \boldsymbol{x})$. The appropriate value of M avoids over-fitting and improves out-of-sample prediction accuracy. We determine the parameters (K, M) via the cross-validation method. The k–fold cross-validation method splits the data into k equal-sized folds. Let $\kappa(i)$:$\{1, \ldots, \theta\} \rightarrow \{1, \ldots, k\}$ be an index function that indicates the fold to which the i^{th} observation is allocated by randomization. We calculate loss of the j^{th} fold data by using functions estimated by the remaining $k - 1$ folds. We repeat this procedure for $j = 1, \ldots, k$. Denote by $(\hat{f}_{-j}(\boldsymbol{x}, K, M), \hat{g}_{-j}(\boldsymbol{x}, N, K, M))$ the functions estimated with the j^{th} fold data removed and with the parameters (K, M). Then, the cross-validation estimate of loss is

$$\text{CV}(K, M) = \sum_{i=1}^{\theta} (\Psi_1(n_i, f_{-\kappa(i)}(\mathbf{x}_i, K, M)) + \Psi_2(s_i, g_{-\kappa(i)}(\mathbf{x}_i, n_i, K, M))). \tag{31}$$

The optimal (K, M) is obtained by minimizing the cross-validation estimate of loss

$$(\hat{K}, \hat{M}) = \arg\min_{K,M} \text{CV}(K, M). \tag{32}$$

Then, we use $(\hat{K}, \hat{M})$ in the D-FSBoost algorithm and finish all estimates.

Simulation study

In this section, we compare the D-FSBoost model with GLM and GAM frequency-severity models in two simulation experiments. We consider the models in the cases that the frequency and severity are independent and dependent. Denote by D-FSBoost, D-GAM and D-GLM the three models in the dependent case and by FSBoost, GAM and GLM the three models in the independent case. We compare the models in prediction accuracy of the claim frequency and severity distributions. We also investigate the impact of the value of δ on estimating $F_{\tilde{Y}|N}(\boldsymbol{x}, N; \boldsymbol{\beta})$ in the D-FSBoost model.

In simulation studies, we use one set of samples for training and another one for testing. Denote by $\{\hat{n}_i, \hat{s}_i, \hat{\boldsymbol{x}}_i\}_{i=1}^{\hat{\theta}}$ the testing sample with known true functions or parameters $\{F_N(\boldsymbol{x}; \boldsymbol{\alpha}), F_{\tilde{Y}|N}(\boldsymbol{x}, N; \boldsymbol{\beta}), \delta\}$. Let $\{\hat{f}(\boldsymbol{x}), \hat{g}(\boldsymbol{x}, N), \hat{\delta}\}$ be the functions or parameters estimated by the model. We use the out-of-sample loss and parameter estimation errors to measure prediction accuracy of the models. Table 1 shows the specific performance measures. In the FSBoost and D-FSBoost models, we use the five-fold cross-validation method to select parameters (K, M) among the combinations of $K \in \{1, 2, 3, 4, 5\}$ and $M \in \{100, 200, 300, 400, 500\}$ and search the optimal δ among 21 equally spaced values $\{1, 1.1, \ldots, 3\}$.

Simple case

In this subsection, we demonstrate that the D-FSBoost model can capture the nonlinear relation between the claim frequency (severity) and predictors, complex interactions among predictors, and the nonlinear dependence between the claim frequency and severity. The sample $\{n_i, s_i, \boldsymbol{x}_i\}_{i=1}^{\theta}$ is generated using the following specifications,

$$n_i \sim \text{Poi}(\lambda_i),\ \ s_i \sim \text{Gamma}(\mu_{n_i}, \delta),\ \ x_{ij} \sim \text{Unif}(0,1),\ \ i = 1, \ldots, \theta,\ \ j = 1, \ldots, 4, \tag{33}$$

where $\lambda_i = \exp(F_1(x_{i1}, x_{i2}))$, $\mu_{n_i} = \exp(F_2(\boldsymbol{x}_{i3}, \boldsymbol{x}_{i4}, n_i))$, $\delta = 2$, and

$$\begin{cases} F_1(x_{i1}, x_{i2}) &= \dfrac{\pi}{15}(3x_{i1}^2 + 2(1 - x_{i2})^2 + 10x_{i1}x_{i2}) \\ F_2(x_{i3}, x_{i4}, n_i) &= \ln(n_i + 3)e^{x_{i3}^2 - 2(1-x_{i4})^2} + \ln(n_i + 5)e^{\frac{1}{2}x_{i3}x_{i4}} \end{cases}. \tag{34}$$

We generate a sample of size 10000 for training and another one of size 10000 for testing. The out-of-sample loss and parameter estimation errors on the testing sample are shown in

Table 1. Performance measures.

Measure	Description	Formula
Frequency Loss	Out-of-sample loss for the claim frequency	$\sum_{i=1}^{\hat{\theta}} \Psi_1(\hat{n}_i, \hat{f}(\hat{\boldsymbol{x}}_i))$
Severity Loss	Out-of-sample loss for the claim severity	$\sum_{i=1}^{\hat{\theta}} \Psi_2(\hat{s}_i, \hat{g}(\hat{\boldsymbol{x}}_i, \hat{n}_i))$
Frequency Error	Average relative error of $F_N(x_i; \boldsymbol{\alpha})$	$\frac{1}{\hat{\theta}} \sum_{i=1}^{\hat{\theta}} \frac{\lvert \hat{f}(\hat{\boldsymbol{x}}_i) - F_N(\hat{\boldsymbol{x}}_i; \boldsymbol{\alpha}) \rvert}{F_N(\hat{\boldsymbol{x}}_i; \boldsymbol{\alpha})}$
Severity Error	Average relative error of $F_{\tilde{Y}\lvert N}(\boldsymbol{x}, N; \boldsymbol{\beta})$	$\frac{1}{\hat{\theta}} \sum_{i=1}^{\hat{\theta}} \frac{\lvert \hat{g}(\hat{\boldsymbol{x}}_i, \hat{n}_i) - F_{\tilde{Y}\lvert N}(\hat{\boldsymbol{x}}_i, \hat{n}_i; \boldsymbol{\beta}) \rvert}{F_{\tilde{Y}\lvert N}(\hat{\boldsymbol{x}}_i, \hat{n}_i; \boldsymbol{\beta})}$
δ Estimation Error	Relative error of δ	$\frac{\lvert \hat{\delta} - \delta \rvert}{\delta}$

https://doi.org/10.1371/journal.pone.0238000.t001

Table 2. Out-of-sample loss and parameter estimation errors.

	Frequency Loss	Severity Loss	Frequency Error	Severity Error	δ Estimation Error
GLM	-	55005.72 (3731.04)	-	1.3651 (0.0893)	6.9003 (3.9905)
D-GLM	18450.51 (89.25)	46753.29 (322.34)	0.1516 (0.0030)	0.4802 (0.0181)	0.6908 (0.1391)
GAM	-	50678.64 (1409.73)	-	1.2669 (0.0740)	2.8192 (1.0205)
D-GAM	18384.86 (84.01)	46087.76 (218.75)	0.1449 (0.0023)	0.3928 (0.0139)	0.3200 (0.0539)
FSBoost	-	46886.52 (166.85)	-	0.9186 (0.0315)	0.2075 (0.0373)
D-FSBoost	18114.45 (74.25)	45541.16 (176.06)	0.0702 (0.0153)	0.1217 (0.0083)	0.0150 (0.0235)

https://doi.org/10.1371/journal.pone.0238000.t002

Table 2, which are averaged over 20 independent replications. Since the independent and dependent models share the same claim frequency model, we only list the claim frequency result for the dependent models. We can find that dependent models perform better than independent ones. In dependent models, the D-FSBoost model has the best performance in terms of the smallest out-of-sample loss and parameter estimation errors.

In contrast to the GLM, D-GLM, GAM and D-GAM models, the FSBoost and D-FSBoost models can capture complex interactions. Denote by c_1 and c_2 the coefficients of cross-product terms $x_{i1}\,x_{i2}$ and $x_{i3}\,x_{i4}$, respectively. In Fig 1, we change c_1 from 8 to 12 and c_2 from 0.3 to 0.7 to increase effects of interaction terms. We can find that the FSBoost and D-FSBoost models have more stable predictive performance. In the GLM, D-GLM, GAM and D-GAM models, the parameter estimation errors show an increasing trend since they can't capture interaction effects. Next, we use values of x_{i3} and x_{i4} in the training sample and change all values of n_i, $i = 1, \ldots, 10000$, from 0 to 20. For each value of n_i, $i = 1, \ldots, 10000$, we calculate

$$s = \frac{1}{10000}\sum_{i=1}^{10000}\hat{g}(\boldsymbol{x}_{i3}, \boldsymbol{x}_{i4}, n_i). \tag{35}$$

Then, we show the change of s with respect to n_i in Fig 2. The D-GLM model can only measure a linear relation between the claim frequency and severity. Both of the D-GAM and D-FSBoost models can capture the nonlinear dependence between the claim frequency and severity. The D-FSBoost model performs better. The results also confirm that the D-FSBoost model can capture the nonlinear relation between the claim frequency (severity) and predictors.

Complex case

In this subsection, we demonstrate the D-FSBoost model in a complex simulation experiment. We compare the models on a variety of randomly generated functions by using the "random function generator" in Friedman [27].

The "random function generator" generates a function as a linear expansion of functions $\{g_k\}_{k=1}^{20}$:

$$F(\boldsymbol{x}) = \sum_{k=1}^{20} a_k g_k(z_k). \tag{36}$$

Each coefficient a_k is generated from a uniform distribution on (0, 1). The variables $\boldsymbol{z}_k$ is a m_k−sized subset of p-input variables $\boldsymbol{x}$ as

$$z_k = \{\boldsymbol{x}_{\phi(k)}\}_{k=1}^{m_k}, \tag{37}$$

where $\phi(k)$ is an independent random permutation of integers $\{1, \ldots, p\}$. The size m_k is randomly selected as $\min(\lfloor 2.5 + r_k \rfloor, p)$, where r_k is generated from an exponential distribution

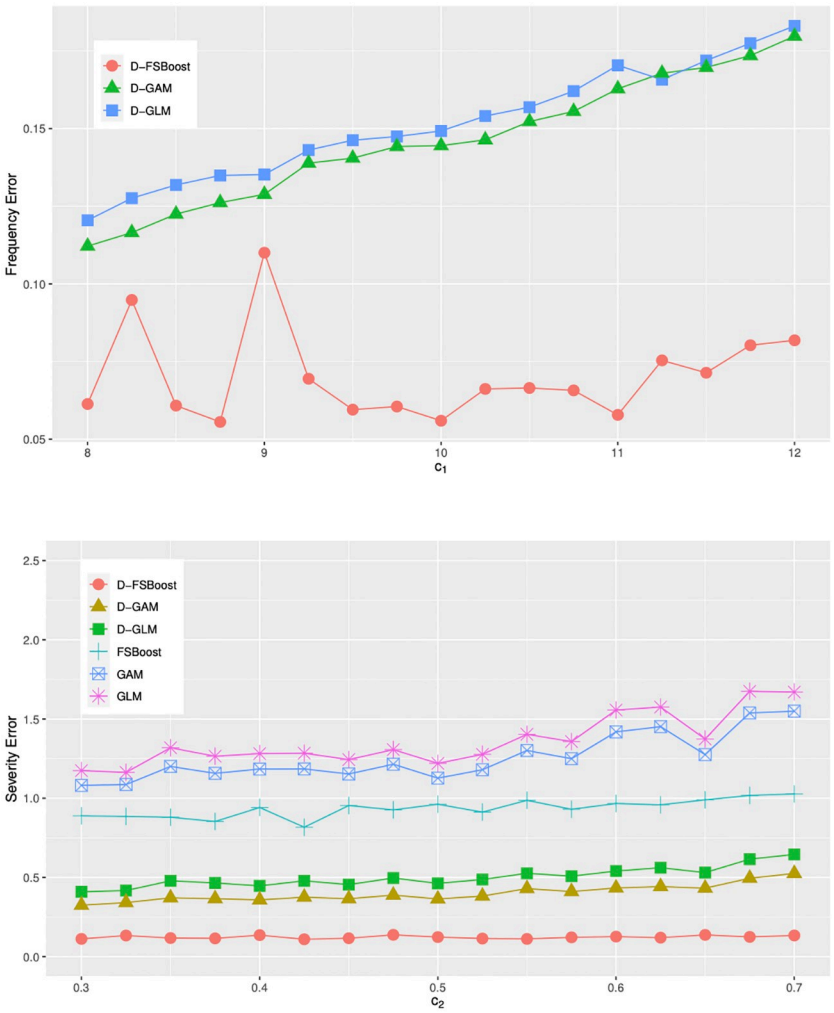

Fig 1. Parameter estimation errors w.r.t. c_1 and c_2.

https://doi.org/10.1371/journal.pone.0238000.g001

with mean 2. Then, the expected number of input variables for each $g_k(\boldsymbol{z}_k)$ is between four and five. Each $g_k(\boldsymbol{z}_k)$ is an m_k–dimensional Gaussian function

$$g_k(z_k) = \exp\left(-\frac{1}{2}(z_k - u_k)^T V_k (z_k - u_k)\right), \tag{38}$$

where each mean vector $\boldsymbol{u}_k$ is generated from the same distribution as $\boldsymbol{z}_k$. The $m_k \times m_k$ covariance matrix V_k is generated by

$$V_k = U_k D_k U_k^T, \tag{39}$$

where U_k is a random orthonormal matrix, $D_k = \text{diag}\{d_1^k, \ldots, d_{m_k}^k\}$, and the square root of each eigenvalue $\sqrt{d_j^k}$ is generated from a uniform distribution on (a, b), where the values of a and b are determined by the distribution of $\boldsymbol{z}_k$. We set the number of predictors $p = 10$ and generate the data $\{n_i, s_i, \boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{\theta}$ using the following specifications,

$$n_i \sim \text{Poi}(\lambda_i), \;\; s_i \sim \text{Gamma}(\mu_{n_i}, \delta), \;\; \boldsymbol{x}_i \sim N(0, \mathbf{I}_p), \;\; y_i \sim N(0, \mathbf{I}_p), \;\; i = 1, \ldots, \theta, \tag{40}$$

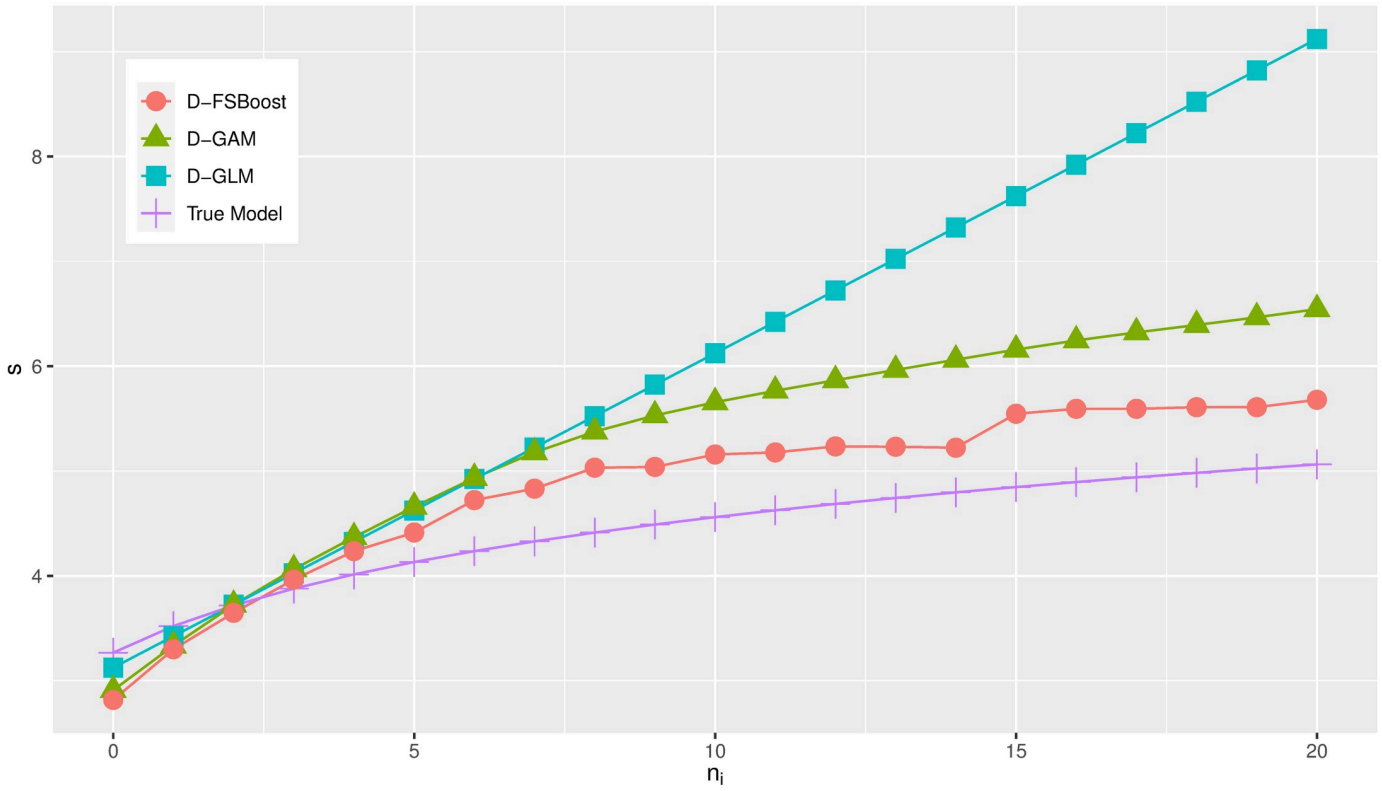

Fig 2. The change of s w.r.t. n_i.

https://doi.org/10.1371/journal.pone.0238000.g002

where $\lambda_i = 1.2\exp(F_1(x_i))$, $\mu_{n_i} = \exp(\log(n_i + 5)F_2(\boldsymbol{y}_i))$, $\delta = 2$, and $F_1(\boldsymbol{x}_i)$ and $F_2(\boldsymbol{y}_i)$ are the functions generated from the "random function generator". The eigenvalue limits are $a = 0.1$ and $b = 2$.

We generate 10000 observations for training and another 10000 for testing. Table 3 reports parameter estimation errors on the testing sample, which are averaged over 10 independent replications. Fig 3 shows out-of-sample loss. The results are the same as in the simple case. Dependent models have more accurate prediction than independent models. The D-FSBoost model performs best in predicting the claim frequency and severity distributions.

The impact of the parameter δ

In this subsection, we investigate the impact of the value of δ on estimating $F_{\tilde{Y}|N}(\boldsymbol{x}, N; \boldsymbol{\beta})$. We generate 20 sets of training samples as in the complex case. Then, we estimate $F_{\tilde{Y}|N}(\boldsymbol{x}, N; \boldsymbol{\beta})$ using the D-FSBoost algorithm for each value of $\delta \in \{1.5, 1.6, \ldots, 2.5\}$. Fig 4 shows parameter estimation errors. We can see that the value of δ has no significant effect on estimation accuracy of $F_{\tilde{Y}|N}(\boldsymbol{x}, N; \boldsymbol{\beta})$.

Application

In this section, we apply the D-FSboost model to analyze a French auto insurance claim data. We compare the models in prediction of the claim frequency and severity distributions. Then, we introduce two important tools, variable importance measures and partial dependence plots, from Friedman [27] to interpret the D-FSBoost model.

Table 3. Parameter estimation errors.

	Frequency Error	Severity Error	δ Estimation Error
GLM	-	4.2459 (1.6405)	6.3718 (2.1039)
D-GLM	0.4529 (0.1768)	2.9406 (0.9037)	2.4299 (0.6622)
GAM	-	1.2852 (0.3275)	1.8402 (0.6285)
D-GAM	0.2175 (0.0500)	0.7398 (0.1398)	0.6570 (0.1870)
FSBoost	-	0.8238 (0.2403)	0.2500 (0.0667)
D-FSBoost	0.2092 (0.0411)	0.4562 (0.0441)	0.0950 (0.0284)

https://doi.org/10.1371/journal.pone.0238000.t003

Data

We consider a French motor third-party liability dataset, where the data "freMTPL2freq" and "freMTPL2sev" are in the R package "CASdatasets". Noll, Salzmann, and Wuthrich [32] use the data "freMTPL2freq" to compare the GLM, regression tree, gradient boosting Poisson model and neural network in predicting the claim frequency. We make the same data

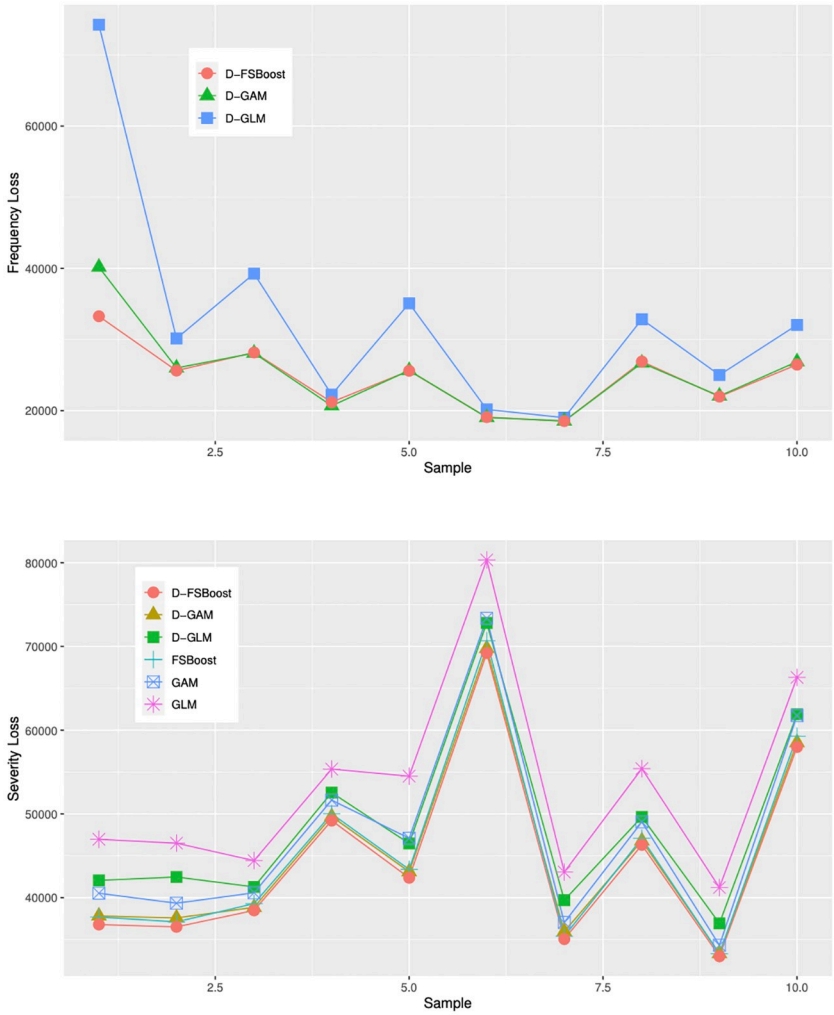

Fig 3. Out-of-sample loss.

https://doi.org/10.1371/journal.pone.0238000.g003

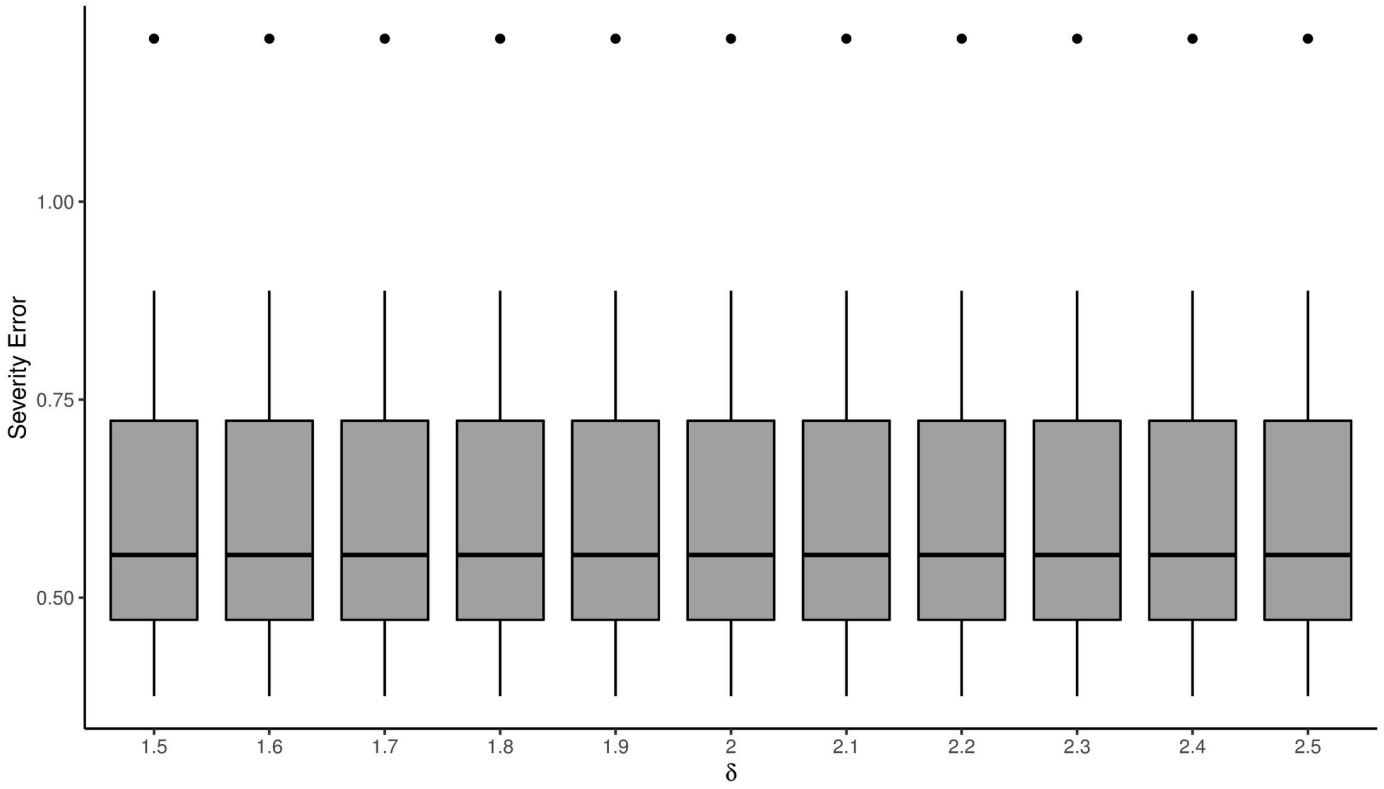

Fig 4. Parameter estimation errors of the D-FSBoost model when varying the value of δ from 1.5 to 2.5.

https://doi.org/10.1371/journal.pone.0238000.g004

preprocess as in Noll, Salzmann, and Wuthrich [32], except deleting records that have positive claim frequency but have no claim severity and also except using different partitions on variables "VehAge" and "DrivAge". After the data preprocess, the dataset contains 668897 records. The dataset is openly available from S1 Dataset. Table 4 shows variables in the dataset. There are 24944 (3.73%) policies that have positive claim frequency. Table 5 reports the distribution

Table 4. Variables.

Variable	Type	Description
ClaimNb	Numeric	The claim frequency during the exposure period
Exposure	Numeric	The period of exposure for a policy in years
ClaimSev	Numeric	The average claim severity
Area	Categorical	The density value of the city community where the policyholder lives: from "1" for rural area to "6" for urban centre (1-6)
VehPower	Categorical	The power of the car (6 classes)
VehAge	Categorical	The vehicle age in years ((0,1], (1,4], (4,10], (10,∞))
DrivAge	Categorical	The driver age in years ([18, 21], (21,25], (25,35], (35,45], (45,55], (55,70], (70,∞))
BonusMalus	Numeric	Bonus/malus: <100 means bonus and >100 means malus in France (50-150)
VehBrand	Categorical	The car brand (B1-B14)
VehGas	Categorical	The car gas (diesel or regular)
LogDensity	Numeric	The log-density of inhabitants of the city where the policyholder lives (number of inhabitants per km^2)
Region	Categorical	The policy region in France based on the 1970-2015 classification (22 classes)

https://doi.org/10.1371/journal.pone.0238000.t004

Table 5. The distribution of the claim frequency and average claim severity.

Claim frequency	0	1	2	3	4	5	6	8	9	11	16
Number of policies	643953	23570	1299	62	5	2	1	1	1	2	1
Average claim severity	0	2177.12	2932.36	4115.35	2203.49	3559.01	1608.93	3103.22	2039.41	1966.92	2220.59

https://doi.org/10.1371/journal.pone.0238000.t005

of the claim frequency and average claim severity. There are only several policies in which the claim frequency is larger than 3. The average claim severity shows an increasing trend when the claim frequency changes from 0 to 3. This implies a positive dependence structure between the claim frequency and severity.

In Figs 5 and 6, we can find that the usage of old cars tend to incur more accidents and higher claim payments. Young drivers have less driving experience than middle-age and old drivers and cause more car crashes and more serious accident loss. In Fig 7, we can find that there are interactions among predictor variables. For young drivers, the vehicle age has a significant effect on the claim frequency. When the driver age increases, the effect gradually decreases. For young and old drivers, there are significant difference in the claim severity between different vehicle age groups. However, for middle-age drivers, the difference is small.

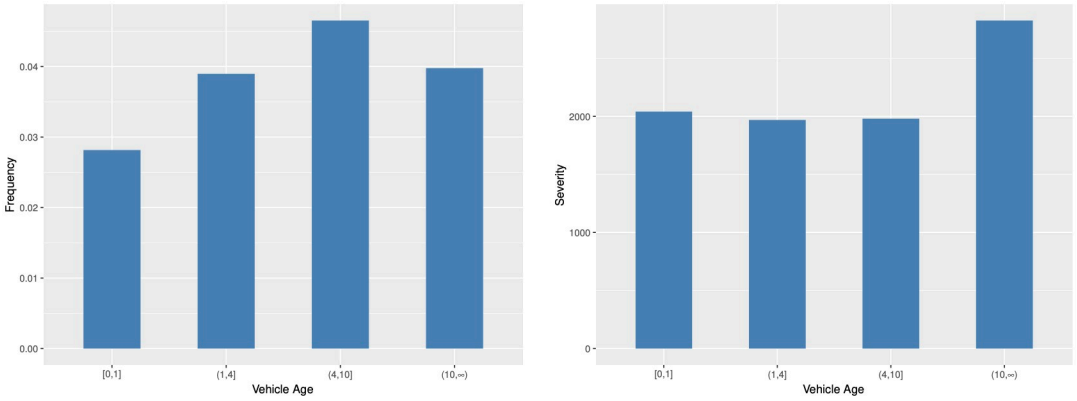

Fig 5. Histogram of the average claim frequency and severity per vehicle age group.

https://doi.org/10.1371/journal.pone.0238000.g005

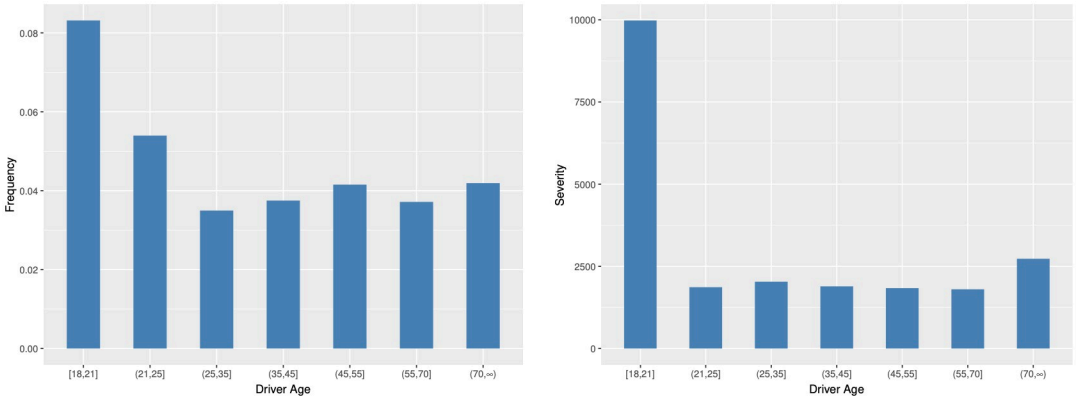

Fig 6. Histogram of the average claim frequency and severity per driver age group.

https://doi.org/10.1371/journal.pone.0238000.g006

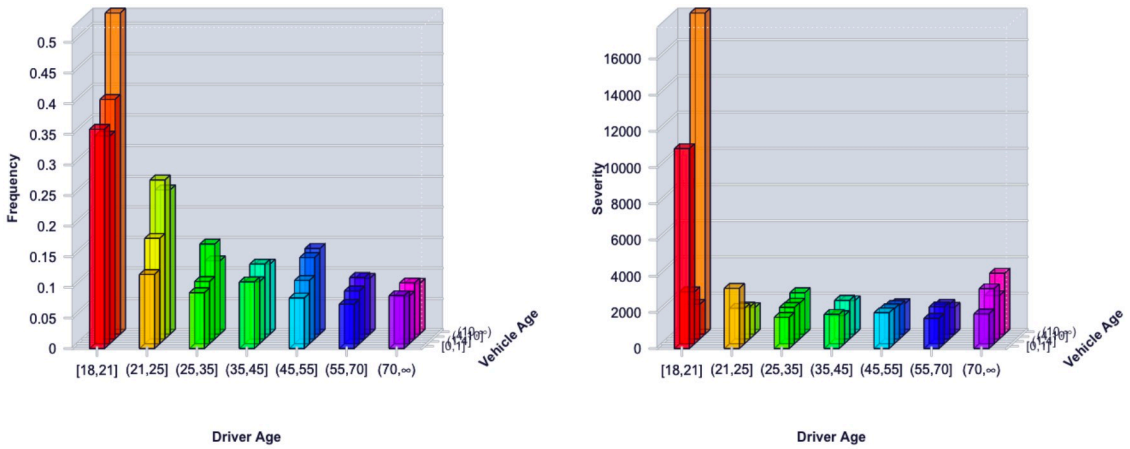

Fig 7. Histogram of the average claim frequency and severity per driver age and vehicle age group.

https://doi.org/10.1371/journal.pone.0238000.g007

Model comparison

We use 445931 observations as training data and the remaining 222966 as testing data. Then, we estimate the GLM, D-GLM, GAM, D-GAM, FSBoost and D-FSBoost models. In dependent models, we take the frequency/exposure instead of the frequency as the predictor variable. The FSBoost and D-FSBoost models can finish automatic feature selection. In the GLM, D-GLM, GAM and D-GAM models, we remove the insignificant variables. Table 6 shows out-of-sample loss for the models. The results indicate that dependent models are more competitive than independent models. The D-FSBoost model is most favorable.

Then, we calculate pure premium prediction from the models on the testing data. We compare the models by using a Gini index to measure the discrepancy between the premium and loss distributions (Frees, Meyers, and Cummings [36, 37]). Let $B(\boldsymbol{x})$ be the base premium and $T(\boldsymbol{x})$ be the alternative premium. Denote by $\Pi(\boldsymbol{x}_i)$ and y_i the pure premium and loss for the i^{th} observation, respectively. Frees, Meyers, and Cummings [36] define a relativity

$$R(\boldsymbol{x}) = \frac{T(\boldsymbol{x})}{B(\boldsymbol{x})} \tag{41}$$

and order observations by relativities $\{R(\boldsymbol{x}_1), \ldots, R(\boldsymbol{x}_\theta)\}$. They define the ordered premium distribution as

$$F_{\Pi}(s) = \frac{\sum_{i=1}^{\theta} \Pi(\boldsymbol{x}_i) \mathbb{1}_{R(\boldsymbol{x}_i) \leq s}}{\sum_{i=1}^{\theta} \Pi(\boldsymbol{x}_i)} \tag{42}$$

Table 6. Out-of-sample loss.

	Frequency Loss	Severity Loss
GLM	-	82595.21
D-GLM	37664.53	82449.21
GAM	-	82108.26
D-GAM	34697.80	82008.89
FSBoost	-	78371.91
D-FSBoost	34155.54	78355.77

https://doi.org/10.1371/journal.pone.0238000.t006

and the ordered loss distribution as

$$F_L(s) = \frac{\sum_{i=1}^{\theta} y_i \mathbb{1}_{R(\boldsymbol{x}_i)\leq s}}{\sum_{i=1}^{\theta} y_i}. \tag{43}$$

The graph of $(F_\Pi(s), F_L(s))$ is an ordered Lorenz curve. When the percentage of losses equals the percentage of premiums, the curve results in a 45-degree line known as "the line of equality". The Gini index is defined as twice the area between the Lorenz curve and the line of equality. Then, the empirical Gini index can be computed by

$$\text{Gini} = 1 - \sum_{i=0}^{\theta-1} (F_\Pi(R(\boldsymbol{x}_{i+1})) - F_\Pi(R(\boldsymbol{x}_i)))(F_L(R(\boldsymbol{x}_{i+1})) + F_L(R(\boldsymbol{x}_i))), \tag{44}$$

where $F_\Pi(R(\boldsymbol{x}_0)) = F_L(R(\boldsymbol{x}_0)) = 0$. A larger Gini index represents more profits for an insurer. Table 7 reports Gini indices calculated by using the prediction from each model as the base premium and using predictions from the remaining models as alternative premiums. Following Frees, Meyers, and Cummings [37] and Yang, Qian, and Zou [33], we use a "minimax" strategy to find the best model. For each base premium, we calculate the maximum Gini index over all alternative premiums. Then, we choose the base premium model that is least vulnerable to alternative premium models, i.e., we select the base premium model that has the smallest maximum Gini index. We find that the maximum Gini index is 0.9432 when using GLM as the base premium model, -0.1300 when using D-GLM, 0.0198 when using GAM, 0.0737 when using D-GAM, 0.0233 when using FSBoost, -0.2855 when using D-FSBoost. Thus, the D-FSBoost model represents the best choice.

Model interpretation

In this subsection, we use variable importance measures and partial dependence plots to interpret the D-FSBoost model. Variable importance measures show the importance of each predictor in predicting the frequency (severity). Partial dependence plots visualize the effect of the predictor on the frequency (severity).

Variable importance. For a single K–terminal node tree T_i, Breiman, Friedman, Olshen, and Stone [38] introduce the following importance measure for the predictor x_j,

$$I_{\boldsymbol{x}_j}(T_i) = \sum_{k=1}^{K-1} \rho_k \mathbb{1}_{v_k}(\boldsymbol{x}_j), \tag{45}$$

Table 7. Gini indices.

Base premium	Alternative premium					
	GLM	D-GLM	GAM	D-GAM	FSBoost	D-FSBoost
GLM	-	0.0861	0.9432	-0.2309	-0.9472	-0.2004
D-GLM	-0.9999	-	-0.9998	-0.2126	-0.9995	-0.1300
GAM	-0.9939	-0.0996	-	0.0198	-0.9912	0.0101
D-GAM	-0.9999	-0.2779	-0.9998	-	-0.9995	0.0737
FSBoost	-0.9999	-0.1052	-0.9989	0.0233	-	0.0202
D-FSBoost	-0.9999	-0.3552	-0.9998	-0.2855	-0.9997	-

https://doi.org/10.1371/journal.pone.0238000.t007

where the sum is taken over all $K-1$ internal nodes, v_k is the splitting variable in the node k, $\mathbb{1}_{v_k}(\boldsymbol{x}_j)$ is an indicator function that equals one when the splitting variable v_k is x_j, and ρ_k denotes the decrease in squared error by splitting the region associated with the node k into two subregions. Friedman [27] generalizes the variable importance measure to the gradient boosting model by taking the average over all trees $\{T_1, \ldots, T_M\}$,

$$\hat{I}_{\boldsymbol{x}_j} = \frac{1}{M}\sum_{m=1}^{M} I_{\boldsymbol{x}_j}(T_m). \tag{46}$$

The variable importance measure is biased since an independent predictor x_j can be selected as a splitting variable and hence $\hat{I}_{\boldsymbol{x}_j}$ can not be zero. See Sandri and Zuccolotto [39, 40] for a bias correction.

In Fig 8, we show variable importance measures for the D-FSBoost model. We can find that VehBrand and BonusMalus are two most important variables in predicting the frequency. The VehBrand dominates the prediction. In predicting the severity, the variables DrivAge, Frequency/Exposure, BonusMalus and LogDensity are most influential. The DrivAge and Frequency/Exposure exert the leading effects. This result also provides further evidence on the dependence between the frequency and severity.

Partial dependence plots. Let $\boldsymbol{z}_k$ be the subset of variables $\boldsymbol{x}$ and $\boldsymbol{z}_{-k}$ be the complement subset of $\boldsymbol{z}_k$ such that

$$z_k \cup z_{-k} = \boldsymbol{x}. \tag{47}$$

The partial dependence of $F_N(\boldsymbol{x}; \boldsymbol{\alpha})$ on $\boldsymbol{z}_k$ can be calculated by

$$\hat{F}(z_k) = \frac{1}{\theta}\sum_{i=1}^{\theta} F_N(z_k, z_{i,-k}; \boldsymbol{\alpha}), \tag{48}$$

where $\boldsymbol{z}_{i,-k}$ is the i^{th} observation of $\boldsymbol{z}_{-k}$. Then, the partial dependence plot of the frequency part is obtained by plotting the function $\hat{F}(\boldsymbol{z}_k)$ against $\boldsymbol{z}_k$. The partial dependence plot of the severity part can be obtained in the same manner.

In Fig 9, we show the partial dependence plots for the D-FSBoost model, indicating the effects of two most important variables on the claim frequency and severity. From the top two panels, we can find that the car with brands B7-B9 causes much more accidents. The frequency is positively associated with the bonus-malus level. In France, the bonus-malus level less than 100 and larger than 100 means bonus and malus, respectively. The change from bonus to

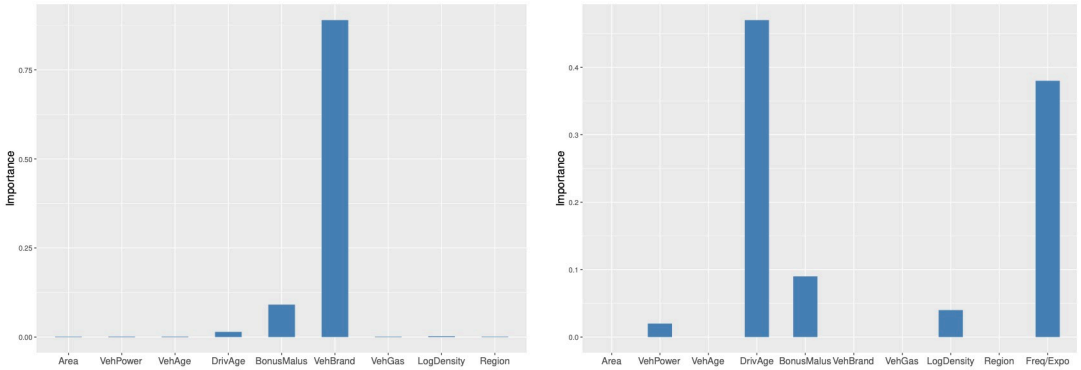

Fig 8. Variable importance measures.

https://doi.org/10.1371/journal.pone.0238000.g008

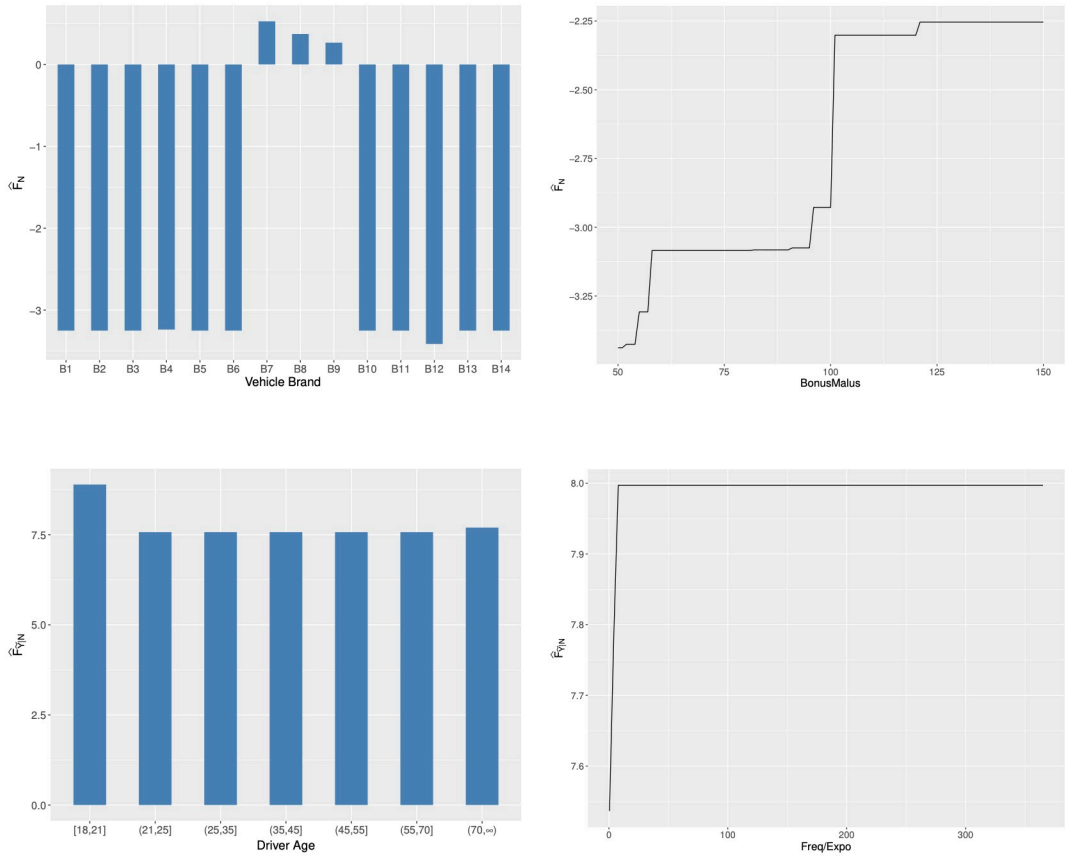

Fig 9. Partial dependence plots.

https://doi.org/10.1371/journal.pone.0238000.g009

malus represents that at least an accident occurs. This explains the sudden change in the frequency at the bonus-malus level 100. The bonus-malus level 60 is the least bonus level to encourage policyholders to drive more carefully, which explains the sudden increase in occurrence of accidents when the bonus-malus level is near to 60. The bottom two panels show that young drivers induce more serious accidents. The severity increases dramatically when the claim frequency is small. This result is consistent with the observation in the distribution of the claim frequency and average claim severity.

Conclusion

This paper develops a stochastic gradient boosting frequency-severity model by using the stochastic gradient boosting algorithm and profile likelihood approach. We demonstrate that the model can flexibly capture the nonlinear relation between the claim frequency (severity) and predictors and complex interactions among predictors, and can also fully capture the nonlinear dependence between the claim frequency and severity. The model is superior to other state-of-the-art models in the sense that it provides more accurate predictions in the claim frequency and severity distributions and pure premium.

In this paper, we illustrate the model with a Poisson distribution for the claim frequency and with a gamma distribution for the average claim severity. In fact, there are more flexible distribution choices. For example, we can use the negative binomial distribution for the claim frequency and the generalized gamma distribution for the average claim severity as in Shi,

Feng, and Ivantsova [18]. The model can also be extended to capture different features of the claim data. For example, our model can combine with the hurdle and zero-inflated modeling framework to accommodate the overdispersion and zero inflation in the claim frequency. We can generalize our model to a random parameters version. We can also assume that the dispersion parameter depends on predictors and model the dispersion parameter with another ensemble of regression trees. These works are left for future research.

Supporting information

S1 Dataset. Motor.
(CSV)

Author Contributions

Conceptualization: Xiaoshan Su.

Data curation: Manying Bai.

Formal analysis: Xiaoshan Su.

Funding acquisition: Manying Bai.

Investigation: Manying Bai.

Methodology: Xiaoshan Su.

Project administration: Manying Bai.

Resources: Manying Bai.

Software: Xiaoshan Su.

Supervision: Manying Bai.

Validation: Manying Bai.

Visualization: Xiaoshan Su.

Writing – original draft: Xiaoshan Su.

Writing – review & editing: Manying Bai.

References

1. Dionne G, Gouriéroux C, Vanasse C. Testing for evidence of adverse selection in the automobile insurance market: A comment. Journal of Political Economy. 2001; 109(2):444–453.
2. Anastasopoulos PC, Shankar VN, Haddock JE, Mannering FL. A multivariate tobit analysis of highway accident-injury-severity rates. Accident Analysis & Prevention. 2012; 45:110–119.
3. Huang H, Song B, Xu P, Zeng Q, Lee J, Abdel-Aty M. Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of transport geography. 2016; 54:248–256.
4. Zeng Q, Wen H, Wong S, Huang H, Guo Q, Pei X. Spatial joint analysis for zonal daytime and nighttime crash frequencies using a Bayesian bivariate conditional autoregressive model. Journal of Transportation Safety & Security. 2020; 12(4):566–585.
5. Aguero-Valverde J. Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. Accident Analysis & Prevention. 2013; 59:365–373.
6. Barua S, El-Basyouny K, Islam MT. Multivariate random parameters collision count data models with spatial heterogeneity. Analytic methods in accident research. 2016; 9:1–15.
7. Zeng Q, Wen H, Huang H, Pei X, Wong S. A multivariate random-parameters Tobit model for analyzing highway crash rates by injury severity. Accident Analysis & Prevention. 2017; 99:184–191.

8. Zeng Q, Guo Q, Wong S, Wen H, Huang H, Pei X. Jointly modeling area-level crash rates by severity: a Bayesian multivariate random-parameters spatio-temporal Tobit regression. Transportmetrica A: Transport Science. 2019; 15(2):1867–1884.
9. Dong B, Ma X, Chen F, Chen S. Investigating the differences of single-vehicle and multivehicle accident probability using mixed logit model. Journal of Advanced Transportation. 2018; 2018.
10. Chen F, Chen S, Ma X. Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data. Journal of safety research. 2018; 65:153–159. PMID: 29776524
11. Chen F, Song M, Ma X. Investigation on the injury severity of drivers in rear-end collisions between cars using a random parameters bivariate ordered probit model. International journal of environmental research and public health. 2019; 16(14):2632.
12. Frees EW, Shi P, Valdez EA. Actuarial applications of a hierarchical insurance claims model. ASTIN Bulletin: The Journal of the IAA. 2009; 39(1):165–197.
13. Hastie T, Tibshirani R. Generalized Additive Models. Statistical Science. 1986; 1(3):297–318.
14. Wood SN. Generalized Additive Models: An Introduction with R. 2006.
15. Gschlößl S, Czado C. Spatial modelling of claim frequency and claim size in non-life insurance. Scandinavian Actuarial Journal. 2007; 2007(3):202–225.
16. Frees EW, Gao J, Rosenberg MA. Predicting the frequency and amount of health care expenditures. North American Actuarial Journal. 2011; 15(3):377–392.
17. Erhardt V, Czado C. Modeling dependent yearly claim totals including zero claims in private health insurance. Scandinavian Actuarial Journal. 2012; 2012(2):106–129.
18. Shi P, Feng X, Ivantsova A. Dependent frequency–severity modeling of insurance claims. Insurance: Mathematics and Economics. 2015; 64:417–428.
19. Garrido J, Genest C, Schulz J. Generalized linear models for dependent frequency and severity of insurance claims. Insurance: Mathematics and Economics. 2016; 70:205–215.
20. Czado C, Kastenmeier R, Brechmann EC, Min A. A mixed copula model for insurance claims and claim sizes. Scandinavian Actuarial Journal. 2012; 2012(4):278–305.
21. Krämer N, Brechmann EC, Silvestrini D, Czado C. Total loss estimation using copula-based regression models. Insurance: Mathematics and Economics. 2013; 53(3):829–839.
22. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences. 1997; 55(1):119–139.
23. Breiman L, et al. Arcing classifier (with discussion and a rejoinder by the author). The annals of statistics. 1998; 26(3):801–849.
24. Breiman L. Prediction games and arcing algorithms. Neural computation. 1999; 11(7):1493–1517. PMID: 10490934
25. Friedman J, Hastie T, Tibshirani R, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics. 2000; 28(2):337–407.
26. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics New York; 2001.
27. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001; p. 1189–1232.
28. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis. 2002; 38(4):367–378.
29. Ridgeway G. The state of boosting. Computing Science and Statistics. 1999; p. 172–181.
30. Ridgeway GK. Generalization of boosting algorithms and applications of bayesian inference for massive datasets; 1999.
31. Kriegler B, Berk R. Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting. The Annals of Applied Statistics. 2010; p. 1234–1255.
32. Noll A, Salzmann R, Wuthrich MV. Case study: French motor third-party liability claims. 2018.
33. Yang Y, Qian W, Zou H. Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. Journal of Business & Economic Statistics. 2018; 36(3):456–470.
34. Zhou H, Yang Y, Qian W. Tweedie Gradient Boosting for Extremely Unbalanced Zero-inflated Data. arXiv preprint arXiv:181110192. 2018.
35. Sigrist F, Hirnschall C. Grabit: Gradient tree-boosted Tobit models for default prediction. Journal of Banking & Finance. 2019; 102:177–192.

36. Frees EW, Meyers G, Cummings AD. Summarizing insurance scores using a Gini index. Journal of the American Statistical Association. 2011; 106(495):1085–1098.

37. Frees EW, Meyers G, Cummings AD. Insurance ratemaking and a Gini index. Journal of Risk and Insurance. 2014; 81(2):335–366.

38. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. 1984.

39. Sandri M, Zuccolotto P. A bias correction algorithm for the Gini variable importance measure in classification trees. Journal of Computational and Graphical Statistics. 2008; 17(3):611–628.

40. Sandri M, Zuccolotto P. Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms. Statistics and Computing. 2010; 20(4):393–407.