Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://www.tandfonline.com/loi/ubes20

Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models

Yi Yang, Wei Qian & Hui Zou

Supplementary materials for this article are available online. Please go to http://tandfonline.com/r/JBES

Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models

Yi YANG
McGill University, Montreal, Canada

Wei QIAN
School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY 14623 (wxqsma@rit.edu)

Hui ZOU
School of Statistics, University of Minnesota, MN 55455 (zoux019@umn.edu)

The Tweedie GLM is a widely used method for predicting insurance premiums. However, the structure of the logarithmic mean is restricted to a linear form in the Tweedie GLM, which can be too rigid for many applications. As a better alternative, we propose a gradient tree-boosting algorithm and apply it to Tweedie compound Poisson models for pure premiums. We use a profile likelihood approach to estimate the index and dispersion parameters. Our method is capable of fitting a flexible nonlinear Tweedie model and capturing complex interactions among predictors. A simulation study confirms the excellent prediction performance of our method. As an application, we apply our method to an auto-insurance claim data and show that the new method is superior to the existing methods in the sense that it generates more accurate premium predictions, thus helping solve the adverse selection issue. We have implemented our method in a user-friendly R package that also includes a nice visualization tool for interpreting the fitted model.

KEY WORDS: Claim frequency and severity gradient boosting; Insurance claims data; Ratemaking; Zero inflation.

1. INTRODUCTION

One of the most important problems in insurance business is to set the premium for the customers (policyholders). In a competitive market, it is advantageous for the insurer to charge a fair premium according to the expected loss of the policyholder. In personal car insurance, for instance, if an insurance company charges too much for old drivers and charges too little for young drivers, then the old drivers will switch to its competitors, and the remaining policies for the young drivers would be underpriced. This results in the *adverse selection* issue (Dionne, Gouriéroux, and Vanasse 2001): the insurer loses profitable policies and is left with bad risks, resulting in economic loss both ways.

To appropriately set the premiums for the insurer's customers, one crucial task is to predict the size of actual (currently unforeseeable) claims. In this article, we will focus on modeling claim loss, although other ingredients such as safety loadings, administrative costs, cost of capital, and profit are also important factors for setting the premium. One difficulty in modeling the claims is that the distribution is usually highly right-skewed, mixed with a point mass at zero. Such type of data cannot be transformed to normality by power transformation, and special treatment on zero claims is often required. As an example, Figure 1 shows the histogram of an auto insurance claim data (Yip and Yau 2005), in which there are 6290 policy records with zero claims and 4006 policy records with positive losses.

The need for predictive models emerges from the fact that the expected loss is highly dependent on the characteristics of an individual policy such as age and motor vehicle record points of the policyholder, population density of the policyholder's residential area, and age and model of the vehicle. Traditional methods used generalized linear models (GLM; Nelder and Wedderburn 1972) for modeling the claim size (e.g., Renshaw 1994; Haberman and Renshaw 1996). However, the authors of the above papers performed their analyses on a subset of the policies, which have at least one claim. Alternative approaches have employed Tobit models by treating zero outcomes as censored below some cutoff points (Van de Ven and van Praag 1981; Showers and Shotick 1994), but these approaches rely on a normality assumption of the latent response. Alternatively, Jørgensen and de Souza (1994) and Smyth and Jørgensen (2002) used GLMs with a Tweedie distributed outcome to simultaneously model frequency and severity of insurance claims. They assume Poisson arrival of claims and gamma distributed amount for individual claims so that the size of the total claim amount follows a Tweedie compound Poisson distribution. Due to its ability to simultaneously model the zeros and the continuous positive outcomes, the Tweedie GLM has been a widely used method in actuarial studies (Mildenhall 1999; Murphy, Brockman, and Lee 2000; Peters, Shevchenko, and Wüthrich 2008).

Despite of the popularity of the Tweedie GLM, a major limitation is that the structure of the logarithmic mean is restricted to a linear form, which can be too rigid for real applications. In auto insurance, for example, it is known that the risk does not

Journal of Business & Economic Statistics
July 2018, Vol. 36, No. 3
DOI: 10.1080/07350015.2016.1200981

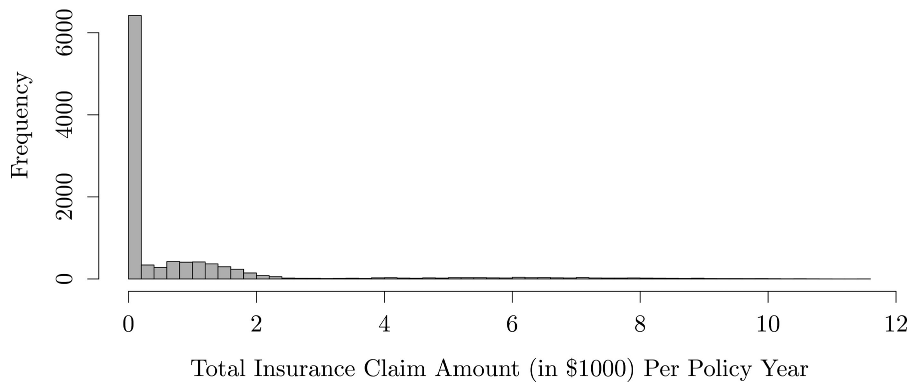

Figure 1. Histogram of the auto-insurance claim data as analyzed by Yip and Yau (2005). It shows that there are 6290 policy records with zero total claims per policy year, while the remaining 4006 policy records have positive losses.

monotonically decrease as age increases (Anstey et al. 2005). Although nonlinearity may be modeled by adding splines (Zhang 2011), low-degree splines are often inadequate to capture the nonlinearity in the data, while high-degree splines often result in the overfitting issue that produces unstable estimates. Generalized additive models (GAM; Hastie and Tibshirani 1990; Wood 2006) overcome the restrictive linear assumption of GLMs and can model the continuous variables by smooth functions estimated from data. The structure of the model, however, has to be determined a priori. That is, one has to specify the main effects and interaction effects to be used in the model. As a result, misspecification of nonignorable effects is likely to adversely affect prediction accuracy.

In this article, we aim to model the insurance claim size by a nonparametric Tweedie compound Poisson model and propose a gradient tree-boosting algorithm (TDboost henceforth) to fit this model. We also implemented the proposed method as an easy-to-use R package, which is publicly available.

Gradient boosting is one of the most successful machine learning algorithms for nonparametric regression and classification. Boosting adaptively combines a large number of relatively simple prediction models called *base learners* into an ensemble learner to achieve high-prediction performance. The seminal work on the boosting algorithm called *AdaBoost* (Freund and Schapire 1997) was originally proposed for classification problems. Later Breiman (1998) and Breiman (1999) pointed out an important connection between the AdaBoost algorithm and a functional gradient descent algorithm. Friedman, Hastie, and Tibshirani (2000) and Hastie, Tibshirani, and Friedman (2009) developed a statistical view of boosting and proposed gradient boosting methods for both classification and regression. There is a large body of literature on boosting. We refer interested readers to Bühlmann and Hothorn (2007) for a comprehensive review of boosting algorithms.

The TDboost model is motivated by the proven success of boosting in machine learning for classification and regression problems (Friedman 2001, 2002; Hastie, Tibshirani, and Friedman 2009). Its advantages are threefold. First, the model structure of TDboost is learned from data and not predetermined, thereby avoiding an explicit model specification. Nonlinearities, discontinuities, complex and higher order interactions are naturally incorporated into the model to reduce the potential modeling bias and to produce high predictive performance, which enables TDboost to serve as a benchmark model in scoring insurance policies, guiding pricing practice, and facilitating marketing efforts. Feature selection is performed as an integral part of the procedure. In addition, TDboost handles the predictor and response variables of any type without the need for transformation, and it is highly robust to outliers. Missing values in the predictors are managed almost without loss of information (Elith, Leathwick, and Hastie 2008). All these properties make TDboost a more attractive tool for insurance premium modeling. On the other hand, we acknowledge that its results are not as straightforward as those from the Tweedie GLM model. Nevertheless, TDboost does not have to be regarded as a black box. It can provide interpretable results, by means of the partial dependence plots, and relative importance of the predictors.

The remainder of this article is organized as follows. We briefly review the gradient boosting algorithm and the Tweedie compound Poisson model in Sections 2 and 3, respectively. We present the main methodological development with implementation details in Section 4. In Section 5, we use simulation to show the high-predictive accuracy of TDboost. As an application, we apply TDboost to analyze an auto-insurance claim data in Section 6.

2. GRADIENT BOOSTING

Gradient boosting (Friedman 2001) is a recursive, nonparametric machine learning algorithm that has been successfully used in many areas. It shows remarkable flexibility in solving different loss functions. By combining a large number of base learners, it can handle higher order interactions and produce highly complex functional forms. It provides high-prediction accuracy and often outperforms many competing methods, such as linear regression/classification, bagging (Breiman 1996), splines, and CART (Breiman et al. 1984).

To keep the article self-contained, we briefly explain the general procedures for the gradient boosting. Let $\mathbf{x} = (x_1, \ldots, x_p)^\intercal$ be a p-dimensional column vector for the predictor variables and y be the one-dimensional response variable. The goal is to estimate the optimal prediction function $\tilde{F}(\cdot)$ that maps $\mathbf{x}$ to y by minimizing the expected value of a loss function $\Psi(\cdot, \cdot)$ over the function class $\mathcal{F}$:

$$\tilde{F}(\cdot) = \underset{F(\cdot)\in\mathcal{F}}{\arg\min} E_{y,\mathbf{x}}[\Psi(y, F(\mathbf{x}))],$$

where Ψ is assumed to be differentiable with respect to F. Given the observed data $\{y_i, \mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$, estimation of $\tilde{F}(\cdot)$ can be done by minimizing the empirical risk function

$$\min_{F(\cdot)\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\Psi(y_i, F(\mathbf{x}_i)). \tag{1}$$

For the gradient boosting, each candidate function $F \in \mathcal{F}$ is assumed to be an ensemble of M base learners

$$F(\mathbf{x}) = F^{[0]} + \sum_{m=1}^{M}\beta^{[m]}h(\mathbf{x}; \boldsymbol{\xi}^{[m]}), \tag{2}$$

where $h(\mathbf{x}; \boldsymbol{\xi}^{[m]})$ usually belongs to a class of some simple functions of $\mathbf{x}$ called base learners (e.g., regression/decision tree) with the parameter $\boldsymbol{\xi}^{[m]}$ $(m = 1, 2, \ldots, M)$. $F^{[0]}$ is a constant scalar and $\beta^{[m]}$ is the expansion coefficient. Note that differing from the usual structure of an additive model, there is no restriction on the number of predictors to be included in each $h(\cdot)$, and consequently, high-order interactions can be easily considered using this setting.

A forward stagewise algorithm is adopted to approximate the minimizer of (1), which builds up the components $\beta^{[m]}h(\mathbf{x}; \boldsymbol{\xi}^{[m]})$ $(m = 1, 2, \ldots, M)$ sequentially through a gradient-descent-like approach. At each iteration stage m $(m = 1, 2, \ldots)$, suppose that the current estimate for $\tilde{F}(\cdot)$ is $\hat{F}^{[m-1]}(\cdot)$. To update the estimate from $\hat{F}^{[m-1]}(\cdot)$ to $\hat{F}^{[m]}(\cdot)$, the gradient boosting fits a negative gradient vector (as the working response) to the predictors using a base learner $h(\mathbf{x}; \boldsymbol{\xi}^{[m]})$. This fitted $h(\mathbf{x}; \boldsymbol{\xi}^{[m]})$ can be viewed as an approximation of the negative gradient. Subsequently, the expansion coefficient $\beta^{[m]}$ can then be determined by a line search minimization with the empirical risk function, and the estimation of $\tilde{F}(\mathbf{x})$ for the next stage becomes

$$\hat{F}^{[m]}(\mathbf{x}) := \hat{F}^{[m-1]}(\mathbf{x}) + \nu\beta^{[m]}h(\mathbf{x}; \boldsymbol{\xi}^{[m]}), \tag{3}$$

where $0 < \nu \le 1$ is the shrinkage factor (Friedman 2001) that controls the update step size. A small ν imposes more shrinkage, while $\nu = 1$ gives complete negative gradient steps. Friedman (2001) has found that the shrinkage factor reduces overfitting and improves the predictive accuracy.

3. COMPOUND POISSON DISTRIBUTION AND TWEEDIE MODEL

In insurance premium prediction problems, the total claim amount for a covered risk usually has a continuous distribution on positive values, except for the possibility of being exact zero when the claim does not occur. One standard approach in actuarial science in modeling such data is using Tweedie compound Poisson models, which we briefly introduce in this section.

Let N be a Poisson random variable denoted by Pois(λ), and let $\tilde{Z}_d$'s $(d = 0, 1, \ldots, N)$ be iid gamma random variables denoted by Gamma(α, γ) with mean $\alpha\gamma$ and variance $\alpha\gamma^2$. Assume N is independent of $\tilde{Z}_d$'s. Define a random variable Z by

$$Z = \begin{cases} 0 & \text{if } N = 0 \\ \tilde{Z}_1 + \tilde{Z}_2 + \cdots + \tilde{Z}_N & \text{if } N = 1, 2, \ldots \end{cases}. \tag{4}$$

Thus, Z is the Poisson sum of independent Gamma random variables. In insurance applications, one can view Z as the total claim amount, N as the number of reported claims and $\tilde{Z}_d$'s as the insurance payment for the dth claim. The resulting distribution of Z is referred to as the compound Poisson distribution (Jørgensen and de Souza 1994; Smyth and Jørgensen 2002), which is known to be closely connected to exponential dispersion models (EDM; Jørgensen 1987). Note that the distribution of Z has a probability mass at zero: $\Pr(Z = 0) = \exp(-\lambda)$. Then based on that Z conditional on $N = j$ is Gamma$(j\alpha, \gamma)$, the distribution function of Z can be written as

$$\begin{aligned} f_Z(z|\lambda, \alpha, \gamma) &= \Pr(N = 0)d_0(z) + \sum_{j=1}^{\infty}\Pr(N = j)f_{Z|N=j}(z) \\ &= \exp(-\lambda)d_0(z) + \sum_{j=1}^{\infty}\frac{\lambda^j e^{-\lambda}}{j!}\frac{z^{j\alpha-1}e^{-z/\gamma}}{\gamma^{j\alpha}\Gamma(j\alpha)}, \end{aligned}$$

where d_0 is the Dirac delta function at zero and $f_{Z|N=j}$ is the conditional density of Z given $N = j$. Smyth (1996) pointed out that the compound Poisson distribution belongs to a special class of EDMs known as Tweedie models (Tweedie 1984), which are defined by the form

$$f_Z(z|\theta, \phi) = a(z, \phi)\exp\left\{\frac{z\theta - \kappa(\theta)}{\phi}\right\}, \tag{5}$$

where $a(\cdot)$ is a normalizing function, $\kappa(\cdot)$ is called the cumulant function, and both $a(\cdot)$ and $\kappa(\cdot)$ are known. The parameter θ is in $\mathbb{R}$ and the dispersion parameter ϕ is in $\mathbb{R}^+$. For Tweedie models the mean $E(Z) \equiv \mu = \dot{\kappa}(\theta)$ and the variance $\mathrm{var}(Z) = \phi\ddot{\kappa}(\theta)$, where $\dot{\kappa}(\theta)$ and $\ddot{\kappa}(\theta)$ are the first and second derivatives of $\kappa(\theta)$, respectively. Tweedie models have the power mean–variance relationship $\mathrm{var}(Z) = \phi\mu^\rho$ for some index parameter ρ. Such mean–variance relation gives

$$\theta = \begin{cases} \frac{\mu^{1-\rho}}{1-\rho}, & \rho \ne 1 \\ \log\mu, & \rho = 1 \end{cases}, \qquad \kappa(\theta) = \begin{cases} \frac{\mu^{2-\rho}}{2-\rho}, & \rho \ne 2 \\ \log\mu, & \rho = 2 \end{cases}. \tag{6}$$

One can show that the compound Poisson distribution belongs to the class of Tweedie models. Indeed, if we reparameterize $(\lambda, \alpha, \gamma)$ by

$$\lambda = \frac{1}{\phi}\frac{\mu^{2-\rho}}{2-\rho}, \qquad \alpha = \frac{2-\rho}{\rho-1}, \qquad \gamma = \phi(\rho-1)\mu^{\rho-1}, \tag{7}$$

the compound Poisson model will have the form of a Tweedie model with $1 < \rho < 2$ and $\mu > 0$. As a result, for the rest of this article, we only consider the model (4), and simply refer to (4) as the Tweedie model (or Tweedie compound Poisson model), denoted by Tw(μ, ϕ, ρ), where $1 < \rho < 2$ and $\mu > 0$.

It is straightforward to show that the log-likelihood of the Tweedie model is

$$\log f_Z(z|\mu, \phi, \rho) = \frac{1}{\phi}\left(z\frac{\mu^{1-\rho}}{1-\rho} - \frac{\mu^{2-\rho}}{2-\rho}\right) + \log a(z, \phi, \rho), \tag{8}$$

where the normalizing function $a(\cdot)$ can be written as

$$a(z, \phi, \rho) = \begin{cases} \frac{1}{z}\sum_{t=1}^{\infty} W_t(z, \phi, \rho) \\ \quad = \frac{1}{z}\sum_{t=1}^{\infty} \frac{z^{t\alpha}}{(\rho-1)^{t\alpha}\phi^{t(1+\alpha)}(2-\rho)^t t!\Gamma(t\alpha)} & \text{for } z > 0 \\ 1 & \text{for } z = 0 \end{cases},$$

and $\alpha = (2-\rho)/(\rho-1)$ and $\sum_{t=1}^{\infty} W_t$ is an example of Wright's generalized Bessel function (Tweedie 1984).

4. OUR PROPOSAL

In this section, we propose to integrate the Tweedie model to the tree-based gradient boosting algorithm to predict insurance claim size. Specifically, our discussion focuses on modeling the personal car insurance as an illustrating example (see Section 6 for a real data analysis), since our modeling strategy is easily extended to other lines of nonlife insurance business.

Given an auto-insurance policy i, let N_i be the number of claims (known as the claim frequency) and $\tilde{Z}_{d_i}$ be the size of each claim observed for $d_i = 1, \ldots, N_i$. Let w_i be the policy duration, that is, the length of time that the policy remains in force. Then, $Z_i = \sum_{d_i=1}^{N_i} \tilde{Z}_{d_i}$ is the total claim amount. In the following, we are interested in modeling the ratio between the total claim and the duration $Y_i = Z_i/w_i$, a key quantity known as the pure premium (Ohlsson and Johansson 2010).

Following the settings of the compound Poisson model, we assume N_i is Poisson distributed, and its mean $\lambda_i w_i$ has a multiplicative relation with the duration w_i, where λ_i is a policy-specific parameter representing the expected claim frequency under unit duration. Conditional on N_i, assume Z_{d_i}'s ($d_i = 1, \ldots, N_i$) are iid Gamma(α, γ_i), where γ_i is a policy-specific parameter that determines claim severity, and α is a constant. Furthermore, we assume that under unit duration (i.e., $w_i = 1$), the mean–variance relation of a policy satisfies $\text{var}(Y_i^*) = \phi[E(Y_i^*)]^\rho$ for all policies, where Y_i^* is the pure premium under unit duration, ϕ is a constant, and $\rho = (\alpha+2)/(\alpha+1)$. Then, it is known that $Y_i \sim \text{Tw}(\mu_i, \phi/w_i, \rho)$, the details of which are provided in Appendix A.

Then, we consider a portfolio of policies $\{(y_i, \mathbf{x}_i, w_i)\}_{i=1}^n$ from n independent insurance contracts, where for the ith contract, y_i is the policy pure premium, $\mathbf{x}_i$ is a vector of explanatory variables that characterize the policyholder and the risk being insured (e.g., house, vehicle), and w_i is the duration. Assume that the expected pure premium μ_i is determined by a predictor function $F : \mathbb{R}^p \to \mathbb{R}$ of $\mathbf{x}_i$:

$$\log\{\mu_i\} = \log\{E(Y_i|\mathbf{x}_i)\} = F(\mathbf{x}_i). \tag{9}$$

In this article, we do not impose a linear or other parametric form restriction on $F(\cdot)$. Given the flexibility of $F(\cdot)$, we call such setting as the boosted Tweedie model (as opposed to the Tweedie GLM). Given $\{(y_i, \mathbf{x}_i, w_i)\}_{i=1}^n$, the log-likelihood function can be written as

$$\begin{aligned} \ell\left(F(\cdot), \phi, \rho | \{y_i, \mathbf{x}_i, w_i\}_{i=1}^n\right) &= \sum_{i=1}^n \log f_Y(y_i|\mu_i, \phi/w_i, \rho), \\ &= \sum_{i=1}^n \frac{w_i}{\phi}\left(y_i \frac{\mu_i^{1-\rho}}{1-\rho} - \frac{\mu_i^{2-\rho}}{2-\rho}\right) \\ &\quad + \log a(y_i, \phi/w_i, \rho). \end{aligned} \tag{10}$$

4.1 Estimating $F(\cdot)$ via TDboost

We estimate the predictor function $F(\cdot)$ by integrating the boosted Tweedie model into the tree-based gradient boosting algorithm. To develop the idea, we assume that ϕ and ρ are given for the time being. The joint estimation of $F(\cdot)$, ϕ, and ρ will be studied in Section 4.2.

Given ρ and ϕ, we replace the general objective function in (1) by the negative log-likelihood derived in (10), and target the minimizer function $F^*(\cdot)$ over a class $\mathcal{F}$ of base learner functions in the form of (2). That is, we intend to estimate

$$\begin{aligned} F^*(\mathbf{x}) &= \operatorname*{argmin}_{F\in\mathcal{F}} \left\{-\ell\left(F(\cdot), \phi, \rho | \{y_i, \mathbf{x}_i, w_i\}_{i=1}^n\right)\right\} \\ &= \operatorname*{argmin}_{F\in\mathcal{F}} \sum_{i=1}^n \Psi(y_i, F(\mathbf{x}_i)|\rho), \end{aligned} \tag{11}$$

where

$$\Psi(y_i, F(\mathbf{x}_i)|\rho) = w_i\left\{-\frac{y_i \exp[(1-\rho)F(\mathbf{x}_i)]}{1-\rho} + \frac{\exp[(2-\rho)F(\mathbf{x}_i)]}{2-\rho}\right\}.$$

Note that in contrast to (11), the function class targeted by Tweedie GLM (Smyth 1996) is restricted to a collection of linear functions of $\mathbf{x}$.

We propose to apply the forward stagewise algorithm described in Section 2 for solving (11). The initial estimate of $F^*(\cdot)$ is chosen as a constant function that minimizes the negative log-likelihood:

$$\begin{aligned} \hat{F}^{[0]} &= \operatorname*{argmin}_{\eta} \sum_{i=1}^n \Psi(y_i, \eta \mid \rho) \\ &= \log\left(\frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}\right). \end{aligned}$$

This corresponds to the best estimate of F without any covariates. Let $\hat{F}^{[m-1]}$ be the current estimate before the mth iteration. At the mth step, we fit a base learner $h(\mathbf{x}; \boldsymbol{\xi}^{[m]})$ via

$$\hat{\boldsymbol{\xi}}^{[m]} = \operatorname*{argmin}_{\boldsymbol{\xi}^{[m]}} \sum_{i=1}^n \left[u_i^{[m]} - h(\mathbf{x}_i; \boldsymbol{\xi}^{[m]})\right]^2, \tag{12}$$

where $(u_1^{[m]}, \ldots, u_n^{[m]})^\mathsf{T}$ is the current negative gradient of $\Psi(\cdot \mid \rho)$, that is,

$$u_i^{[m]} = -\left.\frac{\partial \Psi(y_i, F(\mathbf{x}_i) \mid \rho)}{\partial F(\mathbf{x}_i)}\right|_{F(\mathbf{x}_i)=\hat{F}^{[m-1]}(\mathbf{x}_i)} \tag{13}$$

$$= w_i \left\{-y_i \exp\left[(1-\rho)\hat{F}^{[m-1]}(\mathbf{x}_i)\right] + \exp\left[(2-\rho)\hat{F}^{[m-1]}(\mathbf{x}_i)\right]\right\}, \tag{14}$$

and use an L-terminal node regression tree

$$h(\mathbf{x}; \boldsymbol{\xi}^{[m]}) = \sum_{l=1}^{L} u_l^{[m]} I\left(\mathbf{x} \in R_l^{[m]}\right) \tag{15}$$

with parameters $\boldsymbol{\xi}^{[m]} = \{R_l^{[m]}, u_l^{[m]}\}_{l=1}^L$ as the base learner. To find $R_l^{[m]}$ and $u_l^{[m]}$, we use a fast top-down "best-fit" algorithm with a least-squares splitting criterion (Friedman, Hastie, and Tibshirani 2000) to find the splitting variables and corresponding split locations that determine the fitted terminal regions $\{\widehat{R}_l^{[m]}\}_{l=1}^L$. Note that estimating the $R_l^{[m]}$ entails estimating the $u_l^{[m]}$ as the mean falling in each region:

$$\bar{u}_l^{[m]} = \text{mean}_{i:\mathbf{x}_i \in \widehat{R}_l^{[m]}} \left(u_i^{[m]}\right) \qquad l = 1, \ldots, L.$$

Once the base learner $h(\mathbf{x}; \boldsymbol{\xi}^{[m]})$ has been estimated, the optimal value of the expansion coefficient $\beta^{[m]}$ is determined by a line search

$$\beta^{[m]} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \Psi\left(y_i, \hat{F}^{[m-1]}(\mathbf{x}_i) + \beta h\left(\mathbf{x}_i; \hat{\boldsymbol{\xi}}^{[m]}\right) \mid \rho\right)$$

$$= \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \Psi\Bigg(y_i, \hat{F}^{[m-1]}(\mathbf{x}_i) + \beta \sum_{l=1}^{L} \bar{u}_l^{[m]} I\left(\mathbf{x}_i \in \widehat{R}_l^{[m]}\right) \mid \rho\Bigg). \tag{16}$$

The regression tree (15) predicts a constant value $\bar{u}_l^{[m]}$ within each region $\widehat{R}_l^{[m]}$, so we can solve (16) by a separate line search performed within each respective region $\widehat{R}_l^{[m]}$. The problem (16) reduces to finding a best constant $\eta_l^{[m]}$ to improve the current estimate in each region $\widehat{R}_l^{[m]}$ based on the following criterion:

$$\hat{\eta}_l^{[m]} = \underset{\eta}{\text{argmin}} \sum_{i:\mathbf{x}_i \in \widehat{R}_l^{[m]}} \Psi\left(y_i, \hat{F}^{[m-1]}(\mathbf{x}_i) + \eta \mid \rho\right), \quad l = 1, \ldots, L, \tag{17}$$

where the solution is given by

$$\hat{\eta}_l^{[m]} = \log\left\{\frac{\sum_{i:\mathbf{x}_i \in \widehat{R}_l^{[m]}} w_i y_i \exp[(1-\rho)\hat{F}^{[m-1]}(\mathbf{x}_i)]}{\sum_{i:\mathbf{x}_i \in \widehat{R}_l^{[m]}} w_i \exp[(2-\rho)\hat{F}^{[m-1]}(\mathbf{x}_i)]}\right\}, \quad l = 1, \ldots, L. \tag{18}$$

Having found the parameters $\{\hat{\eta}_l^{[m]}\}_{l=1}^L$, we then update the current estimate $\hat{F}^{[m-1]}(\mathbf{x})$ in each corresponding region

$$\hat{F}^{[m]}(\mathbf{x}) = \hat{F}^{[m-1]}(\mathbf{x}) + \nu \hat{\eta}_l^{[m]} I(\mathbf{x} \in \widehat{R}_l^{[m]}), \qquad l = 1, \ldots, L, \tag{19}$$

where $0 < \nu \leq 1$ is the shrinkage factor. Following (Friedman 2001), we set $\nu = 0.005$ in our implementation. More discussions on the choice of tuning parameters are in Section 4.4.

In summary, the complete TDboost algorithm is shown in Algorithm 1. The boosting step is repeated M times and we report $\hat{F}^{[M]}(\mathbf{x})$ as the final estimate.

Algorithm 1. TDboost

1. Initialize $\hat{F}^{[0]}$

$$\hat{F}^{[0]} = \log\left(\frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}\right).$$

2. For $m = 1, \ldots, M$ repeatedly do steps 2(a)–2(d)
   2. (a). Compute the negative gradient $(u_1^{[m]}, \ldots, u_n^{[m]})^\intercal$

   $$u_i^{[m]} = w_i\{-y_i \exp[(1-\rho)\hat{F}^{[m-1]}(\mathbf{x}_i)] + \exp[(2-\rho)\hat{F}^{[m-1]}(\mathbf{x}_i)]\} \qquad i = 1, \ldots, n.$$

   2. (b). Fit the negative gradient vector $(u_1^{[m]}, \ldots, u_n^{[m]})^\intercal$ to $(\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$ by an L-terminal node regression tree, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$ for $i = 1, \ldots, n$, giving us the partitions $\{\widehat{R}_l^{[m]}\}_{l=1}^L$.
   2. (c). Compute the optimal terminal node predictions $\eta_l^{[m]}$ for each region $\widehat{R}_l^{[m]}$, $l = 1, 2, \ldots, L$

   $$\hat{\eta}_l^{[m]} = \log\left\{\frac{\sum_{i:\mathbf{x}_i \in \widehat{R}_l^{[m]}} w_i y_i \exp[(1-\rho)\hat{F}^{[m-1]}(\mathbf{x}_i)]}{\sum_{i:\mathbf{x}_i \in \widehat{R}_l^{[m]}} w_i \exp[(2-\rho)\hat{F}^{[m-1]}(\mathbf{x}_i)]}\right\}.$$

   2. (d). Update $\hat{F}^{[m]}(\mathbf{x})$ for each region $\widehat{R}_l^{[m]}$, $l = 1, 2, \ldots, L$

   $$\hat{F}^{[m]}(\mathbf{x}) = \hat{F}^{[m-1]}(\mathbf{x}) + \nu \hat{\eta}_l^{[m]} I(\mathbf{x} \in \widehat{R}_l^{[m]}) \quad l = 1, 2, \ldots, L.$$

3. Report $\hat{F}^{[M]}(\mathbf{x})$ as the final estimate.

4.2 Estimating (ρ, ϕ) via Profile Likelihood

Following Dunn and Smyth (2005), we use the profile likelihood to estimate the dispersion ϕ and the index parameter ρ, which jointly determine the mean–variance relation $\text{var}(Y_i) = \phi\mu_i^\rho / w_i$ of the pure premium. We exploit the fact that in Tweedie models the estimation of μ depends only on ρ: given a fixed ρ, the mean estimate $\mu^*(\rho)$ can be solved in (11) without knowing ϕ. Then conditional on this ρ and the corresponding $\mu^*(\rho)$, we maximize the log-likelihood function with respect to ϕ by

$$\phi^*(\rho) = \underset{\phi}{\text{argmax}}\left\{\ell(\mu^*(\rho), \phi, \rho)\right\}, \tag{20}$$

which is a univariate optimization problem that can be solved using a combination of golden section search and successive parabolic interpolation (Brent 2013). In such a way, we have determined the corresponding $(\mu^*(\rho), \phi^*(\rho))$ for each fixed ρ. Then, we acquire the estimate of ρ by maximizing the profile likelihood with respect to 50 equally spaced values $\{\rho_1, \ldots, \rho_{50}\}$ on $(0, 1)$:

$$\rho^* = \underset{\rho \in \{\rho_1, \ldots, \rho_{50}\}}{\text{argmax}} \left\{\ell(\mu^*(\rho), \phi^*(\rho), \rho)\right\}. \tag{21}$$

Finally, we apply ρ^* in (11) and (20) to obtain the corresponding estimates $\mu^*(\rho^*)$ and $\phi^*(\rho^*)$. Some additional computational issues for evaluating the log-likelihood functions in (20) and (21) are discussed in Appendix B.

4.3 Model Interpretation

Compared to other nonparametric statistical learning methods such as neural networks and kernel machines, our new estimator provides interpretable results. In this section, we discuss some ways for model interpretation after fitting the boosted Tweedie model.

4.3.1 Marginal Effects of Predictors. The main effects and interaction effects of the variables in the boosted Tweedie model can be extracted easily. In our estimate we can control the order of interactions by choosing the tree size L (the number of terminal nodes) and the number p of predictors. A tree with L terminal nodes produces a function approximation of p predictors with interaction order of at most $\min(L-1, p)$. For example, a stump ($L = 2$) produces an additive TDboost model with only the main effects of the predictors, since it is a function based on a single splitting variable in each tree. Setting $L = 3$ allows both main effects and second-order interactions.

Following Friedman (2001) we use the so-called partial dependence plots to visualize the main effects and interaction effects. Given the training data $\{y_i, \mathbf{x}_i\}_{i=1}^n$, with a p-dimensional input vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)^\intercal$, let $\mathbf{z}_s$ be a subset of size s, such that $\mathbf{z}_s = \{z_1, \ldots, z_s\} \subset \{x_1, \ldots, x_p\}$. For example, to study the main effect of the variable j, we set the subset $\mathbf{z}_s = \{z_j\}$, and to study the second order interaction of variables i and j, we set $\mathbf{z}_s = \{z_i, z_j\}$. Let $\mathbf{z}_{\backslash s}$ be the complement set of $\mathbf{z}_s$, such that $\mathbf{z}_{\backslash s} \cup \mathbf{z}_s = \{x_1, \ldots, x_p\}$. Let the prediction $\hat{F}(\mathbf{z}_s|\mathbf{z}_{\backslash s})$ be a function of the subset $\mathbf{z}_s$ conditioned on specific values of $\mathbf{z}_{\backslash s}$. The partial dependence of $\hat{F}(\mathbf{x})$ on $\mathbf{z}_s$ then can be formulated as $\hat{F}(\mathbf{z}_s|\mathbf{z}_{\backslash s})$ averaged over the marginal density of the complement subset $\mathbf{z}_{\backslash s}$

$$\hat{F}_s(\mathbf{z}_s) = \int \hat{F}(\mathbf{z}_s|\mathbf{z}_{\backslash s}) p_{\backslash s}(\mathbf{z}_{\backslash s}) d\mathbf{z}_{\backslash s}, \tag{22}$$

where $p_{\backslash s}(\mathbf{z}_{\backslash s}) = \int p(\mathbf{x}) d\mathbf{z}_s$ is the marginal density of $\mathbf{z}_{\backslash s}$. We estimate (22) by

$$\bar{F}_s(\mathbf{z}_s) = \frac{1}{n} \sum_{i=1}^{n} \hat{F}(\mathbf{z}_s|\mathbf{z}_{\backslash s,i}), \tag{23}$$

where $\{\mathbf{z}_{\backslash s,i}\}_{i=1}^n$ are evaluated at the training data. We then plot $\bar{F}_s(\mathbf{z}_s)$ against $\mathbf{z}_s$. We have included the partial dependence plot function in our R package "TDboost." We will demonstrate this functionality in Section 6.

4.3.2 Variable Importance. In many applications, identifying relevant predictors of the model in the context of tree-based ensemble methods is of interest. The TDboost model defines a variable importance measure for each candidate predictor X_j in the set $X = \{X_1, \ldots, X_p\}$ in terms of prediction/explanation of the response Y. The major advantage of this variable selection procedure, as compared to univariate screening methods, is that the approach considers the impact of each individual predictor as well as multivariate interactions among predictors simultaneously.

We start by defining the variable importance (VI henceforth) measure in the context of a single tree. First introduced by Breiman et al. (1984), the VI measure $\mathcal{I}_{X_j}(T_m)$ of the variable X_j in a single tree T_m is defined as the total heterogeneity reduction of the response variable Y produced by X_j, which can be estimated by adding up all the decreases in the squared error reductions $\hat{\delta}_l$ obtained in all $L-1$ internal nodes when X_j is chosen as the splitting variable. Denote $v(X_j) = l$ the event that X_j is selected as the splitting variable in the internal node l, and let $I_{jl} = I(v(X_j) = l)$. Then

$$\mathcal{I}_{X_j}(T_m) = \sum_{l=1}^{L-1} \hat{\delta}_l I_{jl}, \tag{24}$$

where $\hat{\delta}_l$ is defined as the squared error difference between the constant fit and the two subregion fits (the sub-region fits are achieved by splitting the region associated with the internal node l into the left and right regions). Friedman (2001) extended the VI measure $\mathcal{I}_{X_j}$ for the boosting model with a combination of M regression trees, by averaging (24) over $\{T_1, \ldots, T_M\}$:

$$\mathcal{I}_{X_j} = \frac{1}{M} \sum_{m=1}^{M} \mathcal{I}_{X_j}(T_m). \tag{25}$$

Despite of the wide use of the VI measure, Breiman et al. (1984) and White and Liu (1994) among others have pointed out that the VI measures (24) and (25) are biased: even if X_j is a noninformative variable to Y (not correlated to Y), X_j may still be selected as a splitting variable, hence the VI measure of X_j is nonzero by Eq. (25). Following Sandri and Zuccolotto (2008) and Sandri and Zuccolotto (2010) to avoid the variable selection bias, in this article we compute an adjusted VI measure for each explanatory variable by permutating each X_j, the computational details are provided in Appendix C.

4.4 Implementation

We have implemented our proposed method in an R package "TDboost," which is publicly available from the Comprehensive R Archive Network at *http://cran.r-project.org/web/packages/TDboost/index.html*. Here, we discuss the choice of three meta parameters in Algorithm 1: L (the size of the trees), ν (the shrinkage factor), and M (the number of boosting steps).

To avoid overfitting and improve out-of-sample predictions, the boosting procedure can be regularized by limiting the number of boosting iterations M (early stopping; Zhang and Yu 2005) and the shrinkage factor ν. Empirical evidence (Friedman 2001; Bühlmann and Hothorn 2007; Ridgeway 2007) showed that the predictive accuracy is almost always better with a smaller shrinkage factor at the cost of more computing time. However, smaller values of ν usually requires a larger number of boosting iterations M and hence induces more computing time (Friedman 2001). We choose a "sufficiently small" $\nu = 0.005$ throughout and determine M by the data.

The value L should reflect the true interaction order in the underlying model, but we almost never have such prior knowledge. Therefore, we choose the optimal M and L using K-fold cross-validation, starting with a fixed value of L. The data are split into K roughly equal-sized folds. Let an index function $\pi(i) : \{1, \ldots, n\} \mapsto \{1, \ldots, K\}$ indicate the fold to which observation i is allocated. Each time, we remove the kth fold of the data ($k = 1, 2, \dots, K$), and train the model using the remaining $K - 1$ folds. Denoting by $\hat{F}^{[M]}_{-k}(\mathbf{x})$ the resulting model, we compute the validation loss by predicting on each kth fold of the data removed:

$$\mathrm{CV}(M, L) = \frac{1}{n} \sum_{i=1}^{n} \Psi\left(y_i, \hat{F}^{[M]}_{-\pi(i)}(\mathbf{x}_i; L) \mid \rho\right). \quad (26)$$

We select the optimal M at which the minimum validation loss is reached

$$\widehat{M}_L = \underset{M}{\operatorname{argmin}}\, \mathrm{CV}(M, L).$$

If we need to select L too, then we repeat the whole process for several L (e.g., $L = 2, 3, 4, 5$) and choose the one with the smallest minimum generalization error

$$\widehat{L} = \underset{L}{\operatorname{argmin}}\, \mathrm{CV}(L, \widehat{M}_L).$$

For a given ν, fitting trees with higher L leads to smaller M being required to reach the minimum error.

5. SIMULATION STUDIES

In this section, we compare TDboost with the Tweedie GLM model (TGLM: Jørgensen and de Souza 1994) and the Tweedie GAM model in terms of the function estimation performance. The Tweedie GAM model was proposed by Wood (2001), which is based on a penalized regression spline approach with automatic smoothness selection. There is an R package "MGCV" accompanying the work, available at *http://cran.r-project.org/web/packages/mgcv/index.html*. In all numerical examples below using the TDboost model, five-fold cross-validation is adopted for selecting the optimal (M, L) pair, while the shrinkage factor ν is set to its default value of 0.005.

5.1 Case I

In this simulation study, we demonstrate that TDboost is well suited to fit target functions that are nonlinear or involve complex interactions. We consider two true target functions:

- Model 1 (Discontinuous function): The target function is discontinuous as defined by $F(x) = 0.5I(x > 0.5)$. We assume $x \sim \mathrm{Unif}(0, 1)$, and $y \sim \mathrm{Tw}(\mu, \phi, \rho)$ with $\rho = 1.5$ and $\phi = 0.5$.
- Model 2 (Complex interaction): The target function has two hills and two valleys:

$$F(x_1, x_2) = e^{-5(1-x_1)^2 + x_2^2} + e^{-5x_1^2 + (1-x_2)^2},$$

which corresponds to a common scenario where the effect of one variable changes depending on the effect of another. We assume $x_1, x_2 \sim \mathrm{Unif}(0, 1)$, and $y \sim \mathrm{Tw}(\mu, \phi, \rho)$ with $\rho = 1.5$ and $\phi = 0.5$.

We generate $n = 1000$ observations for training and $n' = 1000$ for testing, and fit the training data using TDboost, MGCV, and TGLM. Since the true target functions are known, we consider the mean absolute deviation (MAD) as performance criteria,

$$\mathrm{MAD} = \frac{1}{n'} \sum_{i=1}^{n'} |F(\mathbf{x}_i) - \hat{F}(\mathbf{x}_i)|,$$

where both the true predictor function $F(\mathbf{x}_i)$ and the predicted function $\hat{F}(\mathbf{x}_i)$ are evaluated on the test set. The resulting MADs on the testing data are reported in Table 1, which are averaged over 100 independent replications. The fitted functions from Model 2 are plotted in Figure 2. In both cases, we find that TDboost outperforms TGLM and MGCV in terms of the ability to recover the true functions and gives the smallest prediction errors.

Table 1. The averaged MADs and the corresponding standard errors based on 100 independent replications

Model	TGLM	MGCV	TDboost
1	0.1102 (0.0006)	0.0752 (0.0016)	0.0595 (0.0021)
2	0.3516 (0.0009)	0.2511 (0.0004)	0.1034 (0.0008)

5.2 Case II

The idea is to see the performance of the TDboost estimator and MGCV estimator on a variety of very complicated, randomly generated predictor functions, and study how the size of the training set, distribution settings and other characteristics of problems affect final performance of the two methods. We use the "random function generator" (RFG) model by Friedman (2001) in our simulation. The true target function F is randomly generated as a linear expansion of functions $\{g_k\}_{k=1}^{20}$:

$$F(\mathbf{x}) = \sum_{k=1}^{20} b_k g_k(\mathbf{z}_k). \quad (27)$$

Here, each coefficient b_k is a uniform random variable from $\mathrm{Unif}[-1, 1]$. Each $g_k(\mathbf{z}_k)$ is a function of $\mathbf{z}_k$, where $\mathbf{z}_k$ is defined as a p_k-sized subset of the 10-dimensional variable $\mathbf{x}$ in the form

$$\mathbf{z}_k = \{x_{\psi_k(j)}\}_{j=1}^{p_k}, \quad (28)$$

where each ψ_k is an independent permutation of the integers $\{1, \dots, p\}$. The size p_k is randomly selected by $\min(\lfloor 2.5 + r_k \rfloor, p)$, where r_k is generated from an exponential distribution with mean 2. Hence, the expected order of interactions presented in each $g_k(\mathbf{z}_k)$ is between four and five. Each function $g_k(\mathbf{z}_k)$ is a p_k-dimensional Gaussian function:

$$g_k(\mathbf{z}_k) = \exp\left\{-\frac{1}{2}(\mathbf{z}_k - \mathbf{u}_k)^\intercal \mathbf{V}_k (\mathbf{z}_k - \mathbf{u}_k)\right\}, \quad (29)$$

where each mean vector $\mathbf{u}_k$ is randomly generated from $\mathrm{N}(0, \mathbf{I}_{p_k})$. The $p_k \times p_k$ covariance matrix $\mathbf{V}_k$ is defined by

$$\mathbf{V}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{U}_k^\intercal, \quad (30)$$

where $\mathbf{U}_k$ is a random orthonormal matrix, $\mathbf{D}_k = \mathrm{diag}\{d_k[1], \dots, d_k[p_k]\}$, and the square root of each diagonal element $\sqrt{d_k[j]}$ is a uniform random variable from $\mathrm{Unif}[0.1, 2.0]$. We generate data $\{y_i, \mathbf{x}_i\}_{i=1}^{n}$ according to

$$y_i \sim \mathrm{Tw}(\mu_i, \phi, \rho), \quad \mathbf{x}_i \sim \mathrm{N}(0, \mathbf{I}_p), \quad i = 1, \dots, n, \quad (31)$$

where $\mu_i = \exp\{F(\mathbf{x}_i)\}$.

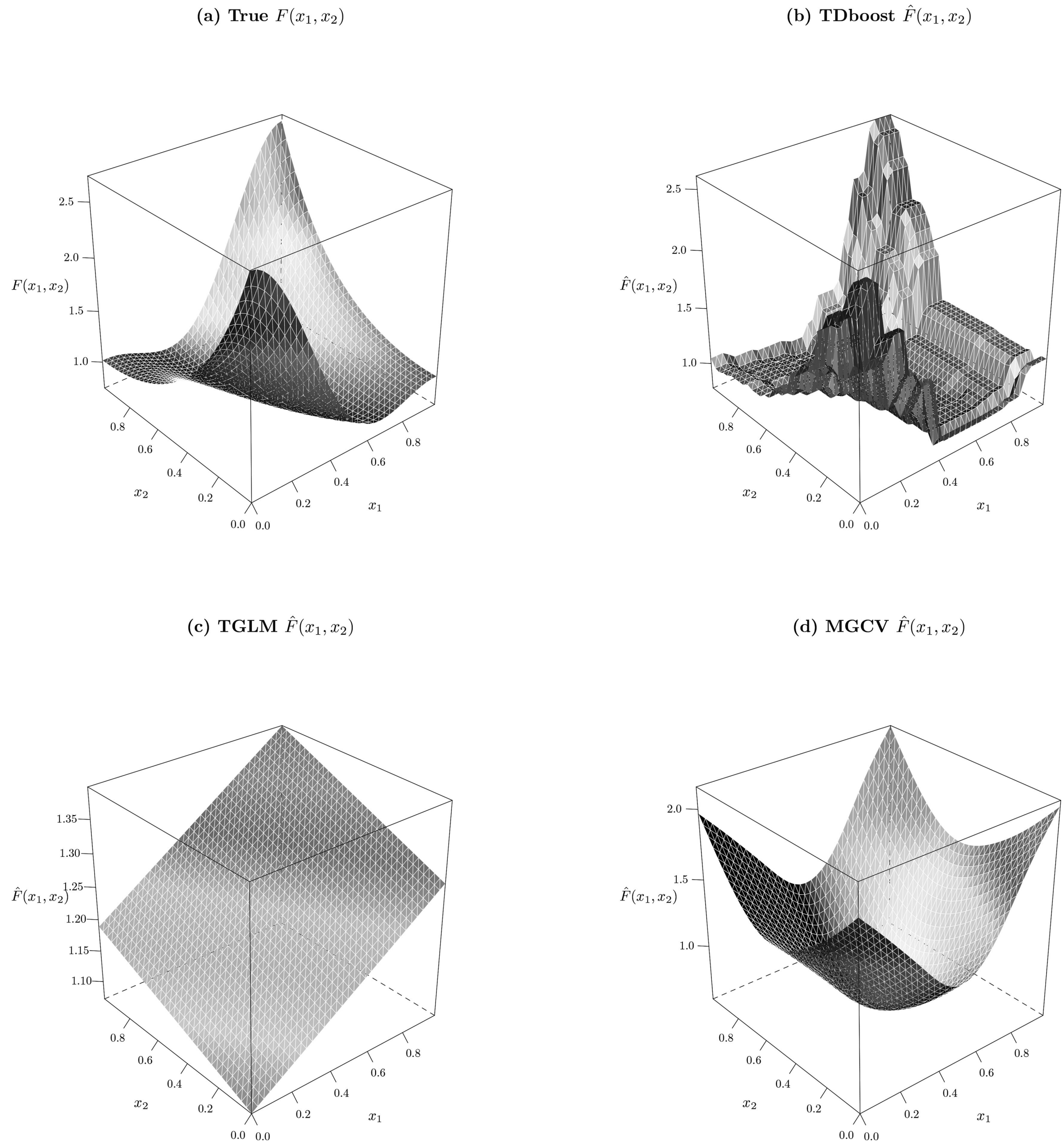

Figure 2. Fitted curves that recover the target function defined in Model 2. The top left figure shows the true target function. The top right, bottom left, and bottom right figures show the predictions on the testing data from TDboost, TGLM, and MGCV, respectively.

Setting I: when the index is known.. First, we study the situation that the true index parameter ρ is known when fitting models. We generate data according to the RFG model with index parameter $\tilde{\rho} = 1.5$ and the dispersion parameter $\tilde{\phi} = 1$ in the true model. We set the number of predictors to be $p = 10$ and generate $n \in \{1000, 2000, 5000\}$ observations as training sets, on which both MGCV and TDboost are fitted with ρ specified to be the true value 1.5. An additional test set of $n' = 5000$ observations was generated for evaluating the performance of the final estimate.

Figure 3 shows simulation results for comparing the estimation performance of MGCV and TDboost, when varying the training sample size. The empirical distributions of the MADs shown as box-plots are based on 100 independent replications. We can see that in all of the cases, TDboost outperforms MGCV in terms of prediction accuracy.

We also test estimation performance on μ when the index parameter ρ is misspecified, that is, we use a guess value ρ differing from the true value $\tilde{\rho}$ when fitting the TDboost model. Because μ is statistically orthogonal to ϕ and ρ, meaning that the off-diagonal elements of the Fisher information matrix are zero (Jørgensen 1997), we expect $\hat{\mu}$ will vary very slowly as ρ changes. Indeed, using the previous simulation data with the true value $\tilde{\rho} = 1.5$ and $\tilde{\phi} = 1$, we fitted TDboost models with nine guess values of $\rho \in \{1.1, 1.2, \ldots, 1.9\}$. The resulting MADs are displayed in Figure 4, which shows the choice of the value ρ has almost no significant effect on estimation accuracy of μ.

Setting II: using the estimated index. Next, we study the situation that the true index parameter ρ is unknown, and we use the estimated ρ obtained from the profile likelihood procedure discussed in Section 4.2 for fitting the model. The same data generation scheme is adopted as in Setting I, except now both

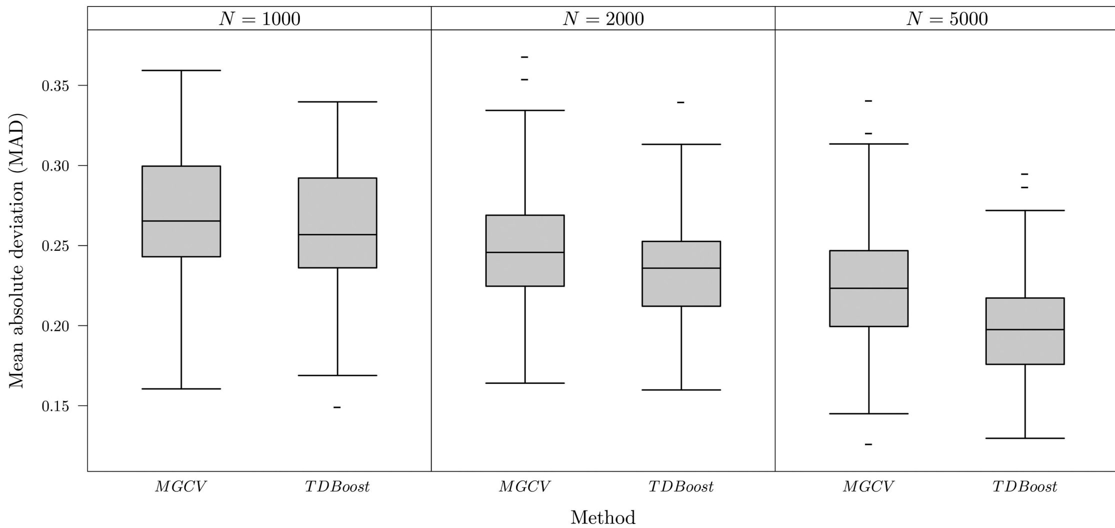

Figure 3. Simulation results for Setting I: compare the estimation performance of MGCV and TDboost when varying the training sample size and the dispersion parameter in the true model. Boxplots display empirical distributions of the MADs based on 100 independent replications.

MGCV and TDboost are fitted with ρ estimated by maximizing the profile likelihood. Figure 5 shows simulation results for comparing the estimation performance of MGCV and TDboost in such setting. We can see that the results have no significant difference to the results of Setting I: TDboost still outperforms MGCV in terms of prediction accuracy when using the estimated ρ instead of the true value.

Finally, we demonstrate our results from the estimation of the dispersion ϕ and the index ρ by using the profile likelihood. A total number of 200 sets of training samples are randomly generated from a true model according to the setting (31) with $\phi = 2$ and $\rho = 1.7$, each sample having 2000 observations. We fit the TDboost model on each sample and compute the estimates ϕ^* at each of the 50 equally spaced values $\{\rho_1, \ldots, \rho_{50}\}$ on $(1, 2)$. The $(\rho_j, \phi^*(\rho_j))$ corresponding to the maximal profile likelihood is the estimate of (ρ, ϕ). The estimation process is repeated 200 times. The estimated indices have mean $\overline{\rho^*} = 1.68$ and standard error $\mathrm{SE}(\rho^*) = 0.026$, so the true value $\rho = 1.7$ is within $\overline{\rho^*} \pm \mathrm{SE}(\rho^*)$. The estimated dispersions have mean $\overline{\phi^*} = 1.82$ and standard error $\mathrm{SE}(\phi^*) = 0.12$. Figure 6 shows the profile likelihood function of ρ for a single run.

6. APPLICATION: AUTOMOBILE CLAIMS

6.1 Dataset

We consider an auto-insurance claim dataset as analyzed in Yip and Yau (2005) and Zhang and Yu (2005). The dataset contains 10,296 driver vehicle records, each record including an individual driver's total claim amount (z_i) in the last five years ($w_i = 5$) and 17 characteristics $x_i = (x_{i,1}, \ldots, x_{i,17})$ for the driver and the insured vehicle. We want to predict the ex-

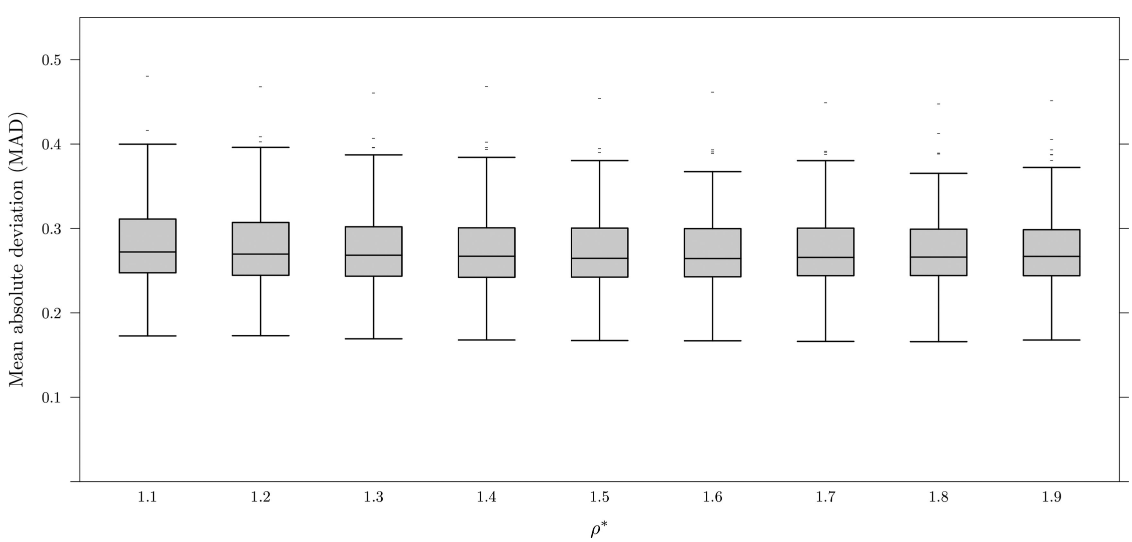

Figure 4. Simulation results for Setting I when the index is misspecified: the estimation performance of TDboost when varying the value of the index parameter $\rho \in \{1.1, 1.2, \ldots, 1.9\}$. In the true model $\tilde{\rho} = 1.5$ and $\tilde{\phi} = 1$. Boxplots show empirical distributions of the MADs based on 200 independent replications.

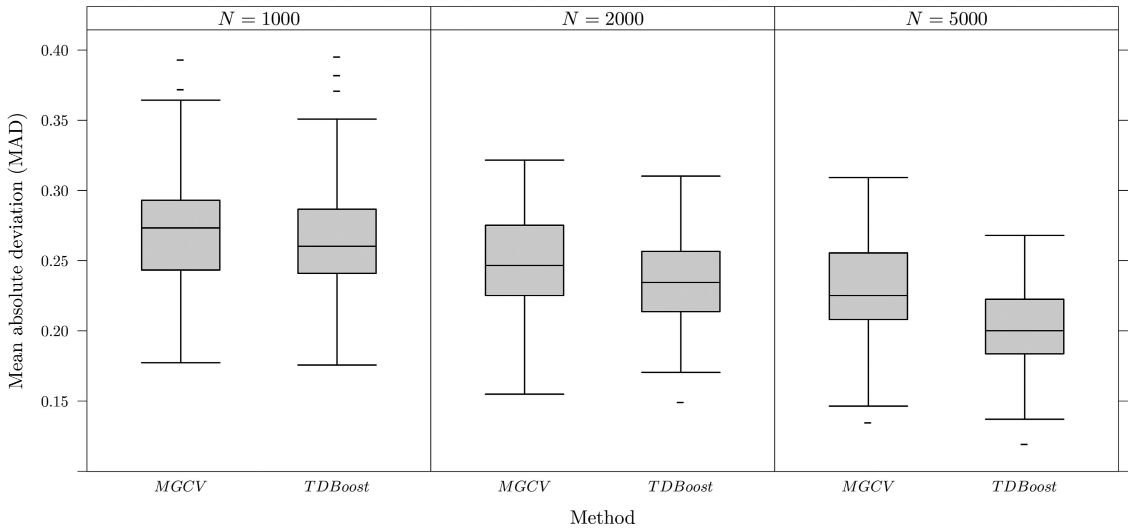

Figure 5. Simulation results for Setting II: compare the estimation performance of MGCV and TDboost when varying the training sample size and the dispersion parameter in the true model. Boxplots display empirical distributions of the MADs based on 100 independent replications.

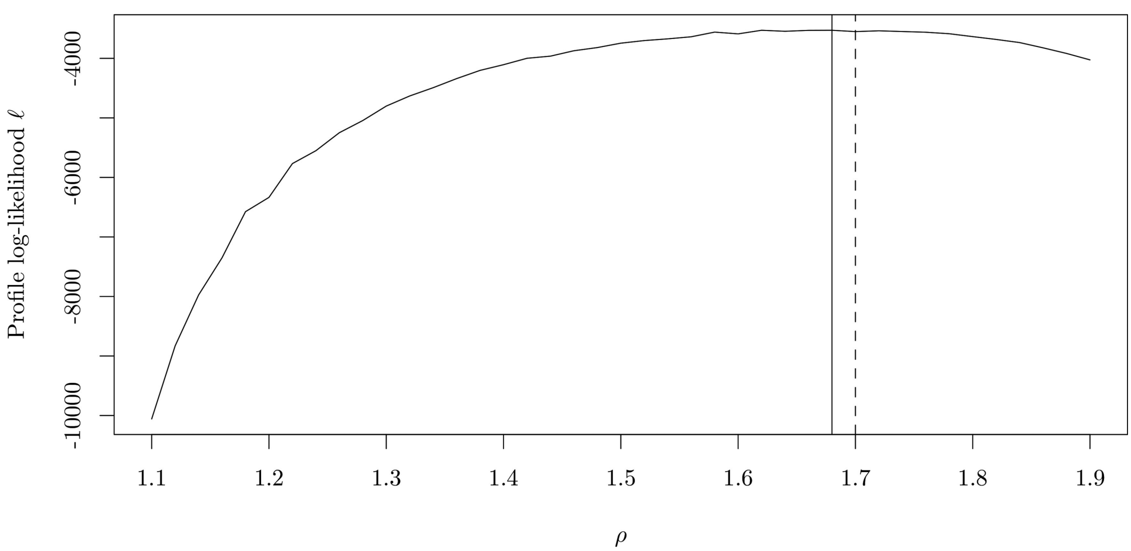

Figure 6. The curve represents the profile likelihood function of ρ from a single run. The dotted line shows the true value $\rho = 1.7$. The solid line shows the estimated value $\rho^* = 1.68$ corresponding to the maximum likelihood. The associated estimated dispersion is ϕ^* =1.89.

pected pure premium based on x_i. Table 3 summarize the dataset. The descriptive statistics of the data are provided in Appendix D. The histogram of the total claim amounts in Figure 1 shows that the empirical distribution of these values is highly skewed. We find that approximately 61.1% of policyholders had no claims, and approximately 29.6% of the policyholders had a positive claim amount up to \$10,000. Note that only 9.3% of the policyholders had a high claim amount above \$10,000, but the sum of their claim amount made up to 64% of the overall sum. Another important feature of the data is that there are interactions among explanatory variables. For example, from Table 2 we can see that the marginal effect of the variable REVOKED on the total claim amount is much greater for the policyholders living in the urban area than those living in the rural area. The importance of the interaction effects will be confirmed later in our data analysis.

6.2 Models

We separate the entire dataset into a training set and a testing set with equal size. Then, the TDboost model is fitted on the training set and tuned with five-fold cross-validation. For comparison, we also fit TGLM and MGCV, both of which are

Table 2. The averaged total claim amount for different categories of the policyholders

		AREA	
		Urban	Rural
REVOKED	No	3150.57	904.70
	Yes	14551.62	7624.36
Difference		11401.05	6719.66

Table 3. Explanatory variables in the claim history dataset. Type N stands for numerical variable, Type C stands for categorical variable

ID	Variable	Type	Description
1	AGE	N	Driver's age
2	BLUEBOOK	N	Value of vehicle
3	HOMEKIDS	N	Number of children
4	KIDSDRIV	N	Number of driving children
5	MVR_PTS	N	Motor vehicle record points
6	NPOLICY	N	Number of policies
7	RETAINED	N	Number of years as a customer
8	TRAVTIME	N	Distance to work
9	AREA	C	Home/work area: Rural, Urban
10	CAR_USE	C	Vehicle use: Commercial, Private
11	CAR_TYPE	C	Type of vehicle: Panel Truck, Pickup, Sedan, Sports Car, SUV, Van
12	GENDER	C	Driver's gender: F, M
13	JOBCLASS	C	Unknown, Blue Collar, Clerical, Doctor, Home Maker, Lawyer, Manager, Professional, Student
14	MAX_EDUC	C	Education level: High School or Below, Bachelors, High School, Masters, PhD
15	MARRIED	C	Married or not: Yes, No
16	REVOKED	C	Whether license revoked in past 7 years: Yes, No

fitted using all the explanatory variables. In MGCV, the numerical variables AGE, BLUEBOOK, HOMEKIDS, KIDSDRIV, MVR_PTS, NPOLICY, RETAINED, and TRAVTIME are modeled by smooth terms represented using penalized regression splines. We find the appropriate smoothness for each applicable model term using generalized cross-validation (GCV; Wahba 1990). For the TDboost model, it is not necessary to carry out data transformation, since the tree-based boosting method can automatically handle different types of data. For other models, we use logarithmic transformation on BLUEBOOK, that is, log(BLUEBOOK), and scale all the numerical variables except for HOMEKIDS, KIDSDRIV, MVR_PTS, and NPOLICY to have mean 0 and standard deviation 1. We also create dummy variables for the categorical variables with more than two levels (CAR_TYPE, JOBCLASS, and MAX_EDUC). For all models, we use the profile likelihood method to estimate the dispersion ϕ and the index ρ, which are in turn used in fitting the final models.

Table 4. The averaged Gini indices and standard errors in the auto-insurance claim data example based on 20 random splits

	Competing premium		
Base premium	TGLM	MGCV	TDboost
TGLM	0	7.833 (0.338)	15.528 (0.509)
MGCV	3.044 (0.610)	0	12.979 (0.473)
TDboost	4.000 (0.364)	3.540 (0.415)	0

6.3 Performance Comparison

To examine the performance of TGLM, MGCV, and TDboost, after fitting on the training set, we predict the pure premium $P(\mathbf{x}) = \hat{\mu}(\mathbf{x})$ by applying each model on the independent held-out testing set. However, attention must be paid when measuring the differences between predicted premiums $P(\mathbf{x})$ and real losses y on the testing data. The mean squared loss or mean absolute loss is not appropriate here because the losses have high proportions of zeros and are highly right skewed. Therefore, an alternative statistical measure—the ordered Lorentz curve and the associated Gini index—proposed by Frees, Meyers, and Cummings (2011) are used for capturing the discrepancy between the premium and loss distributions. By calculating the

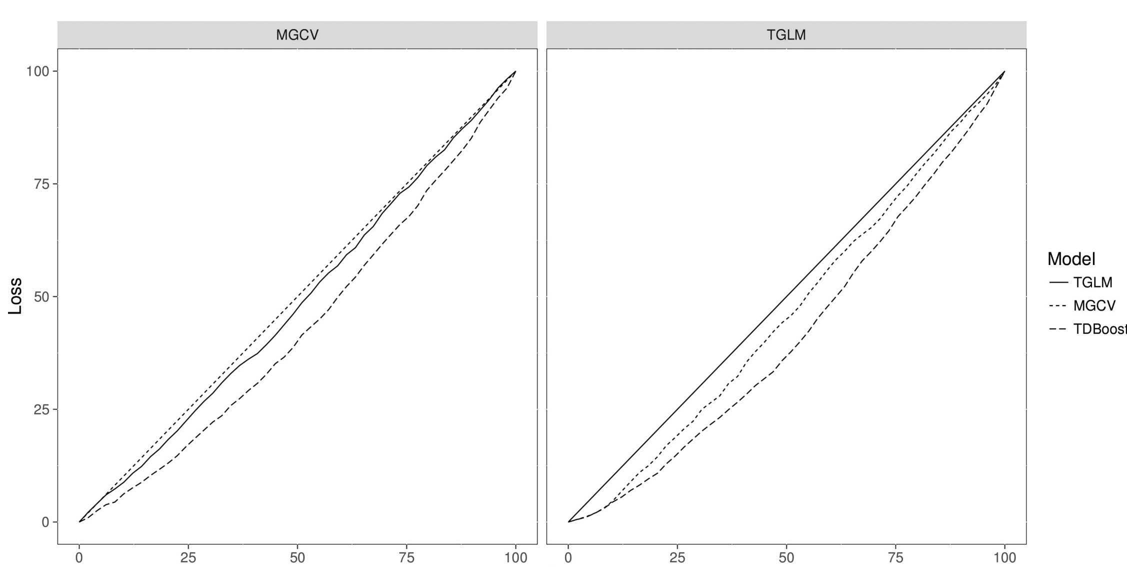

Figure 7. The ordered Lorentz curves for the auto-insurance claim data.

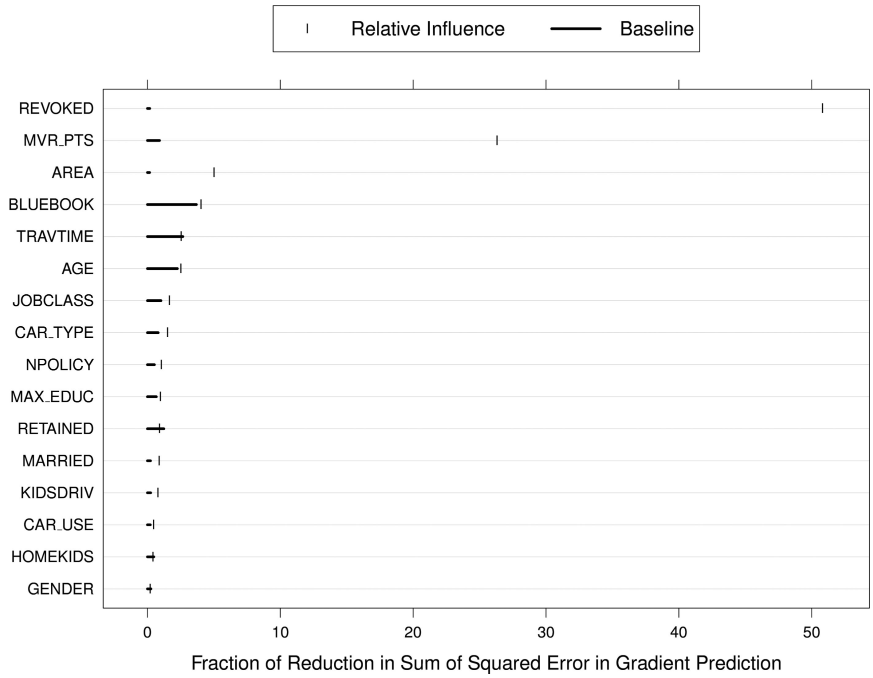

Figure 8. The variable importance measures and baselines of 17 explanatory variables for modeling the pure premium.

Gini index, the performance of different predictive models can be compared. Here, we only briefly explain the idea of the ordered Lorentz curve (Frees, Meyers, and Cummings 2011, 2014). Let $B(\mathbf{x})$ be the "base premium," which is calculated using the existing premium prediction model, and let $P(\mathbf{x})$ be the "competing premium" calculated using an alternative premium prediction model. In the ordered Lorentz curve, the distribution of losses and the distribution of premiums are sorted based on

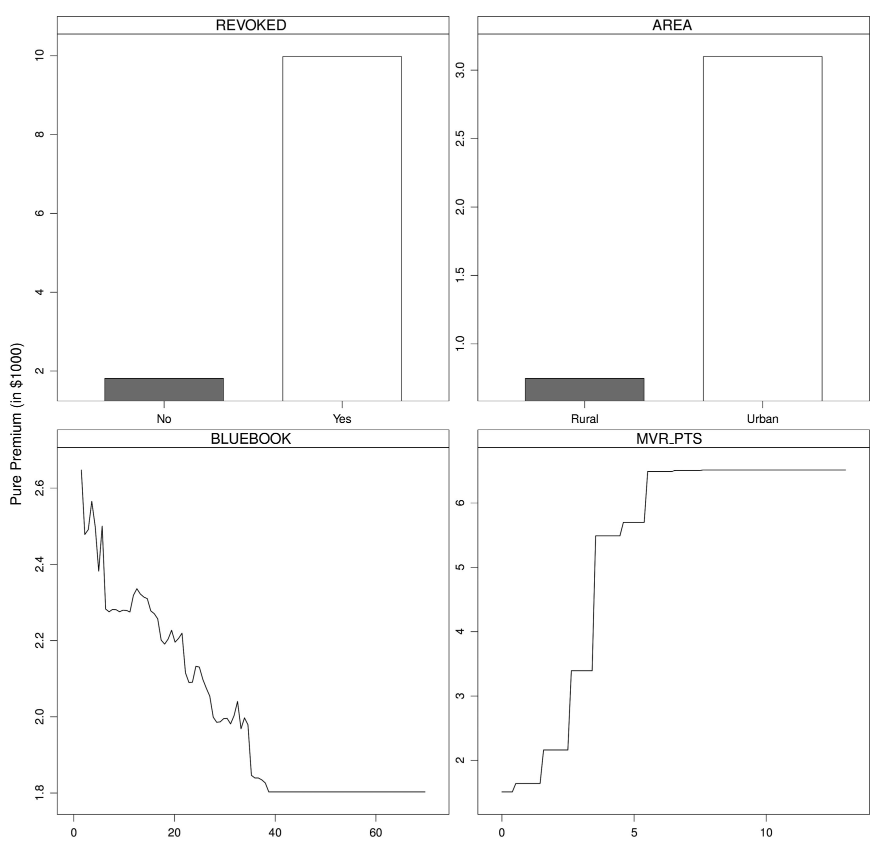

Figure 9. Marginal effects of four most significant explanatory variables on the pure premium.

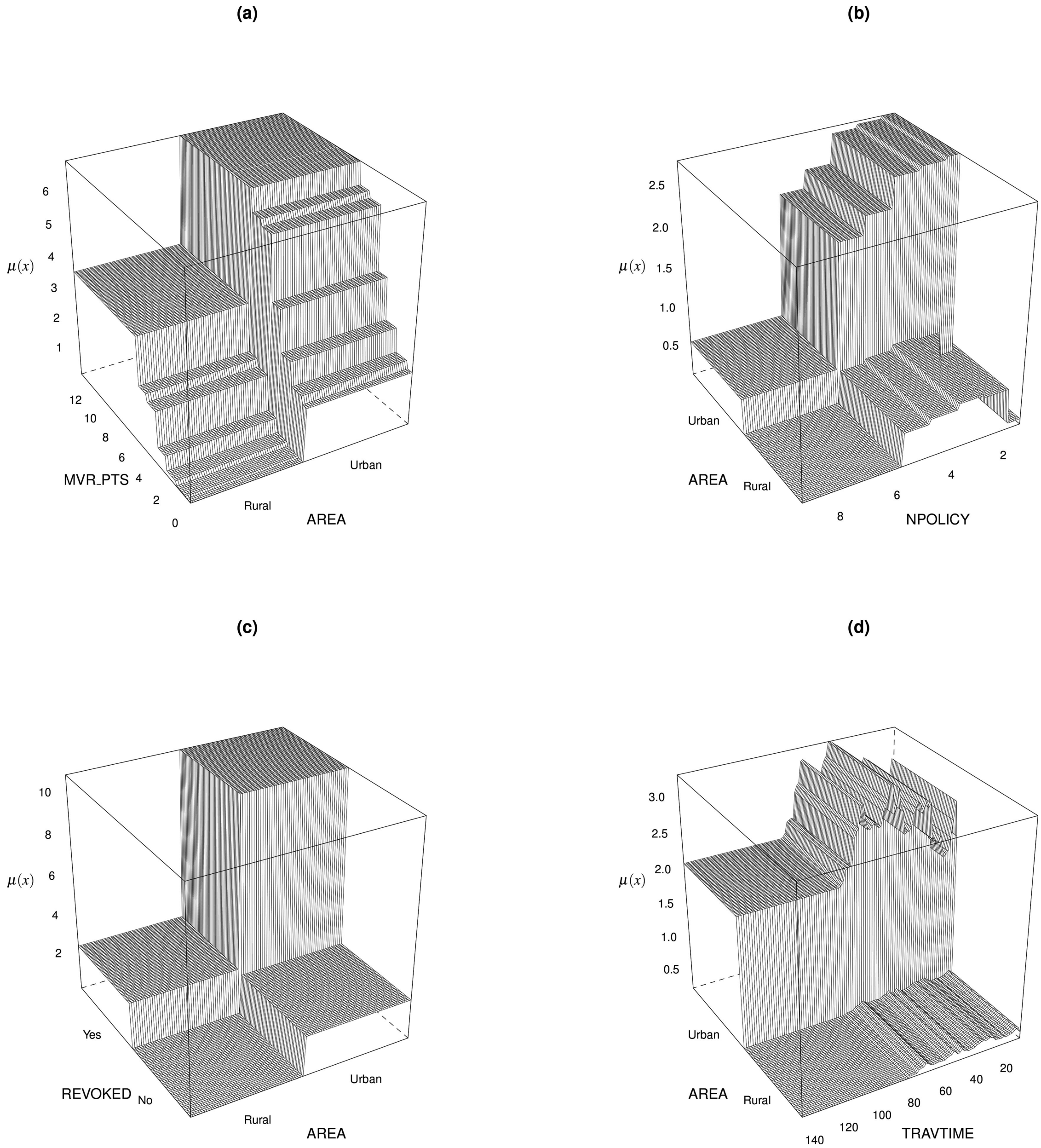

Figure 10. Four strong pairwise interactions.

the relative premium $R(\mathbf{x}) = P(\mathbf{x})/B(\mathbf{x})$. The ordered premium distribution is

$$\hat{D}_P(s) = \frac{\sum_{i=1}^{n} B(\mathbf{x}_i) I(R(\mathbf{x}_i) \le s)}{\sum_{i=1}^{n} B(\mathbf{x}_i)},$$

and the ordered loss distribution is

$$\hat{D}_L(s) = \frac{\sum_{i=1}^{n} y_i I(R(\mathbf{x}_i) \le s)}{\sum_{i=1}^{n} y_i}.$$

Two empirical distributions are based on the same sort order, which makes it possible to compare the premium and loss distributions for the same policyholder group. The ordered Lorentz curve is the graph of $(\hat{D}_P(s), \hat{D}_L(s))$. When the percentage of losses equals the percentage of premiums for the insurer, the curve results in a 45-degree line, known as "the line of equality." Twice the area between the ordered Lorentz curve and the line of equality measures the discrepancy between the premium and loss distributions, and is defined as the Gini index. Curves below the line of equality indicate that, given knowledge of the relative premium, an insurer could identify the profitable contracts, whose premiums are greater than losses. Therefore, a larger Gini index (hence a larger area between the line of equality and the curve below) would imply a more favorable model.

Following Frees, Meyers, and Cummings (2014), we successively specify the prediction from each model as the base premium $B(\mathbf{x})$ and use predictions from the remaining models as the competing premium $P(\mathbf{x})$ to compute the Gini indices. The entire procedure of the data splitting and Gini index computation are repeated 20 times, and a matrix of the averaged Gini indices and standard errors is reported in Table 4. To pick the "best" model, we use a "minimax" strategy (Frees, Meyers, and Cummings 2013) to select the base premium model that are least vulnerable to competing premium models; that is, we select the model that provides the smallest of the maximal Gini indices, taken over competing premiums. We find that the maximal Gini index is 15.528 when using $B(\mathbf{x}) = \hat{\mu}^{\text{TGLM}}(\mathbf{x})$ as the base premium, 12.979 when $B(\mathbf{x}) = \hat{\mu}^{\text{MGCV}}(\mathbf{x})$, and 4.000 when $B(\mathbf{x}) = \hat{\mu}^{\text{TDboost}}(\mathbf{x})$. Therefore, TDboost has the smallest maximum Gini index at 4.000, hence is the least vulnerable to alternative scores. Figure 7 also shows that when TGLM (or MGCV) is selected as the base premium, the area between the line of equality and the ordered Lorentz curve is larger when choosing TDboost as the competing premium, indicating again that the TDboost model represents the most favorable choice.

6.4 Interpreting the Results

Next, we focus on the analysis using the TDboost model. There are several explanatory variables significantly related to the pure premium. The VI measure and the baseline value of each explanatory variable are shown in Figure 8. We find that REVOKED, MVR_PTS, AREA, and BLUEBOOK have high VI measure scores (the vertical line), and their scores all surpass the corresponding baselines (the horizontal line-length), indicating that the importance of those explanatory variables is real. We also find the variables AGE, JOBCLASS, CAR_TYPE, NPOLICY, MAX_EDUC, MARRIED, KIDSDRIV, and CAR_USE have larger-than-baseline VI measure scores, but the absolute scales are much less than aforementioned four variables. On the other hand, although the VI measure of, for example, TRAVTIME is quite large, it does not significantly surpass the baseline importance.

We now use the partial dependence plots to visualize the fitted model. Figure 9 shows the main effects of four important explanatory variables on the pure premium. We clearly see that the strong nonlinear effects exist in predictors BLUEBOOK and MVR_PTS: for the policyholders whose vehicle values are below 40 K, their pure premium is negatively associated with the value of vehicle; after the value of vehicle passes 40 K, the pure premium curve reaches a plateau; additionally, the pure premium is positively associated with motor vehicle record points MVR_PTS, but the pure premium curve reaches a plateau when MVR_PTS exceeds six. On the other hand, the partial dependence plots suggest that a policyholder who lives in the urban area (AREA="URBAN") or with driver's license revoked (REVOKED="YES") typically has relatively high pure premium.

In our model, the data-driven choice for the tree size is $L = 7$, which means that our model includes higher order interactions. In Figure 10, we visualize the effects of four important second order interactions using the joint partial dependence plots. These four interactions are AREA $\times$ MVR_PTS, AREA $\times$ NPOLICY, AREA $\times$ REVOKED, and AREA $\times$ TRAVTIME. These four interactions all involve the variable AREA: we can see that the marginal effects of MVR_PTS, NPOLICY, REVOKED, and TRAVTIME on the pure premium are greater for the policyholders living in the urban area (AREA="URBAN") than those living in the rural area (AREA="RURAL"). For example, a strong AREA $\times$ MVR_PTS interaction suggests that for the policyholders living in the rural area, motor vehicle record points of the policyholders have a weaker positive marginal effect on the expected pure premium than for the policyholders living in the urban area.

7. CONCLUSIONS

The need for nonlinear risk factors as well as risk factor interactions for modeling insurance claim sizes is well-recognized by actuarial practitioners, but practical tools to study them are very limited. In this article, relying on neither the linear assumption nor a prespecified interaction structure, a flexible tree-based gradient boosting method is designed for the Tweedie model. We implement the proposed method in a user-friendly R package "TDboost" that can make accurate insurance premium predictions for complex datasets and serve as a convenient tool for actuarial practitioners to investigate the nonlinear and interaction effects. In the context of personal auto insurance, we implicitly use the policy duration as a volume measure (or exposure), and demonstrate the favorable prediction performance of TDboost for the pure premium. In cases that exposure measures other than duration are used, which is common in commercial insurance, we can extend the TDboost method to the corresponding claim size by simply replacing the duration with any chosen exposure measure.

TDboost can also be an important complement to the traditional GLM model in insurance rating. Even under the strict circumstances that the regulators demand the final model to have a GLM structure, our approach can still be quite helpful due to its ability to extract additional information such as nonmonotonicity/nonlinearity and important interaction. In Appendix E, we provide an additional real data analysis to demonstrate that our method can provide insights into the structure of interaction terms. After integrating the obtained information about the interaction terms into the original GLM model, we can much enhance the overall accuracy of the insurance premium prediction while maintaining a GLM model structure.

In addition, it is worth mentioning that the applications of the proposed method can go beyond the insurance premium prediction and be of interest to researchers in many other fields including ecology (Foster and Bravington 2013), meteorology (Dunn 2004), and political science (Lauderdale 2012). See, for example, Dunn and Smyth (2005) and Qian, Yang, and Zou (2016) for descriptions of the broad Tweedie distribution applications. The proposed method and the implementation tool allow researchers in these related fields to venture outside the Tweedie GLM modeling framework, build new flexible models from nonparametric perspectives, and use the model interpretation tools demonstrated in our real data analysis to study their own problems of interests.

SUPPLEMENTARY MATERIALS

The online supplementary materials contain:

- Part A: A Property of Tweedie Distributions
- Part B: Computational Issues for Profile Likelihood
- Part C: Bias-adjusted Variable Importance Measure

– Part D: Descriptive Statistics for Real Data
– Part E: Identifying Important Interactions

[Received June 2014. Revised April 2016.]

REFERENCES

Anstey, K. J., Wood, J., Lord, S., and Walker, J. G. (2005), "Cognitive, Sensory and Physical Factors Enabling Driving Safety in Older Adults," *Clinical Psychology Review*, 25, 45–65. [457]

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. [457]

——— (1998), "Arcing Classifier" (with discussion and a rejoinder by the author), *The Annals of Statistics*, 26, 801–849. [457]

——— (1999), "Prediction Games and Arcing Algorithms," *Neural Computation*, 11, 1493–1517. [457]

Breiman, L., Friedman, J., Olshen, R., Stone, C., Steinberg, D., and Colla, P. (1984), *CART: Classification and Regression Trees*, London, UK: Chapman and Hall/CRC. [457,461]

Brent, R. P. (2013), *Algorithms for Minimization Without Derivatives*, Englewood Cliffs, NJ: Prentice-Hall. [460]

Bühlmann, P., and Hothorn, T. (2007), "Boosting Algorithms: Regularization, Prediction and Model Fitting," *Statistical Science*, 22, 477–505. [457,461]

Dionne, G., Gouriéroux, C., and Vanasse, C. (2001), "Testing for Evidence of Adverse Selection in the Automobile Insurance Market: A Comment," *Journal of Political Economy*, 109, 444–453. [456]

Dunn, P. K. (2004), "Occurrence and Quantity of Precipitation can be Modelled Simultaneously," *International Journal of Climatology*, 24, 1231–1239. [469]

Dunn, P. K., and Smyth, G. K. (2005), "Series Evaluation of Tweedie Exponential Dispersion Model Densities," *Statistics and Computing*, 15, 267–280. [460,469]

Elith, J., Leathwick, J. R., and Hastie, T. (2008), "A Working Guide to Boosted Regression Trees," *Journal of Animal Ecology*, 77, 802–813. [457]

Foster, S. D., and Bravington, M. V. (2013), "A Poisson–Gamma Model for Analysis of Ecological Non-Negative Continuous Data," *Environmental and Ecological Statistics*, 20, 533–552. [469]

Frees, E. W., Meyers, G., and Cummings, A. D. (2011), "Summarizing Insurance Scores Using a Gini Index," *Journal of the American Statistical Association*, 106. [466]

Frees, E. W. J., Meyers, G., and Cummings, A. D. (2014), "Insurance Ratemaking and a Gini Index," *Journal of Risk and Insurance*, 81, 335–336. [467,468]

Freund, Y., and Schapire, R. (1997), "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 55, 119–139. [457]

Friedman, J. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232. [457,458,460,461,462]

Friedman, J., Hastie, T., and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical View of Boosting" (with discussion and a rejoinder by the authors), *The Annals of Statistics*, 28, 337–407. [457,460]

Friedman, J. H. (2002), "Stochastic Gradient Boosting," *Computational Statistics & Data Analysis*, 38, 367–378. [457]

Haberman, S., and Renshaw, A. E. (1996), "Generalized Linear Models and Actuarial Science," *Statistician*, 45, 407–436. [456]

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models* (Vol. 43), Boca Raton, FL: CRC Press. [457]

Hastie, T. J., Tibshirani, R. J., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Ed.), Springer Series in Statistics, New York: Springer. [457]

Jørgensen, B. (1987), "Exponential Dispersion Models," *Journal of the Royal Statistical Society,* Series B, 49, 127–162. [458]

——— (1997), *The Theory of Dispersion Models* (Vol. 76), Boca Raton, FL: CRC Press. [463]

Jørgensen, B., and de Souza, M. C. (1994), "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data," *Scandinavian Actuarial Journal*, 1994, 69–93. [456,458,462]

Lauderdale, B. E. (2012), "Compound Poisson–Gamma Regression Models for Dollar Outcomes that are Sometimes Zero," *Political Analysis*, 20, 387–399. [469]

Mildenhall, S. J. (1999), "A Systematic Relationship Between Minimum Bias and Generalized Linear Models," *Proceedings of the Casualty Actuarial Society*, 86, 393–487. [456]

Murphy, K. P., Brockman, M. J., and Lee, P. K. (2000), "Using Generalized Linear Models to Build Dynamic Pricing Systems," in *Casualty Actuarial Society Forum, Winter*, ed. D. L. Lange, Landover, MD: Colortone Press, pp. 107–139. [456]

Nelder, J., and Wedderburn, R. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society,* Series A, 135, 370–384. [456]

Ohlsson, E., and Johansson, B. (2010), *Non-Life Insurance Pricing With Generalized Linear Models*, New York: Springer. [459]

Peters, G. W., Shevchenko, P. V., and Wüthrich, M. V. (2008), "Model Risk in Claims Reserving Within Tweedie's Compound Poisson Models," *ASTIN Bulletin,* to appear. [456]

Qian, W., Yang, Y., and Zou, H. (2016), "Tweedie's Compound Poisson Model With Grouped Elastic Net," *Journal of Computational and Graphical Statistics*, 25, 606–625. [469]

Renshaw, A. E. (1994), "Modelling the Claims Process in the Presence of Covariates," *ASTIN Bulletin*, 24, 265–285. [456]

Ridgeway, G. (2007), "Generalized Boosted Regression Models," R package manual. [461]

Sandri, M., and Zuccolotto, P. (2008), "A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees," *Journal of Computational and Graphical Statistics*, 17, 611–628. [461]

——— (2010), "Analysis and Correction of Bias in Total Decrease in Node Impurity Measures for Tree-Based Algorithms," *Statistics and Computing*, 20, 393–407. [461]

Showers, V. E., and Shotick, J. A. (1994), "The Effects of Household Characteristics on Demand for Insurance: A Tobit Analysis," *Journal of Risk and Insurance*, 61, 492–503. [456]

Smyth, G. K. (1996), "Regression Analysis of Quantity Data With Exact Zeros," in *Proceedings of the Second Australia–Japan Workshop on Stochastic Models in Engineering, Technology and Management*, Citeseer, 572–580. [458,459]

Smyth, G. K., and Jørgensen, B. (2002), "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling," *ASTIN Bulletin*, 32, 143–157. [456,458]

Tweedie, M. (1984), "An Index Which Distinguishes Between Some Important Exponential Families," in *Statistics: Applications and New Directions: Proceedings Indian Statistical Institute Golden Jubilee International Conference*, 579–604. [458,459]

Van de Ven, W., and van Praag, B. M. (1981), "Risk Aversion and deductibles in Private Health Insurance: Application of an Adjusted Tobit Model to Family Health Care Expenditures," *Health, Economics, and Health Economics*, 125–148. [456]

Wahba, G. (1990), *Spline Models for Observational Data* (Vol. 59), Philadelphia, PA: SIAM. [466]

White, A. P., and Liu, W. Z. (1994), "Technical Note: Bias in Information-Based Measures in Decision Tree Induction," *Machine Learning*, 15, 321–329. [461]

Wood, S. (2001), "Mgcv: GAMs and Generalized Ridge Regression for R," *R News*, 1, 20–25. [462]

——— (2006), *Generalized Additive Models: An Introduction With R*, Boca Raton, FL: CRC press. [457]

Yip, K. C., and Yau, K. K. (2005), "On Modeling Claim Frequency Data in General Insurance With Extra Zeros," *Insurance: Mathematics and Economics*, 36, 153–163. [456,464]

Zhang, T., and Yu, B. (2005), "Boosting With Early Stopping: Convergence and Consistency," *The Annals of Statistics*, 1538–1579. [461,464]

Zhang, W. (2011), "Cplm: Monte Carlo EM Algorithms and Bayesian Methods for Fitting Tweedie Compound Poisson Linear Models," R package, *http://cran.r-project.org/web/packages/cplm/index.html*. [457]

A Supplemental Material for "Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models"

Yi Yang*, Wei Qian† and Hui Zou‡

April 20, 2016

Part A: A property of Tweedie distributions

For completeness, we give here a known result of the Tweedie distribution and its detailed proof.

Proposition 1. *Let $Z_i = \sum_{d_i=1}^{N_i} \tilde{Z}_{d_i}$ is the total claim amount. Let $Y_i = Z_i/w_i$, where ω_i is the duration. Assume N_i is Poisson distributed* $\mathrm{Pois}(\lambda_i w_i)$*. Conditional on N_i, assume Z_{d_i}'s ($d_i = 1, \dots, N_i$) are i.i.d.* $\mathrm{Gamma}(\alpha, \gamma_i)$*. Assume that under unit duration (i.e., $w_i = 1$), the mean-variance relation satisfies $Var(Y_i^*) = \phi[E(Y_i^*)]^\rho$, where Y_i^* is the pure premium under unit duration, ϕ is a constant, and $\rho = (\alpha+2)/(\alpha+1)$. Then for the pure premium Y_i under duration ω_i*

$$Y_i \sim \mathrm{Tw}(\mu_i, \phi/w_i, \rho).$$

Proof. Note that under unit duration $w_i = 1$,

$$\mu_i^* := E(Y_i^*) = E(E(Y_i^*|N_i)) = \lambda_i \alpha \gamma_i,$$
$$Var(Y_i^*) = E(Var(Y_i^*|N_i)) + Var(E(Y_i^*|N_i)) = \lambda_i \alpha \gamma_i^2 + \lambda_i \alpha^2 \gamma_i^2.$$

*McGill University

†Rochester Institute of Technology

‡Corresponding author, zoux019@umn.edu, University of Minnesota

Similarly, under any duration w_i,

$$\mu_i := E(Y_i) = \frac{1}{w_i} E(Z_i) = \lambda_i \alpha \gamma_i,$$
$$Var(Y_i) = \frac{1}{w_i^2} Var(Z_i) = (\lambda_i \alpha \gamma_i^2 + \lambda_i \alpha^2 \gamma_i^2)/w_i.$$

As a result, we can obtain the mean-variance relation for the pure premium Y_i that

$$Var(Y_i) = \frac{1}{w_i} Var(Y_i^*) = \frac{\phi}{w_i} (\mu_i^*)^\rho = \frac{\phi}{w_i} \mu_i^\rho, \tag{1}$$

where the second equation follows by

$$Var(Y_i^*) = \phi [E(Y_i^*)]^\rho. \tag{2}$$

By the scale-invariance property of Tweedie distribution, the proof is complete. □

Part B: Computational issues for profile likelihood

There are some computational issues, which must be taken care of when evaluating the log-likelihood functions in (20) and (21) of Section 4.2: since in general there are no closed forms for Tweedie densities, in likelihood evaluation one must deal with an infinite summation in the normalizing function $a(y, \phi, \rho) = \frac{1}{y} \sum_{t=1}^{\infty} W_t$. For numerical evaluation of Tweedie densities, Dunn and Smyth (2005) proposed a series expansions approach, which sums an infinite series arising from a Taylor expansion of the characteristic function. Alternatively, Dunn and Smyth (2008) developed a Fourier inversion approach, which consists of an inversion of the characteristic function based on numerical integration methods for oscillating functions. These two numerical methods turn out to be complementary since each has advantages under a certain situation: when only considering the case $1 < \rho < 2$, the series approach performs very well for small y but gradually loses computational efficiency as y increases, whereas the inversion approach performs very well for large y but gradually fails to provide accurate results as y decreases. Hence the inversion approach is preferred for large y and the series approach for small y. Dunn and Smyth (2008) provided a simple guideline to choose between the two methods. In this paper we use their R package "tweedie" (available at `http://cran.r-project.org/web/packages/tweedie/index.html`) for evaluating

Tweedie densities in our profile likelihood computation. For further details regarding their algorithms, the reader may refer to Dunn and Smyth (2005, 2008).

Part C: Bias-adjusted variable importance measure

Following Sandri and Zuccolotto (2008) and Sandri and Zuccolotto (2010), we compute the biased-adjusted VI measure for each explanatory variable in the following way:

(1) For $s = 1, \ldots, S$, repeat steps (2)–(4).

(2) Generate a matrix $\mathbf{z}^s$ by randomly permutating (without replacement) the n rows of the design matrix $\mathbf{x}$, while keeping the order of columns unchanged.

(3) Create an $n \times 2p$ matrix $\tilde{\mathbf{x}}^s = [\mathbf{x}, \mathbf{z}^s]$ by binding $\mathbf{z}^s$ with $\mathbf{x}$ matrix by column.

(4) Use the data $\{y, \tilde{\mathbf{x}}^s\}$ to fit the model, and compute VI measures $\mathcal{I}^s_{X_j}$ for X_j and $\mathcal{I}^s_{Z^s_j}$ for Z^s_j, where Z^s_j (jth column of Z^s) is the pseudo-predictor corresponding to X_j.

(5) Compute the VI measure $\overline{\mathcal{I}}_{X_j}$ as the average of $\mathcal{I}^s_{X_j}$ and the baseline $\overline{\mathcal{I}}_{Z_j}$ as the average of $\mathcal{I}^s_{Z^s_j}$

$$\overline{\mathcal{I}}_{X_j} = \frac{1}{S} \sum_{s=1}^{S} \mathcal{I}^s_{X_j} \qquad \overline{\mathcal{I}}_{Z_j} = \frac{1}{S} \sum_{s=1}^{S} \mathcal{I}^s_{Z^s_j}. \tag{3}$$

(6) Report the adjusted VI measure as $\mathcal{I}^{\text{adj}}_{X_j} = \overline{\mathcal{I}}_{X_j} - \overline{\mathcal{I}}_{Z_j}$ for the variable X_j.

The basic idea of the above algorithm is the following: the permutation breaks the association between the response variable Y and each pseudo-predictor Z^s_j, but still preserves the association between Z^s_j and Z^s_k ($k \neq j$); since Z^s_j is re-shuffled from X_j, Z^s_j has the same number of possible splits as the corresponding predictor X_j and has approximately the same probability of being selected in split nodes. Therefore, $\overline{\mathcal{I}}_{Z_j}$ can be viewed as a bias approximation of the importance of X_j.

Part D: Descriptive statistics for real data

The descriptive statistics of Yip and Yau (2005) data used in Section 6 are provided in Table A1, A2 and A3.

Total Claim Amount	% obs.	% of total sum	Mean	Median
0	61.1	0	0	0
$(0, 10000]$	29.6	36.0	4902	4842
$(10000, 50000]$	9.1	61.5	27144	27679
> 50000	0.2	2.5	52157	51986

Table A1: Description of the individual total claim amount in the last five years.

	AGE	HOMEKIDS	BLUEBOOK	KIDSDRIV
Min.	16.00	0.0000	1500	0.0000
1st Qu.	39.00	0.0000	9200	0.0000
Median	45.00	0.0000	14405	0.0000
Mean	44.84	0.7199	15666	0.1694
3rd Qu.	51.00	1.0000	20900	0.0000
Max.	81.00	5.0000	69740	4.0000

	NPOLICY	RETAINED	TRAVTIME	MVR_PTS
Min.	1.000	1.000	5.00	0.000
1st Qu.	1.000	1.000	22.00	0.000
Median	1.000	4.000	33.00	1.000
Mean	1.695	5.328	33.42	1.709
3rd Qu.	2.000	7.000	44.00	3.000
Max.	9.000	25.000	142.00	13.000

Table A2: Descriptive statistics for the continuous variables in the claim history data set in Section 6.

AREA	MARRIED	REVOKED	GENDER
Rural: 20.2%	No: 39.9%	No: 87.8%	F: 53.8%
Urban: 79.8%	Yes: 60.1%	Yes: 12.2%	M: 46.2%

CAR_USE	MAX_EDUC	CAR_TYPE	JOBCLASS
Private: 63.2%	<High School: 14.6%	Panel Truck: 8.3%	Blue Collar: 22.2%
Commercial: 36.8%	Bachelors: 27.3%	Pickup: 17.3%	Clerical: 15.5%
	High School: 28.7%	Sedan: 26.2%	Professional: 13.6%
	Masters: 20.2%	Sports Car: 11.4%	Manager: 12.2%
	PhD: 9.2%	SUV: 27.9%	Lawyer: 10.0%
		Van: 8.9%	Student: 8.7%
			(Other): 17.8%

Table A3: Descriptive statistics for the categorical variables in the claim history data set in Section 6.

Part E: Identifying important interactions

In this section, we demonstrate that the nonparametric approach described in this paper can serve as an important complement to the traditional GLM model in insurance rating. Even under strict circumstances that the final model must have a GLM structure, our approach can still be quite helpful due to its ability to automatically identify additional information such as important interactions. It is often challenging for a GLM approach alone to capture such information, especially if many explanatory variables are discrete (which is quite common for insurance data sets). For example, if there are eight discrete explanatory variables each with eight different values, there are $\binom{8}{2} \times 7 \times 7 = 1372$ possible two-way interaction terms. Even for data sets with millions of observations, it is in general not practical to fit simultaneously all interaction terms in a GLM model.

We continue using the real data example in Section 6. Suppose one builds a TGLM model with all main effects and applies the stepwise selection for variable selection (the p-values for entering and removal of an variable are set to be 0.05 and 0.10, respectively). The resulting model TGLM1 is showed in Table A4.

We next show that TDboost can provide insights into the structure of interaction terms, which can be subsequently integrated into TGLM1. Elith et al. (2008) proposed a relative importance measure to quantify magnitudes of fitted interaction effects. By adopting this method for TDboost, we can calculate the relative importance of two-way interactions for all possible pairs of predictors in TGLM1. Table A5 provides a summary list of 10 two-way interactions with the highest relative importance. To improve TGLM1, we then add to TGLM1 the two strongest interactions MVR_PTS:AREA and REVOKED:AREA, which account for approximately 88.33% of the total relative importance. We denote the adjusted model with interactions as TGLM2. Table A4 suggests that both interactions in TGLM2 are significant at 0.05 significance level.

To compare TGLM1 and TGLM2, we use the Gini index as the criterion. As shown in Table A6, we find that the maximal Gini index is 9.751 when using TGLM1 as the base premium, and -2.172 when using TGLM2. Therefore, TGLM2 is more favorable than TGLM1. We also compare the TGLM2 model against the TGLM1 model using the likelihood ratio test and get the same conclusion ($\chi^2 = 371.79$, $df = 2$, $p \approx 0$). Therefore, with the help of TDboost, the overall model performance is improved under a GLM model structure.

Variable	TGLM1		TGLM2	
	Estimate	Std.Error	Estimate	Std.Error
Intercept	-2.93**	0.20	-4.61**	0.54
KIDSDRIV	0.10**	0.04	0.10**	0.05
REVOKED	1.54**	0.06	2.47**	0.44
MVR_PTS	0.20**	0.01	0.58**	0.07
MARRIED	-0.17**	0.04	-0.17**	0.05
AREA	1.22**	0.07	2.12**	0.27
CAR_TYPE_2	-0.08	0.10	-0.07	0.11
CAR_TYPE_3	-0.07	0.10	-0.07	0.11
CAR_TYPE_4	0.22*	0.11	0.23*	0.12
CAR_TYPE_5	0.09	0.10	0.10	0.11
CAR_TYPE_6	0.13	0.11	0.13	0.12
JOBCLASS_2	-0.17	0.11	-0.17	0.12
JOBCLASS_3	-0.07	0.12	-0.07	0.13
JOBCLASS_4	-0.47**	0.17	-0.49**	0.19
JOBCLASS_5	-0.07	0.13	-0.06	0.15
JOBCLASS_6	-0.42**	0.13	-0.43**	0.14
JOBCLASS_7	-0.26**	0.12	-0.27**	0.13
JOBCLASS_8	-0.21*	0.11	-0.21*	0.12
JOBCLASS_9	0.05	0.13	0.02	0.14
MVR_PTS:AREA			-0.20**	0.03
REVOKED:AREA			-0.49**	0.23

Note. * $p < 0.10$; ** $p < 0.05$.

Table A4: TGLM1 and TGLM2 model coefficient estimates.

	Variable 1	Variable 2	Importance
1	AREA	MVR_PTS	73.66
2	AREA	REVOKED	14.67
3	AREA	CAR_TYPE	7.34
4	MVR_PTS	REVOKED	6.80
5	CAR_TYPE	REVOKED	2.49

Table A5: Summary of the top 10 most important two-way interactions in the TDboost model in the automobile claims data example.

	Competing Premium	
Base Premium	TGLM1	TGLM2
TGLM1	0	9.751 (0.213)
TGLM2	-2.172 (0.260)	0

Table A6: The averaged Gini indices and standard errors for TGLM1 and TGLM2 in the auto insurance claim data example based on 20 random splits.

References

Dunn, P. K. and Smyth, G. K. (2005), “Series evaluation of Tweedie exponential dispersion model densities,” *Statistics and Computing*, 15, 267–280.

— (2008), “Evaluation of Tweedie exponential dispersion model densities by Fourier inversion,” *Statistics and Computing*, 18, 73–86.

Sandri, M. and Zuccolotto, P. (2008), “A bias correction algorithm for the Gini variable importance measure in classification trees,” *Journal of Computational and Graphical Statistics*, 17.

— (2010), “Analysis and correction of bias in Total Decrease in Node Impurity measures for tree-based algorithms,” *Statistics and Computing*, 20, 393–407.

Yip, K. C. and Yau, K. K. (2005), “On modeling claim frequency data in general insurance with extra zeros,” *Insurance: Mathematics and Economics*, 36, 153–163.

A Supplemental Material for "Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models"

Yi Yang[*], Wei Qian[†] and Hui Zou[‡]

April 20, 2016

Part A: A property of Tweedie distributions

For completeness, we give here a known result of the Tweedie distribution and its detailed proof.

Proposition 1. *Let* $Z_i = \sum_{d_i=1}^{N_i} \tilde{Z}_{d_i}$ *is the total claim amount. Let* $Y_i = Z_i/w_i$*, where* ω_i *is the duration. Assume* N_i *is Poisson distributed* $\mathrm{Pois}(\lambda_i w_i)$*. Conditional on* N_i*, assume* Z_{d_i}*'s* $(d_i = 1, \ldots, N_i)$ *are i.i.d.* $\mathrm{Gamma}(\alpha, \gamma_i)$*. Assume that under unit duration (i.e.,* $w_i = 1$*), the mean-variance relation satisfies* $Var(Y_i^*) = \phi[E(Y_i^*)]^\rho$*, where* Y_i^* *is the pure premium under unit duration,* ϕ *is a constant, and* $\rho = (\alpha+2)/(\alpha+1)$*. Then for the pure premium* Y_i *under duration* ω_i

$$Y_i \sim \mathrm{Tw}(\mu_i, \phi/w_i, \rho).$$

Proof. Note that under unit duration $w_i = 1$,

$$\mu_i^* := E(Y_i^*) = E(E(Y_i^*|N_i)) = \lambda_i \alpha \gamma_i,$$
$$Var(Y_i^*) = E(Var(Y_i^*|N_i)) + Var(E(Y_i^*|N_i)) = \lambda_i \alpha \gamma_i^2 + \lambda_i \alpha^2 \gamma_i^2.$$

[*]McGill University
[†]Rochester Institute of Technology
[‡]Corresponding author, zoux019@umn.edu, University of Minnesota

Similarly, under any duration w_i,

$$\mu_i := E(Y_i) = \frac{1}{w_i} E(Z_i) = \lambda_i \alpha \gamma_i,$$
$$Var(Y_i) = \frac{1}{w_i^2} Var(Z_i) = (\lambda_i \alpha \gamma_i^2 + \lambda_i \alpha^2 \gamma_i^2)/w_i.$$

As a result, we can obtain the mean-variance relation for the pure premium Y_i that

$$Var(Y_i) = \frac{1}{w_i} Var(Y_i^*) = \frac{\phi}{w_i} (\mu_i^*)^\rho = \frac{\phi}{w_i} \mu_i^\rho, \tag{1}$$

where the second equation follows by

$$Var(Y_i^*) = \phi [E(Y_i^*)]^\rho. \tag{2}$$

By the scale-invariance property of Tweedie distribution, the proof is complete. □

Part B: Computational issues for profile likelihood

There are some computational issues, which must be taken care of when evaluating the log-likelihood functions in (20) and (21) of Section 4.2: since in general there are no closed forms for Tweedie densities, in likelihood evaluation one must deal with an infinite summation in the normalizing function $a(y, \phi, \rho) = \frac{1}{y} \sum_{t=1}^{\infty} W_t$. For numerical evaluation of Tweedie densities, Dunn and Smyth (2005) proposed a series expansions approach, which sums an infinite series arising from a Taylor expansion of the characteristic function. Alternatively, Dunn and Smyth (2008) developed a Fourier inversion approach, which consists of an inversion of the characteristic function based on numerical integration methods for oscillating functions. These two numerical methods turn out to be complementary since each has advantages under a certain situation: when only considering the case $1 < \rho < 2$, the series approach performs very well for small y but gradually loses computational efficiency as y increases, whereas the inversion approach performs very well for large y but gradually fails to provide accurate results as y decreases. Hence the inversion approach is preferred for large y and the series approach for small y. Dunn and Smyth (2008) provided a simple guideline to choose between the two methods. In this paper we use their R package "tweedie" (available at `http://cran.r-project.org/web/packages/tweedie/index.html`) for evaluating

Tweedie densities in our profile likelihood computation. For further details regarding their algorithms, the reader may refer to Dunn and Smyth (2005, 2008).

Part C: Bias-adjusted variable importance measure

Following Sandri and Zuccolotto (2008) and Sandri and Zuccolotto (2010), we compute the biased-adjusted VI measure for each explanatory variable in the following way:

(1) For $s = 1, \dots, S$, repeat steps (2)–(4).

(2) Generate a matrix $\mathbf{z}^s$ by randomly permutating (without replacement) the n rows of the design matrix $\mathbf{x}$, while keeping the order of columns unchanged.

(3) Create an $n \times 2p$ matrix $\tilde{\mathbf{x}}^s = [\mathbf{x}, \mathbf{z}^s]$ by binding $\mathbf{z}^s$ with $\mathbf{x}$ matrix by column.

(4) Use the data $\{y, \tilde{\mathbf{x}}^s\}$ to fit the model, and compute VI measures $\mathcal{I}^s_{X_j}$ for X_j and $\mathcal{I}^s_{Z^s_j}$ for Z^s_j, where Z^s_j (jth column of Z^s) is the pseudo-predictor corresponding to X_j.

(5) Compute the VI measure $\overline{\mathcal{I}}_{X_j}$ as the average of $\mathcal{I}^s_{X_j}$ and the baseline $\overline{\mathcal{I}}_{Z_j}$ as the average of $\mathcal{I}^s_{Z^s_j}$

$$\overline{\mathcal{I}}_{X_j} = \frac{1}{S} \sum_{s=1}^{S} \mathcal{I}^s_{X_j} \qquad \overline{\mathcal{I}}_{Z_j} = \frac{1}{S} \sum_{s=1}^{S} \mathcal{I}^s_{Z^s_j}. \tag{3}$$

(6) Report the adjusted VI measure as $\mathcal{I}^{\text{adj}}_{X_j} = \overline{\mathcal{I}}_{X_j} - \overline{\mathcal{I}}_{Z_j}$ for the variable X_j.

The basic idea of the above algorithm is the following: the permutation breaks the association between the response variable Y and each pseudo-predictor Z^s_j, but still preserves the association between Z^s_j and Z^s_k ($k \neq j$); since Z^s_j is re-shuffled from X_j, Z^s_j has the same number of possible splits as the corresponding predictor X_j and has approximately the same probability of being selected in split nodes. Therefore, $\overline{\mathcal{I}}_{Z_j}$ can be viewed as a bias approximation of the importance of X_j.

Part D: Descriptive statistics for real data

The descriptive statistics of Yip and Yau (2005) data used in Section 6 are provided in Table A1, A2 and A3.

Total Claim Amount	% obs.	% of total sum	Mean	Median
0	61.1	0	0	0
$(0, 10000]$	29.6	36.0	4902	4842
$(10000, 50000]$	9.1	61.5	27144	27679
> 50000	0.2	2.5	52157	51986

Table A1: Description of the individual total claim amount in the last five years.

	AGE	HOMEKIDS	BLUEBOOK	KIDSDRIV
Min.	16.00	0.0000	1500	0.0000
1st Qu.	39.00	0.0000	9200	0.0000
Median	45.00	0.0000	14405	0.0000
Mean	44.84	0.7199	15666	0.1694
3rd Qu.	51.00	1.0000	20900	0.0000
Max.	81.00	5.0000	69740	4.0000

	NPOLICY	RETAINED	TRAVTIME	MVR_PTS
Min.	1.000	1.000	5.00	0.000
1st Qu.	1.000	1.000	22.00	0.000
Median	1.000	4.000	33.00	1.000
Mean	1.695	5.328	33.42	1.709
3rd Qu.	2.000	7.000	44.00	3.000
Max.	9.000	25.000	142.00	13.000

Table A2: Descriptive statistics for the continuous variables in the claim history data set in Section 6.

AREA	MARRIED	REVOKED	GENDER
Rural: 20.2%	No: 39.9%	No: 87.8%	F: 53.8%
Urban: 79.8%	Yes: 60.1%	Yes: 12.2%	M: 46.2%

CAR_USE	MAX_EDUC	CAR_TYPE	JOBCLASS
Private: 63.2%	<High School: 14.6%	Panel Truck: 8.3%	Blue Collar: 22.2%
Commercial: 36.8%	Bachelors: 27.3%	Pickup: 17.3%	Clerical: 15.5%
	High School: 28.7%	Sedan: 26.2%	Professional: 13.6%
	Masters: 20.2%	Sports Car: 11.4%	Manager: 12.2%
	PhD: 9.2%	SUV: 27.9%	Lawyer: 10.0%
		Van: 8.9%	Student: 8.7%
			(Other): 17.8%

Table A3: Descriptive statistics for the categorical variables in the claim history data set in Section 6.

Part E: Identifying important interactions

In this section, we demonstrate that the nonparametric approach described in this paper can serve as an important complement to the traditional GLM model in insurance rating. Even under strict circumstances that the final model must have a GLM structure, our approach can still be quite helpful due to its ability to automatically identify additional information such as important interactions. It is often challenging for a GLM approach alone to capture such information, especially if many explanatory variables are discrete (which is quite common for insurance data sets). For example, if there are eight discrete explanatory variables each with eight different values, there are $\binom{8}{2} \times 7 \times 7 = 1372$ possible two-way interaction terms. Even for data sets with millions of observations, it is in general not practical to fit simultaneously all interaction terms in a GLM model.

We continue using the real data example in Section 6. Suppose one builds a TGLM model with all main effects and applies the stepwise selection for variable selection (the p-values for entering and removal of an variable are set to be 0.05 and 0.10, respectively). The resulting model TGLM1 is showed in Table A4.

We next show that TDboost can provide insights into the structure of interaction terms, which can be subsequently integrated into TGLM1. Elith et al. (2008) proposed a relative importance measure to quantify magnitudes of fitted interaction effects. By adopting this method for TDboost, we can calculate the relative importance of two-way interactions for all possible pairs of predictors in TGLM1. Table A5 provides a summary list of 10 two-way interactions with the highest relative importance. To improve TGLM1, we then add to TGLM1 the two strongest interactions MVR_PTS:AREA and REVOKED:AREA, which account for approximately 88.33% of the total relative importance. We denote the adjusted model with interactions as TGLM2. Table A4 suggests that both interactions in TGLM2 are significant at 0.05 significance level.

To compare TGLM1 and TGLM2, we use the Gini index as the criterion. As shown in Table A6, we find that the maximal Gini index is 9.751 when using TGLM1 as the base premium, and -2.172 when using TGLM2. Therefore, TGLM2 is more favorable than TGLM1. We also compare the TGLM2 model against the TGLM1 model using the likelihood ratio test and get the same conclusion ($\chi^2 = 371.79$, $df = 2$, $p \approx 0$). Therefore, with the help of TDboost, the overall model performance is improved under a GLM model structure.

Variable	TGLM1		TGLM2	
	Estimate	Std.Error	Estimate	Std.Error
Intercept	-2.93**	0.20	-4.61**	0.54
KIDSDRIV	0.10**	0.04	0.10**	0.05
REVOKED	1.54**	0.06	2.47**	0.44
MVR_PTS	0.20**	0.01	0.58**	0.07
MARRIED	-0.17**	0.04	-0.17**	0.05
AREA	1.22**	0.07	2.12**	0.27
CAR_TYPE_2	-0.08	0.10	-0.07	0.11
CAR_TYPE_3	-0.07	0.10	-0.07	0.11
CAR_TYPE_4	0.22*	0.11	0.23*	0.12
CAR_TYPE_5	0.09	0.10	0.10	0.11
CAR_TYPE_6	0.13	0.11	0.13	0.12
JOBCLASS_2	-0.17	0.11	-0.17	0.12
JOBCLASS_3	-0.07	0.12	-0.07	0.13
JOBCLASS_4	-0.47**	0.17	-0.49**	0.19
JOBCLASS_5	-0.07	0.13	-0.06	0.15
JOBCLASS_6	-0.42**	0.13	-0.43**	0.14
JOBCLASS_7	-0.26**	0.12	-0.27**	0.13
JOBCLASS_8	-0.21*	0.11	-0.21*	0.12
JOBCLASS_9	0.05	0.13	0.02	0.14
MVR_PTS:AREA			-0.20**	0.03
REVOKED:AREA			-0.49**	0.23

Note. * $p < 0.10$; ** $p < 0.05$.

Table A4: TGLM1 and TGLM2 model coefficient estimates.

	Variable 1	Variable 2	Importance
1	AREA	MVR_PTS	73.66
2	AREA	REVOKED	14.67
3	AREA	CAR_TYPE	7.34
4	MVR_PTS	REVOKED	6.80
5	CAR_TYPE	REVOKED	2.49

Table A5: Summary of the top 10 most important two-way interactions in the TDboost model in the automobile claims data example.

	Competing Premium	
Base Premium	TGLM1	TGLM2
TGLM1	0	9.751 (0.213)
TGLM2	-2.172 (0.260)	0

Table A6: The averaged Gini indices and standard errors for TGLM1 and TGLM2 in the auto insurance claim data example based on 20 random splits.

References

Dunn, P. K. and Smyth, G. K. (2005), "Series evaluation of Tweedie exponential dispersion model densities," *Statistics and Computing*, 15, 267–280.

— (2008), "Evaluation of Tweedie exponential dispersion model densities by Fourier inversion," *Statistics and Computing*, 18, 73–86.

Sandri, M. and Zuccolotto, P. (2008), "A bias correction algorithm for the Gini variable importance measure in classification trees," *Journal of Computational and Graphical Statistics*, 17.

— (2010), "Analysis and correction of bias in Total Decrease in Node Impurity measures for tree-based algorithms," *Statistics and Computing*, 20, 393–407.

Yip, K. C. and Yau, K. K. (2005), "On modeling claim frequency data in general insurance with extra zeros," *Insurance: Mathematics and Economics*, 36, 153–163.