ADDRESSING IMBALANCED INSURANCE DATA THROUGH ZERO-INFLATED POISSON REGRESSION WITH BOOSTING

BY

SIMON C.K. LEE

ABSTRACT

A machine learning approach to zero-inflated Poisson (ZIP) regression is introduced to address common difficulty arising from imbalanced financial data. The suggested ZIP can be interpreted as an adaptive weight adjustment procedure that removes the need for post-modeling re-calibration and results in a substantial enhancement of predictive accuracy. Notwithstanding the increased complexity due to the expanded parameter set, we utilize a cyclic coordinate descent optimization to implement the ZIP regression, with adjustments made to address saddle points. We also study how various approaches alleviate the potential drawbacks of incomplete exposures in insurance applications. The procedure is tested on real-life data. We demonstrate a significant improvement in performance relative to other popular alternatives, which justifies our modeling techniques.

KEYWORDS

Boosting trees, predictive modeling, insurance, machine learning, imbalanced data, zero-inflated Poisson.

JEL codes: C14

1. IMBALANCED DATA IN MACHINE LEARNING APPLICATIONS

Imbalanced data, in which one class dominates all others, are typical in insurance. Research on ratemaking, underwriting, reinsurance, and fraud generally involves data with occurrences of less than 10%. Usually, fraud rates range from 3 to 7% across all industries (Gee and Button, 2019), and from 2 to 8% among short-term retail insurance claims, according to Lee and Lin (2018) and

Astin Bulletin 51(1), 27-55. doi:10.1017/asb.2020.40 © 2020 by Astin Bulletin. All rights reserved.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

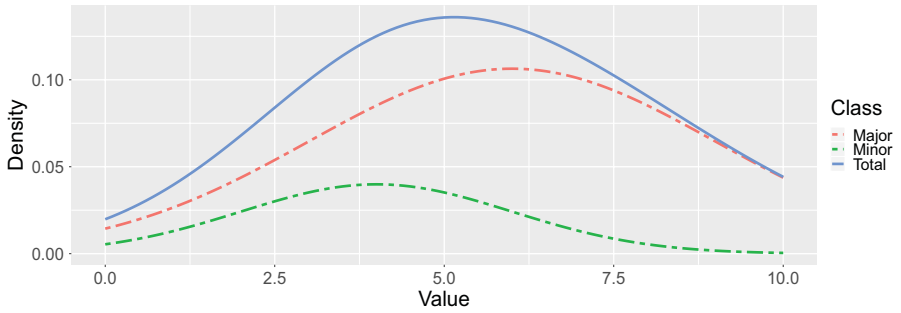


FIGURE 1: Class imbalance illustration, with dashed lines representing the unconditional density of imbalanced classes.

the Insurance Information Institute.[1] Despite this low frequency, the minor class has the biggest financial impact; for example, fraud cases around the globe were estimated to cause damages of approximately 5 trillion USD in 2018 (Gee and Button, 2019).

Counterintuitively, the challenge of exactly modeling imbalanced data is largely due to the primary goal of most statistical and machine learning models, which is to optimize overall predictive accuracy (He and Ma, 2013; Haixiang *et al.*, 2017; Fernández *et al.*, 2018). Figure 1 shows a simple scenario in which the major class (red dashed line) accounts for 90% of the observations. The total density (blue solid line) is then calculated by incorporating the minor class (green dashed line).

Although the distribution of the classes is notably varied, the observations at each value are dominated by the major class. Accuracy-driven models would classify all observations as the major class. However, these models ignore information from the minority class and cannot be trusted for practical purposes.

The class imbalance in retail insurance is more severe. On average, less than 5% of all policies result in claims. Moreover, insurance data generally involve over-dispersed distributions, which indicates higher variances than that in the distributions used for modeling (Boucher *et al.*, 2007; Klein *et al.*, 2015; Lee and Lin, 2018). The accurate evaluation of underlying insurance risks presents a challenge to actuaries and underwriters. Even in countries with mature markets, such as the UK,[2] US,[3] and Canada,[4,5] the retail property and casualty insurance industry suffer from underwriting losses. When the quality of the pricing models is impacted by imbalanced data, some client segments are over-priced, while others are under-priced. For a fixed customer base, these errors can be offset if the overall price level is adequate. However, in competitive insurance markets, customer demand is elastic to price change. Customers receiving a price increase are more likely to terminate their insurance policies, whereas competitive prices attract more clients. Thus, pricing errors contribute to the phenomena of adverse selection and underwriting losses.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

The imbalanced data phenomenon has been actively investigated, and some benchmark studies are listed here for the reader's reference. Fernández *et al.* (2018), He and Ma (2013) offer a solid review of influential research and an end-to-end walkthrough of problems, algorithms, and applications related to this topic. Haixiang *et al.* (2017) demonstrate how the latest research addresses practical problems in the fields of engineering, finance, and information technology, among others.

Three major research areas address imbalanced data. Cost-sensitive learning leverages domain expertise to modify learning objectives and steer predictive models to search for more suitable parameters. For instance, in financial and medical applications, false positives are penalized far less than false negatives. As modeling results are highly susceptible to the cost matrix, users can design an appropriate matrix to reflect their level of domain understanding and risk preference.

The reliability of any single model may erode and give rise to an overfitting phenomenon. Ensembling aggregates multiple models to mitigate risk and hence improves the generalization power. This method, however, suffers from elevated training costs and poor applicability to practical tasks and remains prone to noisy data sets.

Modification of data distribution tempers the underlying data distribution to achieve higher precision. The most trivial method involves either the creation of additional observations for the minor class or the removal of observations from the major class through resampling. Rather than duplicating the existing minor class, SMOTE synthesizes extra observations from the minor class by interpolating explanatory variates with their neighbors (Chawla *et al.*, 2002). While there have been many successes in this area, the selection of resampling and distance thresholds remains an exercise in trial and error. Thus, this methodology injects an additional judgmental variance that impedes the quality of the models.

In the actuarial profession, established standards of practices (e.g. the Actuarial Standard of Practices[6] from the Actuarial Standard Board, the Statements of Principles[7] from the Casualty Actuarial Society) serve actuaries throughout the insurance modeling life cycle. The standards cover aspects including data collection, processing, modeling, and calibration to address analytical problems arising from commonly imbalanced insurance data. Actuarial credibility theory also quantifies the adequacy of interested minorities and derives credibility factors to calibrate the forecasts from predictive models (Bühlmann and Gisler, 2006; Wuthrich and Buser, 2019).

Motivated by the above introduction, we propose a boosting (Schapire, 1990) implementation of the ZIP regression to model insurance data. The contributions of this paper are three-fold. First, according to our knowledge of actuarial research, we are the first to implement a non-parametric two-parameter estimation of zero-inflated regression. We derive the full boosting algorithm in detail. Previous researches, including Boucher *et al.* (2009, 2007), Klein *et al.* (2015), Calderín-Ojeda *et al.* (2019), demonstrate

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

the strength of zero-inflated regression with parametric (GLM) and semi-parametric (Bayesian GAMLSS) estimation. We introduce additional procedures to address the potential violation of convexity, as the two-parameter zero-inflated family does not belong to the exponential family. Second, we establish the ZIP regression in boosting as an adaptive re-weighting mechanism to downweigh the major class. The algorithm is one-shot and loss-optimizing, as it does not require additional pre-processing, sampling, or post-modeling calibration. Third, we introduce several strategies for incorporating an exposure offset into the regression. We identify a criterion for applying exposure offsets that avoids a vanishing probability of a perfect state, and we show that an appropriate exposure handling strategy is crucial in achieving good prediction quality.

The rest of this paper is organized as follows. We first define the ZIP and delta boosting in Sections 2 and 3, respectively. In Section 4, we derive the algorithm for the ZIP regression in detail. In Section 5, we address the existence of incomplete exposures of insurance data. In Section 6, we compare the results of delta boosting implementation of the ZIP to other distributions and predictive modeling techniques on a real-life data through practical metrics. Finally, we conclude and propose future research directions in Section 7.

2. Zero-Inflated Poisson distribution

Despite criticism of its weakness in modeling thick-tailed distributions, the Poisson distribution remains the main claim frequency regression model due to its simplicity and time homogeneity (Teugels and Vynckie, 1996). Several studies have demonstrated considerable improvements in precision by expanding the Poisson variance to the family of compound or mixed Poisson distributions, including the negative binomial, quasi-Poisson, double Poisson, and generalized Poisson distributions, and their zero-inflated extensions (Boucher *et al.*, 2007, 2009; Klein *et al.*, 2015; Lee, 2020).

We propose the ZIP (Lambert, 1992) regression as an extension to address the imbalanced data problem while maintaining the simplicity of parameter interpretation. $Y_i \sim ZIP(p_i, \lambda_i)$ is a mixture of a perfect state and a Poisson distribution:

$$Y_i \overset{\text{def}}{=} (1 - \mathbb{1}_i) Y_i^*, \tag{2.1}$$

$$P(Y_i = k) = \begin{cases} p_i + (1 - p_i) \exp^{-\lambda_i}, & \text{if } k = 0, \\ (1 - p_i) \frac{\exp^{-\lambda_i} \lambda_i^k}{k!}, & \text{if } k = 1, 2, 3, \ldots \end{cases} \tag{2.2}$$

where $\mathbb{1}_i \sim Bernoulli(p_i)$ is the indicator random variable of the perfect state, and $Y_i^* \sim Poisson(\lambda_i)$ represents the counting distribution of the minor class for observation i, with $Y_i^* \perp\!\!\!\perp \mathbb{1}_i$. λ_i and p_i are defined to be positive and between

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

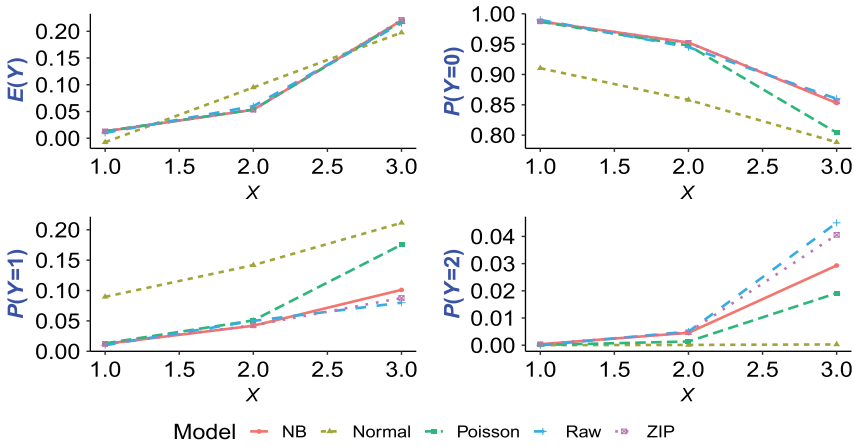


FIGURE 2: Illustration of model divergence over synthetic data. The top left shows the mean of each distribution, while the remainder portrays the probabilities of $Y = 0$, 1, and 2, respectively.

0 and 1. Representation (2.2) explicitly accounts for the excessive zero phenomenon that is common in non-life insurance. In auto, home, and other non-life insurance products, typically most policies are claim-free.

To illustrate why the selection of the proper distribution can be meaningful when modeling imbalanced data, we create synthetic data with 80%, taking a value of 0 and 20% from a binomial distribution with expectations that are linear to a variate among 1, 2, and 3. The normal, Poisson, negative binomial, and ZIP regressions are chosen for comparison.

From Figure 2, it is clear that the ability to track the underlying distribution improves significantly as we shift from heavier tail distributions. In particular, the discrepancy is already notable for Poisson and Gaussian regression at $P(Y = 0)$. As we measure the difference in probability of a larger Y, the ZIP exhibits good consistency with the imbalanced sample data, even though the underlying distribution is not ZIP. A more complex real-life data regression with a large set of covariates and a more imbalanced distribution is likely to exacerbate the difference in performance.

3. TWO-PARAMETER DELTA BOOSTING IMPLEMENTATION

First introduced in Schapire (1990), boosting is a popular machine learning algorithm based on ensembles. Empirical studies suggest that boosting is a highly competitive technique among machine learning candidates in terms of predictive accuracy (Chen *et al.*, 2015; Ke *et al.*, 2017). Two fundamental elements are common in boosting:

Strength of weak learnability: Weak learners are defined as predictive rules that are slightly better than coin flips. By combining many predictions of weak learners, a highly predictive model is built.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

Adaptive modeling: At every iteration, the base learner derived is based on the residuals from previous iterations. This is also known as forward stagewise modeling.

By combining the predictions of many base learners, an ensemble is created. In this paper, we adopt CART (Breiman *et al*., 1984) as the base learners, as these offer a straightforward representation of high-dimensional dependence.

The formula of the model output is as follows:

$$\hat{F}_T(\mathbf{x}_i) = \sum_{t=1}^{T} \beta_t \cdot h(\mathbf{x}_i; \mathbf{a}_t) = \sum_{t=1}^{T} \sum_{j=1}^{J_t} \beta_{t,j} h(\mathbf{x}_i; \mathbf{a}_{t,j}), \tag{3.1}$$

where $\mathbf{x}$ represents the co-variate matrix, $\mathbf{x}_i$ represents the co-variate vector for observation i, and $\mathbf{a}_t = \{\mathbf{a}_{t,1}, \mathbf{a}_{t,2}, \ldots, \mathbf{a}_{t,J_t}\}$ are the collection of J_t disjoint partitions induced by the base learner created at iteration t. In the case where trees are used as the base learners, the partitions are generated by looping through the co-variate set to identify split points that minimize square errors. $h(\mathbf{x}_i, \mathbf{a}_t)$ is an indicator function of whether $\mathbf{x}_i$ is in the partition $\mathbf{a}_t$, and β_t are coefficients of the adjustment from the tree-based estimation.

In a ZIP regression with data set $\mathcal{D} = \{(y_i, \mathbf{x}_i) : i = 1, \ldots, M\}$, the response $y_i \sim ZIP(p_i, \lambda_i)$. The estimation of p_i and λ_i varies by the algorithm; for example, GLM assumes $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \alpha_0 + \sum_{j=1}^{M} \alpha_j x_{i,j}$, whereas GAM assumes $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \sum_{j=1}^{M} f_j(x_{i,j})$. α(coefficients of input variables) and f(smooth functions that approximate the dependence between individual variables and the response) are estimated to optimize the corresponding objectives. Interested readers can find these estimation algorithms in Boucher *et al*. (2007, 2009) and Klein *et al*. (2015).

Due to the forward stagewise characteristics of boosting, the estimation of multiple parameters is considerably simplified. In the rest of this section, we explain the functional forms of p_i and λ_i in delta boosting and the tool needed to derive the estimations.

The boosting family includes many variations. Adaboost (Freund and Schapire, 1995) is considered the first major success in this family; more weights are assigned to wrongly classified observations in each iteration, such that the base learner can more effectively correct mistakes. Gradient boosting adopts gradient descent to build decision trees (Friedman, 2002). Chen *et al*. (2015) introduce xgboost which allows to incorporate regularization factors. Furthermore, instead of using a pure gradient descent approach, the authors employ a second-degree approximation to derive the estimates. Ke *et al*. (2017) present lightgbm as a highly efficient tree estimation approach to speed up the creation of the ensemble.

Delta boosting, which was introduced in Lee and Lin (2018); Wuthrich and Buser (2019), consists of three main steps: basis, regression, and adjust.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

The authors illustrate that the boosting mechanism is the optimal boosting candidate for many prevalent regressions, including Poisson, Gamma, Tweedie, normal, and Huber. Compared with other members of the boosting family, delta boosting has two primary distinguishing features. First, delta boosting uses loss minimizers, instead of gradients, as the basis for regression. Second, the regression and adjustment steps are integrated to obtain the optimal partitions and base learners' parameters simultaneously.

Lee (2020) extends delta boosting to simultaneously estimate the shape and scale parameters in the negative binomial regression model, with treatments to address concavity. Other common libraries, including lightgbm (Ke *et al.*, 2017) and xgboost (Chen *et al.*, 2015), also offer multinomial regressions that estimate more than one parameter. Judging from the code architecture, both libraries update all parameters while assuming that the remainders are constant and, thus, ignore mixed partial derivatives. However, Lee (2020) shows that the magnitudes of the mixed partial derivatives can be large enough to render this simplified estimation unreliable for the negative binomial regression. In contrast, the cyclic coordinate descent approach updates one parameter at a time according to the most recent updated estimates. This approach implicitly considers dependence among variables in a simple and stable manner, with finite-time convergence (Saha and Tewari, 2010). Motivated by established research that two-parameter zero-inflated regression has a strong potential to enhance predictive performance, we propose to implement ZIP regression through delta boosting with a cyclic coordinate descent.

4. Algorithms under full exposure

The mathematical definitions and proofs involve specific notation. To facilitate this discussion, we summarize the key notation used in this paper in Table 1.

In regression, common practice is to impose a log-link transformation to λ_i and a logit transformation to p_i so that the transformed parameters can take any real value. We denote the transformed parameters by F_i and G_i, respectively, so that

$$\lambda_i = \exp^{F_i}, \quad \hat{\lambda}_{i,t} = \exp^{\hat{F}_{i,t}},$$

$$p_i = \frac{\exp^{G_i}}{1+\exp^{G_i}}, \quad \hat{p}_{i,t} = \frac{\exp^{\hat{G}_{i,t}}}{1+\exp^{\hat{G}_{i,t}}}.$$

In this paper, capped parameters always indicate estimated values. A subscript t represents the corresponding estimates at iteration t in the boosting implementation of ZIP. We drop subscripts when the meaning is clear from the context.

We set the loss function of interest, $\Phi(\mathbf{y}, F(\mathbf{x}), G(\mathbf{x}))$, as the negative log-likelihood of ZIP.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

TABLE 1

KEY NOTATION AND DEFINITIONS.

Notation	Description
$\mathcal{D}, \mathbf{y}, \mathbf{x}$	$\mathcal{D}$ denotes the dataset, $\mathcal{D} \stackrel{\text{def}}{=} \{(y_i, \mathbf{x}_i) : i = 1, \ldots, M\}$, where $\mathbf{y}$, $\mathbf{x}$ are the response vector and co-variate matrix, respectively. y_i and $\mathbf{x}_i$ represent the response and co-variate at observation i, respectively. M indicates the total number of observations
$h(\mathbf{x}_i; \mathbf{a}_t)$	An indicator function to show whether $\mathbf{x}_i$ is in partition $\mathbf{a}_t$
$\mathbf{N}_{t,j}$	An index set of partition j at t-th iteration. $\mathbf{x}_i \in \mathbf{a}_{t,j} \Leftrightarrow i \in \mathbf{N}_{t,j} \Leftrightarrow h(\mathbf{x}_i; \mathbf{a}_{t,j}) = 1$
J_t	Number of partitions at iteration t. In the case of tree stumps, $J_t = 2$ and we use N_L and N_R to represent the index set, where N_L (N_R) represents the data indices left (right) of the split point
$\hat{p}_t(\mathbf{x}_i)$ or $\hat{p}_{i,t}$	The estimated probability of a perfect state in ZIP for observation i, $p(\mathbf{x}_i)$, at iteration t, as a function of underlying features $\mathbf{x}_i$
$\hat{q}_{i,t}$	$\hat{q}_{i,t} \stackrel{\text{def}}{=} 1 - \hat{p}_{i,t}$
$\hat{G}_t(\mathbf{x}_i)$ or $\hat{G}_{i,t}$	$\exp^{\hat{G}_t} \stackrel{\text{def}}{=} \frac{\hat{p}_t}{1-\hat{p}_t}$, common logit transformation of $\hat{p}_t$
$\hat{\lambda}_t(\mathbf{x}_i)$ or $\hat{\lambda}_{i,t}$	Estimated Poisson mean in ZIP for observation i, λ_i, at iteration t
$\hat{F}_t(\mathbf{x}_i)$ or $\hat{F}_{i,t}$	Define $\exp^{\hat{F}_t} \stackrel{\text{def}}{=} \hat{\lambda}_t$, common log-transformation of positive $\hat{\lambda}_t$
$\Phi(y_i, \hat{F}_t(\mathbf{x}_i), \hat{G}_t(\mathbf{x}_i))$	The loss function of observation i for DB regression with two underlying parameters
$\Phi(\mathbf{y}, \hat{F}_t(\mathbf{x}), \hat{G}_t(\mathbf{x}))$	$\sum_{i=1}^{M} \Phi(y_i, \hat{F}_t(\mathbf{x}_i), \hat{G}_t(\mathbf{x}_i))$
$\Phi'_{F,i,t}$ and $\Phi'_{G,i,t}$	$\Phi'_{F,i,t}$ as an abbreviation of $\frac{\partial}{\partial \hat{F}_t(\mathbf{x}_i)} \Phi(y_i, \hat{F}_t(\mathbf{x}_i), \hat{G}_t(\mathbf{x}_i))$. Similarly for $\Phi'_{G,i,t}$
$\Phi''_{FF,i,t}$, $\Phi''_{FG,i,t}$, $\Phi''_{GG,i,t}$	$\Phi''_{GG,i,t}$ as an abbreviation of $\frac{\partial^2}{\partial \hat{G}_t^2(\mathbf{x}_i)} \Phi(y_i, \hat{F}_t(\mathbf{x}_i), \hat{G}_t(\mathbf{x}_i))$. Similarly for $\Phi''_{FF,i,t}$ and $\Phi''_{FG,i,t}$.

$$\begin{aligned}\Phi(\mathbf{y}, F(\mathbf{x}), G(\mathbf{x})) &= -\sum_{i=1}^{M} \log\left(P(Y_i = y_i|\mathbf{x}_i)\right) \\ &= -\sum_{i=1}^{M} \log\left(\frac{\exp^{G_i} I_{y_i=0} + \lambda_i^{y_i} \exp^{-\lambda_i}}{\exp^{G_i} + 1}\right) + C,\end{aligned}$$

where $I_{y_i=0}$ is an indicator function for $\{y_i = 0\}$, and C is a constant that can be ignored. At iteration t, our goal is to identify $\Delta_{\hat{G}_{j,t}}$ and $\Delta_{\hat{F}_{j,t}}$ for each node $N_{j,t}$ such that the following quantity is minimized:

$$\begin{aligned}&\Phi(\mathbf{y}, \hat{F}_{t-1} + \Delta_{\hat{F}_t}, \hat{G}_{t-1} + \Delta_{\hat{G}_t}) \\ &= -\sum_{j=1}^{J} \sum_{i \in N_{j,t}} \log\left(\frac{\exp^{\hat{G}_{i,t-1} + \Delta_{\hat{G}_{j,t}}} I_{y_i=0} + \hat{\lambda}_i^{y_i} \exp^{-\exp^{\hat{F}_{i,t-1} + \Delta_{\hat{F}_{j,t}}}}}{1 + \exp^{\hat{G}_{i,t-1} + \Delta_{\hat{G}_{j,t}}}}\right).\end{aligned}$$

By differentiating the above function with respect to $\Delta_{\hat{F}_{j,t}}$ and $\Delta_{\hat{G}_{j,t}}$, respectively, and setting the differentiated functions to 0 for each node $N_{j,t}$, we obtain

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

score equations

$$\begin{cases} \sum\limits_{i\in N_{j,t}} \Phi'_{F,i,t-1}(y_i, \hat{F}_{i,t-1} + \Delta_{\hat{F}_{j,t}}, \hat{G}_{i,t-1} + \Delta_{\hat{G}_{j,t}}) = 0, \\ \sum\limits_{i\in N_{j,t}} \Phi'_{G,i,t-1}(y_i, \hat{F}_{i,t-1} + \Delta_{\hat{F}_{j,t}}, \hat{G}_{i,t-1} + \Delta_{\hat{G}_{j,t}}) = 0. \end{cases}$$

that is,

$$\begin{cases} \sum\limits_{i\in N_{j,t}} \left(\dfrac{\exp^{\hat{F}_{i,t-1}+\Delta_{\hat{F}_{j,t}}} \exp^{-\exp^{\hat{F}_{i,t-1}+\Delta_{\hat{F}_{j,t}}}}}{\exp^{\hat{G}_{i,t-1}+\Delta_{\hat{G}_{j,t}}} I_{y_i=0} + \exp^{-\exp^{\hat{F}_{i,t-1}+\Delta_{\hat{F}_{j,t}}}}} - y_i \right) = 0, & (4.1) \\ \sum\limits_{i\in N_{j,t}} \left(\dfrac{\exp^{\hat{G}_{i,t-1}+\Delta_{\hat{G}_{j,t}}}}{\exp^{\hat{G}_{i,t-1}+\Delta_{\hat{G}_{j,t}}}+1} - \dfrac{\exp^{\hat{G}_{i,t-1}+\Delta_{\hat{G}_{j,t}}} I_{y_i=0}}{\exp^{\hat{G}_{i,t-1}+\Delta_{\hat{G}_{j,t}}} + \exp^{-\exp^{\hat{F}_{i,t-1}+\Delta_{\hat{F}_{j,t}}}}} \right) = 0. & (4.2) \end{cases}$$

4.1. ZIP as an adaptive re-weighting mechanism for imbalanced data

In boosting algorithms, we update the running estimation by adding an adjustment factor. For $i \in N_{j,t}$, the adjustment factors are, respectively, $\Delta_{\hat{F}_{j,t}}$ and $\Delta_{\hat{G}_{j,t}}$:

$$\hat{F}_{i,t} = \hat{F}_{i,t-1} + \Delta_{\hat{F}_{j,t}}, \tag{4.3}$$

$$\hat{G}_{i,t} = \hat{G}_{i,t-1} + \Delta_{\hat{G}_{j,t}}. \tag{4.4}$$

By setting

$$\begin{aligned} \tilde{p}_{i,t} &\overset{\text{def}}{=} \frac{\hat{p}_{i,t}}{\hat{p}_{i,t} + \hat{q}_{i,t} P(Y_i^* = 0)} I_{y_i=0} \\ &= \frac{\exp^{\hat{G}_{i,t}}}{\exp^{\hat{G}_{i,t}} + \exp^{-\hat{\lambda}_{i,t}}} I_{y_i=0}, \end{aligned} \tag{4.5}$$

we have

$$\tilde{q}_{i,t} = 1 - \tilde{p}_{i,t} = \begin{cases} \frac{\exp^{-\hat{\lambda}_{i,t}}}{\exp^{\hat{G}_{i,t}} + \exp^{-\hat{\lambda}_{i,t}}} & \text{if } y_i = 0, \\ 1 & \text{otherwise.} \end{cases} \tag{4.6}$$

By applying Equations (4.3), (4.4), and (4.6) to Equations (4.1) and (4.2), we arrive at two important relations of this paper,

$$\sum_{i\in N_{j,t}} \tilde{q}_{i,t} = \sum_{i\in N_{j,t}} \hat{q}_{i,t}, \tag{4.7}$$

$$\sum_{i\in N_{j,t}} \tilde{q}_{i,t} \hat{\lambda}_{i,t} = \sum_{i\in N_{j,t}} y_i. \tag{4.8}$$

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

We first study Equation (4.8). From Lee and Lin (2018), the loss optimizing $\hat{\lambda}_i$ for a Poisson regression model is $\sum_{i\in N_{j,t}} \hat{\lambda}_{i,t} = \sum_{i\in N_{j,t}} y_i$. Intuitively, the expectation is set to equal the sample mean. In ZIP, the expectation $E(Y_i)$ of $Y_i \sim ZIP(p_i, \lambda_i)$ is $q_i\lambda_i$. Equation (4.8) closely resembles this relationship, with updated weights $\tilde{q}_{i,t}$ as explained below.

From Equation (4.6), we observe that $\tilde{q}_{i,t} = 1 > \hat{q}_{i,t}$ for $y_i > 0$ and $\tilde{q}_{i,t} < \hat{q}_{i,t}$ otherwise. Our proposal thus suggests a log-likelihood that maximizes the re-weighting mechanism, which maintains the overall weights (Equation (4.7)), while the weights of the interested minority ($\{y_i \neq 0\}$) are increased to 1. Additionally, the weights of the major class are reduced, according to the credibility from Equation (4.6). Comparing $\tilde{q}_{i,t}$ and $\hat{q}_{i,t}$ for $y_i = 0$, the magnitude of $\hat{\lambda}_{i,t}$ dictates the weight downsizing. Thus, the optimal re-weighting approach marked by our proposed algorithm modifies the weights most rigorously when the observations are close to the region with a high expected occurrence of the interested minority. The observations in the minor class receive a theoretically maximized up-weighting, and the increase in weight is compensated more heavily for those observations with negative responses.

As mentioned in Section 1, SMOTE synthesizes artificial observations near the interested minority with the users' artificially defined distance functions. Other data pre-processing strategies down-sample the major class and/or up-sample the minor class with a judgmentally selected probability. These approaches temper the natural distribution of data and, thus, oblige parameter calibration after modeling is completed. In contrast, the re-weighting approach suggested by our proposal is adaptive and optimally maximizes the log-likelihood. Hence, no further re-calibration is necessary.

4.2. Extension to generic zero-inflated family

Equations (4.7) and (4.8) can be generalized for generic zero-inflated counting distributions. Assuming

$$P(Y_i = y_i) = \begin{cases} p_i + q_i P(Y_i^* = 0) & \text{if } y_i = 0, \\ q_i P(Y_i^* = y_i) & \text{otherwise,} \end{cases}$$

where Y_i^* is any counting distribution, then the same definition of $\tilde{p}_i$ as in Equation (4.5) yields

$$\sum_{i\in N_{j,t}} \hat{p}_{i,t} = \sum_{i\in N_{j,t}} \tilde{p}_{i,t}, \tag{4.9}$$

$$\sum_{i\in N_{j,t}} \tilde{q}_{i,t} \nabla_{\hat{F}_{i,t}} \log\left(\hat{P}_t(Y_i^* = y_i)\right) = 0, \tag{4.10}$$

where $\hat{P}_t(Y_i^* = y_i)$ is the estimate of $P(Y_i^* = y_i)$ at iteration t and is represented by a vector of parameters $\hat{F}_t(\mathbf{x}_i)$.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

TABLE 2

ADAPTIVE RE-WEIGHTING FORMULAE FOR OTHER COMMON ZERO-INFLATED DISTRIBUTIONS.

Distribution	$P(Y_i^* = y_i)$	$\hat{F}_t(\mathbf{x}_i)$	Equivalence of Equation (4.8)
Negative Binomial	$\binom{r+y_i-1}{y_i}\frac{\theta^r}{(1+\theta_i)^{r+y_i}}$	$\log(\hat{\theta}_i)$	$\sum_{i\in N_{j,t}} \tilde{q}_{i,t}\frac{r}{1+\hat{\theta}_{i,t}} = \sum_{i\in N_{j,t}} \frac{\hat{\theta}_{i,t}}{1+\hat{\theta}_{i,t}} y_i$
Generalized Poisson	$\frac{\theta_i(\theta_i+\eta y_i)^{y_i-1}}{\exp^{\theta_i+\eta y_i} y_i!}$	$\log(\hat{\theta}_{i,t})$	$\sum_{i\in N_{j,t}} \tilde{q}_{i,t}\hat{\theta}_{i,t} = \sum_{i\in N_{j,t}} \frac{\hat{\theta}_{i,t}+\eta}{\hat{\theta}_{i,t}+\eta y_i} y_i$
Double Poisson	$\left(\frac{\theta^{1/2} y_i^{y_i}}{\exp^{\theta\lambda_i+y_i} y_i!}\right)\left(\frac{\exp\lambda_i}{y_i}\right)^{\theta y_i}$	$\log(\hat{\lambda}_{i,t})$	$\sum_{i\in N_{j,t}} \tilde{q}_{i,t}\hat{\lambda}_{i,t} = \sum_{i\in N_{j,t}} y_i$

Negative binomial, generalized Poisson, and double Poisson are commonly adopted counting distributions for zero-inflated regression. Insights similar to Equation (4.8) can be deduced in these cases and are listed in Table 2.

Thus, we observe that zero-inflated regression serves as the weighted regression for the minority, as the algorithm becomes the standard regression when $\tilde{q}_{i,t}$ is set to 1.

Hurdle-at-zero models are competitive alternatives to zero-inflated models for regressions with excessive zeros (Boucher *et al.*, 2007). The core distinction of these models is their explicit separation of the positive class from zeros, which simplifies parameter estimation in some scenarios.

A hurdle-at-zero random variable can still be represented in the form of Equation (2.1) with $Y_i^* = Y_i^{**} | Y_i^{**} > 0$ for any underlying counting distributions Y_i^{**}. The probability function for hurdle-at-zero Poisson is

$$P(Y_i = k) = \begin{cases} p_i, & \text{if } k = 0, \\ \frac{1-p_i}{1-\exp(-\lambda_i)}\frac{\exp^{-\lambda_i}\lambda_i^k}{k!}, & \text{if } k = 1, 2, 3, \ldots. \end{cases}$$

We focus on ZIP model in the remainder of the paper.

4.3. Delta estimation

We now proceed to solving Equations (4.7) and (4.8). Two approaches are considered. The first approach, chosen by Lee (2020), utilizes a bi-variate Taylor expansion such that for all $N_{j,t}$

$$\sum_{i\in N_{j,t}} \left(\Phi''_{FF,i,t-1}\Delta_{\hat{F}_{j,t}} + \Phi''_{FG,i,t-1}\Delta_{\hat{G}_{j,t}} + \Phi'_{F,i,t-1}\right) = 0, \tag{4.11}$$

$$\sum_{i\in N_{j,t}} \left(\Phi''_{FG,i,t-1}\Delta_{\hat{F}_{j,t}} + \Phi''_{GG,i,t-1}\Delta_{\hat{G}_{j,t}} + \Phi'_{G,i,t-1}\right) = 0. \tag{4.12}$$

Despite its complexity, solving this set of linear equations is generally easy.

For the sake of simplifying computationally intensive calculations, we adopt a cyclic coordinate descent optimization (Saha and Tewari, 2010;

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

Wright, 2015). In essence, we cyclically minimize the loss function with respect to each coordinate direction, one at a time, while fixing the other coordinates. Similar to many optimization techniques, coordinate descent optimization can become trapped in saddle point regions when the function to optimize is non-smooth. In ZIP regression, the loss function is smooth with everywhere-bounded second derivatives and hence avoids this problem.

Under cyclic coordinate descent optimization, the estimation of $\lambda_{i,t}$ and $p_{i,t}$ can be reduced to

$$\Delta_{\hat{G}_{j,t}} = -\frac{\sum_{i\in N_{j,t}} \Phi^{'}_{G,i,t-1}}{\sum_{i\in N_{j,t}} \Phi^{''}_{GG,i,t-1}}, \tag{4.13}$$

$$\Delta_{\hat{F}_{j,t}} = -\frac{\sum_{i\in N_{j,t}} \Phi^{'}_{F,i,t-1}}{\sum_{i\in N_{j,t}} \Phi^{''}_{FF,i,t-1}}. \tag{4.14}$$

At each iteration, only one of the above two equations is triggered. Contrarily, popular boosting libraries such as xgboost (Chen *et al.*, 2015) and lightgbm (Ke *et al.*, 2017) update $\hat{F}_{i,t}$ and $\hat{G}_{i,t}$ simultaneously using the same equations. This is equivalent to ignoring the impact of the mixed partial derivative $\Phi^{''}_{FG,i,t-1}$ in Equations (4.11) and (4.12). In the case where F and G are separable, both strategies converge to the same optimum. However, this is not the case for the negative binomial distribution (Lee, 2020) or the ZIP regression. Readers can verify that $\Phi^{''}_{FG,i,t-1}$ is generally not negligible and can be even bigger than $\Phi^{''}_{GG,i,t-1}$. This leads to solutions derived from Equations (4.13) and (4.14) that are materially different to those from Equations (4.11) and (4.12).

Cyclic coordinate descent implicitly addresses the dependence effect by taking parameter updates in turns. Compared to the full approach, it offers a viable alternative with quick turnaround time.

4.4. Concavity

The Hessian of the negative log-likelihood for distributions in the exponential family is always positive-definite and implies that the loss function is always convex.

However, the Hessian of ZIP is not necessarily positive-definite.

$$\Phi^{''}_{GG,i,t-1} = \hat{p}_{i,t-1}\hat{q}_{i,t-1} - \tilde{p}_{i,t-1}\tilde{q}_{i,t-1}, \tag{4.15}$$

$$\implies \Phi^{''}_{GG,i,t-1} = \begin{cases} \geq 0 & \text{if } |\tilde{p}_{i,t-1} - 0.5| \geq |\hat{p}_{i,t-1} - 0.5|, \\ < 0 & \text{otherwise.} \end{cases} \tag{4.16}$$

As $\tilde{p}_{i,t-1} > \hat{p}_{i,t-1}$ for $y_i = 0$, we have $\Phi^{''}_{GG,i,t-1} < 0$ if $\tilde{p}_{i,t-1} < 0.5$ for $y_i = 0$. The possibility of negative denominators in Equation (4.13) cannot be neglected. In the empirical study in Section 6, we observe more than 800 partitions in 3000 iterations having $\sum_{i\in N_{j,t}} \Phi^{''}_{GG,i,t-1} < 0$.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

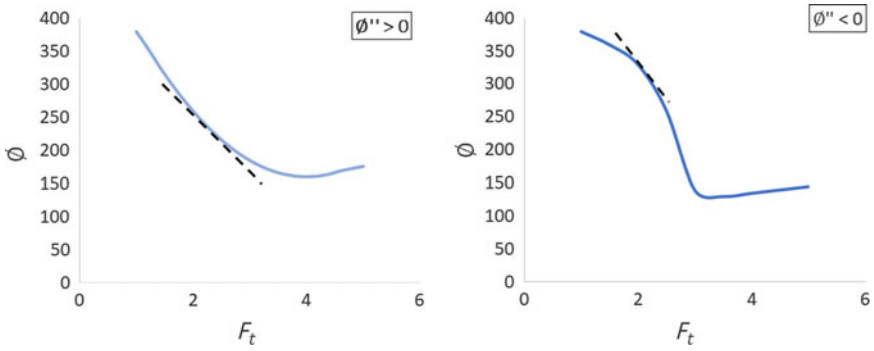


FIGURE 3: Illustration of the convexity challenge on parameters' estimation.

The concavity of the loss function implies that solving Equations (4.13) and (4.14) may result in saddle points or even local maxima of the loss function, which is in opposition to our goal.

The Hessian is employed to adjust for the magnitude indicated by the gradient (Chen *et al.*, 2015; Ke *et al.*, 2017). Hence, we can rely solely on the sign/direction from the gradient according to the gradient descent approach (Friedman, 2002) and Figure 3. As a result, Equations (4.13) and (4.14) are modified to:

$$\Delta_{\hat{G}_{j,t}} = -\frac{\sum_{i\in N_{j,t}} \Phi'_{G,i,t-1}}{|\sum_{i\in N_{j,t}} \Phi''_{GG,i,t-1}|}\ ,\ \Delta_{\hat{F}_{j,t}} = -\frac{\sum_{i\in N_{j,t}} \Phi'_{F,i,t-1}}{|\sum_{i\in N_{j,t}} \Phi''_{FF,i,t-1}|}. \tag{4.17}$$

Lee (2020) adopts another practical procedure to add positive constants to $\Phi''_{GG,i,t-1}$, which effectively reduces the possibility of negative denominators. In deep learning and regression, this handling is treated as a regularization because increasing the sizes of the denominators can effectively temper the adjustment's magnitude.

We conclude this section by summarizing the algorithm for ZIP regressions under full exposure in Algorithm 1. The algorithm uses the basis, regression, and adjust steps in Lee and Lin (2018). The δ in the basis step is the responses used for CART regression, weighted by the second derivatives Φ''. Recall that gradient boosting chooses the negative gradient $-\Phi'$ as the response. This differs from the traditional weighted the CART, where the prediction Δ is $\sum_i w_i\delta_i/\sum_i w_i$; we apply the absolute values to the denominator $\sum_i w_i\delta_i/|\sum_i w_i|$ for the above reasons.

5. INCOMPLETE EXPOSURE ANALYSIS

Lee (2020) elaborates the presence of incomplete exposures in underwriting, pricing, and claim functions and their impact to the quality of models if common handling is used without proper scrutiny.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

Algorithm 1 Delta boosting for ZIP regression under full exposure

(1) Initialize $\hat{F}_0(\mathbf{x})$, $\hat{G}_0(\mathbf{x})$ to be constants that satisfy $\{\hat{F}_0(\mathbf{x}), \hat{G}_0(\mathbf{x})\} = \underset{s,v}{\text{argmin}}\, \Phi(\mathbf{y}, v, s)$

(2) **For** $t = 1$ to $2T$ **Do**

(3) **If** t is **odd, Do**

(a) **Basis for** $\hat{F}_{i,t}$**:** Compute the working responses

$$\delta_{\hat{F}_{i,t}} = -\frac{\Phi'_{F,i,t-1}}{\Phi''_{FF,i,t-1}}.$$

(b) **Regression for** $\hat{F}_{i,t}$**:** Create t-th CART-based tree on data $\{\delta_{\hat{F}_{i,t}}, \mathbf{x}_i : i = 1, \ldots, M\}$ with weights $\Phi''_{FF,i,t-1}$.

(c) **Adjust for** $\hat{F}_{i,t}$**:** The adjustment for node $\mathbf{N}_j$ is

$$\Delta_{\hat{F}_{j,t}} = -\frac{\sum_{i\in N_{j,t}} \Phi'_{F,i,t-1}}{|\sum_{i\in N_{j,t}} \Phi''_{FF,i,t-1}|}.$$

Then select partition $\mathbf{a}_t = \underset{\mathbf{a}}{\text{argmin}} \sum_{i=1}^{M} \sum_{i\in \mathbf{N}_j} \Phi(y_i, F_{t-1}(\mathbf{x}_i) + \Delta_{F_{j,t}}, G_{t-1}(\mathbf{x}_i))$.

(d) Update $\hat{F}_t(\mathbf{x}_i) = \hat{F}_{t-1}(\mathbf{x}_i) + \Delta_{\hat{F}_{j,t}} h(\mathbf{x}_i; \mathbf{a}_t)$.

(4) **If** t is **even, Do**

(a) **Basis for** $\hat{G}_{i,t}$ **:** Compute the working responses

$$\delta_{\hat{G}_{i,t}} = -\frac{\Phi'_{G,i,t-1}}{\Phi''_{GG,i,t-1}}.$$

(b) **Regression for** $\hat{G}_{i,t}$**:** Create t-th CART-based tree on data $\{\delta_{\hat{G}_{i,t}}, \mathbf{x}_i : i = 1, \ldots, M\}$ with weights $\Phi''_{GG,i,t-1}$.

(c) **Adjust for** $\hat{G}_{i,t}$**:** The adjustment for node $\mathbf{N}_j$ is

$$\Delta_{\hat{G}_{j,t}} = -\frac{\sum_{i\in N_{j,t}} \Phi'_{G,i,t-1}}{|\sum_{i\in N_{j,t}} \Phi''_{GG,i,t-1}|}.$$

Then select partition $\mathbf{a}_t = \underset{\mathbf{a}}{\text{argmin}} \sum_{i=1}^{M} \sum_{i\in \mathbf{N}_j} \Phi(y_i, F_t(\mathbf{x}_i), G_{t-1}(\mathbf{x}_i) + \Delta_{G_{j,t}})$.

(d) Update $\hat{G}_t(\mathbf{x}_i) = \hat{G}_{t-1}(\mathbf{x}_i) + \Delta_{\hat{G}_{j,t}} h(\mathbf{x}_i; \mathbf{a}_t)$.

(5) **End For**

(6) Output $\hat{F}(\mathbf{x}_i) = \hat{F}_{2T}(\mathbf{x}_i)$ and $\hat{G}(\mathbf{x}_i) = \hat{G}_{2T}(\mathbf{x}_i)$.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

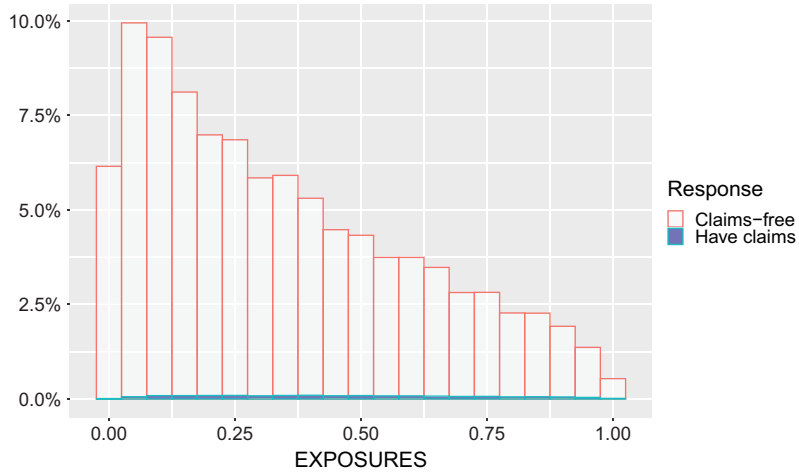


FIGURE 4: Exposure distribution of the empirical data.

Figure 4 demonstrates a standard distribution of frequency with less than 1% of the data having full exposure. This suggests to investigate a suitable method to properly address the incomplete exposure problem.

In this section, we first explain that ZIP is not additive and study four plausible options to handle exposure offsets. We also develop a necessary condition under which the probability of zero claims will not vanish due to split exposures. Actuarial research has acknowledged the need to incorporate exposure as an offset (Boucher *et al.*, 2007; De Jong and Heller, 2008; Klein *et al.*, 2015; Lee and Lin, 2018; Wuthrich and Buser, 2019), and our work aims to set a direction beyond the Poisson model to study the optimal method for addressing incomplete exposures in a regression analysis.

5.1. Time homogeneity as a model selection factor

According to the common pricing assumption in motor and home insurance, the propensity to claims and their prices should be proportional to the length of the exposure. The commonly adopted Poisson regression model addresses this situation by default, as it is the only counting distribution that inherits a time homogeneity property (Kingman, 2005).

Assume that $Y_{a,b}$ is a Poisson random variable which captures the number of claims between times a and b. Then, there exists $Y_{l_{i-1},l_i} \overset{ind}{\sim} Poisson((l_i - l_{i-1})\lambda)$ with $a = l_0 < l_1 < \cdots < l_n = b$ such that

$$Y_{a,b} = \sum_{i=1}^{n} Y_{l_{i-1},l_i}.$$

However, the above cannot be applied to other counting distributions.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

Now, we study the case for ZIP. Assume $Y_{a,b} \sim ZIP(p, (b-a)\lambda)$ is a ZIP random variable that captures the number of claims between time a and time b. In the simplest scenario, where $n = 2$ and $Y_{0,0.5}$, $Y_{0.5,1} \overset{iid}{\sim} ZIP(p, 0.5\lambda)$, we have

$$P(Y_{0,0.5} + Y_{0.5,1} = 0) = (p + (1-p)\exp^{-0.5\lambda})^2$$
$$P(Y_{0,0.5} + Y_{0.5,1} = 1) = 2(p + (1-p)\exp^{-0.5\lambda})(1-p)\lambda \exp^{-\lambda}.$$

Thus, $P(Y_{0,0.5} + Y_{0.5,1} = 0) \neq P(Y_{0,1} = 0)$, in general. Even worse, $Y_{0,0.5} + Y_{0.5,1}$ does not follow ZIP. Thus, ZIP is not additive.

5.2. Importance of exposure treatment

In this subsection, we study various options for addressing incomplete exposures, including their analytical and asymptotic behaviors.

To facilitate the discussion, we define $Y = \sum_{i=1}^{n} Y_{l_{i-1},l_i}$ with $0 = l_0 < l_1 < \cdots < l_n = 1$, $\grave{Y} = \sum_{i=1}^{\grave{n}} Y_{\grave{l}_{i-1},\grave{l}_i}$ with $0 = \grave{l}_0 < \grave{l}_1 < \cdots < \grave{l}_{\grave{n}} = 1$, $\grave{n} > n$, $l_i \in \{\grave{l}_0, \grave{l}_1, \ldots, \grave{l}_{\grave{n}}\}$. That is, $\grave{Y}$ represents the scenario with a finer split. $Y_{a,b}$ are independently distributed ZIP when the exposure periods do not overlap and $Y_{0,1} \sim ZIP(p, \lambda)$ with $0 < p < 1$ and $\lambda > 0$. All $Y_{a,b}$ have the same underlying explanatory variables. The estimation error $Y_{0,1} - \sum_{i=1}^{n} Y_{l_{i-1},l_i}$ for the optimal option should be low. As the fundamental feature of ZIP is the ability to capture excessive zeros, we focus on the behavior of $P(Y = 0)$.

Option 1 – multiplying exposure to $\exp^F$ or λ ($\exp^F \to w\exp^F$):
This option is usually chosen in ZIP research when exposures are considered (Boucher *et al.*, 2009) and is inspired by the conventional treatment in Poisson regressions. The option has the desirable feature that the mean is proportional to the exposure, $E(Y_{a,b}) = (1-p)(b-a)\lambda$. Under this option, $Y_{a,b} \sim ZIP(p, (b-a)\lambda)$ and we define $f_1(w) = \log(p + q\exp^{-w\lambda})$ as the log-probability of zero for ZIP with exposure length w. Proposition 5.1 proves that the probability of zero decreases if a policy period is split finer, and Theorem 5.2 derives the theoretical maximum of the probability divergence by proving that ZIP converges to Poisson as the exposure is split infinitesimally.

Proposition 5.1. *Let* $\grave{Y} = \sum_{i=1}^{\grave{n}} Y_{\grave{l}_{i-1},\grave{l}_i}$ *with* $0 = \grave{l}_0 < \grave{l}_1 < \cdots < \grave{l}_{\grave{n}} = 1$, $\grave{n} > n$, $l_i \in \{\grave{l}_0, \grave{l}_1, \ldots, \grave{l}_{\grave{n}}\}$ *and* $Y_{a,b} \overset{iid}{\sim} ZIP(p, (b-a)\lambda)$. *Then* $P(Y = 0) > P(\grave{Y} = 0)$.

Proof. Without loss of generality, we can assume $\grave{n} = n + 1$, $\grave{l}_i = l_i$ for $i < n$, $\grave{l}_{n+1} = l_n$ and $l_{n-1} < \grave{l}_n < l_n$. This is equivalent to splitting the last component into two parts.

$$\frac{\partial f_1(w)}{\partial w} = -\frac{q\lambda\exp(-w\lambda)}{p + q\exp(-w\lambda)}$$
$$\frac{\partial^2 f_1(w)}{\partial w^2} = \frac{pq\lambda^2\exp(-w\lambda)}{(p + q\exp(-w\lambda))^2} > 0.$$

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

TABLE 3

PROBABILITY OF ZIP IN SPLITTING SCENARIOS UNDER OPTION 1.

	(Split) 1	2	6	12	∞
0	0.91353	0.87757	0.84153	0.83056	0.81873
1	0.02707	0.06893	0.12411	0.14280	0.16375
2	0.02707	0.03582	0.02831	0.02315	0.01638
3	0.01804	0.01284	0.00509	0.00307	0.00109
4	0.00902	0.00366	0.00082	0.00036	0.00005

Thus, $f_1(w)$ is convex on w. In addition, it is trivial that $f_1(0) = 0$. Thus, f_1 is superadditive, which implies $P(Y_{0,1} = 0) \geq P(Y = 0)$ and

$$\begin{aligned}\log(P(\grave{Y} = 0)) &= \sum_{i=1}^{n-1} f_1(l_i - l_{i-1}) + f_1(\grave{l}_n - l_{n-1}) + f_1(1 - \grave{l}_n) \\ &< \sum_{i=1}^{n-1} f_1(l_i - l_{i-1}) + f_1(1 - l_{n-1})). \\ &= \log(P(Y = 0)).\end{aligned}$$

□

We can determine the upper bound of the discrepancy $P(Y_{0,1} = 0) - P(Y = 0)$ by studying the asymptotic distribution of Y.

Proposition 5.2. *$Y \to Poisson(q\lambda)$ weakly as $n \to \infty$ given $\max(l_i - l_{i-1}) = O(1/n), \forall i$.*

Proof.

$$\begin{aligned}E(\exp^{tY}) &= \prod_{i=1}^{n} \left(p + q \exp^{(l_i - l_{i-1})\lambda(\exp^t - 1)}\right) \\ &= \prod_{i=1}^{n} \left(1 + q(\exp^{(l_i - l_{i-1})\lambda(\exp^t - 1)} - 1)\right) \\ &= \prod_{i=1}^{n} \left(1 + q(l_i - l_{i-1})\lambda(\exp^t - 1) + o(1/n)\right) \\ &\to \exp^{q\lambda(\exp^t - 1)} \quad \text{as } n \to \infty.\end{aligned}$$

□

The condition of $\max(l_i - l_{i-1}) = O(1/n)$ ensures that the quadratic error is negligible asymptotically. Propositions 5.1 and 5.2 suggest that the effect of zero-inflation wears off as the exposure is increasingly granular and disappears completely asymptotically. This impact is detailed numerically in Table 3, which illustrates the probability of $Y = 0, 1, 2, 3, 4$ under various equal splittings of a random variable with $\lambda = 2$ and $p = 0.9$.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

TABLE 4

RATIO OF PROBABILITY IN SPLITTING SCENARIOS.

	(Split) 1	2	6	12	∞
0	1.00000	0.96063	0.92118	0.90918	0.89622
1	1.00000	2.54650	4.58540	5.27590	6.04960
2	1.00000	1.32320	1.04600	0.85542	0.60496
3	1.00000	0.71161	0.28212	0.17036	0.06050
4	1.00000	0.40581	0.09038	0.04008	0.00605

The column with a splitting of ∞ refers to the Poisson random variable with mean $0.1 * 2 = 0.2$, and the choice of this terminology is explained by Proposition 5.2.

To help explain the tail divergence, Table 4 provides the calculated ratio between each split combination to the full year ZIP (first column). The ratios suggest that as the splitting becomes finer, the impact of exposure on accuracy bifurcates and rapidly deteriorates as responses increases. The divergence is detrimental in general insurance, as the tail risk is the focus of interest. For example, given the same mean, if the risk exposure is artificially split too fine, the estimated probability of the occurrence of over 4 claims can be less than 1% of the true potential. Thus, addressing partial exposures is indeed found to be decisive for the overall performance of ZIP.

Option 2 – multiplying exposures to $\exp^G$ ($\exp^G \to w \exp^G$):
In contrast to Option 1, which applies the proportional offset to λ, Option 2 explores the application of offset to $\exp^G$. Under this option, $Y_{a,b} \sim ZIP\left(\frac{(b-a)\exp^G}{1+(b-a)\exp^G}, \lambda\right)$ and we define $f_2(w) = \log\left(\frac{w\exp^G + \exp^\lambda}{w\exp^G + 1}\right)$ to capture the log-probability of zero for ZIP with exposure length w. Inspecting $P(Y = 0)$, we have

$$\prod_{i=1}^{n} P(Y_{l_{i-1},l_i} = 0) = \prod_{i=1}^{n} \left(\frac{(l_i - l_{i-1})\exp^G}{(l_i - l_{i-1})\exp^G + \exp^{-\lambda_{i,t-1}}}\right) \to 0 \quad \text{as } n \to \infty.$$

Thus, this option drives $P(Y = 0)$ to 0 quickly as the split becomes finer, making this handling of $\exp^G$ undesirable. From the above derivation, it is trivial that the necessary condition for $\prod_{i=1}^{n} P(Y_{l_{i-1},l_i} = 0)$ not to vanish is to have $P(Y_{l_{i-1},l_i} = 0)$ decrease as $l_i - l_{i-1}$ increases. Otherwise, the smaller probability coupled with compounding guarantees $P(Y = 0) = 0$ asymptotically. Following this logic, we rule out this common option of applying the offset to p ($p \to wp$) suggested in De Jong and Heller (2008).

Option 3 – multiplying exposure to q ($q \to wq$):
Because applying exposure to reduce p is not desirable, we shift our attention to applying the offset to reduce q. Under this option, $Y_{a,b} \sim ZIP(1 - (b - a)q, \lambda)$

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

TABLE 5

PROBABILITY OF ZIP IN SPLITTING SCENARIOS UNDER OPTION 3.

	1	2	6	12	∞
0	0.91353	0.91540	0.91659	0.91688	0.91717
1	0.02707	0.02590	0.02517	0.02500	0.02482
2	0.02707	0.02608	0.02546	0.02531	0.02516
3	0.01804	0.01763	0.01736	0.01729	0.01723
4	0.00902	0.00906	0.00907	0.00907	0.00907

and we define $f_3(w) = \log(1 - wq + wq\exp^{-\lambda})$ to capture the log-probability of zero for ZIP with exposure length w. As $E(Y_{l_{i-1},l_i}) = (l_i - l_{i-1})q\lambda$, this option also has the desirable feature of retaining the expectation, in common with Option 1. Proposition 5.3 proves that the probability of zero is overestimated as we split the exposure.

Proposition 5.3. *Let* $\grave{Y} = \sum_{i=1}^{\grave{n}} Y_{l_{i-1},l_i}$ *with* $0 = \grave{l}_0 < \grave{l}_1 < \cdots < \grave{l}_{\grave{n}} = 1$, $\grave{n} > n$, $l_i \in \{\grave{l}_0, \grave{l}_1, \cdots, \grave{l}_{\grave{n}}\}$ *and* $Y_{a,b} \overset{iid}{\sim} ZIP(1-(b-a)q, \lambda)$. *Then* $P(Y=0) < P(\grave{Y}=0)$.

Proof.

$$\frac{\partial f_3(w)}{\partial w} = \frac{q(\exp^{-\lambda} - 1)}{1 + wq(\exp^{-\lambda} - 1)}$$

$$\frac{\partial^2 f_1(w)}{\partial w^2} = -\left(\frac{q(\exp^{-\lambda} - 1)}{1 + wq(\exp^{-\lambda} - 1)}\right)^2 < 0.$$

Thus, $f_3(w)$ is concave on w. Additionally, it is trivial that $f_3(0) = 0$. Thus, f_3 is subadditive. The proposition can then be proved similarly to Proposition 5.1. □

$P(Y=0)$ also converges asymptotically.

Proposition 5.4. $\lim_{n\to\infty} P(Y=0) = \exp^{q(\exp^{-\lambda}-1)}$ *given* $\max(l_i - l_{i-1}) = O(1/n), \forall i$.

Proof.

$$P(Y=0) = \prod_{i=1}^{n} \left(1 + (l_i - l_{i-1})q(\exp^{-\lambda} - 1)\right)$$

$$\lim_{n\to\infty} P(Y=0) = \exp^{q(\exp^{-\lambda}-1)}.$$ □

Similar to Option 1, we conclude this option with a numerical illustration in Table 5. This option can be seen to have a more satisfactory performance than Option 1 by comparing Tables 3 and 5, and the probability difference is still acceptable for larger responses. The main source of error comes from the probability for $Y = 1, 2$.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

TABLE 6

PROBABILITY OF ZIP IN SPLITTING SCENARIOS UNDER OPTION 4.

	1	2	6	12	∞
0	0.91353	0.91105	0.90930	0.90885	0.90840
1	0.02707	0.02720	0.02728	0.02730	0.02732
2	0.02707	0.02740	0.02762	0.02768	0.02773
3	0.01804	0.01854	0.01887	0.01895	0.01904
4	0.00902	0.00954	0.00990	0.00999	0.01008

Option 4 – applying the offset to $\exp^{-G}$ ($\exp^{-G} \to w\exp^{-G}$):
This is another practicable option to reduce q. Under this option, $Y_{a,b} \sim ZIP\left(\frac{\exp^G}{\exp^G+(b-a)}, \lambda\right)$ and we define $f_4(w) = \log\left(\frac{\exp^G + w\exp^{-\lambda}}{\exp^G + w}\right)$ to capture the log-probability of zero for ZIP with exposure length w.

Proposition 5.5. *Let* $\grave{Y} = \sum_{i=1}^{\grave{n}} Y_{\grave{l}_{i-1},\grave{l}_i}$ *with* $0 = \grave{l}_0 < \grave{l}_1 < \cdots < \grave{l}_{\grave{n}} = 1$, $\grave{n} > n$, $l_i \in \{\grave{l}_0, \grave{l}_1, \cdots, \grave{l}_{\grave{n}}\}$ *and* $Y_{a,b} \overset{iid}{\sim} ZIP\left(\frac{\exp^G}{\exp^G+(b-a)}, \lambda\right)$. *Then* $P(Y=0) > P(\grave{Y}=0)$.

Proof. The proposition can be proved the same way as Proposition 5.1, as $f_4(w)$ is convex on w and $f_4(0) = 0$, making f_4 superadditive. □

Proposition 5.6. $\lim_{n\to\infty} P(Y=0) = \exp^{\frac{\exp^{-\lambda}-1}{\exp^G}}$ *given* $\max(l_i - l_{i-1}) = O(1/n), \forall i$.

Proof.

$$P(Y=0) = \prod_{i=1}^{n}\left(1 + \frac{l_i - l_{i-1}}{\exp^G + l_i - l_{i-1}}(\exp^{-\lambda} - 1)\right)$$

$$\Longrightarrow \lim_{n\to\infty} P(Y=0) = \exp^{\frac{\exp^{-\lambda}-1}{\exp^G}}.$$ □

Thus, $P(Y=0)$ converges asymptotically under this option. Compared with Option 3 (Table 5), this option (Table 6) offers a similarly close approximation for probability of $Y = 1, 2$ in addition to satisfactory performance for the rest. From the previous propositions, we have $\sum_{i=1}^{n} f_1(l_i - l_{i-1}) < \log(p + q\exp(-\lambda))$, $\sum_{i=1}^{n} f_4(l_i - l_{i-1}) < \log(p + q\exp(-\lambda))$, and $\log(p + q\exp(-\lambda)) < \sum_{i=1}^{n} f_3(l_i - l_{i-1})$. Thus, $f_3(w)$ is the largest among the chosen approximations. However, the approximation error differs in each problem. In particular, there is no definite ordering between the approximation errors from Option 1 and Option 4, although the error for Option 4 is generally smaller as G increases. Simply by inspecting the above possibilities, actuaries should benefit from a careful study of exposure handling. We have also considered a few more options, including $p \to p^w$ and a combination of the above options. Some practical candidates were excluded because they do not exhibit an intuitive rationale. In implementation, Algorithm 1 can be applied for each of the

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

above-mentioned options by modifying $\hat{p}_{i,t}$, $\tilde{p}_{i,t}$, and $\hat{\lambda}_{i,t}$ according to the chosen option. Using Option 4 as an example, we have

$$\hat{p}_{i,t} = \frac{\exp^{\hat{G}_{i,t}}}{\exp^{\hat{G}_{i,t}} + w_i}, \quad \hat{q}_{i,t} = 1 - \hat{p}_{i,t},$$
$$\tilde{p}_{i,t} = \frac{\exp^{\hat{G}_{i,t}}}{\exp^{\hat{G}_{i,t}} + w_i \exp^{-\hat{\lambda}_{i,t}}} I_{y_i=0}, \quad \tilde{q}_{i,t} = 1 - \tilde{p}_{i,t}.$$

6. Application to insurance data

In this section, we compare the performance of various competitive regression models with an empirical test of insurance claim data from a Canadian insurer. GLM, GAM, gradient boosting, and delta boosting implementations are considered as the modeling candidates, with Poisson, negative binomial, and ZIP as the distributions to be regressed. We also apply some common sampling techniques for unbalanced data on Poisson for illustration. Not every combination is tested due to the availability of relevant libraries. The data consist of policy and claim information at vehicle level for collision coverage in Ontario. Collision coverage protects the insured from the cost of repairing or replacing their vehicles if the covered vehicles collide with another vehicle or any object in or on the ground. The data set includes the claim experience for accident years 2001 to 2005. The response to be predicted is the claim frequency, expressed as the number of claims per vehicle year. More than 1000 variables are available in the data warehouse in which the data are stored, including the characteristics of policyholders, drivers, and vehicles. Fifty variables, including those created via feature engineering, are eventually selected for modeling. The data include 290,147 vehicle-years and an overall claim frequency of 4.414%.

6.1. Model comparisons

Lee and Lin (2018), Lee (2020) use the same data, with the deletion of an insignificant portion of observations with missing values, to compare the performance of a few competing predictive modeling techniques. We randomly select 80% of the data for training, and the remaining 20% is used for independent holdout data to validate the model performance, as shown in Table 7. Interested readers can find a comprehensive treatment of the data processing and modeling in Lee and Lin (2018). We select commonly used diagnostics to assess the model performance. The measures include the loss based on Poisson distribution Φ_{Poisson}, the loss based on underlying distribution Φ_{Others} (negative binomial, hurdle-at-zero Poisson, and ZIP for the respective models), the lifts, and the Gini index. The winners in each diagnostic in holdout are in bold. For models created by negative binomial, hurdle-at-zero, or ZIP regression, Φ_{Poisson} is calculated based on treating the derived mean for Poisson. We

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

TABLE 7

MODEL DIAGNOSTICS FOR ALL CANDIDATES IN HOLDOUT DATASET.

Model setting	Legend in plots	$\Phi_{Poisson}$	Φ_{Others}	Lift (one-way)	Lift (two-way)	Gini index
Poisson in GLM	GLM	0.0000		2.3849	5.5750	0.2700
Poisson in GAM	GAM	−39.7881		2.4979	6.0089	0.3193
Poisson in DBM with 100pc upsample		−53.1643		2.4090	6.3501	0.2991
Poisson in DBM with 400pc upsample		−50.7689		2.4294	6.3368	0.3173
Poisson in DBM with 900pc upsample		−39.5844		2.4374	6.4941	0.3036
Poisson in DBM with 25pc downsample		−54.9272		2.3890	6.1660	0.3204
Poisson in DBM with 50pc downsample		−51.6156		2.3930	6.3081	0.3145
Poisson in DBM with 75pc downsample		−51.5940		2.4170	6.2387	0.2987
Poisson in DBM with SMOTE 100pc		−4.0290		2.4090	6.4885	0.3048
Poisson in DBM with SMOTE 400pc		−53.9116		2.4373	6.5645	0.3114
Poisson in DBM with SMOTE 900pc		−55.2752		2.4253	6.1956	0.3136
ZIP in GLM	ZIPGLM	−8.2091	−17.9507	2.3971	5.2101	0.2948
Hurdle Poisson in GLM		61323521	160.2337	2.5625	6.9775	0.2985
Hurdle Poisson in GLM (adjusted)		165.6324	159.7341	2.5624	6.9766	0.2909
ZIP in GAM	ZIPGAM	−139.0122	−151.64	2.6957	6.8861	0.3414
Negative Binomial in GLM (shape fixed)		6.1623	-9.2600	2.4091	5.5268	0.2669
Negative Binomial in GAM (shape fixed)		−39.1224	−53.2532	2.5058	6.0874	0.2870
Negative Binomial in DBM (shape fixed)		−61.4535	−68.5992	2.5502	6.5824	0.3334
Negative Binomial in DBM	NBDBM(2)	−62.7674	−75.5533	2.5665	6.6939	0.3562
ZIP in DBM with Option 1	ZIP1	−60.2006	−74.3183	2.5464	6.6420	0.3440
ZIP in DBM with Option 3	ZIP3	−60.3271	−58.9072	2.5383	6.9118	0.3427
ZIP in DBM with Option 4	ZIP4	**−191.9117**	**−198.1252**	**2.7763**	**8.0933**	**0.3819**

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

should compare the losses of the models' corresponding distributions because they represent the ex-ante belief of the underlying distribution of the data. This metric should carry the highest weight because it dictates that the search of parameters during the training due to loss minimization is equivalent to maximizing the log-likelihood in this experiment. Both metrics are captured as the difference between their losses and the corresponding loss from GLM Poisson for ease of comparison. We can draw several insights from Table 7.

First, the choice of incomplete exposure handling dictates the performance of machine learning modeling. Although, to the best of our knowledge, it has not been proposed in the literature, Option 4 wins with respect to the selected metrics. As shown in Figure 4, policy periods are generally split granularly, driving high values of p_i given G_i in all options discussed. We conjecture that because the treatment in Option 3 further overestimates the value of $P(Y_i = 0)$ given p_i, this partially explains why the results of this option are less favorable. In addition, we observe that the approximation error of Option 4 is less than that of Option 1 as G_i increases. The low q_i for the delta boosting in Figure 6(b) suggests a high portion of perfect states or high G_i. In the rest of the study, we focus on the model associated with Option 4 given its strength in all metrics considered.

Second, the hurdle-at-zero regression produces a highly unsatisfactory Φ_{Poisson} because the weight offset assignment in this distribution structure is not appropriate in our problem setting. The expected response and the probability of zero with an incomplete exposure are $(1 - p_i)(w_i\lambda_i)/(1 - \exp(-w_i\lambda_i))$ and p_i, respectively. The first number indicates that the denominator counterbalances the weight effort. Because the second number is constant with respect to exposure, it faces the same problem as Option 2 in which $P(Y = 0)$ goes to zero as an observation is split into finer periods, forcing G_i to be large to offset this effect. Conversely, as G_i becomes larger, λ_i must also increase to match the given sample means. Further inspection of the empirical results shows that a small fraction ($< 0.1\%$) of the testing observations have predictions bigger than 1 (the sample mean being 0.04), driving the loss to become extraordinarily high. It is also interesting that those observations impact the Φ_{Others} much less than the other observations, as almost all of those outliers are no-claims observations with loss equal to $\log(p_i)$. Large λ_i values thus only impact Φ_{Poisson} but not Φ_{Others} in those observations. Another version of hurdle Poisson is run by removing 0.5% of training observations that have high predictions. Although still worse than the standard Poisson regression, the predictions are considerably better fit to the Poisson distributions. This again confirms the presence of incomplete exposure, if not handled with caution, posts a threat to accurate predictive modeling.

Third, Table 7 indicates that Φ_{Poisson} is lowest for the delta boosting implementation of ZIP, which is counterintuitive at first glance. The ZIP regression aims to minimize Φ_{Others}, and we artificially multiply q_i and λ_i as the mean parameter and plug it into Φ_{Poisson}. In contrast, the mean and Φ_{Poisson} are directly optimized by the Poisson regression and, thus, should perform better

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

in this metric. The result suggests that the additional flexibility due to the zero inflation may lead to a more effective search of optimal parameters, which in turn generates superior performance over a wider array of loss functions. We notice the same phenomenon in GLM and GAM, where the Poisson loss for ZIP is likewise better than that for Poisson.

Fourth, machine learning implementations outperform simpler candidates. We notice that the models' accuracy improves as we move from GLM to GAM to boosting. This suggests that the assumption of linearity and independence among explanatory variables may be too restrictive.

Fifth, the choice of distribution plays a major role in predictive performance. From one-parameter Poisson to two-parameter ZIP or two-parameter negative binomial regression, the increase in the number of parameters enhances the explanatory power over the data. However, we also note degradation in the single-parameter negative binomial, in which the shape parameter is held constant for all observations. The assumption may restrict the search for optimal mapping. We also observe that data distribution modification through up-sampling, down-sampling, or SMOTE is not effective in this empirical study, possibly because the assumption of (close-to) uniform sampling rate is too strong. In Equation (4.6), we prove that the optimal sampling rate should differ for each observation.

6.2. The lift and Gini index

The lift and the Gini index are auxiliary metrics that evaluate the models without an assumption on the underlying distribution. The lift is the ratio of the average response of the top decile prediction over the average response of the bottom decile (for two-way) or population (for one-way). In simpler terms, the lift can be interpreted as a measure of the ability of a model to differentiate observations. A higher lift illustrates that the model is more capable of separating the extreme values from the average. As the lift focuses on the tails of the distribution, the measure helps actuaries effectively devise risk mitigation plans with a variety of tools beyond pricing, including but not limited to underwriting, reinsurance, and mandating the implementation of safety measures. ZIP demonstrates a remarkable ability to capture observations with high claim frequency. The top 10% risk classified by ZIP contributes around 28% of the total claims according to the one-way lift. This discriminating power allows insurance companies to take more targeted actions to manage risks. In highly simplified circumstances, where insurers remove those risks through underwriting, the remaining 90% of the insured represent only 72% of the original loss, allowing insurance companies to charge 20% (1-72/90) lower on average. This additional room for discounts will enable insurers to attract more favorable risks. For better illustration, we capture the lift plots for a few selected candidates in Figure 5. From the figure, the boosting implementations exhibit good lifts and alignment with the $y = x$ straight line, which benchmarks the perfect predictions. In contrast to the lift measure, the Gini index coefficient assesses

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

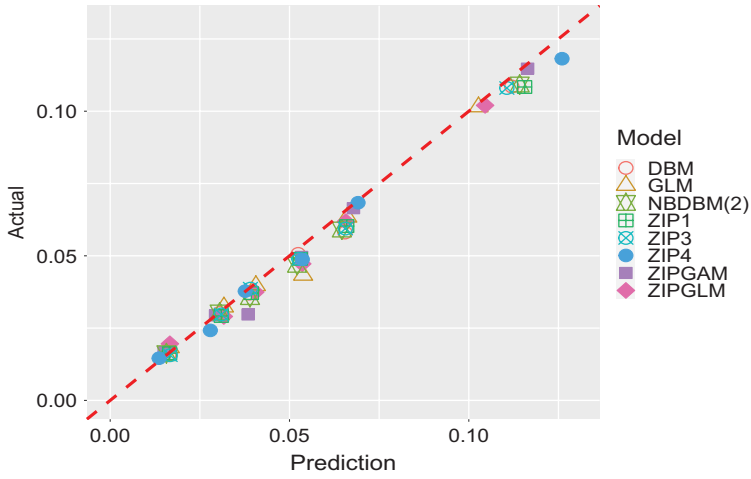


FIGURE 5: Lift plot for competing candidates.

the discriminatory power of machine learning models using the full data set and is therefore more robust against the fluctuation of predictive performance in both tails (Wuthrich and Buser, 2019). In both measures, ZIP exhibits the best performance.

6.3. Partial dependence plots

Actuaries can use partial dependence plots to assess the predictive contribution of individual variables. Defined in Friedman (2001), the partial dependence plot is a visualization tool that views the marginal contribution of a selected small subset, usually one, of the input variables. With proper transformation and normalization, the plots can be presented in a manner similar to that of differential (or relativity) plots. This type of low-dimensional visualization allows actuaries to understand the predictive patterns at variable levels and hence make models more explainable to decision-makers and regulators. As an illustration, the plots for years licensed and renewal counts are depicted in Figure 6(a). The shaded bar indicates the portion of risk at that level, whereas the line indicates the differential, with the values normalized such that the differential at the mid-point of all levels is 1. We observe that the differences among all candidates are visible even at the levels with significant exposures. These differences contribute to the diverging predictions between both models. The curves for GLM and GAM are usually steeper because of the extrapolation of linearity/spline, whereas boosting implementations tend to have a flattened curve due to tree-based learning. In addition to the standard plots for the expected response, we drill further to analyze the contributions of q and λ. Figure 6 illustrates a surprising phenomenon. Whereas all of the zero-inflated

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

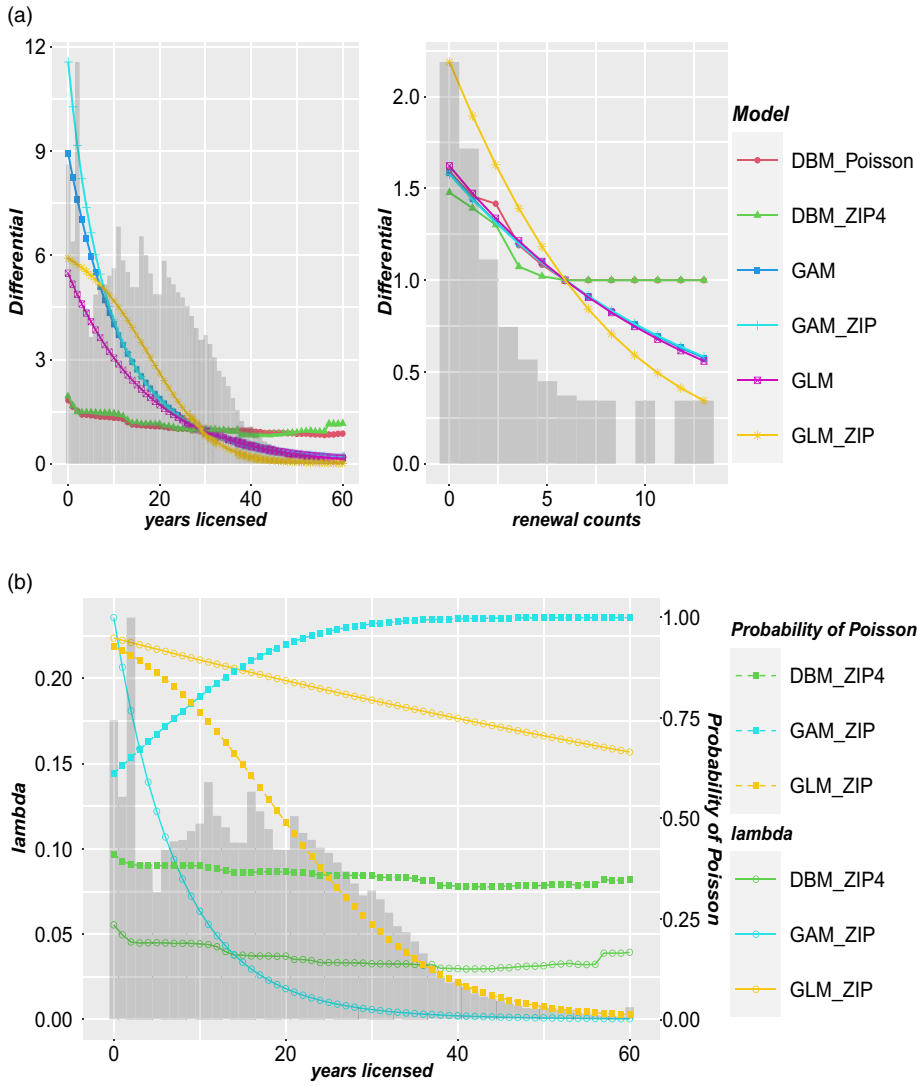


FIGURE 6: Differential plots to assess marginal contribution of variables. (a) Differential plots for years licensed and renewal count on frequency. (b) Differential plots, split analysis on p and lambda.

Poisson models demonstrate a reasonable fit to the data, with a positively correlated trend in the plot of partial dependence on the expected frequency $q\lambda$, the contributions of q and λ vary significantly. The curves for the parameters in delta boosting are stable and similar, whereas the probability of Poisson in GLM drops substantially as the input variable increases. This implies that q contributes more to the reduction of frequency $q\lambda$. Unlike the pattern observed in GLM and delta boosting, q increases along with the input variable value and λ drops at a faster rate to offset the increase in GAM.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

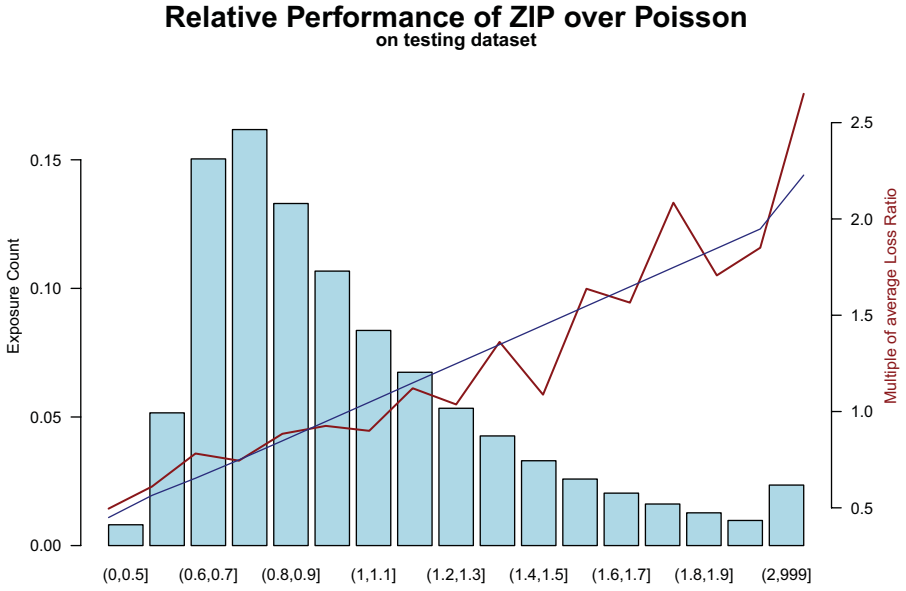


FIGURE 7: Double lift plot for ZIP over Poisson.

6.4. Double lift plot

We conclude this section with a double lift plot. Following Lee and Lin (2018), we use this plot to directly compare the predictive performances of the ZIP and Poisson models, both implemented through delta boosting. We explain the concept further through Figure 7. The buckets in the X-axis indicate the ratio of ZIP prediction over Poisson prediction, with the blue line indicating the bucket's average. The red line indicates the ratio of the actual value over the Poisson predictions. From an actuarial perspective, we can consider the ZIP regression as a new costing algorithm, whereas the Poisson model is the current algorithm. The blue line represents the dislocation, that is, the change in rates charged to individual policyholders. Correspondingly, the red line is the loss ratio.

From Figure 7, we witness a significant positive relationship between the blue line (indicated rate change) and the red line (loss ratio) with a correlation of 0.95, suggesting that ZIP can effectively capture the residuals coming from the Poisson model. Thus, our proposed algorithm is more equitable to the market. Compared to the double lift plot in Lee (2020), we observe a wider spread of exposure across the dislocation range. In particular, a consequential proportion of the insured receives over 20% premium change if we move from Poisson to ZIP. While this implies that the ZIP is capable of distinguishing individual loss exposures, as supported by a high correlation with the actual claims, insurers will likely curb the magnitude of change to avoid customer shock at a significant price swing.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

7. Concluding remarks

This paper contributes to two main aspects of research on data science in insurance applications.

ZIP as an adaptive re-weighting mechanism to handle imbalanced financial data: Most of the data-handling techniques developed to address imbalanced data involve the judgmental selection of re-sampling portions. Not only does this introduce additional noise, but it also requires post-modeling calibration. Additionally, the assumption of a uniform sampling rate for all observations is too strong for insurance data. Our proposal eliminates this complication because it iteratively re-weights each observation to minimize the underlying loss function. Our algorithm automatically reverts the weights to unconditional weights and does not require further manual calibration.

Examination of various incomplete exposure handling: In insurance, the characteristics of the insured parties or the macro-environment can change within the same policy period. Common data management involves splitting the full period into several smaller pieces. The proportion of observations impacted by partial exposures is not negligible. As such, determining a proper handling of this scenario can have a positive influence on the predictive accuracy. We examine possible options and their strengths and weaknesses using mathematics. Through this empirical study, we confirm that the choice of handling has a decisive impact on the results.

Notes

1. www.iii.org/fact-statistic.
2. https://www.ey.com/en_uk/news/2019/12/uk-motor-insurance-market-to-record-underwriting-loss-in-2019.
3. https://www.iii.org/sites/default/files/docs/pdf/insurance_factbook_2019.pdf.
4. https://www.insuranceinstitute.ca/en/cipsociety/Articles/Items/2019/03/30/Auto-loss-ratios-unsustainable-in-these-provinces
5. https://www.canadianunderwriter.ca/earnings-ratings/how-profitable-have-canadian-pc-insurers-been-so-far-this-year-1004171955/.
6. http://www.actuarialstandardsboard.org/standards-of-practice/.
7. https://www.casact.org/professionalism/standards/princip/.

References

Boucher, J.-P., Denuit, M. and Guillén, M. (2007) Risk classification for claim counts: A comparative analysis of various zeroinflated mixed poisson and hurdle models. *North American Actuarial Journal*, **11**(4), 110–131.

Boucher, J.-P., Denuit, M. and Guillen, M. (2009) Number of accidents or number of claims? An approach with zero-inflated poisson models for panel data. *Journal of Risk and Insurance*, **76**(4), 821–846.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984) *Classification and Regression Trees*. Boca Raton, Florida, USA: CRC Press.

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press

BÜHLMANN, H. and GISLER, A. (2006) *A Course in Credibility Theory and Its Applications*. Berlin, Germany: Springer Science & Business Media.
CALDERÍN-OJEDA, E., GÓMEZ-DÉNIZ, E. and BARRANCO-CHAMORRO, I. (2019). Modelling zero-inflated count data with a special case of the generalised poisson distribution. *ASTIN Bulletin: The Journal of the IAA*, **49**(3), 689–707.
CHAWLA, N.V., BOWYER, K.W., HALL, L.O. and KEGELMEYER, W.P. (2002) Smote: A ynthetic minority over-sampling technique. *Journal of artificial intelligence research*, **16**, 321–357.
CHEN, T., HE, T., BENESTY, M., KHOTILOVICH, V. and TANG, Y. (2015) Xgboost: Extreme gradient boosting. R package version 0.4-2, 1–4.
DE JONG, P. and HELLER, G.Z. (2008) *Generalized Linear Models for Insurance Data*. Cambridge, UK: Cambridge University Press.
FERNÁNDEZ, A., GARCÍA, S., GALAR, M., PRATI, R.C., KRAWCZYK, B. and HERRERA, F. (2018) *Learning from Imbalanced Data Sets*. Springer.
FREUND, Y. and SCHAPIRE, R.E. (1995) A desicion-theoretic generalization of on-line learning and an application to boosting. *European Conference on Computational Learning Theory*, pp. 23–37. Springer.
FRIEDMAN, J.H. (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, **29**(5), 1189–1232.
FRIEDMAN, J.H. (2002) Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**(4), 367–378.
GEE, J. and BUTTON, M. (2019) The financial cost of fraud 2019: The latest data from around the world. Tech. rep., Crowe UK.
HAIXIANG, G., YIJING, L., SHANG, J., MINGYUN, G., YUANYUE, H. and BING, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, **73**, 220–239.
HE, H. and MA, Y. (2013). *Imbalanced learning: Foundations, Algorithms, and Applications*. New York, USA: John Wiley & Sons.
KE, G., MENG, Q., FINLEY, T., WANG, T., CHEN, W., MA, W., YE, Q. and LIU, T.-Y. (2017) Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pp. 3146–3154.
KINGMAN, J.F.C. (2005) Poisson processes. Encyclopedia of biostatistics 6.
KLEIN, N., KNEIB, T. and LANG, S. (2015) Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, **110**(509), 405–419.
LAMBERT, D. 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.
LEE, S.C. (2020) Delta boosting implementation of negative binomial regression in actuarial pricing. *Risks*, **8**(1), 19.
LEE, S.C. and LIN, S. (2018) Delta boosting machine with application to general insurance. *North American Actuarial Journal*, **22**(3), 405–425.
SAHA, A. and TEWARI, A. (2010) On the finite time convergence of cyclic coordinate descent methods. arXiv preprint arXiv: 1005.2146.
SCHAPIRE, R.E. (1990) The strength of weak learnability. *Machine Learning*, **5**(2), 197–227.
TEUGELS, J.L. and VYNCKIE, P. (1996). The structure distribution in a mixed poisson process. *International Journal of Stochastic Analysis*, **9**(4), 489–496.
WRIGHT, S.J. (2015) Coordinate descent algorithms. *Mathematical Programming*, **151**(1), 3–34.
WUTHRICH, M.V. and BUSER, C. (2019). Data analytics for non-life insurance pricing. Swiss Finance Institute Research Paper 2019 (16-68).

SIMON CK LEE
Department of Statistics and Actuarial Science
The University of Hong Kong
E-Mail: slee2016@hku.hk

https://doi.org/10.1017/asb.2020.40 Published online by Cambridge University Press