Contents lists available at *Dergipark*

Journal of Scientific Reports-A

journal homepage: *https://dergipark.org.tr/tr/pub/jsr-a*

***E-ISSN: 2687-6167*** ***Number 59, December 2024***

***RESEARCH ARTICLE***

*Receive Date: 02.09.2024* *Accepted Date: 25.12.2024*

# Comparison of regression and tree-based methods for the prediction of zero-inflated claim data

Aslıhan Şentürk Acar*

**Hacettepe University, Department of Actuarial Sciences, Beytepe, Ankara, Turkey*
*aslihans@hacettepe.edu.tr, ORCID: 0000-0002-1708-2028*

**Abstract**

Pricing non-life insurance products is based on the prediction of two components; claim frequency and claim severity. In this study we focus on claim frequency data that has a zero-inflated structure. Although zero-modified regression models such as zero-inflated and hurdle models are used for data sets with excess zeros, machine learning (ML) methods are also preferred for this type of data sets in recent years. When the objective is the prediction, ML methods generally provide more accurate results than regression models especially for large and complex datasets. Tree-based ML methods run decision trees as the base of the algorithm and improve performance by using the predictions of multiple trees. Combining the traditional methods with ML methods is a current popular approach for prediction tasks. Objective of this study is to compare the predictive performance of regression methods and tree-based ML methods for zero-inflated claim frequency data using a real insurance dataset. Motor third party liability insurance claim data from an insurance company in Turkey is used for the case study. To predict claim frequency, generalized linear models (GLM), zero-inflated model and hurdle model are used under Poisson distribution as regression models and regression trees, boosting and GLM-Boost that is a combination of GLM and Boosting algorithm are used as ML methods. Predictive performances of candidate models are compared using both average in-sample and average out-of-sample losses. According to the case study results, ML methods performed better predictive performance than zero-modified models. Specially, GLM-Boost method performed best among others and that is a promising result for the approaches that are combinations of GLM and ML methods.



*Keywords:* non-life insurance, claim frequency, zero-inflated data, machine learning, predictive modeling

## 1. Introduction

Actuarial pricing methods in non-life insurance is generally based on generalized linear models due to the ease of implementation and interpretation. Beside the practical usage, GLMs assume a linear relationship between the transformed response and explanatory variables on the basis of link function, need an exponential dispersion family distribution for the response variable, has sensitivity to multicollinearity and don't take into account interactions and

non-linear relationships. Also, GLMs may struggle with high-dimensional data with regards to matrix operations in maximum likelihood estimation. Because of these drawbacks, machine learning methods become popular since they can handle interactions and non-linearities automatically, can control overfitting, have flexibility with data types and are generally robust to outliers. Due to the data driven property of insurance business, machine learning methods became popular also for insurance data modelling in recent years [1-5]. Wuthrich and Buser [1] use various ML methods including regression trees, ensemble methods and neural networks for claim frequency data. Liu et al. [2] compare the predictive performance of AdaBoost algorithm with GLM, neural network and support vector machines to predict claim frequency. Dal Pozzolo [3] used various ML methods to estimate claims in Kaggle competition. Noll et al. [4] used GLM, regression trees, boosting, GLM-Boost and neural network methods to predict claim frequency data. Clemente et al. [5] compared GLM and gradient boosting model for claim frequency data.

Classification and regression trees are simple and visually practical methods that provide basis for other ML methods such as boosting. Boosting is an iterative ensemble learning method that is a combination of many weak learners such as regression trees. For the predictive purposes, boosting algorithm performs well due to the iterative process that combines weak learners into a strong learner by minimizing error. Various boosting algorithms are used to estimate both claim frequency and claim severity data [2, 6-9]. All these studies emphasise the promising performance of boosting algorithms for claim data modelling. In recent years, combination of ML algorithms and traditional methods is also popular for predictive issues in insurance data [1, 4]. One of these methods is GLM-Boost [4] that is a combination of GLM and boosting algorithm.

From the other side, main characteristic of motor insurance claim data is excess zeros due to the No Claim Discount (NCD) system and deductible modification. Policyholders don't report low-cost claims not to pay over-premium in the next policy year or when the claim size is below the deductible amount. Zero-inflated and hurdle models are popular for imbalanced data that is frequently observed property in insurance [10, 11]. Those zero-modified methods are used as an alternative to GLMs for zero-inflated claim frequency data [12, 13].

In this study, we aim to compare predictive performance of traditional GLM, zero-inflated model, hurdle model, regression trees, boosting and GLM-Boost approach for the prediction of a zero-inflated motor insurance claim frequency data. Different from other studies in the literature, zero-inflated and hurdle models are compared with tree-based ML methods for zero-inflated claim frequency data in this study.

In the second part of the paper, statistical methods are summarized, in the third part, dataset is explained and the statistical analysis is performed. The paper ends with the conclusion part.

## 2. Statistical Models

### *2.1. Generalized linear models*

Generalized linear models are purposed by Nelder and Wedderburn [14]. The distribution of response variable in GLMs can be any distribution from exponential family and the mean of response variable can be a linear function of explanatory variables on different scales depending on the ink function.

For $i = 1,2,\dots,n$ let $N_i$ be the number of claims of policy $i$. Under GLM, $N_i$'s are assumed to be independent. Mean function of $N_i$ is defined as,

$$E(N_i) = h^{-1}(\boldsymbol{x_i'\beta}) \tag{1}$$

where $\boldsymbol{x_i'}$ and $\boldsymbol{\beta}$ are the vector of covariates and the vector of regression coefficients respectively. *h(.)* is the link function that specifies the relation between the linear predictor and the response variable.

Under Poisson assumption, distribution of $N_i$ is defined as,

$$N_i \sim Poi(\lambda(x_i)\, v_i) \tag{2}$$

where, $\lambda(x_i)$ is the regression function and $v_i$ is the exposure for policy *i*.

### *2.2. Zero-inflated and hurdle models*

Zero-inflated and hurdle models are used to model datasets that have excess zeros. Zero-inflated models are mixed models that consist of a point mass at zero and a positive count distribution. Hurdle models are the combinations of left-truncated count and right-censored hurdle components [15].

Let the probability of observing zero and the probability density function of counts are denoted by $\pi$ and $f_2(n)$ respectively. Probability function of zero-inflated distribution is defined as,

$$P(N=j) = \begin{cases} \pi + (1-\pi)f_2(0), & j=0 \\ (1-\pi)f_2(j), & j>0 \end{cases} \tag{3}$$

In hurdle model, let the probability of observing zero is denoted by $f_1(0)$. Probability of observing $j$ claims with hurdle model is defined as follows [16],

$$P(N=j) = \begin{cases} f_1(0), & j=0 \\ \frac{1-f_1(0)}{1-f_2(0)} f_2(j), & j>0 \end{cases} \tag{4}$$

### *2.3. Regression trees*

Regression trees are the popular non-parametric, simple and flexible methods for regression tasks. The objective is to construct trees in a way that feature space is partitioned into homogenous subsets. For the partitioning, binary tree growing algorithm can be used. The algorithm is repeated until a stopping rule is applied. For the goodness of binary splits, optimal split is chosen such that deviance loss is minimized. Regression trees are the fundamental methods for other ensemble algorithms that rely on iterative process such as boosting.

Let's assume that there are *p* explanatory variables and we will split data into *M* regions $(R_1, R_2, \dots, R_M)$. In each region, claim frequency parameter is $\lambda_m$. So, expected frequency is,

$$\hat{\lambda}(x) = \sum_{m=1}^{M} \hat{\lambda}_m\, I(x \epsilon R_m) \tag{5}$$

If we minimize the sum of squares of the difference between the responses and predictions, $\hat{\lambda}_m$ is the average of response in region $R_m$ [17]. For more details about classification and regression trees we refer [1].

### *2.4. Boosting*

In boosting method, forward stage-wise algorithm of Friedman [18] is applied to solve optimization problem by fitting weak learners and adding it to the previous fitted terms sequentially. We try to find optimal parameters by adaptively minimizing loss function in each iteration adding a weak learner to the present predictor. This algorithm is called gradient boosting machine. In our case, weak learners are the regression trees and the objective is to minimize Poisson deviance.

Assume that $L(.)$ is the objective function and we try to minimize in-sample loss over class of functions $f$ given as,

$$\hat{f} = \underset{f}{argmin} \frac{1}{n} \sum_{i=1}^{n} L(N_i, f(\boldsymbol{x_i}), v_i) \quad (6)$$

Let $\hat{f}_{m-1}(.)$ be the minimizer of Eq. (6) and $\hat{g}_m(\boldsymbol{x})$ is a regression model that acts as base learner in the algorithm. If we define working weights,

$$w_i^{(m)} = v_i \, e^{\hat{f}_{m-1}(\boldsymbol{x_i})} \quad (7)$$

and $f$ functions as the logged frequency ($log\lambda$), the steps of Poisson regression tree boosting machine with logarithmic link function are given as follows,

1. Choose a constant shrinkage parameter, $\alpha \in (0,1]$. This parameter makes weak learner even weaker.
2. Calculate $\hat{f}_0(\boldsymbol{x}) = \log \hat{\lambda}_0(\boldsymbol{x}) = \log\left(\frac{\sum_{i=1}^{n} N_i}{\sum_{i=1}^{n} v_i}\right)$
3. For *m=1,2,..,M* repeat,
    a) Calculate working weights, $w_i^{(m)}$
    b) Fit a Poisson regression tree to working data (learning dataset)
    c) Update: $\hat{f}_m(\boldsymbol{x}) = \hat{f}_{m-1}(\boldsymbol{x}) + \alpha \hat{g}_m(\boldsymbol{x})$
4. Obtain the estimator, $\hat{f}(\boldsymbol{x}) = \hat{f}_M(\boldsymbol{x})$ [1]

*2.5. GLM-Boost*

In this method, optimal GLM estimates are used as initial values ($\hat{f}_0(\boldsymbol{x})$) that is set into the exposure of the boosted regression trees and then boosting algorithm is processed iteratively [4].

## 3. Case Study

In this study, objective function is the average Poisson deviance loss function that is defined as follows,

$$L(D, \hat{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} 2N_i \left[\frac{\lambda(\boldsymbol{x_i})\, v_i}{N_i} - 1 - \log\left(\frac{\lambda(\boldsymbol{x_i}) v_i}{N_i}\right)\right] \geq 0 \quad (8)$$

where *D* is training data set. We try to minimize this loss function that provides maximum likelihood estimate of $\lambda$ [4]. For test dataset, we put *T* instead of *D* in Eq.(8).

*3.1. Data set*

Dataset consists of information related to motor third party liability insurance, started/renewed in years 2009-2012. Dataset is taken from a private insurance company in Turkey. Only the policies of private automobiles are taken into account. Response variable is the number of claims in one policy year.

Risk factors do not change during the policy period. For each policy, we have information about policy number, novation number, rider number, rider type, policy year, province where vehicle is registered, no claim discount (ncd) code, age of vehicle, age and gender of policyholder, cubic capacity of vehicle, previous claim number of policy and binary code that indicates whether policy has just started (new) or not (old). Exposure is calculated using policy number and rider numbers in one accounting year. One policy number belongs to only one year, repeated information of the same policy in unknown since policy number changes in each renewal process. Any deductible modification is

not applied in these policies. We made an exposure based calculation of correction factors for each accident quarter, using the Bornhuetter-Ferguson and Cape-Cod methods and multiplied those factors with the observed claim numbers to prevent distortion due to the reporting process.

Provinces are clustered into seven regions where clusters are generated using *k*-means clustering algorithm based on the 2010 year accident statistics of accident numbers, death numbers and injured party numbers for each province. Accident statistics are taken from the web site of Turkish Statistical Institute.

In preliminary analysis of data, we observed some values for age of driver, age of vehicle and previous claim number were pointless. So, we assumed upper limit of 90 for the age of driver, 50 for the age of vehicle and 5 for the previous claim number. All analyses of case study are performed using R Studio software, version 4.3.0.

There are 1,246,990 automobile insurance policies between the years 2009-2012. Explanatory variables are given below,

- Policy year (*year*)
- Previous claim number (*prev_cnum*), number of claim that occurred in previous policy year
- Age of driver (*age*),
- Age of vehicle (*ageveh*)
- Cubic capacity of vehicle (*cc*)
- No claim discount (*ncd*) level of policy, categorical variable with 7 levels where 4 is entrance level, 7 is highest discount and 1 is highest over premium level.
- Region, cluster of provinces where the vehicle is registered. Categorical with 7 levels
- Gender of policyholder, binary variable
- New_old, binary variable that indicates whether the policy is new or renewed (old). (new:0, old:1)

Statistics of continuous features, exposure and claim numbers are given in Table 1,

Table 1. Summary statistics of continuous features, exposure and claim numbers

| | Age of policyholder | Age of vehicle | Cubic capacity | Year | Prev_cnum | Exposure | Claim number |
|---|---|---|---|---|---|---|---|
| Min | 18 | 0 | 0 | 2009 | 0 | 0.0027 | 0 |
| Median | 42 | 13 | 1600 | 2011 | 0 | 1 | 0 |
| Mean | 43.31 | 13.13 | 1555 | 2011 | 0.0264 | 0.8317 | 0.0661 |
| Max | 90 | 50 | 7000 | 2012 | 5 | 1 | 8 |

Data is highly zero inflated since the median value of claim number is zero. Mean exposure value is close to one that shows most of the policies are in force for the whole policy year. Histogram of claim numbers is given in Figure 1. In accordance with Table 1, we see that data has a highly zero-inflated structure.

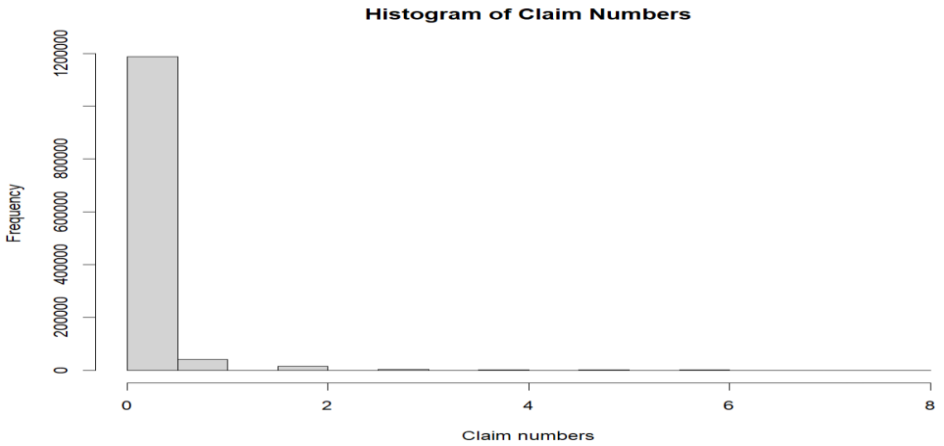


Figure 1. Histogram of claim numbers

Number of policies in each level of categorical variables are given in Table 2,

Table 2. Policy numbers in levels of categorical features

| Gender | | | | | | |
|---|---|---|---|---|---|---|
| Female | | | | Male | | |
| 172,264 | | | | 1,074,726 | | |
| New_old | | | | | | |
| New | | | | Old | | |
| 931,168 | | | | 315,822 | | |
| Ncd | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1,257 | 7,499 | 23,933 | 452,874 | 214,223 | 157,528 | 389,676 |
| Region | | | | | | |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 165,754 | 258,567 | 93,757 | 261,859 | 189,234 | 84,586 | 193,233 |

Most of the policy holders are male, are new in the system, have entrance ncd level (4) and from Region 3. Claim frequency of general portfolio is 0.0795 that is an indication of zero-inflation.

Dummy coding is applied for categorical variables and chose reference level that has the biggest volume. To test the collinearity between the variables we calculated Pearson's correlation coefficients for continuous features and Cramer's V measures for categorical variables are given in Table 3 and Table 4.

Table 3. Pearson's correlation coefficients

| | Year | Prev_cnum | Age | Ageveh | Cc |
|---|---|---|---|---|---|
| Year | 1 | 0 | 0.08 | 0.03 | 0.02 |
| Prev_cnum | 0 | 1 | 0 | -0.01 | 0 |
| Age | 0.08 | 0 | 1 | 0.11 | -0.02 |

| | | | | | |
|---|---|---|---|---|---|
| Ageveh | 0.03 | -0.01 | 0.11 | 1 | -0.03 |
| Cc | 0.02 | 0 | -0.02 | -0.03 | 1 |

Table 4. Cramer's V measures

| | Region | Ncd | Gender | New_old |
|---|---|---|---|---|
| Region | 1 | 0.0397 | 0.0793 | 0.0603 |
| Ncd | 0.0397 | 1 | 0.0215 | 0.4270 |
| Gender | 0.0793 | 0.0215 | 1 | 0.0046 |
| New_old | 0.0603 | 0.4270 | 0.0046 | 1 |

Based on correlation coefficients, there is a high correlation between ncd level and new_old status of policies and a slight correlation between the age of driver and the age of vehicle. For now, we will use all the features in the models.

### *3.2. Analysis*

We fit Poisson GLM by assuming the age of driver, age of vehicle and the cubic capacity both continuous and categorical while other features are same in the models. We compared results based on AIC and the significance of the variables. In each case, variables are significant at 95% significance level but AIC is smaller when the variables are categorical. As a result, we used these three variables categorical in regression models. Categories are binned as follows:

Cubic capacity, with 3 levels [0-1300], (1300-1600], (1600-7000]

Age of driver, with 6 levels [18,25], (25,30], (30,40], (40,50], (50,70], (70,90]

Age of vehicle, with 5 levels [0,1], (1,5], (5,10], (10,20], (20,50]

We applied dummy coding for categorical variables. Design matrix has full rank under dummy coding and this means linearly independence of columns [19]. To compare the predictive performance of candidate models, dataset is partitioned into two parts, training dataset (80%) and the test dataset (20%). In-sample loss is calculated on training data set while out-of-sample loss is calculated on test data set using Eq. (8). Reason of using deviance loss function is that this loss function evaluated on a different test dataset provides a good predictive performance indicator [4].

First model is GLM under Poisson distribution with logarithmic link function. We used all explanatory variables in Poisson GLM only taking into account main effects. We call this model as *GLM1*. According to the model results, new_old was not statistically significant with p-value 0.5261. Analysis of deviance table is given in Table 5,

Table 5. Anova results of GLM1

| | Df | Deviance | Resid. Df | Resid. Dev |
|---|---|---|---|---|
| Null | | | 997591 | 407221 |
| ageglm | 5 | 1824 | 997586 | 405396 |
| agevehglm | 4 | 1101 | 997582 | 404295 |
| cc | 2 | 188 | 997580 | 404106 |
| gender | 1 | 86 | 997579 | 404020 |
| region | 6 | 2425 | 997573 | 401594 |
| ncd | 6 | 4370 | 997567 | 397223 |
| new_old | 1 | 0 | 997566 | 397223 |
| prev_cnum | 1 | 197 | 997565 | 397026 |
| year | 1 | 7 | 997564 | 397019 |

A decrease in residual deviance from a simple model to a complex model indicates that the additional parameters provide a better fit. Based on deviance results, new_old variable is excluded from the model since there is no difference in residual deviance after new_old variable is added to the model. Summary of refitted model without new_old (*GLM2*) is given in Table 6,

Table 6. Results of GLM2

| Parameter | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | 17.5549 | 7.5406 | 2.3280 | 0.0199 |
| ageglm1 | 0.3402 | 0.0173 | 19.6990 | < 2e-16 |
| ageglm2 | 0.1599 | 0.0128 | 12.4940 | < 2e-16 |
| ageglm4 | 0.0259 | 0.0105 | 2.4660 | 0.0136 |
| ageglm5 | -0.0608 | 0.0111 | -5.4840 | 0.0000 |
| ageglm6 | -0.0785 | 0.0297 | -2.6410 | 0.0083 |
| agevehglm1 | 0.1418 | 0.0165 | 8.5850 | < 2e-16 |
| agevehglm2 | 0.1281 | 0.0113 | 11.2870 | < 2e-16 |
| agevehglm3 | 0.0516 | 0.0108 | 4.7890 | 0.0000 |
| agevehglm5 | -0.0436 | 0.0126 | -3.4590 | 0.0005 |
| cc1 | -0.1138 | 0.0150 | -7.5660 | 0.0000 |
| cc2 | -0.0300 | 0.0120 | -2.4970 | 0.0125 |
| gendermale | -0.0738 | 0.0108 | -6.8610 | 0.0000 |
| region1 | -0.5689 | 0.0133 | -42.8340 | < 2e-16 |
| region2 | -0.7020 | 0.0196 | -35.7280 | < 2e-16 |
| region3 | -0.3588 | 0.0125 | -28.7530 | < 2e-16 |
| region4 | -0.3691 | 0.0136 | -27.1010 | < 2e-16 |
| region5 | -0.1587 | 0.0162 | -9.7980 | < 2e-16 |
| region6 | -0.4954 | 0.0139 | -35.6550 | < 2e-16 |
| ncd1 | 0.2658 | 0.0816 | 3.2560 | 0.0011 |
| ncd2 | -0.1077 | 0.0443 | -2.4290 | 0.0151 |
| ncd3 | -0.0877 | 0.0259 | -3.3840 | 0.0007 |
| ncd5 | -0.2886 | 0.0109 | -26.4110 | < 2e-16 |
| ncd6 | -0.4449 | 0.0130 | -34.2660 | < 2e-16 |
| ncd7 | -0.6396 | 0.0106 | -60.1430 | < 2e-16 |
| prev_cnum | 0.3003 | 0.0203 | 14.7890 | < 2e-16 |
| year | -0.0096 | 0.0038 | -2.5720 | 0.0101 |

According to GLM results, all features are statistically significant at 95% confidence level. Mean claim frequency is decreasing after driver's age 50, vehicle's age 20 and at lower cubic capacity when compared to the base level. Claim frequency is higher in female drivers than males and in region 0. While previous claim number increases the claim frequency, policy year has a decreasing effect on it. Comparison of two GLMs based on loss values and AICs is given in Table 7.

Table 7. In-sample losses, out-of-sample losses and AIC values of GLMs

| Model | Average in-sample loss | Average out-of-sample loss | AIC |
|---|---|---|---|
| GLM1 | 0.3979 | 0.4011 | 501404 |
| GLM2 | 0.3979 | 0.4011 | 501403 |

There is not a significant difference based on in-sample and out-of-sample loss of two GLMs. According to AIC values *GLM2* performs better with a small difference. To compare other methods, we use same explanatory variables, new_old variable is excluded.

### *3.2.1. Zero-inflated and hurdle Poisson model*

Due to the excess zero structure of claim numbers, we fit both zero-inflated Poisson (ZIP) and hurdle Poisson model to claim numbers. Logistic regression is used for the zero component of the models. First, for the ZIP model, we assumed the probability of excess zero part is independent of exposure and did not use exposure in logit part. Then exposure is used in both parts of the model. Average losses are lower when exposure is used only in count part. Similar with ZIP model, we used exposure only in Poisson part of hurdle model. Average in-sample and out-of-sample losses of models are given in Table 8,

Table 8. Average in-sample and out-of-sample losses of zero-modified models

| Model | Average in-sample loss | Average out-of-sample loss |
|---|---|---|
| ZIP | 0.3980 | 0.4010 |
| Hurdle Poisson | 0.4047 | 0.4074 |

Based on average losses, ZIP model shows better predictive performance than the hurdle model. Hurdle model may not capture the complex structure of excess zeros as effective as the zero-inflated model since zero-inflated models incorporate a separate process for zeros. Average portfolio frequencies of predicted claim numbers from ZIP is 0.0794 and hurdle model is 0.0761. In accordance with prediction performances, hurdle model underestimates average portfolio frequency.

### *3.2.2. Regression trees*

We did not make any feature pre-process for regression trees but used age of driver, age of vehicle and the cubic capacity as continuous variables. 5000 policies are assumed at each leaf of the tree. We also fitted tree with 10000 policies at each leaf but deviances were higher. According to the analysis results, optimal cost complexity (cp) parameter that controls the size of the tree is 0.01. We call this regression tree with cp 0.01 as *RT1*. To be an alternative, we also fitted tree with cp=0.00001 value (*RT2*) since smaller cp value provides a larger tree. Average in-sample and out-of-sample loss values are given in Table 9,

Table 9. Average in-sample and out-of-sample losses of regression trees

| Model | Average in-sample loss | Average out-of-sample loss |
|---|---|---|
| RT1 (cp=0.01) | 0.4036 | 0.4061 |
| RT2 (cp=0.00001) | 0.3971 | 0.4009 |

According to the model results, when cp is smaller (0.00001), tree is constructed on all explanatory variables except previous claim number. When cp is 0.01, the tree is constructed on only ncd level of policy. We can conclude that regression trees perform better predictive performance when they have larger cp value that provides larger tree and more features.

### *3.2.3. Boosting and GLM-Boost methods*

We used regression trees as weak learners for boosting. Similar to regression trees, we assumed cp parameter 0.0001, 50 iterations for the construction of weak learners and 5000 policies at each leaf of the tree. We did not apply shrinkage in the boosting algorithm. We choose depth of the tree as J=2 (*Boost2*) and J=3 (*Boost3*), that shows the number of levels from the root node to a leaf node. Average loss values of both boosting methods are given in Table 10,

Table 10. Average in-sample and out-of-sample losses of boosting algorithms

| Model | Average in-sample loss | Average out-of-sample loss |
|---|---|---|
| Boost2 (J=2) | 0.3966 | 0.4001 |
| Boost3 (J=3) | 0.3956 | 0.4000 |

We can say that predictive performance of boosting algorithm increases when the depth of the tree is higher.

In *GLM-Boost* methods, estimates of *GLM2* model are used as initial values that is set into the exposure of the boosted regression trees. Under the same parameters with *Boost3*, at the end of 50th iteration, average in-sample loss is **0.3955** and out-of-sample loss is **0.3999** that are the smallest loss values among all candidate models. These results support the combination of GLM and boosting approach that iteratively improves the predictive performance of the model is a good alternative for the prediction purposes. In this approach, parameter estimates of best performing GLM are set into the exposure of the regression trees and boosting algorithm works. Combining optimal initial values from GLM and a strong boosting algorithm improves the predictions.

Since average out-of-sample loss is calculated using test data, it is the main indicator for the prediction when compared with in-sample loss. To sum up, average out-of-sample loss of Poisson GLM, ZIP, regression trees and boosting methods are given in Table 11. Although data has a zero-inflated structure, machine learning methods based on regression trees, perform better predictive performance than zero-inflated model. Specially, GLM-Boost approach that combined GLM and boosting is a promising point for the future studies that combine traditional and machine learning methods.

Table 11. Average out-of-sample losses of candidate methods

| Model | Average out-of-sample loss |
|---|---|
| GLM2 | 0.4011 |
| ZIP | 0.4010 |
| RT2 | 0.4009 |
| Boost3 | 0.4000 |
| GLM-Boost | 0.3999 |

To see the similarity between the observed claim numbers and predicted (test data) ones, we give statistics in Table 12.

Table 12. Statistics of observed and predicted claim numbers

| Model | Min | Median | Mean | Max |
|---|---|---|---|---|
| Observed | 0.00000 | 0.00000 | 0.06660 | 8.00000 |
| GLM2 | 0.00009 | 0.06164 | 0.06603 | 0.42789 |
| ZIP | 0.00009 | 0.06199 | 0.06611 | 0.35541 |
| RT2 | 0.00009 | 0.06257 | 0.06604 | 0.20690 |

| | | | | |
|---|---|---|---|---|
| Boost3 | 0.00008 | 0.06062 | 0.06601 | 0.45123 |
| GLM-Boost | 0.00007 | 0.06064 | 0.06605 | 0.43006 |

Because of zero-inflated structure of the data set, predictions have distribution around zero. Predictions of each model have mean value close to the mean of observed claim numbers. In general, we can say that the predictions of different models have similar values and that supports the small differences between the average loss values of the models.

## 4. Conclusion

In this study, we compared the predictive performances of GLM, zero-modified models, regression trees and two boosting approaches using a zero-inflated claim frequency data. Main result of this study is, boosted regression trees and GLM-Boost performed better than both zero-inflated Poisson and hurdle Poisson model in addition to Poisson GLM based on average in-sample and average out-of-sample losses. These results are in accordance with the studies in the literature in that tree-based ML methods (specially boosting) show better predictive performance for claim data when compared with GLM [4, 5]. Although GLMs are easy to implement and interpret, steps such as feature pre-processing, feature selection, interactions affect the predictive performance of the model. Instead, ML methods provide flexibility for these processes and handle interactions and non-linearities automatically. When the predictive accuracy is the subject, ensemble methods provide a good choice since they combine predictions from learners that reduce overfitting and variance. As an ensemble method, boosting strengthens the predictions by sequentially fitting weak learners that are regression trees in this study. GLM-Boost that showed best predictive performance is applied by boosting GLM with regression trees. This approach is a simple example for the combination of a classical regression model and a machine learning method. This type of combinations look bright for the predictive purposes of actuarial data. As a future work, combination of neural networks with a classical regression model can be used to predict claim data [20]. Also, approaches that combine zero-inflated models with boosting can be used to predict imbalanced claim frequency data [21].

A statistical method that shows good predictive performance on a data, may not show the same performance on a different dataset. Additively, average loss values between the predictive models have small differences due to the zero-inflated structure of the data in this study. So, it may be a good idea to perform related models on different datasets that don't have imbalanced distribution and to predict different response variables such as claim amount or probability of a claim.

## Acknowledgements

I would like to thank Ronald Richman (actuary) for all his support on the analysis and other theoretical questions.

## References

[1] M. V. Wuthrich and C. Buser, "Data analytics for non-life insurance pricing," *Swiss Finance Institute Research Paper*, no. 16-68, 2023.
[2] Y. Liu, B.-J. Wang, and S.-G. Lv, "Using multi-class AdaBoost tree for prediction frequency of auto insurance," *Journal of Applied Finance and Banking*, vol. 4, no. 5, p. 45, 2014.
[3] A. Dal Pozzolo, G. Moro, G. Bontempi, and D. Y. A. Le Borgne, "Comparison of data mining techniques for insurance claim prediction," *Universita degli Studi di Bologna*, 2011.
[4] A. Noll, R. Salzmann, and M. V. Wuthrich, "Case study: French motor third-party liability claims," *Available at SSRN 3164764*, 2020.
[5] C. Clemente, G. R. Guerreiro, and J. M. Bravo, "Modelling Motor Insurance Claim Frequency and Severity Using Gradient Boosting," *Risks*, vol. 11, no. 9, p. 163, 2023.
[6] A. Ferrario and R. Hämmerli, "On boosting: Theory and applications," *Available at SSRN 3402687*, 2019.
[7] L. Guelman, "Gradient boosting trees for auto insurance loss cost modeling and prediction," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3659-3667, 2012.

[8] R. Henckaerts, M.-P. Côté, K. Antonio, and R. Verbelen, "Boosting insights in insurance tariff plans with tree-based machine learning methods," *North American Actuarial Journal*, vol. 25, no. 2, pp. 255-285, 2021.
[9] B. So, "Enhanced gradient boosting for zero-inflated insurance claims and comparative analysis of CatBoost, XGBoost, and LightGBM," *Scandinavian Actuarial Journal*, pp. 1-23, 2024.
[10] P. Zhang, D. Pitt, and X. Wu, "A new multivariate zero-inflated hurdle model with applications in automobile insurance," *ASTIN Bulletin: The Journal of the IAA*, vol. 52, no. 2, pp. 393-416, 2022.
[11] I. N. El-Saeiti and G. Alomair, "Comparative Evaluation of Zero-Inflated and Hurdle Models for Balanced and Unbalanced Data: Performance Assessment and Model Fit Analysis," *European Journal of Science, Innovation and Technology*, vol. 3, no. 6, pp. 192-199, 2023.
[12] K. C. Yip and K. K. Yau, "On modeling claim frequency data in general insurance with extra zeros," *Insurance: Mathematics and Economics*, vol. 36, no. 2, pp. 153-163, 2005.
[13] J.-P. Boucher, M. Denuit, and M. Guillén, "Risk classification for claim counts: a comparative analysis of various zeroinflated mixed Poisson and hurdle models," *North American Actuarial Journal*, vol. 11, no. 4, pp. 110-131, 2007.
[14] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 135, no. 3, pp. 370-384, 1972.
[15] J. Mullahy, "Specification and testing of some modified count data models," *Journal of econometrics*, vol. 33, no. 3, pp. 341-365, 1986.
[16] A. Zeileis, C. Kleiber, and S. Jackman, "Regression models for count data in R," *Journal of statistical software*, vol. 27, no. 8, pp. 1-25, 2008.
[17] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, "The elements of statistical learning: data mining, inference, and prediction". *Springer*, 2009.
[18] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189-1232, 2001.
[19] M. V. Wüthrich and M. Merz, "Statistical foundations of actuarial learning and its applications". *Springer* Nature, 2023.
[20] M. V. Wüthrich and M. Merz, "Yes, we CANN!," *ASTIN Bulletin: The Journal of the IAA*, vol. 49, no. 1, pp. 1-3, 2019.
[21] S. C. Lee, "Addressing imbalanced insurance data through zero-inflated Poisson regression with boosting," *ASTIN Bulletin: The Journal of the IAA*, vol. 51, no. 1, pp. 27-55, 2021.