Prediction of Claim Probability with Excess Zeros

Aslıhan Şentürk Acar

Abstract Non-life insurance pricing is based on two components: claim severity and claim frequency. These components are used to estimate expected pure premium for the next policy period. Generalized linear models (GLM) are widely preferred for the estimation of claim frequency and claim severity due to the ease of interpretation and implementation. Since GLMs have some restrictions such as exponential family distribution assumption, more flexible Machine Learning (ML) methods are applied to insurance data in recent years. ML methods use learning algorithms to establish relationship between the response and the predictor variables as an intersection of computer science and statistics. Because of some insurance policy modifications such as deductible and no claim discount system, excess zeros are usually observed in claim frequency data. In the presence of excess zeros, prediction of claim probability can be a good alternative to the prediction of claim numbers since positive numbers are rarely observed in the portfolio. Excess zeros create imbalance problem in the data. When the data is highly imbalanced, predictions will be biased toward majority class due to the priors and predicted probabilities may be uncalibrated. In this study, we are interested in claim occurrence probability in the presence of excess zeros. A Turkish motor insurance dataset that is highly imbalanced is used for the case study. Ensemble methods that are popular ML approaches are used for the probability prediction as an alternative to logistic regression. Calibration methods are applied to predicted probabilities and results are compared.

Keywords Claim probability · Imbalanced data · Non-life insurance · Machine learning

A. Ş. Acar (✉)
Department of Actuarial Sciences, Hacettepe University, Ankara, Turkey
e-mail: aslihans@hacettepe.edu.tr

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,
https://doi.org/10.1007/978-3-030-85254-2_32

1 Introduction

Insurance companies guarantee to compensate policyholder against unpredictable losses during a certain time period by charging premium for the assurance. Basic objective of ratemaking is to determine fair premiums for the policyholders that have different characteristics. For this purpose, actuaries use statistical models and rating factors to determine premiums. Statistical approach depends on the observed data in a certain accounting year. Observed responses of each policy can be only aggregate losses, both total claim numbers and aggregate losses or detailed information for each claim event (Frees et al. 2014). According to actuarial equivalence principle, pure premium is equal to expected total claim size that depends on two components: expected claim frequency and expected claim severity. When observed responses are only aggregate losses in a policy period, we need two components to estimate pure premium: estimates of total claim size and claim probability.

Non-life insurance data have some characteristics such as excess of zeros in claim numbers due to NCD system and deductible modification. Excess of zeros leads to imbalanced data structure. In such a case, predictive models perform poorly because of few instances of minority class and they classify most of the observations as majority class. When the data is highly imbalanced, predictions will be biased toward majority class due to the priors (Kuhn and Johnson 2013). Since we are interested in claim (occurrence) probability that constitutes minority class, we have to deal with imbalance structure of data and the bias to get accurate predictions. To deal with imbalanced data, a common approach is to apply resampling methods (He and Garcia 2009; Japkowicz and Stephen 2002). Various resampling methods are used to rebalance the data such as oversampling, undersampling, and synthetic minority oversampling technique. In addition to resampling methods, feature selection, cost sensitive learning, and hybrid ensemble learning methods are other alternatives to deal with class imbalance (Guo et al. 2008).

From the insurance perspective, zero-inflated and hurdle models are used to deal with excess of zeros (Yip and Yau 2005; Boucher et al. 2007). These models include a component for structural zeros that has to be estimated using generally logistic regression (LR). Although LR is easy to understand and implement, it is constrained to a specified form that creates consistency problems when model is not correctly specified. As an alternative to LR, flexible nonparametric machine learning algorithms are used in many studies in recent years. Although LR can directly predict calibrated probabilities due to the optimization of log loss, some ML methods don't produce calibrated probabilities specially for imbalanced data and they need calibration (Fernández et al. 2018). Uncalibrated probabilities may induce bias in probability scores. If probabilities are calibrated they will represent the likelihood of true classes. Platt scaling (Platt 1999) and isotonic regression (Zadrozny and Elkan 2001) are two popular calibration methods in the literature (Niculescu-Mizil et al. 2012).

In this study, we are interested in claim occurrence probability in the presence of excess zeros. To the best of our knowledge there are very few studies in actuarial literature that use ML methods for the prediction of claim probability. Dal Pozzolo

(Pozzolo 2010) used decision trees, random forest, Naïve Bayes, K-nearest neighbors, neural networks (NN), support vector machine (SVM), and linear discriminant analysis to classify claims whether they are greater than zero or not using claim probability estimates based on different thresholds. Frempong et al. (Frempong et al. 2017) used decision trees to predict probability of making a claim. Pijl (Tim Pijl 2017) used decision trees, random forest (RF), LR, and SVM to predict probability of issuing a claim.

We aim to compare predictive performances of LR and ensemble methods to predict claim probability in the presence of excess zeros. Calibration methods are applied to predicted probabilities and the results are compared using Brier score (BS) (Glenn 1950).

2 Methods

2.1 Logistic Regression

Response variable is binary ($Y = 0$ or $Y = 1$) in classification tasks. Assuming $p = P(Y = 1)$, logistic regression equation is expressed as

$$\log it(p) = \log\left(\frac{p}{1-p}\right) = x'\beta \tag{1}$$

where x is the vector of predictors and β is the vector of regression parameters. Predictions of LR are probabilities of binary event.

2.2 Bagging

Ensemble methods are designed to improve the predictive performance of decision trees. These methods compile the information related to the predictions of base models. They have no distributional restrictions and they handle interactions between variables easily. Bagging (bootstrap aggregating) is the simplest ensemble method in that bootstrap samples are drawn randomly from the study sample with replacement to reduce the variance. Different training samples are created using bootstrap, generally decision trees are chosen as base learning algorithm and the predictions are averaged over bootstrap samples (Breiman 1996). Predicted probability of binary event for a unit is obtained as the ratio of units that have the event among all units in related subset (Austin et al. 2013).

2.3 Random Forest

Random forest approach (Breiman 2001) also uses bootstrap samples similar as bagging but considers binary splits of tree on a random sample of predictor variables instead of all candidate predictor variables to decorrelate the trees and increase the accuracy (Austin et al. 2013; James et al. 2013). When the number of randomly selected predictors is equal to the total number of predictors, random forest algorithm reduces to the bagging algorithm.

2.4 Boosting

Boosting works in similar way with bagging but with boosting method each tree uses the information from previous tree and trees are grown sequentially by applying weak learner to the reweighted data (James et al. 2013). Objective is to reduce the error of a weak learner and get a strong predictor (Freund and Schapire 1996). There are several boosting algorithms in the literature but generalized boosted model (GBM) (Friedman 2001) is used in this study.

3 Case Study

3.1 About Dataset

Case study is implemented using motor insurance dataset from an insurance company in Turkey. There are 376,719 individual automobile policies in the portfolio. All policies are started or renewed in 2010 and each policy has 1 year of exposure. Very few policies had more than one claim during the policy period. Frequency of claim numbers is given in Table 1.

There are only four individuals that have reported four claims and %0.15 of policies had more than one claim during 1-year policy period. Data is highly imbalanced since %95.43 of policyholders did not report any claim during the policy year. Therefore, we prefer to model claim probability instead of claim numbers. Predictor

Table 1 Frequency of claim numbers

Number of claims	Number of records
0	359,487
1	16,660
2	540
3	28
4	4

variables are age of policyholder (18–90), gender of policyholder (0:female, 1:male), province in which number plate of vehicle is registered, age of vehicle (0–64), and horse power of vehicle. Based on frequency information, we cut age of policyholder at 90. Provinces are clustered into seven clusters using 2010 year accident statistics that are published by Turkish Statistical Institute. Statistics related to the continuous predictors are given in Table 2. Frequencies related to categorical variables are given in Table 3.

As can be seen from Table 3, most of the policyholders are the males.

Table 2 Statistics of continuous predictor variables

Claim (1) No Claim (0)	Min	1st Qu	Median	Mean	3rd Qu	Max
Age of policyholder						
0	18	32	40	42.13	50	90
1	18	31	39	40.74	49	90
Age of vehicle						
0	0	7	13	13.23	18	88
1	0	6	12	12.57	17	62
Horse power						
0	20	75	80	88.69	100	1001
1	26	75	80	89.86	100	445

Table 3 Frequencies of categorical variables (percentage)

Province	No claim	Claim
0	42,845 (%11)	3030 (%0.8)
1	71,641 (%19)	2897 (%0.8)
2	26,795 (%7)	965 (%0.3)
3	76,647 (%20)	3628 (%1)
4	53,139 (%14)	2550 (%0.7)
5	31,622 (%8)	1860 (%0.5)
6	56,798 (%15)	2302 (%0.6)
Gender	No claim	Claim
Female	48,747 (%13)	2761 (%1)
Male	310,740 (%82)	14,471 (%4)

Table 4 LR estimation results

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
intercept	−2.153	0.054	−40.196	< 2e-16
Male	−0.141	0.024	−5.868	4.41e-09
Age	−0.008	0.003	−17.204	< 2e-16
province1	−0.524	0.030	−17.012	< 2e-16
province2	−0.698	0.043	−16.252	< 2e-16
province3	−0.357	0.029	−12.474	< 2e-16
province4	−0.364	0.031	−11.678	< 2e-16
province5	−0.181	0.034	−5.276	1.32e-07
province6	−0.520	0.032	−16.254	< 2e-16
vage	−0.007	0.001	−5.888	3.92e-09
horse power	0.0004	0.0003	1.334	0.182

3.2 Analysis

Data is randomly partitioned into two parts: training dataset (%80) and test dataset (%20). Models are fitted to training data and validated using test data. No interaction and nonlinear terms are used in LR. Parameter estimates of LR are given in Table 4. From Table 4, we can say that all predictor variables except horse power are statistically significant at %95 confidence level.

For the ease of computation, fivefold cross-validation (CV) method is used for tuning the hyperparameters of ML methods. There is no hyperparameter for bagging algorithm. A tree is constructed for each of drawn 50 bootstrap samples. Random forest algorithm has three hyperparameters: the number of predictors selected at each split, split rule, and minimal node size (Wright et al. xxxx). Finally, generalized boosted regression model has four hyperparameters: number of trees, maximum depth of trees, learning rate, and minimum observation number in terminal nodes.

Statistics related to the predicted probabilities of policyholders in test dataset are given in Table 5. Interval of probability predictions of LR and GBM is too narrow that reflect the imbalance structure of the dataset. We can easily say that predictions

Table 5 Statistics related to the predicted claim probabilities

Statistics	LR	Bagging	RF	GBM
Min	0.0172	0.0000	0.0015	0.0304
1st Qu	0.0380	0.0000	0.0269	0.0414
Median	0.0424	0.0000	0.0383	0.0448
Mean	0.0456	0.0326	0.0466	0.0456
3rd Qu	0.0510	0.0000	0.0573	0.0490
Max	0.0980	0.9500	0.4471	0.0698

Table 6 Brier score values of predictive models

Model	BS
LR	0.0439
Bagging	0.0523
RF	0.0445
GBM	0.0440

are biased toward majority class. Bagging predicts too many zeros (median is zero) compared to other methods because it is a weaker learner compared to RF and GBM.

To deal with imbalanced structure of dataset, we planned to use resampling methods to investigate effect of resampling on predictive performance. Racing algorithm (Birattari et al. 2002) of unbalanced R package (Pozzolo and Caelen xxxx) is applied to select the best resampling method. According to the result, none of the resampling algorithm is suitable for our dataset. Eventually, we did not apply any resampling scheme.

We compared predictive performance of candidate models using BS that is the mean squared loss between the predicted probabilities and actual responses. Lower BS means better predictive performance. BS values of each prediction method related to test dataset are given in Table 6.

According to Table 6, LR and GBM have best predictive performance with lowest BS values. But there is a very little difference with RF method. Bagging performs worse as expected from prediction values given in Table 5.

We used Platt scaling and isotonic regression for the calibration of probabilities predicted. In Platt scaling, a sigmoid function is used for the probabilities and gradient descent algorithm is used to find two parameters of sigmoid function (Platt 1999). Isotonic regression approach fits a nonparametric monotonic regressor. It minimizes the squared error between the true class labels and the outputs (Zadrozny and Elkan 2002). BS values after calibration methods are given in Table 7.

As seen from Table 7, both calibration methods are not effective on the predictive performances. Specially, with isotonic regression, BS values increased in bagging and GBM methods. We can conclude that calibration methods did not work for this highly imbalanced dataset.

Table 7 BS values after calibration methods

Calibration methods	LR	RF	Bagging	GBM
Platt scaling	0.04391	0.04399	0.04400	0.04393
Iso Reg	0.04422	0.04424	0.25620	0.16706

4 Conclusion

Claim probability is an important measure about the uncertainty of claim occurrence. It also constitutes one part of two-part models such as ZIP that is frequently used for claim frequency modeling. In this study, main objective is to compare predictive performance of logistic regression and ensemble methods for the prediction of claim probability in presence of excess zeros. A Turkish motor insurance dataset is used for the case study. According to case study results, predicted probabilities were biased to majority class (zero) and calibration methods did not improve predictive performance of the methods based on Brier score values. Another result is that RF and GBM performed similar predictive performance with logistic regression. Bagging method performed worst among all predictive models since it neither has random variable selection nor works sequentially like RF and GBM to increase accuracy. Consequently, ML methods are good alternatives to classical approaches for the prediction.

As a future study, more complex ML methods such as neural networks can be used for the prediction of claim probability. Another idea is the comparison of claim severity*frequency approach with the claim probability*total claim size approach for the prediction of total claim amount.

References

Frees EW, Derrig RA, Meyers G (2014) Predictive modeling applications in actuarial science. Cambridge University Press, p 565

Kuhn M, Johnson K (2013) Applied predictive modelling, vol 26. Springer

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284

Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study. Intell Data Anal 6(5):429–449

Guo X, Yin Y, Dong C, Zhou G (2008) On the class imbalance problem. IEEE Conf Publ 4:192–201

Yip KCH, Yau KKW (2005) On modeling claim frequency data in general insurance with extra zeros. Insur Math Econ 36(2):153–163

Boucher JP, Denuit M, Guillén M (2007) Risk classification for claim counts: a comparative analysis of various zeroinflated mixed poisson and hurdle models. North Am Actuar J 11(4):110–131

Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets, vol 11. Springer, Berlin

Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classif 10(3):61–74

Zadrozny B, Elkan C (2001) Obtaining calibrated probability estimates from decision trees and Naive Bayesian classifiers. In: Proceedings of the Eighteenth International Conference on Machine learning [Internet]. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp 609–616. (ICML '01). Available from: http://dl.acm.org/citation.cfm?id=645530.655658

Niculescu-Mizil A, Caruana RA (2012) Obtaining calibrated probabilities from boosting. Jul 4 [cited 2021 May 29]; Available from: https://arxiv.org/abs/1207.1403v1

Pozzolo AD (2010) Comparison of data mining techniques for insurance claim prediction [Master of Science]. University of Bologna

Frempong NK, Nicholas N, Boateng MA (2017) Decision tree as a predictive modeling tool for auto insurance claims. Int J Stat Appl 7(2):117–120
Tim P (2017) A framework to forecast insurance claims [Master of Econometrics and Management Science]. Erasmus University Rotterdam
Glenn W (1950) Brier, verification of forecasts expressed in terms of probability. Mon Weather Rev 78(1):1–3
Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
Austin PC, Tu JV, Ho JE, Levy D, Lee DS (2013) Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol 66(4):398–407
Breiman L (2001) Random forests. Mach Learn 45(1):5–32
James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning [Internet], vol 6. Springer. Available from: https://doi.org/10.1007/978-1-4614-7138-7.pdf
Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In 1996. pp 148–56
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
Marvin NW, Wager S, Probst P (2018) "ranger" package
Birattari M, Stützle T, Paquete L, Varrentrapp K (2002) A racing algorithm for configuring metaheuristics. In: Proceedings of the 4th Annual Conference on Genetic and evolutionary computation [Internet]. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp 11–18. (GECCO'02)
Pozzolo AD, Caelen O, Bontempi G (2015) Package "unbalanced."
Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. In 2002 [cited 2021 Jun 4]. Available from: https://doi.org/10.1145/775047.775151