InNova Auto Insurance Company Modeling Project Report

Zhaoliang Zhou

Due: Monday 12/14/2020 12pm

Abstract

InNova data set contains information on one-year vehicle insurance policies from 2004-2005. Our goal for this project is to develop statistical models to predict the total claim cost for each policy. For this project, we implemented many statistical methods and machine learning algorithms. For our final model, however, we selected a simpler but more practical and interpretable approach which we model total claim cost using claim indicator, severity, and frequency models.

Introduction

In a competitive insurance market, it is more important and advantageous for an insurance company to charge the policyholder a fare premium according to the policyholder's expected loss. For instance, if a vehicle insurance company charges too much for an older driver and too little for a younger driver, this would result the older driver switch to the vehicle insurance company's competitor and the young driver would have under-priced remaining policies (Yang, Qian, Zou, 2016)[4]. In order to accurately determine the premium for the customers, first it is important to predict the actual claim cost using a set of predictors such as vehicle's age, driver's age, and driver's gender etc.

One of the major difficulties of modeling claim cost is that the data is highly right skewed with a large amount of zeros. Figure 1 shows the distribution of claim cost in the training data set from the InNova data set which we used for this project. It has similar distribution of claim cost compare that from other studies. Such data cannot be transformed to normal using methods such as log transformation and etc. Thus, it is required to model the right-skewed distribution with some special treatment of those large amount of zeros(Yang, Qian, Zou, 2016)[4].

According to many studies from actuarial science, Tweedie GLM seems to be the most popular methods in dealing with insurance type data (Yang, Qian, Zou, 2016)[4]. Dunn and Smyth (2005)[2] had a detailed discussion and mathematical explanation and derivation on Tweedie family. In general, if we model the the total number of claims (N) using a Poisson distribution with parameter λ, where:

$$N \sim Poisson(\lambda),\ \lambda > 0$$

Then, if define severity to be the total claim cost divided by the number of claims made and denote severity as Z, then we can model severity using gamma distribution with parameter α and β:

$$Z \sim Gamma(\alpha, \beta),\ \alpha, \beta > 0$$

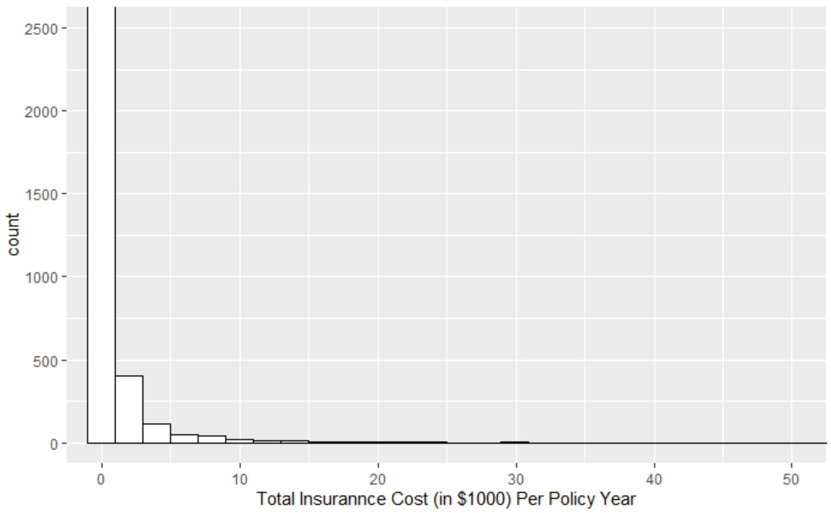

Figure 1: Histogram of claim cost from the InNova training data set. There are total 22,610 policies, and 21,076 policies have 0 claim costs (counts above 2500 are omitted for illustration purpose)

Then we can model total claim cost Y using Compound Poisson distribution $CPG(\lambda, \alpha, \beta)$:

$$Y = \sum_{i=1}^{N} Z_i$$

where Z_i are identically independent distributed $Gamma(\alpha, \beta)$ sequence independent from number of claims (N) (Quijano and Garrido, 2014)[3]. For this project, we used the Tweedie regression model as our benchmark model and further develop and improve our prediction models. For the remaining of this paper, we will introduce the data, the methods or models we considered and our final model, our final results, and discussion on limitations and future directions.

Materials and Methods

The Data

The analysis data set, InNova data set was provided by the Travelers Companies, which is based on one-year vehicle insurance policies from 2004 to 2005. There are total 45,239 policies included in the data set, and at least 6.8% of them has at least one claim. The data was split into test set and training set. The training set has all the variables, and the testing set has all the 3 responses variables omitted for which we need to make predictions on. There are total 3 response variables that we might potentially need in order to predict the claim cost. Those are claim indicator (whether the policy has a claim or not), claim counts which is the total number of claims made, and claim cost which is the total cost of all the claims made. For potential explanatory variables, the data set included the market value of the vehicle, vehicle types, the age of the vehicle, the gender of the driver, driving area of residence, driver's age which categorized into 6 categories from young to old, exposure which is the basic unit of risk underlying an insurance premium.

For our analysis, motivated by Quijano and Garrido (2014)[3] and based on the given variables, we created additional response variables: pure premium which is the claim cost divided by exposure, severity (average cost) which is the total claim cost divided total number of claims, and frequency which is number of claims divided by exposure.

Methods

We first began with Tweedie GLM to model the pure premium, and since pure premium is the claim cost divided by exposure, we set the weight parameter to be "exposure" so we could model claim cost directly. We used 10 folds cross-validation to find the best set of Tweedie parameters (α and power) that would result in the highest Gini score. For this model. For this mode, we had a Gini score of 0.09082 on the testing data set.

In addition, we tried stacking model approach which is common in machine learning for improving prediction accuracy. The general idea of model stacking is first train various machine learning algorithms on the first level training data (original training data) and make the predictions. Each of the models would produce predictions of the outcomes. Then, we could use those predictions as second level training data. Then, we would train our second level models on the second level training data set in order to produce the final prediction. For our stacking models, we first applied Box-Cox transformation on the response variables pure premium, severity, and frequency. For our level one and level two models we included following algorithms: Tweedie regression, Random Forrest (RF), XGboost, Light Gradient Boosting Machine (LightGBM). Using those algorithms, we first made predictions on the claim indicator (whether the policy has a claim or not), then we made predictions on frequency with Poisson objective, and lastly we made predictions on the severity with Gamma objective. We combined those predictions and used them as second level data set in order to produce the final predictions. In order to get the final predictions for the total claim cost, we just multiplied our predictions for claim indicator, claim severity, claim frequency, and exposure. For our model stacking approach, we had a Gini score of 0.09551 which is an improvement over the Tweedie benchmark model.

Though model stacking approach seem to have a better performance on the prediction, this approach contains models that are often considered "black box" methods which lack of interpretability and transparency of model results. Also, this approach requires relative longer time for computation and tuning process. Thus, we decided to aim for a model that is more practical and interpretable which resulted in our final model.

For our final model, we created additional features that might potentially be helpful. For instance, we categorized vehicles into expensive and cheap vehicles according to there vehicle values. We also categorized drivers into young driver and old driver according to driver's age. We also created features with interactions. For instance, we created an indicator variable representing younger driver with old vehicles, young male/female driver, and young male driver with old vehicle. Thus, for our final model, we have variables: exposure, vehicle value, vehicle age, area, driver's age, young male driver, gender, expensive vehicle, frequent area, older driver, young driver with old vehicle, young male driver with old vehicle. Based on previous model building for this project, we found that target encoding of categorical variables would result in better prediction than one-hot encoding. Our final model is consist of 3 models. First model is the indicator model where we used logistic regression to predict whether the policy has claim or not. The second model is the frequency model, which we used Poisson regression to predict the frequency. The last part of the model is severity model where we used Gamma regression to predict the severity of each policy. Our prediction for claim cost is the product of predicted claim indicator, the predicted severity, predicted frequency, and the exposure. In order to improve risk segmentation, we calculated relativities using loss ratio and count ratio based on risk segmentation score then multiply the relitivities and the prediction of the policies that falls into the category. Incorporating relativities in our final model allows us to account for selection bias in claims data and helps us better classify risks.

Results

For our final model, we had a Gini score of 0.10672 which is a significant improvement over the Tweedie regression benchmark model and the stacking models. The relitivities is summarized in the Table 1. For instance, if a driver is at age category 1 (young drivers), the predicted total claim amount will be about 1.8 times higher. If a driver is male, then the predicted total claim amount will be 1.6 times higher. In addition, area F has a very large relativity when comparing with other areas. Our model also separates low and high risks well, and this is illustrated in the lift chart in Figure 2 where we could see the predicted loss and actual loss are very close for the very left end and right end.

Table 1: Features and corresponding Relativities

Features	Relativities
Driver age 1	1.794*
Driver age 5	0.569
Driver age 6	0.715
Male	1.559*
Vehicle age 4	0.971
Area F	2.88*
Area D	0.745
Vehicle type Station Wagon	1.12*
Vehicle type Sedan	0.803
Expensive vehicle	0.5922
Cheap vehicle	0.802

* Relitivities > 1 (Higher risk).

Discussions and Limitations

There are few potential improvements we could make to our final model to increase our predicting power. Clark (2011)[1] has shown credibility is an important feature when predicting expected loss in insurance type of data. Thus, we could incorporate credibility into relativity calculations. In addition, we could use mixed models to better capture the ”area” effect. For our model validation, we could further examine the stability of our model under changing conditions. There are some aspects of the data we would like more explanations. For instance, the driver's age variable takes value from 1 (young) to 6 (old), it would be better to have driver's actual age in years. In addition, the ”area” variable is categorized into from A to F, having more information on how the area is categorized would help us during the modeling process. In addition, having information on driver's education levels, job types, and marital status etc would help us to make better predictions on the total claim amount. Lastly, we found that incorporating policy ID variable into our model increases the performance by a lot (Gini score from 0.10672 to 0.45531). After some investigation we found that the ID variable is predictive of total claim cost only through how the data is sorted, which is by area and loss and after a certain ID within an area, it is all losses. For ethical reasons in Statistics, we did not use the ID variable for our final model.

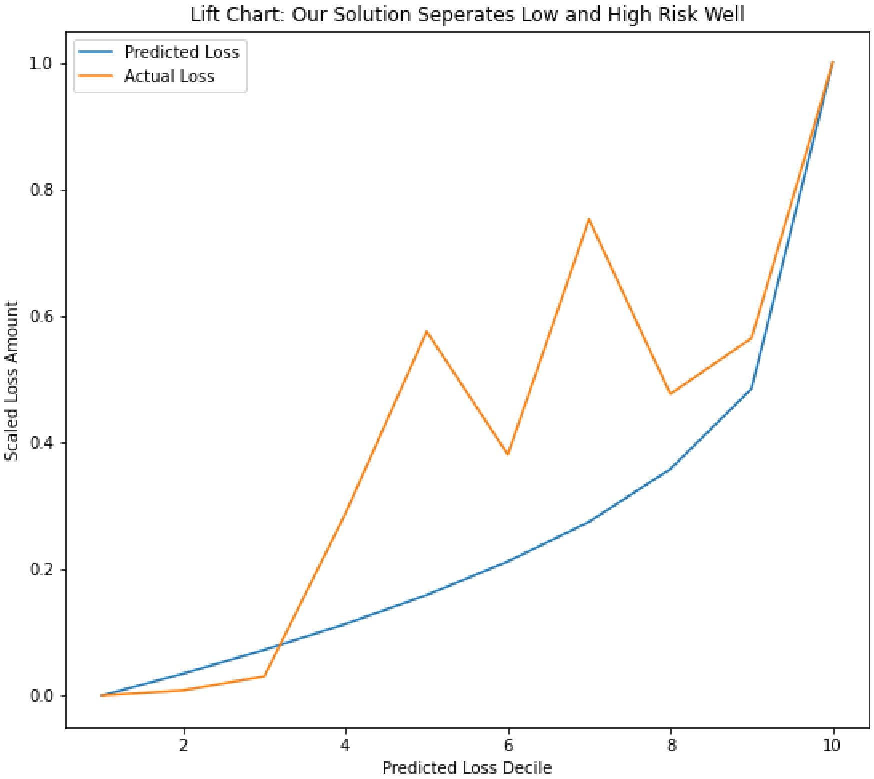

Figure 2: Lift chart shows our model separates low and high risk well

References

[1] David R Clark. Credibility for a tower of excess layers. In *Casualty Actuarial Society E-Forum, Winter 2011 Volume 2*, 2011.

[2] Peter K Dunn and Gordon K Smyth. Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing*, 15(4):267–280, 2005.

[3] Oscar Alberto Quijano Xacur and José Garrido. Generalised linear models for aggregate claims: to tweedie or not? *European Actuarial Journal*, 5(1):181–202, 2015.

[4] Yi Yang, Wei Qian, and Hui Zou. Insurance premium prediction via gradient tree-boosted tweedie compound poisson models. *Journal of Business & Economic Statistics*, 36(3):456–470, 2018.