2016 Travelers Model Competition Report

Jiani Chen, Tongan Liu, Zhiyu Quan, Zhiguo Wang

November 27, 2016
1 Introduction

1. Data Description: The Kangaroo data set is based on one-year vehicle insurance policies from 2004 to 2005. There are 67856 policies, of which 4624 (6.8%) had at least one claim. The data will be split to three parts, training data, validation data and hold out data.

Variable information in the data:
ID: policy identity  
$Veh\_value$: market value of the vehicle  
Exposure: The basic unit underlying an insurance premium  
Clm: Occurrence of claim (0=no, 1=yes)  
Numclaims: The number of claims  
Claimcst0: Claim amount (our response variable)  
$Veh\_body$: Type of vehicles  
$Veh\_age$: Age of vehicles (1=youngest, 4=oldest)  
Gender: Gender of driver  
Area: Driving area of residence  
Agecat: Driver’s age category from young (1) to old (6)

In hold out data, claimcst0, clm and numclaims will be set to NA. Building model on training data and testing the model on validation data. In the end, using best model to score the hold out data.

2. Summary Statistics for the Dataset:

Assuming Numclaims following poisson distribution is reasonable since the mean and variation is same. Only 5% of ID has claim. The $Veh\_body$, Area and Agecat are not balanced variables. Especially, the $Veh\_body$ has 13 categories and very unbalanced. Making several boxplots using log of Claimcst0 versus all the explanatory variables, we found that $Veh\_value$, $Veh\_body$, Gender, Area and Agecat are possibly significant variables.

3. Model Evaluation: The model is evaluated by gini index.
2 Models Considered

1. Tweedie:
   Tweedie distributions are a family of probability distributions which include the purely continuous normal and gamma distributions, the purely discrete scaled Poisson distribution, and the class of mixed compound Poisson – Gamma distributions which have positive mass at zero, but are otherwise continuous. Since Tweedie distribution is the most popular model for predicting auto claim size, we select Tweedie as our favorite model. We using Maximum likelihood estimation to find the Tweedie index parameter p which turned out to be 1.58.

2. Possion and Gamma:
   In order to check our work with Tweedie model, we fit the data by two steps. Using Possion GLM to model number of claims and using Gamma GLM to model claim cost. We found that the two steps model has worse gini index.

3. Inverse Gaussian:
   Directly apply to claimcst0, because we believe multi-step fit should not be better than one step. Since each step we have a prediction error which would dilate for next step prediction.

4. Binomial and Gamma:
   Since the more complexity of model the lower gini index. We try to use Gamma to fit claim amount and use Binomial to predict the probability of having claims or not.

5. Association Rules:
   To identify whether there exists a particular pattern of having claims or not, we used association rules. We use categorical variables to make item set and with 5% support and 85% confident level to find the association rules for having claim or not. Then try to make adjustment to prediction value which based on Tweedie model. The adjustments barely improves gini index. This make sense since characteristics of gini index. If we using other evaluation method, the association rules will be very good remedy for the GLM prediction.

6. Quantile regression:
   Quantile regression is a type of regression analysis which using the method of least squares results in estimates that approximate the conditional mean of the response variable given certain values of the predictor variables, quantile regression aims at estimating either the conditional median or other quantiles of the response variable. Quantile regression is desired if conditional quantile functions are of interest. One advantage of quantile regression, relative to the ordinary least squares regression, is that the quantile regression estimates are more robust against outliers in the response measurements. However, the main attraction of quantile regression goes beyond that. Different measures of central tendency and statistical dispersion can be useful to obtain a more comprehensive analysis of the relationship between variables. In our dataset, since only 5% of ID has claim, the quantile regression has advantages. We used 95% quantile for quantile regression whose gini index achieves similar high value compares to other GLM methods.
3 Model Selected - Tweedie and Quantile

1. Using the GLM with Tweedie family, we built model and made a prediction. We also used the quantile regression. Finally, we made a product of two prediction which increased 10% of gini index compares to the single GLM with Tweedie family.

```r
tweedie_fit <- glm(claimcst0 ~ veh_value +
   relevel(veh_body, 'SEDAN') +
   gender + relevel(area, 'C') +
   relevel(agecat, '4'),
   family = tweedie(var.power = 1.58, link.power = 0),
   data = rbind(train, valid), offset = log(exposure),
   weights = (exposure)^0.58)

hold_tweedie <- predict.glm(tweedie_fit, hold , type = "response")

quantile_fit <- rq(claimcst0 ~ veh_value + gender + relevel(area, 'C'),
   data = rbind(train, valid), tau = 0.95)

hold_quantile <- predict.rq(quantile_fit, hold , type = "response")
py <- hold_tweedie*hold_quantile
```

2. Our interpretation on this product: GLM regresses the mean and using quantile regression giving a weight of the prediction.

4 Variable Selection

4.1 Select Variable by Loop.

1. We rewrote the Gini function for better use during model development.
2. Created a binary addition function which is used to give all combinations of variables.
3. Loop all combinations of variables, record the formula and the corresponding Gini score.
4. Choose the combination of variables which yields the highest Gini score.

4.2 Double Check Significance

By common sense and summary of the data, we believe some variable, like “veh_body”, should be significant. Thus, we tried the following approaches to verify significance.

1. Summary “fit” to check variables’ significance.
2. Change the base of categorical variables and view significance.
3. Group or decrease categorical variables’ levels.
5 Helpful Variables for Explaining Pure Premium

To demonstrate this question, we provide several plots at the end. With the data and common sense, we believe veh_value, agecat, and area are the three most significant elements relating to Pure premium.
We believe that luxury vehicles have better safety features and drivers would pay more attention, such that less likely have severe accidents. Normally, a luxury vehicle driver is older and maturer. Area is a factor contributes to likelihood of the occurrence of accidents. A heavier traffic location or worse climate location could have more claims or worse accidents.

6 Other Useful Information

If possible, the annual or monthly-average mileage could be a useful variable. In the current data, we use exposure to describe the duration of the risk. However, for example, it is vary rare to see a convertible running in the winter. Although this convertible vehicle is covered in a policy but a whole year, it actually may only expose to the risk as exposure = 0.3 depending on usage.