*ISOM 5610, Group 11*

**Risk Management Of Insurance Claims**

*Dec. 17, 2020*

ABHINANDANA KUMAR, Jijitha Neervarath

TANG, Lanrui

TANG, Lanyu

YAO, Yuqian

ZHOU, Jianhua

Executive Summary:

Exposure has the highest influence in response so insurance companies should pay more attention to the exposure of the car in real business cases. Besides, specific risk profiles for customers with different probabilities to issue a claim will be an efficient method for risk management. Based on the probability of making a claim, we are able to anticipate rare customer behavior(making a claim in this case), and be prepared to cope with those high risk clients in order to minimize the cost.

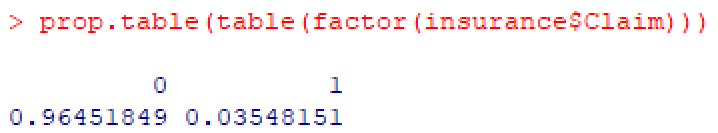
# **1.** **Introduction**

Nowadays, there is a dramatic development of the driving technologies; as a result, the number of claims decrease while the compensation cost for each claim is increased to a large extent. Therefore we, as an insurance company, aim to develop a logistic regression model for predicting the probability of making insurance claims with some useful predictors, and from which we can identify high-risk clients. After that we can conduct efficient claim management by filtering out future applications with high risk, or as an alternative, charging a higher insurance premium to them.

# **2. Analysis & Results**

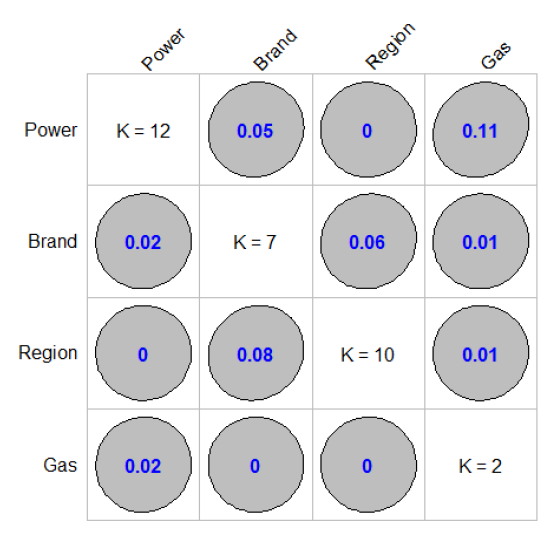
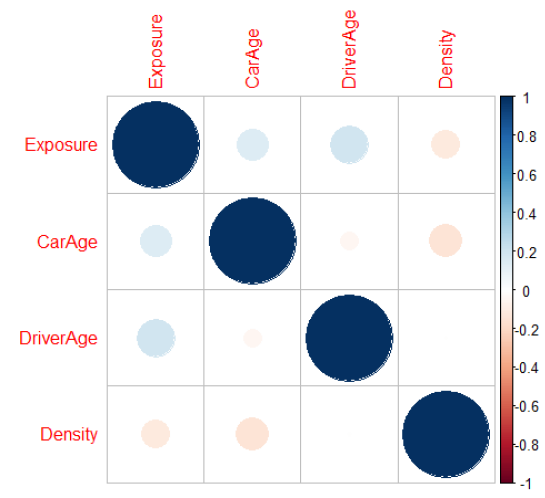
## 2.1 Descriptive analysis

We have 412412 observations, one response (Claim, YES=1, NO=0) and 8 explanatory variables, 4 are continuous variables and 4 are categorical variables. We use split ratio = 0.7 to split our dataset into training and testing sets, build the model based on the training dataset and evaluate the predictive performance based on the testing dataset.

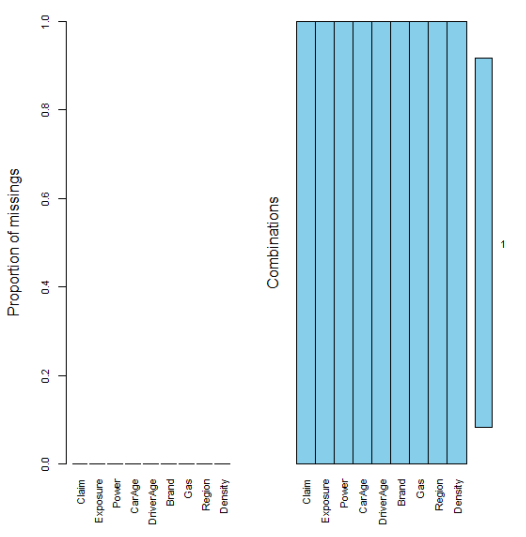
1. The proportion of claim cases is only 3.5%, which is natural in insurance claims since the probability of an accident is relatively small and most people will not make claims. Such a feature indicates the data is highly unbalanced.   
   

In addition, for an insurance company, it is more important to correctly identify a customer who will make a claim than a customer who will not. Thus, we want to have a better control on the TRUE POSITIVE case, and should apply higher cost to it in cost sensitivity classification.

1. The correlation between each two continuous variables are not larger than 0.75, and the relationship between each two categorical variables are also not strong. It means our data is clean.

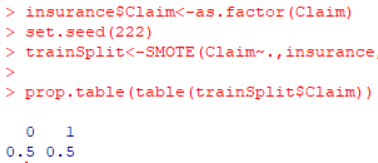


1. There is no missing value.



## 2.2 Sample spliting

Since our data is extremely unbalanced, which is named as a rare event. In order to get better fitness and predictive power, we use SMOTE method to split data and make it balance.

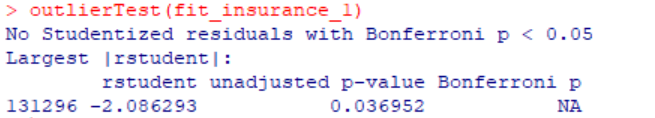
****

## 2.3 Model Building

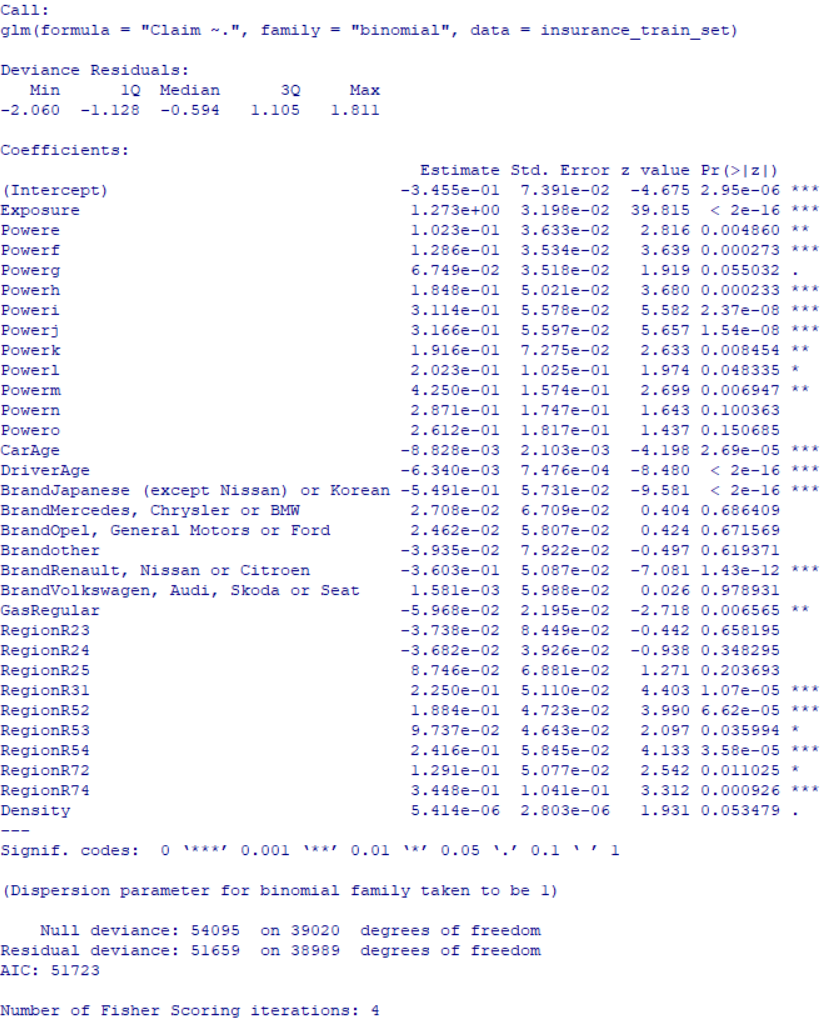
In this case, response is binary so we need to build a logistic model:****, where p is the probability of a client to make an insurance claim (i.e. P(Y=1)).

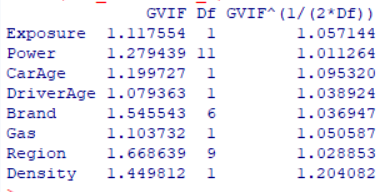
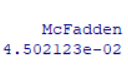
### **Step1: Full model**

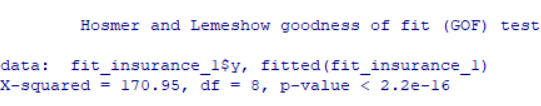
First we fit a model based on full data, and check whether there are outliers. The result shows our dataset do not have outliers.



Here is the estimation of the full model.

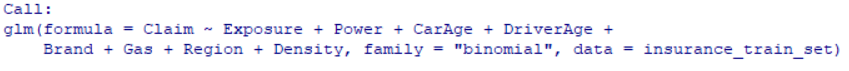
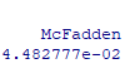
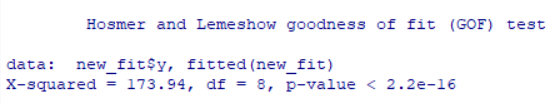
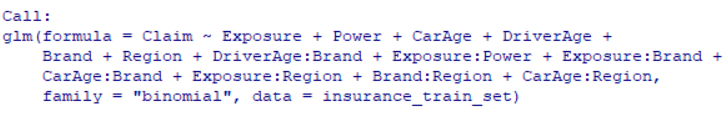
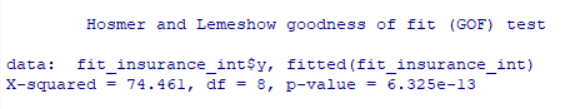


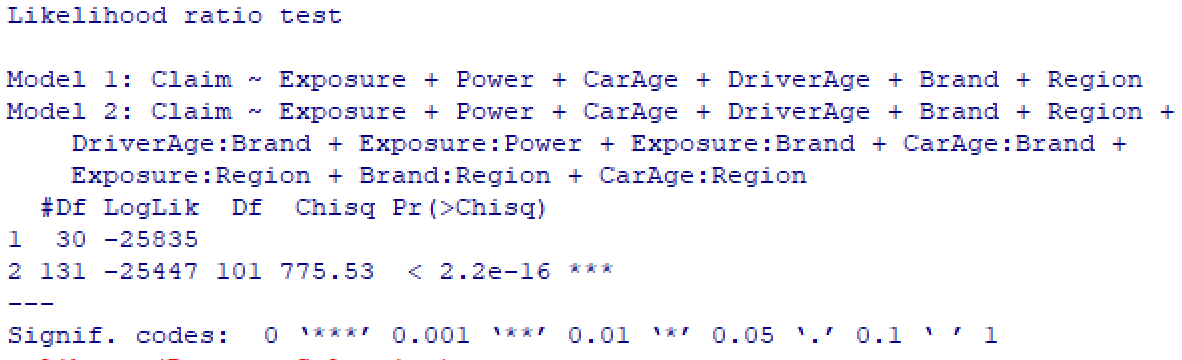


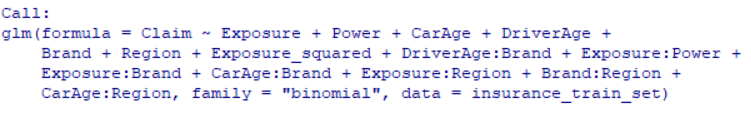
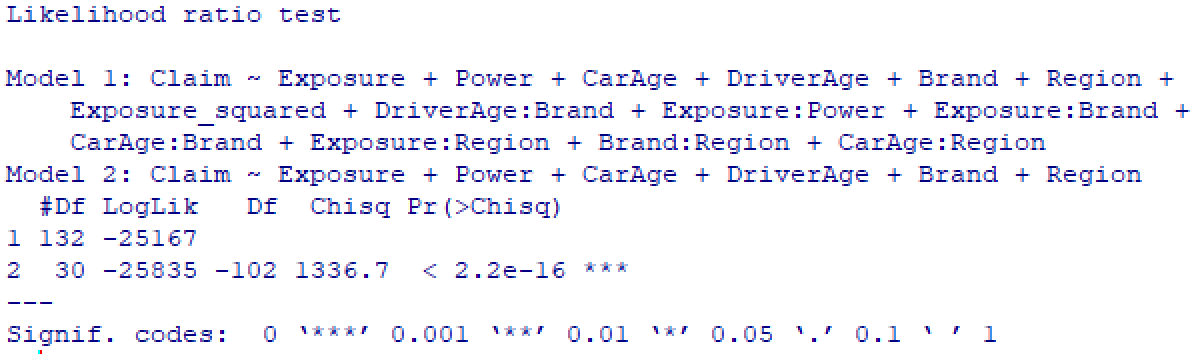
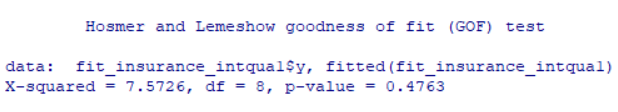


1. Individual coefficient: most predictors have a p-value smaller than 0.05, implying these predictors are useful, given the significant level of 0.05.
2. Goodness of the fit: the pseudo R squared is 0.045. It means only 4.5% of the variation of the response can be explained by the model. It seems that our current model does not have good fitness.
3. Predictive power: AIC is 51723. Since our objective is to predict and to classify, we will mainly focus on minimizing AIC to get a model which has the best predictive power.
4. Multicollinearity: no multicollinearity because the VIF are all lower than 10.
5. Check the outliers and leverage point: We don’t have outliers after conducting the outlier test. Because of that, we do not need to check influential points. Then we move on to check leverage points and get the plots. We computed the leverage points by using this formula: 2\*(k+1)/n where k is the number of predictors.
6. Improvement for current model: p-value of Hosmer-Lemeshow test is lower than 0.05, indicating the fitted model is not adequate and needs to be improved.

### **Step2: Improvement**

1. Stepwise selection: we drop “Gas” and “Density”. The new model has a smaller AIC(51730), and pseudo R squared does not change a lot. P-value is lower than 0.05 in the Hosmer-Lemeshow test, meaning we still need improvement.  
     
     
    
2. Interaction terms: we check the significance of interaction terms and found seven interaction terms are significant, so we add them into our model. The new model has a smaller AIC (51156), and pseudo R squared increases to 5.92%, implying a better fitness. P-value of likelihood test is lower than 0.05, meaning the model as a whole is useful. P-value is lower than 0.05 in the Hosmer-Lemeshow test, meaning we still need improvement.  
     
     
    

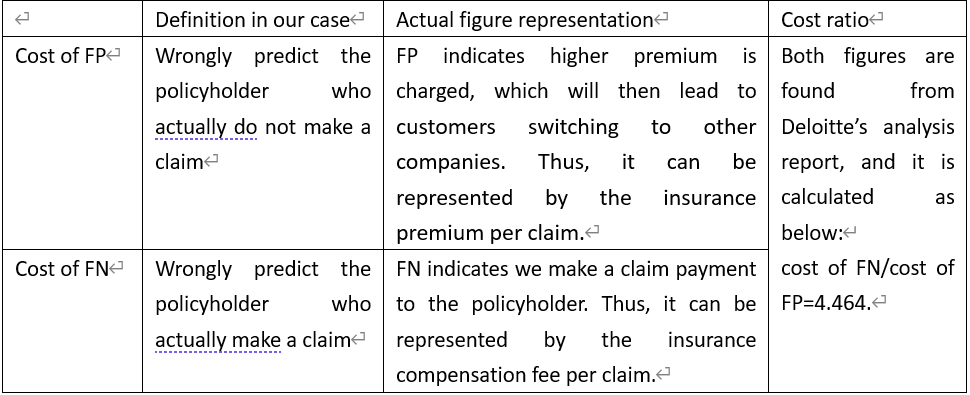


1. Adding the quadratic term“Exposure\_squared”: AIC reduced to 50597, pseudo R squared increases to 6.954%, implying a better fitness. P-value of likelihood test is lower than 0.05, meaning the model as a whole is useful. P-value is larger than 0.05 in the Hosmer-Lemeshow test, meaning the current model is adequate and does not need improvement. Thus, this is our final model.  
     
     
     
   

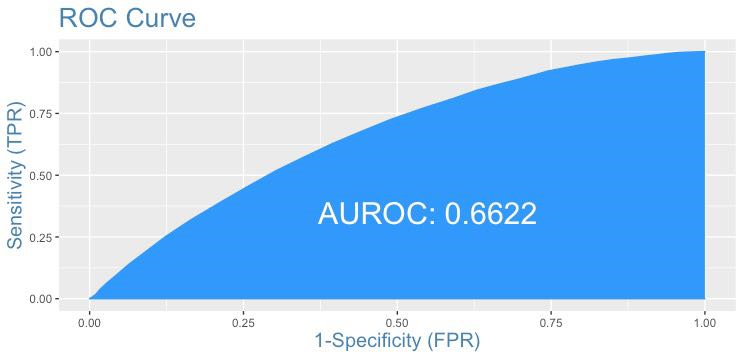
## 2.4 Classification and Prediction

**Classification**

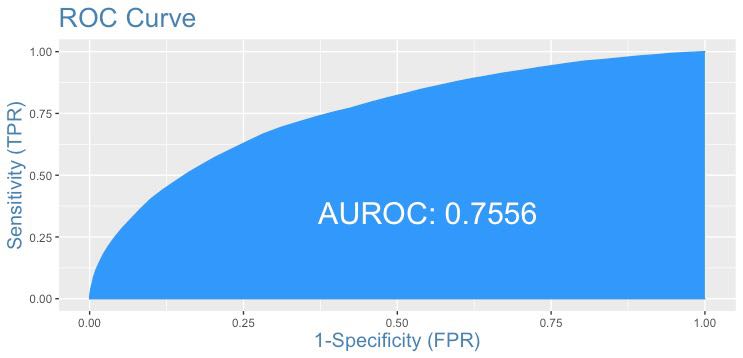
To classify our dataset into two groups, made a claim or no claim, we need to first determine the prediction rule. Firstly, our data is highly unbalanced data, which suggests simply using 0.5 cutoff is unreasonable. Furthermore, in our case, costs of incorrectly predicting 0 and 1 are extremely different. Hence, we should use a cost-sensitive approach. To compute the cost, we should first determine the cost ratio for assigning different costs for FN and FP. The process is shown in the following table:



Then we use the function: to find the best cutoff value that minimizes the cost. The best cutoff based on the cost sensitive approach is 0.17. Then we output the ROC curve.



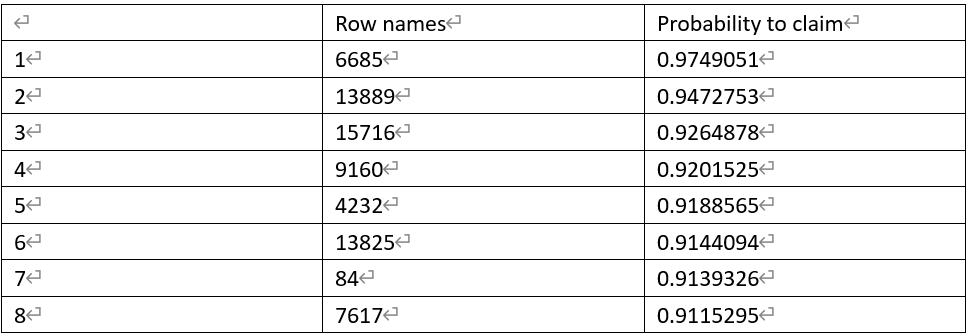
The AUROC is 0.6622 which is considered as poor. In addition, the confusion matrix for the cost sensitive approach is also problematic. Although FN is very low, FP is extremely high which indicates sensitivity is extremely low. Therefore, as an alternative, we use a random forest method to get the best cutoff. And then the AUC is improved to 0.75.



**Prediction**

Our essential focus is to make the prediction so that we can prioritize our clients based on the results.

We use AUC under ROC to evaluate the predictive power of our model, which is 0.665.



We then forecast according to the predict function to prioritize the policyholders. The policyholders with the top 100 largest probability to make claims (the 100th observation = 82.35%) will be our primary focus. Especially for the first eight policyholders who have the probability of making a claim over 90%. These clients are regarded as high risk and we should consider charging them higher premium.

# 3. Conclusion

## 3.1 Model summary

Based on our analysis, one of the most vital steps is to balance our data before building. We have a rare event(Y=0 around 97% while Y=1 only 3%) in our dataset, so using this highly unbalanced data can cause serious problems. During the model building, we use different selection methods and add different types of terms to find the optimal model. A striking feature of our final model is it includes seven interaction terms, showing our predictors have significant interaction effects. Adding significant interaction terms means better fit to the data, and result in better prediction from the model. For our final model, Pseudo R-square improves a lot compared with the initial model. Also, it passes the Hosmer-Lemeshow Test which confirms our final model is good enough. Look at our final model in-depth, we detect the relationship between predictors in our model and the response. From that, we can also gain more insight into the classification of policyholders. Based on the z-statistic of the final model. Exposure has the highest influence on response since it has the highest absolute value of z-statistic. It is natural that higher degree of exposure results in more cases of car accidents and then higher probability of insurance claim. It inspires us, in order to identify high-risk clients, the insurance company should pay more attention to the exposure of the car in real business cases.

## 3.2 Business implication

Our company is in the business of risk-taking, so identifying high risk clients is critical for long-term operation. Through constructing the logistic model with relevant predictors, we can gain business insights; and therefore better arrange our resources and time.

Through classification, we divide each individual into separate groups. And from that, individual claims can be accurately estimated. Based on these results, we can then take actions in advance to mitigate potential risk. Simultaneously, the cost can also be remarkably reduced.

Furthermore, the logistic model has provided the probability of making a claim which is one of the most critical determinations of the risk. By prioritizing our clients, we can then perform in-depth analysis to understand the factors leading to high probability of making a claim. After that, we can use it as a benchmark to help filter out future applications with high risk.

On the basis of the results obtained from classification and prediction, we can then arrange different risk profiles with competitive premium for customers with different probabilities to issue a claim. By specifying each customers’ risk profile, we can then strike a balance to minimize the cost and to maintain the competitive premium.

## 3.3 Extra business reminder

In a real world situation, there is a kind of unpredictable event that is beyond what is normally expected of a situation and has potentially severe consequences. We called it “Black swan events” and it is crucial for risk management. However, due to the limitation of the logistic model itself, the black swan event cannot be presented in the model (no matter how perfect it fits the data). So here we want to give insurance companies an additional risk warning, even though important factors of making a claim or not have been found out. They still need to be aware of the potential loss of extreme events.

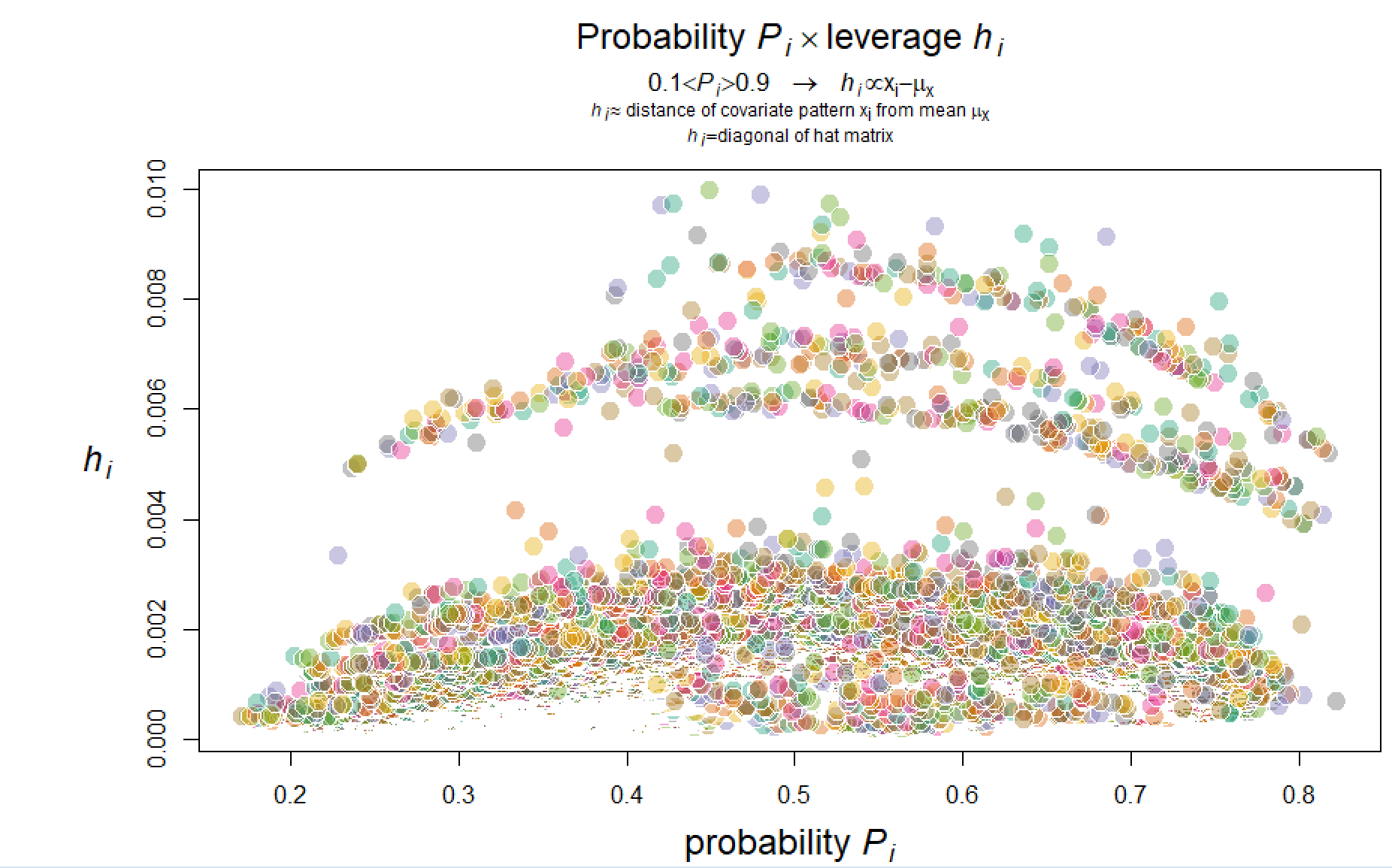
***References***

Deloitte(CN):Analysis of china compulsory third party liability insurance report(2018)

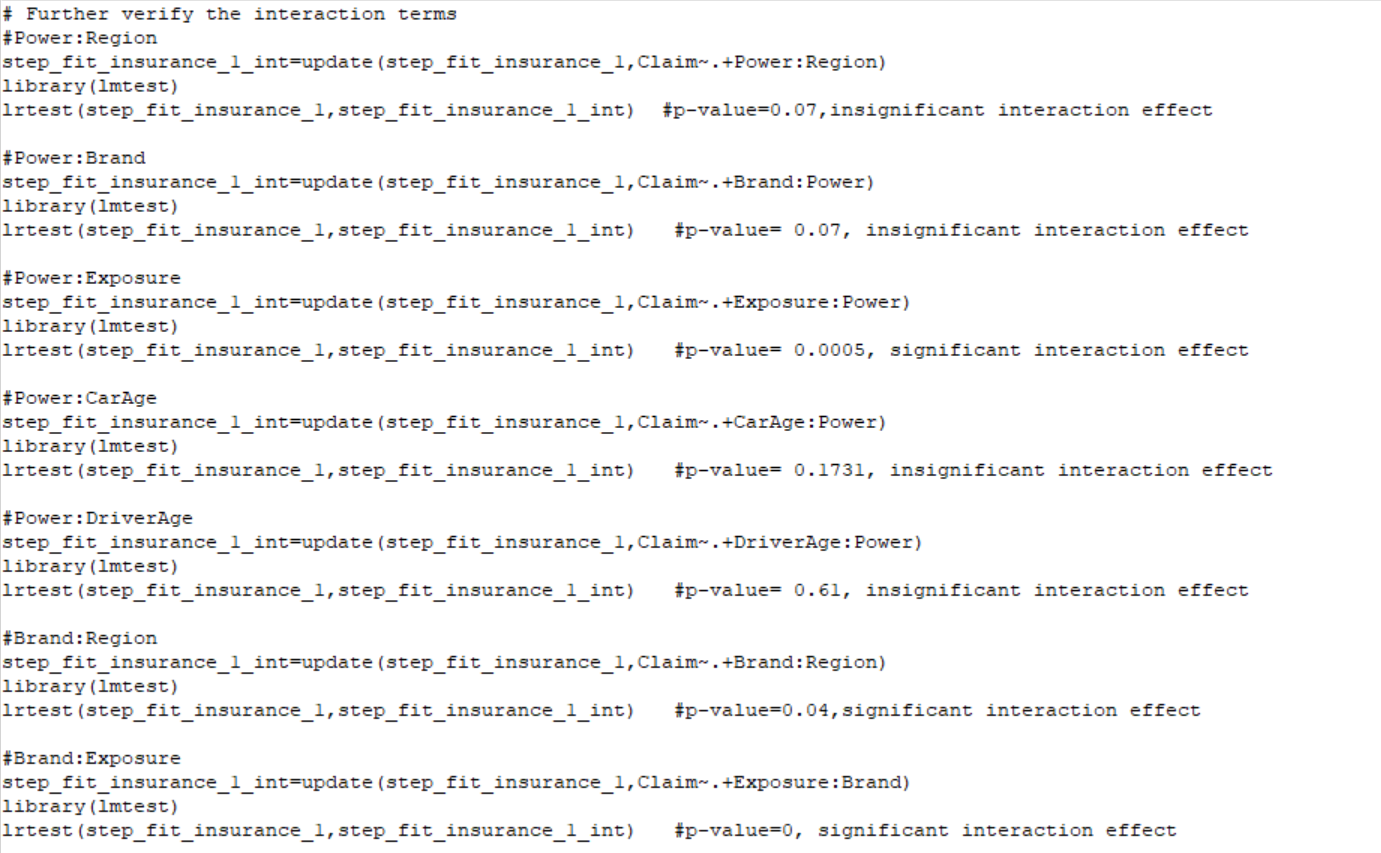
Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences,* *505*, 32-64. doi:10.1016/j.ins.2019.07.070

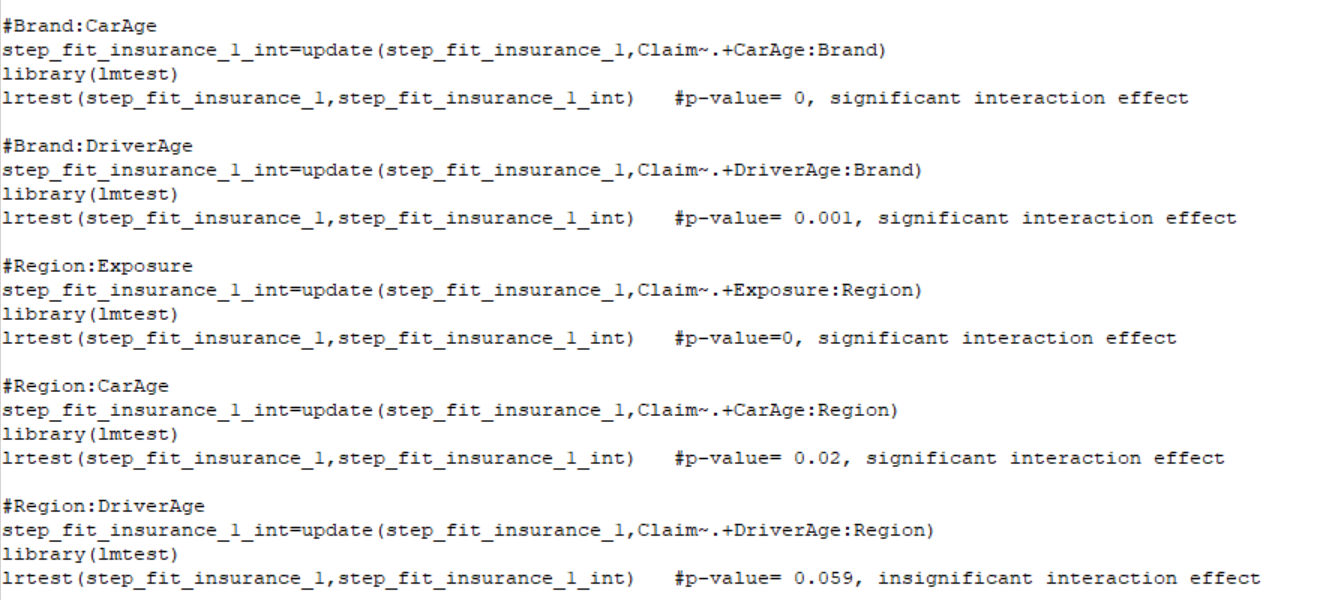
Malhotra, R., Jha, M., Poss, M., & Acharya, R. (2017). A random forest classifier for detecting rare variants in NGS data from viral populations. *Computational and Structural Biotechnology Journal*, *15*, 388–395. https://doi.org/10.1016/j.csbj.2017.07.001

Appendix 1: Check leverage points



Appendix 2: Interaction term effect





Appendix 3: Z-statistic and P-value of final model