Porto Seguro's Safe Driver Prediction Individual Report

ID5059 Knowledge Discovery and Datamining

Group: **G05**

Student ID: **180025784**

Contents

ABSTRACT 2

INTRODUCTION 2

METHODS 3

-DATA MANIPULATION 3
-GENERALIZED LINEAR MODELLING WITH H2O 4

RESULTS 5

-MODEL 1 5
-MODEL 2 6
-MODEL 3 7

DISCUSSION 8

CONCLUSION 9

REFERENCES 9

APPENDICES 10

PLOT OF MISSING VALUES IN SOME VARIABLES 10
PLOT OF CORRELATION AMONG ALL THE VARIABLES 11

Abstract

Porto Seguro, one of Brazil's largest auto and homeowner insurance companies, pays much attention to the significance of the accuracies in car insurance company's claim predictions. Due to the fact that if this fails to be done, the cost of insurance for good drivers will probably increases as well as that for bad drivers will decreases. The paper initially uses a generalized linear modeling (GLM) as a logistic regression to be fitted with the dataset provided by Porto Seguro, and then predicts the probability that a driver will file an insurance claim next year. The improvement of the GLM can be made by changing different *k* value for the *k*-fold cross validation and containing lambda search. Nevertheless, the best GLM still turns out to show a poor performance with public Gini score of 0.2366516. Compared with other models, such as XGBoost, the GLM should not be used for this dataset.

Introduction

The present report analyses the result of GLM fitting the Porto Seguro dataset to discuss if GLM could be used to predict the probability of driver filing an insurance claim next year well. This report aims to fit a GLM and see how it performs in the prediction. The train and test dataset used in the report are provided by Porto Seguro, which is available on Kaggle[1]. The train dataset have 57 explanatory variables and 1 response variable called 'target', that is, the probability of driver filing the insurance claim next year. There are features of explanatory variables tagged in the name of these variables, such as *ind*, *reg*, *car*, *calc*. Plus, the postfixes of the variables contain *bin* and *cat*, which indicates binary and categorical features respectively. In this connection, variables without postfixes are either continuous or ordinal. Value of -1 refers to the missing value.

The results of GLM demonstrates it not suitable for fitting noisy data. Also, GLM overfits the data because of the large explained deviance 0.9693 and the AUC score 0.6181. The normalized Gini score of the best GLM is 0.2308, which is not so good compared with other models, e.g. XGBoost. By this, in order to predict better for Porto Seguro test dataset, GLM should not be used. The target audience of this report are people who have statistical and machine learning knowledge. The h2o package in R is primarily used in this analysis. Data

[1] https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data

manipulation, containing data cleaning and data analysis, are conducted in R programming language using R-Studio version 3.5.2 (R Core Team, 2018).

Methods

-Data Manipulation

The train dataset totally has 59 columns, including id number, 57 explanatory variables and 1 response variable 'target'. Amongst the 57 explanatory variables, there are 17 binary, 10 continuous, 16 ordinal and 14 categorical variables. After calculating the percentage of missing values of each explanatory variable, it is easy to find that *ps_car_03_cat* and *ps_car_05_cat* have 69.09% and 44.78% missing values respectively. Thus, these two variables are dropped for better prediction in the future. In addition, when using the H2O package in R Studio, we have to eliminate the high correlations among the variables, especially those in the same group. The below Figure 1 shows clearly the strong correlation between some of the explanatory variables. Thus, *ps_car_13*, *ps_reg_03*, *ps_car_14*, *ps_ind_12_bin*, *ps_ind_17_bin* and *ps_ind_18_bin* are also dropped. However, the pairwise comparison indicates there's no correlation between the response variable and explanatory variables.

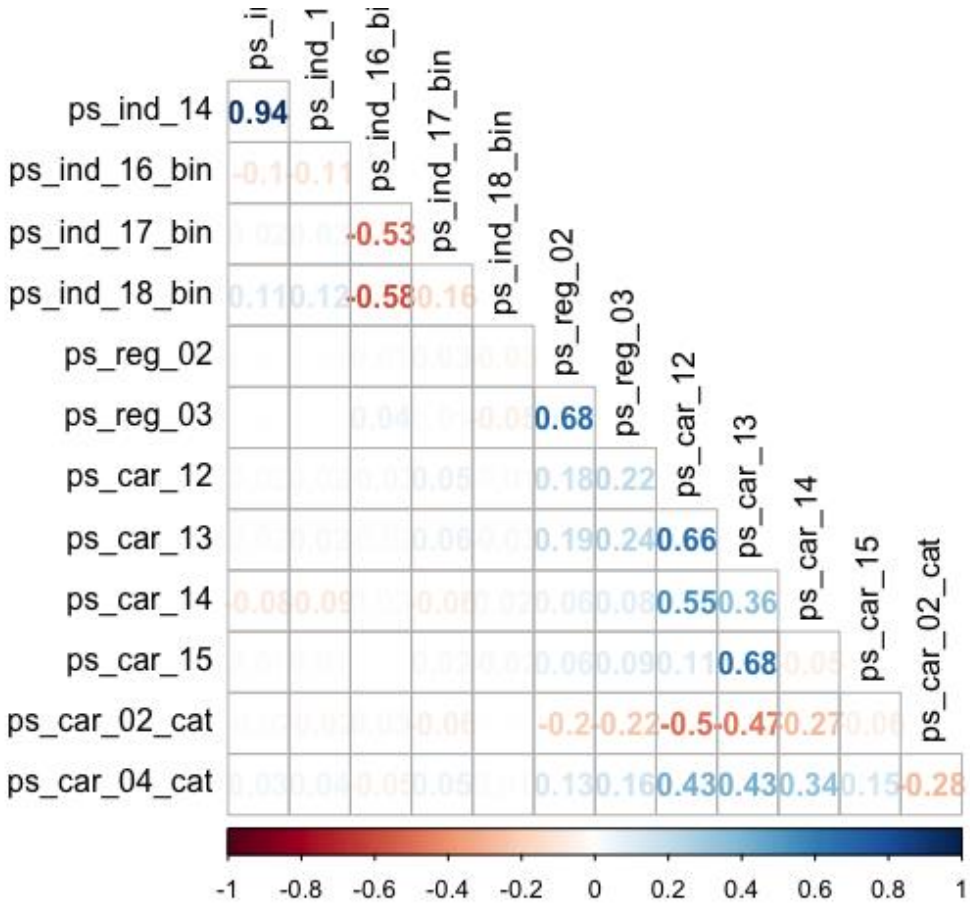


Figure 1 Strong positive and negative correlation between explanatory variables. The deeper the colour of the number, the stronger correlation between variables. The number tends to blue displays the two variables are positive correlated whilst that tends to red is negative.

The imputation of dataset contains replacing missing continuous values with mean, ordinal values with mode, categorical values with new factor level -1. Also, those variables with postfixes '*bin*' are set to be logical while '*cat*' are set to be factor. Furthermore, because of the dataset is unbalanced, which may contribute to the bias in the model, Unbalanced package is used to balance the under sampling train dataset, which makes the dataset contain 43388 rows.

-Generalized Linear Modelling with H2O

H2O, as a software, has platform contains interfaces of Python, R, Java etc. H2O can be used to make many machine learning algorithms and show the output for predictions, one of which is GLM (Nykodym et al., 2019). In this report, H2O cluster version 3.22.1.1 is used in R Studio Version 3.5.2.

There are generally two ways to optimize the model for better results, including changing the *k*-value of the *k*-fold cross validation and accepting the lambda search in the model. Initially, I did a *k*-fold cross validation without lambda search on fitting the model. It is widely known that *k*-fold cross validation first divides the dataset into *k* sections with equal observations. The *k*-1 sections are called train data that are used to be trained for computing validation on the remaining section, which is called test data (Krstajic et al., 2014). The starting value chosen for *k* is 5, which was later modified to 10 as some researchers (Coelho, Infante and Santos, 2013) believe that 10-fold cross validation are widely used. However, the result of the model did not change.

Another way to find the optimal GLM is to choose whether use lambda search or not in the model. Lambda search refers to efficient and automatic search for the best value of lambda parameter (Nykodym et al., 2019). When using lambda search, GLM will first fit a model with the largest regularization and reduce it little by little until the model become overfitted (Nykodym et al, 2019). Due to the fact that lambda search can move out the noise, it may be able to deal with wide datasets and run with a dataset with many predictors like Porto Seguro's dataset. Therefore, my job is to focus on finding the optimal lambda parameter so that the model can be improved.

Results

-Model 1

The first model is GLM model using 5-fold cross validation without lambda search. Figure 2 is the confusion matrix of this model. From the result, we can calculate the proportion of misclassified observations, i.e., (1244 + 19024) / (2670 + 19024 + 1244 + 20450) = 46.71%. This means 46.71% possibility of the model prediction might be wrong, in other words, 53.29% of the probability the case will be correctly predicted.

Confusion Matrix		Actual	
		0	**1**
Predicted	**0**	2670	19024
	1	1244	20450

Figure 2 Confusion Matrix of Model 1

The output (Figure 3) of the model also gives some other scores. The public Gini score of the model is calculated with 30% of the test data whilst the private Gini score is calculated with 70% of the test data. Due to the fact that the model will be better if Gini score tends to be 0.5, this model may not be a good model. The explained deviance of the model is 0.9690, which also shows that the model is overfitted. The AIC of the model is 58381.85, which can be compared with other model later. The AUC is not good as well. However, the MSE of this model is quite small.

Public Gini	Private Gini	Explained Deviance	MSE	AIC	AUC
0.22984	0.23681	0.9689986	0.2394116	58381.85	0.6183258

Figure 3 Results for Model 1

Figure 4 shows the most significant covariates of Model 1 with their coefficients. Further interpretations are hard to made because there is only a few description of the explanatory variables in the dataset.

Names	Coefficients	Sign
ps_ind_05_cat	0.158701	POS
ps_car_07_cat	0.123772	NEG
ps_car_15	0.114806	POS
ps_car_12	0.100132	POS
ps_ind_16_bin	0.095095	NEG

Figure 4 Top 5 Features for Model 1

-Model 2

Model 2 is a GLM using 10-fold cross validation with lambda search. From the confusion matrix, we can calculate the misclassification rate of the model (20089 + 711) / (1605 + 20089 + 711 + 20983) = 47.94%. It is clear that the probability of wrong prediction goes up from 46.71% to 47.94%, which means that Model 2 may not be better than Model 1. There is still 52.06% possibility getting correct predictions though.

Confusion Matrix		Actual	
		0	1
Predicted	0	1605	20089
	1	711	20983

Figure 5 Confusion Matrix of Model 2

Although the explained deviance is still quite large, public Gini and private Gini are both comparatively smaller than those in Model 1, which further proves that Model 2 is better than Model 1. As a result, the reason why the sensitivity become small could be the regularization shows less penalty on misclassification rate because of lambda search. Plus, the AIC also raises from 58381.85 to 58379.42, which indicates Model 2 is better than Model 1 despite of the higher MSE and lower AUC.

Public Gini	Private Gini	Explained Deviance	MSE	AIC	AUC
0.23079	0.23761	0.9692907	0.2395167	58379.42	0.6181039

Figure 6 Results for Model 2

The important covariates of Model 2 with the coefficients are displayed in the Figure 7. The covariates are the same five as those in Model 1, but the coefficients are all smaller than the previous ones.

Names	Coefficients	Sign
ps_ind_05_cat	0.150335	POS
ps_car_07_cat	0.117719	NEG
ps_car_15	0.108642	POS
ps_car_12	0.094756	POS
ps_ind_16_bin	0.09169	NEG

Figure 7 Top 5 Features for Model 2

-Model 3

Model 3 is on the basis of Model 1, but it enables lambda search, which also indicates that the difference between Model 2 and 3 is the reduction of *k*-value from 10 to 5. From the confusion matrix of Model 3, the misclassification rate can be computed as (20132 + 690) / (1562 + 20132

+ 690 + 21004) = 47.99%. Compared with Model 2, it continues going up, but as mentioned before, it does not mean Model 3 is definitely poorer than Model 2.

Confusion Matrix		Actual	
		0	**1**
Predicted	**0**	1562	20132
	1	690	21004

Figure 8 Confusion Matrix of Model 3

From the output of Model 3, the public Gini and private Gini both decreases compared with Model 2, which means Model 2 is better. The increase of explained deviance, MSE and AIC and the decrease of AUC also prove this. From these results, it can be sure that Model 3 is not better than Model 2.

Public Gini	Private Gini	Explained Deviance	MSE	AIC	AUC
0.23074	0.23758	0.9693422	0.2395344	58380.52	0.618069

Figure 9 Results for Model 3

Below are the top five features for Model 3. Compared with Model 1 and 2, the coefficients become smaller again.

Names	Coefficients	Sign
ps_ind_05_cat	0.149504	POS
ps_car_07_cat	0.117119	NEG
ps_car_15	0.108031	POS
ps_car_12	0.094175	POS
ps_ind_16_bin	0.091348	NEG

Figure 10 Top 5 Features for Model 3

Discussion

According to the calculated Gini Index, it is clear that Model 2 is the best model among these three, followed by Model 3 and the worst one is Model 1. Thus, we can concluded that in the process of analysing GLM to do prediction, Model 2, a GLM using 10-fold cross validation

with lambda search is the optimal model with public Gini 0.23079 and private Gini 0.23761. The probability of correct prediction the using this model is 52.06%, which is only 2.06% better than random guessing. Nonetheless, the model can be further improved by trying a grid search for the value of alpha in that the value of alpha has not been considered. However, due to the fact that XGBoost model performs much better than GLM, for the case of Porto Seguro, it is unnecessary to spend much time doing GLM grid search.

Conclusion

All in all, the primary finding of this report is that the GLM using 10-fold cross validation with lambda search is the best model that can predict the probability of driver filing an insurance claim next year for Porto Seguro. Also, another finding is that GLM can be improved to do prediction by changing the value of *k* in k-fold cross validation and enabling lambda search in the model. If conducting further investigation on GLM in this case, a grid search of alpha value can be considered. However, public Gini and private Gini of the best GLM is only 0.23079 and 0.23761, which means GLM does not fit the dataset quite well when predicting. Therefore, other algorithms such as XGBoost should be considered. Actually, public and private Gini of the XGBoost model, the best model among all the algorithms my group has worked on, is 0.28243 and 0.28641, which is comparatively much better than GLM.

References

Coel, R., Infante, P. and Santos, M.N. (2013) 'Application of generalized linear models and generalized estimation equations to model at-haulback mortality of blue sharks captured in a pelagic longline fishery in the Atlantic Ocean', Fisheries Research, 145, pp. 66-75.

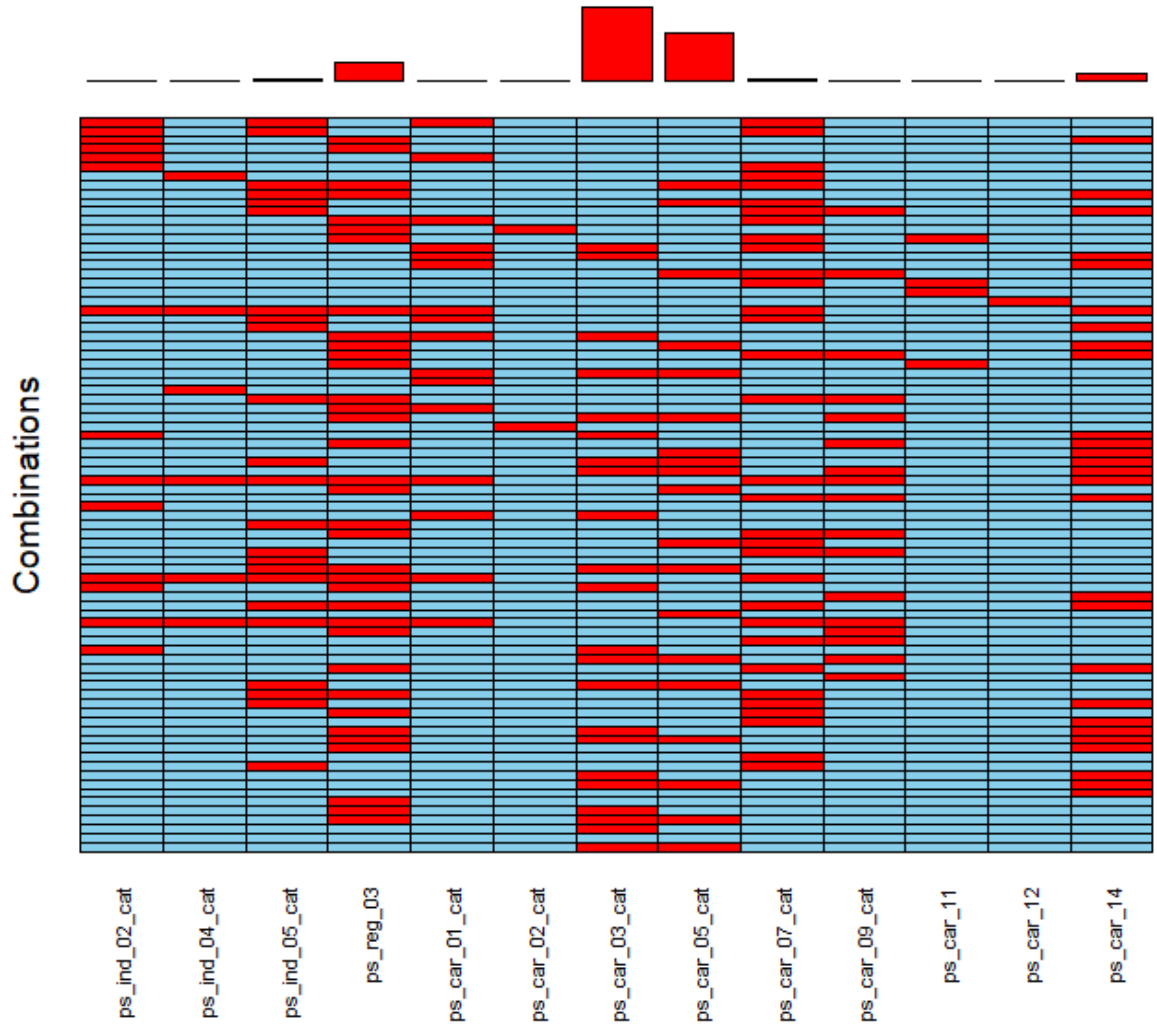
Krstajic, D., Buturovic, L.J., Leahy, D.E. and Thomas, S. (2014) 'Cross-validation pitfalls when selecting and assessing regression and classification models', *Journal of Chemintormatics*, 6(1), pp. 1-15. *BMC* [Online]. Available at: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-6-10/ (Accessed: 10 April 2019).

Nykodym, T., Kraljevic, T., Wang, A. and Wong, W. (2019) *Generalized Linear Modeling with H2O (7th Edition)* [Online]. Available at: http://h2o-release.s3.amazonaws.com/h2o/master/4616/docs-website/h2o-docs/booklets/GLMBooklet.pdf/ (Accessed: 8 April 2019).

R Core Team (2018) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/ (Accessed: 2 April 2019).

Appendices

Plot of missing values in some variables



Plot of correlation among all the variables