Porto Seguro's Safe Driver Prediction

Team : S.A.S.

Suvam Das (MT2020022)
International Institute of Information Technology, Bangalore
(suvam.das@iiitb.org)

Aditya Saha (MT2020093)
International Institute of Information Technology, Bangalore
(aditya.saha@iiitb.org)

Subhodeep Sahoo (MT2020129)
International Institute of Information Technology, Bangalore
(subhodeep.sahoo@iiitb.org)

Abstract **--- In todays world, automobiles specially cars and luxury cars become a important part of human life because it saves our precious time in travelling and also provides status and the opportunity for personal control and autonomy. For purposes of safety and security, Law has made it compulsory that all the owners of vehicles should possess Car Insurance. But nnaccuracies in car insurance company's claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones. So, Porto Seguro, one of Brazil's largest auto and homeowner insurance companies, is looking for a powerful machine learning methods that will allow to further tailor their prices for the good drivers, and hopefully make auto insurance coverage more accessible to more drivers. The purpose of this study was to predict safe driving behaviors among car drivers by building a model that predicts the probability that a driver will initiate an auto insurance claim in the next year. Through this work, we have tried to extract important columns via feature engineering and then predict the probability using different traditional machine learning models. Such work is a very challenging task and the outcomes may lead to decrease road accidents by ensuring less cost for insurance among good drivers.**

Keywords **--- Feature Engineering, Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost**

I. Introduction

Improving the accuracy of insurance claims benefits both customers and insurance companies. Incorrect predictions effectively raise insurance costs for safe drivers and lower costs for risky drivers, and can be costly to insurance companies. Better predictions increase car-ownership accessibility for safer drivers and allow car insurance companies to charge fair prices to all customers. Better predictions also lead to improved profits for insurance companies.

The general aim of this project was to investigate the possibilities of developing statistical models to predict individual driver crash involvement based on driving style, demographic, and behavioral history data. Such models have a range of applications, in particular in the areas of fleet safety management and insurance. Although the relationship between individual drivercharacteristics and road safety has a long research history, the advent of large sets of naturalistic driving data, which include a significant number of crashes as well as driver behavior and demographics data, allows for exciting new research possibilities.

It is well known that a small proportion of drivers often account for a major proportion of crashes (Sagberg et al., 2015), a phenomenon often referred to as the Pareto principle or the 80– 20 rule, that has also been observed in many other domains. Thus, it is of great value to be able to identify these risky drivers before crashes happen. For example, is a driver who is regularly speeding and/or tailgating, and/or has a history of traffic violations, more likely to crash than a driver adopting a less aggressive driving style who has received no tickets in the past? If so, can such risky drivers be reliably identified based on individual driver characteristics, such as observed driving style, demographics, personality screening, or behavioral history. In any case, it would be premature to dismiss the possibilities of predicting crash involvement from enduring personal factors solely based on the present results, and there are several ways the classification models may be improved and there are other ways of analyzing these data that may shed further light on the relationship between enduring personal factors and crash involvement.

The problem is as follows: "given a series of unlabeled features collected by an insurance company about a customer, can we predict whether the customer will file an insurance claim during a period of interest?" This paper is divided into different sections discussing our work. Section II discusses about the related works which has been done. Section III contains a breif overview of the dataset. Section IV and section V discusses about the observation and preprocessing done on the available dataset. Section VI and Section VII discusses about the model selection, model building and results. Finally the paper concludes with a conclusion in Section VIII and challenges, future scope, acknowledgement and references.

II. Related work

Insurance claim prediction has rarely been studied in past 229 projects. The project that is most similar to this one was conducted by Diveesh Singh in 2016 using images of drivers while operating their vehicle to predict driver behavior (Singh 2016).

However, insurance is a topic that has been researched more broadly in the field. Drivers are the contributing factor in the majority of road crashes and understanding the relationship between individual driver characteristics and crash involvement has been a

longstanding goal in road safety research (e.g., Elander et al., 1993; Guo et al., 2010; McKenna,1983). Smith et al. (2000) use several techniques such as decision trees and neural networks to study which customer characteristics affect retention in their policy. Yeo et al. (2001) employ clustering techniques to group customers according to their risk classification and then regression methods to model the expected claim costs within a risk group. Outside of automobile insurance, researchers have also studied health insurance (Browne 1992), non-life insurance (Salcedo-Sanz et al. 2005), and many other situations where risk plays a substantial factor in pricing (Lee & Urrutia 1996, Cummins & Phillips 1999). Klauer et al.(2006) and Victor et al.(2015) also employ some decision tree based bagging and boosting algoritms to study the drivinganalyses of driver inattention and crash/near crash involvement.

III. Dataset

The dataset used in this project was provided by Porto Seguro, one of Brazil's largest auto and homeowner insurance companies. This dataset contains different attributes which affects the probability that an auto insurance policy holder files a claim. The columns names are not specified in dataset. All the columns are mainly separated by names as reg, ind, and car. Also postfix, such as bin and cat are appended to denote binary and categorical columns. If there are no postfix appended to columns then these columns are either continuous or ordinal columns. Also null values are provided with -1. The dataset is splitted into train and test set. Training data initially contains 59 columns and 416648 unique rows. Out of 59 columns, one column named as 'target' denotes ground truth. Test set contains 58 columns and 178564 unique rows. Output of test set is not revealed. Both set contains id to unquely identify each row present in column.

IV. Observations

In this section, a few of the observations on the dataset are discussed. This is important to get a better insight of the data for further preprocessing. Ground truth in train set contains 0 and 1 indicating 'insurance not claimed' and 'insurance claimed' respectively. Hence this is a binary classification problem.

1. Distribution of data between claimed and not claimed

On observing the distribution of datapoints in between the target variable, it is evident that data points are unevenly distributed between infected vs non-infected. Datapoint distribution between claimed and not claimed is 15221, and 401427, which leads to the approximate ratio of claimed to not claimed to 26.37:1. This is visually evident from the plot provided in Figure 1. Hence, the problem is highly imbalance binary classification problem.

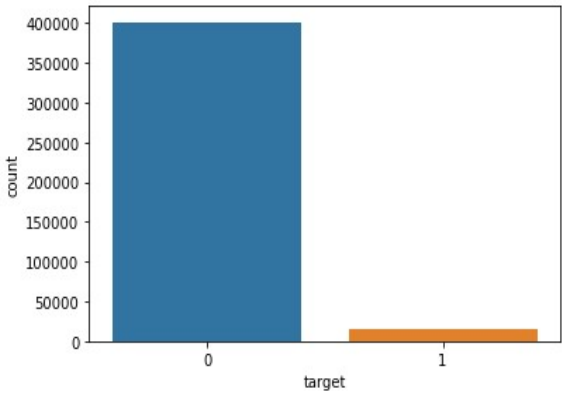
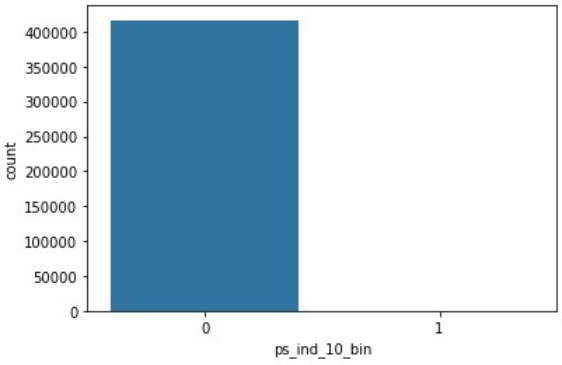
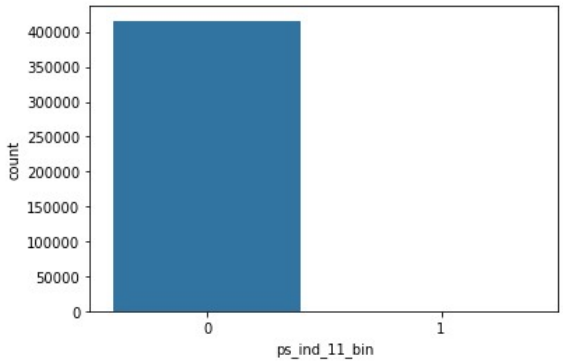


Figure 1 : Distibution of datapoints among not claimed and claimed, represented as 0 and 1 respectively.

2. Skewness in features

Upon carefull observation and going through features, it has been observed that there are few features which are highly skewed. There are few columns which have high amount of skewed data. The columns are ps_ind_10_bin, ps_ind_11_bin, ps_ind_12_bin, ps_ind_13_bin. The distrubution of data of the columns are given in figure 2.





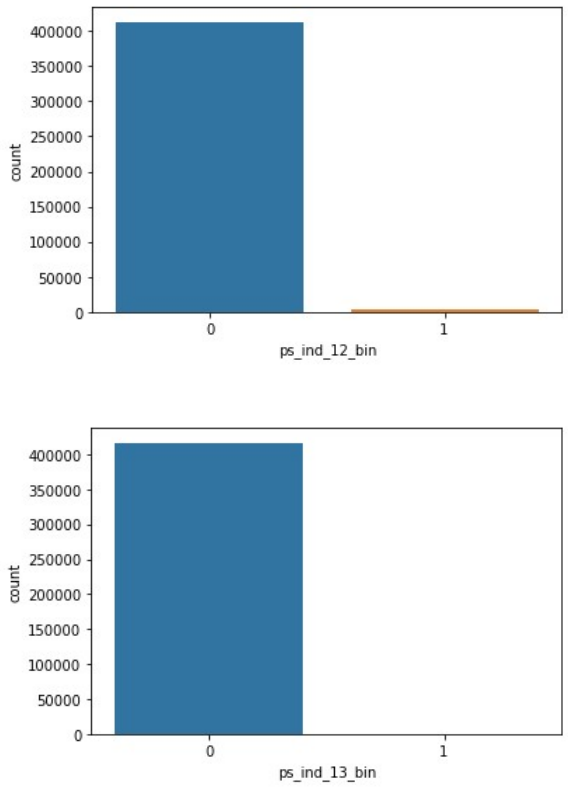


Figure 2 : Columns with highly imbalance data.

3. Missing data

The dataset contains few columns with high null values. Rest of the few columns contains little amount of null columns. Percentage of amount of null values present in the dataset are provided in Table1. ps_car_03_cat and ps_car_05_cat contains 69.17% and 44.79% respectively.

Table1 : Columns and Percentage of missing values

Column name	Percentage
ps_car_03_cat	69.17
ps_car_05_cat	44.79
ps_reg_03	18.12
ps_car_14	07.16
ps_car_07_cat	01.94
ps_ind_05_cat	00.97
ps_car_09_cat	00.09
ps_ind_02_cat	00.04
ps_car_01_cat	00.02
ps_ind_04_cat	00.01

V. Data Preprocessing And Feature Extraction

Data Preprocessing and feature Extraction is a very important step to clean the data and extract information out of data. This extracted information is used in further processing, to train the model and predict outcome. From the observation, it can be concluded that many number of columns contains missing values and many columns are also highly skewed. The dataset being imbalance in nature, makes it a very challenging task to properly process the data to extract the information properly for both the classes. Reducing the dataset size by removing unnecessary columns and preserving important columns is a very crucial in case of our dataset. Rest of the section discusses about the data preprocessing and feature extraction performed in our work.

1. Removal of columns having high missing values

From Section IV part, it is evident as it is evident that ps_car_03_cat and ps_car_05_cat contain high amount of null values. These features do not contribute to training of model. These missing values can be replaced by any statistical methods such as mean, median or mode. But replacing such high number of missing values with statistical measure may distort the original distrubution. Hence it is safe to drop such features from both training and testing dataset. This also reduces the dimension of feature set, which further leads to faster training of model.

2. Removal of highly imbalanced features

From the obsevation Section IV part 3, it is evident that there are few features which are highly imbalanced in nature. Using these feature for learning purpose does not adds any value to the model but only increases feature dimension. Hence these features can be safely removed from the dataset.

3. Filling of missing data

There are few columns in the dataset which consists of small amount of missing values. These missing values can safely be filled using a statistical measure applied over the dataset. Mainly mean and mode have been used to fill the missing values based on values are continuous or categorical respectively.

VI. Model selection

In this work, we are trying to predict the probability of a customer availing for insurance. Hence, the problem is two class classifier problem. During our early stages model, we have tried Logistic Regression and Random Forest. Plain Logistic regression is highly over fitting the densely populated class. Accuracy score was too high so we had to reject it. Next we used logistic regression with class weights. We used “balanced” class weight. The “balanced” class weight gives higher weight to sparse class so that learning can be balanced. But it was difficult to achieve a good score using these models as these are not primarily imbalance class predictors. We tried KMeans cluster classifier as well. Kmeans algorithm is

an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. We tried Kmeans clustering as well. The philosophy was to find the cluster centers for the two clusters and assign new data points a probability based on how close they were to the two cluster centroids. But there were no well-defined clusters in the data so this method did not turn out well.

After this, we tried different ensembling models such as Easy Ensemble, Xtreme Gradient Boost and Light Gradient Boost classifiers to improve. EasyEnsembling techniques with RandomForest classfier uses a bag of balanced boosted learners. The balancing is done by random under sampling of the highly populated class. For the base estimator we used RandomForestClassifier of 5 estimators. Using 200 samples of random forest classification makes EasyEnsemble Classifier a good ensemble learner. But we could not make the model larger because it would overload the system. Random Forest is a bagging algorithm that we tried. So next we tried boosting algorithms. XGBClassifier was our first choice being a very well knows classifier and for unique boosting technique different from normal gradient boosting. XGB is a an advanced form of gradient boost decision tree algorithm. So just like gradient boost XGBoost is used to predict the residuals using decision trees. This means after every new tree is constructed the overall residuals tend to be smaller than previous. So this helps generalize the model well and converge faster with better accuracy. The score was the best for this boosting technique. We also used a newer algorithm LGBMClassifier by Microsoft. Both of these algorithms come under the common umbrella of GBDTs (Gradient Boosting Decision Trees). But they use tricks to improve their efficiency. Although lightGBM gave fairly good model but for our case XGboost out performed. We also explored at CatBoost classifier from Yandex. This classifier handles categorical data using its own encoding technique. It also give us very good results with default parametre. So it saves lots of user time which is used in preprocessing and also in parameter tuning. CatBoost is best fit for categorical features. But for the dataset given CatBoost doesn't perform as expected probably because of parameter value selection and produces poor result than LGBM and XGBoost.

VII. Result

For evaluation of our model, we have split the available train data into training data and testing data in a ratio of 4:1. To evaluate our model, we are using f1 score, recall, precission, accuracy, and AUC ROC score as our evaluation metric for the model. Here we are only presenting the metrics of ensembling techniqies used in our work. Table 3 shows the score of different ensembling techniques used in our work. The best of the results among three models are highlighted in the table2.

Table 2 : Results of Models

Models	Precision	Recall	F1 Score	ROC_ AUC	Accuracy
LR	0.0538	0.56	0.098	0.59	0.6215
ER	0.0535	0.60	0.098	0.60	0.5933
XGB	0.0611	0.43	0.107	0.59	0.7346
LGBM	0.0562	0.56	0.102	0.60	0.6405
CBoost	0.0580	0.46	0.103	0.58	0.7028

LR : Logistc Regression
KM : Kmean
ER : Easy Ensemble using Random Forst
XGB : Extreme Gradient Boosting
LGBM : Ligh Gradient Boosting Model
Cboost : Cat Boost

VIII. Conclusion

The results from the project clearly demonstrated an association between individual enduring personal factors and the involvement in crashes and near crashes, while the prediction of was less successful. This is likely due to a combination of the outlier and missing value in the training data and a weak association between the currently used predictor variables and individual crash involvement. However, the current driving style variables were relatively simple and, in a follow-up project, there is clearly much room for exploring whether other types of metrics representing individual characteristics, such as close following and speeding behaviors and “inattention proneness” may be more strongly associated with crash involvement. It would also be interesting to analyze more thoroughly why the present models were able to predict individual involvement in near crashes, but not crashes.

This work is a very challenging task because of missing values, imbalanced features and imbalance classes. In our work, we have performed various preprocessing to clean the data and extract information. After processing the data, we have trained different traditional models and used ensembling techniques to improve the AUC score. Proper tuning of hyperparameters were required in order to determine a model which will produce a reliable prediction. We would like to conclude that we were able to come up with an efficient model to predict wheather the customer will file an insurance claim during a period of interest based on his driving style. This work obviously enabled us to come up with a machine learning model which would reliably predict the probability of a individual will file an insurance claim during a period of interest.

To conclude, this project represented an initial exploration of applying predictive analytics models to identify unsafe drivers based on naturalistic driving data. The project generated some interesting and promising results but barely scratched the surface with regard to the possibilities of performing this type of analysis. These possibilities could be explored in a follow-up project with a larger budget and scope.

Challenges And Future Scope

The dataset provided by Porto Seguro is imbalanced in nature. Identification of features which will accurately classify the datapoints becomes a major challenge. Besides, there are many feature having high null values and highly imbalanced data. Presence of these features makes it a very challenging task to accurately predict the probability of a individual will file an insurance claim during a period of interest.

In our work, we hava tried to compare the performance of different model such as Random Forest using under sampling, XGBoost, LightGBM and CatBoost for our dataset. LightGBM model gave us the best result among all of these. As future scope, one can tune the hyperparameters of the model to acheive better results for our dataset. Also, advanced machine learning techniques using neural networks can be used to acheive better results.

Acknowledgement

We would like to thank Professor G. Srinivas Raghavan and Teaching Assistants of the course. Vibhav Agarwal, Tejas Kotha, Tanmay Jain, Shreyas Gupta, Arjun Verma, Nikhil Sai Bukka,, Divyanshu Khandelwal, for giving us the opportunity to work on the project and help us whenever we were struck by giving us ideas and resources to learn from. We would also like to thank all other teams in Kaggle for being a great competitor and setting a benchmark time by time for the rest of us which acted as a driving fuel for us to constantly work hard and surpass them. We would gladly say that we had a great learning experience while working on the project. Leaderboard was great motivation to work on the project and the competition forced us to read up various articles and papers which gave us ideas and enthusiasm for the project.

References

[1] Browne, M. (1992). Evidence of Adverse Selection in the Individual Health Insurance Market. The Journal of Risk and Insurance, 59(1), 13-33. doi:10.2307/253214

[2] Cummins, J., Grace, M., & Phillips, R. (1999). Regulatory Solvency Prediction in PropertyLiability Insurance: Risk-Based Capital, Audit Ratios, and Cash Flow Simulation. The Journal of Risk and Insurance, 66(3), 417-458. doi:10.2307/253555

[3] Lee, S., & Urrutia, J. (1996). Analysis and Prediction of Insolvency in the Property-Liability Insurance Industry: A Comparison of Logit and Hazard Models. The Journal of Risk and Insurance, 63(1), 121-130. doi:10.2307/253520

[4] Bagdadi, O., & Varhelyi, A., 2011. Jerky driving—An indicator of accident proneness? Accid. Anal. Prev. 43 (4), 1359-1363. doi:10.1016/j.aap.2011.02.009

[5] Breiman, L. 1984. Classification and Regression Trees. Chapman & Hall/CRC

[6] Dahlen, E. R., White, R. P., 2006. The Big Five Factors, Sensation Seeking, and Driving Anger In the Prediction of Unsafe Driving. Personality and Individual Differences, 41 (5), 903-915. doi:10.1016/j.paid.2006.03.016

[7] De Winter, J.C.F., Dreger, F.A., Huang, W., Miller, A., Soccolich, S., Machiani, S.G., Engström, 2018. The relationship between the Driver Behavior Questionnaire, Sensation Seeking Scale, and recorded crashes: A brief comment on Martinussen et al. (2017) and new data from SHRP 2. Accid. Anal. Prev. 118, 54–56. doi:10.1016/j.aap.2018.05.016

[8] "RandomForestClassifier Documentation". [Online]. Available: "https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html".

[9] Random Forests in XGBoost, "XGBoost Documentation". [Online]. Available: "https://xgboost.readthedocs.io/en/latest/tutorials/rf.html"

[10] LightGBM using Python API "LightGBM Documentation". [Online]. Available: "https://lightgbm.readthedocs.io/en/latest/Python-API.html"

[11] CatBoostClassifier using Python Package. [Online]. Available : "https://catboost.ai/docs/concepts/python-reference_catboostclassifier.html"