Semester Project

Soutri Mukherjee[1], Ashay Sawant[2]

Abstract
Most of the insurance providers calculate the rate based on the demographics and accident history. Failure in accurately predicting insurance claims, results in extravagant insurance cost. Porto Seguro, one of the largest auto insurance companies in Brazil, has collaborated with Kaggle's Machine Learning Community to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year.

Keywords
Regression — XGBoost — Random Forest

[1] *Computer Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
[2] *Data Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
***Corresponding author**: soumukh@iu.edu, ahsawant@umail.iu.edu

Contents

Introduction

Car insurance costs way too much now-a-days. Car insurance providers unable to estimate the possible claims for next year, hence car owners end up paying huge amount. Porto Seguro, one of the largest auto insurance companies in Brazil, has collaborated with Kaggle's Machine Learning Community to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year. More perfect model will help in accurately estimating the insurance rate which will be much less than what insurance provider currently charge. We can find kaggle page for the competition over here.

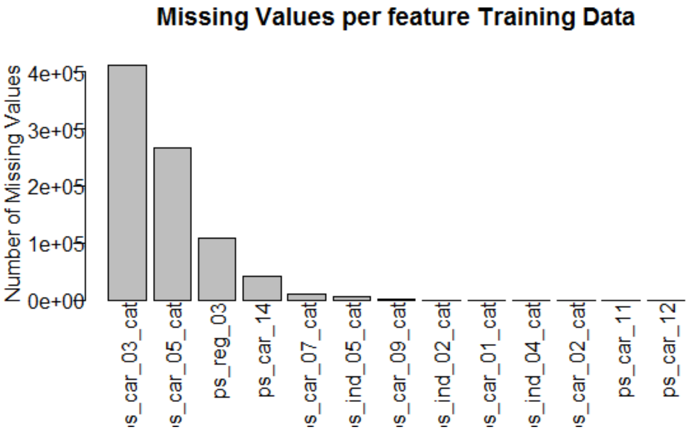
Kaggle has provided each one of training dataset, testing dataset, sample submission file. The training dataset has 59 variables which includes columns like id -an identifier of every row, target -prediction variable for training dataset and 57 binary, categorical and continuous variables. Training dataset contains 58 variables, everything except target variable (prediction). We have used two kernels from Kaggle, which helped us with data processing. Both of the Kernels use different approaches to preprocess data and estimate the target variable which tells us about whether a driver will file a claim next year.

1. Algorithm and Methodology

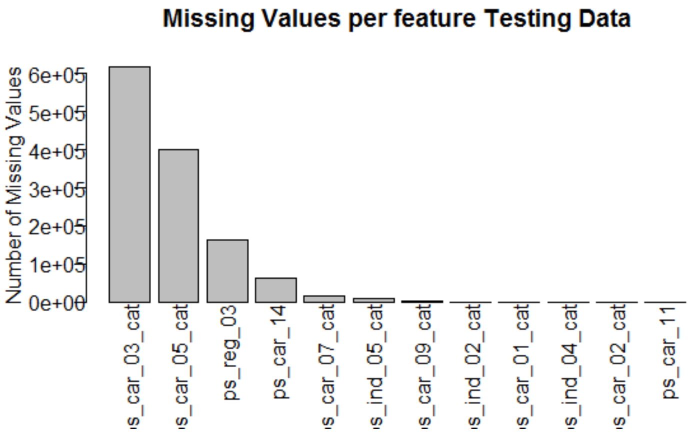
1.1 Logistic Regression

To predict the target variable using the logistic regression, we will have to pre-process the data by converting all the features to categorical values and removing NULL values.
Our pre-processing includes the following steps:

- Initially, we have calculated the Number of NA's per feature of Training and Testing Dataset. pscar03cat, pscar05cat, psreg03 are top three features with most number of NA's. We have removed these columns from the dataset that contains maximum NA values. Following histogram plot shows the features with NA values and respective number of NA fields for training and test dataset. From the above plot we conclude that the first three columns have the maximum number of NAs. Thus, Since the relevance of such kind of columns in our data set in bare minimum, we remove these columns from training data as well as our testing data.



- At the second stage, we are replacing the rest of the NA values in our new dataset. To do so, we have con-



sidered the MODE value for categorical features and mean value for continuous features. We are using mode and mean values as a replacement because, we were looking for some central value that might represent our column without majorly impacting the overall behavior of our dataset.

We have implemented Logistic Regression over the processed data. Since, we have converted our data into categorical dataset, we have decided to use Logistic Algorithm. We are trying to fit the model over training dataset and then predict the target variable for testing dataset using the fitted model.
After submitting the prediction on Kaggle, we have received 0.250 public score and ranked in top 3000 submissions.

1.2 Random Forest

Initially, we have used XGBoost Algorithm. XGBoost stands for Extreme Gradient Boosting algorithm. We started of with calculating the correlation among all features in dataset. Using this correlation, we concluded that there is strong correlation between, all the individual binary columns, calc binary columns. We combine them as a single feature by summing it up. Then we have applied the XGBoost algorithm on processed data. On the processed data, we have applied Random Forest Algorithm to make prediction on testing Data.
After submitting the prediction on Kaggle, we have received 0.267 public score and ranked in top 3000 submissions.

2. Summary and Conclusions

Therefore, we have partially solved our Porto Seguero's problem. From the above, we summarize our work as follows: We have analyzed the dataset and have performed Data Cleaning and modeling at the initial stage.Then we have applied Logistic Regression and Random Forest Algorithm to predict the target. In the end, we concluded that Random Forest performed better than Logistic Regression.