**HackStat 2.0 — Kaggle Competition**
**Team: fastmosquitonet150**
Team Members: Ramith Udara Hettiarachchi (Team Leader), Mohamed Afham, Kithmini Kaumadi Herath, Hasindu Kariyawasam, Udith Haputhanthri
**University of Moratuwa** Submission: 22/09/2019

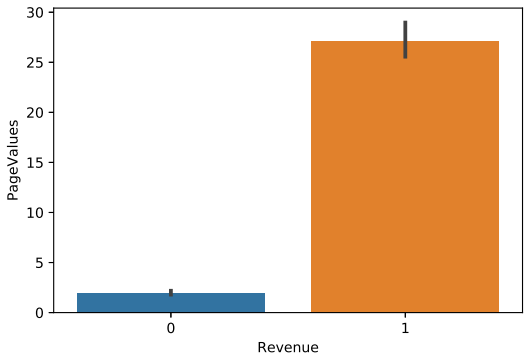
# 1 Introduction

This is a supervised learning binary classification problem. We were provided an Online Shoppers Purchasing Intention Data set which contains 8858 'Revenue:0' labels and 1622 'Revenue:1' labels.

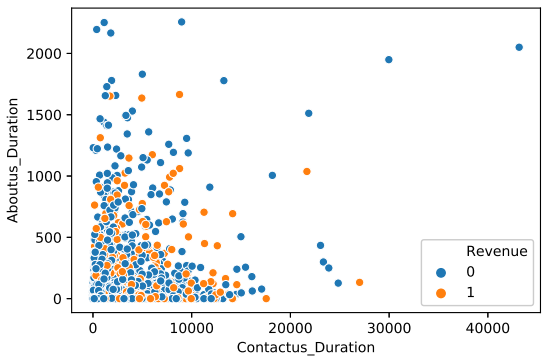
# 2 Methodology

## 2.1 Exploratory Analysis and Feature Engineering

During the exploratory analysis stage we observed that, PageValues variable had the most correlation with the Revenue and this is elaborated in Figure 1(a) followed by BounceRates and ExitRates. Therefore, we decided to discretize those features according to the predicted probabilities of Decision Tree Classifier Method. As seen in Figure 1(b). there are a considerable amount of outliers. We decided that the main reason for this behaviour is the variance of those features, since we observed a similar distribution from the test set as well. From this observation we compressed the distribution by taking the logarithmic value of those features.




(a) Page Values Vs. Revenue




(b) Distribution of Revenue w.r.t Contactus_Duration and Aboutus_Duration

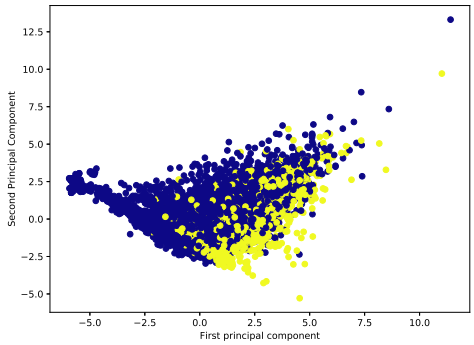
Figure 1

We also used the "Label Encoding" technique to encode the categorical data since there were 7 categorical features.

## 2.2 Model Selection

We decided to consider a tree based classifier. Out of such classifiers we chose RandomForest classifier. The reason was, after plotting data points using Principal Component Analysis(PCA) we obtained the plot Figure 2(a). It was clear that a distance based classifier such as Support

Vector Machines(SVM) and K-nearest Neighbours is not suitable for this task. Also it was observed that a cross validation score higher than 0.90 cannot be achieved.



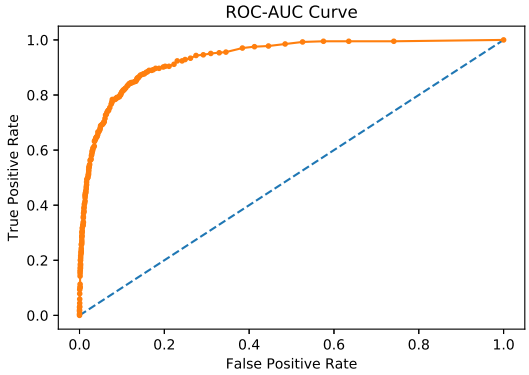

# 3 Results

## 3.1 Confusion Metrics

Table 1: Confusion Matrix

| | |
|---|---|
| 2126 | 90 |
| 158 | 246 |

## 3.2 Class-wise error rate & ROC-AUC Curve




AUC: 0.926

| | precision | recall | f1-score | classwise error rate |
|---|---|---|---|---|
| 0 | 0.93 | 0.96 | 0.94 | 4.0614% |
| 1 | 0.73 | 0.61 | 0.66 | 39.1089% |

# 4 Conclusion

After following the above methods we were able to obtain an accuracy of 0.90972 from our model in the test data predictions. We have included our source code here.