

# hâckStat

Team Leader's Name

- Nishan Chathuranga

Team Members Names

- Hansani Wathsala
- Hirushi Ekanayaka
- Sachini Wijesinghe
- Thihari Wijesinghe

University / Institute

- University of Moratuwa

Team Name, Kaggle User Name and Display Name

AsyncWave



## Introduction

There are several supervised learning classifiers to predict class variables according to a given set of dependent variables. Those classifiers give different AUC values (Area under the ROC curve) and in this study, we found the best classifier by selecting the classifier which has the highest AUC. We predicted the class using data of visitors of a website. We used 28 independent variables for the prediction. The class variable explains whether the customer would be a revenue generation customer or not. We used 16 968 data points for the prediction and 67% of the data set was used to train the models and 33% of the data set was used to test the models. Also, there are class imbalance in the data set. As per the correlation matrix for training data, it has been observed that majority of the provided features do not support the prediction.

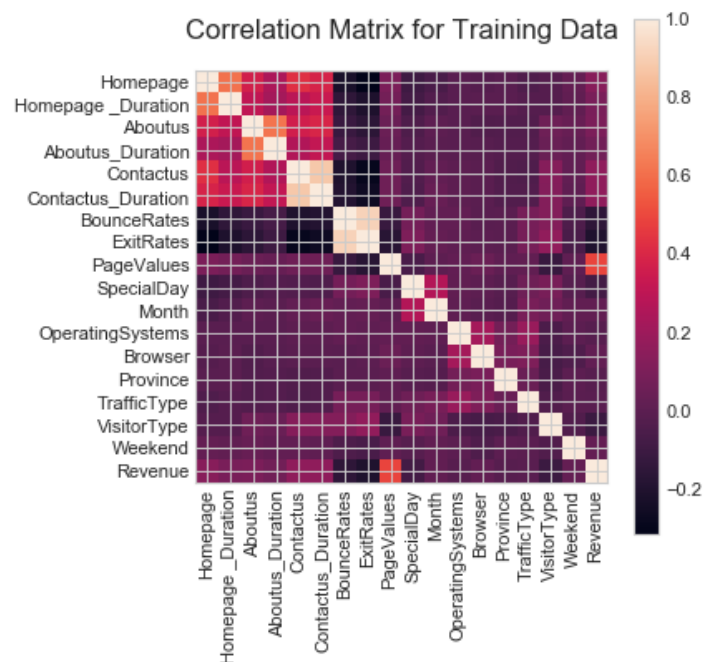


Figure 1. Correlation matrix for training data

## Methodology

Initially we tried K-Nearest Neighbors, Gradient Boost and Random Forest Classifiers, out of those, Random Forest Classifier performed best giving us **0.936** accuracy. There after decision tree-based models were chosen to move forward. We were able to get a clear idea about the features after looking at the feature importance provided by random forest classifier. Visit <https://github.com/nishanc/hackstat-2.0> for code.

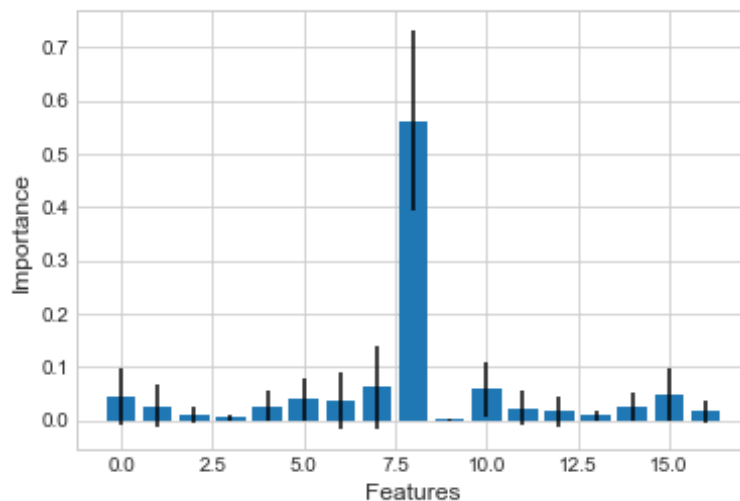


Figure 2. Feature impotence before normalization

We have tried multiple models in Azure ML Studio to see which one performs best, such as, linear regression, two-class decision forest, two-class logistic regression, two-class support vector machine and many others, out of those most promising results were given by **Two-Class Boosted Decision Tree**, with parameters,

- Maximum number of leaves per tree - 36
- Minimum number of samples per leaf node - 7
- Learning rate - 0.333
- Number of trees constructed - 182

A boosted decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction. Visit <https://gallery.cortanaintelligence.com/Experiment/First-round-competition-for-hackStat-2-0-by-team-AsyncWave> to see the published experiment created using Azure ML studio.

Also, there is a class imbalance in the dataset, number of positive data points are substantially less than negative data points.

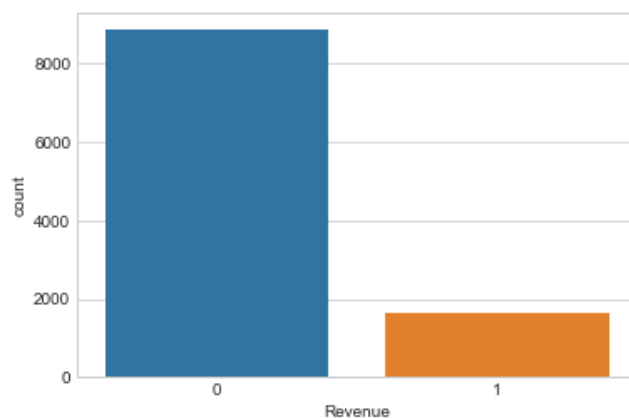


Figure 3. Count of class variable

To overcome this problem SMOTE - Synthetic Minority Over-sampling Technique has been used with 1 nearest neighbor as parameter, replaced missing values with mean, converted **Month** and **VisitorType** to indicator values and normalized all other features using logistic transformation method before feeding the dataset into the model.

Homepage	Homepage_Duration	Aboutus	Aboutus_Duration	Contactus	Contactus_Duration	BounceRates
1	10	0	0	9	700	0
2	15	0	0	10	894.666667	0
1	85	0	0	14	306.5	0
5	175.1	0	0	26	615.559524	0
2	25	0	0	5	40	0.066667
1	113.75	0	0	26	1905.078571	0.022667
0	0	0	0	53	1624.320952	0.009804
9	102.866667	3	339.5	29	693.254167	0
4	49.3	0	0	12	524.969231	0

Figure 4. Before normalizing

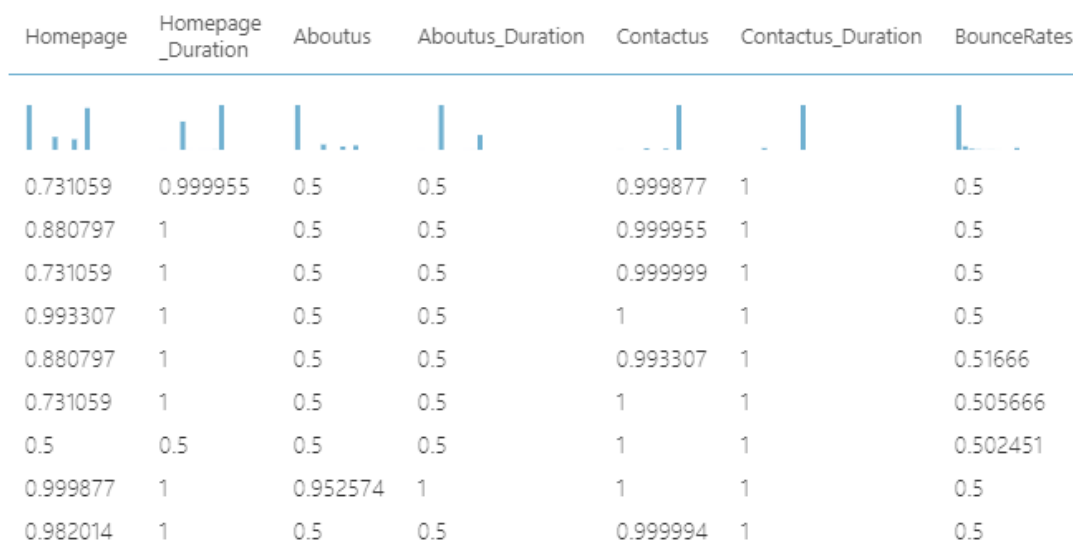


Figure 5. After normalizing

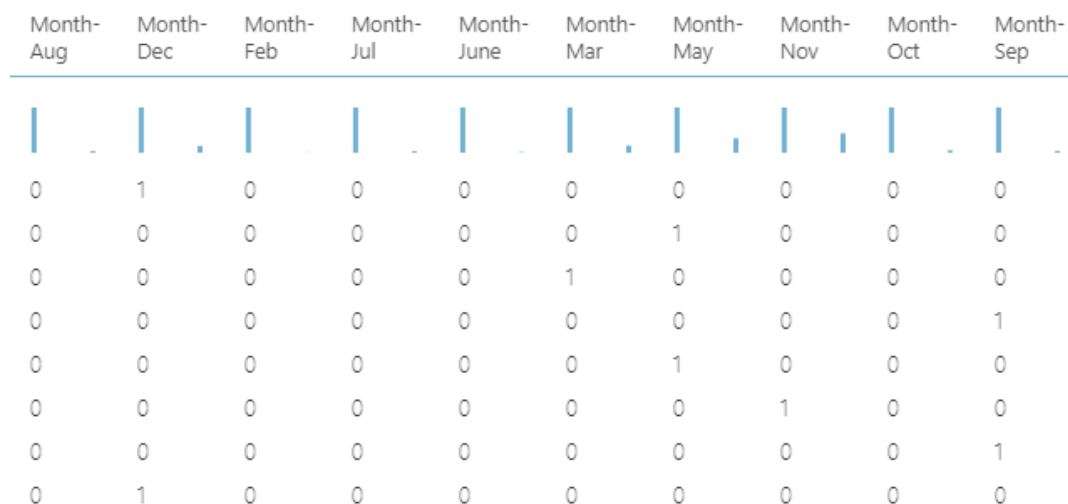


Figure 6. Converting month to indicator values

## Results

We got **0.912** as accuracy and **0.969** as AUC (Area under the ROC Curve). Because SMOTE is used, we were able to increase the number of true positives and true negatives, which in terms helped us to increase accuracy and recall.

True Positive	False Negative	Accuracy	Precision
1857	132	0.912	0.884
False Positive	True Negative	Recall	F1 Score
243	2010	0.934	0.908

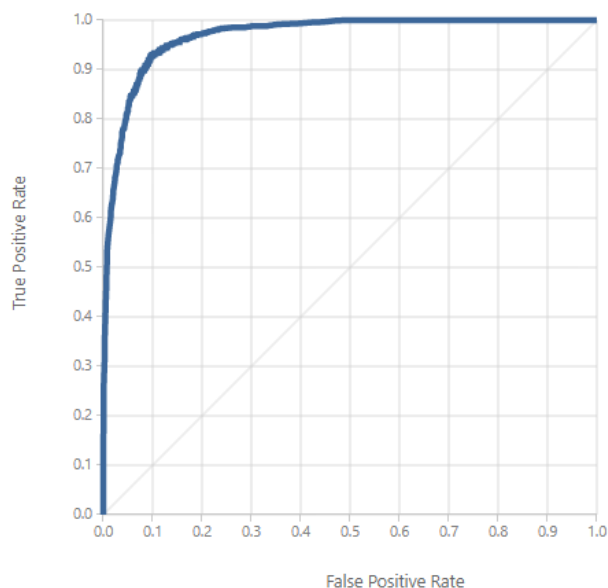


Figure 7. ROC Curve

## Conclusion

According to the online shoppers purchasing intention data set, it would be interesting to understand what the important factors that affect to generate a revenue are. Linear regression, two-class decision forest, two-class logistic regression, two-class support vector machine and others are used as classification models. Based on the overall performance, we find the best model is two-class boosted decision tree, which gives an accuracy as 0.912 and AUC (Area under the ROC curve) as 0.969