Predictive Analysis - Marketing Campaigns Of A Banking Institution

Submitted by: Saad Waseem (saad.waseem@abo.fi)

Introduction

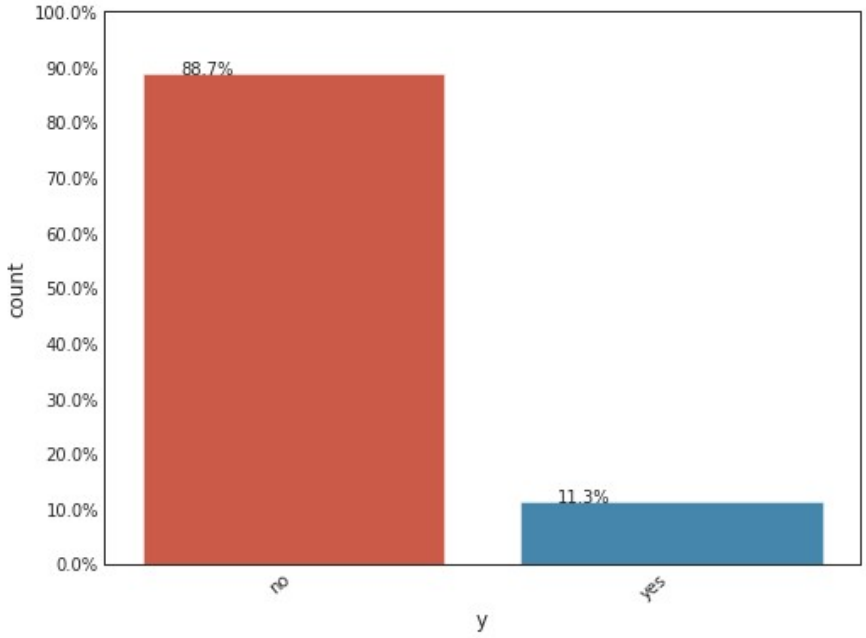
Today, predictive analytics applications became an urgent desire in banking institutions. Predictive analytics use advanced analysis techniques that encompass machine learning implementation to extract valuable insights that could uncover a potential revenue growth for businesses.

This study undertakes a neural network (Scikit MLPClassifer) based approach to devise a predictor that could answer potential outcomes of a banking campaign before it takes place. Concretely, I performed exploratory data analysis of the historical campaigns data to understand relationships between attributes and their distribution. In order to benefit historical data I transformed the data into more meaningful units that could be fed into machine learning model.

The developed model achieved above 90% accuracy.

Exploratory Data Analysis

It is evident from the below graph that the historical dataset of campaigns is imbalanced as it contains more than 88% samples when the customers said ‘no’ and only 11% samples when customers said ‘yes’.



As it is mentioned in the given dataset duration has a strong correlation with output variable it is essential that different features should be plotted against duration to understand contributing attributes towards success or failure of a campaign. Graph in Figure -1 shows shorter duration calls are not contributing positively towards a campaign’s success. Similarly self employed and management job

roles are spending more time on call and eventually contributing positively towards a campaign's success.

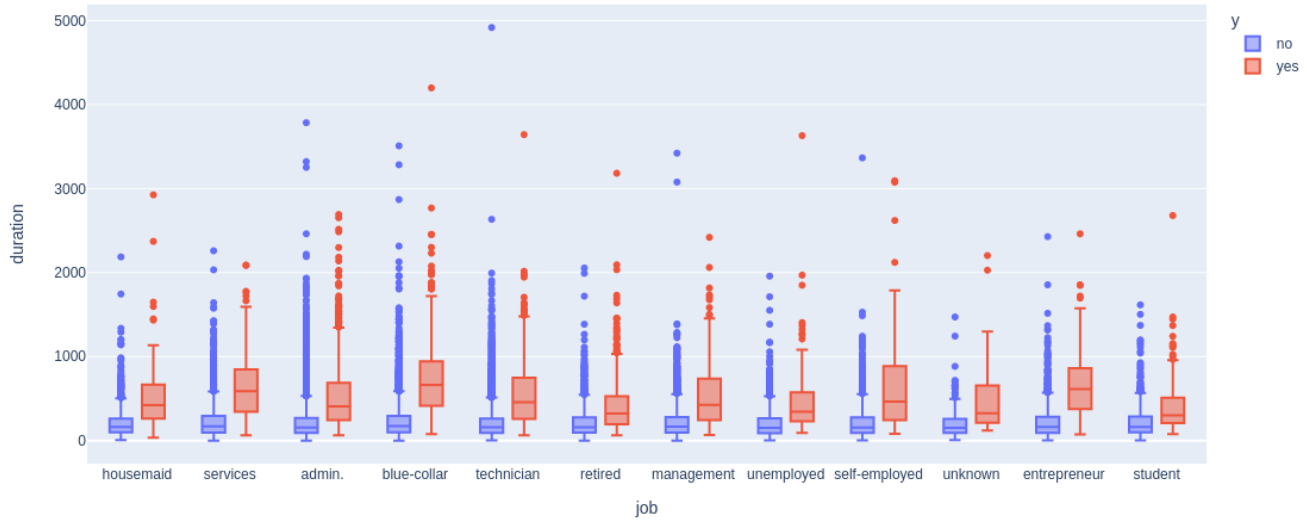


Figure -1

As different age groups respond differently to telemarketing campaigns so I plotted the graph in Figure-2 that shows people in age group 30 – 45 has positive trend towards success of a campaign it needs further investigation because the same age group responding negatively in close proportions.

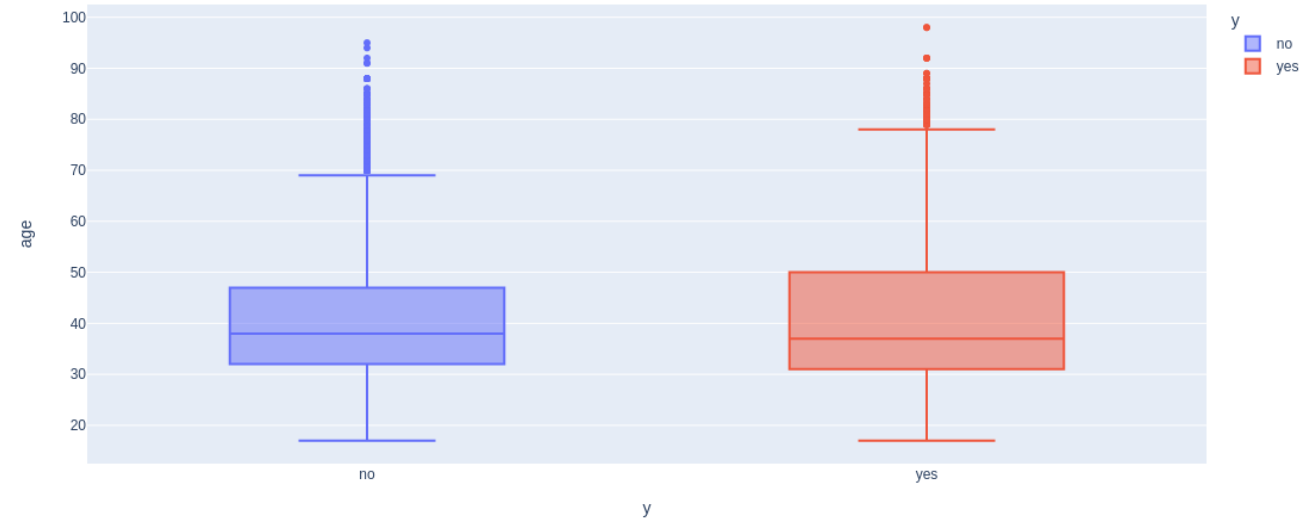


Figure - 2

Time is also crucial when the campaign was launched hence I considered plotting a graph between when the campaigns were launched as shown in Figure - 3.

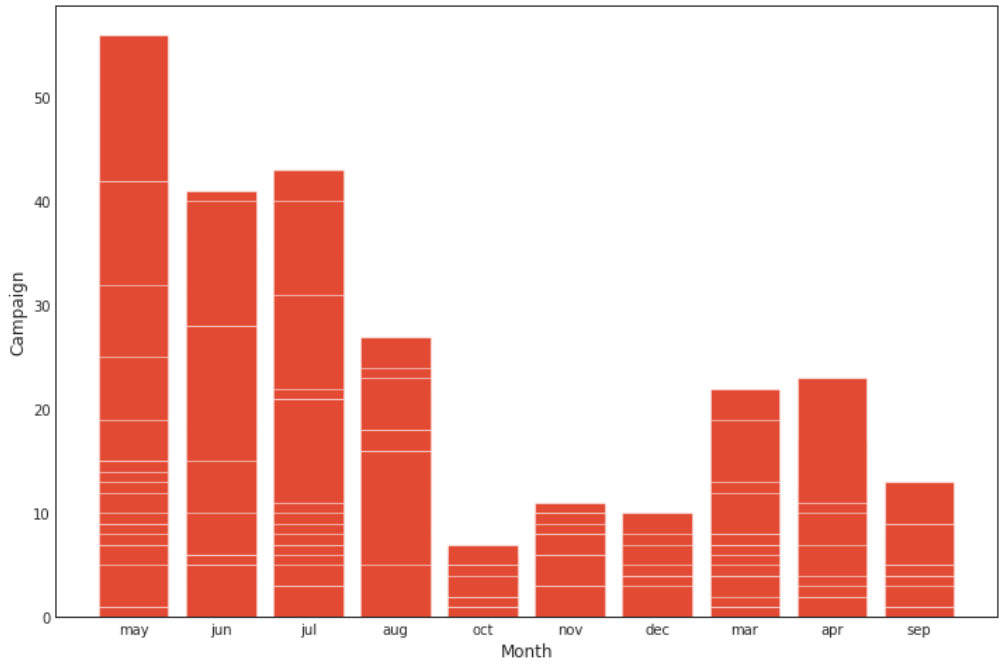


Figure - 3

The scatter plot in Figure – 4 explains the distribution of campaign's success and failure over all campaigns and how call duration started to drop after first 10 campaigns.

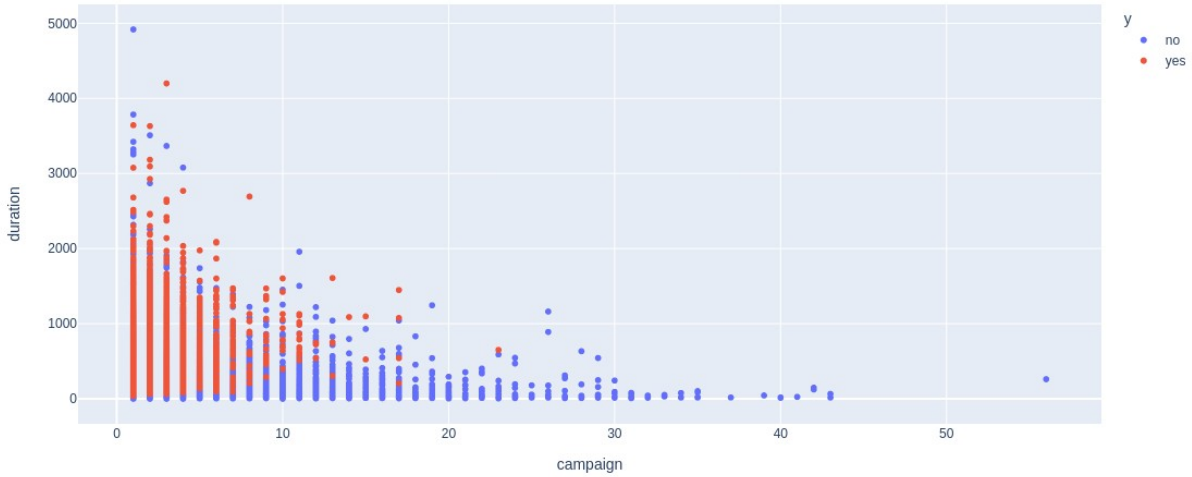


Figure – 4

As the number of features are significantly large so we plot below bar charts to determine which attribute values are showing a positive trend towards campaigns success as shown in Figure - 5

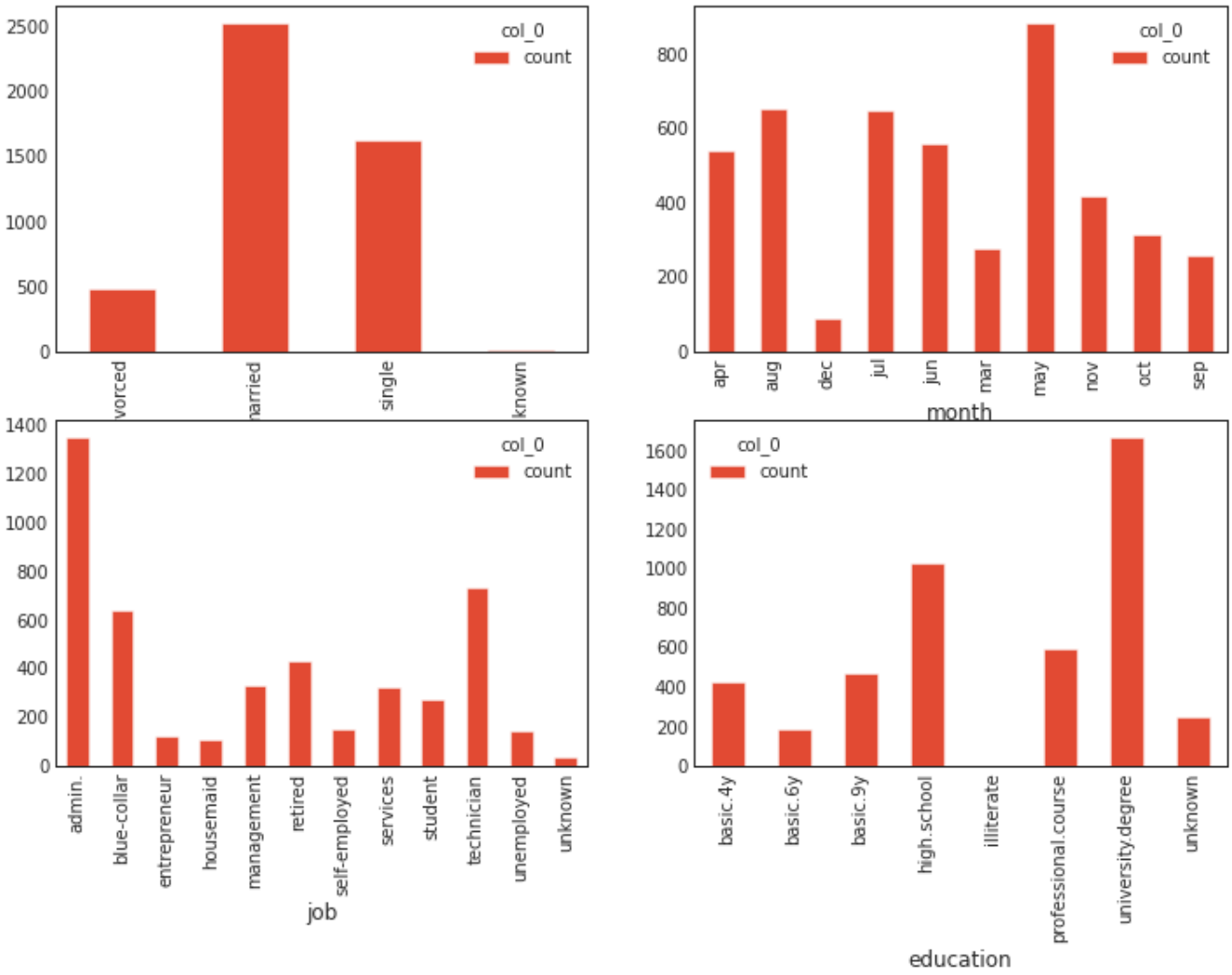


Figure - 5

Data Preparation

During data processing phase I performed boxplot analysis to detect outliers. Three of the features namely Age, Duration and Campaign have outliers. In order to fix these outlier values I employed Interquartile range (IQR) method to replace their values.

Standardization

Numeric feature of the dataset have been processed using zero-mean unit-variance standardization to keep numerical features on the same scale.

Data Encoding

The data contains various categorical variables that must be treated before they go into modeling phase. I applied two types of encoding schemes as explained below:

Label Encoding

Month, Day of week, housing, default, loan, target variable ‘y’ job, education and marital status were label encoded using custom values in order not to introduce more features and use the existing ones.

One-Hot Encoding
contact and poutcome variables were one-hot encoding hence more features were introduced to evaluate their importance and to train model.

Feature Selection
Feature selection is performed using sklearn ensemble tree based estimator ExtraTreesClassifier that provided me with importance score for each feature. Fourteen (14) most important features except duration were selected for model training.

Importance Scores

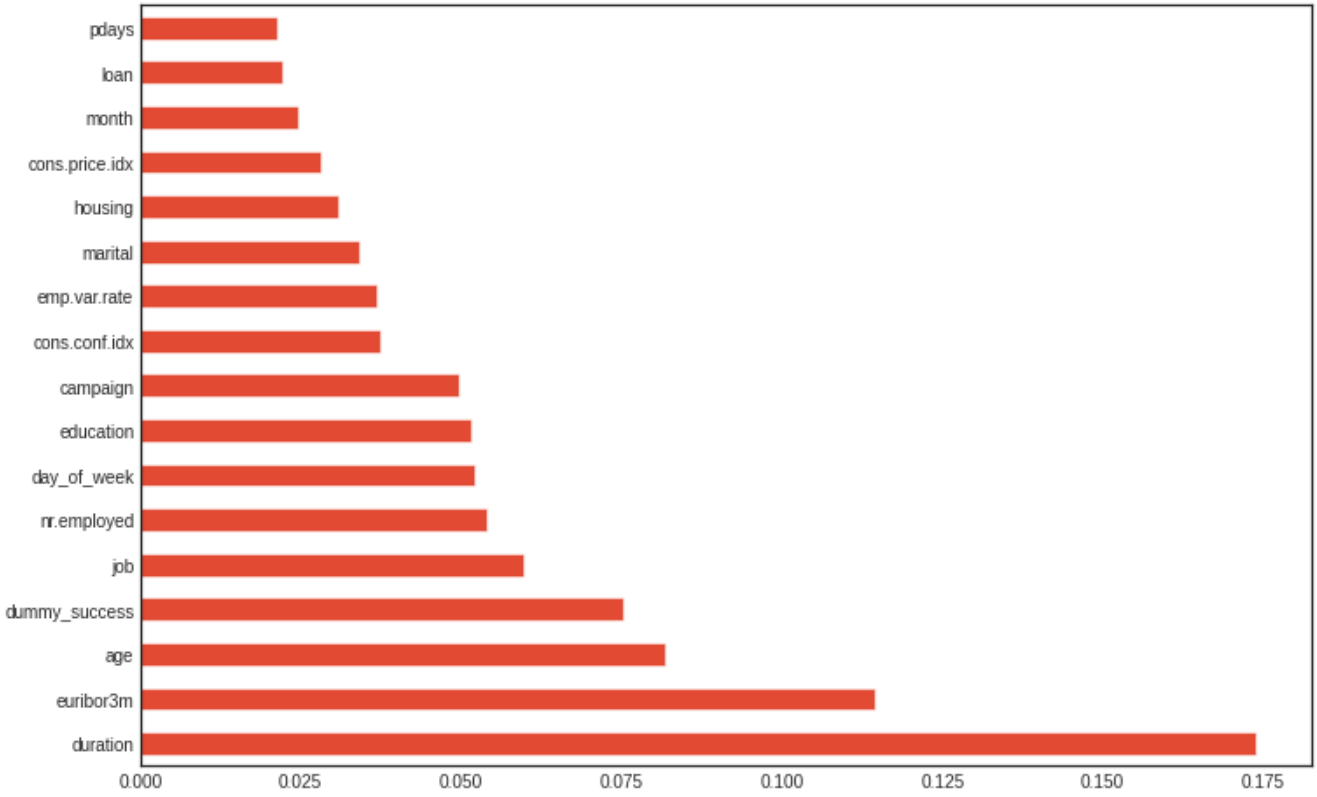


Figure - 6

Modeling
I chose sklearn's Multi Layer Perceptron (MLP) with 2 hidden layers containing 5 and 2 nodes respectively as shown in Figure – 7

Network Architecture

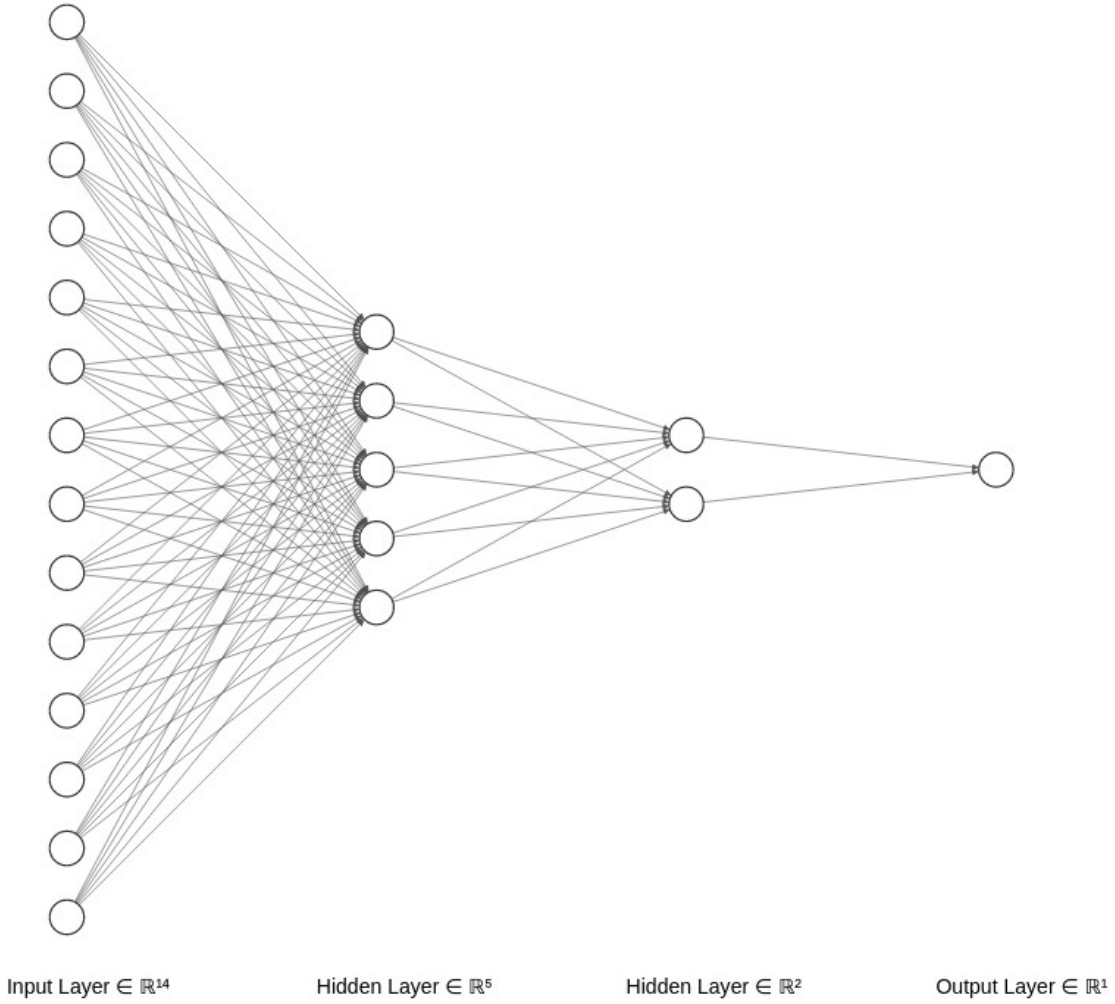


Figure - 7

Split Data

After preprocessig data, it has been divided into training and testing sets using 80 – 20 rule so that model can be trained on different data and unseen samples will be used for model evaluation.

Performance Evaluation

The trained model achieved more than 90% accuracy score with an average precision of 0.92 as captured in classification report.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      6538
           1       0.74      0.26      0.39       575

    accuracy                           0.93      7113
   macro avg       0.84      0.63      0.67      7113
weighted avg       0.92      0.93      0.92      7113
```

ROC Curve

The area under the ROC curve also supports the accuracy scores reported by classification report.

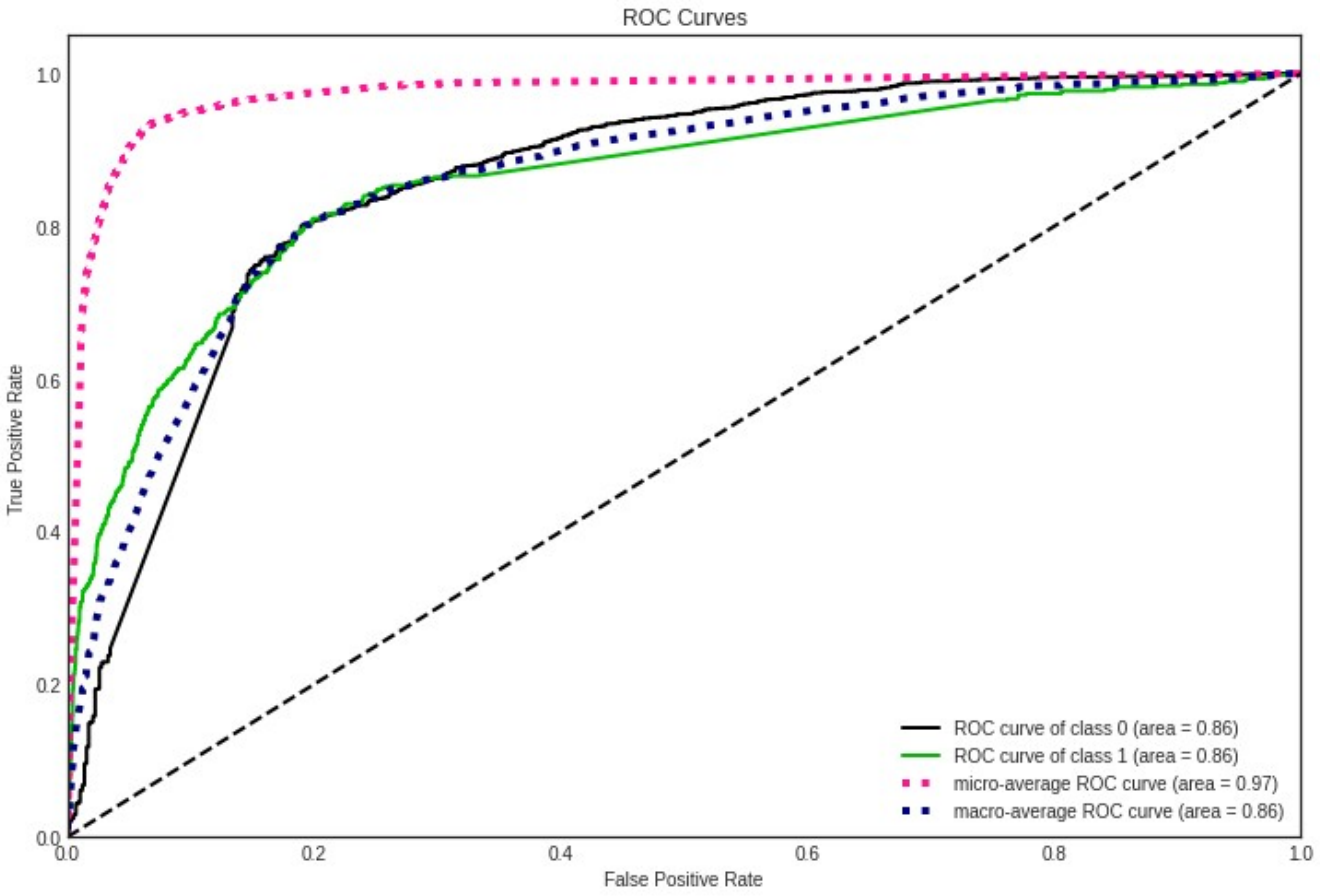


Figure - 8

Precision Recall Curve

Figure-9 shows model is behaving ideally for declined calls (represented by 0) where as the model's response is limiting because of the data availability (the number of samples with y=1).

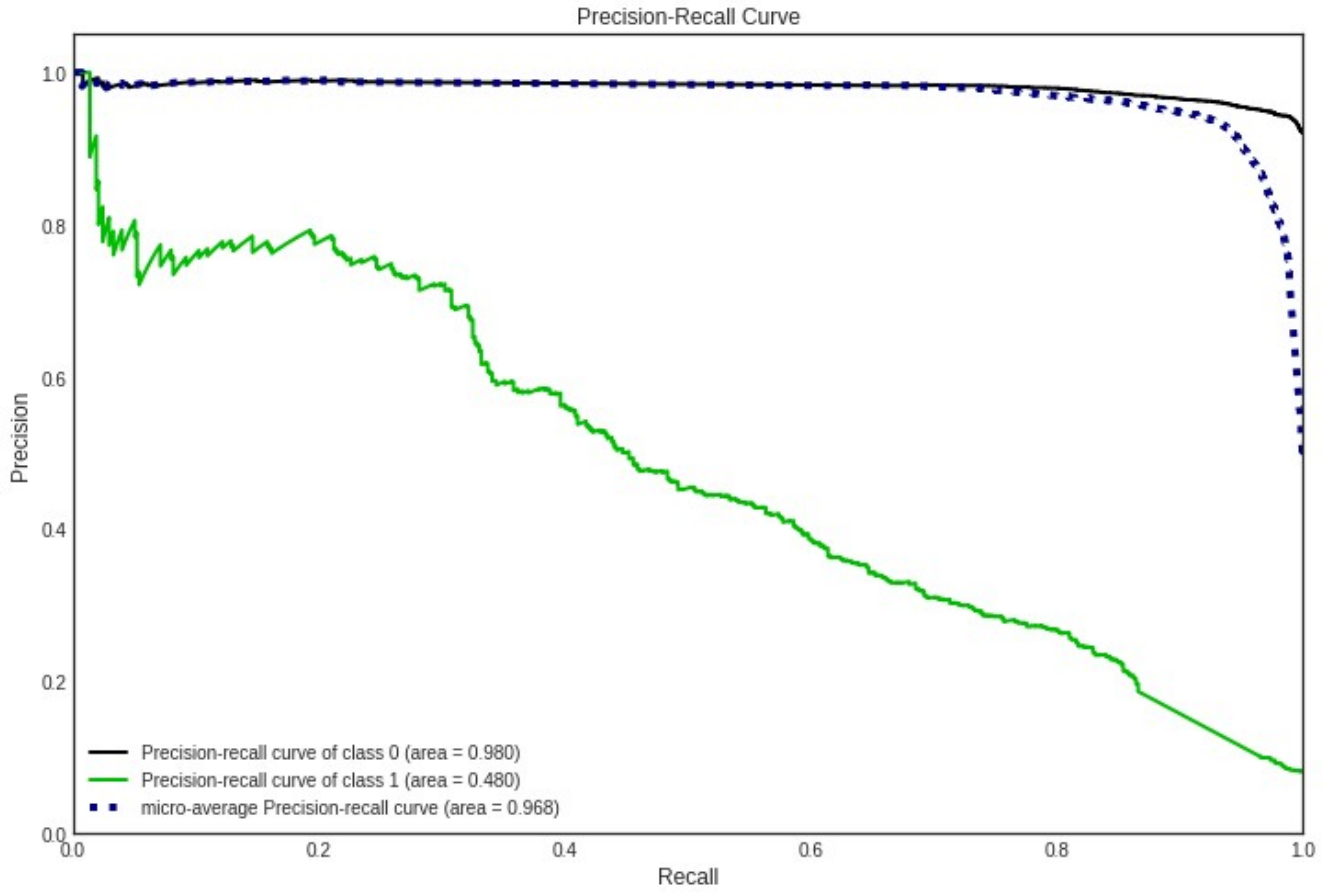


Figure - 9

Conclusion

Predicting campaign outcome with machine learning depends on the historic data and precise feature selection. In the presence of large number of features it became hard to choose the right set of features for training the model. I recommend applying tree based estimator for feature selection in order to obtain better results. It is also essential to split the provided dataset into training and test datasets for training and evaluation phases. For performance evaluation of a classifier classification report and ROC curve would be ideal in explaining the trained model behavior.