MGT-415 Executive Summary (PS3)

Authors: Sascha Frey, Kuber Malhorta

Predicting Churn amongst customers

Introduction

The aim of this project is to predict the customer churn amongst the customer base using different machine learning techniques/models. This is because they are necessary to keep the costs down in a Telecom company.

That is, these insights can be useful in focusing on retention efforts to keep track the most valuable customers, improving on resource allocation; because acquiring new customers can be much more expensive than retaining existing ones. We will see that a number of different analyses can be applied to identify the customers.

Data Loading and Pre-processing for Customer Segmentation

Since, it is a Telecom company, it provides contractual services, that is customers avail their services for intervals of time on a prepaid/postpaid contract. In our analysis, we see that 'Tenure' and 'Monthly Charges' are the most important in understanding the customer churn amongst our total customers. We also see that the Total Charges Column is related to both the Tenure and the Monthly Charges since TotalCharges = Tenure*MonthlyCharges.

Models and Cross Validation

We used two models Boosting Classifier Model and Random Forest Classifier Model and ran Cross-validation over them. The two models help in understanding the likelihood of churn in order to understand why customers choose to leave, which can help in initiating long-term retention strategies. Since, retention strategy exercises are expensive, not as much as gaining customers, we try on reducing the number of false-positive hits, that is minimizing the number of low-reaping customers which happen to fall under our strategy.

(XG) Boosting Classifier Model

This technique is used for the implementation of effective strategy building across multiple different customer segments for reducing churn and increasing retention. This classification model is a gradient boosting technique that typically uses multiple iterations of a decision tree classifier, combining the results of iterations. On each iteration, weaker decision trees are assigned less weight, thereby reducing their importance.

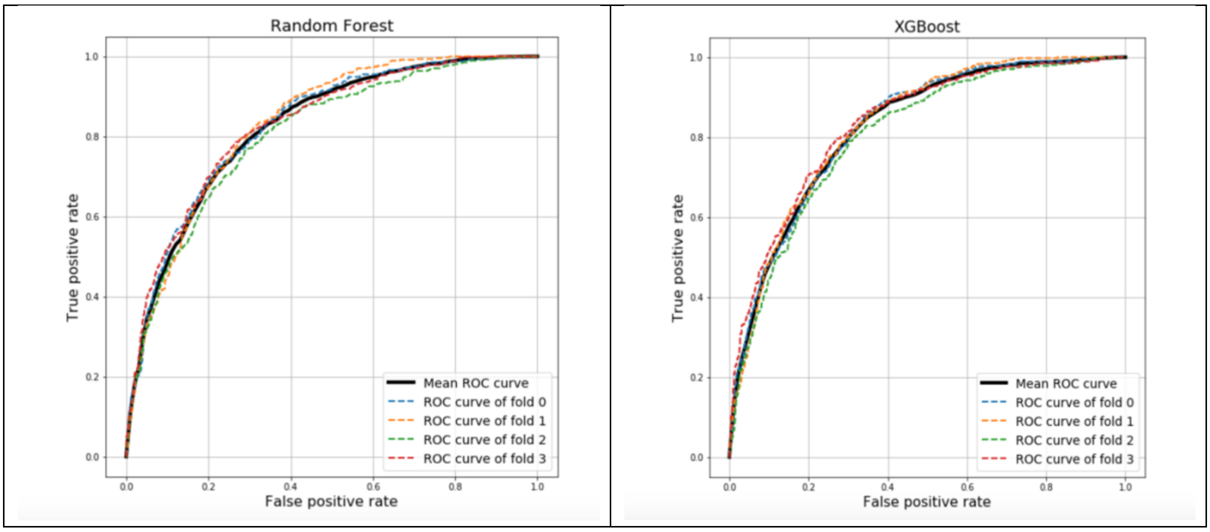

Figure 1: ROC Curves for Random Forest classifier (left) and XGBoost (right).

Random Forest Classifier

This model is better at fitting non-linear data (as we can see from Figure 1). The technique again makes use of a number of decision trees, each voting for a given class. The main difference is that there is no iterative process with “weaker” trees given a smaller vote. The main strength of this technique is dealing with features that may be correlated.

Comparison of the two Models

The 2 models have the following accuracies in predicting the churn of customers, which are as follows:-

1) XGBoost Model – 78.70%
2) Random Forest Model – 78.99%

This is quite a good result when it comes to analyzing all related tags such as Phone service, dependents, partner characteristics.

Feature Engineering/Indexing

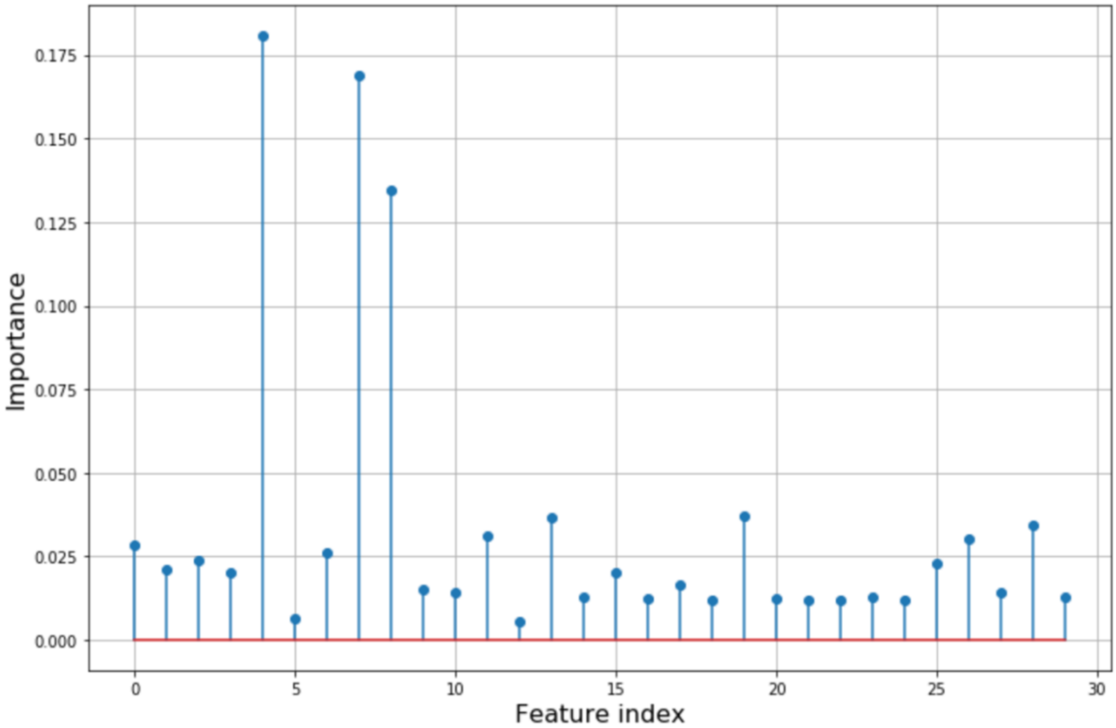

Figure 2: Importance of each of the different features in classification.

Figure 2 helps us in understanding the most important factors to consider while predicting the churn of customers:-

1) The most important factor is Tenure.
2) The second most important factor is Monthly Charges.

Conclusion

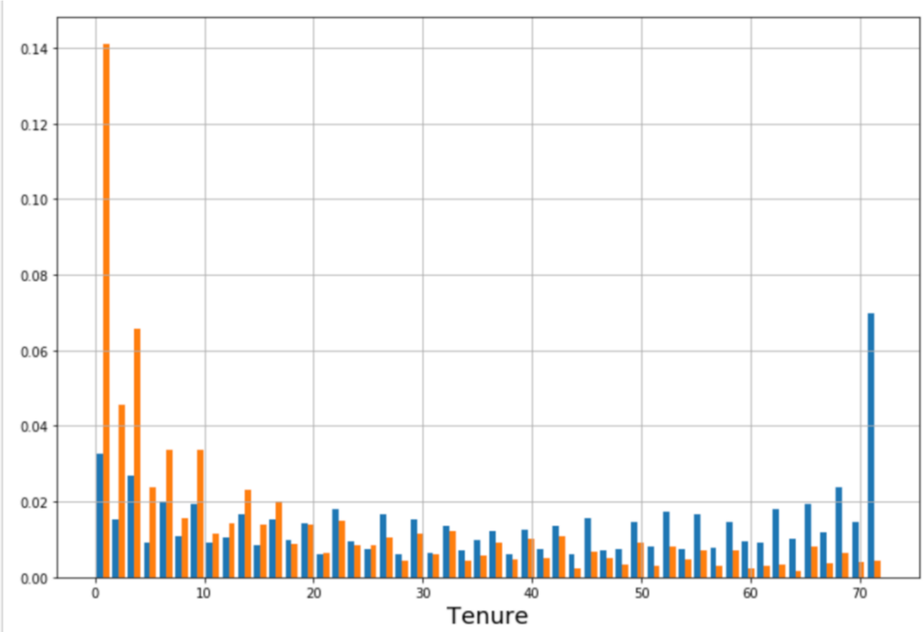

Figure 3: Frequency histogram of tenure. Churning customers given in orange, remaining customers given in blue.

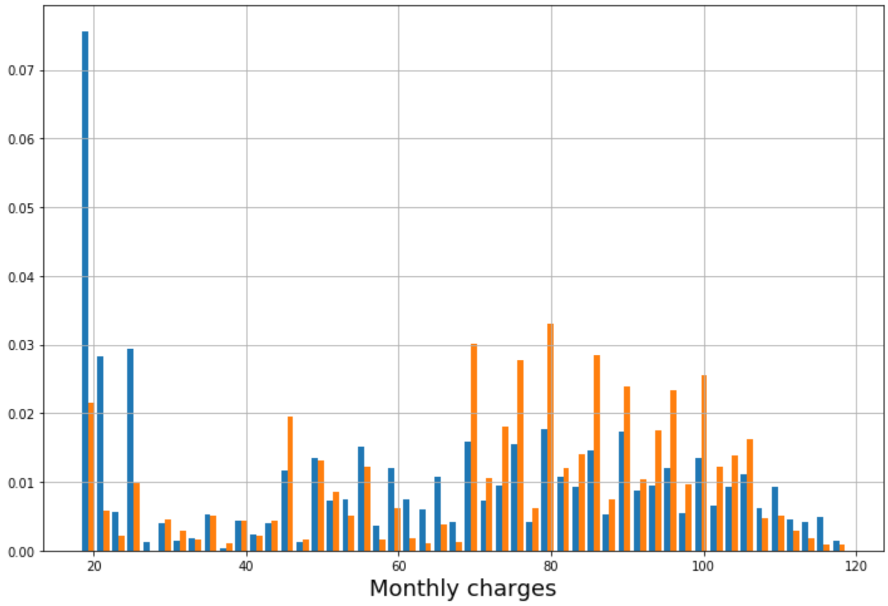

Figure 4: Frequency histogram of monthly charges. Churning customers are given in orange, remaining customers in blue.

From the histograms we conclude that:-

(1) Customers with smaller tenure are more likely to churn than long-term customers.
(2) Customers with higher monthly charges are more likely to churn.

We see that the retention intervention for different customer segments is

a) Continuous AND
b) Systemic