**Objective**:

Customer churn is one of the most important metrics used to account the percentage of users who have stopped doing business with a company within a certain timeframe. The metric is having paramount importance in most B2C companies as it costs more to acquire a new customer than it does to retain existing customers. In order to acquire new customers, the business would end up burning their cash reserves through marketing and sales activities, instead, retaining the top customers are more effective as the company has earned the trust and loyalty of old customers. Telecom industry is one such industry which is extremely sensitive to customer churn as the competing companies tend to attract phone users of competitors by aggressively pricing the data/call plans to pull the customers. Predicting the customer churn is very much essential in retaining the old customers.

**Problem Statement:**

To predict customers who are more likely to exit the service so that the company can make the necessary arrangements to stop clients from leaving their service. Analyze the data and identify factors which can contribute significantly to customer churn.

**Dataset:**

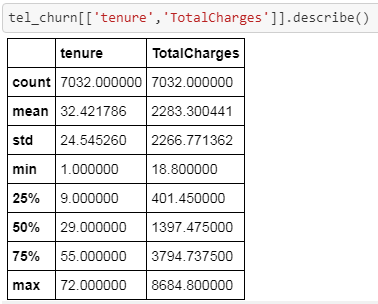
This project is based on IBM sample datasets from Kaggle. There are around 7000 customer records with 21 features (customer attributes). We will make use of the target variable ‘Churn’ to predict whether the customer will churn or not.

**Data Manipulation:**

The data was loaded to Python notebook. While analyzing the data types for each attributes/columns, it was found that the total charges column was of object data type. Upon closing reviewing the records in the total charges column, it was found that there were around 11 NAN values. Customer records having NAN populated for the total charges column was dropped from the dataframe as it was not significant when compared to around 7000 records. The total charges was converted to float just to make sure the data types is in line with the data that has been captured under that column.

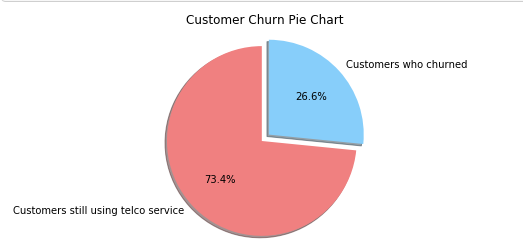
**Exploratory Data Analysis:**

The main goal of this section is to understand how the data is distributed and identify correlation between the variables. Initially, variables such as tenure and total charges were analyzed to check how the data was distributed.



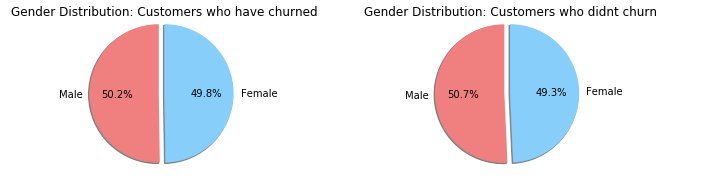
From the descriptive analysis, it was clear that the average tenure of customer was between 2-3 years and they were expected to shell out an average of 2283 dollars for the service.

As the next step, pie chart was created to visualize the percentage of customers who have ceased the service.

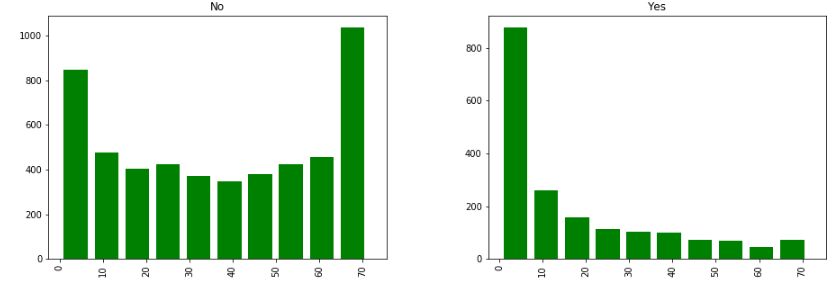


More than a quarter of customers from dataset had stopped using service of telco company. The number is still significant as the company has to invest more to compensate for the lost revenue.

The Pie chart created for the gender distribution as shown below didn’t appear to affect the churn rate for the company.



During the EDA, tenure was the variable which could provide some insights on the churn. The histogram shows the distribution of customer tenures based on churn. The bulk of customers who didn't churn appears to be having a tenure of 70 and 10 months. Whereas, the customers who did churn is bound to stop using telco service within 10 months of service



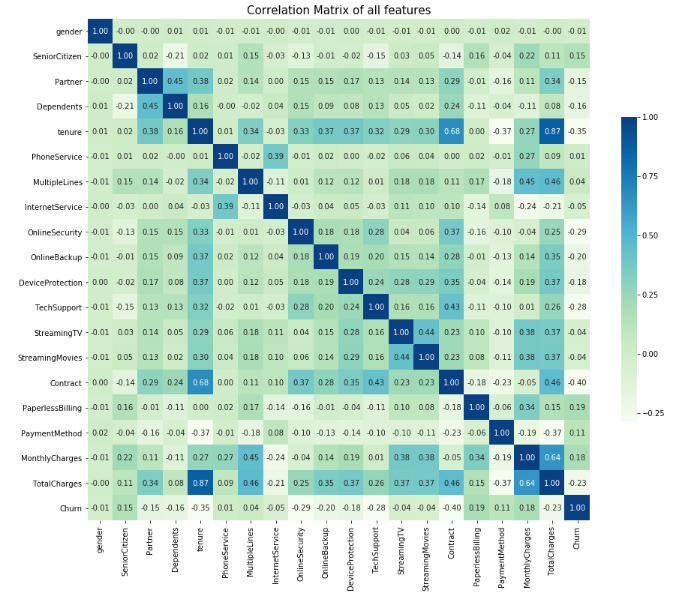
**Convert categorical to numerical data:**

For future analysis, categorical data had to be converted to numerical values with the help of label encoder library as the machine learning models will work only with numerical values.

**Correlation matrix**:

The correlation matrix was created to identify any correlation between all the features. From the matrix it was clear that there wasn’t much redundant variables in the sample. However, the tenure and total charges appeared to be highly correlated with a correlation value of 0.87.

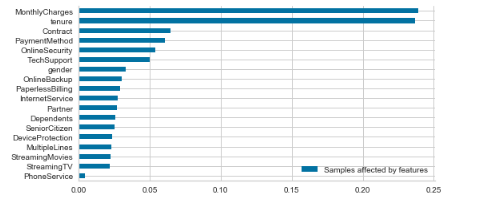
For the ML models, the total charges variable was excluded just avoid reducing the power of classification model to identify independent variables that are statistically significant



**Feature Selection:**

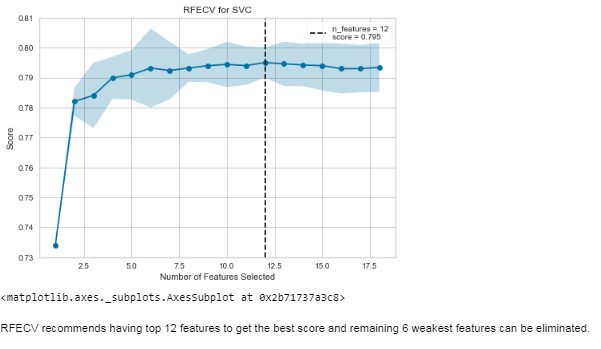
Feature selection is the important process of selecting all useful features in the data. Redundant features will unnecessarily cause performance issue ,and hence we should make sure only those features which directly affects the prediction of the model should be used for the analysis.

The features were sorted based on the importance by fitting the random forest classifier model against the training set.



Recursive Feature Elimination:

It is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. The recursive feature elimination method was used to identify whether the weakest feature can be eliminated. During the elimination process, the number of weakest features were identified and eliminated from the analysis.



**Machine Learning Techniques:**

# Three different classification models were used to identify which model will make a best fit.

# The models that will be used are

# 1) Logistic Regression

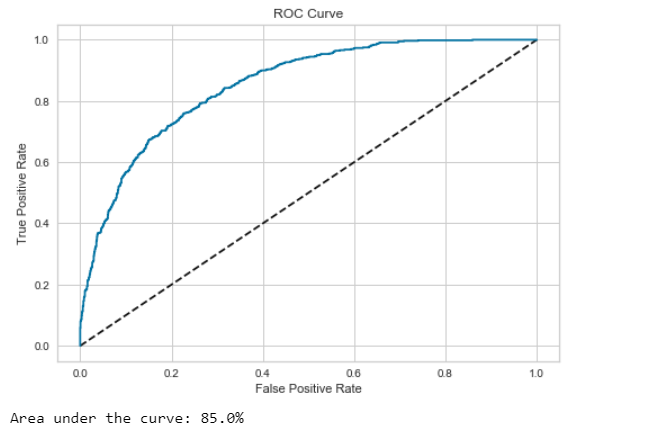
# 2) Random Forest

# 3) Decision Tree

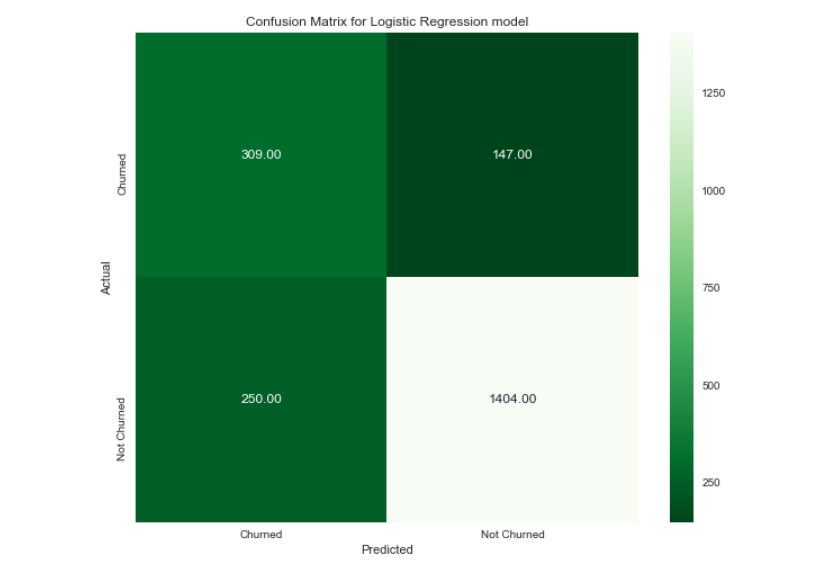
# **Logistic Regression:**

As we are working on a classification problem with a binary target variable, the logistic regression was first used to predict the customer churn. The model was fit against the train data and the accuracy for the test set was found to be around 81%. The model evaluation was done using K-fold cross validation with K=5, the CV score was found to be around 80% which confirmed the effectiveness of the model. AUC-ROC curve was plotted to evaluate the model and found 85% of the area under the curve. We did create confusion matrix to precision, recall and f-score to check the model accuracy.

AUC-ROC Curve:



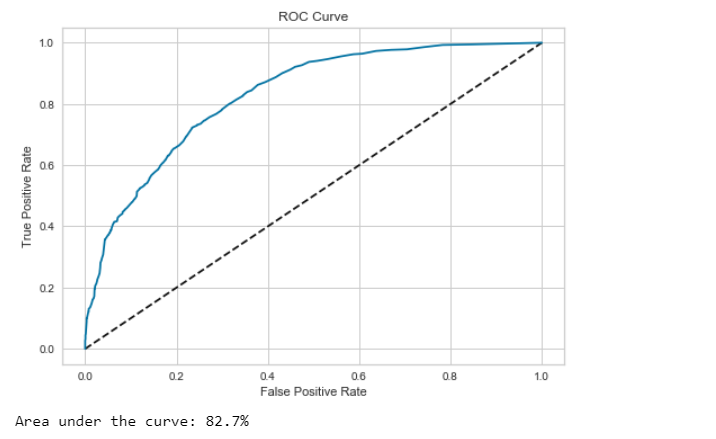
Confusion Matrix:



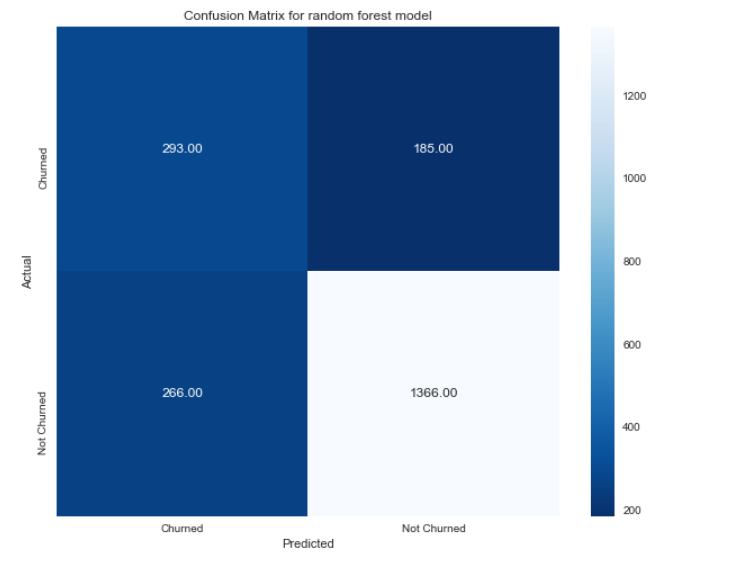
**Random Forest**:

The next model we went ahead was the Random Forest model as it uses ensemble learning and weighs certain features more important than others. Cross validation was not used as it it randomizes the variable selection during each tree split and it's not prone to overfit unlike other models. The model didn’t perform better than regression model with accuracy of 78.6% for the test set and the AUC score of 82.7%.

AUC-ROC Curve:



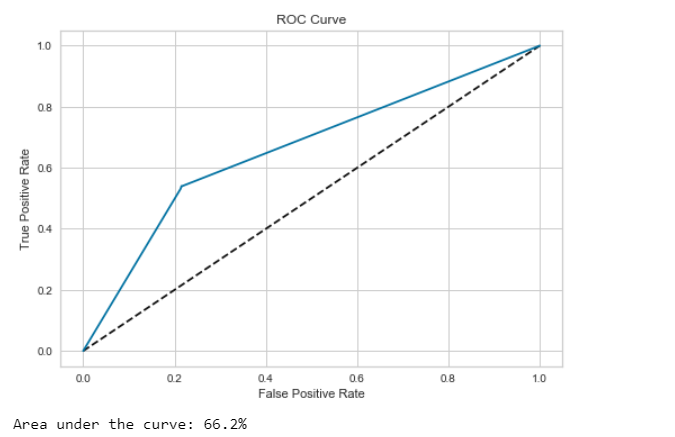
Confusion Matrix:



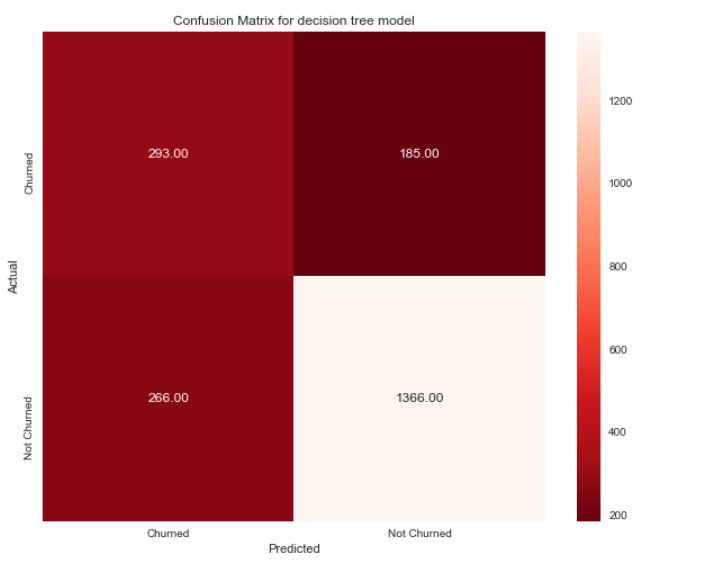
**Decision Tree:**

The Decision tree mdeo was finally used as it is known to bisect spaces to smaller region in order to get the best decision boundary. The model accuracy of the test set was 71.9% which was the least among the other two models. The model evaluation was done using K-fold cross validation with K=5, the CV score was found to be around 72.2% which confirmed the effectiveness of the model. The Area under the curve was found to be 66.2% which suggested the model may not be the right fit for the analysis.

AUC-ROC Curve:



Confusion Matrix:



**Summary:**

1) It was clear from the analysis that the features such as tenure, monthly charges, contract, payment method, online security and tech support played a major role in deciding the customer churn.

2) Customers having a tenure of less than 10 months were more vulnerable to churn.

3) The Recursive Feature Elimination method helped us in eliminating six weakest features.

4) The logistic regression model was found to be the best classification model to predict the customer churn among Random forest and Decision tree models. However, the accuracy score was not the convincing one as it had only managed to score around 81%.

5) The tuning of hyper parameter was not successful as we focused on only one parameter. Tuning needs to be performed for other hyper parameters as well to check the capability of the algorithm.