Predicting Churn Rate of Telecom Customers

Final Project Report

Group: 14

Surya Teja Karri
Sai Sruthi Kalidindi

857-348-7800 (Surya Teja Karri)
513-496-4660 (Sai Sruthi Kalidindi)

karri.s@northeastern.edu
kalidindi.sa@northeastern.edu

Percentage of Effort Contributed by Surya: 50%

Percentage of Effort Contributed by Sruthi: 50%

Signature of Surya: 

Signature of Sruthi: 

Submission Date: 12/09/2022

PROBLEM SETTING:

The principal problem we are tackling in this project is to help the telecom company retain its customers in the long run. No matter the quality of its products or customer service, every company experiences churn. Churn rate, sometimes known as attrition rate, is the rate at which customers stop doing business with a company over a given period. A persisting issue that the sellers frequently face is that sometimes they may not be aware of the features customers want. With the help of churn rate, a company can assess its weak areas and work on optimizing them thereby running the business effectively.

PROBLEM DEFINITION:

This project predicts the churn rate of customers based on several attributes such as gender, age, tenure in months, number of phone lines, etc. Making sure that customers still use our services is an important factor in a company's success and this may vary from time to time. For example, We can consider the situation where a user may or may not extend his phone plan under different circumstances such as an increase in prices. This is beneficial in identifying long-term customers for the company. This data can be used for analysis to find insights and take corresponding business decisions to retain customers.

DATA SOURCE CITATION:

The dataset was procured from a website called as Maven Analytics.
Data source citation: https://www.mavenanalytics.io/data-playground

DATA DESCRIPTION:

There are 7043 samples and 38 columns in the dataset including the Customer Status column which is the target variable. Out of the total 38 columns, 15 columns are numerical and 23 are characters. There is a slight imbalance in the target class with majority of the records belonging to true class and some belonging to false class.

Data Description:

The following columns are available in the dataset

- CustomerID - A unique ID that identifies each customer
- Gender - The customer's gender: Male, Female
- Age - The customer's current age, in years, at the time the fiscal quarter ended (Q2 2022)
- Married - Indicates if the customer is married: Yes, No
- Number of Dependents - Indicates the number of dependents that live with the customer (dependents could be children, parents, grandparents, etc.)
- City - The city of the customer's primary residence in California
- Zip Code - The zip code of the customer's primary residence
- Latitude - The latitude of the customer's primary residence
- Longitude - The longitude of the customer's primary residence
- Number of Referrals - indicates the number of times the customer has referred a friend or family member to this company to date
- Tenure in Months - Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above
- Offer - Identifies the last marketing offer that the customer accepted: None, Offer A, Offer B, Offer C, Offer D, Offer E
- Phone Service - Indicates if the customer subscribes to home phone service with the company: Yes, No
- Avg Monthly Long-Distance Charges - Indicates the customer's average long distance charges, calculated to the end of the quarter specified above
- Multiple Lines - Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No
- Internet Service - Indicates if the customer subscribes to Internet service with the company: Yes, No
- Internet Type - Indicates the customer's type of internet connection
- Avg Monthly GB Download - Indicates the customer's average download volume in gigabytes, calculated to the end of the quarter specified above
- Online Security - Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No
- Online Backup - Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No
- Device Protection Plan - Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No
- Premium Tech Support - Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No
- Streaming TV - Indicates if the customer uses their Internet service to stream television programing from a third-party provider at no additional fee: Yes, No

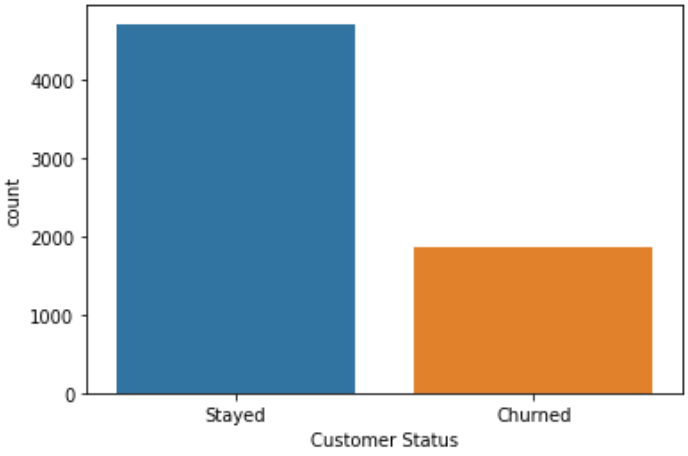
- Streaming Movies - Indicates if the customer uses their Internet service to stream movies from a third-party provider at no additional fee: Yes, No
- Streaming Music - Indicates if the customer uses their Internet service to stream music from a third-party provider at no additional fee: Yes, No
- Unlimited Data - Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No
- Contract - Indicates the customer's current contract type: Month - to - Month, One Year, Two Year
- Paperless Billing - Indicates if the customer has chosen paperless billing: Yes, No
- Payment Method - Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check
- Monthly Charge - Indicates the customer's current total monthly charge for all their services from the company
- Total Charges - Indicates the customer's total charges, calculated to the end of the quarter specified above
- Total Refunds - Indicates the customer's total refunds, calculated to the end of the quarter specified above
- Total Extra Data Charges - Indicates the customer's total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above
- Total Long-Distance Charges - Indicates the customer's total charges for long-distance above those specified in their plan, by the end of the quarter specified above
- Total Revenue - Indicates the company's total revenue from this customer, calculated to the end of the quarter specified above
- Customer Status - Indicates the status of the customer at the end of the quarter: Churned, Stayed, or Joined
- Churn Category - A high-level category for the customer's reason for churning, which is asked when they leave the company: Attitude, Competitor, Dissatisfaction, Other, Price
- Churn Reason - A customer's specific reason for leaving the company, which is asked when they leave the company

- Checking the descriptive statistics of the dataset.

	Age	Number of Dependents	Zip Code	Latitude	Longitude	Number of Referrals	Tenure in Months	Avg Monthly Long Distance Charges	Avg Monthly GB Download	Monthly Charge	Total Charges	Total Refunds	Total Extra Data Charges
count	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	6361.000000	5517.000000	7043.000000	7043.000000	7043.000000	7043.000000
mean	46.509726	0.468692	93486.070567	36.197455	-119.756684	1.951867	32.386767	25.420517	26.189958	63.596131	2280.381264	1.962182	6.860713
std	16.750352	0.962802	1856.767505	2.468929	2.154425	3.001199	24.542061	14.200374	19.586585	31.204743	2266.220462	7.902614	25.104978
min	19.000000	0.000000	90001.000000	32.555828	-124.301372	0.000000	1.000000	1.010000	2.000000	-10.000000	18.800000	0.000000	0.000000
25%	32.000000	0.000000	92101.000000	33.990646	-121.788090	0.000000	9.000000	13.050000	13.000000	30.400000	400.150000	0.000000	0.000000
50%	46.000000	0.000000	93518.000000	36.205465	-119.595293	0.000000	29.000000	25.690000	21.000000	70.050000	1394.550000	0.000000	0.000000
75%	60.000000	0.000000	95329.000000	38.161321	-117.969795	3.000000	55.000000	37.680000	30.000000	89.750000	3786.600000	0.000000	0.000000
max	80.000000	9.000000	96150.000000	41.962127	-114.192901	11.000000	72.000000	49.990000	85.000000	118.750000	8684.800000	49.790000	150.000000

EXPLORATORY DATA ANALYSIS:

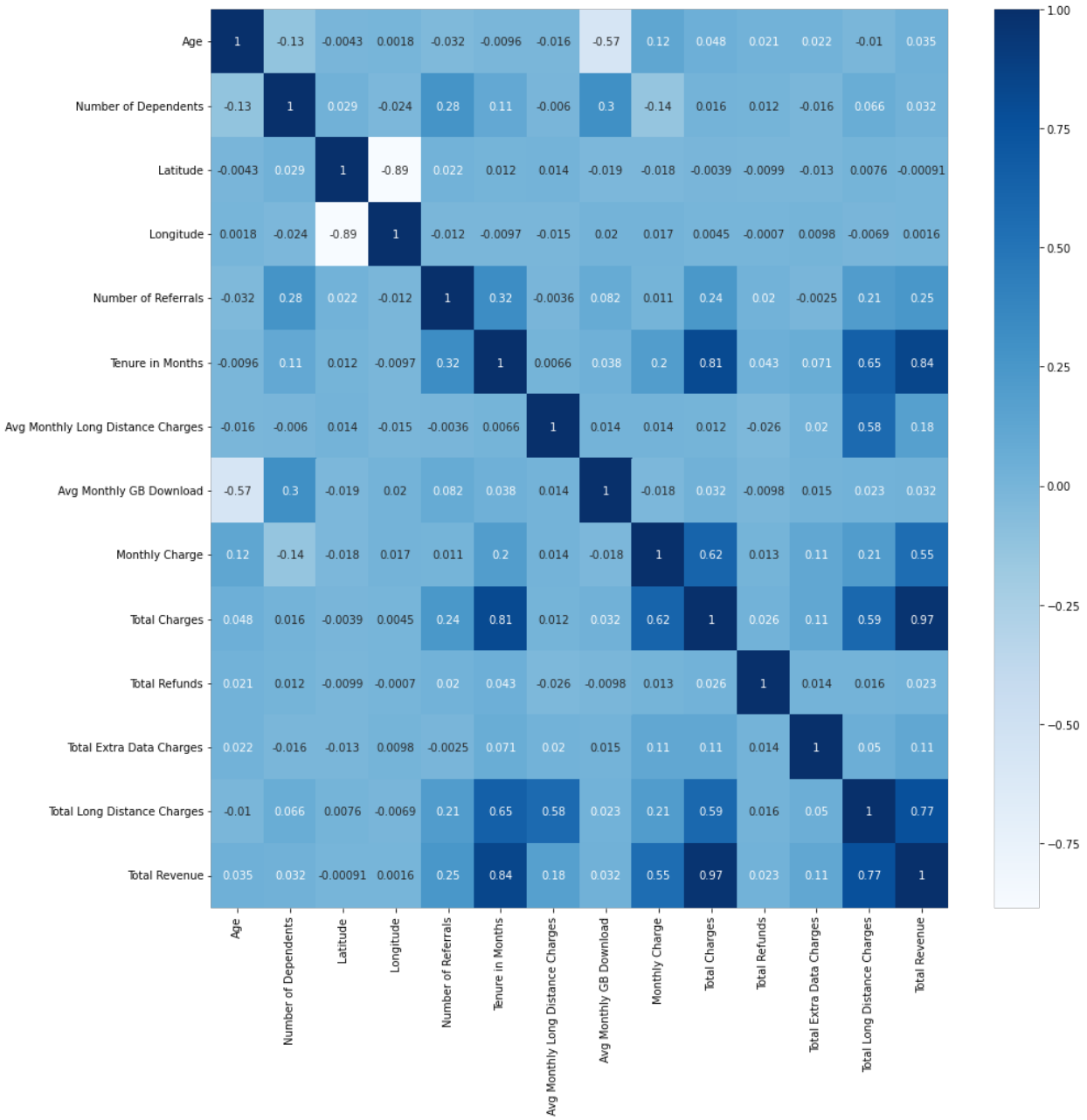
First and foremost, we have plotted the distribution of the target variable. This was done using the countplot function of the seaborn library. From the below countplot, it can be said that the target variable is imbalanced and this needs to be balanced for the model performance using balancing techniques.



CORRELATION HEATMAP:

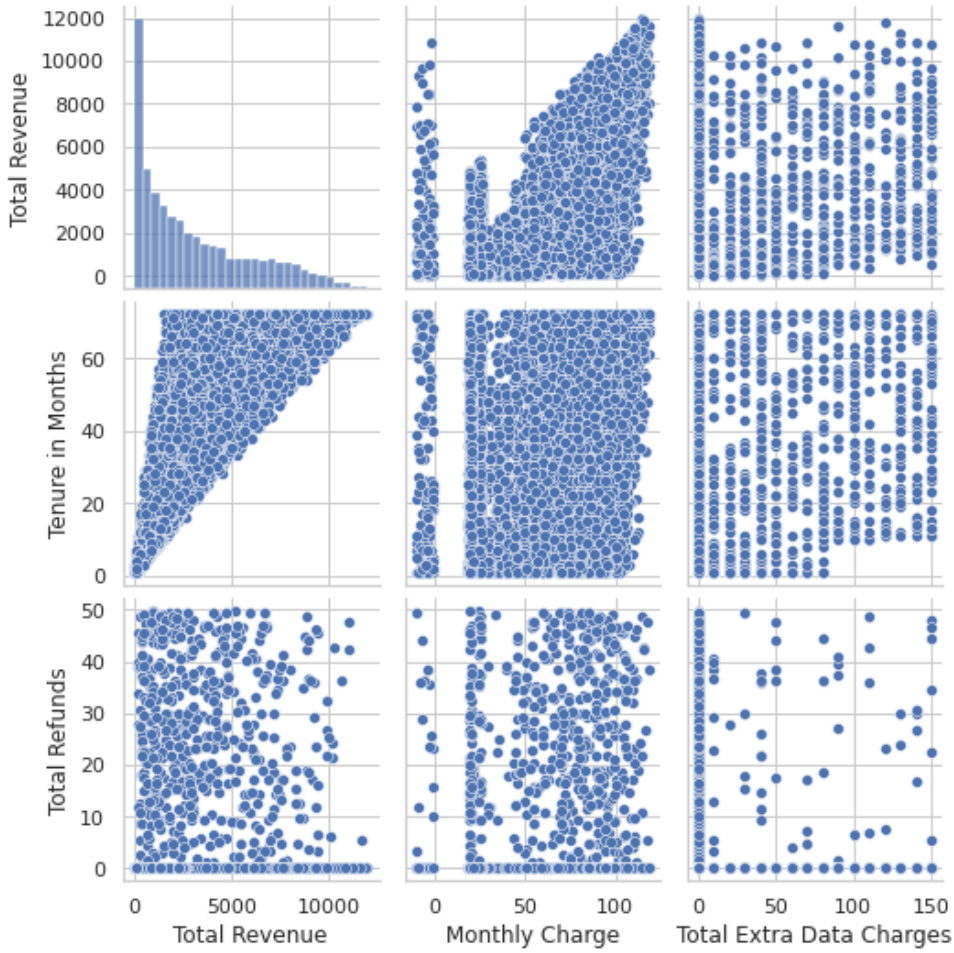
- In the following step, we have plotted a heatmap of all the variables in the dataset to plot the correlation between them.
- From the correlation heatmap, it can be observed that Total Charges and Tenure in Months have a very high correlation.

- Similarly, Total Revenue and Total Long Distance charges also have a high correlation.



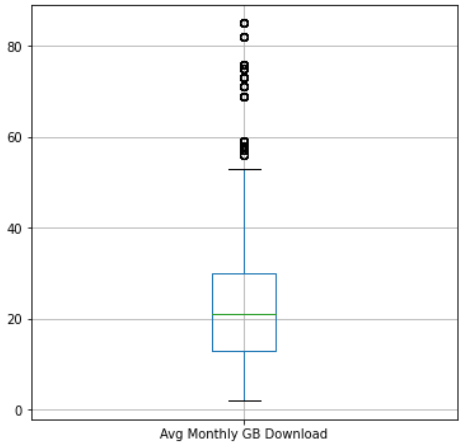
SCATTERPLOT MATRIX:

- We have utilized pairplots to visually represent the correlation between various attributes.
- It can be observed that Total Revenue and Total Refunds don't have a proper correlation
- Total revenue and Tenure in Months have a positive correlation
- Monthly charge and total revenue also have a positive correlation



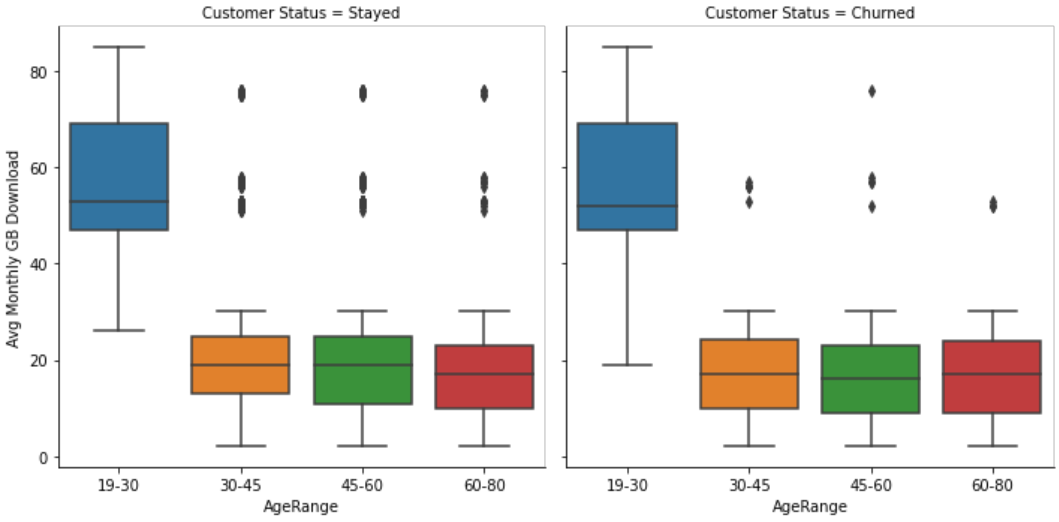
BOX PLOT:

Plotting a box plot for Avg monthly GB download.
It can be seen from the box plot that there are several outliers.
We further need to analyze if these outliers have a valid reason to be in such extremes.



We then try to analyze if age has any effect on Avg monthly GB download.

- As expected, people in the younger age groups (Ages: 19 to 30) have the highest downloads in GB.
- This is observed in all the classes of the target variable.



Dimension Reduction & Variable Selection:

Principal Component Analysis (PCA):

The decision to not apply PCA to the dataset is taken because of the following reasons:

- PCA is usually performed on variables which have very high correlation. But from our correlation heatmap, it can be observed that apart from 2-3 similar variables, most of the other variables have the correlation coefficient of less than 0.3. During such cases, PCA will not be very helpful.
- The algorithm is heavily biased in datasets with extreme outliers. Although it is recommended to remove the outliers, in some cases outliers play an extremely important role in raising questions about the specific behavior of the data which can lead to new insights through analysis.
- PCA is sensitive to the scale of features in our dataset. Irrespective of the highest variance in the data, PCA will heavily favor the first feature as the first principle component.
- Only when the most significant variables also happen to have the greatest degree of variance can PCA be considered a reliable technique for feature selection.
- The result of PCA is our original features transforming into principal components which is the linear combination of original features. These principal components are less interpretable as compared to the original features.

MODEL SELECTION:

Model selection is a crucial step in the process of machine learning, where we select the ideal model that offers the best performance from an initial set of models to solve the problem that we are dealing with.

Since our goal to predict the customer churn is a regarding classification, we have chosen the following classification models to address the issue.

In the subsequent sections, each of the model's descriptions along with the advantages and disadvantages are explained.

1. Logistic Regression:

- Logistic Regression is one of the most common used supervised algorithms used for classification.
- Logistic regression estimates the probability of an event occurring based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.
- It categorizes the data into discrete classes from the labeled data. Logistic regressions are used when the dependent variables are binary

Advantages:

- The results of the model are accurate when the data is linearly separable.
- It is less prone to over-fitting on low-dimensional data.
- It is easy to implement and very efficient to train.
- It indicates the relevance (coefficient size) of a predictor and gives its direction of association (positive or negative).

Disadvantages:

- When are there less number of observations and more number of features, the model tends to overfit
- It cannot be applied on continuous data
- Non-linear problems can't be solved with logistic regression because it has a linear decision surface

2. Support Vector Machine (SVM):

- SVM is a supervised learning algorithm that classifies data based on the hyperplane.
- The hyper plane can be described as a decision boundary that segregates dimensional space into classes so that the new data point can be correctly put in the correct class.
- SVM chooses extreme cases/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Advantages:

- SVM works comparably well when there is an understandable margin of dissociation between classes.
- SVM is efficient in high dimensional spaces
- SVM has L2 Regularization feature. So, it has good generalization capabilities which prevent it from over-fitting.

Disadvantages:

- In the scenario where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
- The training time for large datasets is relatively high due to the complexity of the algorithm.
- SVM does not execute very well when the data set has more sound i.e. target classes are overlapping.

3. Random Forest classifier:

- Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble.
- Each individual tree in the random forest gives out a class prediction and the class with the most votes becomes the model's prediction
- The three main hyperparameters, which need to be set before training are node size, the number of trees, and the number of features sampled.

Advantages:

- It is robust to outliers and can handle missing values
- It works well with both categorical and continuous variables.
- Feature scaling is not needed as it uses rule-based approach instead of distance calculation.
- It reduces overfitting and variance.

Disadvantages:

- Because of its complexity, it requires a lot of computational power and resources
- They are not easily interpretable. They provide feature importance but it does not provide complete visibility into the coefficients as linear regression

4. Gradient Boosting:

- It's used primarily for classification and regression tasks.
- It is known as a greedy algorithm as its prone to overfit data quickly.
- The main elements of this algorithm are a loss function, a weak learner and an additive model that is used to add weak models to reduce the loss function.

Advantages:

- It optimizes on many loss functions and provides the option of hyper parameter tuning
- Data preprocessing is not required
- Imputation is not required to do as it has the ability to handle missing data

Disadvantages:

- They are prone lo overfitting as they go on to improve with the intention to reduce the error
- It requires a lot of time and memory to run
- It is difficult to interpret in some cases of data

EVALUATION METRICS USED:

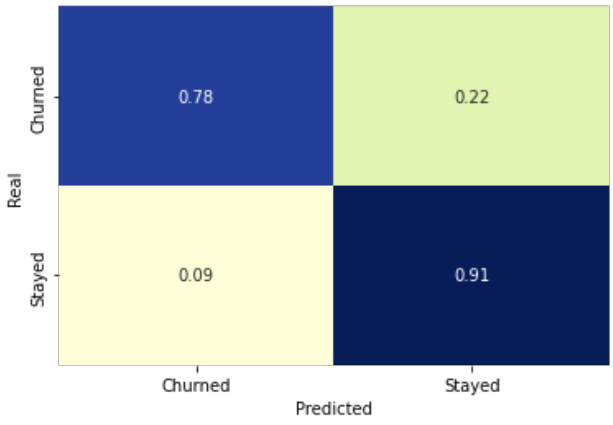
Recall and precision were used as the metrics to evaluate our models. We did not use accuracy as an evaluation metric the reason metric the reason being the imbalance in our target class. Recall can be defined as the value of true positives over the summation of true positives and false negatives. The false negative in our case would be that the customer has left the telecom service, but the model has predicted otherwise.

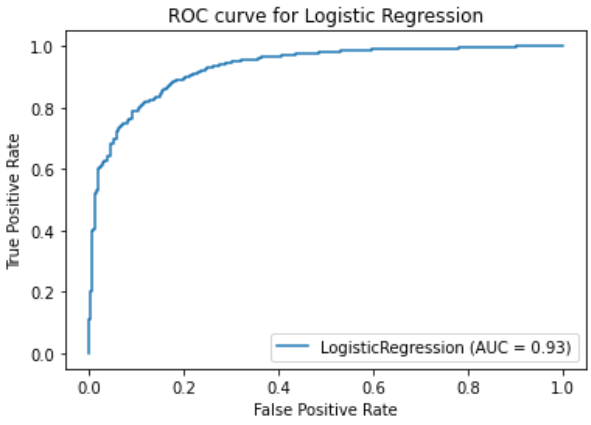
Implementation results of the models:

Logistic Regression:

- This is our first model as logistic regression is widely used for classification tasks while having the benefits of easy to implement nature and training efficiency.
- We have also plotted the learning curves to get additional information about the model's performance. We have achieved a recall of 0.78 and an F1-score of 0.77.
- We have obtained an AUC of 0.93.

	precision	recall	f1-score	support
Churned	0.77	0.78	0.77	368
Stayed	0.91	0.91	0.91	927
accuracy			0.87	1295
macro avg	0.84	0.84	0.84	1295
weighted avg	0.87	0.87	0.87	1295

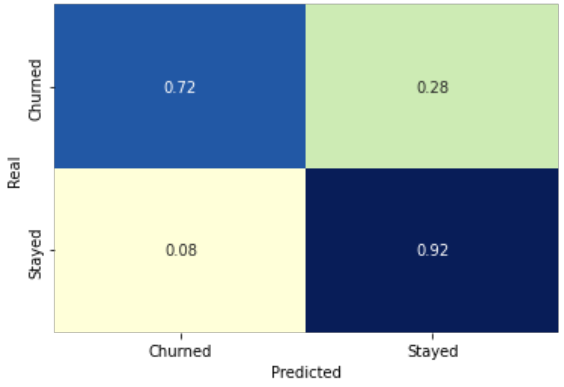


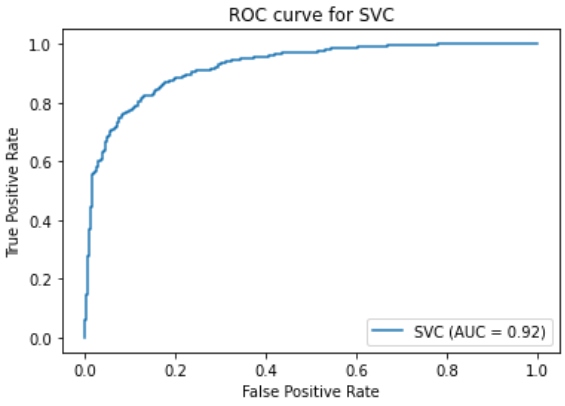


Random Forest:

For the random forest model, we have used stratified sampling because we have an imbalanced data. We have achieved a recall of 0.72 and an F1-score of 0.75. Below are the classification report and the confusion matrix.

	precision	recall	f1-score	support
Churned	0.77	0.72	0.75	368
Stayed	0.89	0.92	0.90	927
accuracy			0.86	1295
macro avg	0.83	0.82	0.83	1295
weighted avg	0.86	0.86	0.86	1295

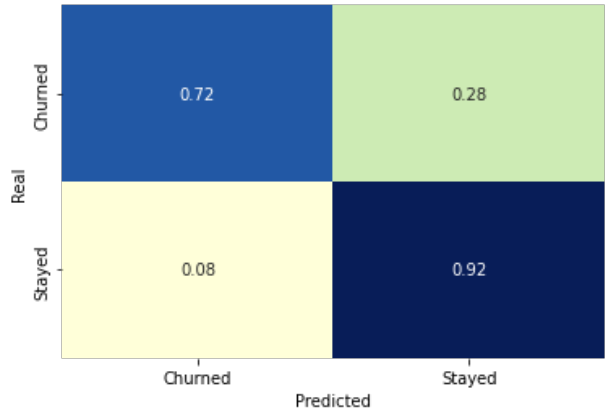


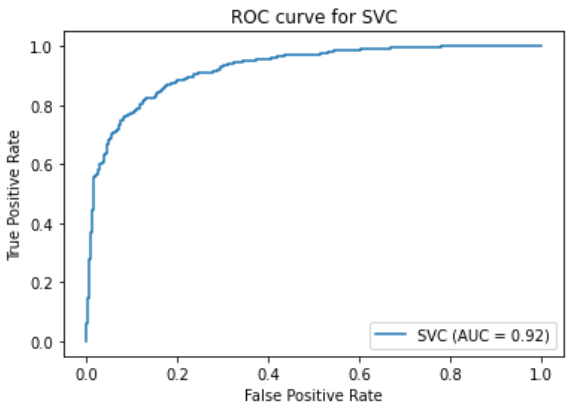


Support Vector Classifier (SVC):

We have used default gamma parameter scale for implementing the SVC. The gamma parameter defines how far the influence of a single training example reaches. We have achieved a recall of 0.72 and a F-1 score of 0.75.

	precision	recall	f1-score	support
Churned	0.77	0.72	0.75	368
Stayed	0.89	0.92	0.90	927
accuracy			0.86	1295
macro avg	0.83	0.82	0.83	1295
weighted avg	0.86	0.86	0.86	1295

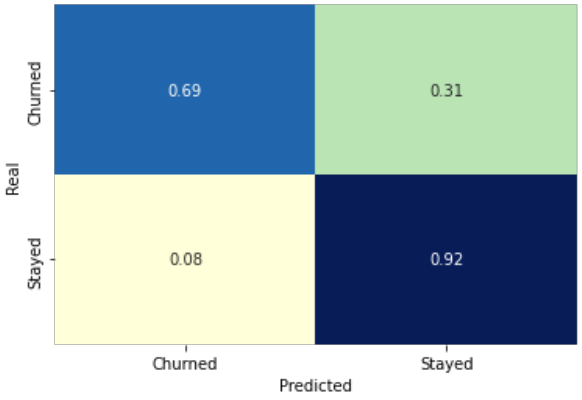


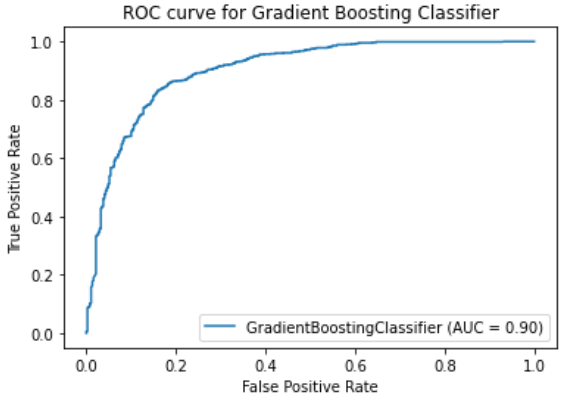


Gradient Boost:

For gradient boosting we have used the parameters such as learning rate, n_estimators and max_depth. Learning rate, denoted as α, simply means how fast the model learns and max_depth bounds the maximum depth of regression tree. n_estimators is the number of trees or weak learners that are included in the model. We have achieved a recall of 0.68 and F1-score of 0.73.

	precision	recall	f1-score	support
Churned	0.77	0.69	0.73	368
Stayed	0.88	0.92	0.90	927
accuracy			0.85	1295
macro avg	0.83	0.81	0.82	1295
weighted avg	0.85	0.85	0.85	1295





PERFORMANCE EVALUATION:

To evaluate the performance of the various models implemented, we have displayed all the performance metric values in a dataframe in the form of a summary table for the category ‘churned’.

	F1 Score	Precision	Recall
Logistic Regression	0.77	0.77	0.78
SVC	0.75	0.77	0.72
Random Forest	0.75	0.77	0.72
XG Boost	0.73	0.77	0.69

- We have also tabulated the above dataframe as a table in order to make the presentation of data even better to the user.

```
+---------------------+------------+-------------+----------+
|                     |   F1 Score |   Precision |   Recall |
|---------------------+------------+-------------+----------|
| Logistic Regression |       0.77 |        0.77 |     0.78 |
| SVC                 |       0.75 |        0.77 |     0.72 |
| Random Forest       |       0.75 |        0.77 |     0.72 |
| XG Boost            |       0.73 |        0.77 |     0.69 |
+---------------------+------------+-------------+----------+
```

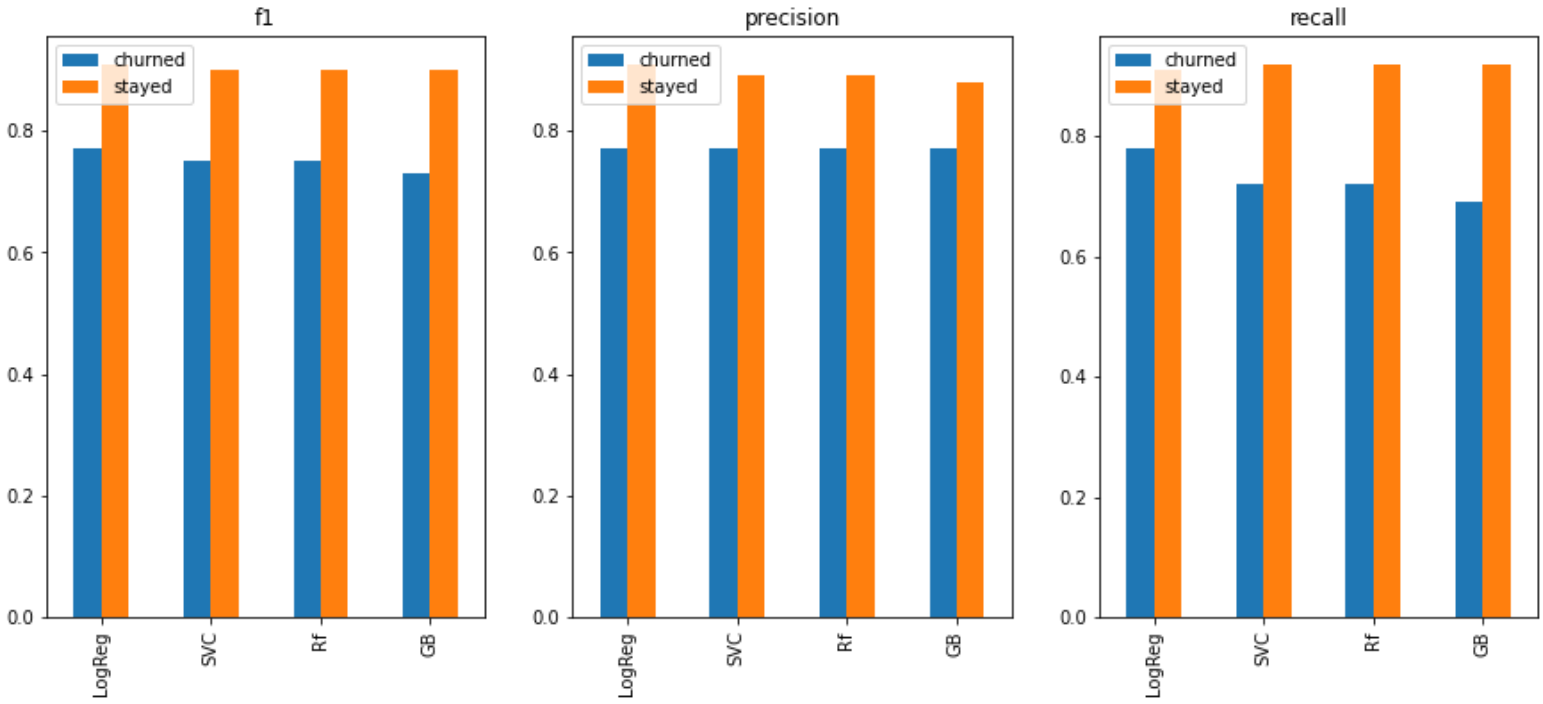
- To get the best performing model, we have also created a table which shows the highest score, a model has achieved in our analysis.

Maximum scores table

	model	f1	model	precision	model	recall
churned	LogReg	0.77	LogReg	0.77	LogReg	0.78
stayed	LogReg	0.91	LogReg	0.91	SVC	0.92

- Out of all the models, Logistic Regression has performed the best with the highest scores of Precision: 0.77, F1 score: 0.77, recall: 0.78 and with an accuracy of 0.87.

Bar charts have also been plotted to accentuate the performance visualization.



CONCLUSION:

One of the major reasons for customer has turned out to be high monthly average prices. If the monthly prices for customers are priced according to the customers profile and the competitors' price plans, it can be major step in retaining customers in the long run.