Spring 2019

BUAN 6340 Programming for Data Science

A Report of

Group Project
on
‘Telecom Customer Churn Analysis’

Submitted by Group 9

Ashwin Kondapalli Reddy
Vaibhav Shrivastava
Pragati Mishra
Ishan Jain

Under the Guidance of
Prof. Ling Ge

Table of contents

1 Executive Summary

Our project mainly deals with the analysis of Customer data of telecom organization facing sever customer attrition rate. As part of our analysis firstly we have preprocessed the data to make it suitable for exploratory data analysis. Then we have further explored the dataset by basic visualizations using Seaborn and Plotly visualization libraries, before moving to advanced predictive analytics of the dataset. We have carried out exploratory data analysis to understand the influence of individual predictors on our target variable (Customer churn). Then we proceeded to feature Selection by analyzing how each independent variable will influence customer attrition rate and thereby decide the top predictors or key drivers of Customer churn.

The latter portion of our project examines the impact different parameters have on Customer Churn using some of the major statistical learning algorithms that we have learnt as part of our Programming for Data Science course. We have built predictive models for customer churn using some of the major statistical methods like logistic regression, KNN Classifier, Decision trees, Random Forest Classifier, Support Vector Machine and Naïve Bayes Classifier. Then we did a comparison of performance metrics (MSE, Confusion matrices, ROC curves) for all the predictive models built as part of our analysis. Using the predictive models to identify customers who fit the churn profile we have suggested the telecom organization with few strategies so that we can proactively target them with marketing and retention programs.

2 Project background (Problem statement):

Customer churn is known as loss of customer:
Service company often uses customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Long term customers can be worth much more to a company than newly recruited clients.

There are two types of churn: voluntary churn and involuntary churn. **Voluntary churn** occurs due to a decision by the customer to switch another company or service provider, and **Involuntary churn** occurs due to circumstances such as customer's relocation to a long-term care facility.

Our project use churn prediction models that predict customer churn by assessing their propensity of risk to churn. Since these models generate a small prioritized list of potential predictors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn.

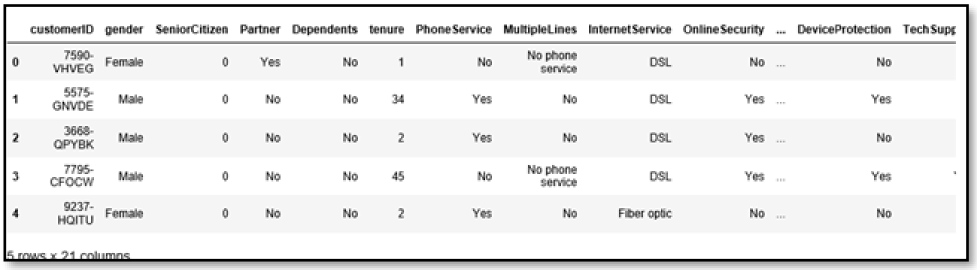
Our objective is to find relation between potential defectors and churn rate, quantify their relation. Summary of this study can be used to make strategy and decrease churn rate. Following relations are important to understand churn rate:

- ✓ Effect of Gender with respect to churn rate;
- ✓ Senior Citizen with respect to churn rate;
- ✓ Effect of tenure period on churn rate;
- ✓ Relation of Customer contract type and churn rate;
- ✓ Relation of Fiber optics service with churn rate;
- ✓ Relation of phone service, online security and multiple line with churn rate;
- ✓ Relation of Online backup, device protection and tech support with churn rate; and
- ✓ Effect of monthly charges and total charges on churn rate

3 Data Description

Data being used: We have worked on the “Telco Customer Churn” data set taken from IBM Watson Analytics community https://community.watsonanalytics.com/resources/. IBM Watson Analytics team posted this dataset of a telecommunications company which is troubled by the number of customers leaving their landline business for other competitors. They need to understand who is leaving. This data set provides information to help us predict customer behavior to retain customers. We can analyze all relevant customer data and develop focused customer retention programs.

Each row represents a customer, each column contains customer’s attributes described on the column Metadata. The raw data contains 7043 rows (customers) and 21 columns (features). The “Churn” column is our target. This data set contains 21 columns (features), but we will be selecting a subset of the most important columns for analysis purposes.



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupp
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	

5 rows x 21 columns

Fig 1 data description

Rows and columns

```
(7043, 21)
```

Variable overview:

Variable name	Type of Variable	Explanation
Customer ID	Numerical	Unique ID of customers
Gender	Categorical	Male / Female
Senior Citizen	Categorical	Yes/No
Partner	Categorical	Yes/No
Dependents	Categorical	Yes/No
Tenure	Numerical	Number of months customer used service
Phone Service	Categorical	Yes/No
Multiple Lines	Categorical	Yes/No/No phone services
Internet Services	Categorical	DSL/Fiber optics/No
Online Security	Categorical	No/Yes/No internet service
Online Backup	Categorical	No/Yes/No internet service
Device protection	Categorical	No/Yes/No internet service
Tech Support	Categorical	No/Yes/No internet service
Streaming TV	Categorical	No/Yes/No internet service
Streaming Movies	Categorical	No/Yes/No internet service
Contract	Categorical	Month to Month/1Year/2Years
Monthly Charges	Numerical	Amount charged in USD
Total Charges	Numerical	Amount charged in USD

Table 1: Variable details

4 Exploratory Data Analysis

4.1 Understanding churn problem: Customer attrition rate in Y-Mobile

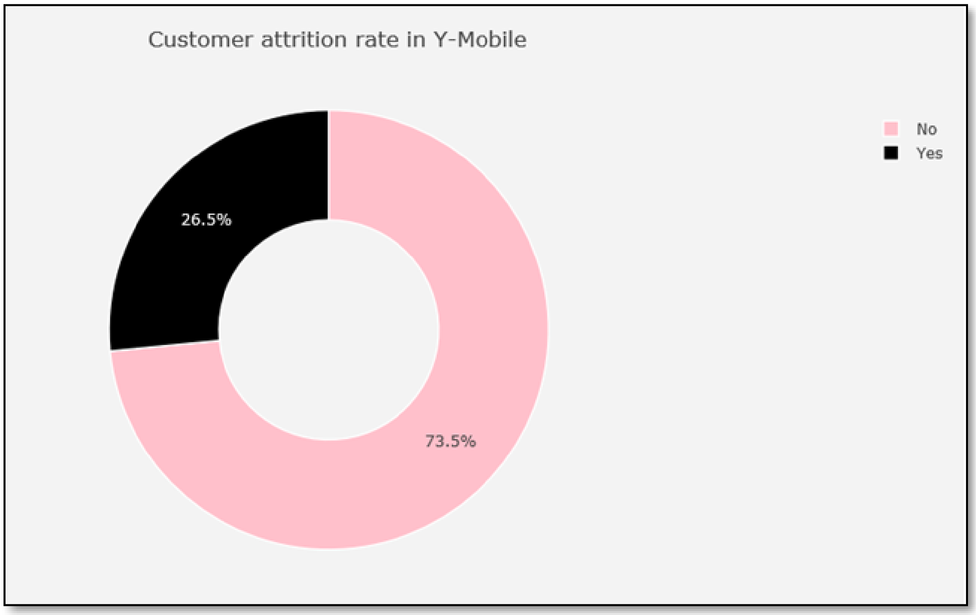


Fig 2 Customer attrition

The churn rate is the percentage of subscribers to a service who discontinue their subscriptions to the service within a given time. So, current attrition rate is 26.5%. Objective of project is to determine bottle neck issue in the business to decrease churn rate.

Exploratory data analysis:

Following relations have been studied to understand which variables are more related to churn rate:

1. Frequency distribution of Gender with respect to churn rate.
2. Frequency Distribution of Senior Citizen with respect to churn rate.
3. Understanding the effect of tenure period on churn
4. Relation of Customer contract type and churn rate
5. Relation of Fiber optics service with churn rate
6. Relation on phone service, online security and multiple line with churn rate
7. Online backup, device protection and tech support with churn rate
8. Effect of monthly charges and total charges on churn rate

4.2 Frequency distribution of Gender with respect to churn rate:

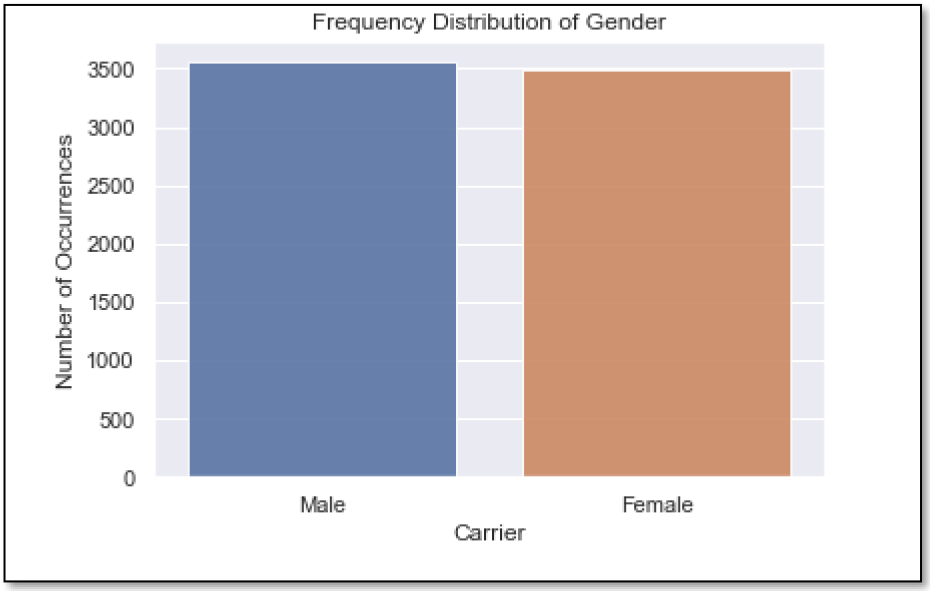
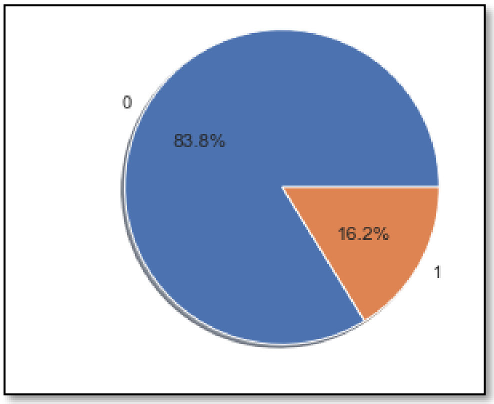
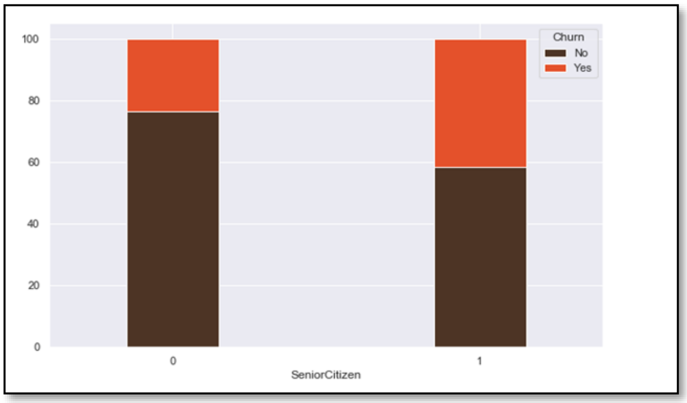


Fig 3 Frequency distribution of Gender

- Number of customers in dataset has almost 50:50 distribution of male and female. So, there no effect of gender biasedness

4.3 Relative frequency distribution of customer as per they are senior citizen or not





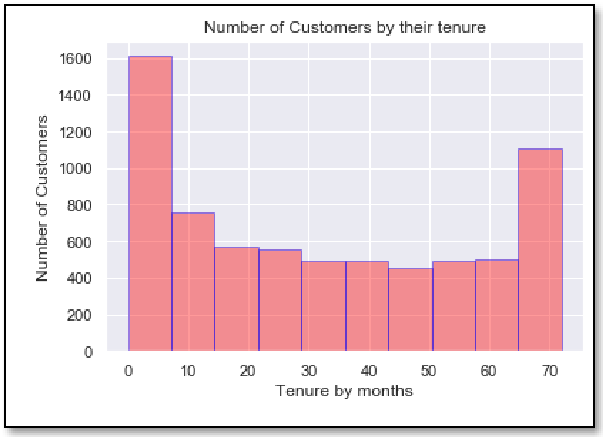
0- Young Citizen
1- Senior Citizen

Fig 4 Senior citizen pie chart

Fig 5 Bar chart of churn rate in senior citizen

- Out of total customers, senior citizens are 16.2%. In senior citizen, churn rate is higher than in non-senior citizen group.

4.4 Study of effect of tenure period on churn rate:



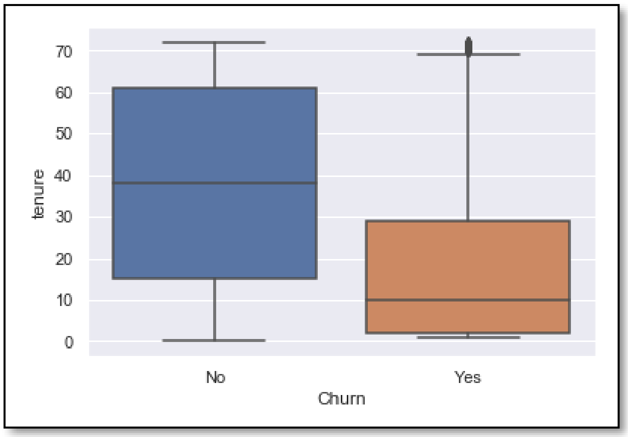


Fig 6 Tenure distribution

Fig 7 Box plot of tenure

- As customer is in initial period of tenure, chances of leave service are higher.

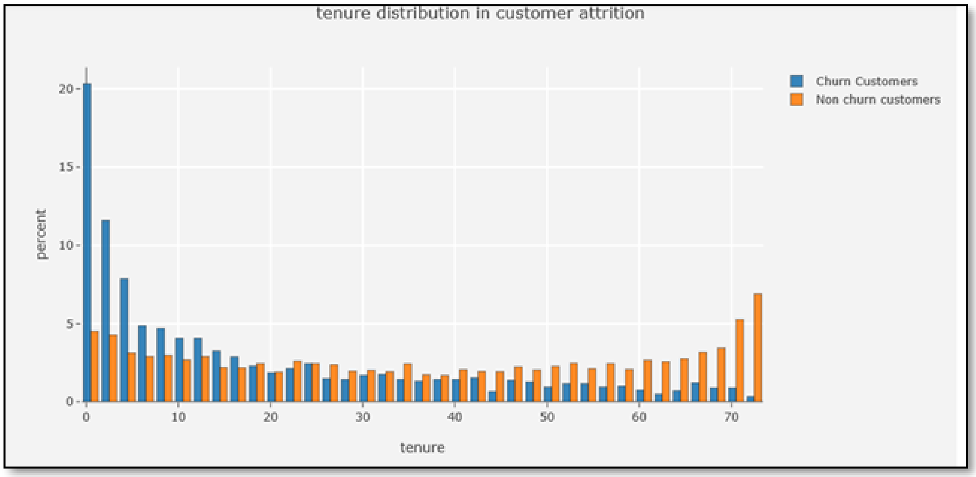


Fig 8 Bar chart of churn and non-churn customer

- As tenure period increases, number of customers leaving services decreases.

4.5 Relation of Customer contract type and churn rate:

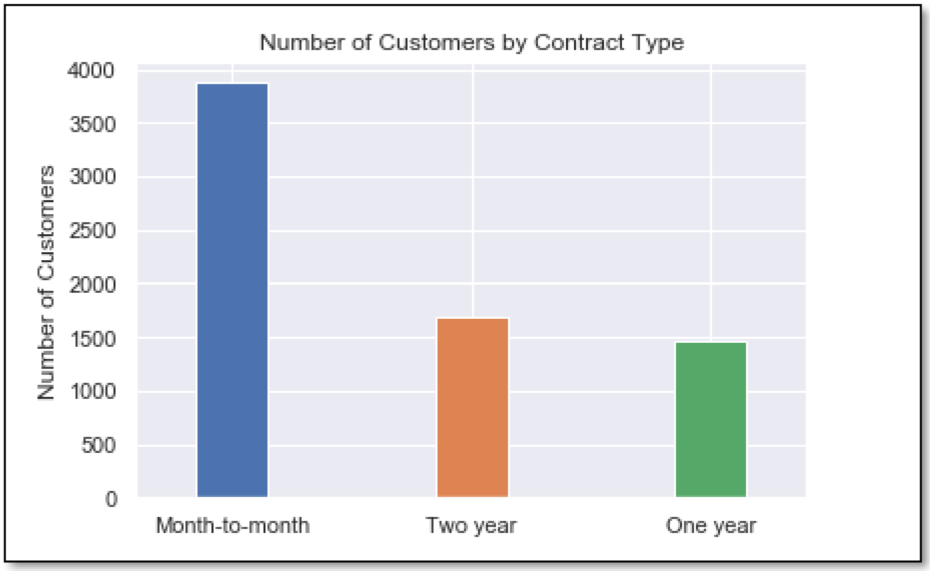


Fig 9 Bar chart of distribution of Customer by contract type

- Month-to-month contract type is having highest number of customers following by two years and one year.

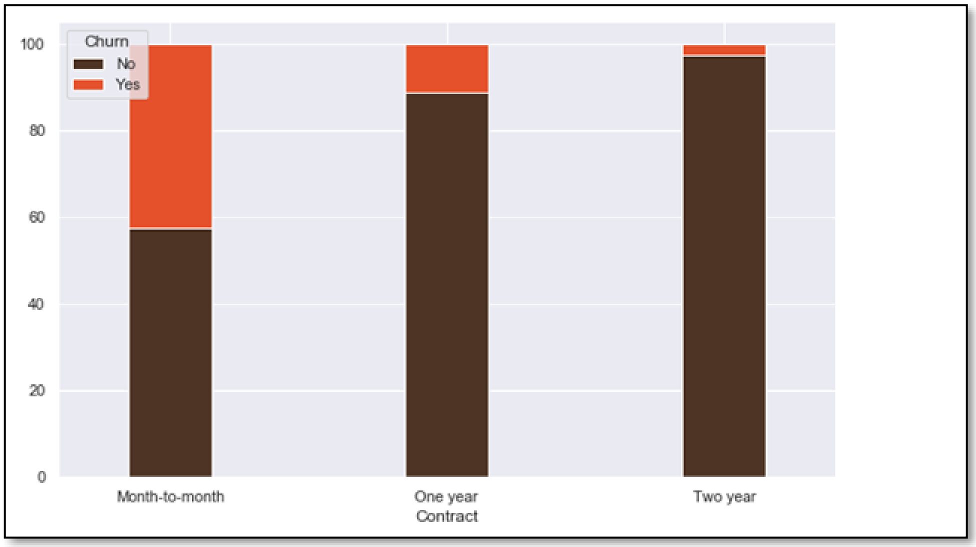


Fig 10 Stacked plot of contract type w.r.t to churn rate

- Month to month is having highest churn.

4.6 Study of effect of services on churn rate

4.6.1 Relation of Fiber optics service with churn rate:

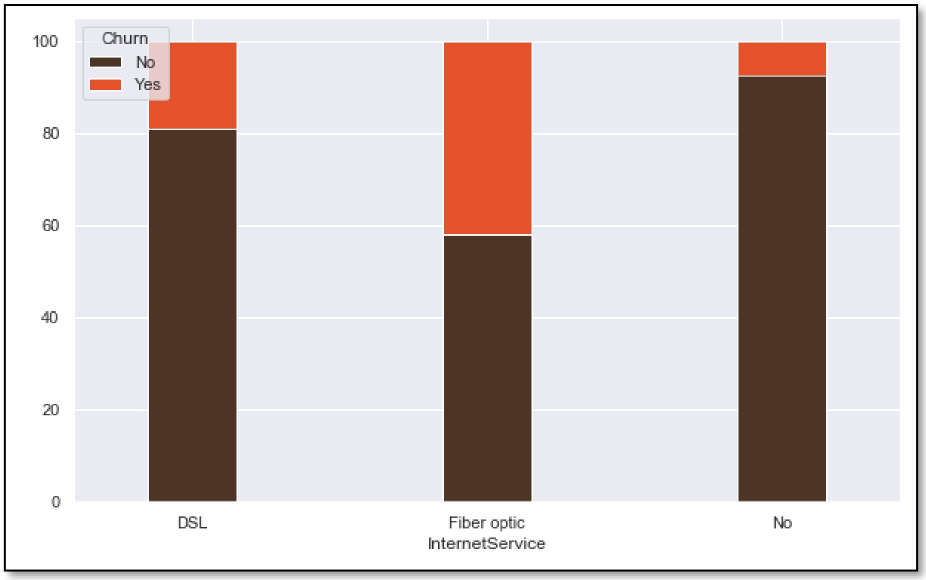


Fig 11 Stacked plot of internet service

Observation: Customers using Fiber optics churn more than other group customers.

4.6.2 Relation of phone service with churn rate:

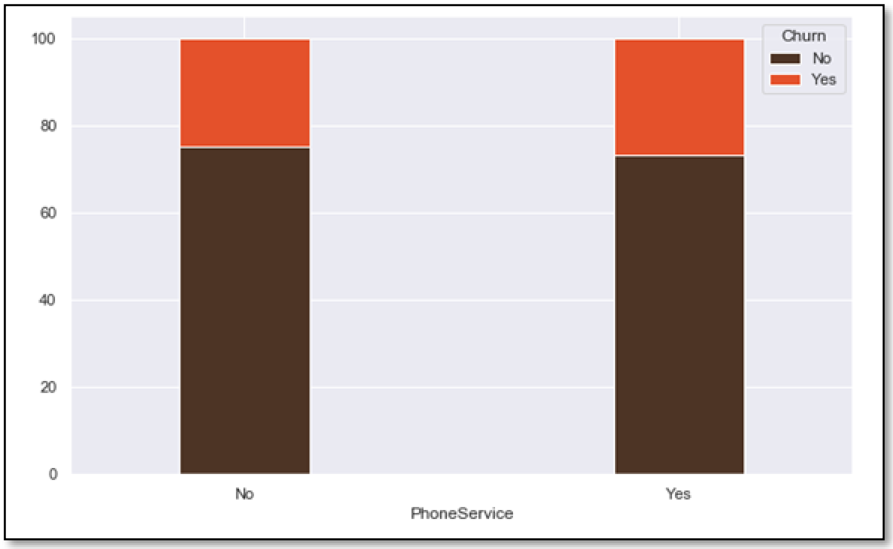


Fig 12 Stacked plot of phone service

- Churn rate is approximately same for customer using phone service vs those not using.

4.6.3 Relation of internet service subscription with online security:

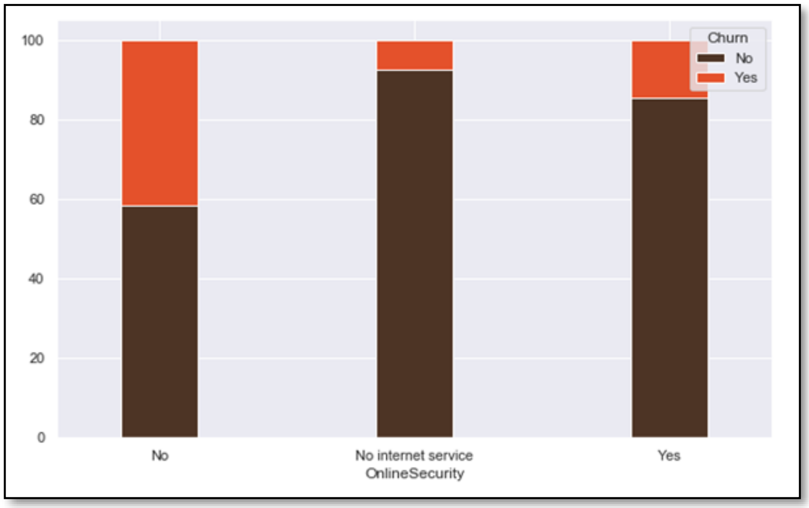


Fig 13 Stacked plot of online security

- Customer who doesn't use internet services churn more than other group customers.

4.6.4 Relation of multiple line service with churn rate:

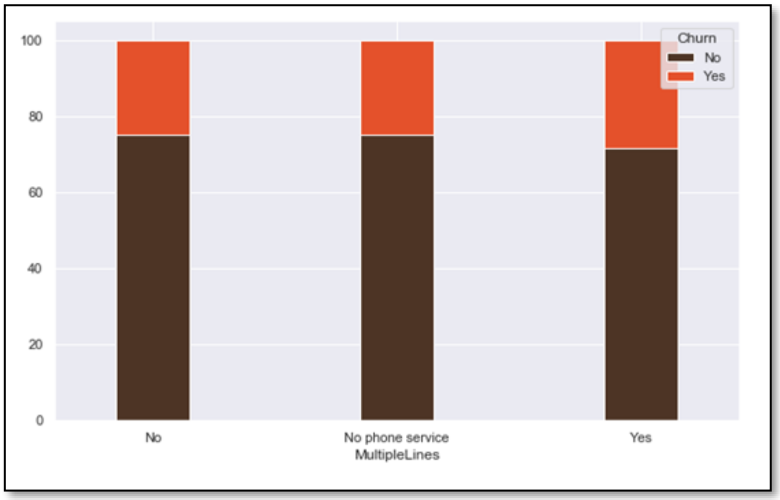


Fig 14 Stacked plot of multiple lines

- Multiple line service doesn't have any relation with churn rate.

4.6.5 Relation of online back service with churn rate:

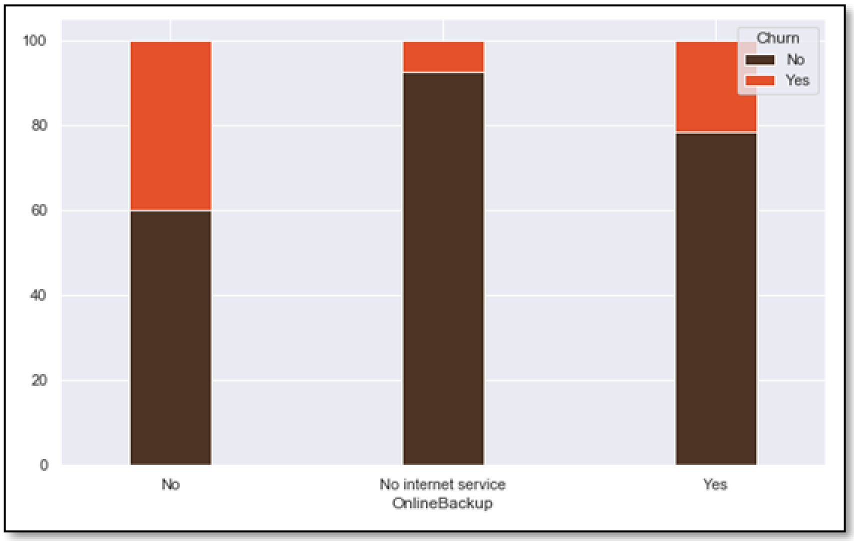


Fig 15 Stacked plot of online backup

- Customers who doesn't use online backup option, churn more than other group of customers

4.6.6 Relation of device protection service with churn rate:

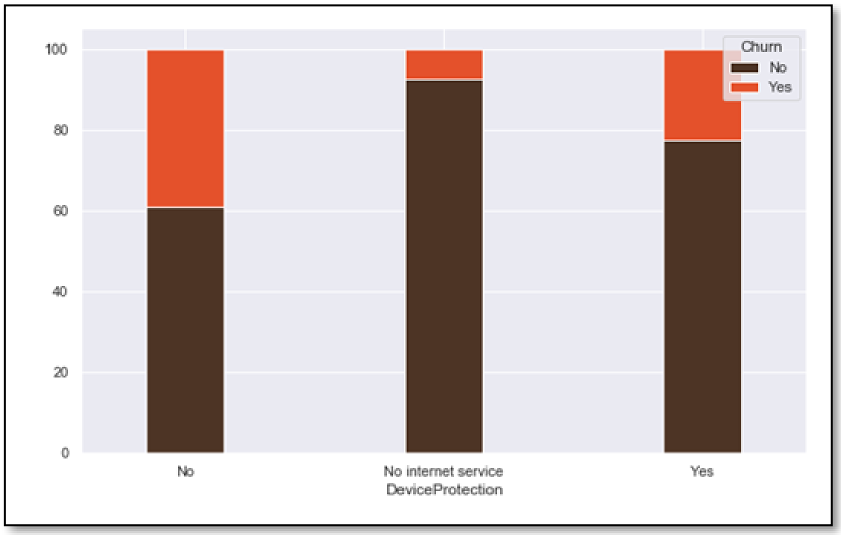


Fig 16 Stacked plot of device protection

- Customer with internet services doesn't use device protection, churn more than other group of customers

4.6.7 Relation of tech support service with churn rate:

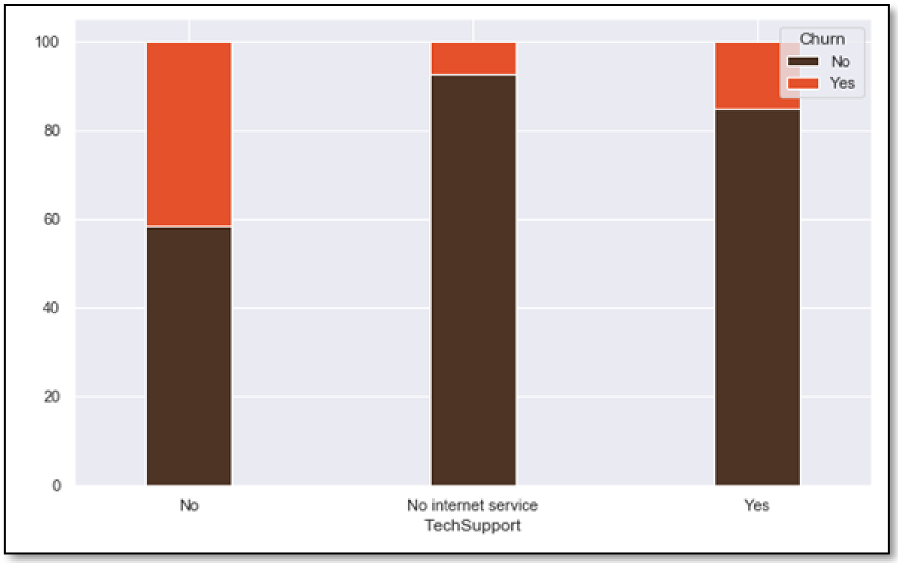


Fig 17 Stacked plot of tech support

- Customer with internet services doesn’t use tech support, churn more than other group of customers

4.6.8 Relation of streaming TV with churn rate:

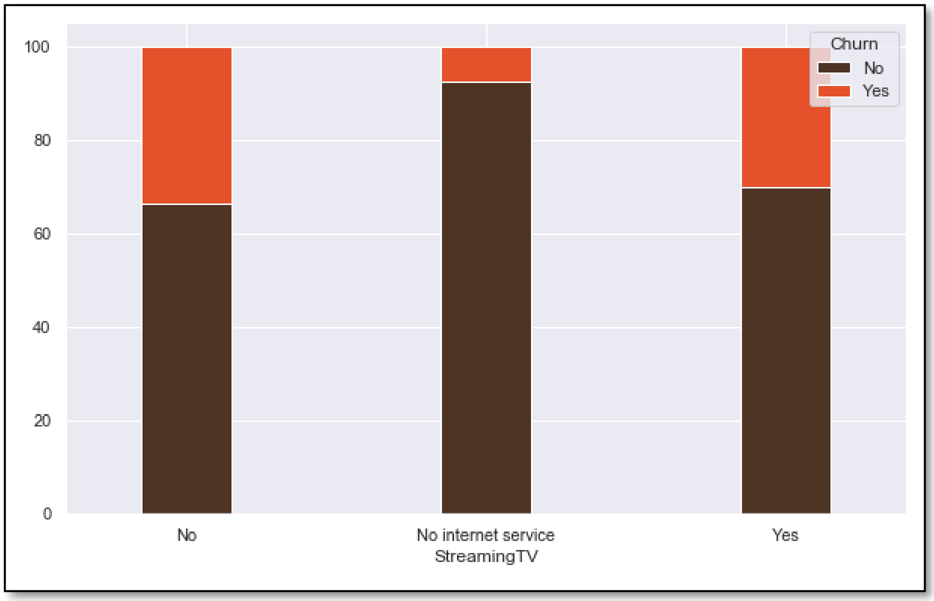


Fig 18 Stacked plot of streaming TV

- Customer with internet services who use or doesn’t use streaming TV, churn more than other group of customers.

4.7 Study of churn rate as per monthly charges and total monthly charges

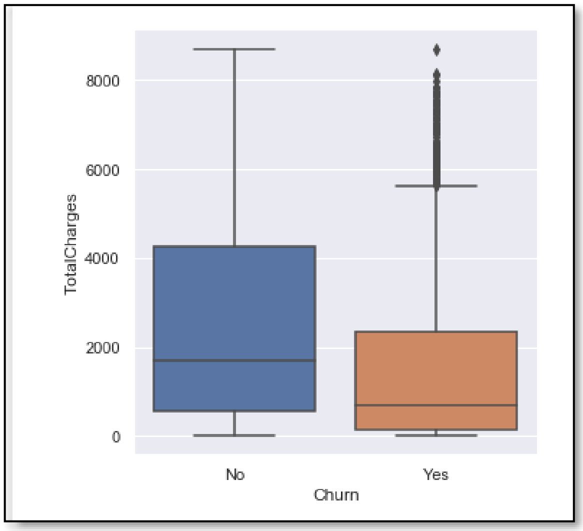


Fig 19 Box plot of total charges

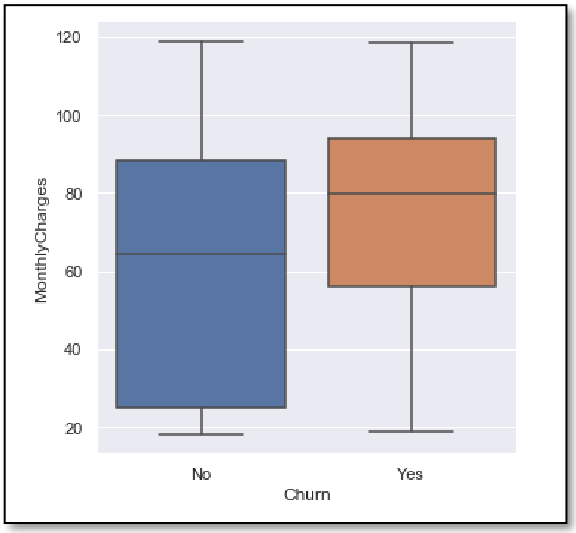


Fig 20 Box plot of Monthly charges

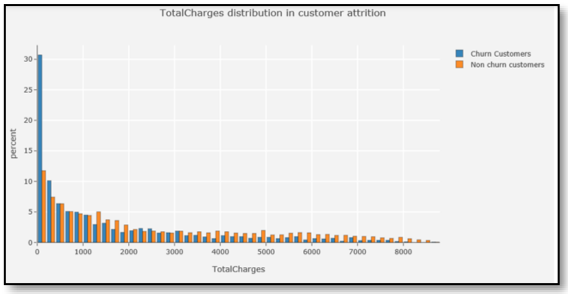


Fig 21 Bar chart of total charges

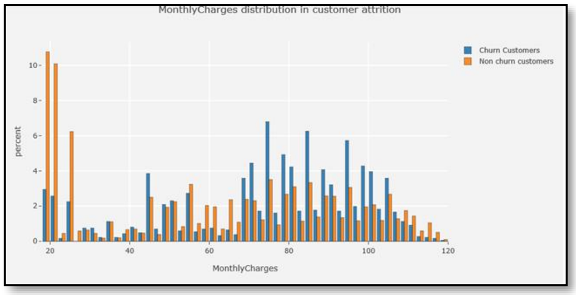


Fig 22 Bar chart of monthly charges

- Median of Monthly charges of customer who churn is higher than who doesn’t churn.
- Median of Total charges of customers who churn is lower than who doesn’t churn.

4.8 Conclusion of EDA:

- Senior citizen has higher churn rate.
- Churn rate is higher in initial period of tenure.
- Month-to month contract has highest churn rate.
- Fiber optics customer churn more frequently than other group customer.

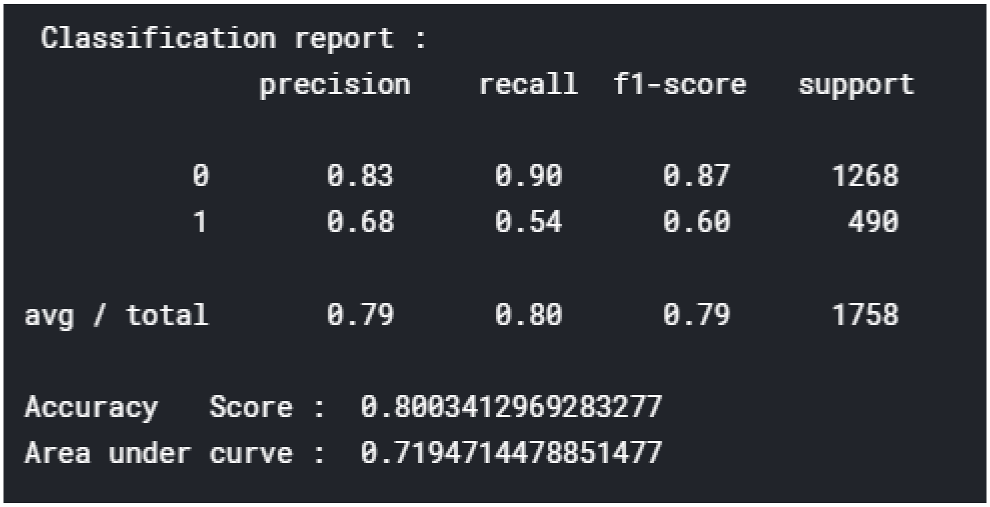
5 Models and Analysis

5.1 Logistic Regression

- The probability of the response taking a value is modeled based on a combination of values taken by the predictors.
- The advantage of Logistic Regression model is that it gives the confidence of prediction as a probability.
- The disadvantage is that it assumes that the classes are linearly separable in feature space.
- For all of the models that we have built in this section, we have divided our Train and test data into 75-25 ratio.

Table 2: Confusion Matrix

	Not Churn	Churn
Churn	227	263
Not Churn	1144	124



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

          0       0.83      0.90      0.87      1268
          1       0.68      0.54      0.60       490

avg / total       0.79      0.80      0.79      1758

Accuracy   Score :  0.8003412969283277
Area under curve :  0.7194714478851477
```

Classification Report for Logistic Model

The model has a very good accuracy i.e. 80% and AUC of 71.9%

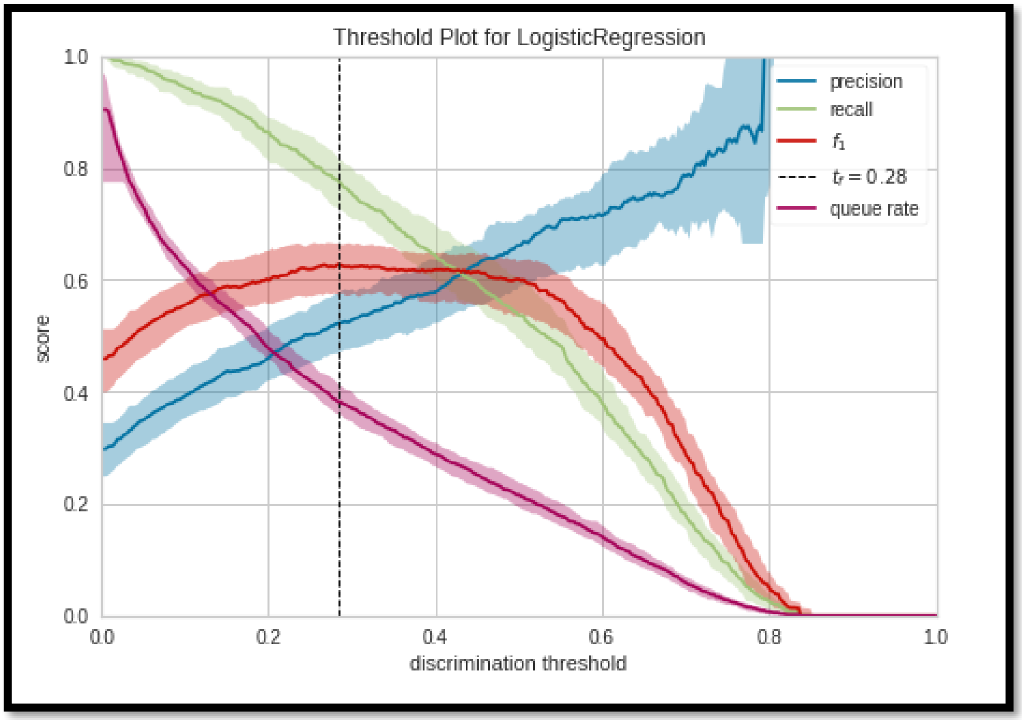


Fig 23 Threshold Plot for Logistic Regression

We see from the plot that the threshold is 0.28 which tells that customers below the threshold are less likely to churn whereas the customers above the threshold are more likely to churn

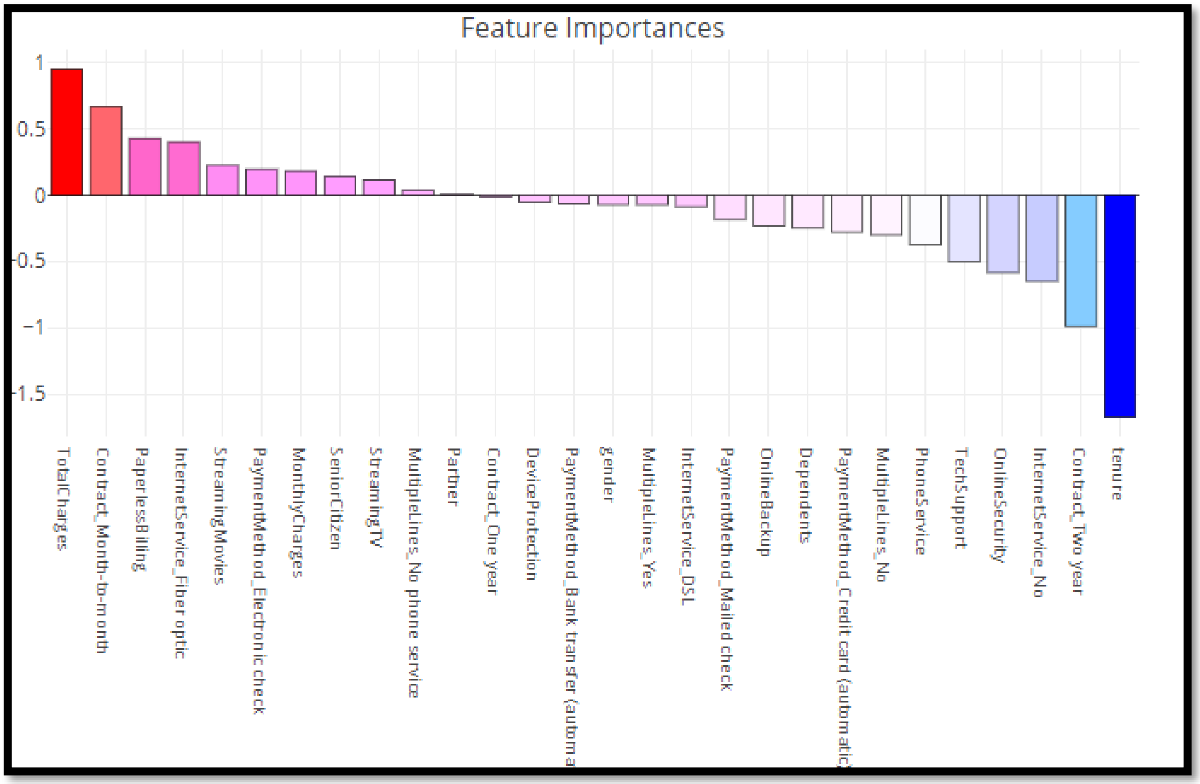


Fig 24 Feature Importance’s for Logistic Model

5.2 Logistic Regression (RFE)

Recursive Feature Elimination (RFE) is based on the idea to repeatedly construct a model and choose either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features.

Table 2: Confusion Matrix

	Not Churn	Churn
Churn	227	12
Not Churn	17702	291

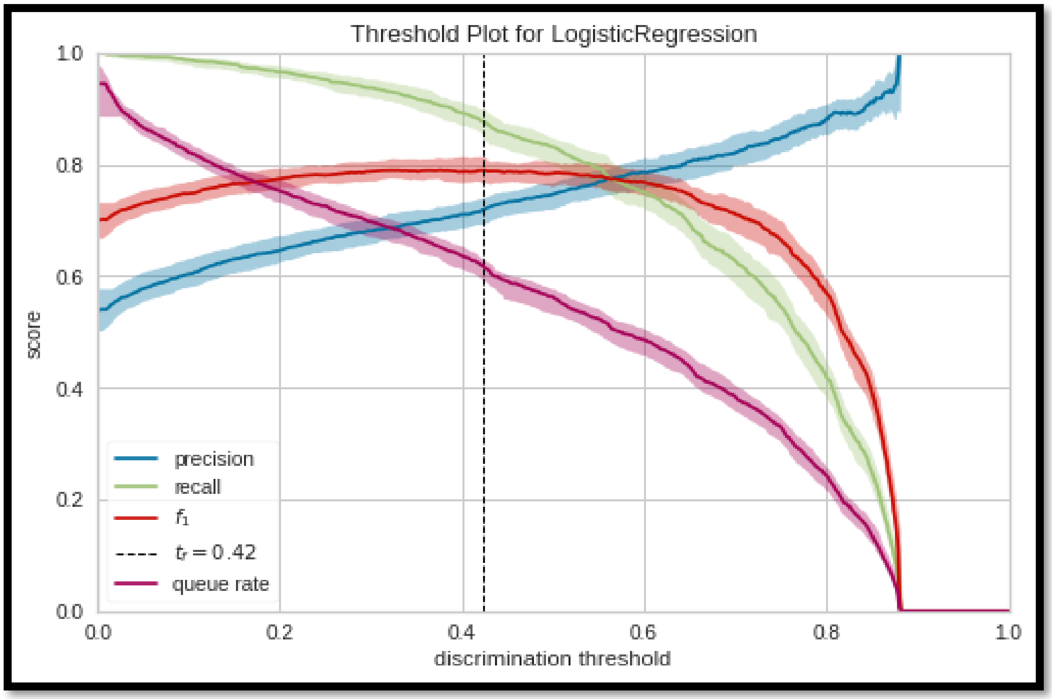


Fig 25 Threshold Plot for RFE in Logistic Regression (Threshold= 0.42)

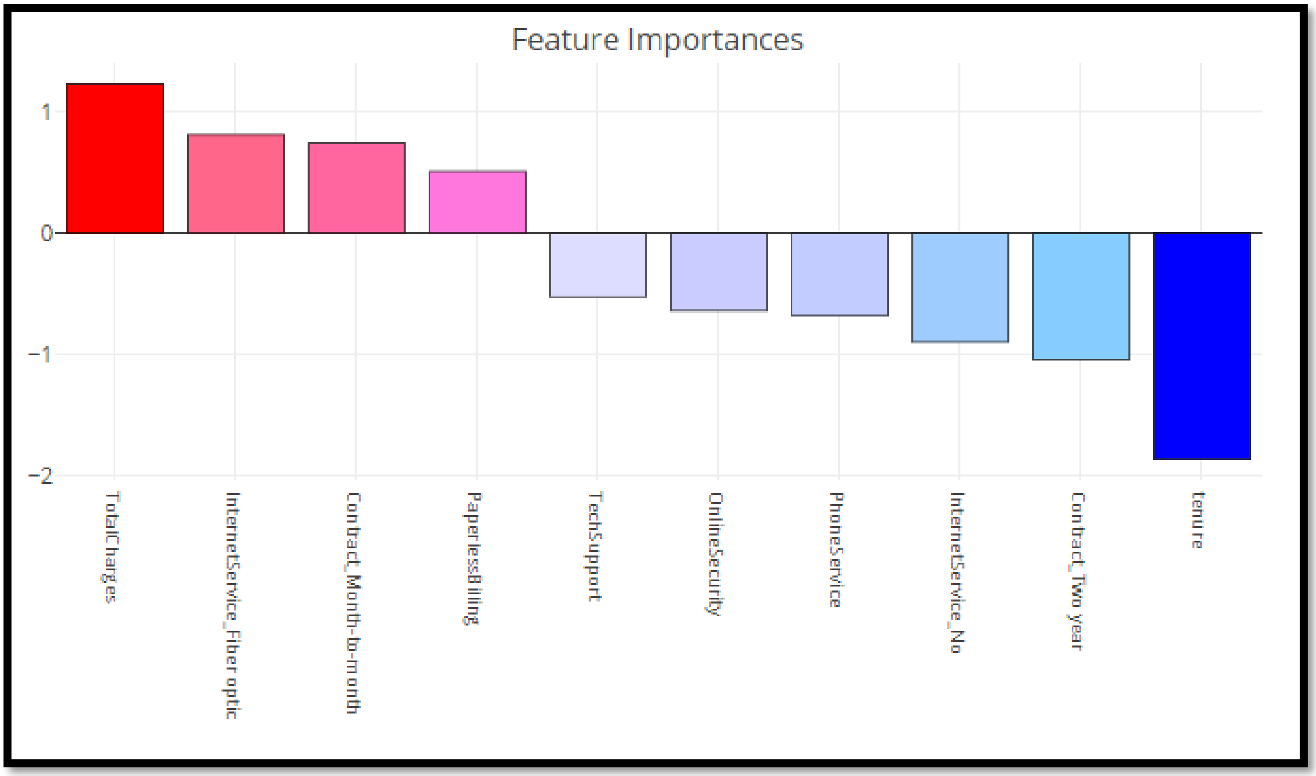


Fig 26 Feature Importance's for RFE Model

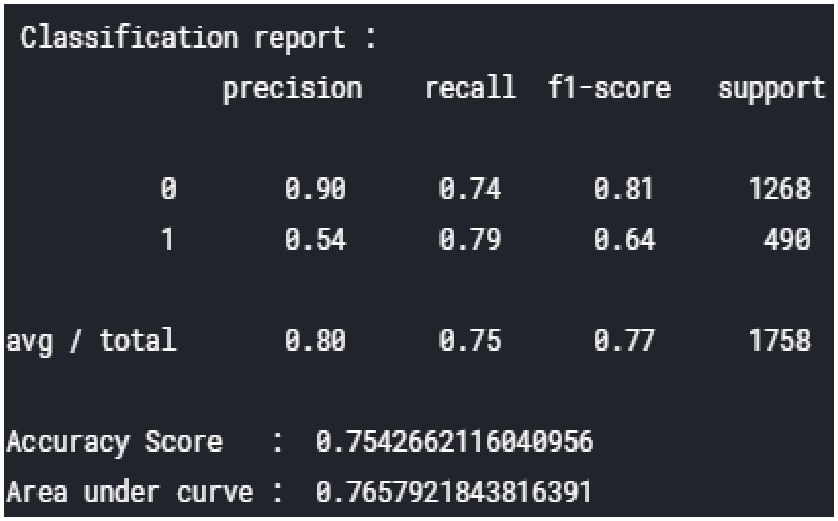
These are the most important predictors in Customer attrition after applying RFE in Logistic Regresssion

5.3 Naïve Bayes Classifier

- It performs well in case of categorical input variables compared to numerical variables. For numerical variable, normal distribution is assumed.
- One of its limitation is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

Table 3: Confusion Matrix

	Not Churn	Churn
Churn	102	388
Not Churn	938	330



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

          0       0.90      0.74      0.81      1268
          1       0.54      0.79      0.64       490

avg / total       0.80      0.75      0.77      1758

Accuracy Score   :  0.7542662116040956
Area under curve :  0.7657921843816391
```

Fig 27 Classification Report for Naïve Bayes

The model has 75.4 % accuracy and AUC of 76.5 %

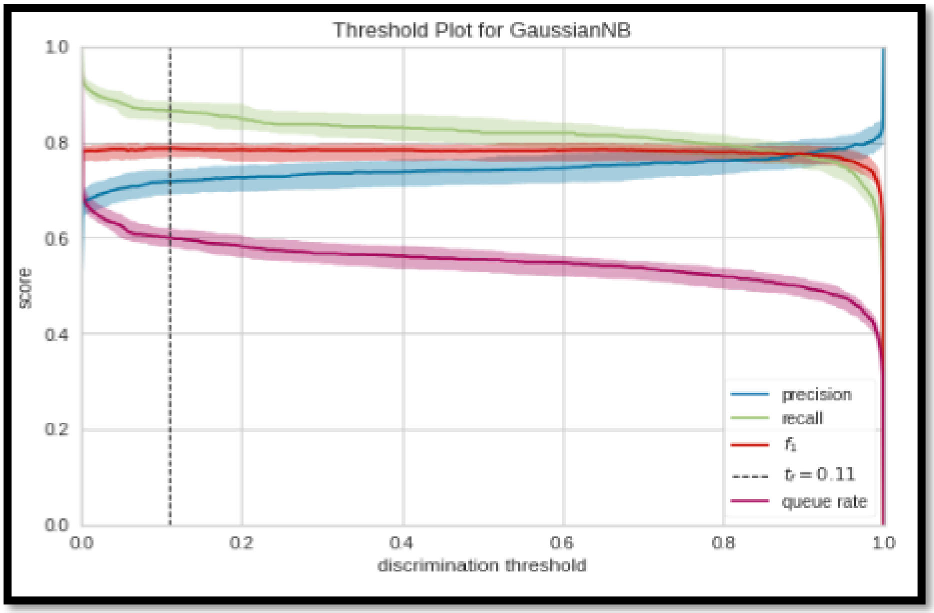


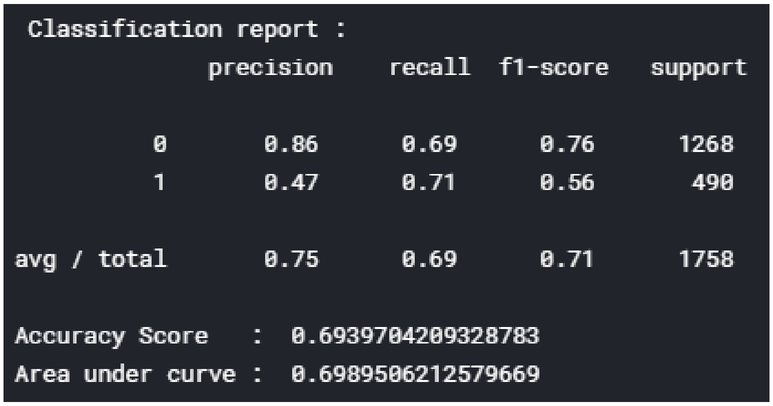
Fig 28 Threshold Plot for Naïve Bayes Classifier (Threshold= 0.11)

5.4 KNN Classifier

- KNN is a non-parametric algorithm, which means that it does not make any assumptions on the underlying data distribution.
- It is based on feature similarity**,** which means how closely out-of-sample features resemble our training set determines how we classify a given data point.

Table 4: Confusion Matrix

	Not Churn	Churn
Churn	142	348
Not Churn	872	396



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

          0       0.86      0.69      0.76      1268
          1       0.47      0.71      0.56       490

avg / total       0.75      0.69      0.71      1758

Accuracy Score   :  0.6939704209328783
Area under curve :  0.6989506212579669
```

Fig 29 Classification Report for KNN Algorithm

The model has 69.3% accuracy and AUC of 69.8% which is a bit low compared to Logistic Regression and KNN

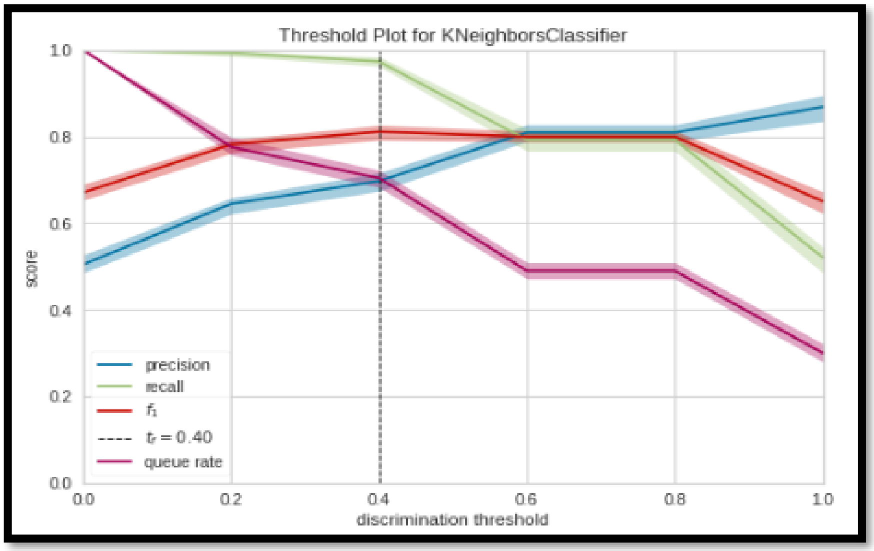


Fig 30 Threshold Plot for KNN Algorithm (Threshold= 0.40)

5.5 Decision Trees

- A decision tree is a flowchart-like tree structure where an internal node represents feature, the branch represents a decision rule, and each leaf node represents the outcome. The flow chart like structure helps you in classifying your decision-making process and is one of the best models for easy interpretability.
- We have built a decision Tree Classifier using the best three predictors we have found in the previous models i.e., we have used “Tenure”, “Monthly Charges” and “Total Charges” to build the Decision Tree Classifier.

Table 5: Confusion Matrix

	Not Churn	Churn
Churn	223	267
Not Churn	1071	197

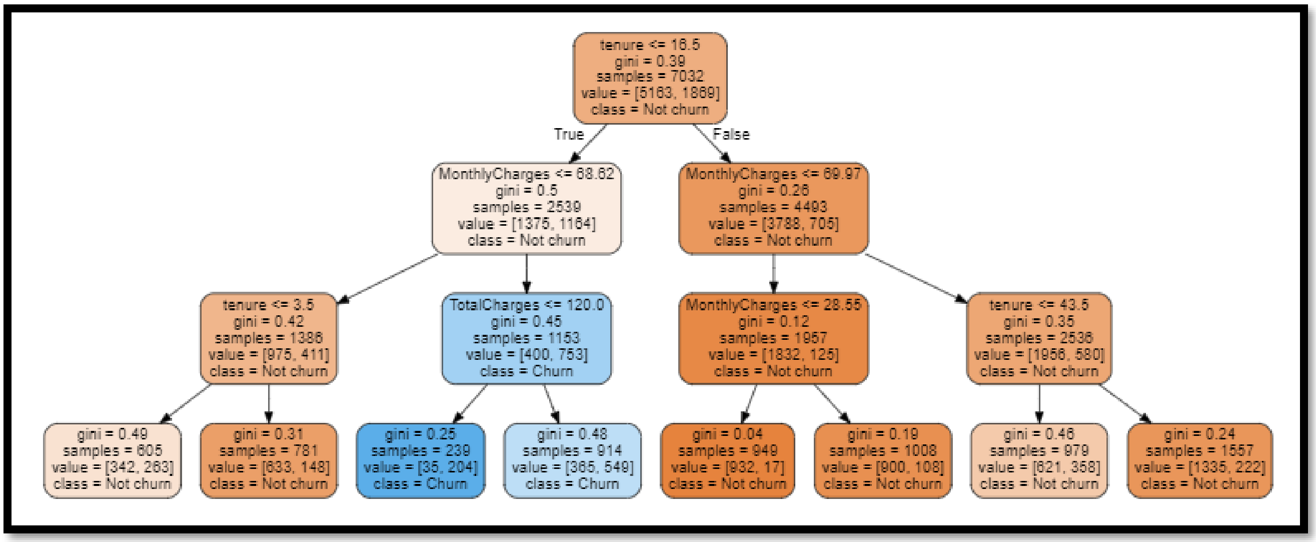
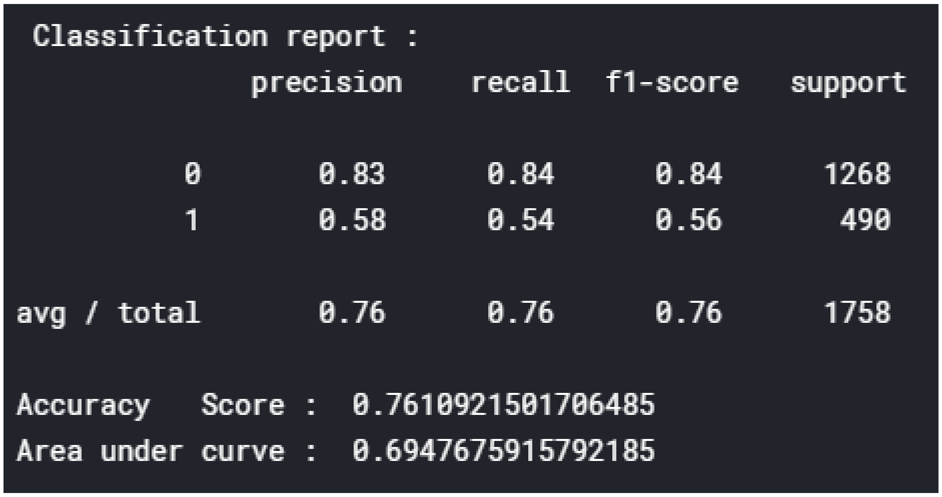


Fig 31 Visual Representation of Decision Tree Model



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

          0       0.83      0.84      0.84      1268
          1       0.58      0.54      0.56       490

avg / total       0.76      0.76      0.76      1758

Accuracy   Score :  0.7610921501706485
Area under curve :  0.6947675915792185
```

Fig 32 Classification Report for Decision Tree Model

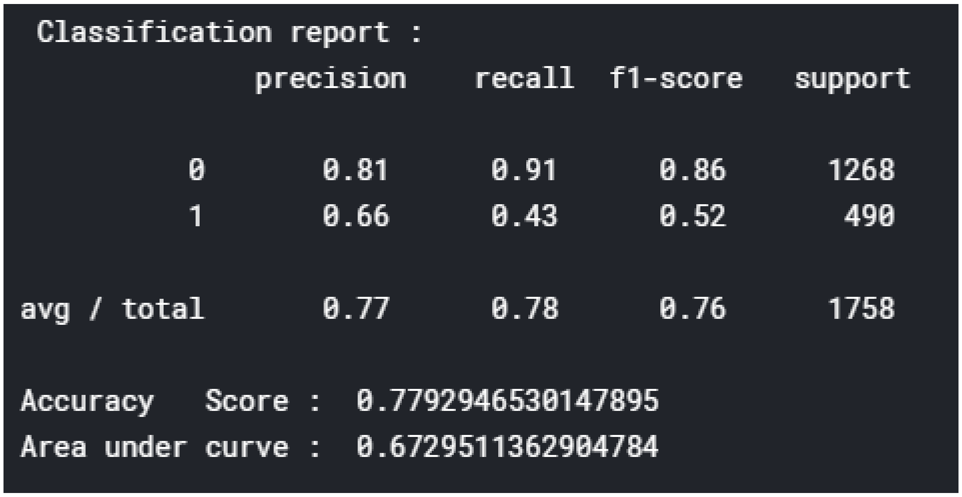
The model has 76.1% accuracy and AUC of 69.4% which is a bit higher compared to Logistic Regression and KNN.

5.6 Random Forest Classifier

- A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn with replacement.

Table 6: Confusion Matrix

	Not Churn	Churn
Churn	102	388
Not Churn	938	330



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

          0       0.81      0.91      0.86      1268
          1       0.66      0.43      0.52       490

avg / total       0.77      0.78      0.76      1758

Accuracy   Score :  0.7792946530147895
Area under curve :  0.6729511362904784
```

Fig 33 Classification Report for Random Forests Model

The model has 77.9% accuracy and AUC of 67.3% which is a better compared to other models.

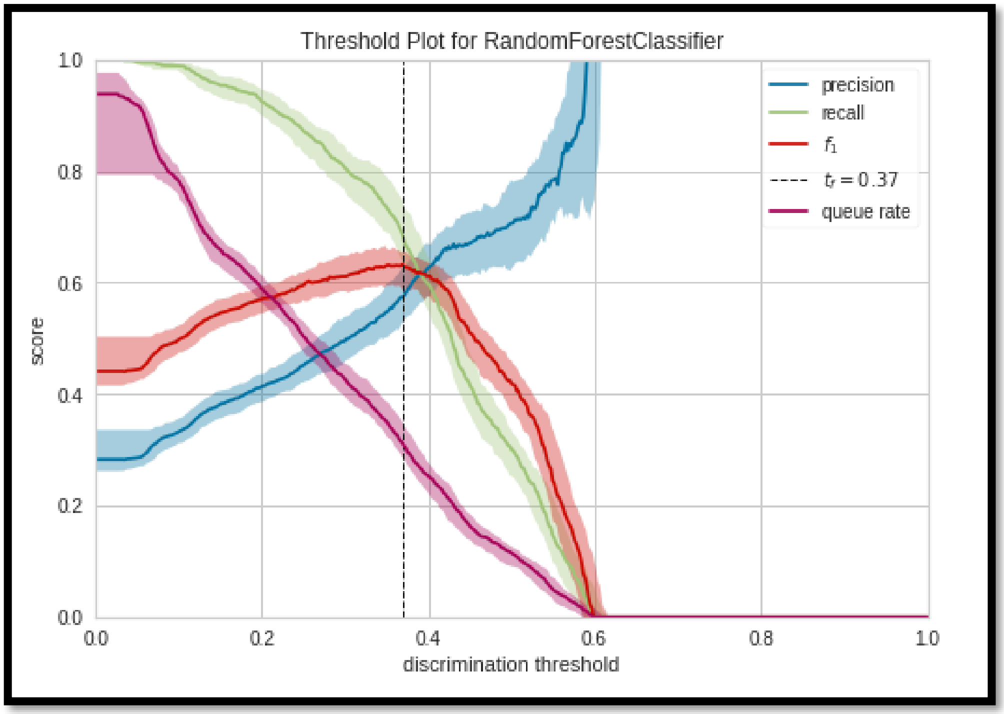


Fig 34 Threshold Plot for Random Forests Model

(Threshold= 0.37)

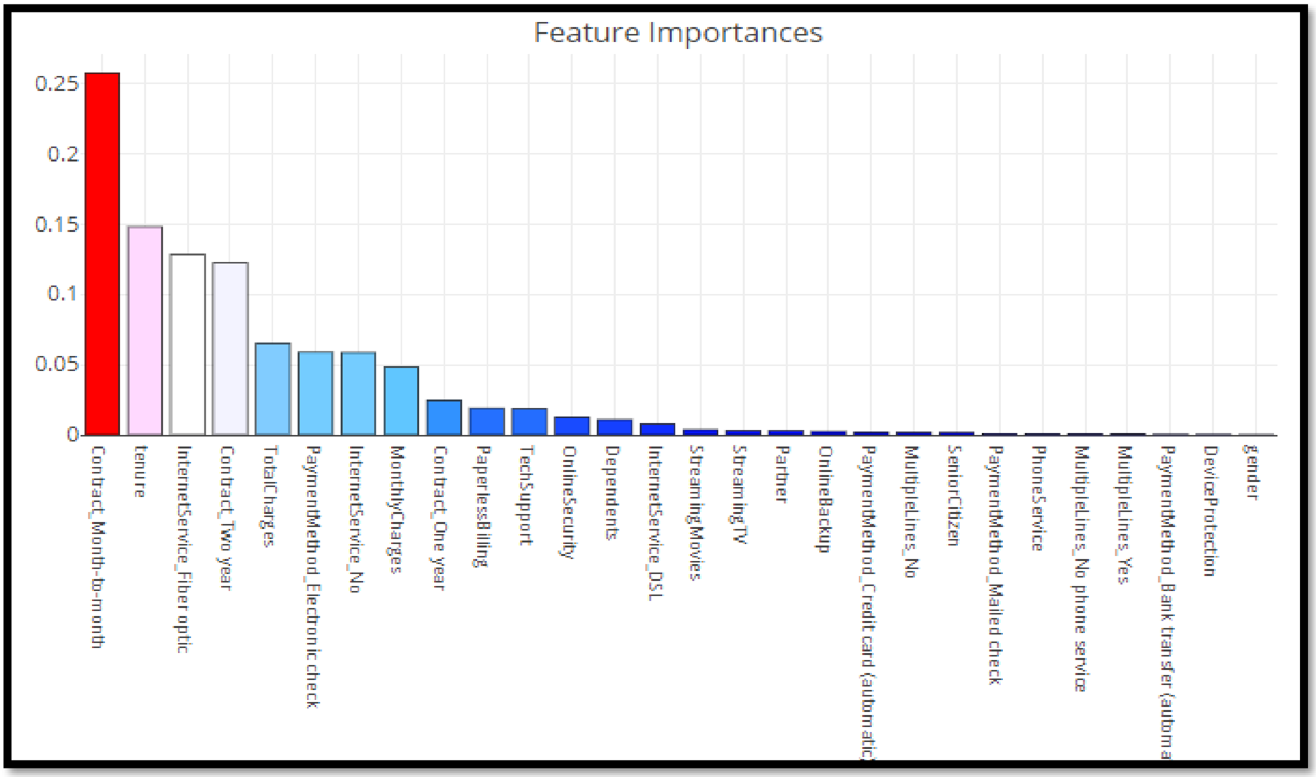


Fig 35 Feature Importance's for Random Forests Model

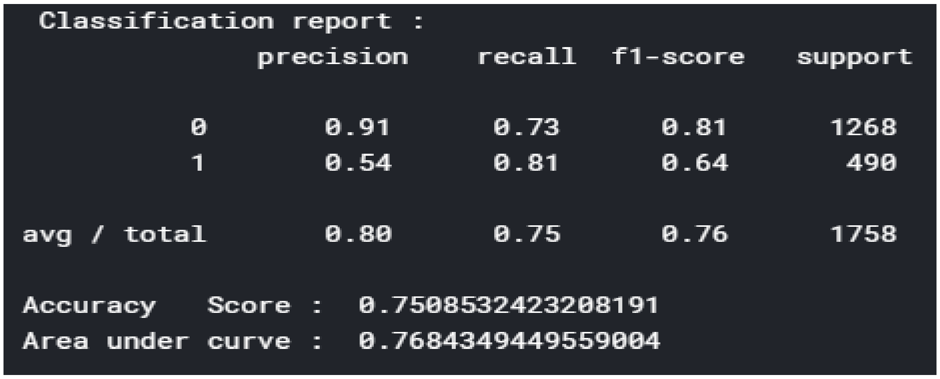
The most important predictors in predicting Customer attrition used in the Random Forest Classifier model are shown in the above Feature Importance plot.

5.7 Support Vector Machine (SVM)

- “Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space. where n is number of features you have) with the value of each feature being the value of a coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes

Table 3: Confusion Matrix

	Not Churn	Churn
Churn	94	396
Not Churn	924	344



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

          0       0.91      0.73      0.81      1268
          1       0.54      0.81      0.64       490

avg / total       0.80      0.75      0.76      1758

Accuracy   Score :  0.7508532423208191
Area under curve :  0.7684349449559004
```

Fig 36 Classification Report for SVM Model

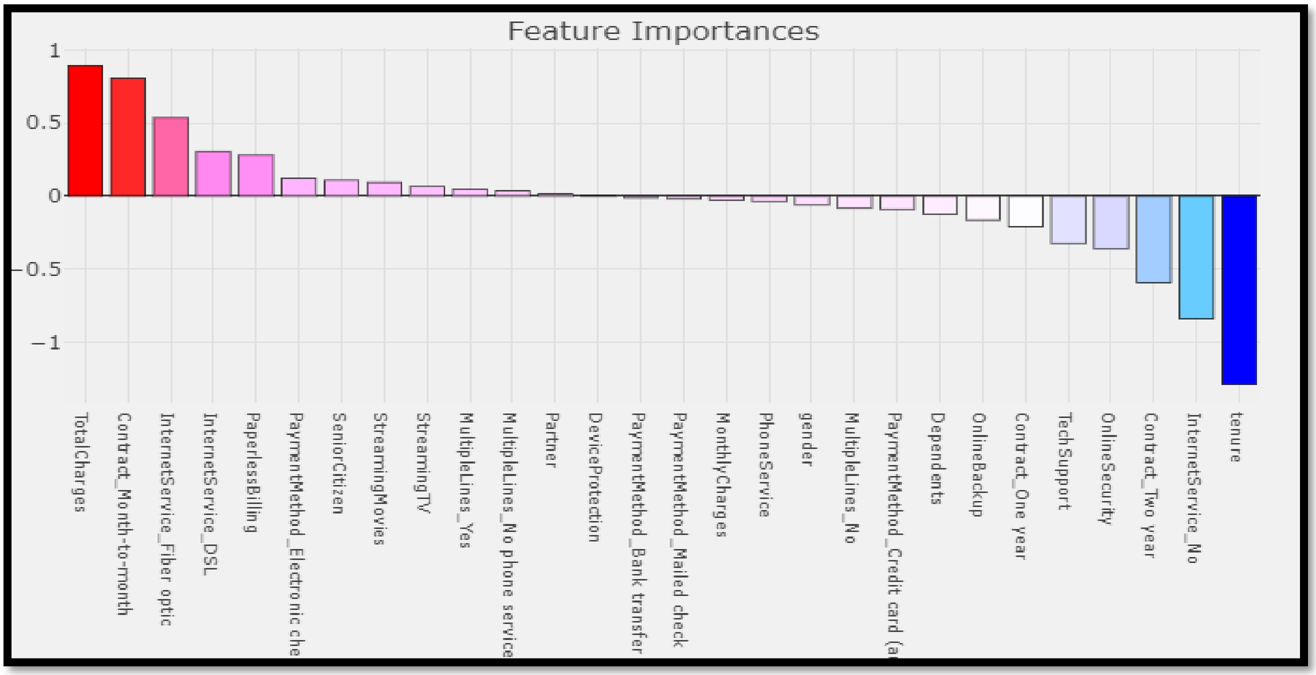


Fig 37 Feature importance’s for SVM Model

The most important predictors in predicting Customer attrition used in the Support Vector Machine model are shown in the above Feature Importance plot.

6 Comparison of Models and Managerial implications

- A comparative plot of the confusion matrices for all the models we have built as part of our predictive analysis is given below.

- We have assumed that the cost of False Positives is equal to the cost of False Negatives while calculating our classification Threshold and other performance parameters.

- In cases where the cost of False Negative is more, we can use the "Random Forest Classifier" because it has the lowest number of False Negatives whereas in the case where the cost of False Positive is more, we can use the "Support Vector Machine Classifier" because it has the lowest number of False Positives.

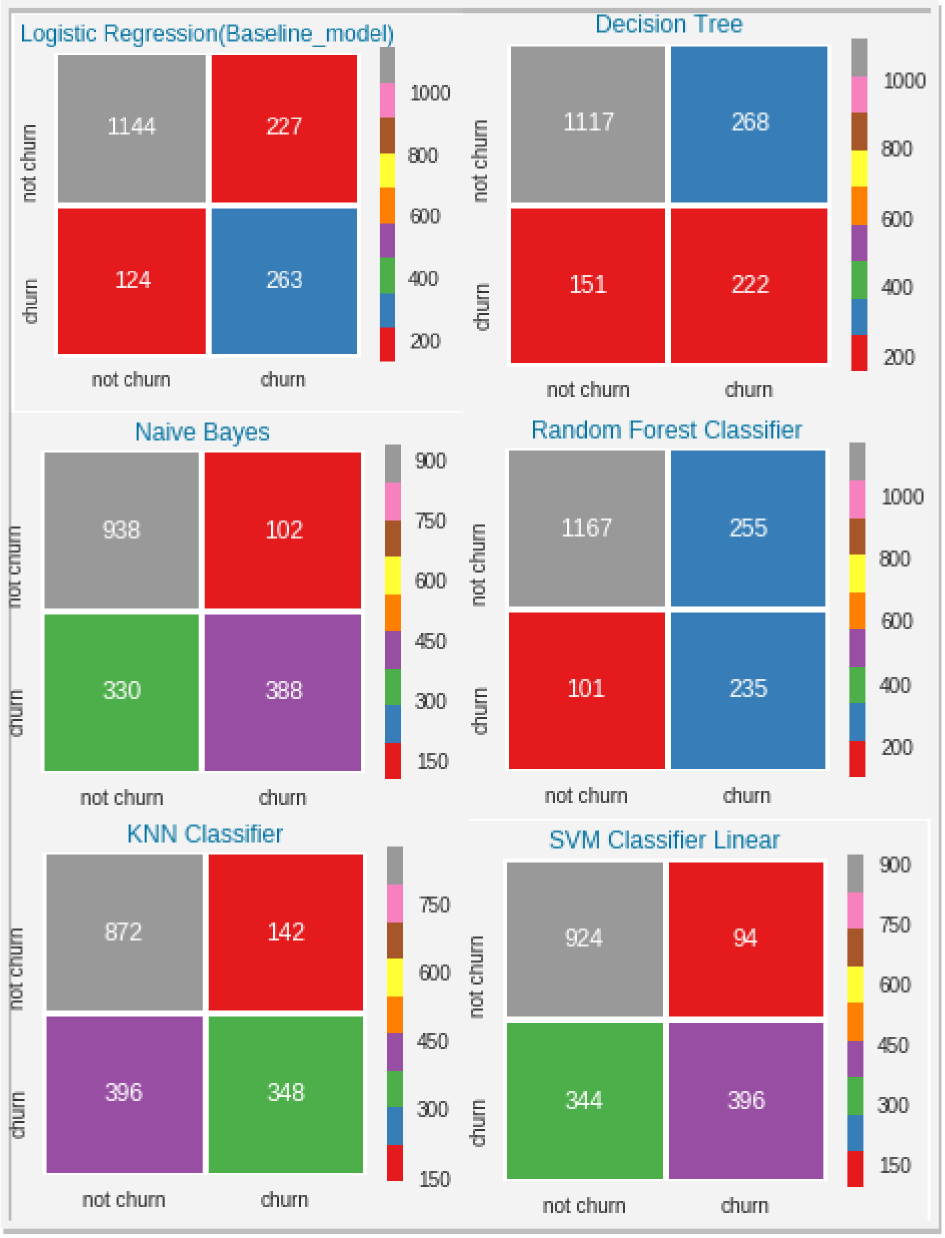


Fig 38 Visual Comparison of Confusion Matrices of all Models

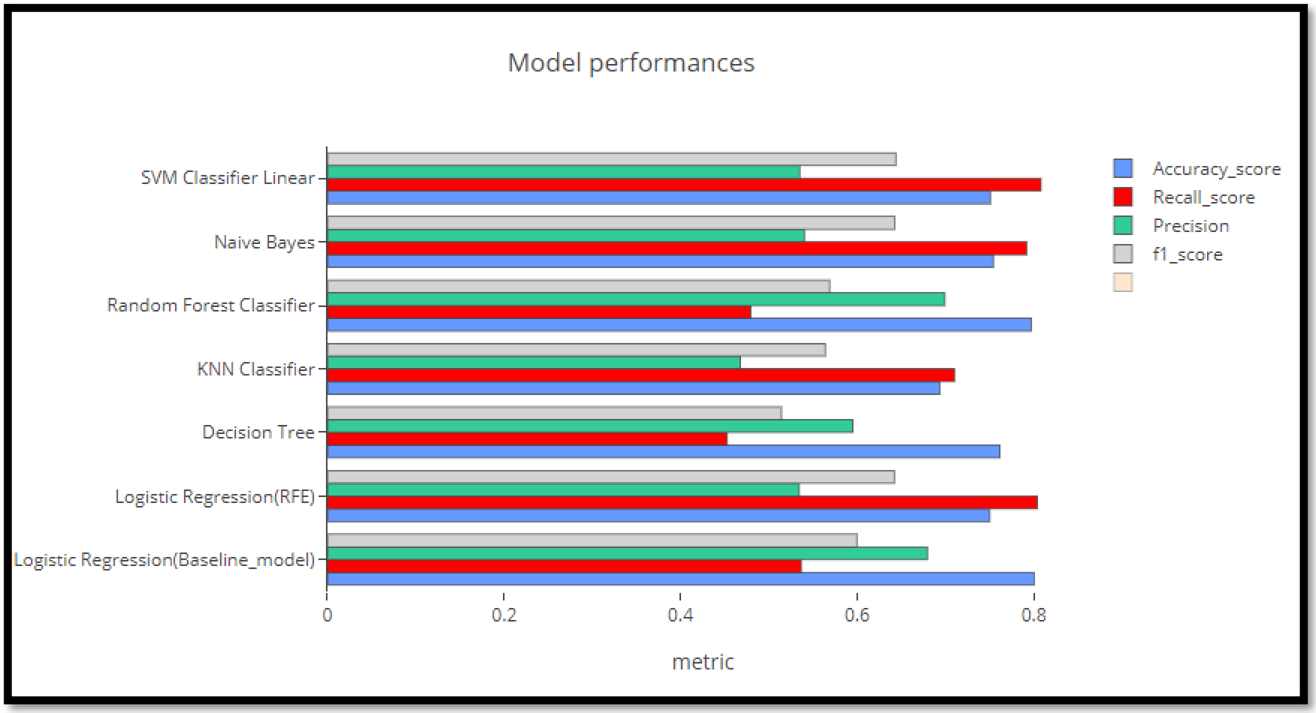


Fig 39 Visual Comparison of Performance Metrics of all Models

After a thorough analysis and comparing the results obtained from modeling methods such as logistic regression, KNN Classifier, Decision trees, Random Forest Classifier, Support Vector Machine and Naïve Bayes Classifier we can conclude that there is no major difference in their performance for this dataset. By comparing the confusion matrices of all the above we can see that Logistic regression (80% accuracy) performs slightly better than Random Forest Classifier (79% accuracy). Also, Logistic Regression (RFE), Random Forest Classifier and Support Vector Machines Classifier have the highest AUC, Precision Score and F1-Score. Hence, we can say that these three models provide the best predictive capabilities depending on our business use case. In conclusion, it can be said that without considering the business use case into perspective, logical conclusions cannot be made which is the best model. Different model performances will vary depending on the business objective of the analysis.

7 Conclusions

The initial part of our project mainly dealt with exploratory data analysis of the telecom company’s customer data. We found that senior citizens have higher churn rate. We also observed that Churn rate is higher in initial period of tenure of customers. This led us to give one of the most important recommendation targeting customers with low tenure. We also found that customers with Month-to month contract have the highest churn rate. Finally, we observed that Fiber optics customer churn more frequently than other group customer.

Using the predictive models, we have developed in Section 5 of our project one can find out which customer is most likely to churn and with the help of this information, the telecom company can target that customer with a focused customer retention strategy. Based on our analysis throughout this project we also offer the final marketing strategies to the telecom company as follows:

1. Target new customers to onboard them with the company on a 1 year or 2-year contract with special incentives, since getting new customers to sign a long-term contract will make them less likely to leave the company.
2. Pitch special bundle package offers to new customers in order to minimize future chance of churn since we have found out that customers with Fiber Optic internet are leaving the company at a higher rate and hence it would be beneficial if we offer them a bundled internet package.
3. Minimize total charges levied on customers with less tenure, since we found that very high costs levied on customers is a major factor leading customers to leave the company.
4. Offer high discount packages to customers most likely to churn as of today in order to stop the immediate attrition of customers and retain its existing customer base.

REFERENCES

1. https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/

2. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An **Introduction to Statistical Learning** : with Applications in R. New York :Springer, 2013. Print.

3. VanderPlas, J. (2016). Python Data Science Handbook. O'Reilly Media, Inc.

4. Beazley, D. and B.K. Jones (2013). Python cookbook, 3rd Edition. O'Reilly Media, Inc.

5. https://www.kaggle.com/pavanraj159/telecom-customer-churn-prediction

6. https://www.kaggle.com/bandiatindra/telecom-churn-prediction