Customer Churn Prediction in Telecom using Machine Learning

DISSERTATION

Submitted in partial fulfillment of the requirements of

MTech Software Engineering Degree Programme

By

Samarth Malhotra

2018AP04535

Under the supervision of

Mayank Jain

Principal Software Developer

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

Aug, 2021

DSE CL ZG628T DISSERTATION

Customer Churn Prediction in Telecom using Machine Learning

Submitted in partial fulfillment of the requirements of the

M. Tech. Data Science and Engineering Degree Programme

By

Samarth Malhotra

2018AP04535

Under the supervision of

Mayank Jain

Principal Software Developer

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

Aug, 2021

ACKNOWLEGDEMENTS

I would like to take this opportunity to express my gratitude to my supervisor **Mayank Jain**, who provided me the opportunity to do this wonderful project. His experience and guidance were great help in completing this research work.

My thanks are due to the Internal Supervisor **Prof M J Shankar Raman** for guiding me throughout the course of this project. Their constant support and direction helped me reach the target.

I owe a word of thanks to my family and friends for their support.

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the Dissertation entitled **"Customer Churn Prediction in Telecom using Machine Learning"** and submitted by Mr. **Samarth Malhotra** ID No. **2018AP04535** in partial fulfillment of the requirements of DSE CL ZG628T Dissertation, embodies the work done by him/her under my supervision.

Signature of the Supervisor

Place: Pune

Date: 6th Aug 2021

Name: Mayank Jain

Designation: Principal Software Developer

BITS ID No. 2018AP04535   **Name of Student:** Samarth Malhotra

Name of Supervisor: Mayank Jain

Designation of Supervisor: Principal Software Developer

Qualification and Experience: B.E. Computer Science

E- mail ID of Supervisor: Mayank.nirmal@gmail.com

Topic of Dissertation: Customer Churn in Telecom using Machine Learning

Name of First Examiner: Prof. M. J. Shankar Raman

Designation of First Examiner: _____

Qualification and Experience: _____

E- mail ID of First Examiner: _____

Name of Second Examiner: _____

Designation of Second Examiner: _____

Qualification and Experience: _____

E- mail ID of Second Examiner: _____

(Signature of Student)   (Signature of Supervisor)
Date: 26th May 2021   Date: 26th May 2021

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

Work Integrated Learning Programmes Division

I SEMESTER 2020-21

DSE CL ZG628T **DISSERTATION**

(EC-2 Mid-Semester Progress Evaluation Sheet)

Scheduled Month August:

NAME OF THE STUDENT:	**Samarth Malhotra**
ID NO.:	**2018AP04535**
Email Address:	**malhotra.samarth@hotmail.com**
NAME OF SUPERVISOR:	**Mayank Jain**
PROJECT TITLE:	**Customer Churn in Telecom using Machine Learning**

Evaluation Details

EC No.	Component	Weightage	Comments (Technical Quality, Originality, Approach, Progress, Business value)	Marks Awarded
1	Dissertation Outline	10%		
2.	Mid-Sem Progress			
	Seminar	10%		
	Viva	5%		
	Work Progress	15%		

	Supervisor	Additional Examiner
Name	Mayank Jain	
Qualification	BE Computer Science	
Designation & Address	Principal Software Developer E804, Palsh Society, Wakad, Pune, Maharashtra, 411057	
Email Address	Mayank.nirmal@gmail.com	
Signature		
Date	26th May 2021	

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

Work Integrated Learning Programmes Division

I SEMESTER 2020-21

Supervisor's Rating of the Technical Quality of this Dissertation Outline

EXCELLENT / GOOD / FAIR/ POOR (Please specify): EXCELLENT

Supervisor's suggestions and remarks about the outline (if applicable).

Date: 26th May 2021   (Signature of Supervisor)

Name of the supervisor: Mayank Jain

Email Id of Supervisor: Mayank.nirmal@gmail.com

Mob # of supervisor: +91 8805413880

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes Division
I SEMESTER 2015-16

DSE CL ZG628T **DISSERTATION**

(Final Evaluation Sheet)

NAME OF THE STUDENT:	**Samarth Malhotra**
ID NO.:	**2018AP04535**
Email Address:	**malhotra.samarth@hotmail.com**
NAME OF THE SUPERVISOR:	**Mayank Jain**
PROJECT TITLE:	**Customer Churn in Telecom using Machine Learning**

(Please put a tick (α) mark in the appropriate box)

S.No.	Criteria	Excellent	Good	Fair	Poor
1	Work Progress and Achievements	✓			
2	Technical/Professional Competence	✓			
3	Documentation and expression	✓			
4	Initiative and originality	✓			
5	Punctuality	✓			
6	Reliability	✓			
	Recommended Final Grade	✓			

EVALUATION DETAILS

EC No.	Component	Weightage	Marks Awarded
1	Dissertation Outline	10%	
2	Mid-Sem Progress Seminar Viva Work Progress	10% 5% 15%	
3	Final Seminar/Viva	20%	
4	Final Report	40%	
	Total out of	100%	

	Supervisor	**Additional Examiner**
Name	Mayank Jain	
Qualification	BE Computer Science	
Designation & Address	Principal Software Developer E804, Palsh Society, Wakad, Pune, Maharashtra, 411057	
Email Address	Mayank.nirmal@gmail.com	
Signature		
Date	6th Aug 2021	

NB: Kindly ensure that recommended final grade is duly indicated in the above evaluation sheet. **POSTAL ADDRESS FOR ALL FUTURE CORRESPONDENCE. FILL IT UP NEATLY IN CAPITAL LETTER WITH PIN CODE ETC.**

Address: 202 Gautam Buddha Niwas, Banasthali Vidyapith, Dist. Tonk, Rajasthan

Pin Code: 304022

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

SECOND SEMESTER 2020-21

DSE CL ZG628T **DISSERTATION**

Dissertation Title   :   Customer Churn Prediction in Telecom using Machine Learning

Name of Supervisor   :   Mayank Jain

Name of Student   :   Samarth Malhotra

ID No. of Student   :   2018AP04535

Abstract

Customer churn is the likelihood of a customer to leave a brand, stop using its services and switching over to other providers. It is a major challenge in businesses with subscription-based model and has direct impact on the revenue of the company, especially in the telecom field. The cost of churn includes both the loss of revenue and the marketing costs involved in replacing those customers with new ones, therefore, predicting and preventing customer churn has a potential revenue source therefore the telecom companies must make an effort to retain their customers.

In the face of stiff competition in the market, the customers have very wide choice and often they switch over from one product to another, there is always a search for better options.

There can be several factors responsible for customer churn, including:

- The availability of quality services
- Low-cost alternatives
- Better features and content
- Customer experience
- Availability of self-service options
- Easy access to the maintenance staff
- Network coverage

Customer churn prediction modelling aims to understand the customer's behavior and attributes (gender, age, dependents, financial status), also the likelihood to switching of the brand, possible reasons and the remedial measures to retain the customer.

With a better understanding and an insight into potential customers leaving the brand in the volatile market condition, the brand can take a suitable action after the analysis which will lead to most retention impact on the customer.

Contents

List of Symbols and Abbreviations

ML   : Machine Learning
EDA   : Exploratory Data Analysis
RFC   : Random Forest Classifier
FE   : Feature Engineering

List of Tables

List of Figures

Chapter 1

Introduction

Customer churn is the likelihood of a customer to leave a brand, stop using its services and switching over to other providers. It is a major challenge in businesses with subscription-based model and has direct impact on the revenue of the company, especially in the telecom field. The cost of churn includes both the loss of revenue and the marketing costs involved in replacing those customers with new ones, therefore, predicting and preventing customer churn has a potential revenue source therefore the telecom companies must make an effort to retain their customers.

In the face of stiff competition in the market, the customers have very wide choice and often they switch over from one product to another, there is always a search for better options.

There can be several factors responsible for customer churn, including:

- The availability of quality services
- Low-cost alternatives
- Better features and content
- Customer experience
- Availability of self-service options
- Easy access to the maintenance staff
- Network coverage

The above-mentioned list is not exhaustive, the inventory could vary depending upon service provider and would require domain knowledge.

Customer churn prediction modelling aims to understand the customer's behavior and attributes (gender, age, dependents, financial status), also the likelihood to switching of the brand, possible reasons and the remedial measures to retain the customer.

With a better understanding and an insight into potential customers leaving the brand in the volatile market condition, the brand can take a suitable action after the analysis which will lead to most retention impact on the customer.

Objective

The main objective of the present project is to design a churn prediction model that could help telecom operators to foresee the customer behavior and accurately predict the customers who are likely to churn.

In order to know the customer behavior, the relevant historical data will be used and as the current research in the field confirms machine learning could be efficiently applied to predict the customer churn and take the retention measures.

Principal objectives:

a) Create visualizations to showcase how each feature is affecting the target class
b) Create multiple machine learning models to predict the target variable and evaluate them with multiple metrics (AUC Score, precision, recall, f1 score)
c) Identify the features which are important for the chosen model

Uniqueness of the project

In many organizations the customer churn is reactive in the sense that when customer calls to end the subscription only then offers are rolled out to retain the customers.

In this project we are aiming to make this process proactive by actively predicting unhappy customers in advance and making necessary adjustments to retain them.

Benefit to the organization

Oracle provides end to end cloud solutions to telecommunication providers, it spans everything from capturing the network calling data to billing and processing payments to generating audit reports.

The project will directly benefit Oracle in providing up to date information about customer churn to its client i.e., telecom operators and operators in turn will ensure that the customer churn could be prevented in time by opting for retention strategies which will have direct impact on their revenue.

Chapter 2

Data Acquisition

Here we are using sample dataset provided by IBM community, the dataset contains information about a fictional telecommunication vendor that provides phone and internet services in California. It provides information about customers who have left, or stayed, or signed up for the service. Multiple important demographics are included for each customer, as well as Customer Lifetime Value (CLTV) index.

The dataset includes information about:
- Consumer who left within the last month – the attribute is Churn
- Services that each customer has opted for – phone, multiple lines etc.
- protection, tech support, and streaming TV and movies
- Consumer account information – how long they've been a customer, contract, payment etc.
- Demographic information – gender, partners, dependents etc.

Data Preprocessing
Handling Missing data

The dataset has some missing values for features Total Charges and Churn Reason, for records where Tenure Months is 0 there is no value for Total Charges, also for feature Churn Reason for all the data points where Churn Label is No. i.e. The customers who have not left the company there will be no Churn Reason. Before imputing any value for Total Charges, we are changing the data type of the feature from object to float.

Table 1 : Features with missing values

Name	dtype	Missing	Unique
Total Charges	object	11	6531
Churn Reason	object	5174	20

We can impute the values for Churn Reason for the missing values as Not Available as the churn didn't happen also upon closer observation, we can observe that there is a strong correlation between the numerical features Tenure Months, Monthly Charges and Total Charges. If we calculate correlation coefficient for (Tenure Months x Monthly Charges) and Total Charges it is 0.9995605537972277 and therefore we can impute the missing values of the feature Total Charges with (Tenure Months x Monthly Charges).

Removing Unnecessary Features

We are removing unnecessary features such as latitude longitude, zip code, country, state and churn score as the data is only for United States of America also its for state of California, we have removed latitude, longitude and information as we will be using City to identify the location. We are removing the churn score as its not part of actual data but generated by IBM SPSS tool.

Data Exploration
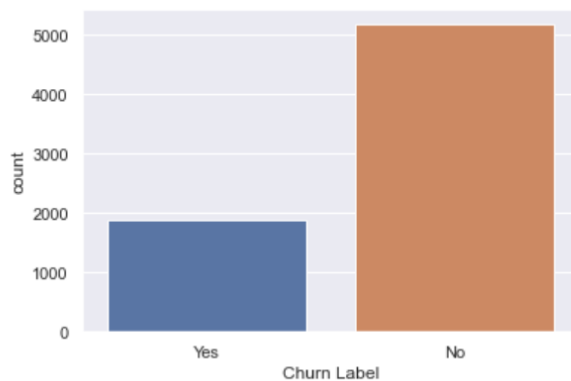We are firstly performing Univariate Analysis

Figure 1: Distribution of Target Variable

Churn Label is the target feature and as we can observe from above plot the dataset is imbalanced, there are more data points for customers who have not churned while performing model creation and any analysis we have to keep this in mind   .
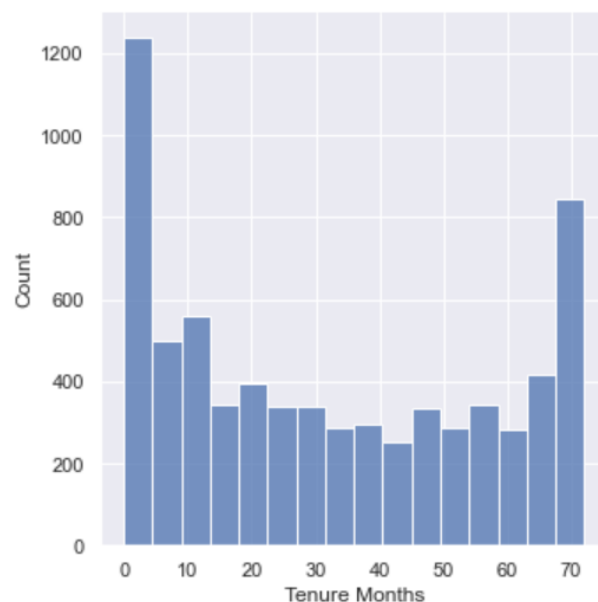

Figure 2: Distribution of Tenure in Months

From the above numerical feature Tenue Months, we can observe that it is a bimodal distribution, which means there are two different kinds among customers and we can find out what services are kept by those who stay more than 70 months.

Secondly, we will perform bivariate analysis of numerical features, we will see how different numerical features are distributed in terms of target variable.

Figure 3: Kernel Density Estimation of Tenure Months

From above kernel density estimation plot we can see the probability density function of numerical feature Tenure Months. It can be observed that customers who have recently joined are more likely to churn.
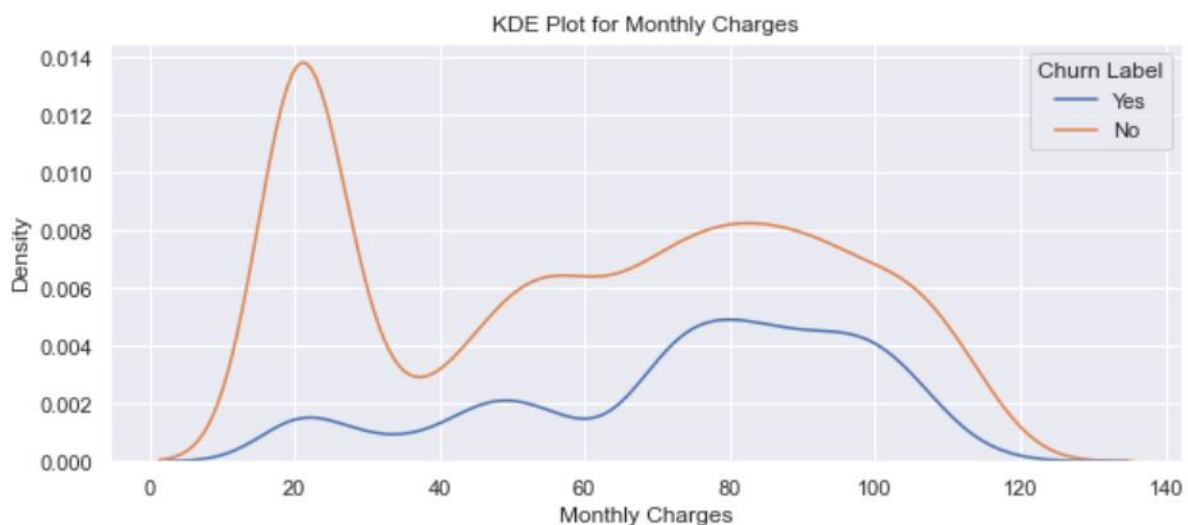

Figure 4: Kernel Density Estimation of Monthly Charges

From above kernel density estimation of numerical feature Monthly charges, we can observe that customers with higher monthly charges are more likely to churn than those with lower monthly charges.

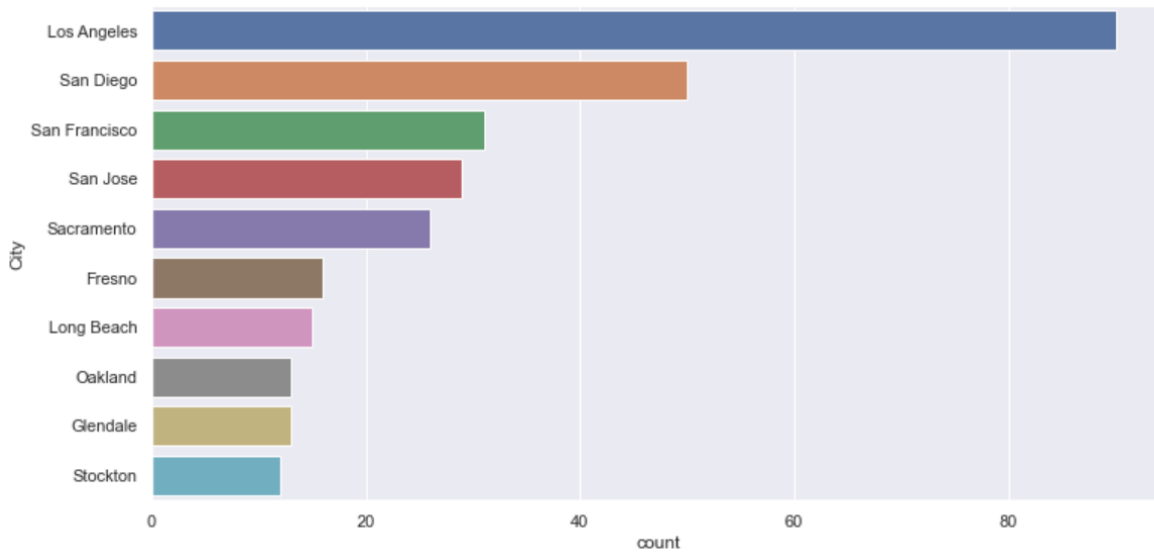Now, we will see how categorical features are distributed in terms of target feature

Figure 5: Distribution of customer who churned for different cities

From the above plot it is easily visible that city Los Angeles, San Diego, San Francisco, San Jose, Sacramento accounts for most churn customers and therefore we need to investigate further why so many customers are leaving from these particular locations   .
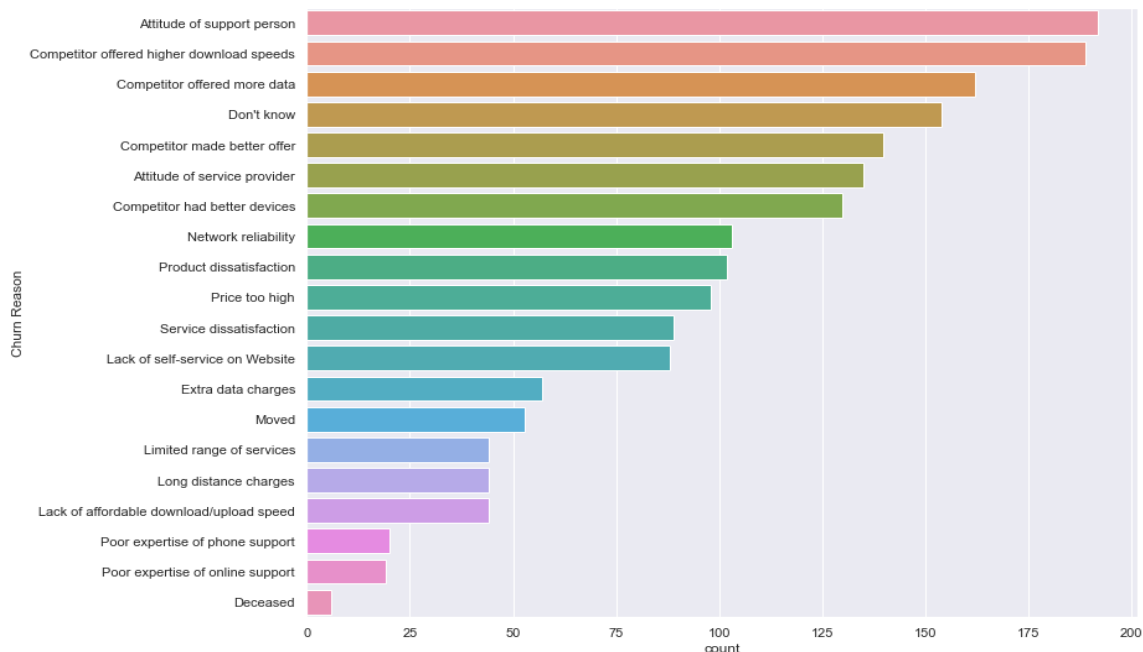

Figure 6: Distribution of customer who churned for different reasons

From the above distribution plot of Churn Reason, it can be observed that the reason of highest churn among customers is dissatisfaction from support services and internet data and speed and therefore remedial actions can be taken to improve support service and internet speed also we can come up with better internet plans.

Figure 7: Distribution of customer for gender
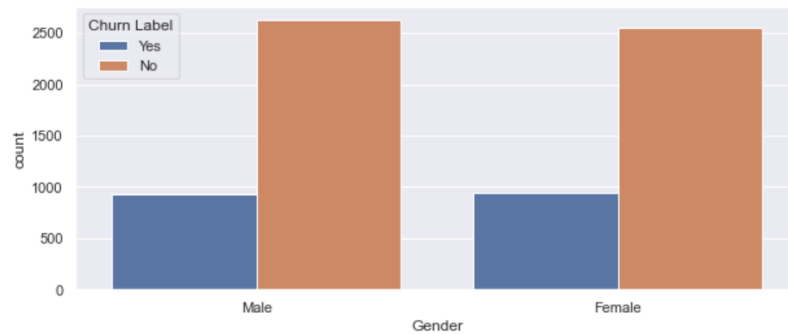
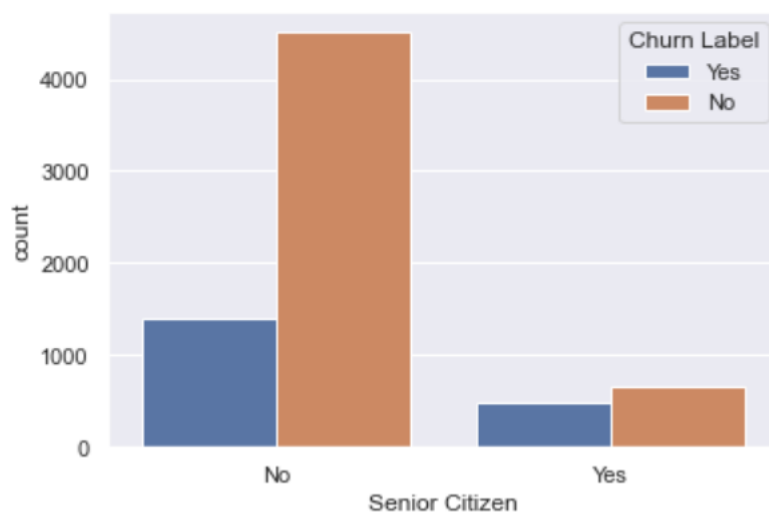From the above figure we can see that feature gender has no influence on customer churn

Figure 8: Distribution of customers for senior citizens

From the above distribution plot, we can observe that, even though there are only 16 % senior citizen among total customers but the churn rate among senior citizens in 41.6 % compared to 23.6 % in younger customers.

Similarly,
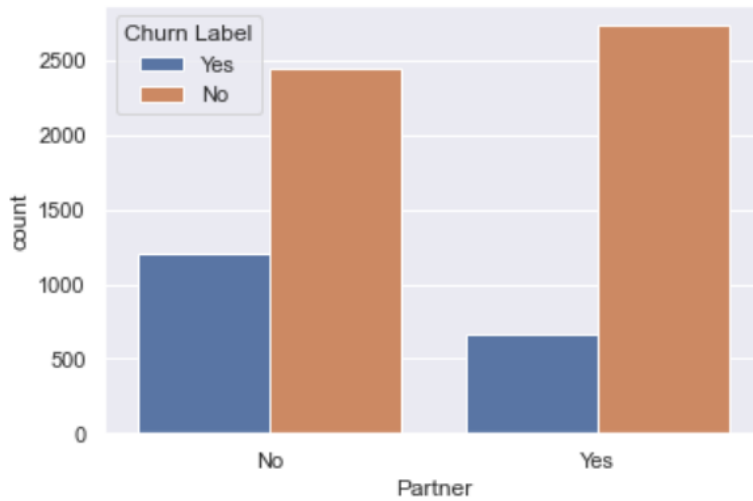
Figure 9: Distribution of customers for customer with partners

From above distribution plot it is evident that customer without partners is more likely to churn in comparison to customers with partners.
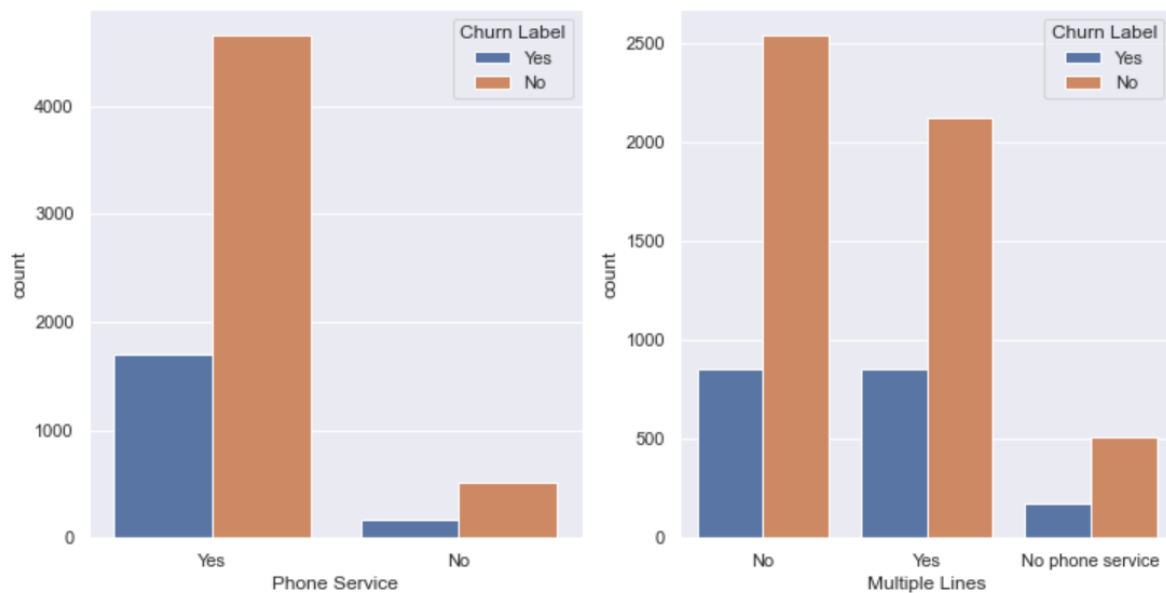

Figure 10: Distribution of customers for Phone Service and Multiple Lines

From the above distribution plots, it is evident that customer with no phone service is less and customers with multiple lines have slightly higher churn.
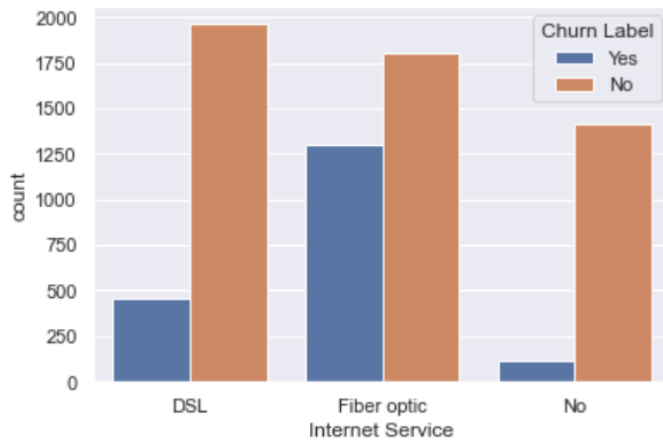
Figure 11: Distribution of customers for fiber optic service

It can be observed from above plot that customers without internet have very low churn, also customers with fiber optic cable are more likely to churn than customers with DSL connection.
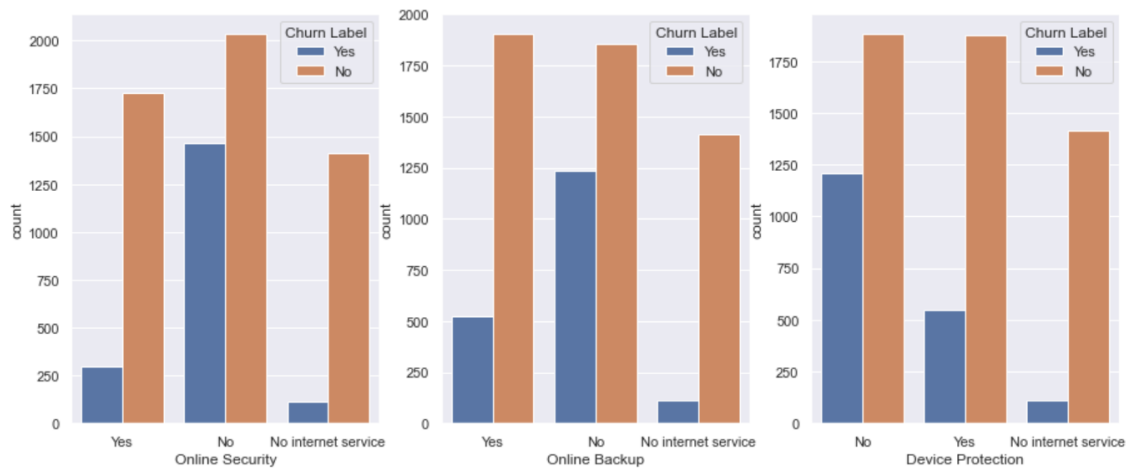
Figure 12: Distribution of customers for online security, online backup and device protection

From above plot it is evident that

- customers without internet are less likely to churn
- Customer with online security is less likely to churn
- Customers with online backup are less likely to churn
- Also, customers with device protection are less likely to churn
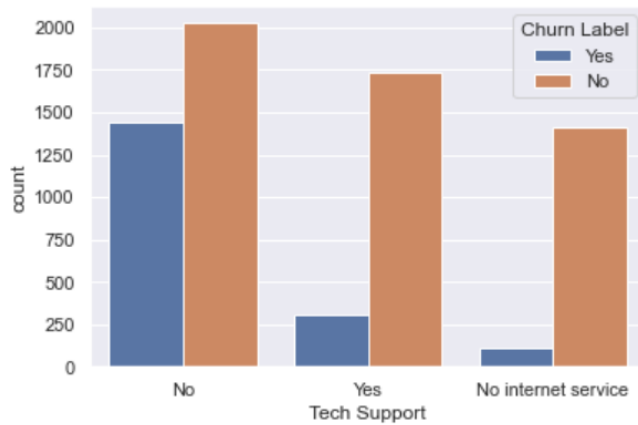
Figure 13: Distribution of customers for tech support

From the above plot we can see that customers with tech support are less likely to churn

Similarly,
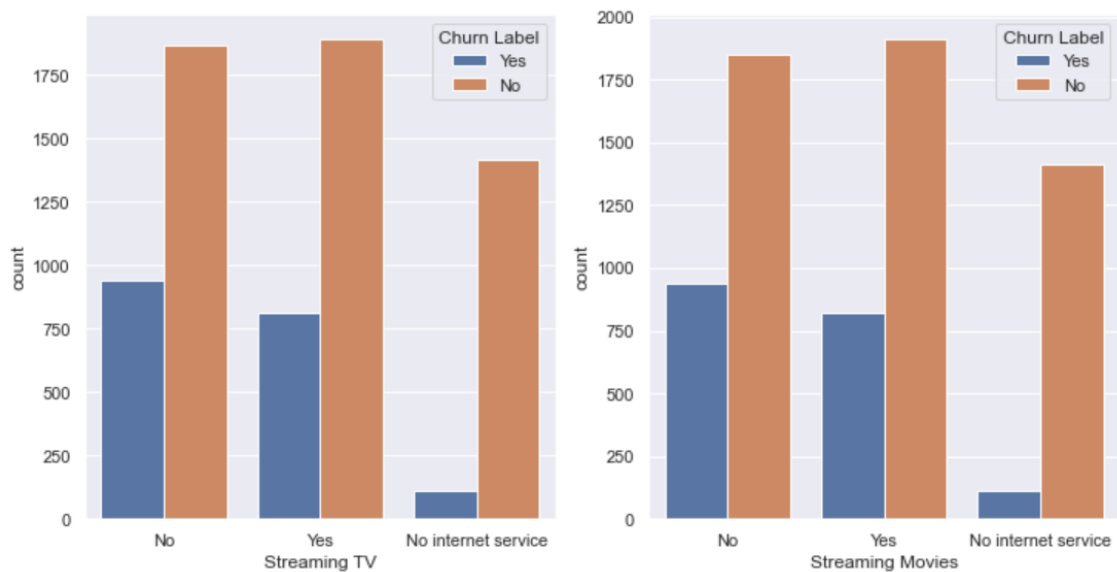
Figure 14: Distribution of customers for steaming tv and streaming movies

From above plot it is very evident that customers with Streaming TV and Streaming Movies are less likely to churn.
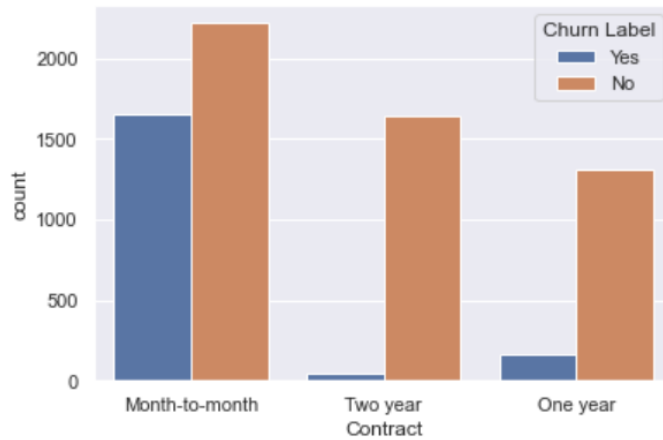
Figure 15: Distribution of customers for contract

From the above graph we can observe that customers one-year and two-year contracts are less likely to churn.
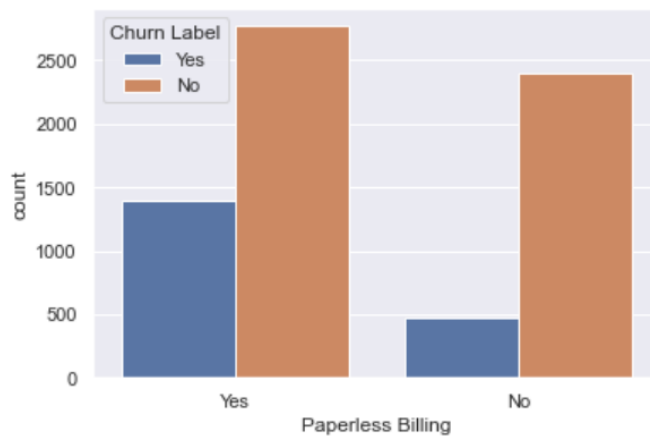
Figure 16: Distribution of customers for paperless billing

From above plot we can observe that customers with paperless billing are more likely to churn as compared to other customers.
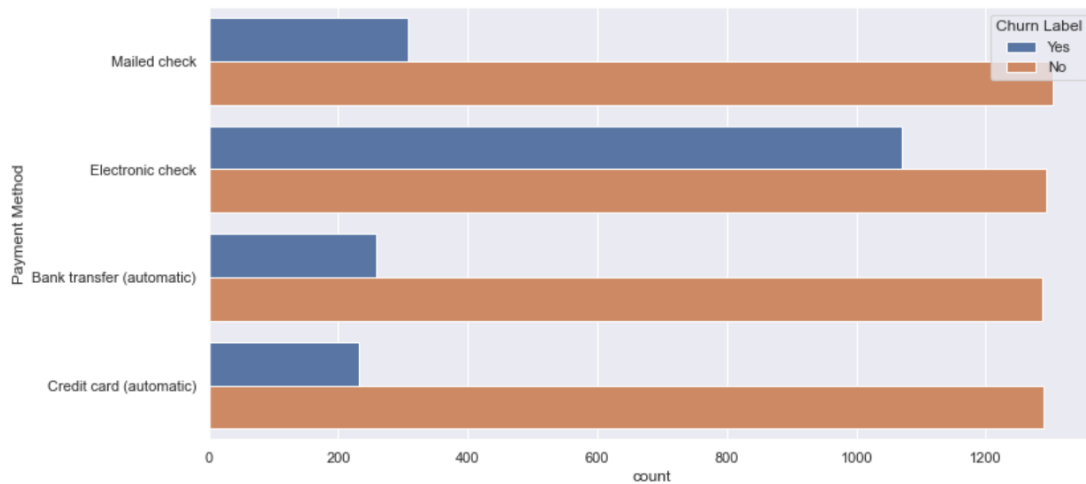
Figure 17: Distribution of customers for payment method

From the above plot we can observe that customers with electronic check are more likely to churn compared to other payment methods.

Now we will check for correlation between features, checking correlations is an important part of the exploratory data analysis process. The following analysis is used to decide which features affect the target variable the most, and are used in predicting target variable. It is a commonly used method for feature selection. Here we have divided the dataset into two sets one for numerical features and one for categorical features. For numerical features Tenure Months, Monthly Charges, Total Charges and CLTV we have created correlation matric and plotted heatmap.
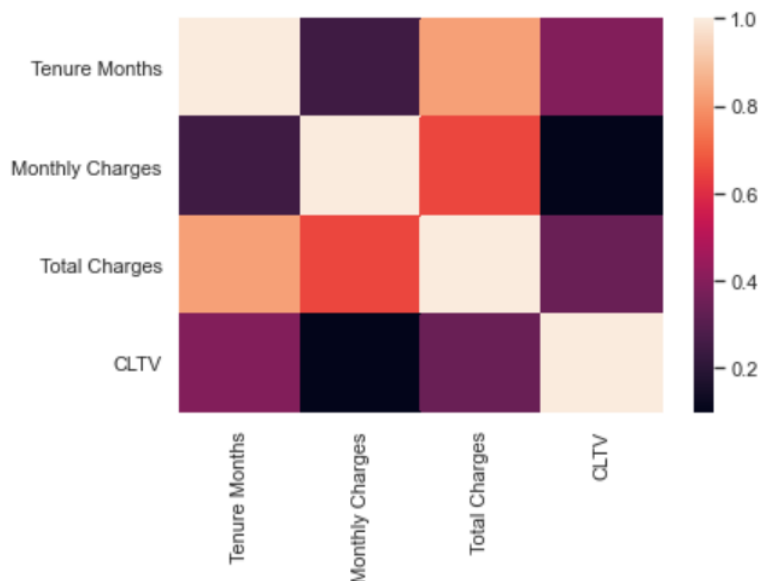
Figure 18: Correlation between numeric features

From the above heatmap it is very clear that Tenure Months and Total charges are highly correlated, similarly for categorical features we have created a separate dataset and created correlation matrix.
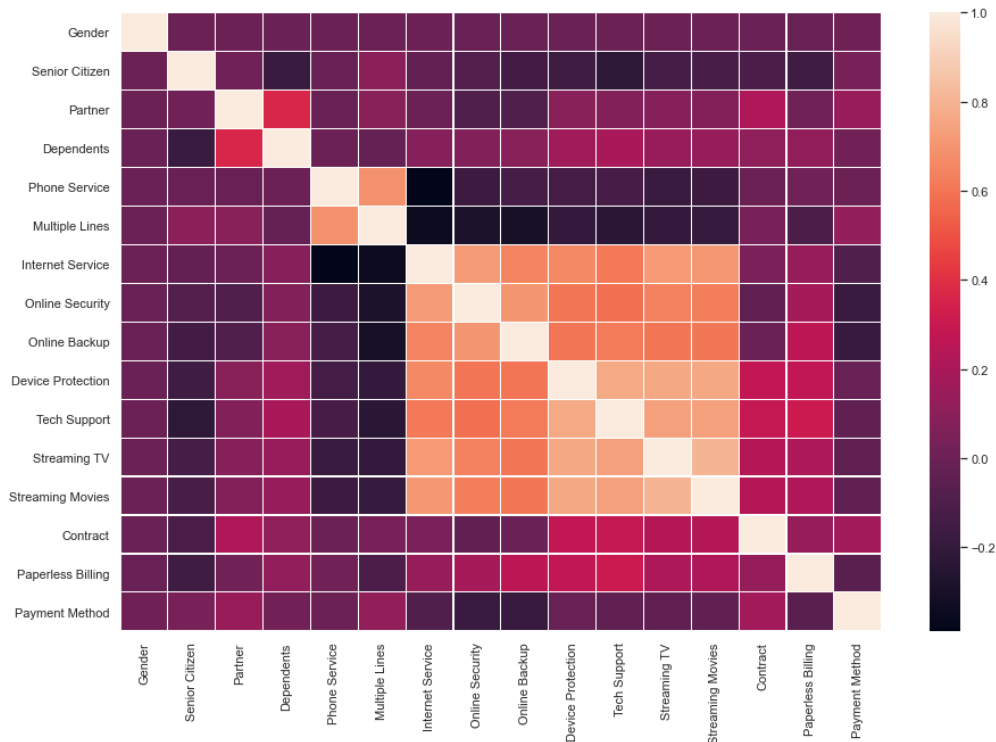
Figure 19: Correlation between categorical features

Form the above heatmap we can observe that Phone Service and Multiple Lines are correlated, similarly, internet Service, Online Security, Device Protection, Tech Support, Streaming TV and Streaming Movies are correlated.

Now we try to find the feature importance using Random Forest Classifier here, we have one-hot encoded the categorical features and dropped the features which are not required for analysis and also performed hyper parameter tuning by applying model in various settings.
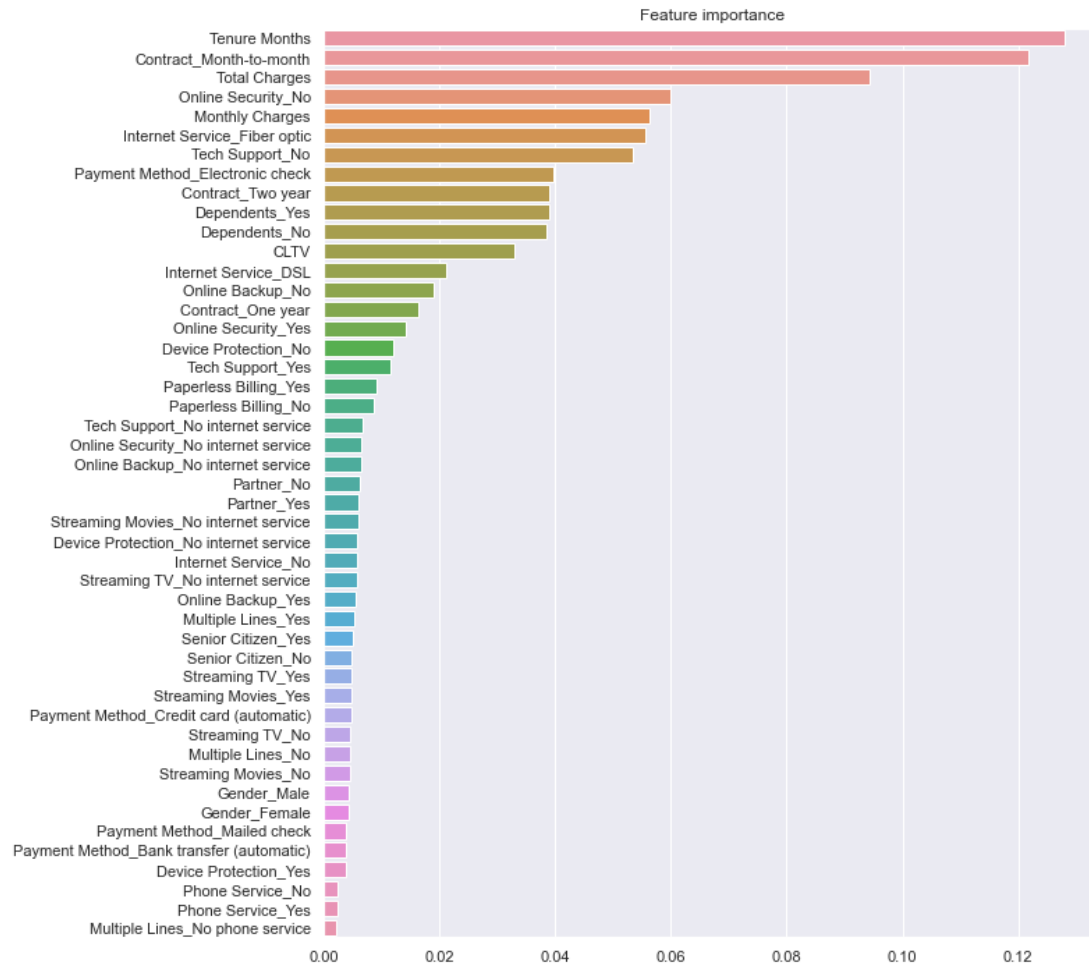
Figure 20: Feature Importance using Random Forest Classifier

As we have also observed in exploratory data analysis numerical features Tenure Months, Monthly Charges, and Total Charges are very important features to predict customer churn also categorical features Contract Mouth-to-mouth which are highly likely to churn. These results obtained are in line with the results obtained in exploratory data analysis.

Model creation and evaluation

For this project we have created multiple machine learning model and also performed comparative study.

Logistic Regression

We started with the most commonly used model for classification i.e., Logistic Regression, Logistic function comes from Sigmoid family of functions and assign probability for the class for the given set of inputs. The Logistic Regression is a discriminative classifier means it learns the

boundaries between the classes. It makes predictions based on conditional probability. The learning algorithm works as we train the model across the data points and adjust the parameters using the training labelled data and then again test the model using a separately held out data called the test data.

Before training the model, we scaled (standardized) the features of the dataset using min max scaler, it helps normalize the data within a particular range it also helps in speeding up the calculations.

The train-test data is split in 70 - 30 proportion and then the model is trained with training data. As the distribution of target feature is un-balanced, we have used class_weight='balanced' (It penalizes mistakes in sample of class[i] with class_weight[i] instead of 1. So higher class-weight means we want to put more emphasis on a class which is less in proportion.

We get accuracy score of approximately 0.755, Here accuracy score is not a good measure as the dataset is imbalanced means that there are more records which are not customer churn and therefore model is trained more to classify customer which did not churn.

```
In [72]:    1  # Confusion matric
            2  confusion_matrix(y_test, y_pred_test)

Out[72]:  array([[1123,  401],
                 [ 117,  472]], dtype=int64)

In [73]:    1  # View confusion matrix for test data and predictions
            2  metrics.plot_confusion_matrix(model, X_test, y_test, cmap="Blues")

Out[73]:  <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x2658a50cca0
          >
```
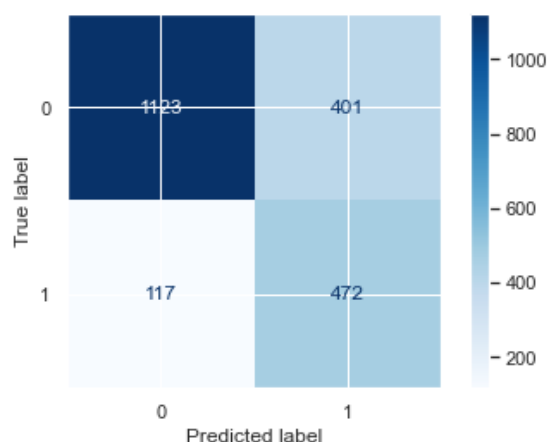

Figure 21: Confusion Matrix of Logistic Regression Classifier

Therefore, finding confusion matrix is a way to show how many of a classifier's predictions were correct and incorrect, here the rows represent the true labels and columns represents the predicted labels. values on the diagonal represent the number of times where the predicted label matches.

Also, the classification report is shown below:

Table 2: Classification Report of Logistic Regression Classifier

	precision	recall	f1-score	support
0	0.91	0.74	0.81	1524
1	0.54	0.80	0.65	589
accuracy			0.75	2113
macro avg	0.72	0.77	0.73	2113
weighted avg	0.80	0.75	0.77	2113

The classification report provides various metrices to evaluate the classifier, like precision, recall, F1 Score etc.

Similarly, we have also computed ROC – AUC ~ 0.855
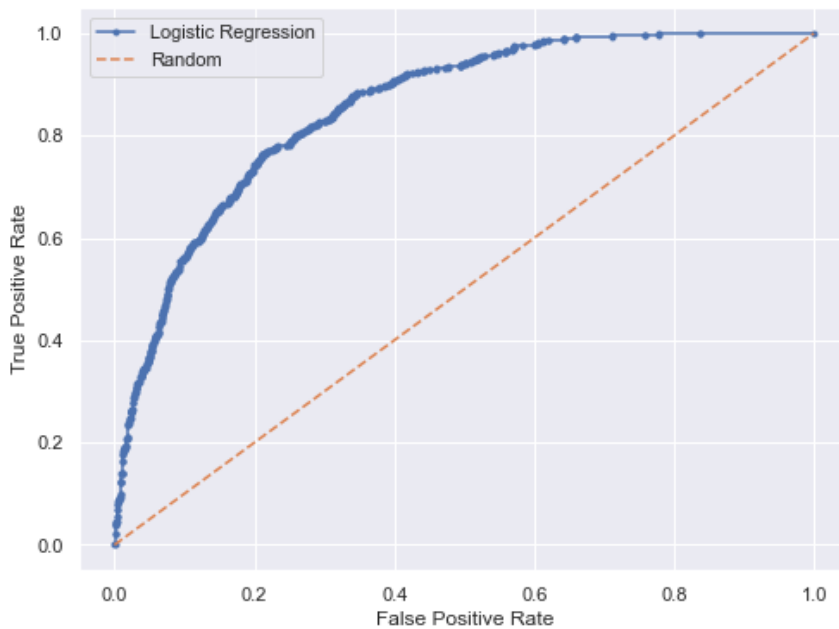
Figure 22: Receiver Operating Characteristics of Logistic Regression Classifier

Here the function inputs the true outcomes (0,1) from the test data and the predicted probabilities for the 1 class. The function outputs false positive rates for each threshold, true positive rates for each threshold and also the threshold values.

Naïve Bayes

Naïve Bayes classifiers in machine learning are a family of probabilistic classifiers based on Bayes Theorem with naive independence assumption between attributes. They are among the simplest Bayesian network models and with KDE they can achieve very good accuracy scores.

These are generative classifiers i.e.; it tries to model class or features of class means it models how a particular class would generate input data. When a fresh observation is given to the classifier, it predicts which class have most likely generated the given observation.

As with Logistic regression classifier before we have trained the classifier, we are normalizing the data and splitting it into train test and then fitting the model, for this project we have used Gaussian Naïve Bayes Classifier as it was providing better scores. The accuracy score we have received from this classifier is ~ 0.72. Also, the confusion matrix, classification report and the ROC-AUC is mention below.

```
1  # Confusion matric
2  confusion_matrix(y_test, y_pred_test)
```

```
array([[1038,  486],
       [ 105,  484]], dtype=int64)
```

```
1  # View confusion matrix for test data and predictions
2  metrics.plot_confusion_matrix(model, X_test, y_test, cmap="Blues")
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x2658a531250
>
```
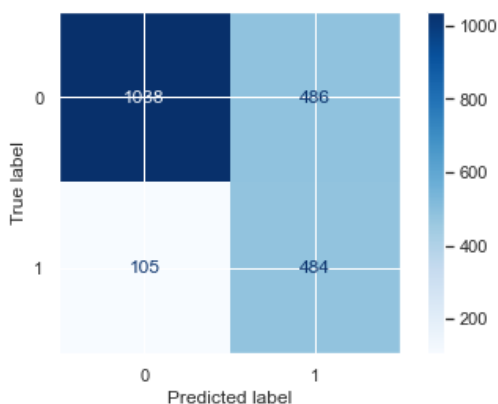

Figure 23: Confusion Matrix of Naive Bayes Classifier

Similarly, the classification report and ROC-AUC are shown below:

Table 3: Classification Report of Naïve Bayes Classifier

	precision	recall	f1-score	support
0	0.91	0.68	0.78	1524
1	0.50	0.82	0.62	589
accuracy			0.72	2113
macro avg	0.70	0.75	0.70	2113
weighted avg	0.79	0.72	0.73	2113
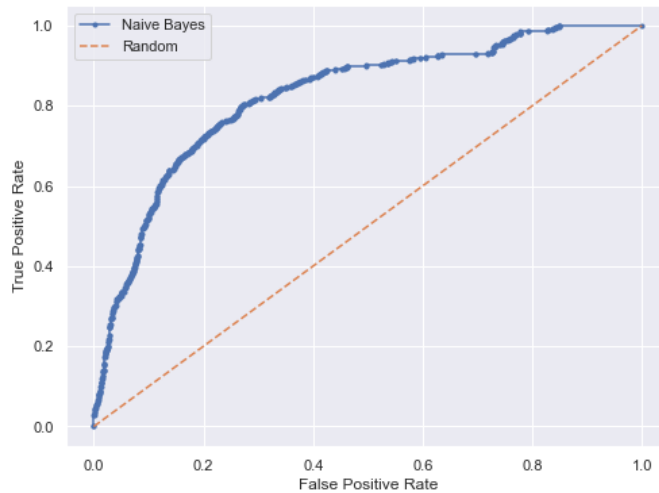
Figure 24: Receiver Operating Characteristics of Naive Bayes Classifier

We have also included some ensemble methods which use bagging and boosting to see what performance we get for our dataset, to start with we used Random Forest Classifier.

Random Forest

Random Forest is an ensemble classifier which basically means it uses many base classifiers, here the base classifier mostly used is decision tree classifier. Random Forest uses bagging technique, it trains many classifiers in parallel, there is no interaction between these trees while building the trees. Once all the trees are built then voting or average is taken across all the trees.

Similar to previous methods before training random forest classifier we have normalized the data and split it into train and test. The model is trained on training data and is evaluated using a

separately held out test data. Following are the performance metrics we are getting starting with accuracy score of ~ 0.766. Further metrics like precision, recall, F1 Score and ROC-AUC are shown below:

```
1  # View confusion matrix for test data and predictions
2  confusion_matrix(y_test, y_pred_test)
```

```
array([[976, 318],
       [ 94, 373]], dtype=int64)
```

```
1  metrics.plot_confusion_matrix(model, X_test, y_test, cmap="Blues")
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x26589d644f0
>
```


Figure 25: Confusion matrix for Random Forest Classifier

Similarly, Classification report for Random Forest classifier is shown below

Table 4: Classification Report of Random Forest Classifier

	precision	recall	f1-score	support
0	0.91	0.75	0.83	1294
1	0.54	0.80	0.64	467
accuracy			0.77	1761
macro avg	0.73	0.78	0.73	1761
weighted avg	0.81	0.77	0.78	1761

```
Random: ROC AUC=0.500
Random Forest Classifier: ROC AUC=0.854
```
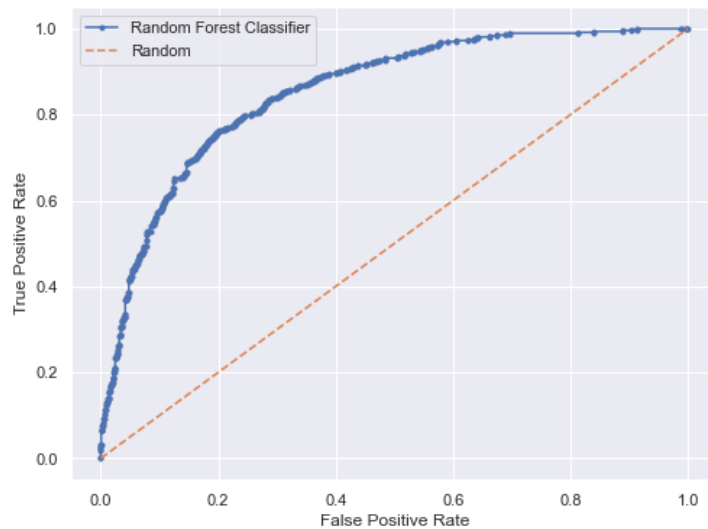

Figure 26: Receiver Operating Characteristics of Random Forest Classifier

At last, we have also trained a boosting based ensemble method called Gradient Boosting Classifier.

Gradient Boosting

Gradient Boosting is an ensemble classifier which basically means it uses many base classifiers, here and the base classifier mostly used is decision tree. Gradient Boosting uses boosting technique, it runs many classifiers in sequence i.e., the output of one is the input to other, this is achieved by creating a model from the training data, and then creating a second model to correct the errors from the first model. The models are added until the training set is predicted or a maximum number of models are added.

Similar to previous methods before training random forest classifier we have normalized the data and split it into train and test. The model is trained on training data and is evaluated using a separately held out test data. Following are the performance metrics we are getting starting with accuracy score of ~ 0.81. Further metrics like precision, recall, F1 Score and ROC-AUC are shown below:
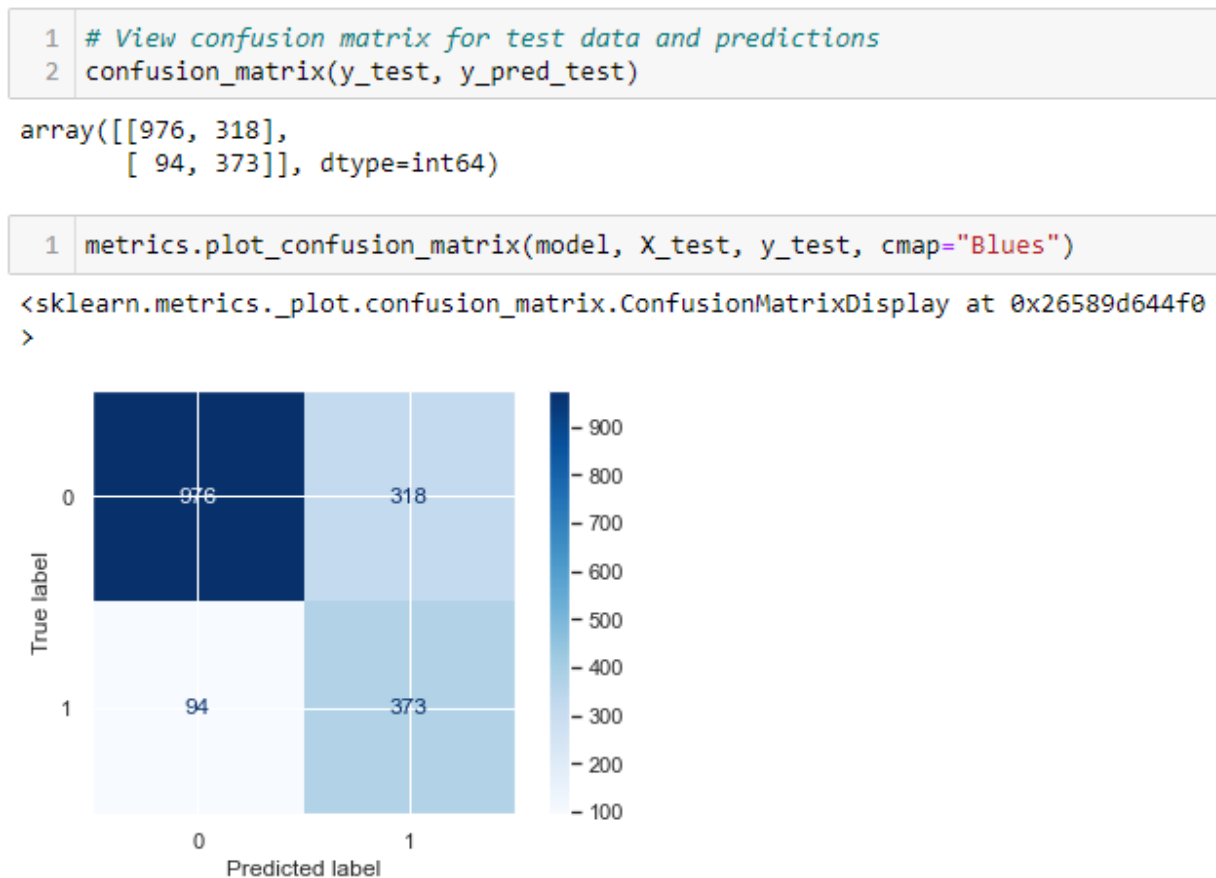
```
1  # View confusion matrix for test data and predictions
2  confusion_matrix(y_test, y_pred_test)
```

```
array([[1168,  126],
       [ 206,  261]], dtype=int64)
```

```
1  metrics.plot_confusion_matrix(model, X_test, y_test, cmap="Blues")
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x2659463d850
>
```
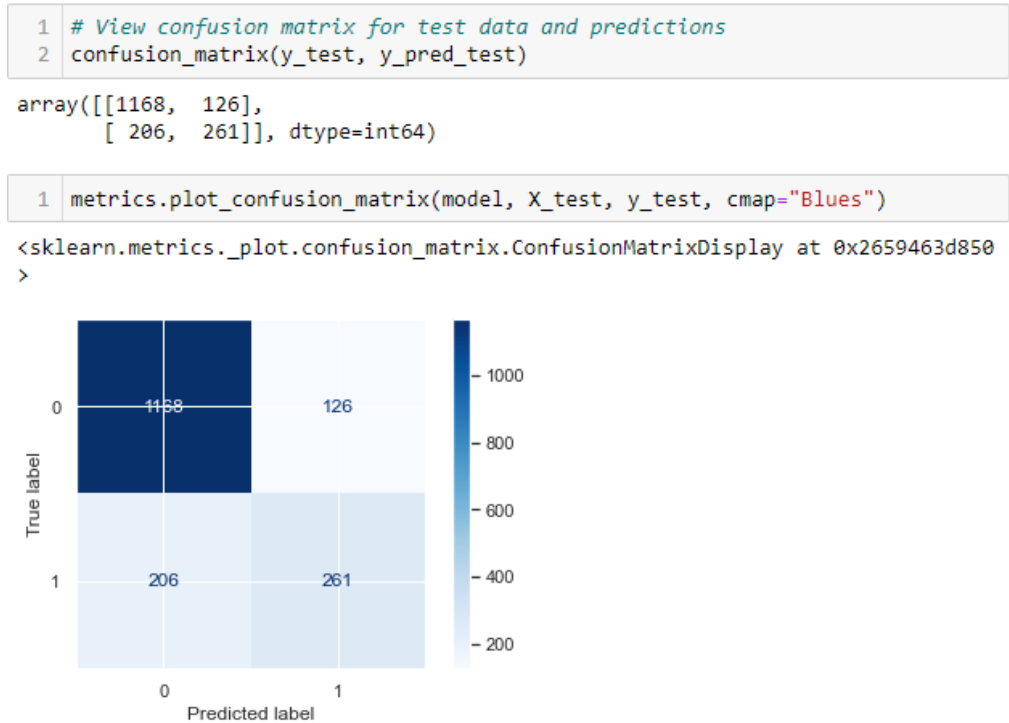

Figure 27: Confusion matrix for Gradient Boosting Classifier

Similarly, Classification report for Gradient Boosting classifier is shown below

Table 5: Classification Report of Gradient Boosting Classifier

	precision	recall	f1-score	support
0	0.85	0.90	0.88	1294
1	0.67	0.56	0.61	467
accuracy			0.81	1761
macro avg	0.76	0.73	0.74	1761
weighted avg	0.80	0.81	0.81	1761

```
Random: ROC AUC=0.500
Gradient Boosting Classifier: ROC AUC=0.857
```


Figure 28: Receiver Operating Characteristics of Gradient Boosting Classifier

Comparing Models

Comparison of ROC-AUC for different classifiers

```
Random: ROC AUC=0.500
Logistic Regression Classifier: ROC AUC=0.855
Naive Bayes Classifier: ROC AUC=0.825
Random Forest Classifier: ROC AUC=0.854
Gradient Boosting Classifier: ROC AUC=0.857
```
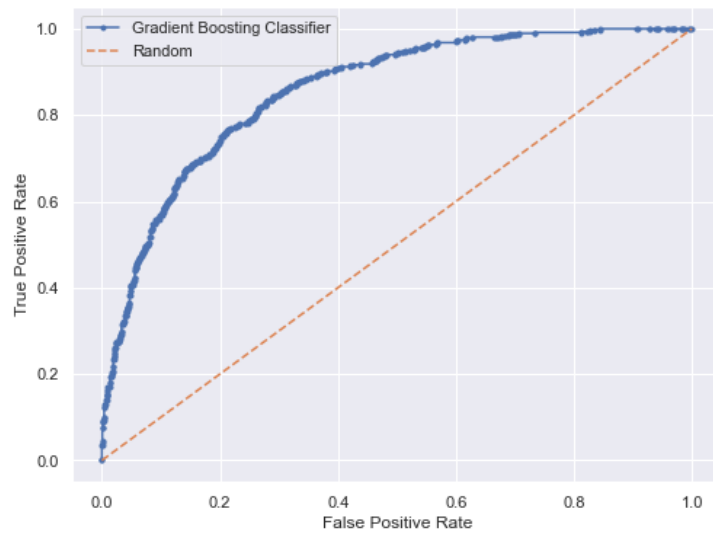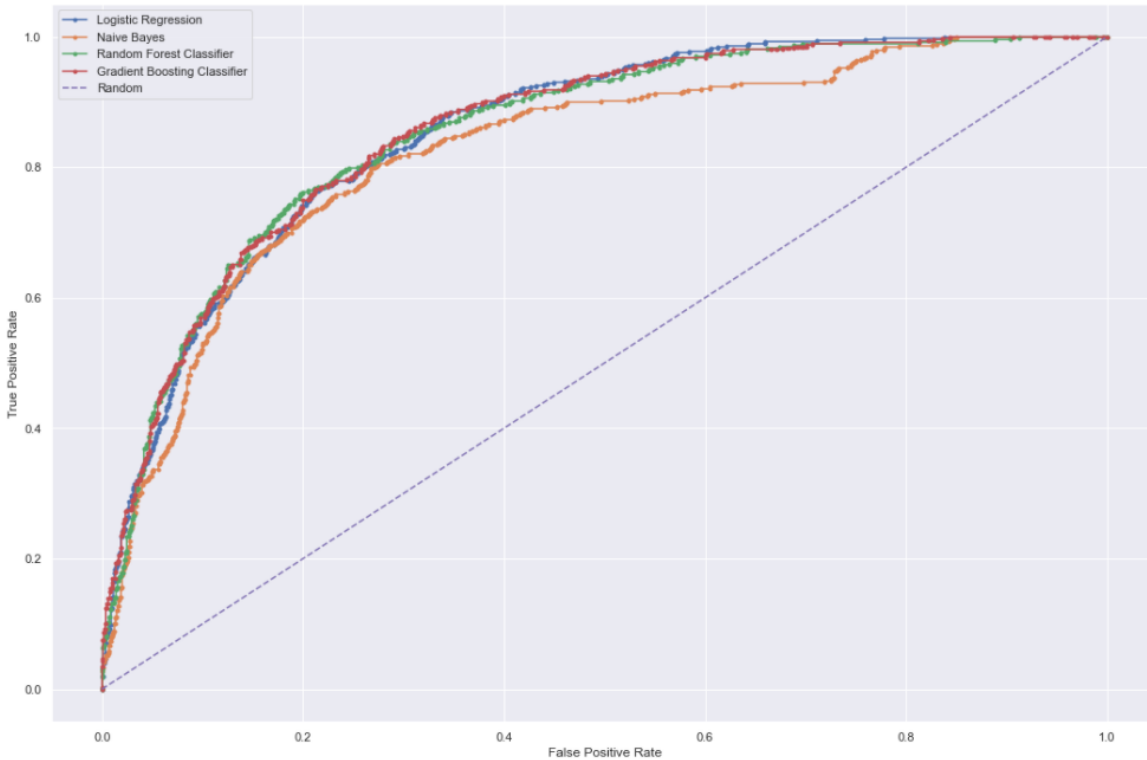

Figure 29: Receiver operating characteristics of different classifiers

The above chart shows receiver operating characteristics for different classifiers, The ROC curve, is a visual chart that describes the diagnostic ability of a binary classifier as its discrimination threshold alters. Similarly, Accuracy, Precision, Recall, F1 Score of different classifiers are shown in below table.

Table 6: Comparison of Matrix for different classifiers

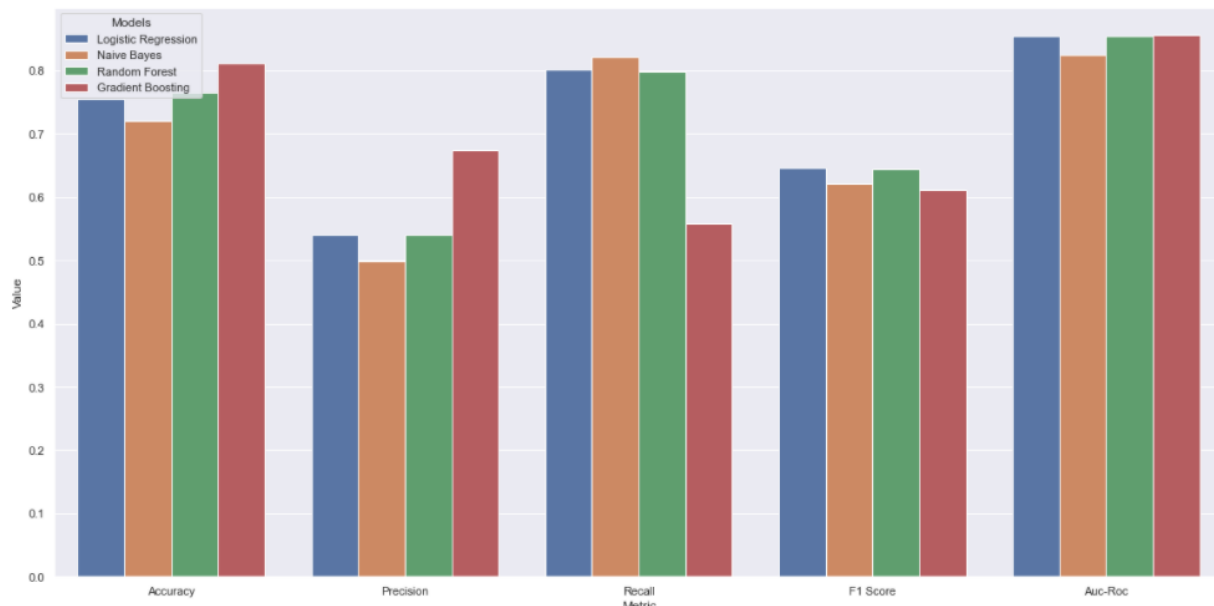|	Models	Accuracy	Precision	Recall	F1 Score	Roc-Auc
1	Logistic Regression	0.754851	0.540664	0.80136	0.64569	0.85487
2	Naive Bayes	0.720303	0.498969	0.82173	0.62091	0.82484
3	Random Forest	0.766042	0.539797	0.79872	0.64421	0.85404
4	Gradient Boosting	0.811471	0.674419	0.55889	0.61124	0.85697

Figure 30: Comparison of Accuracy, Precision, Recall, F1 Score and Roc-Auc

The precision and recall measures are also widely used in classification. Precision is measure of exactness (i.e., what percentage of rows labeled as positive are actually positive), whereas recall is a measure of completeness (what percentage of positive rows are labeled as positive). recall is same as sensitivity (or the true positive rate). These measures can be computed as

precision = TP / (TP + FP) recall = TP / (TP + FN) = TP / P.

An ideal figure of 1.0 for a class means that every tuple that the classifier labeled as belonging to that class does indeed belong to the same class. An ideal recall figure of 1.0 for a class means that every item from the class was labeled correctly, but it does not inform us about other tuples that were incorrectly labeled as belonging to the same class. There is an inverse relationship between precision and the recall. From the above data we can see that even though the accuracy and ROC-AUC is higher for ensemble methods like Random Forest and Gradient Boosting Algorithm but the recall is highest in Logistic Regression and Naive Bayes Recall = TP/ (TP + FN) = TP/P, An ideal recall figure of 1.0 for a class means that every item from that class was labelled correctly, but it does not tell us how many other tuples were incorrectly labelled as belonging to that class i.e., it does not tell us about false positive.

Here we can tolerate false positives but not false negatives as it will lose the revenue of the organization. Here we have taken an arbitrary cutoff recall score and ROC-AUC of 0.8 to select our models.

Therefore, from above data we can conclude that in our dataset the Logistic Regression and Naive Bayes classifier are providing good recall score and good roc-auc score.

Chapter 3

Scope of work
- Acquisition of data
- Preprocessing of data
- Exploration of data
- Implementation of model to predict churn using python
- Visualizations
- Report creation

Chapter 4

Resources needed for the project
- Telecom customer data
- Visualization libraries
- Windows machine
- Python libraries
- Jupyter Notebook

Conclusion / Recommendations

We can now conclude the research work highlighting that for given dataset Logistic Regression Classifier and Naïve Bayes Classifier provide the overall best scores and also, they are much easier to interpret.

In this study we have created multiple models which are Logistic Regression, Naïve Bayes, Random Forest and Gradient Boosting and evaluated these models on multiple metrices like accuracy, precision, Recall, F1 Score and ROC-AUC (Receiver Operating curve).

We have also performed comparison of these models based on above metrices, here as the distribution of the target variable is imbalanced, the accuracy only tells us how good the model is on predicting customers which are not churning and therefore to shortlist we have instead used Recall and ROC-AUC.

In this research work we have created prediction models for customer churn in telecom sector, we can integrate these models with live customers data and give details above which customers are likely to churn to telecom vendors, with this information telecom vendors can use retention strategies.

The live details of the customer churn can be shown in the business dashboard with proper visualizations.

Directions for future work

In future we can improve the models with more data points and attributes also we use. We can design UI to visualize the churn data effectively.

Also, we can take feedback from telecom vendors and provide more details and analysis.

Appendices

Appendices not required.

Bibliography / References

1. Dataset and Metadata

https://community.ibm.com/accelerators/catalog/content/Telco-customer-churn

https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

2. Model References

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://scikit-learn.org/stable/modules/naive_bayes.html

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

https://scikit-learn.org/stable/modules/model_evaluation.html

3. Visualization References

https://matplotlib.org/

https://seaborn.pydata.org/

4. Books

https://www.google.co.in/books/edition/Data_Mining_Concepts_and_Techniques/pQws07tdpjoC?hl=en&gbpv=0

List of Publications/Conference Presentations

The features listed above were presented to the

- Team members during weekly demo

- Other teams willing to implement the features

Duly Completed Checklist

a)	Is the Cover page in proper format?	**Y** / N
b)	Is the Title page in proper format?	**Y** / N
c)	Is the Certificate from the Supervisor in proper format? Has it been signed?	**Y** / N
d)	Is Abstract included in the Report? Is it properly written?	**Y** / N
e)	Does the Table of Contents page include chapter page numbers?	**Y** / N
f)	Does the Report contain a summary of the literature survey?	**Y** / N

i. Are the Pages numbered properly? **Y** / N
ii. Are the Figures numbered properly? **Y** / N
iii. Are the Tables numbered properly? **Y** / N
iv. Are the Captions for the Figures and Tables proper? **Y** / N
v. Are the Appendices numbered? **Y** / N

g)	Does the Report have Conclusion / Recommendations of the work?	**Y** / N
h)	Are References/Bibliography given in the Report?	**Y** / N
i)	Have the References been cited in the Report?	**Y** / N
j)	Is the citation of References / Bibliography in proper format?	**Y** / N