Telecom Customer Churn

Ravindra Neralla

DSC680- Spring 2021

Bellevue University

Milestone-3

Professor: Fadi Alsaleem

Abstract

Customer churn is a major problem and one of the most important concern for every company. Retaining an existing customer is many times more effective than gaining new customers. [2] Companies usually have a greater focus on customer acquisition and keep retention as a secondary priority. However, it can cost five times more to attract a new customer than it does to retain an existing one. Increasing customer retention rates by 5% can increase profits by 25% to 95%, according to research done by Bain & Company. Generally, people only switch to a different company only when the service is not to the level of expectation in one or the other factor when compared to competitors, more than getting attracted to the specials and offers which are offered by the other companies. [1] It is pretty common to hear people express frustration with some aspect of their communications provider — be it convoluted billing, unwanted marketing emails, hard-to-navigate customer service or high plan prices. As the number of peoples who use a phone or a telecom product does not increase in huge volume, any company looking to improve the profits is trying to attract other company customers. At the same time retaining existing customers is very important. As retaining a current customer is ten times more productive than gaining a new one.

Objective:

As customer churn directly effects the revenues of the companies, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn and retain the customers.

Data:

The dataset, I am planning to work on in this project is [3] Telco Customer Churn. Dataset consists of 21 variables with 7043 observations. Each Observation in the dataset represents a customer. Below is the detailed list of variables.

1. customerID
2. gender
3. SeniorCitizen
4. Partner
5. Dependents
6. tenure
7. PhoneService
8. MultipleLines
9. InternetService
10. OnlineSecurity
11. OnlineBackup
12. DeviceProtection
13. TechSupport
14. StreamingTV
15. StreamingMovies
16. Contract
17. PaperlessBilling
18. PaymentMethod
19. MonthlyCharges
20. TotalCharges
21. Churn

Method

Data Analysis:

Customer churn data is loaded into a data frame using python and performed initial analysis to check if any duplicate observations or null values present in the dataset. Generate multiple charts on the variables to see the spread of the values in the variables. Perform Univariant analysis by using histograms and box plots. Perform Bivariant analysis by using correlation matrix and heat maps.

a) Load data into data frame:

Out[2]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	Tec
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 × 21 columns

b) Verify the number of variables and observations:

In [3]:
```
# Retrive the number of rows and columns in data frame
df.shape
```

Out[3]: (7043, 21)

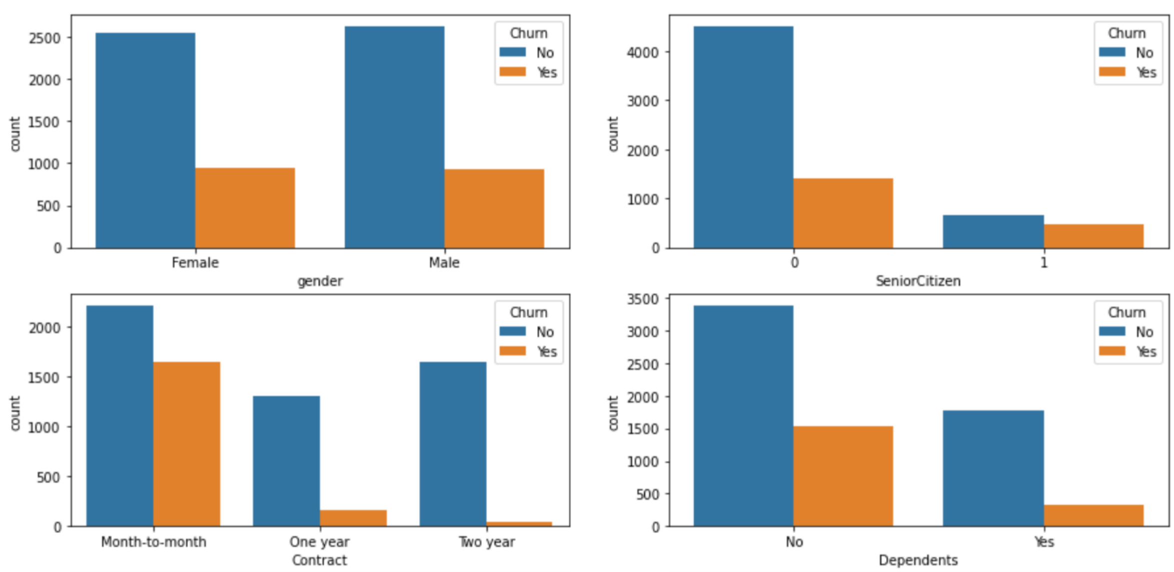
c) Get the stats on numerical variables:

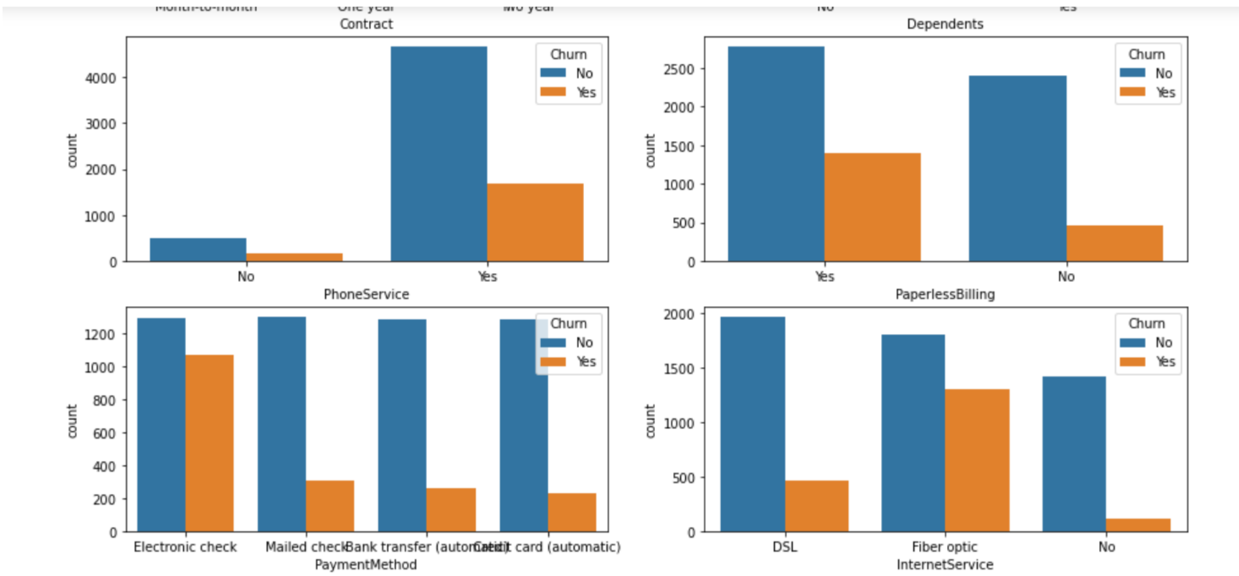
```
Describe Telecom Customer Data
       SeniorCitizen       tenure  MonthlyCharges
count    7043.000000  7043.000000     7043.000000
mean        0.162147    32.371149       64.761692
std         0.368612    24.559481       30.090047
min         0.000000     0.000000       18.250000
25%         0.000000     9.000000       35.500000
50%         0.000000    29.000000       70.350000
75%         0.000000    55.000000       89.850000
max         1.000000    72.000000      118.750000
```

d) <u>Get the stats on categorical variables:</u>

```
Summarized Data
       customerID gender Partner Dependents PhoneService MultipleLines  \
count        7043   7043    7043       7043         7043          7043
unique       7043      2       2          2            2             3
top    0230-UBYPQ   Male      No         No          Yes            No
freq            1   3555    3641       4933         6361          3390

      InternetService OnlineSecurity OnlineBackup DeviceProtection  \
count            7043           7043         7043             7043
unique              3              3            3                3
top       Fiber optic             No           No               No
freq             3096           3498         3088             3095

      TechSupport StreamingTV StreamingMovies       Contract  \
count        7043        7043            7043           7043
unique          3           3               3              3
top            No          No              No Month-to-month
freq         3473        2810            2785           3875

      PaperlessBilling    PaymentMethod TotalCharges Churn
count             7043             7043         7043  7043
unique               2                4         6531     2
top                Yes Electronic check                 No
freq              4171             2365           11  5174
```

<u>g) Generate Charts on Churn vs Categorical Variables:</u>





Observations:

Gender vs Customer_Churn:
We do not see any difference in Male vs Female customers in terms of Customer Churn.

Contract_Type vs Customer_Churn :
'Month-on-month' type Contract has highest Customer Churn compared to other contract Types.

Payment_Method vs Customer_Churn :
'Electronic Check' payment method has the highest Customer Churn.

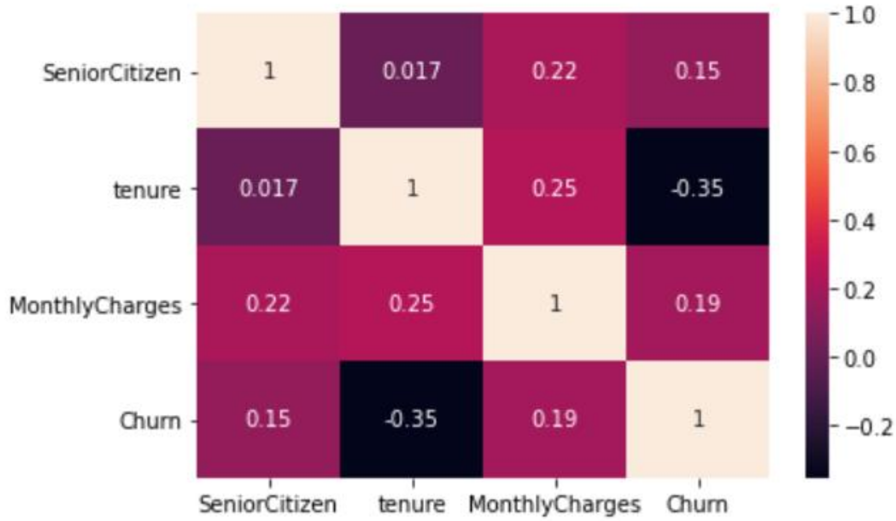
Paperless_Billing vs Customer_Churn : 'Paperless Billing' has highest Customer Churn.

Type_Of_Internet_Service vs Customer_Churn : 'Fiber optic' Internet service has highest Customer Churn.

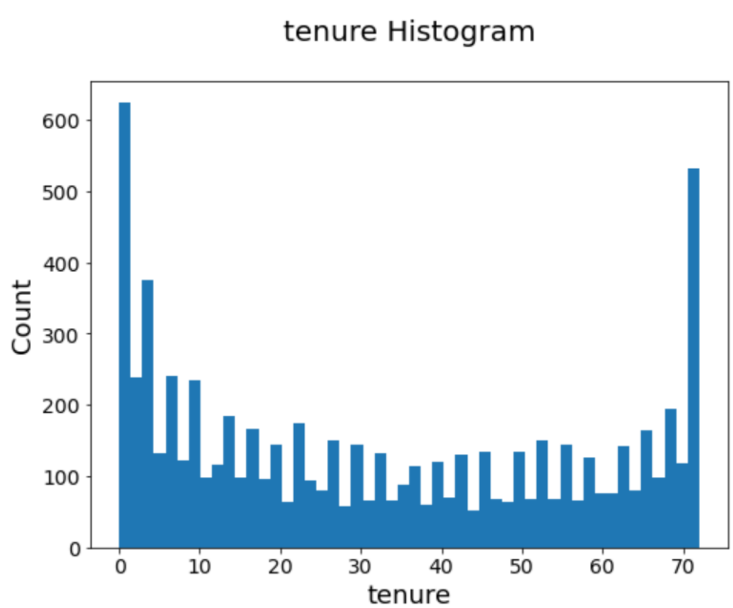
Phone_Service vs Customer_Churn : Customer who has Phone Service has highest Customer Churn.

h) Heatmap Analysis:
From the below heatmap we can observe that tenure and monthly variables are better correlated with churn. So, a histogram is built to understand the spread of tenure data.

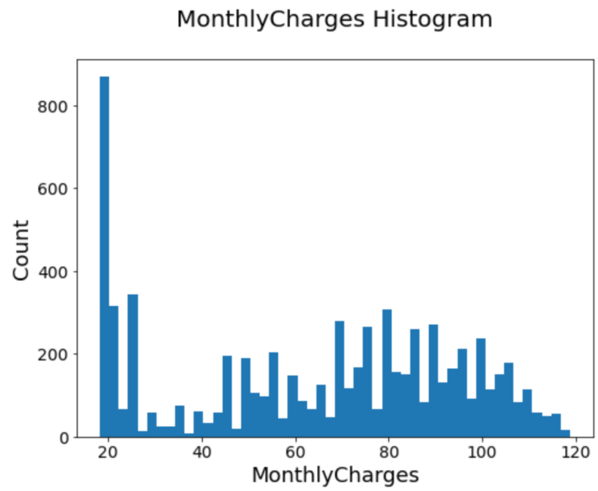
Out[12]: <AxesSubplot:>



<u>Tenure Histogram:</u>



Monthly Charges Histogram:



i) Encoding Method:

Applied Encoding method to convert Categorical variables to numerical values.

Applied Scalar method to convert the Numerical variable to reduce the value and maintain consistent value range, i.e., Tenure in months value ranging from 10 to 70 by applying encoding, values are in consistent range.

In [16]:
```
# Transform Scaled data
df_numerical=scaler.transform(df_numerical)
```

In [17]:
```
# Convert the data to DataFrame
df_numerical = pd.DataFrame(df_numerical)
df_numerical.columns = ['tenure']
df_numerical.head()
```

Out[17]:

	tenure
0	0.013889
1	0.472222
2	0.027778
3	0.625000
4	0.027778

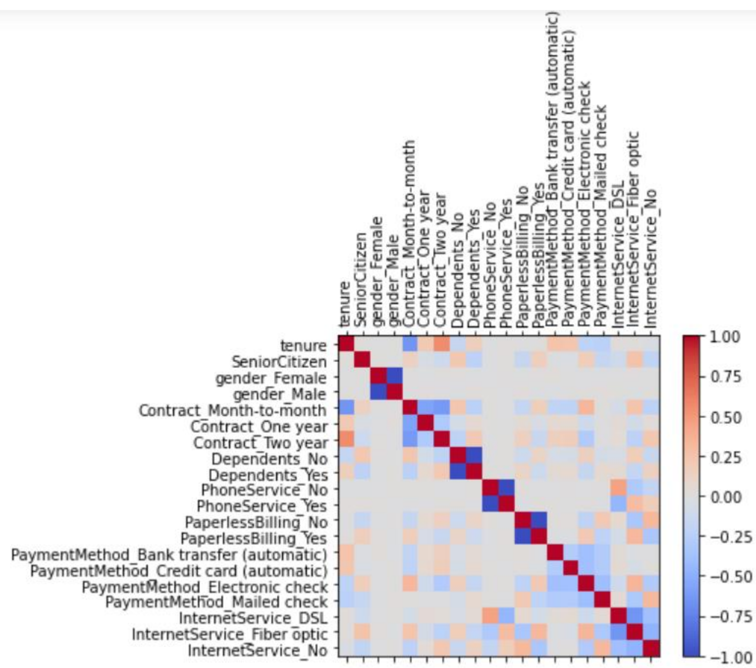
```
n [19]:  # convert the Categorical data to Numerical data
         # One Hot Encoding
         df_categorical = pd.get_dummies(df_categorical)
         # check the data
         df_categorical.head()
```

Out[19]:

	SeniorCitizen	gender_Female	gender_Male	Contract_Month-to-month	Contract_One year	Contract_Two year	Dependents_No	Dependents_Yes	PhoneService_No	PhoneS
0	0	1	0	1	0	0	1	0	1	
1	0	0	1	0	1	0	1	0	0	
2	0	0	1	1	0	0	1	0	0	
3	0	0	1	0	1	0	1	0	1	
4	0	1	0	1	0	0	1	0	0	

j) Dimensionality Reduction:

After converting categorical variables to numeric values, the number of input variables have increased. So, dimensionality reduction was performed by applying Principal Component Analysis [4], reducing the number of variables to 5.



```
In [24]: # Transform the data after applying PCA
df_input_PCA = pca.transform(df_input)
```

```
In [25]: print('Number of elements in the data frame after applying PCA ')
df_input_PCA.shape
```

```
Number of elements in the data frame after applying PCA
```

```
Out[25]: (7043, 5)
```

```
In [26]: # Display the input data which is converted to 5 components using PCA
df_input_PCA = pd.DataFrame(df_input_PCA)
df_input_PCA.columns = ['PCA_Component_1','PCA_Component_2','PCA_Component_3','PCA_Component_4','PCA_Component_5']
df_input_PCA.head()
```

Out[26]:

	PCA_Component_1	PCA_Component_2	PCA_Component_3	PCA_Component_4	PCA_Component_5
0	-0.742041	0.624381	-0.588428	1.473460	0.243152
1	0.942647	-0.733343	-0.675357	0.348005	-0.572353
2	-0.310807	-0.762840	-0.528639	0.858478	0.117877
3	0.963369	-0.689557	-0.338968	0.958506	-0.883852
4	-1.346429	0.659860	-0.188579	-0.207400	0.223515

k) Split the data:

```
# split data into training and validation and check the details of the datasets
X_train, X_test, y_train, y_test = train_test_split(df_input_PCA, df_churn, test_size =0.3, random_state=11)

# number of samples in each set
print("No. of samples in training set: ", X_train.shape[0])
print("No. of samples in validation set:", X_test.shape[0])
```

```
No. of samples in training set:  4930
No. of samples in validation set: 2113
```

Modeling:

A baseline model was established by training a **Logistic Regression Model**, after splitting the data into test and train in the ratio of 70:30

```
Testing Set Confusion Matrix 'LogisticRegression':
 [[1362  179]
 [ 319  253]]

Testing Set Classification_Report 'LogisticRegression':
              precision    recall  f1-score   support

   Churn_YES       0.81      0.88      0.85      1541
    Churn_NO       0.59      0.44      0.50       572

    accuracy                           0.76      2113
   macro avg       0.70      0.66      0.67      2113
weighted avg       0.75      0.76      0.75      2113

Test Set Accuracy LR:  0.7643161381921438
Test Set Sensitivity LR:  0.4423076923076923
Test Set Specificity LR:  0.8838416612589228
```

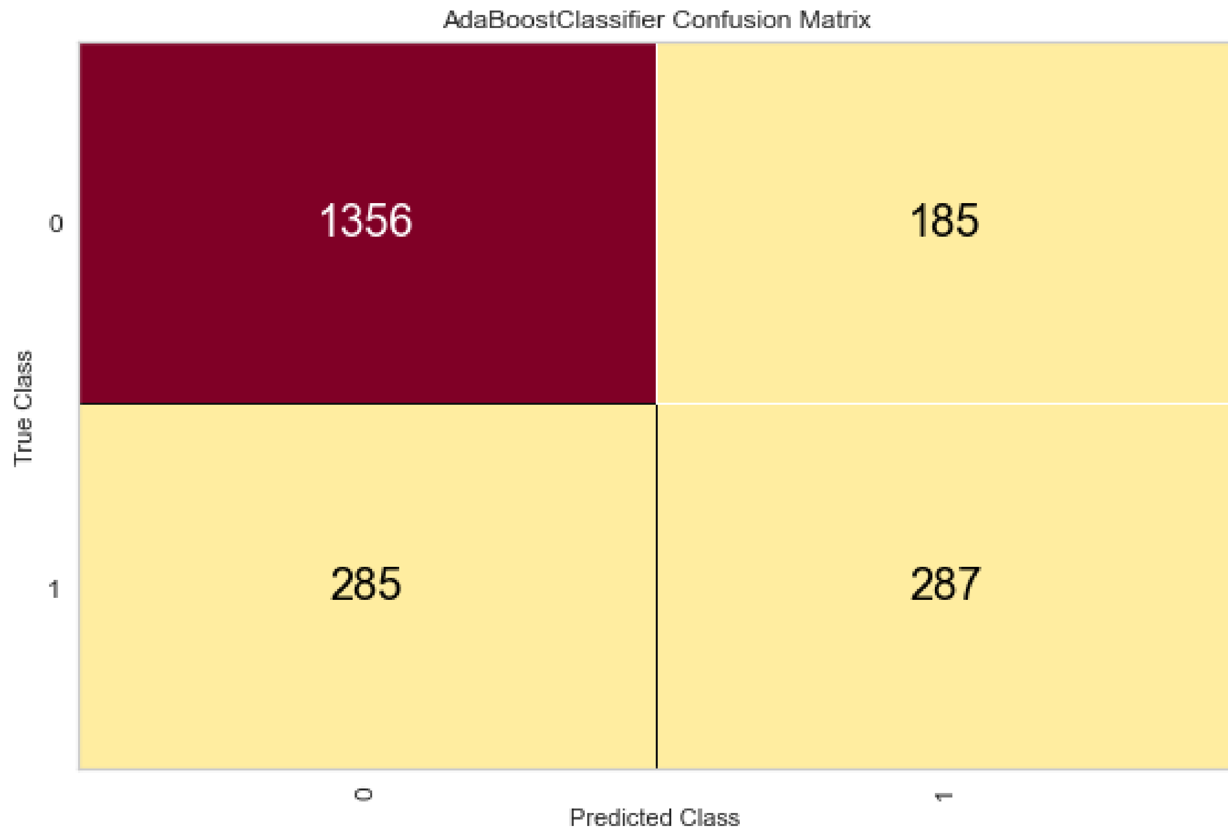
Random Forest Classifier:

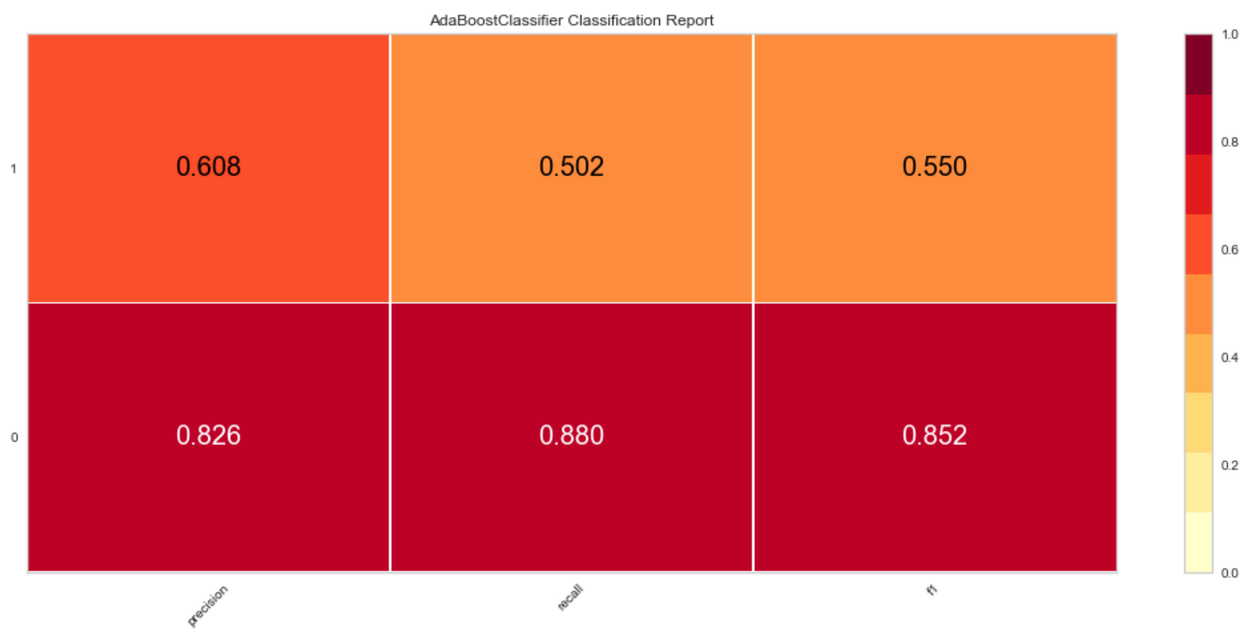
```
Testing Set Confusion Matrix 'RandomForestClassifier':
 [[1378  163]
 [ 299  273]]

Testing Set Classification_Report 'RandomForestClassifier':
              precision    recall  f1-score   support

   Churn_YES       0.82      0.89      0.86      1541
    Churn_NO       0.63      0.48      0.54       572

    accuracy                           0.78      2113
   macro avg       0.72      0.69      0.70      2113
weighted avg       0.77      0.78      0.77      2113

Test Set Accuracy LR:  0.7813535257927118
Test Set Sensitivity LR:  0.4772727272727273
Test Set Specificity LR:  0.8942245295262816
```

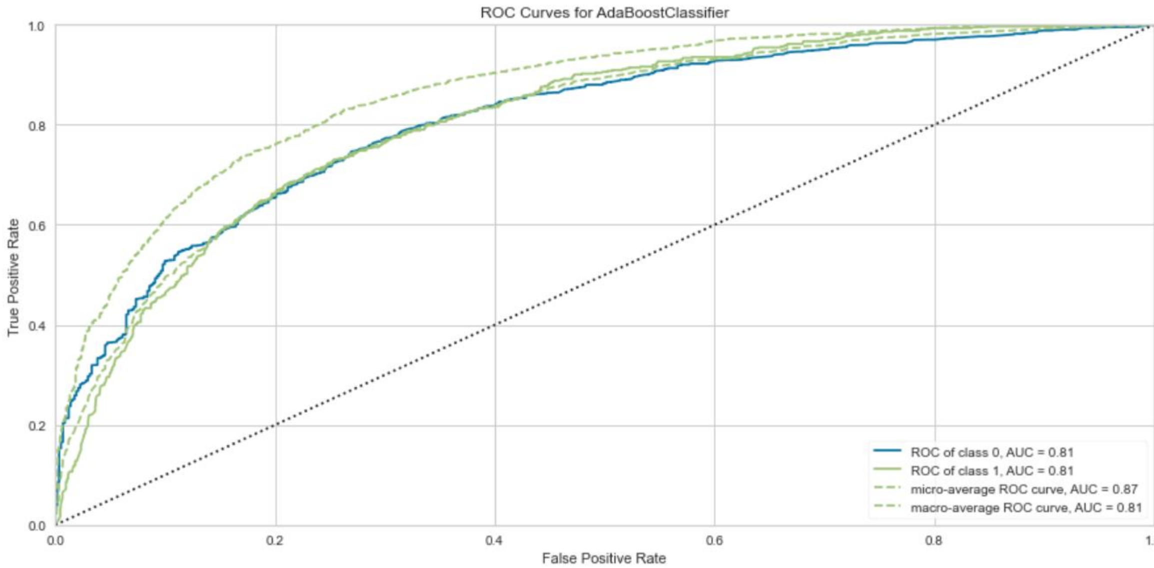
AdaBoost:

```
Testing Set Confusion Matrix 'AdaBoostClassifier':
 [[1356  185]
 [ 285  287]]

Testing Set Classification_Report 'AdaBoostClassifier':
              precision    recall  f1-score   support

   Churn_YES       0.83      0.88      0.85      1541
    Churn_NO       0.61      0.50      0.55       572

    accuracy                           0.78      2113
   macro avg       0.72      0.69      0.70      2113
weighted avg       0.77      0.78      0.77      2113

Test Set Accuracy AB:  0.7775674396592522
Test Set Sensitivity AB:  0.5017482517482518
Test Set Specificity AB:  0.8799480856586632
```

Results:


AdaBoostClassifier Confusion Matrix
True Class
0
1
1356
185
285
287
0
1
Predicted Class


AdaBoostClassifier Classification Report
1
0
0.608
0.502
0.550
0.826
0.880
0.852
precision
recall
f1
1.0
0.8
0.6
0.4
0.2
0.0



Conclusion:

Logistic Regression Model gave accuracy of 76% and Random Forest and ADA Boost gave 78% accuracy. Area Under the Curve (AUC) is 0.81, which indicates that the model is reliable and is accurately working in 81% of cases.

References:

[1]: [How Costly Is Customer Churn in the Telecom Industry? - EBR]

[2]: Customer Churn in Telecom Segment - Curi

https://towardsdatascience.com/customer-churn-in-telecom-segment-5e49356f39e5

[3]: Telco Customer Churn - BlastChar

https://www.kaggle.com/blastchar/telco-customer-churn

[4]: PCA using Python (scikit-learn) - Galarnyk

https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

[5]: Churn Reduction in Telecom Industry – Arthur Middleton Hughes

http://www.dbmarketing.com/telecom/churnreduction.html