Group Assignment – Exploratory Data Analysis

ALY 6040 Data Mining and Application

Northeastern University

College of Professional Studies, Northeastern University, Boston, MA 02115

Data Mining Project 3/3/2023

Submitted to Professor: Justin Grosz

SARVESH YOGESH THORVE

Introduction

"Tele-communications" company as a company is looking to understand the customer Churn. Churning is basically when customers or the subscribers are opting out from the services which the business provides. (*Churn Rate*, 2021). Now, the average annual churn rate in tele-communications business is around 15-20% because of the competition in the market and the fluctuating rates of services provided. So, to stand tall in such competition companies must put their hard work to retain their customers.

So, our main objective is to understand the significance and look for various patterns in the data by looking at the various features which will help us in classification model that will help us to predict which customers are more likely to churn?

Research Question

This analysis has certain steps involved which starts with defining the problem. Now, we are going to construct some hypotheses and will answer those questions in the further analysis. So, we have multiple information available with us we will start by looking at the demographical aspects of the customers then we will look at the operational aspects of the business

---

RESEARCH QUESTIONS: DEMOGRAPHICAL ASPECTS

- Are male more likely to churn than females?
- Are senior citizens more likely to churn than those who are young?
- Are customers with partners more likely to churn than those who are single?

RESEARCH QUESTIONS: OPERATIONAL & FINANCIAL ASPECTS

- Are customers who are paying through electronic check more likely to churn than others?
- Are customers with fiber optic internet service more likely to churn?
- Are customer opted-out for tech-support service more likely to churn

This way we can generate multiple hypotheses and can solve those problems by exploring and visualizing the data. So, the next step is to extract, clean and process the data for exploration and visualization.

Data Preprocessing and Cleaning

The initial step for the analysis is exploring the dataset. For initial analysis we check the structure of the data.

```
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
```

Figure 1: Basic information of the variables.

From the figure 1 we can obtain information about the variables such as their Data type. While exploration we found out that the Variable "TotalCharges" had numerical values, but the datatype was of object, so we converted the datatype of the variable from object to number.

After converting the datatype of variable "TotalCharges", we found out that it had 11 missing values. Since the tenure value of those observation was "0" and the number of values of missing value is lower than **70%** of the observation we decided to drop the missing value.

Below Table is the Descriptive statistics of the numerical variables.

	Senior Citizen	tenure	Monthly Charges	Total Charges
count	7043	7043	7043	7032
mean	0.162147	32.371149	64.761692	2283.300441
std	0.368612	24.559481	30.090047	2266.771362
min	0	0	18.25	18.8
25%	0	9	35.5	401.45
50%	0	29	70.35	1397.475
75%	0	55	89.85	3794.7375
max	1	72	118.75	8684.8

Table: Summary statistics of numerical variables

From the above table we can obtain count, mean, standard deviation, min, max of the numerical variable. For example, we can see that the mean for tenure variable is **32.37** with standard deviation of **24.559481**.

	count	unique	top	freq
customerID	7043	7043	7590-VHVEG	1
gender	7043	2	Male	3555
Partner	7043	2	No	3641
Dependents	7043	2	No	4933
PhoneService	7043	2	Yes	6361
MultipleLines	7043	3	No	3390
InternetService	7043	3	Fiber optic	3096
OnlineSecurity	7043	3	No	3498

OnlineBackup	7043	3	No	3088
DeviceProtection	7043	3	No	3095
TechSupport	7043	3	No	3473
StreamingTV	7043	3	No	2810
StreamingMovies	7043	3	No	2785
Contract	7043	3	Month-to-month	3875
PaperlessBilling	7043	2	Yes	4171
PaymentMethod	7043	4	Electronic check	2365
Churn	7043	2	No	5174

Table: Summary statistics of Categorical variables

From the above table we can obtain the count, unique, top and frequency count of categorical variable. For example, “Gender” variable has count of 7043, with unique 2 values. The value “Male” tops with 3555 frequency count.

We also checked for outlier using boxplot and found that there is no outlier as of now for our dataset.

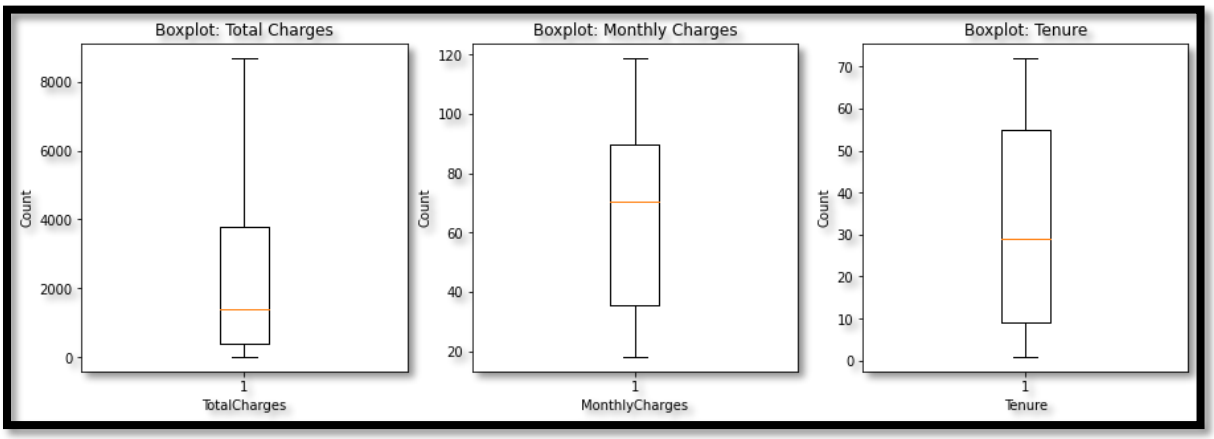


Figure 2: Boxplot of variable with numerical Datatype.

Analysis

Data Visualization:

1) The comparison of the ratio between senior citizens and the number of churns and the number of non-senior citizens and the number of churns can be seen in the graph above. We can observe that the ratio of non-senior citizens to turnover is much higher than the ratio of senior citizens to churn. This might be because people in their senior years don't like to try new things and prefer to stick with whatever plan they have.

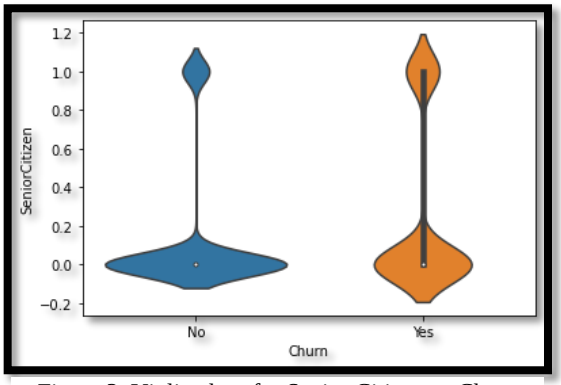


Figure 3: Violin chart for Senior Citizen vs Churn distribution

2) From the below figure we can understand the Churn distribution based on various Services provided by the Telco company. People who have opted for the “Device Protection” services do not churn out of the services. It can also be observed that people who don’t opt for Device Protection services are more likely to Churn out of the services. People who have not opted for the “Tech Support” services are more susceptible churn out of the services. People not subscribing to Online Security Services also opt out of telco service. People who don’t use Online Backup services don’t continue the services with Telco Company.

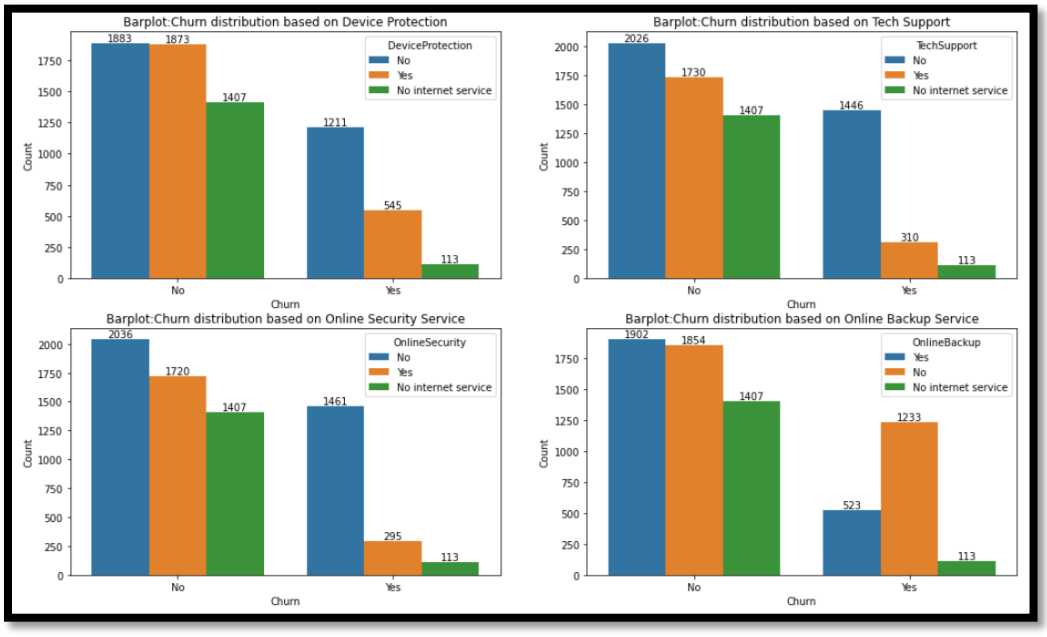


Figure 4

3) The graph below shows the relationship between the variable churn and the type of contract each client is on. Plotting this graph will assist us in identifying the contract types that customers like and attempting to shift them to contract types with lower churn rates. We can

see from the graph below that there is a considerably larger difference in the number of individuals who would churn if they signed long-term contracts like one year or two years. This suggests that those who sign long-term contracts have a lower chance of churning. As a result of the above plot, we can deduce and advocate those additional offers and other forms of advertising are necessary to transfer customers from month-to-month contracts to long-term contracts, resulting in lower churn and consequently more profits for the firm.

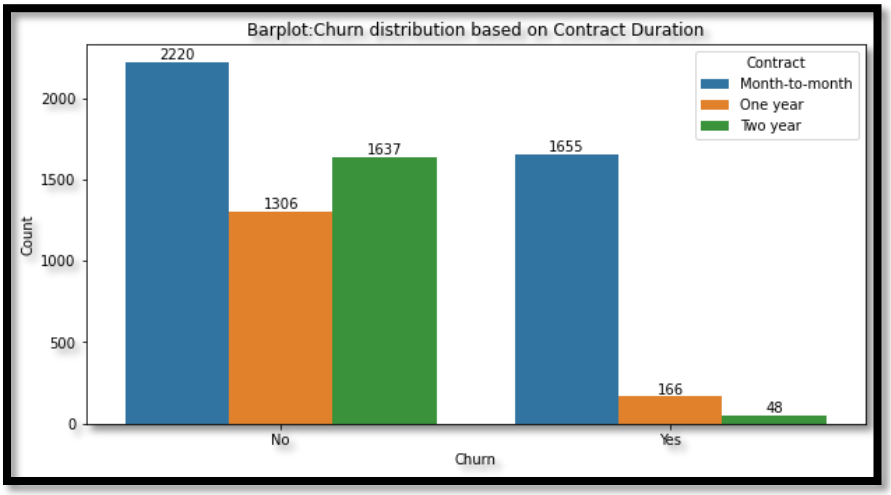


Figure 5

4) The accompanying graphs show the relationship between the Type of Line variable and the average monthly and total charges for consumers to see how important charges are in customer churning. We can observe from the left graph that the churn rate is higher wherever the charges are higher. This means that customers are far less likely to stay with the company and spend more than $80 per month for its services. We can see from the data that consumers who spend more than 79.27 per month are churning, implying that the plans are too expensive for them. People who have used other services are willing to pay a greater price and are less likely to churn. As a result, we should concentrate on signing customers up for long-term contracts and enticing them with other services.

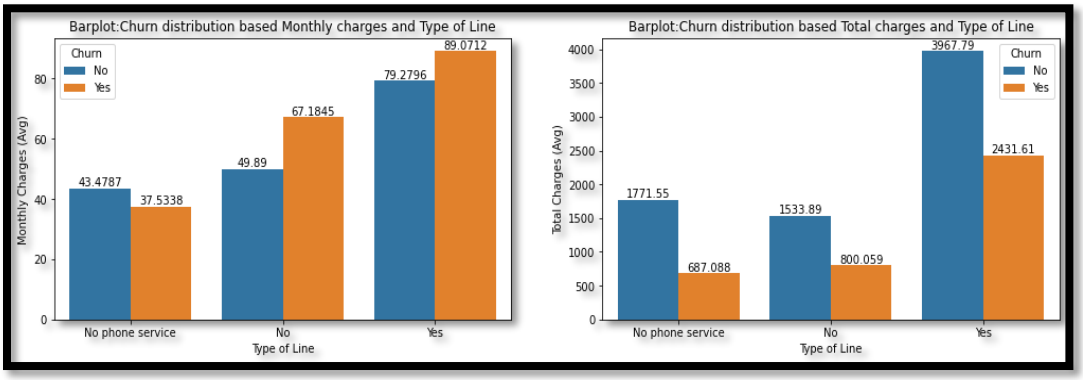


Figure 6

5) Correlation plot

We will primarily focus on the Churn variable as our point of study.
Variables Senior Citizen, Monthly charges, Internet Service Type Fiber optics, Paperless billing subscriber, and Payment method electronic check have a weak positive correlation with the Churn variable which depicts that the people would churn the services.
Variables Tenure, Total Charges, Partner customers, Dependent customers, not an internet subscriber, and No services subscribed have weaker negative correlation which depicts that the people would not churn the services.

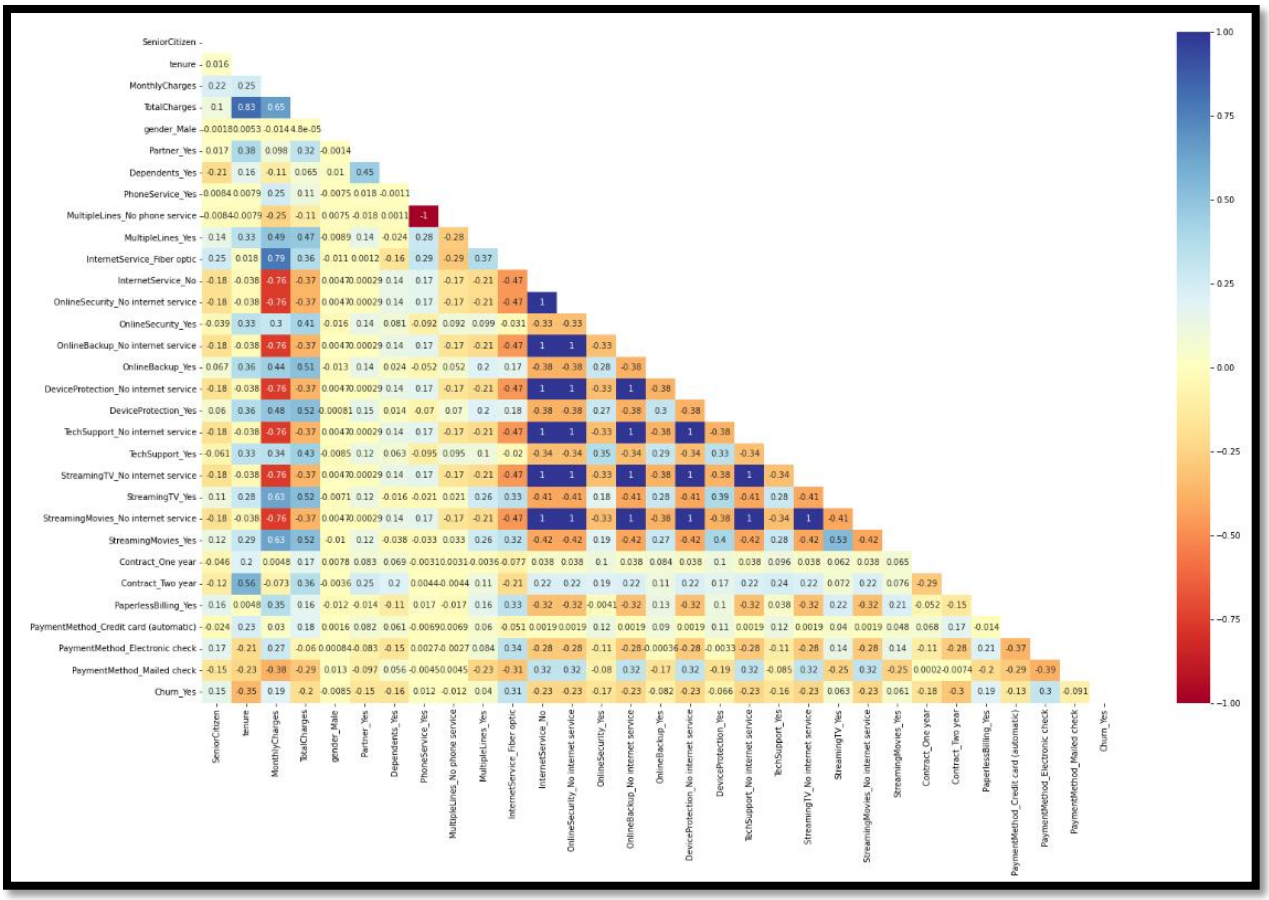


Figure 6

Data Preparation:

We made a ViF test to understand feature multicollinearity in our dataset. As a Result of ViF test we drop multiple columns like tenure','MonthlyCharges','PhoneService_Yes','OnlineSecurity_No internet service' and so on since the ViF score of this variable was more than 10. We also encoded the dataset and divided the dataset into 80:20 as training set and test set.

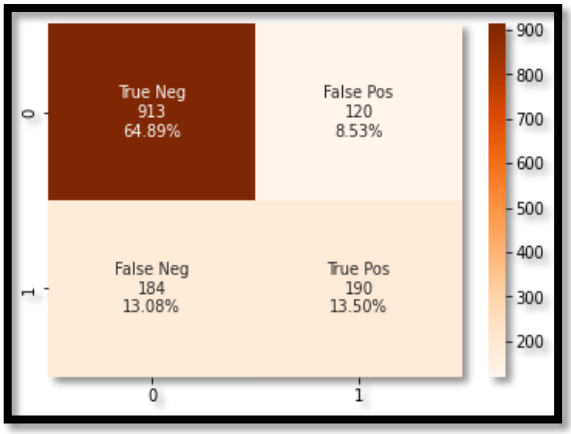
Data Modeling

1) Logistic Regression

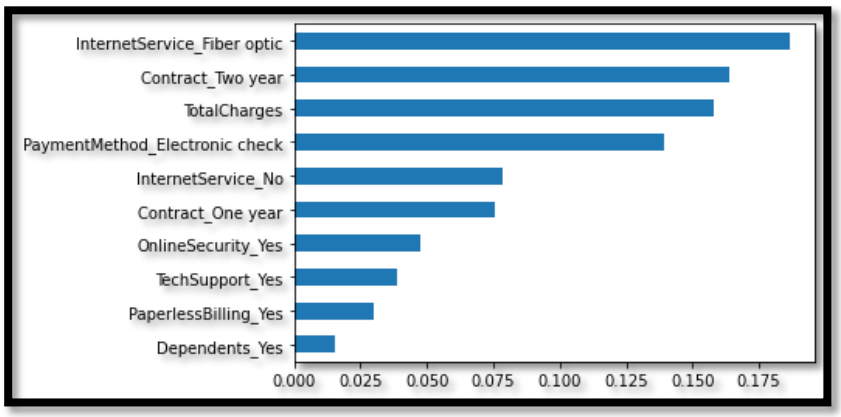
We built logistic model to understand the statistical significance of our features with the target variable. Next, we fit the logistic regression model, to predict whether a person will churn based on the major features in our model. Customers with a two-year contract, no internet service, a one-year contract, and tech support, are less likely to churn. From the model summary we could see that the customers opting for Internet services of Fiber Optics, and other services like Streaming and multiple lines are more likely to Churn.

	a_coef	coef	pval
Contract_Two year	1.702677	-1.702677	7.71E-24
InternetService_No	1.352493	-1.352493	2.27E-28
Contract_One year	0.890382	-0.890382	6.77E-18
InternetService_Fiber optic	0.775636	0.775636	5.21E-20
OnlineSecurity_Yes	0.495386	-0.495386	1.37E-09
PaymentMethod_Credit card (automatic)	0.425585	-0.425585	2.59E-05
TechSupport_Yes	0.40187	-0.40187	1.38E-06
PaymentMethod_Mailed check	0.317358	-0.317358	5.61E-04
StreamingMovies_Yes	0.302081	0.302081	1.31E-04
StreamingTV_Yes	0.302053	0.302053	1.33E-04
Dependents_Yes	0.234248	-0.234248	7.87E-03
MultipleLines_Yes	0.191809	0.191809	1.29E-02
OnlineBackup_Yes	0.190082	-0.190082	1.18E-02
gender_Male	0.151218	-0.151218	1.44E-02
PaperlessBilling_Yes	0.139588	0.139588	4.26E-02
TotalCharges	0.000303	-0.000303	1.45E-28

Overall, the model accuracy obtained is 78.3%. Model predicts that 64.89% of people who would not churn. It could precisely predict 13.51 percent of customers who will stop using the company services.



2) Decision Tree

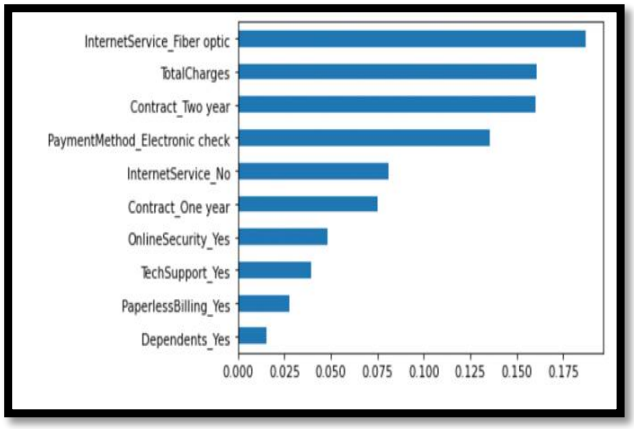


The accuracy of the Decision Tree model we built was **78.3**%. According to Decision Tree "InternetService_ Fiber optic", "Contract_Two year", and "TotalCharges" are the top 3 important features for the model.

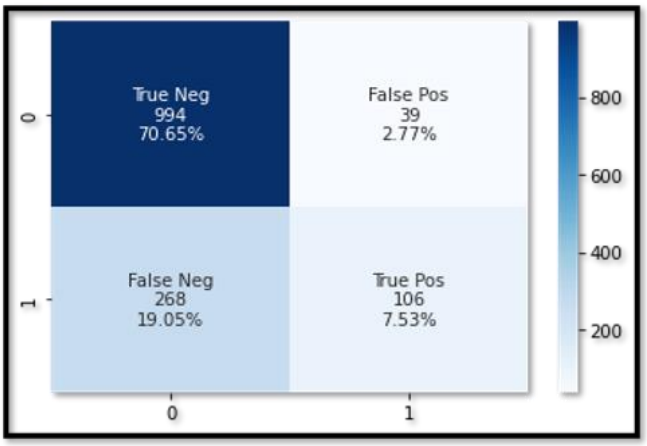
We can see the model's performance from the Confusion matrix below, **68.30**% of observations are of customers who did not churn were predicted correctly by the model, **10.73**% of the observations are of customers who churn were predicted correctly by the model, **5.12**% of the observations are of customers who churn were wrongly predicted by the model and **15.85**% of the observations are of customers who did not churn were wrongly predicted by the model

3) Random Forest

We used the training data to build the Random Forest Classification model and then used the Random Forest Classifier to predict the variables using the test data. It turns out that Internet Service Fiber Optic consumers have the highest occurrences in the model. Total Charges and Users with a 2-year contract are also important factors in classifying the Churn. As a result, we urge that any feature s with a substantial impact on churn rate, such as those listed above, be thoroughly examined. This is because we need to know why persons who have the following variable trait are leaving.



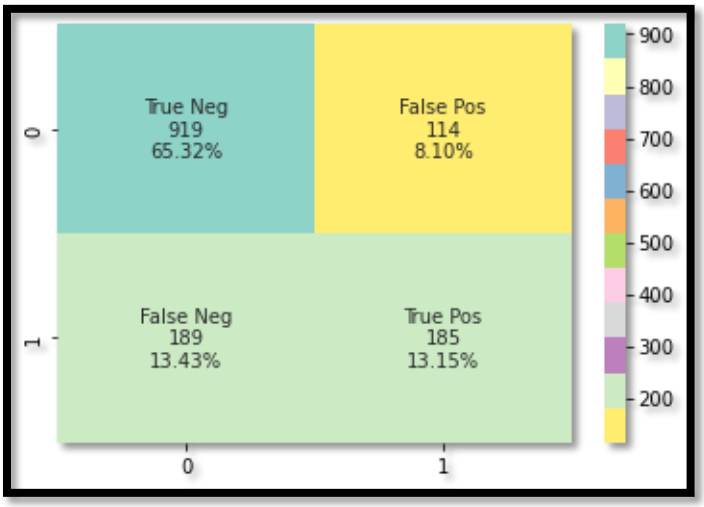
Despite the model's high accuracy of 78.32 percent, we would not advocate using it since it has a high False Negative rate of 19.05%, which implies that a significant portion of the predicted data is incorrectly forecasted, potentially affecting our research.



4) XGBoost:

In this assignment we have also used boosting technique which is basically a technique where we do sequential modelling. Where the algorithm creates different models which are known as Weak learners and together the models are known as Strong learners.

Based on the problem if it is a Regression problem than the mean output of all models will be considered and in the case of Classification problem the majority of binary variable will be considered as the output.



So, the overall accuracy of model is **78.46**% where the model has predicted false negative of around **13.43**% which is less than decision tree classifier and random forest classifier. But we must optimize this square of the confusion matrix because it means that model predicted the customers won't churn but they actually did.

Benchmarking Metrics

Metrics	Logistic Regression	Decision Tree	Random Forest	XGBoost
Accuracy	78.39%	79.03%	78.32%	78.46%
Precision	61.29%	67.71%	73.47%	61.87%
Recall	50.80%	40.37%	28.88%	49.47%
MSE	21.61%	20.97%	20.97%	21.54%
ROC_AUC	69.59%	66.70%	62.55%	69.21%

Decision tree had the highest accuracy, as seen in the table. However, the model with the best combination of high Precision and Recall should be preferred. We can observe from the table that, while Decision Tree and Random Forest have good Precision, they have lower Recall when compared to Logistic Regression. When we carefully examine the Precision and Recall figures, we can see that Logistic regression has excellent accuracy, as well as good Precision and the highest Recall. As a result, we'd continue in the same direction.

Conclusion and Recommendations

The goal of the study is to understand about Telco customer churn. We are assigned as Data Analysts to go over the dataset with many features and investigate the research questions specified in our problem description. Also, with this we must understand the significance of various features with respect to our target variable i.e., Churn. Exploratory Data Analysis and Various models were used to help us reach our goal. During our research, we discovered the following:

From the study we suggest the telco company to use the Logistic regression model to predict the Churn because in our analysis we found, The Prediction for False negative section in all the confusion matrix of all the models built the Logistics Regression and XgBoost model had lowest count of invalid predictions. The company and as an analyst it is our job to minimize the False Negative to increase our model prediction and to increase our customer retention lowering the Churn.

From the logistic model study, we understand that the feature Internet Service Fiber optics users are more likely to churn hence, it is advised to the company that they should work in providing better services with best prices if they want to avoid customer churn.

It is also observed that people who engage in 2 year contract are less likely to Churn hence we should promote this service which will benefit in customer retention.

It is also advised to the Telco Company to use the logistic regression model because it helps us understand the influence and significance of various features on the target variable.

References

1) *Telco Customer Churn*. (2018, February 23). Kaggle. https://www.kaggle.com/datasets/blastchar/telco-customer-churn?datasetId=13996&sortBy=voteCount
2) GeeksforGeeks. (2022, March 8). *Box Plot in Python using Matplotlib*. https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib/
3) Real Python. (2021, June 19). *Python Histogram Plotting: NumPy, Matplotlib, Pandas & Seaborn*. Histogram. https://realpython.com/python-histograms/#visualizing-histograms-with-matplotlib-and-pandas
4) S. (2021, October 16). *Group and Aggregate your Data Better using Pandas Groupby*. Shane Lynn. https://www.shanelynn.ie/summarising-aggregation-and-grouping-data-in-python-pandas/
5) *Churn Rate*. (2021, November 27). Investopedia. https://www.investopedia.com/terms/c/churnrate.asp

6) G. (2020, December 29). *Decision Tree Implementation in Python From Scratch*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/all-about-decision-tree-from-scratch-with-python-implementation/

7) R, S. E. (2021, June 24). *Random Forest | Introduction to Random Forest Algorithm*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/