Capstone Project - 1

Telecom Churn Analysis

Team Members

Akshada

Nikita

Introduction

We'll be working on a telecom churn analysis, where we'll go through the entire EDA process to determine if customers from that particular telecom industry will leave that telecom service or not. In the meantime, we'll use data visualization and analysis to get some insights into the factors that will affect the output, i.e. customer churn.

Introduction(Continued)

Nowadays, the telecom business faces tough competition in satisfying its customers. With the advancement of technology, the services offered by telecom companies have increased. Therefore, there is a constant struggle to strike a perfect balance between the price and services, and to survive in this market, telecom companies must innovate, provide better services, and grow their customer base.

Customer churn, also known as customer attrition, refers to the process of subscribers (either prepaid or postpaid) switching from one service provider to another.

Objective:

The Orange S.A. churn dataset consists of cleaned customer activity data (features). We were provided with a churn label that specifies whether or not a customer has cancelled their subscription, which will help the company to retain their customers quickly and efficiently.

The objective is to analyze the data and extract some insights from it to discover the key factors responsible for customer churn and come up with ways/recommendations to ensure customer retention, before performing further evaluations. We want to know as much information about the data as possible at the beginning of EDA (Exploratory Data Analysis).

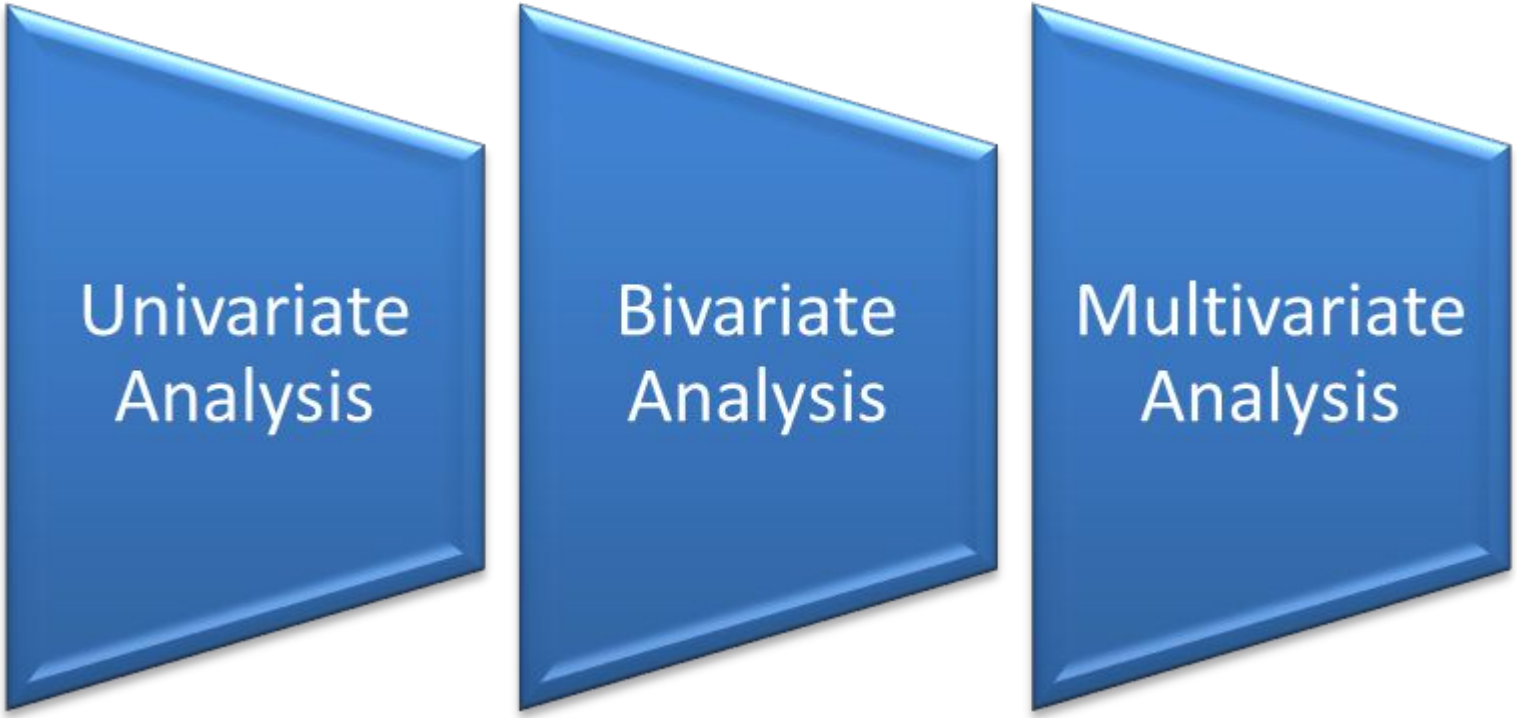
What is exploratory data analysis?

- EDA (Exploratory Data Analysis) is a technique that data professionals can use to understand a dataset before they start to model it. The goal of conducting EDA is to determine the characteristics of the dataset.
- EDA is used to extract information from the data. Data scientists and analysts can take insight from the data using some data visualization techniques such as creating graphs like histograms, scatter plots, and box plots.
- The main goal of EDA is to detect any errors or outliers in the data, as well as to understand different patterns. It allows analysts to have a better understanding before making any assumptions.
- The outcomes of EDA helps telecom businesses to know their customers, expand their business and take decisions accordingly.

Data Visualization

- Data Visualization(or visualisation) is the process of analyzing data in the form of graphs or maps, making it a lot easier to understand the trends or patterns in the data.
- This method translates information into a visual representation, such as a map or graph, in order to make data easier for the human brain to comprehend and extract insights from the data.
- The main purpose of data visualization is to make it easier to recognise patterns, trends, and outliers in huge datasets.
- Data visualisation is one of the processes in the data science process, according to which data must be represented after it has been collected, processed, and modelled in order to draw conclusions.

Different types of analysis



Different types of analysis (Continued)

To explore or analyse the data either graphs or python functions can be used. There are three types of analysis:

1. Univariate analysis: There is only one variable in this type of data. It is not concerned with causes or relationships, and the analysis' primary goal is to explain the data and identify patterns that exist within it.
2. Bivariate analysis: There are two variables involved in this type of data. This type of data analysis is concerned with causes and relationships, and the goal is to determine the relationship between the two variables.
3. Multivariate analysis: More than two variables are involved in multivariate analysis. This type of data analysis is also concerned with causes and relationships, and the goal is to determine the relationship among the variables.

Steps Involved

Step 1:

Importing required libraries using the ‘import’ statement. Libraries imported in this project are numpy, pandas ,seaborn and matplotlib.pyplot.

Step 2:

Loading the dataset to read and store the data using pandas.read_csv.

Step 3:

Exploratory data analysis, the very first step of EDA is to collect all the basic information about the dataset using pandas.DataFrame.info and pandas.DataFrame.describe methods which prints a concise summary of the dataframe and generate descriptive statistics.

Steps Involved(Continued)

Step 4:

Handling missing values(if any we replace the missing value with the mean, median, mode or constant value and another alternative is to remove the entry from the dataset itself). In our dataframe, we don't have any missing values thus we are not performing any operations on it.

Step 5:

After getting the basic information about the data, we further moved deep into each feature of our dataset to figure out its various aspects using data visualization (such as pie chart, bar chart, count plot, box plot, histogram, heat map) and python functions.

Steps Involved(Continued)

Step 6:

We used bar charts, count plots, and box plots to analyze the influence of independent variables on the churn rate.

Step 7:

We used the histogram to check the frequency distribution of each variable in the dataframe.

Step 8:

We created a heat map using the correlation function (pandas.DataFrame.corr) to find out whether the correlation between all the features of the dataframe is positive, negative or perfect.

Variable Breakdown

- STATE: 51 Unique States.
- Account Length: Length of the account.
- Area Code: 415 relates to San Francisco, 408 is of San Jose and 510 is of City of Oakland.
- International Plan: 'Yes' indicates International Plan is present and 'No' indicates no subscription for International Plan.
- Voice Mail Plan: 'Yes' indicates Voice Mail Plan is present and 'No' indicates no subscription for Voice Mail Plan.
- Number vmail messages: Number of Voice Mail Messages ranging from 0 to 51.

Variable Breakdown(Continued)

- **Total day minutes: Total number of Minutes spent by customers in morning.**
- **Total day calls: Total number of Calls made by customer in morning.**
- **Total day charge: Total charge to the customers in morning.**
- **Total eve minutes: Total number of Minutes spent by customers in evening.**
- **Total eve calls: Total number of Calls made by customer in evening.**
- **Total eve charge: Total Charge to the customers in evening.**
- **Total night minutes: Total number of Minutes spent by customers in the night.**

Variable Breakdown(Continued)

- Total night calls: Total number of Calls made by customer in night.
- Total night charge: Total Charge to the customers in night.
- Total intl minutes: Total number of Minutes spent by customers for International Calls.
- Total intl calls: Total number of International Calls.
- Total intl charge: Total Charge to the customers for International Calls.
- Customer service calls: Total number of Customer Service Calls.
- Churn: Target Variable

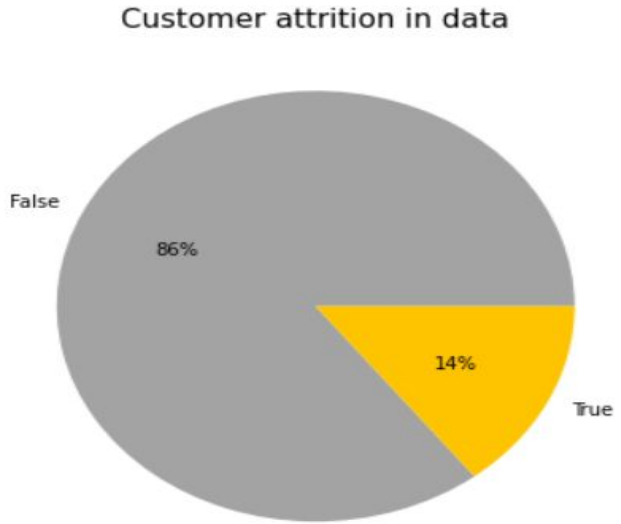
Descriptive statistics

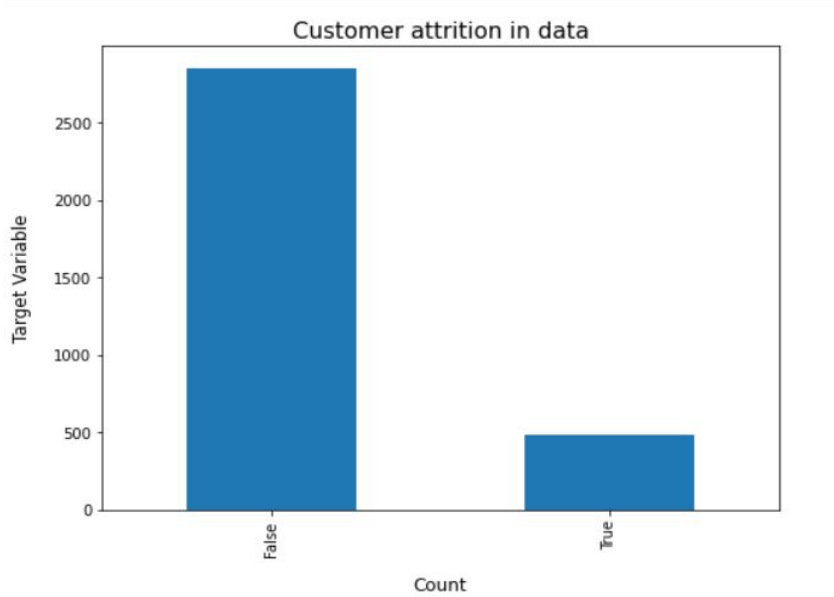
1. Number voice mail messages ranges from 0 to 51.
2. Only 25% of customers are using voice mail messages.
3. Average call minutes in the evening and night are higher than the average call minutes in the day, but the average charge in the day is higher than the average charge in the evening and night.
4. Maximum international call minutes are 20 only.

	Account length	Area code	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	Total night charge	Total intl minutes	Total intl calls	Total intl charge	Customer service calls
count	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000
mean	101.064806	437.182418	8.099010	179.775098	100.435644	30.562307	200.980348	100.114311	17.083540	200.872037	100.107711	9.039325	10.237294	4.479448	2.764581	1.562856
std	39.822106	42.371290	13.688365	54.467389	20.069084	9.259435	50.713844	19.922625	4.310668	50.573847	19.568609	2.275873	2.791840	2.461214	0.753773	1.315491
min	1.000000	408.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	23.200000	33.000000	1.040000	0.000000	0.000000	0.000000	0.000000
25%	74.000000	408.000000	0.000000	143.700000	87.000000	24.430000	166.600000	87.000000	14.160000	167.000000	87.000000	7.520000	8.500000	3.000000	2.300000	1.000000
50%	101.000000	415.000000	0.000000	179.400000	101.000000	30.500000	201.400000	100.000000	17.120000	201.200000	100.000000	9.050000	10.300000	4.000000	2.780000	1.000000
75%	127.000000	510.000000	20.000000	216.400000	114.000000	36.790000	235.300000	114.000000	20.000000	235.300000	113.000000	10.590000	12.100000	6.000000	3.270000	2.000000
max	243.000000	510.000000	51.000000	350.800000	165.000000	59.640000	363.700000	170.000000	30.910000	395.000000	175.000000	17.770000	20.000000	20.000000	5.400000	9.000000

Checking the Customer Attrition

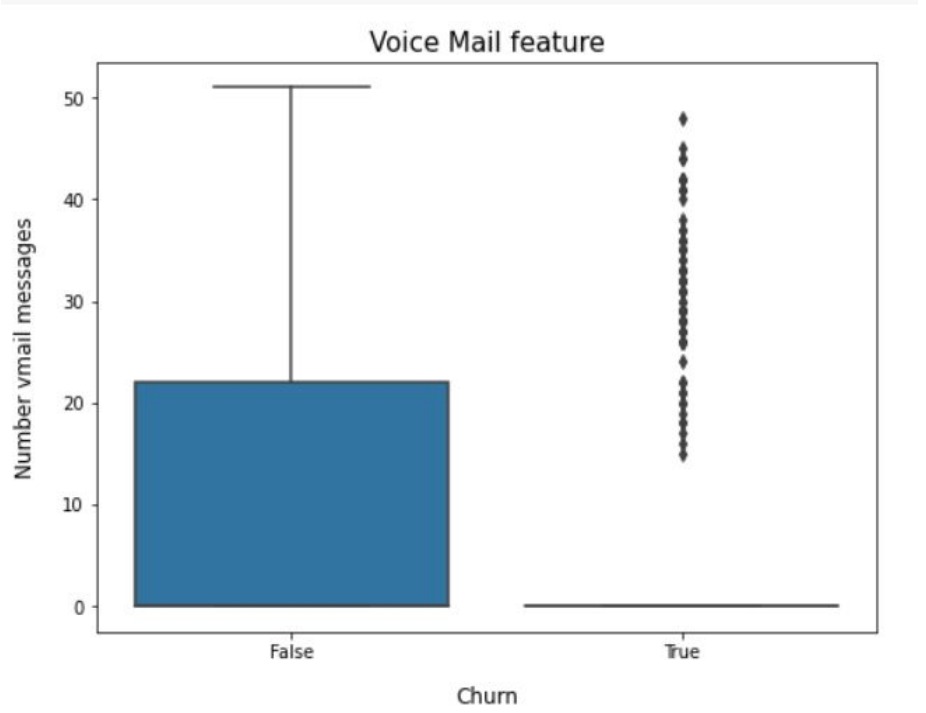
- We can conclude from the pie chart that our dataset is unbalanced, with true values of 14% and false values of 86%.
- As per analysis, the churn rate of the telecom company is 14%.





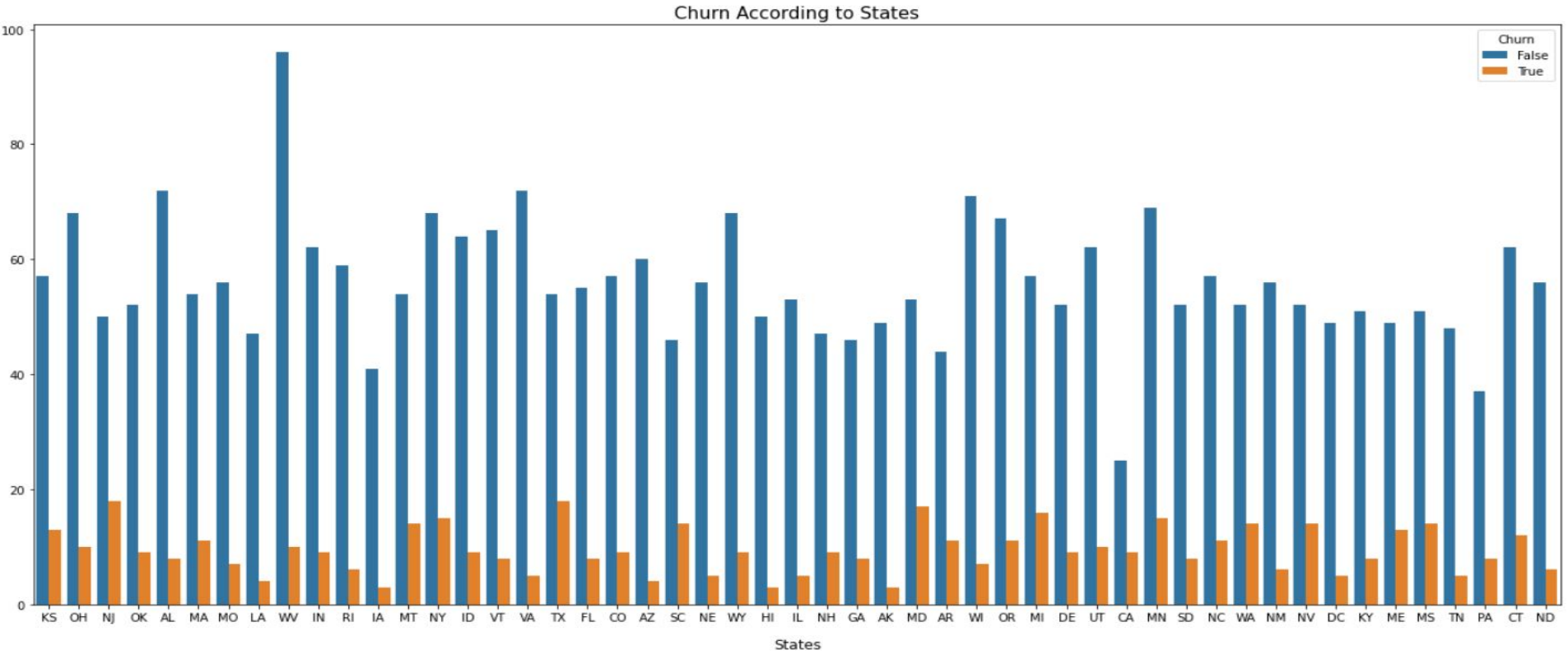
Checking Voice-Mail Feature

- From the boxplot, we can observe that there is a higher churn for customers who are using more than 20 voice-mail messages.
- Quality drop in voice mail after 22 voice mails.
- Certainly, the churn is indicating that we need to improve our voice-mail feature or set a limit and then check whether a customer is retained.



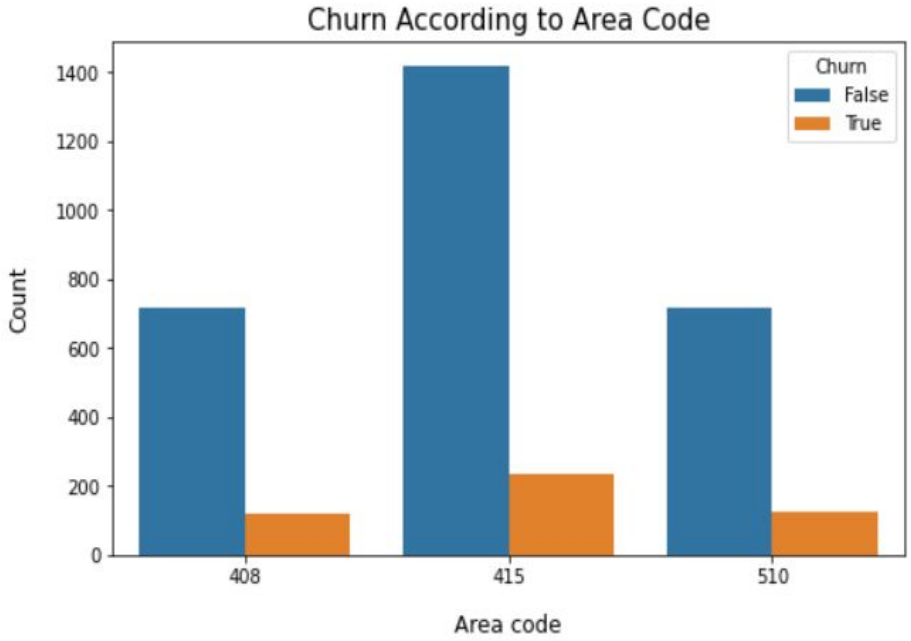
Churn According to States

- **New Jersey, Texas, and Maryland have a bit more churn rate than usual.**



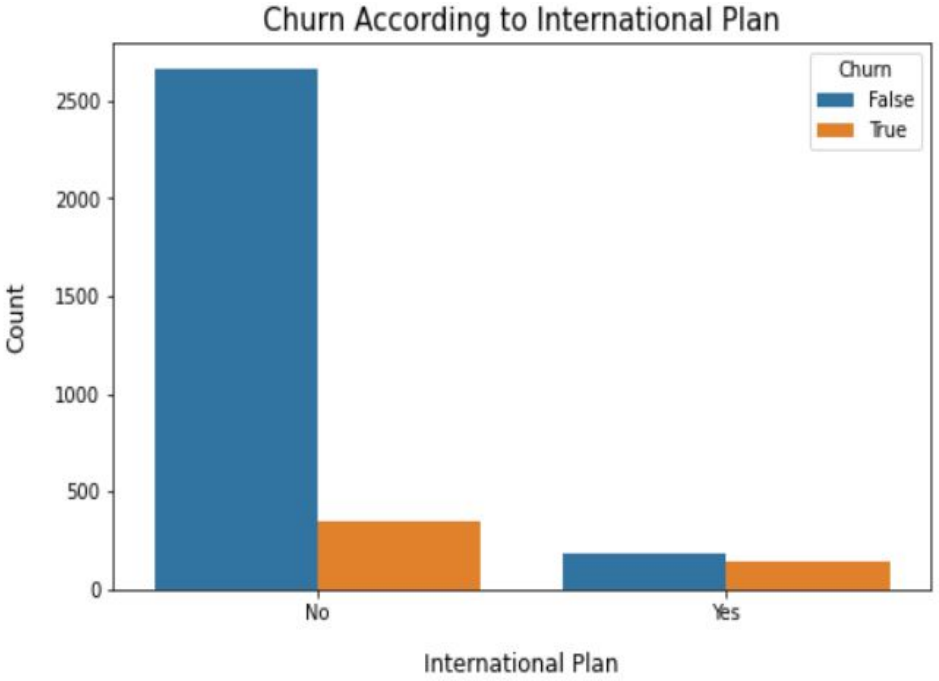
Churn According to Area Code

- Orange S.A. has more customers in San Francisco as compared to San Jose and the City of Oakland.
- Network upgrading is suggested in San Jose and the City of Oakland to increase customers.



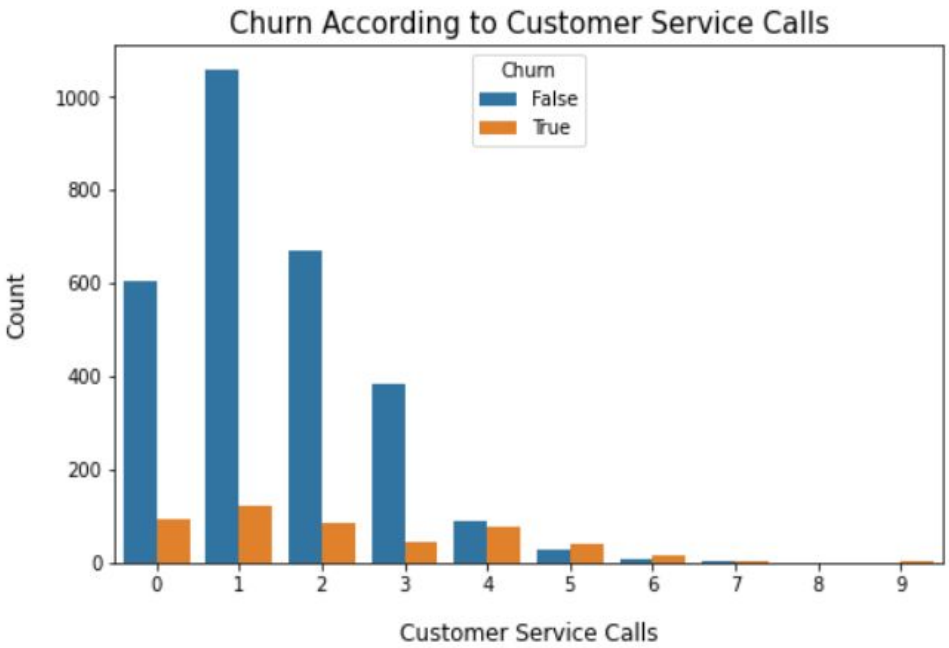
Churn According to International Plan

- Customers with an international plan have a higher churn rate than customers who do not have an international plan.
- To reduce churn, a network upgrade is recommended.

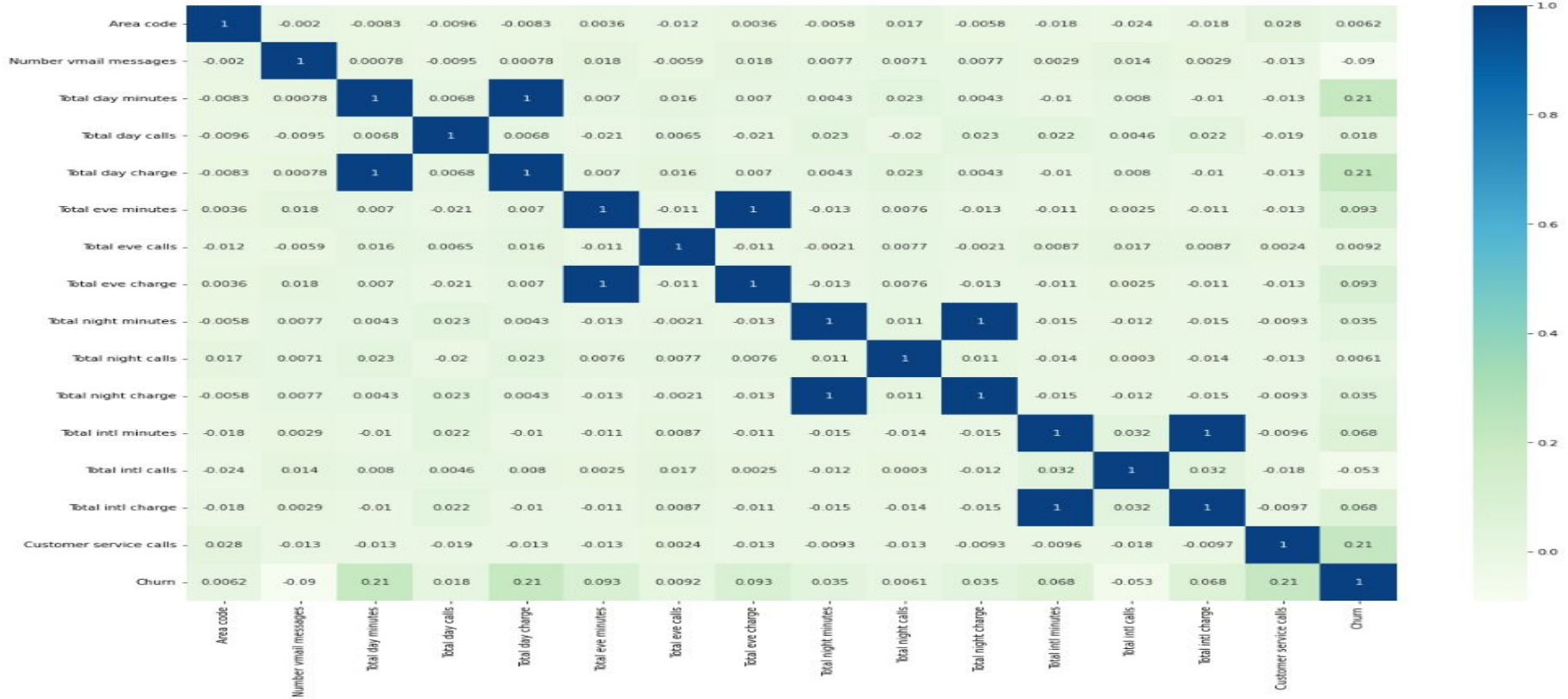


Churn According to Customer Service Calls

- Customers with more number of service calls (4–9 calls) are more likely to churn.
- Some customers have switched to other network operators without resolving their issue (i.e. 0 service calls), while customers who have called once also have a high churn rate, indicating that their issue was not resolved in the first attempt.
- Customer queries and problems should be resolved in minimum calls to avoid churn.



Heat map (to compare the correlation between features)



Heat map (Continued)

1. Perfect correlation between the following features of our dataframe:

- ‘Total day charge' and 'Total day minutes'
- 'Total eve charge' and 'Total eve minutes'
- ‘Total night charge' and 'Total night minutes'
- ‘Total intl charge' and 'Total intl minutes'

2. ‘Total day charge’, ’ Total day minutes’ and ‘Customer service calls’ are more positively correlated with churn among given features.

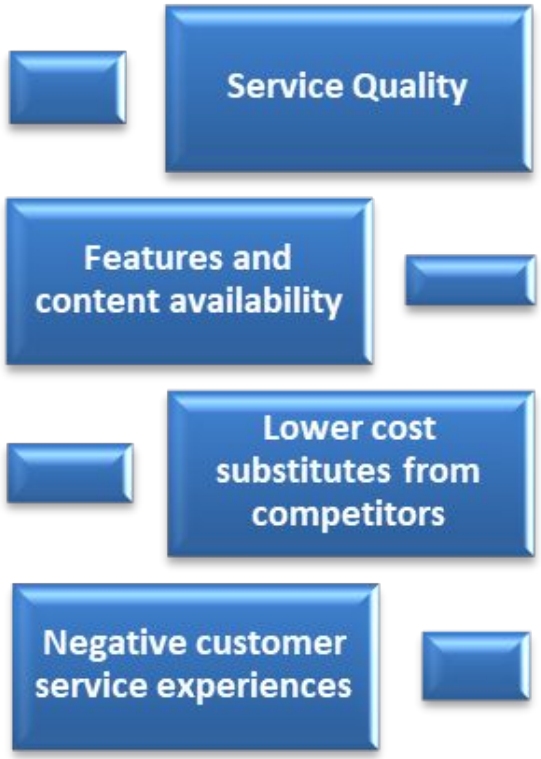
3. ‘Area code’, ‘Total day calls’, ‘Total eve calls’ and ‘Total night calls’ are positively correlated with churn.

4. ‘Number vmail messages’ and ‘Total intl calls’ are negatively correlated with churn.

Factors affecting churn

To avoid customer churn and retain existing customers in the telecom industries, companies should look for the factors affecting the churn rate. Although there are many factors that affect churn rate but following are some major factors:

1. Service Quality
2. Features and content availability
3. Lower cost substitutes from competitors
4. Negative customer service experiences



Why it is important to reduce customer churn?

- Customer churn is one of the major concerns for any industry, as we know that the cost of acquiring a new customer is much more than retaining an existing customer.
- Customers switch to different telecom operators for a variety of reasons, including service dissatisfaction, high subscription charges, and better options.
- There is tough competition due to an increase in the number of telecom industries. So, to sustain this competition, they try to retain their customers rather than acquire new ones, as it has proved to be much costlier. Hence, it is very important to predict and then reduce customer churn.
- To survive in this market, telecom companies must innovate, provide better services, and grow their customer base.

Conclusion:

In this project, we conclude that key factors responsible for customer churn in telecom industries are higher call charges, call drop, quality drop in voice-mail, network disturbance, and delay in resolving customer queries and problems.

To ensure customer retention, we recommend revising the pricing strategy, implementing a better network infrastructure in high churn areas (i.e., New Jersey, Texas, and Maryland), improving voice-mail features or setting a limit, updating and optimizing international call rates, upgrading the network to improve service for long call duration customers, and assuring customers that their queries and problems will be resolved in the first attempt or as soon as possible.


THANK YOU