Telecom Churn Prediction

Adityasingh Thakur

1.Abstract

Customers plays an import role in industry to run industry. Churn of the customer may lead many consequences. Customer churn prediction must the important aspect of any company. This helps in the detection of customers who are likely to cancel a subscription to a service. Recently, the mobile telecommunication market has changed from a rapidly growing market into a state of saturation the focus of telecommunication companies is to shift from growing of large customer into keeping customers in house. For that reason, it is valuable to know which customers are likely to switch to a competitor in future. We have used a raw data set of Telecom Churn Prediction. This data set contains 20different features that can be used for discover key factors responsible for customer churn and come up with recommendations to ensure customer retention.

2.Problem statement

Orange S.A., formerly France Telecom S.A., is a French multinational telecommunications corporation. The Orange Telecom's Churn Dataset, consists of cleaned customer activity data (features), along with a churn label specifying whether a customer cancelled the subscription.

Explore and analyse the data to discover key factors responsible for customer churn and come up with ways/recommendations to ensure customer retention.

3.Introduction

With the aspects of development and the advancements around the world communication is one of the main required entities in the development. Communication needs to fast and reliable for the people for their works and necessities to be done. In developed countries telecom industries post a major role and have become a part of necessities required for people to live. The developments and technological progress and the steady increase in the operating network the competition between them on peaked. Companies need to thrive to survive in the competitive market by implementing new strategies and policies for acquiring the customer base around the world. The main theme behind these strategies to develop and generate more income to the companies. There are different strategies followed by the marketing team in the companied to attract new customers, making the existing customers to buy or upgrade to new services within the same company and finally to keep the customer base for longer period.

4. Methodology

Before we start exploring our dataset, look at our analysis approach and steps that are involved in performing EDA:

Importing libraries- First, we imported all the python libraries required for this, which include NumPy for numerical calculations, Pandas for preparing data and Matplotlib and Seaborn for Data Visualization.

Loading the data into data frame- We read the CSV into a data frame and pandas data frame does the work for us. This is one of the most important steps in EDA.

Discover and access the data- After loading the data, it is important to discover the dataset. This step is about knowing the data and understanding what has to be done before the data becomes useful. Checking the first rows, last rows, shape of the dataset, columns and their data types are the basic things to look for.

Data Cleaning- Cleaning up the data is the most challenging and time-consuming part of EDA, but it's a crucial step for removing faulty data and filling up missing values. It is important to handle missing values effectively, as they can lead to inaccurate inferences and conclusions.

Data Visualization- Data Visualization is the graphical representation of information and data. It can help in interpreting and understanding the data, identifying trends, highlighting important relations with the help of charts and graphs.

5. Data Overview

Before performing any operation on the dataset, it is important to understand the data at a high level. Depending on size and type of data, understanding and interpreting data sets can be challenging. After loading data, we observed the dataset by checking a few of the first and last rows. We checked the shape of the dataset and identified that there are 3333 rows and 20 features(columns) in our dataset. We observed that there are different types of data present in the dataset such as float, string, object. There are categorical variables as well as numeric variables present in the dataset.

State: - 51 Unique States in United States of America.

Account length: - Length of The Account.

Area Code: - Certain unique code of area.

International Plan: -

- Yes (Indicates International Plan is Present).
- No (Indicates no subscription for International Plan).

Voice Mail Plan: -

- Yes (Indicates Voice Mail Plan is Present).
- No (Indicates no subscription for Voice Mail Plan).

Number v-mail messages: -Number of Voice Mail Messages ranging from 0 to 50.

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 evening minutes: -Total Number of Minutes Spent by Customers in Evening.

Total evening calls: -Total Number of Calls made by Customer in Evening.

Total evening charge: -Total Charge to the Customers in Evening.

Total night minutes: -Total Number of Minutes Spent by Customers in the Night.

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 international minutes used.

Total into calls: -Total international calls made.

Total into charge: -Total international charge.

Customer service calls: -Number of service calls made.

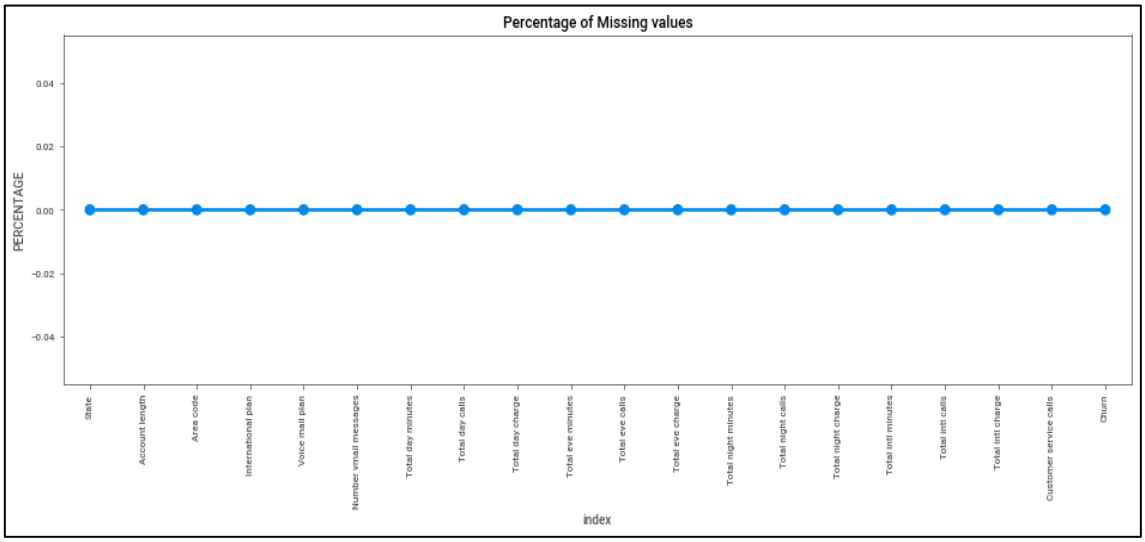
Churn: - Customer churn (Dependent Feature True=1, False=0).

6. Data Preparation

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is the most important step before performing any analysis. Data preparation usually includes handling missing values, standardizing data formats, enriching data and/or removing outliers. Good data preparation allows for efficient analysis, limiting errors and inaccuracies that can occur to data during processing.

7.Data Insights

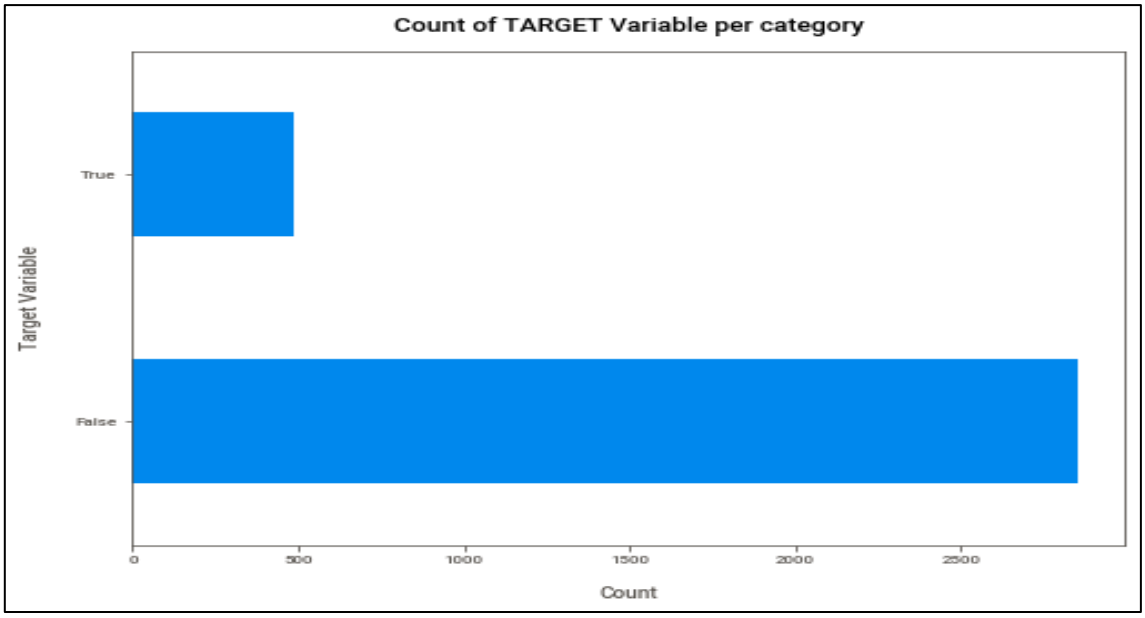
7.1 Find Null Values in Dataframe Using the Point-plot?



Conclusion Drawn: -

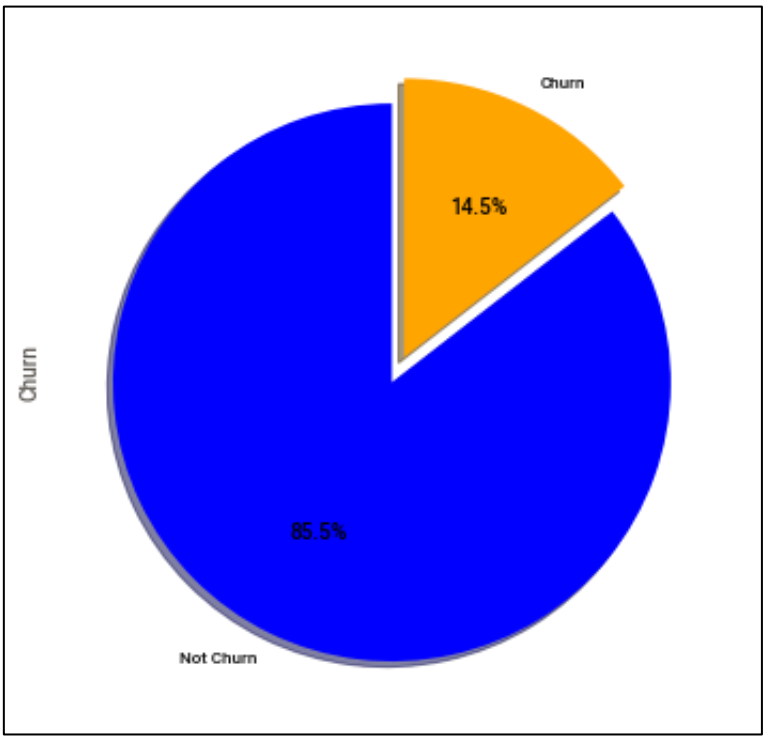
- Here we don't have any missing values and as we clearly see from the graph that the percentage of missing values lies in a horizontal line.
- We can see that there in not a single missing value in our dataset

7.2 Find the Total Number of Churn Data?



- From the above graph we can conclude that there are more False Values compared to true Values as using the value counts () method there are: - False Value Counts: -2850 And True Value Counts: -483

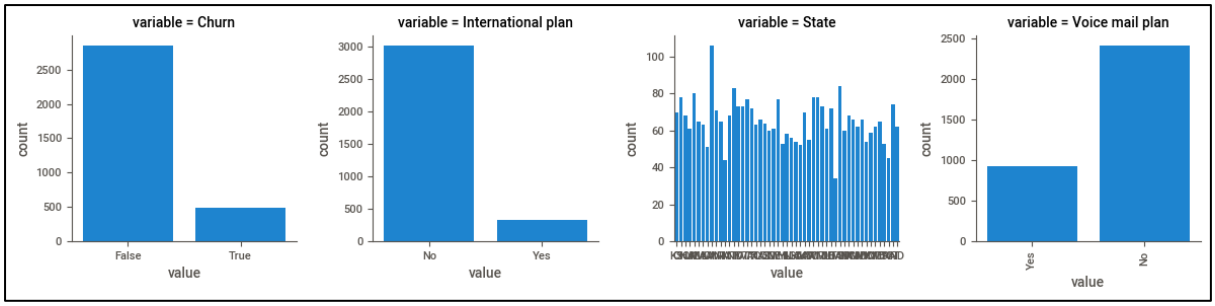
7.3 Find the Churn Rate and Plot the Pie Chart for Churn or Not Churn Data?



Conclusion Drawn: -

- From the above graph we conclude that there is 14.5% Customer Churn And 85.5% Not a Customer Churn. Data is highly imbalanced, ratio = 85:15, So we analyse the data with other features.

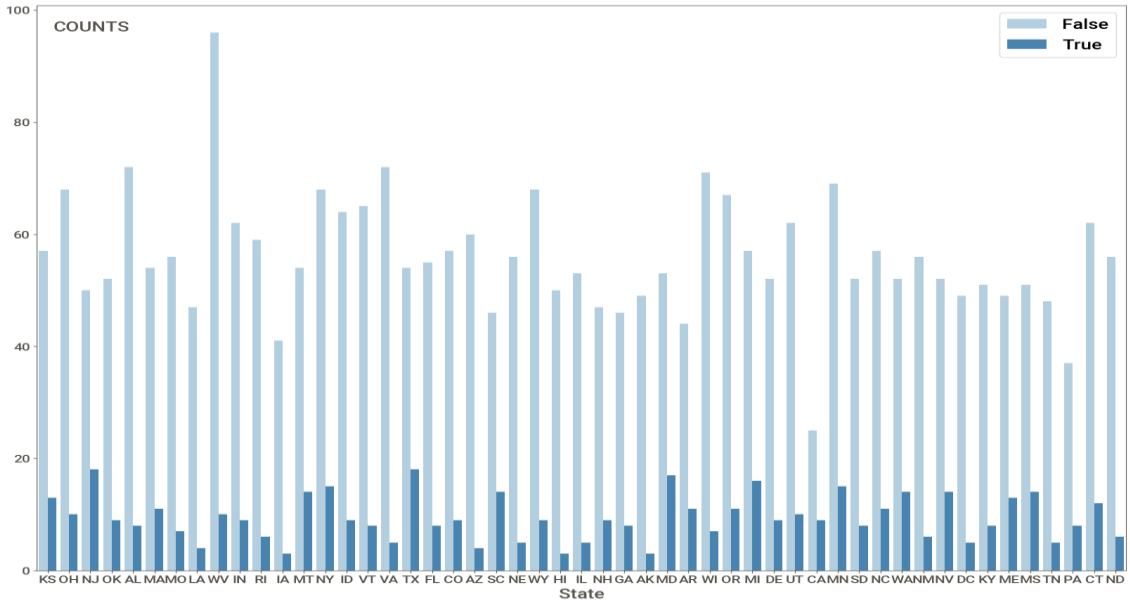
7.4 Plot the FacetGrid for Categorical Feature 'variable' And Determine Its Nature with Other Categorical Features?



Conclusion Drawn: -

- We have used pd. melt function to change the DataFrame format from wide to long. It's used to create a specific format of the DataFrame object where one or more columns work as identifiers. All the remaining columns are treated as values and unpivoted to the row axis and only two columns – variable and value
- We have also used facet grid for visualizing the distribution of variables of a dataset and the relationship between multiple variables.

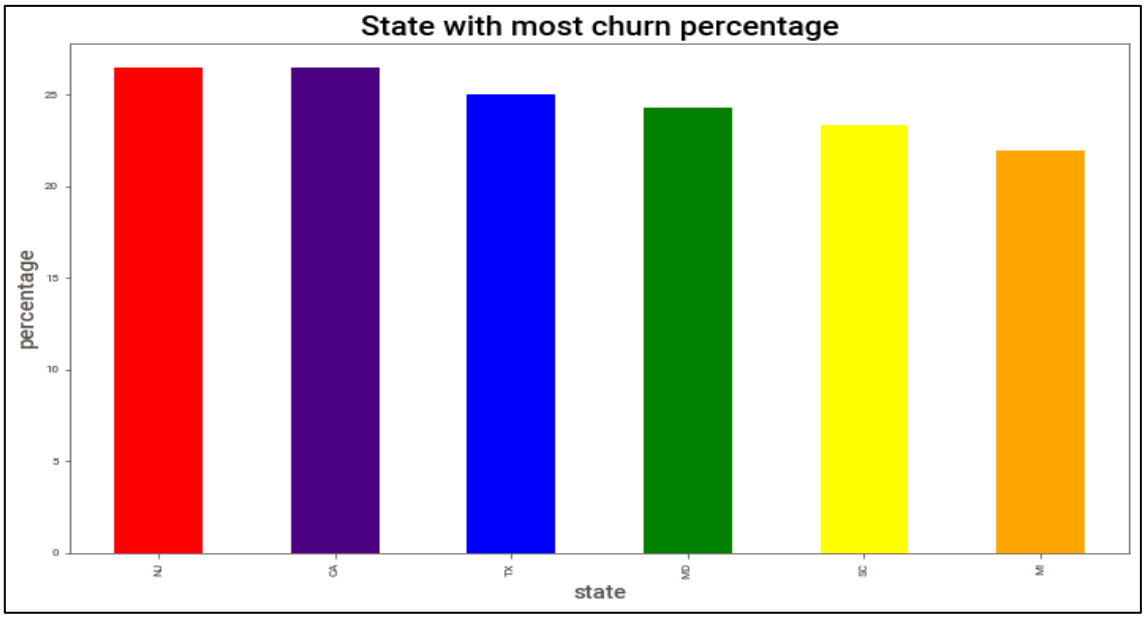
7.5 Plot A Countplot for State Vs Counts and Determine the Values of Respective States?



Conclusion Drawn: -

- We have used Countplot to show the counts of observations in each categorical bin using bars. We have also used Subplot to create a figure and a grid of subplots with a single call, while providing reasonable control over how the individual plots are created.
- From the count plot we can easily say that State 'WV' contains large number of False values as compared to other states in the graph. The State 'CA' Contains least number of false values.
- The States 'NJ' And 'TX' contains large number of true values from All other states in USA. While 'LA' , 'IA' ,'HI' And 'AK' this following states contains least number of true values.

7.6 Using the Bar Graph, Find the Churn Percentage of Various States in Our Dataset?

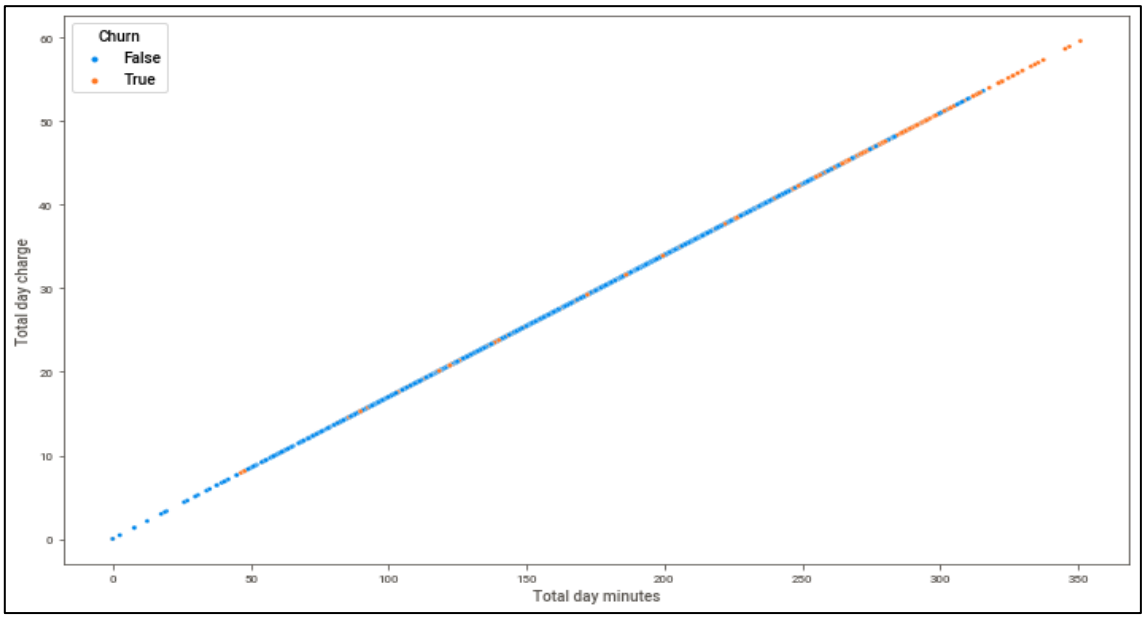


Conclusion Drawn: -

- By importing matplotlib we have also imported the sub-module name rcParams for handling default matplotlib values.
- From the graph we can easily say that the states "NJ" And "CA" are the highest with most churn percentage [Both states more than 25%] While the state "MI" have least churn percentage [Which is approx. 22% to 23%].

7.7 Plot Scatterplot For

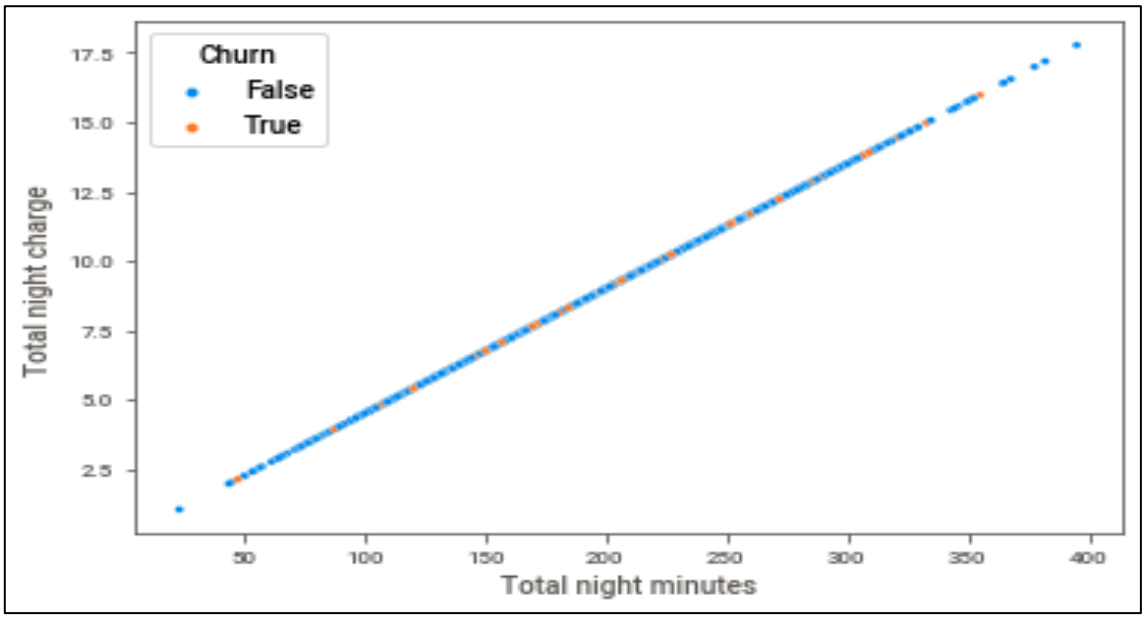
1. "Total day minutes" Vs "Total day charge"



Conclusion Drawn: -

- From the graph we can conclude that there are more false values compared to true values.
- We can also see that there is a linear relationship between total day minutes and total day charge [It means if the total day minutes increases total day charge also increases].

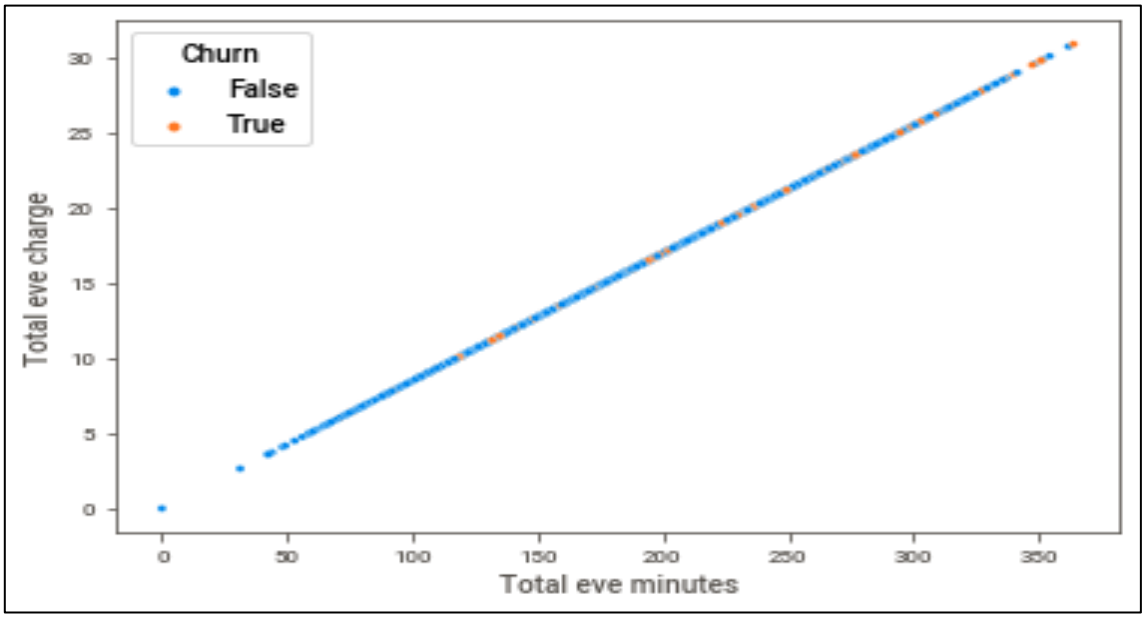
2. Total night minutes' Vs 'Total night charge'



Conclusion Drawn: -

- From the graph we can conclude that there are more false values compared to true values.
- We can also see that there is a linear relationship between total night minutes and total night charge [It means if the total night minutes increases total night charge also increases].

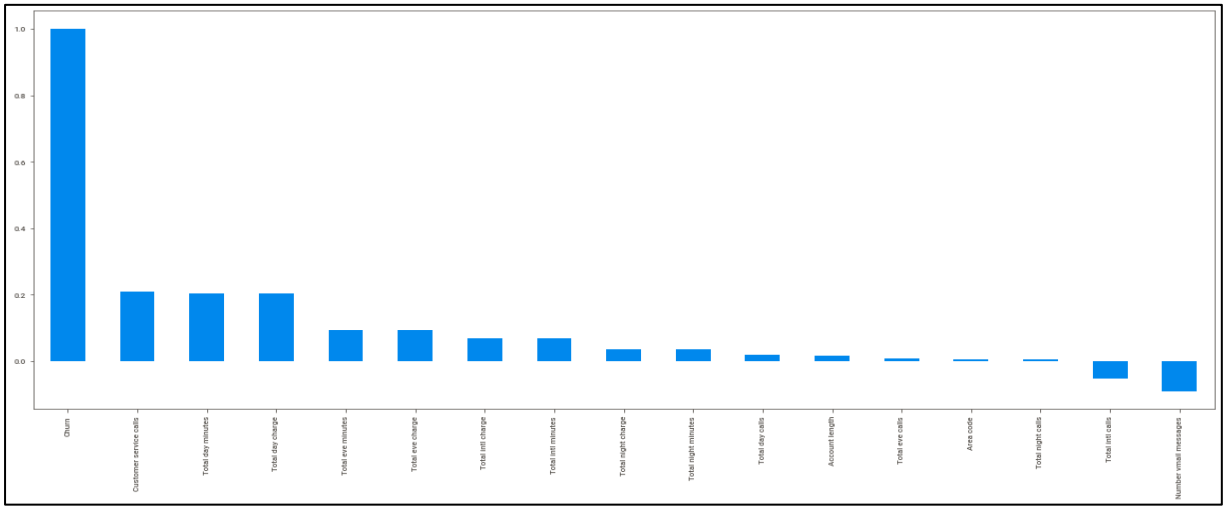
3. "Total eve minutes" Vs " Total eve charge"



Conclusion Drawn: -

- From the graph we can conclude that there are more false values compared to true values.
- We can also see that there is a linear relationship between total eve minutes and total eve charge [It means if the total eve minutes increases total eve charge also increases].

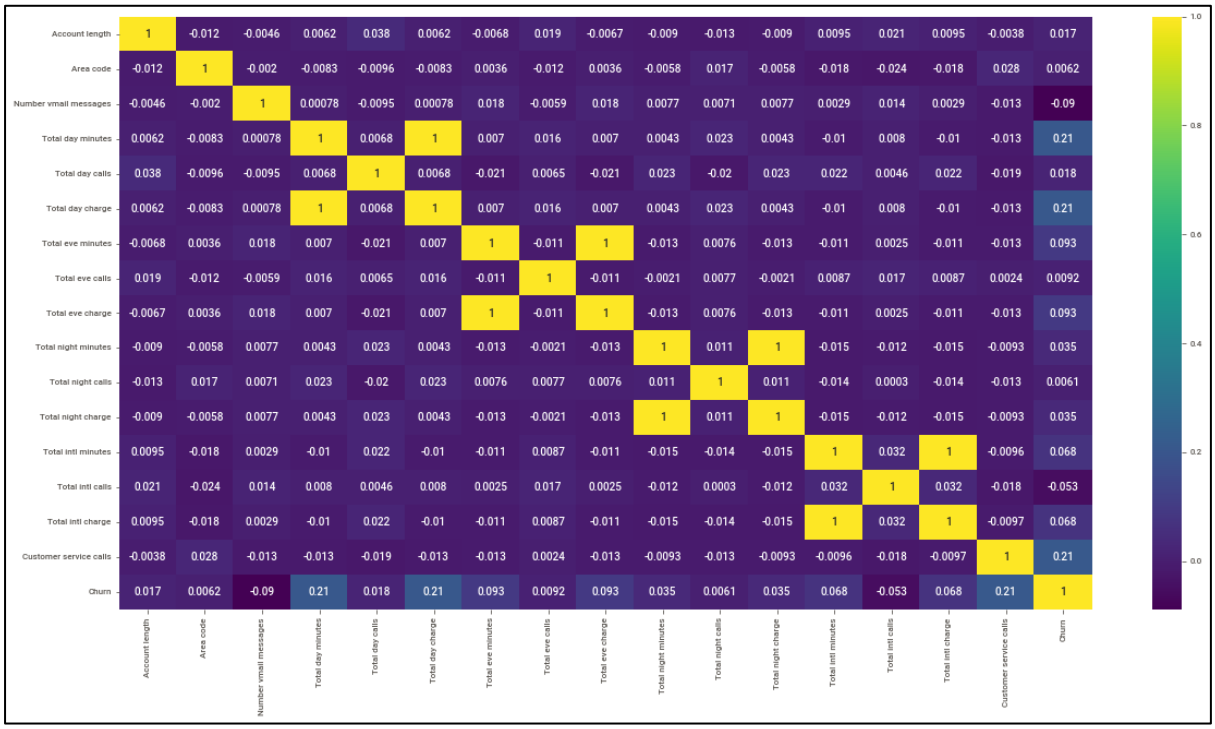
7.8 Plot Correlation Bar Graph of Feature 'Churn' With Other Features in Our Dataset?



Conclusion Drawn: -

- From the graph we can easily say that the column 'Churn' occupies a very large value [1.0] as compared to other columns.
- Columns such as 'Customer service calls', 'Total day minutes', and 'Total day charge' occupies equal value of [0.2].
- Columns such as 'Total eve minutes' And 'Total eve charge' occupies an equal value of [0.1].
- Column 'Total night calls' consumes a very less value in the entire graph.
- Columns such as 'Total intl charge', 'Total intl minutes', 'Total night charge', 'Total night minutes', 'Total day calls' and 'Account length' occupies an equal value of each other.

7.9 Plot Correlation Heatmap of All the Categorical Features Present in Our Dataset?



Conclusion Drawn: -

- Each square shows the correlation between the variables on each axis. Values closer to zero means there is no linear trend between the two variables. value close to 1 the correlation is the more positively correlated. The diagonals are all 1 because those squares are correlating each variable to itself (so it's a perfect correlation). For the rest the larger the number and darker the colour the higher the correlation between the two variables. The plot is also symmetrical about the diagonal since the same two variables are being paired together in those squares.
- Squares containing a grey colour having the value of [0.21] are least located in the graph.
- We can also see that square containing light as well dark purple colour have some negative values whereas some of purple colour squares contains positive values as well as shown in figure above.

8.Solutions to Avoid Customer Churn: -

- We can set up a limit on voice mail service strictly no more than 25 voice mail.
- Solve the network problem.
- The clients who have high call minutes and calls need a discount in the end.
- Different pricing strategy and international calling rate optimization would lead to lower churn rate.
- Stay competitive.

9. Final Conclusion: -

- It can also be observed that most people who use the service in the morning speak for shorter amounts of time but make more calls.
- International plan users are more consistent with their churn w.r.t the ones who do not have the service.
- Customers with the International Plan tend to churn more frequently.
- Customers with four or more customer service calls churn more.
- Customers with high day minutes and evening minutes tend to churn at a higher rate.