AlmaBetter

Project-EDA on Telecom Churn

Sumit Jadhav, Rushikesh Shinde,
Aakash Jadhav, Aniket Okate

Abstract:

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom field, 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.

Customer churn is a big concern for telecom service providers due to its associated costs. This short paper briefly explains our ongoing work on customer churn prediction for telecom services. We are working on data mining methods to accurately predict customers who will change and turn to another provider for the same or similar service.

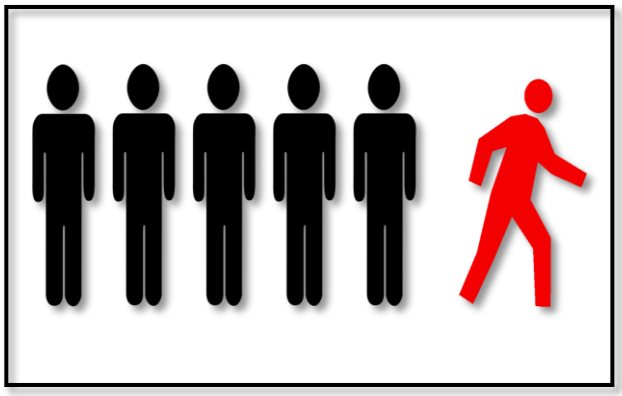
On this dataset we find out behavior of the churn as well as non churn customers. Work on the problems faced by the customers and try to solve them.

Keywords:

EDA, Data analysis, Visualization,, Behavior of dataset, Data preprocessing, Data cleaning.

Introduction:

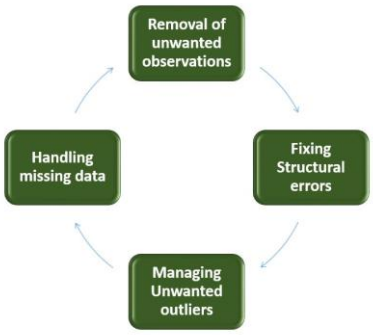
Formerly France Telecom S.A., is a French multinational telecommunications corporation. The Orange Telecom's Churn Dataset, consists of cleaned customer activity data (features), along with a churn label specifying whether a customer canceled the subscription. Explore and analyze the data to discover key factors responsible for customer churn and come up with ways/recommendations to ensure customer retention.



Rapid improvements and dynamics in technology market place make customer retention a competitive effort. Especially in saturated telecommunications market, there are incumbent service providers and newcomers offering deals and packages for consumers who would like to churn to their services. On the defending end, strategies and counter offers have to be made for potential churners as it is more expensive earning a customer back once s/he churns. According to the SAS Institute report, the annual rate of customer churn in telecommunications industry is currently at about 30% with an upward trend in correlation with the growth of the market.

Objectives:

- Data Summery.
- Data Cleaning.
- Visualization to find out behavior of data.
- Other visualization by various features.
- Conclusion



Actual Work:

1.Data summary

Find brief information about dataset. Checkingdata types of each feature. If there is any kind of problem in data type then change data type. Separate out numerical and categorical features. Find unique values in each feature and also find count of unique entries in the feature.

Also find mean, mode, max, min, SD values of each features. At the end give proper overview of the dataset.

2.Data Cleaning

Checking null values are present in the dataset. If present then null value cells are replaced by mean, mode value or if this is a categorical variable then replaced with another word.

In our case there is no null values present in the dataset so there is no need to apply any operation to clean the dataset.

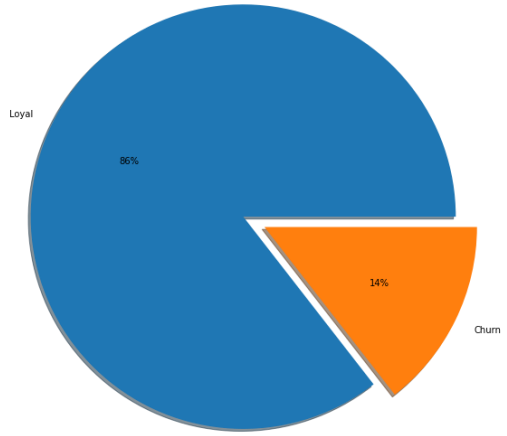
3.Data visualization

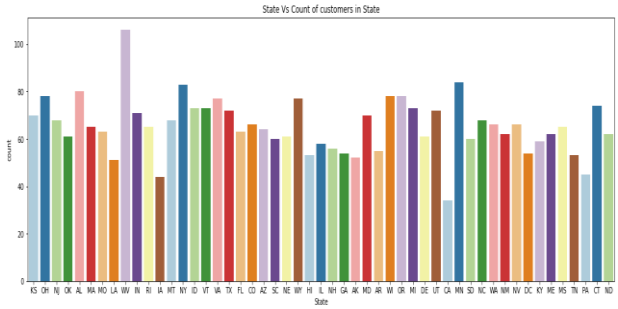
-For this dataset we select “Churn” as a target variable. This contains Boolean values (TRUE/FALSE).
-As well as we have 19 independent variables which can used for find out behavior of dataset.
-For visualization we used matplotlib and seaborn library.

A)Visualization for finding data behaviour.

1.Number of churners

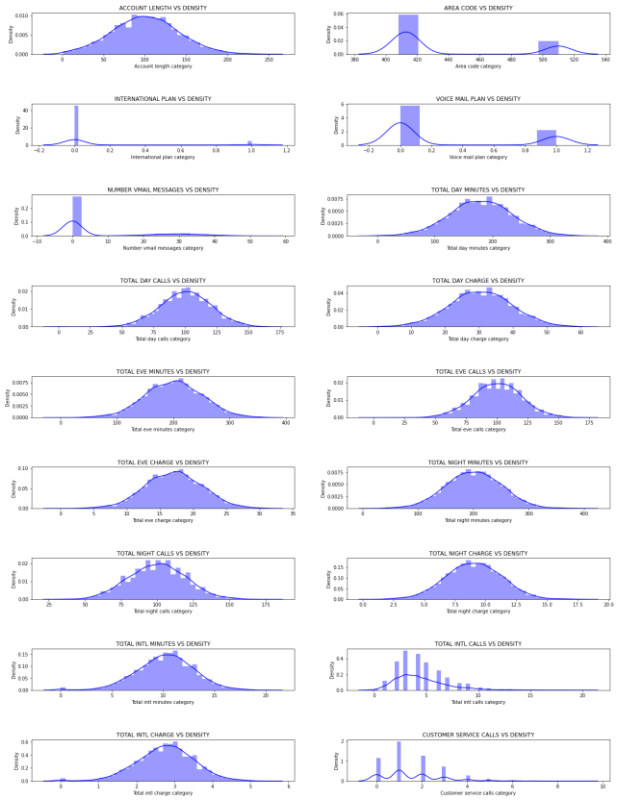
This chart shows percentages of loyal and churn customers.
Loyal Customers- 86%
Churn Customers- 14%





2.Distribution of Numerical variables by density plot

Most of the plot show normal distribution curve. So we can say that, maximum number of features can show normal distributed data.

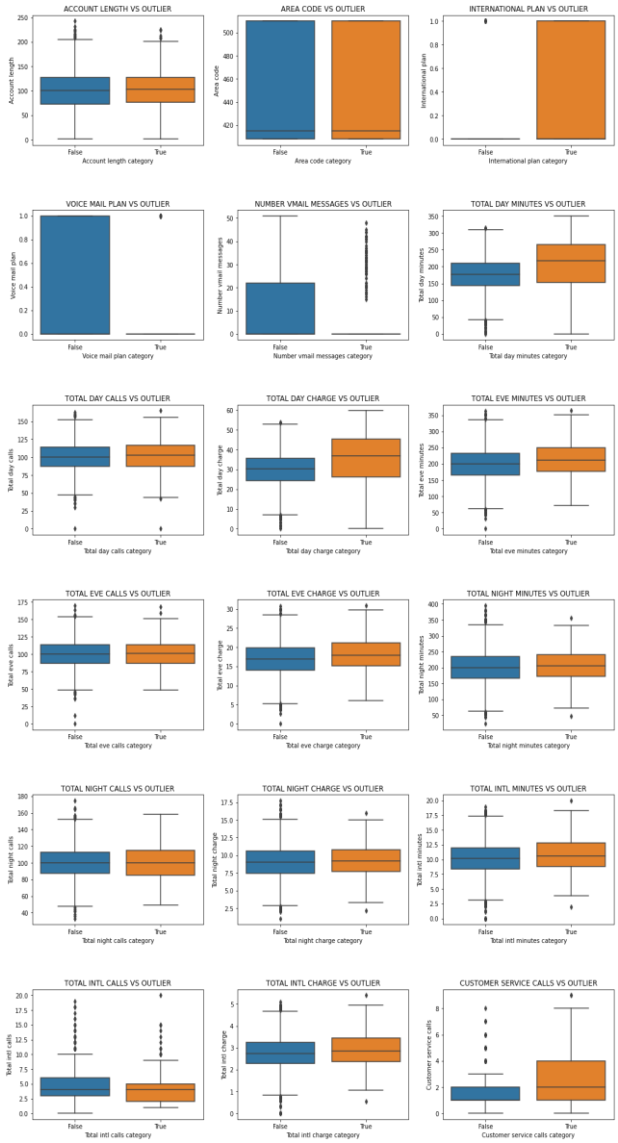


3.Distribution of Categorical variables by count plot

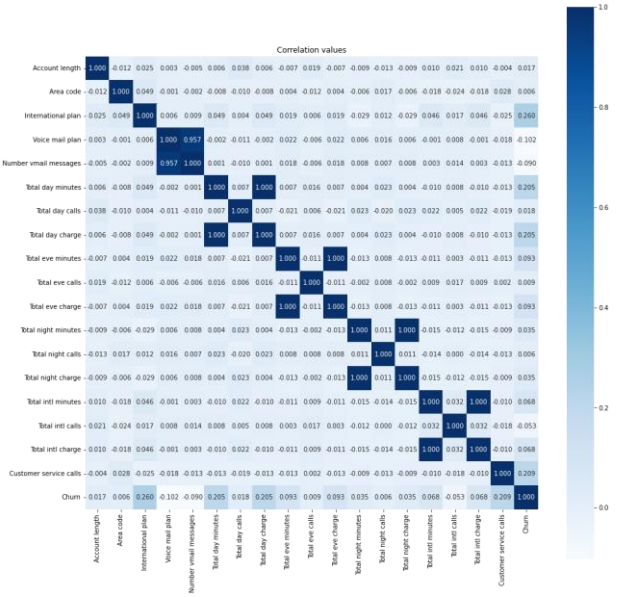
This plot can shows the information about each state contains how many number of customers are present

4.Check outliers in the dataset

This boxplot show us, In some cases outliers are less and in some cases number of outliers are more.



5.To check correlation with target variable

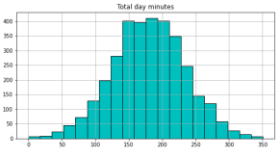


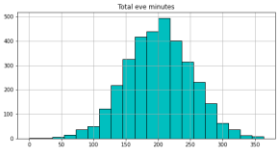
There is no as such correlation in dependent and independent features. International plan have high positive correlation value(0.26) and voice mail plan have high negative correlation value(-0.102)

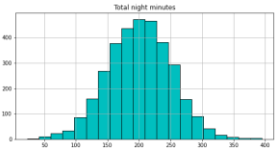
B) Visualization for Other Analysis

1.Checking minutes, calls and charge by day, evening and night

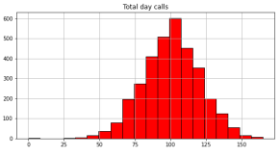
-Minutes

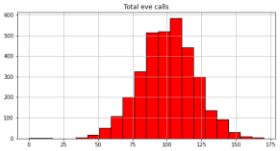


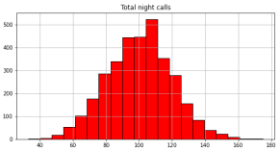




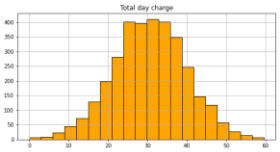
-Day

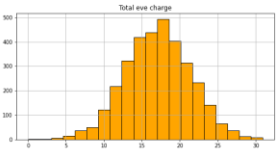


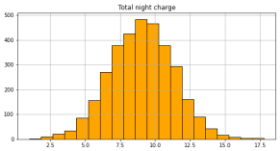




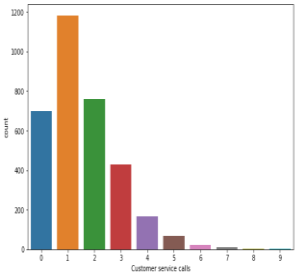
-Charges

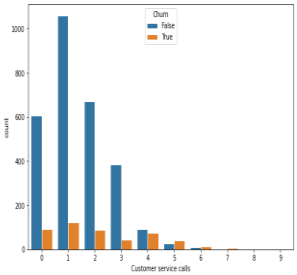






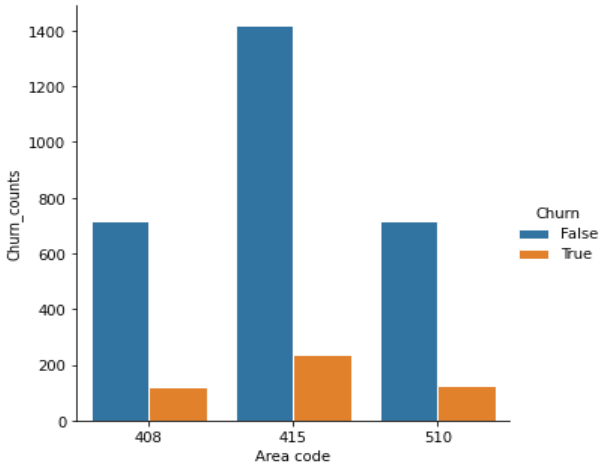
2.Impact of customer service calls on Churn

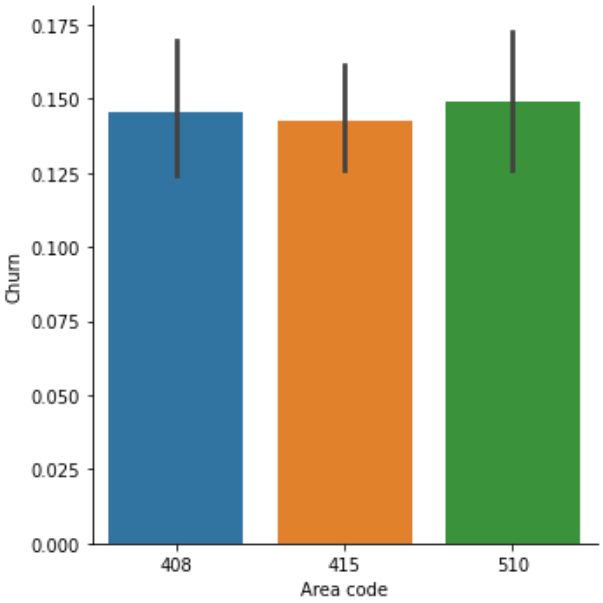




As customer increases customer service calls might they are not satisfying with the solution given by the resolution of the problem so customers frustrate and getting churn in this case while Some customers are lazy and hence without resolving the issue they have jumped to other network operator, while the customers who have called once also have high churn rate indicating their issue was not solved in first attempt

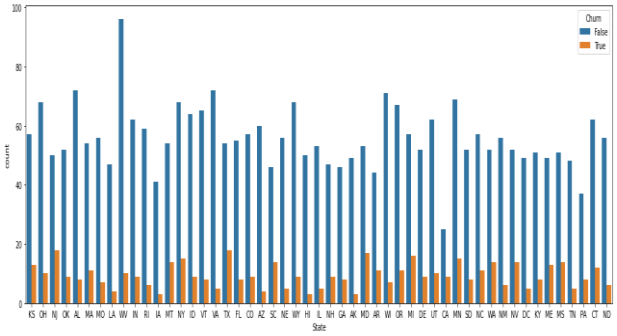
3.Churned and non churned data in specific Areas using their area codes





We conclude that area code 415 have more number of customers and number of churn customers are also more.

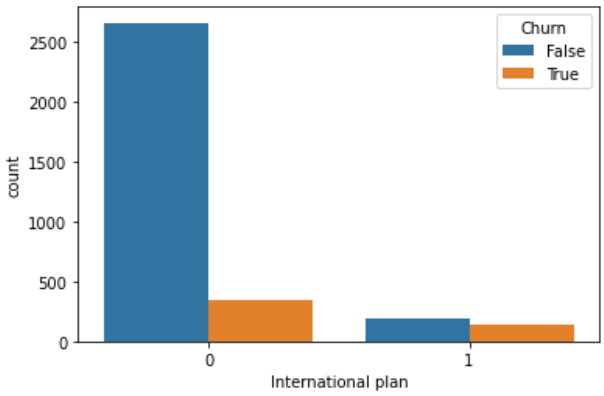
4.In which state have more number of churners

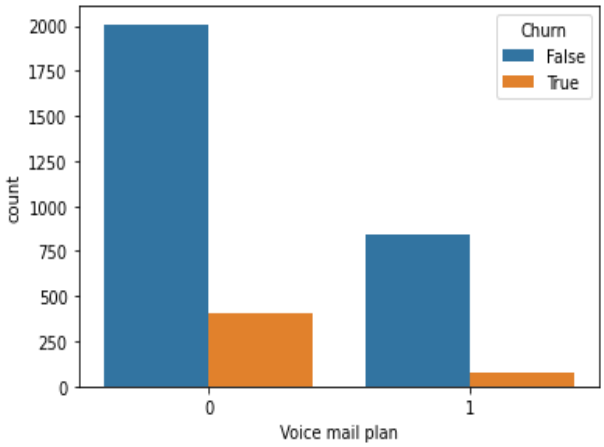


"WV" state has more number non churn(Loyal) customers present and "TX" & "NJ" state has more number of churn customers.

5.find churn count according to their plans

International plans and voice mail plan gives complex output.Means customers having international plan or voice mail plan both can either churn or non churn according to there convinence





Conclusion:

1.Find out data summary.

2.After checking null values and data summary it shows that the given dataset is cleaned dataset.

3.Find unique values in each feature.

4.Saperate ot numerical and categorical features.

5.Some datatypes of feature are changed for our convenience.

Visualization:

1.Persentage of churn and loyal customers

-Loyal Customers= 86%

-Churn Customers=14%

2.Most of the plot show normal distribution curve.So we can say that,maximum number of features can show normal distributed data.

3.Find out count of information about each state contains how many number of customers are present.

4.In some cases outliers are less and in some cases number of outliers are more.

5.There is no as such correlation in dependent and independent features. International plan have high positive correlation value(0.26) and voice mail plan have high negative correlation value(-0.102). There is no any strong relationship between output data that is churn and other features.

6.Data of minutes, calls and charge by day, evening and night is almost normally distributed.

7.After analysis of the data it is conclude that the as customer service call getting increases the percentage of that account getting churned is increases So Company should give promise that We will probably resolve your problem in first attempt if not possible then politely take some time and resolve that problem in that time so customer will stay with company.

8.We conclude that area code 415 have more number of customers and number of churn customers are also more.

9."WV" state has more number non churn(Loyal) customers present and "TX" & "NJ" state has more number of churn customers.

10.The customer having International plans are getting churned as compared to the customers having no international plan so it might be due to high charges on international calls so company should try to minimise the international calls charges.

11.International plans and voice mail plan gives complex output.Means customers having international plan or voice mail plan both can either churn or non churn according to there convinence.