TELECOM CHURN ANALYSIS

Nikita Wangmare, Adarsh Kumar, Shobhit Bahuguna
Data science trainees,
AlmaBetter, Bangalore

Abstract:

One common problem with the telecom industry is many of the customer withdraw their services after a particular duration which is commonly known as churn. We are provided with labelled data of a telecom Industry.

In this project, we perform exploratory data analysis using some python library like pandas, matplotlib and seaborn. Here, we experiment with the features and try to perform a comparative study on various feature to the customer who churn and who do not.

Keywords: python, pandas, matplotlib, seaborn

1.Problem Statement

Dataset provided by Orange S.A., formerly France Telecom S.A., which is a French multinational telecommunications corporation. The dataset consists of 3333 rows and 20 columns i.e., 20 features.

In this analysis, we try to solve these two problem statements:

- Identification of the difference in the feature values of the customers who churn and who does not by exploring various features.
- Identify some KPI's (Key performance indicators) which can help in reducing the churn rate of the company.

2. Introduction

Telecom industry faces a major business problem that many of its customer withdraw their services after a certain period of time, to indicate this rate a technical term churn rate is used.

Churn rate can be defined as the percentage of the total customer withdrawn the service during a particular period out of the total customer at the beginning of that period.

3.Dataset

To perform this analysis, we have been provided with dataset of 20 features. These features are following:

- State: States of the customers.
- Account length: length of accounts of customers.
- Area code: Area code of customer where he uses the service.
- International plan: Does customer have activated international plan subscription.
- Voice mail plan: Does customer have activated voice mail plan subscription.
- Number Vmail messages: Number of voice mail messages customer has received.
- Total day minutes: Total minutes spend by customer in daytime on call.

- Total day calls: Number of calls by customer during day time.
- Total day charge: Total charge of phone calls by customer during day time.
- Total eve minutes: Total minutes spend by customer on call at evening.
- Total eve calls: Number of calls by customer during evening.
- Total eve charge: Total cost of phone calls due to calls at evening.
- Total night minutes: Total minutes spend by the customer on call during night.
- Total night calls: Number of calls by customer during night.
- Total night charge: Total cost of phone calls due to calls at night.
- Total intl minutes: Minutes spend by customer on international calls.
- Total intl calls: Total number of calls made internationally by customer.
- Total intl charges: Total charges due to the international calls by the customer.
- Customer service calls: Total number of times customer made call to the customer service.
- Churn: The value is true if customer leaves during the specified period and false if not.

Although, In EDA there is no feature called target variable but here we are trying to identify the behavior of the customer who churns so we can say that Churn feature is our target variable which of Boolean type and we try to implement our EDA around this target variable. Apart from that, 3 object type features, 8 float and 8 int type features are present in the dataset.

4.Python Libraries:

The libraries used to complete this project are pandas, matplotlib and seaborn.

Pandas are used to perform the data wrangling and analysis. Matplotlib and seaborn is used for plotting the various graphs for visualization purposes.

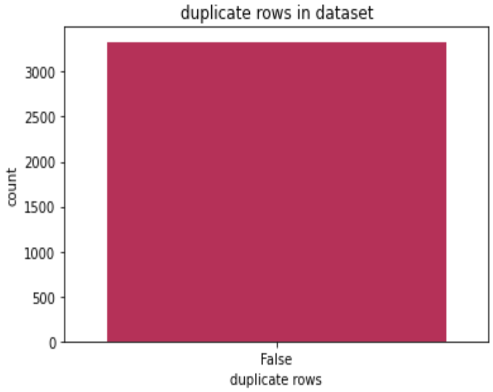
5. Steps involved:

There are two major steps involved in this project-

- Data cleaning: we checked for the null values and the duplicates present in the data.
- Exploratory data analysis: We will start our exploration with categorical feature then we will move towards the numerical feature.

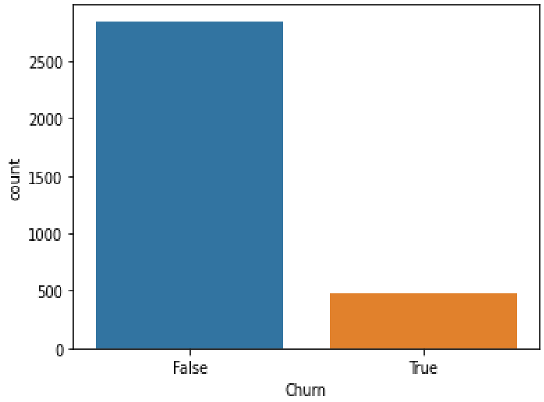
5.1 Data cleaning:

- No null values present the dataset, so no need for any kind of imputation.
- No duplicates found in the dataset as we can see below that all values are false.



5.2 Exploratory data analysis:

We started with calculating the churn rate of the company which is defined as the ratio of the total customer left the services during a particular period to the total customer at beginning of the period. The histogram below shows the total customer churns and those who does not.

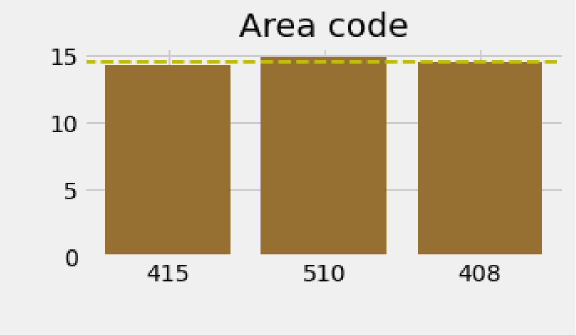


Out of 3333 customers, total 483 customers churns and rest do not. So, the churn ratio of the company is approximately 14.49 %.

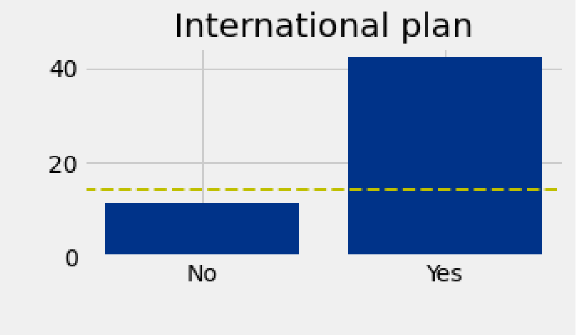
5.2.1 Categorical feature:

In categorical feature, we calculate the churn ratio for various category of features like Area code, international plan, voice mail plan and State.

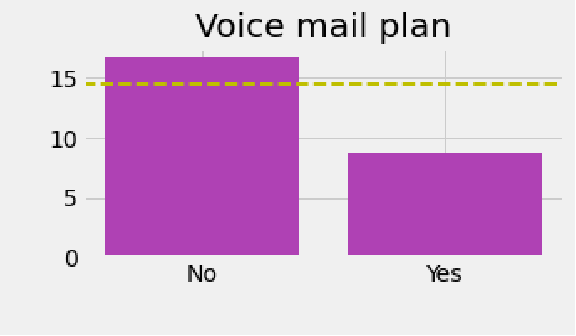
Area code 510,408 have the churn rate greater than the company churn rate. Area code 510 have the maximum equals to 14.88% i.e., out of 840 people 225 churns. However, the maximum number of people who churns are from 415 (236 out of 1655, churn rate equals to 14.26%).



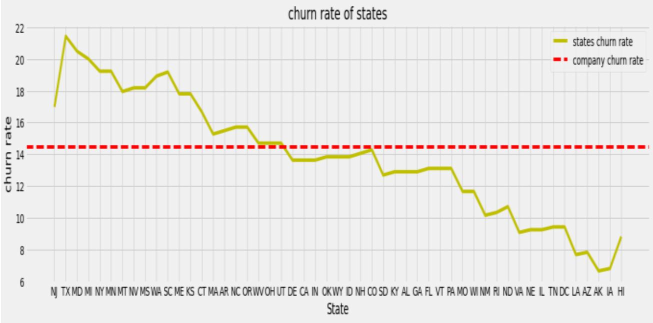
For International plan feature, the customer who have subscribed for the churn rate have the highest value of churn rate (137 out of 323 making it equal to 42.41%) among all the categorical features.



For voice mail plan, people who does not have subscribed for voice mail plan have highest churn rate of approximately 16.715%.



For state features, the highest churn rate is from state TX equals to 21.43% approximately, however the maximum count of people belongs to the state who churns is NJ. Total 21 states have higher churn rate than the company churn rate. The figure below shows the various states having churn rate higher than company churn rate.

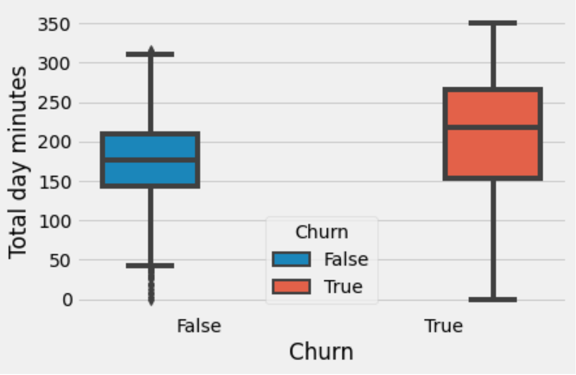


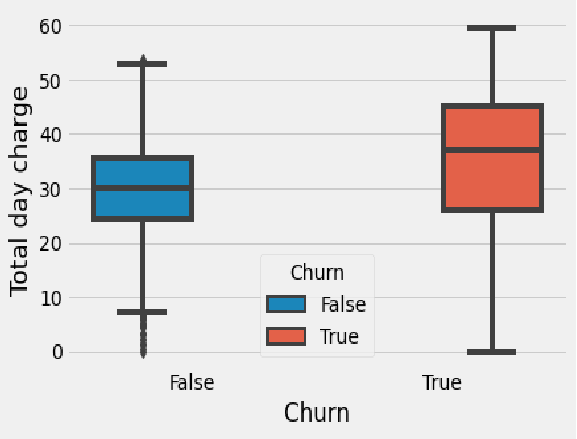
5.2.2 Numerical Feature:

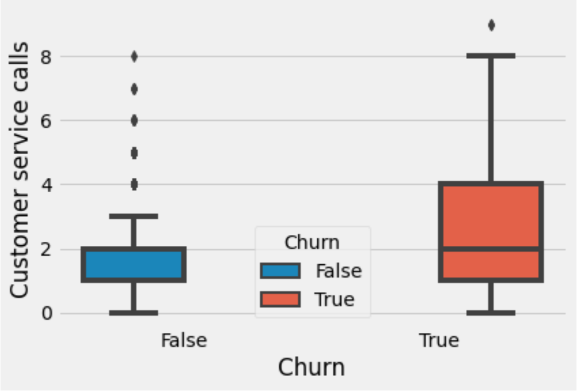
We started the exploration of numerical feature by plotting the boxplot of each numerical feature and comparing the mean, inter quantile range (25% to 75% percentile) value of people wo churn and who does not.

By doing that we found the three most important feature (Total day minutes, Total day charge, Customer service calls) which also have higher correlation value comparison with others for the target value. So, we plotted them individually and made some observations.

From total day minutes plot, it can be seen that more than 50% of customer who churns have total minutes equal 217, but less than 25 % customers who do not the churn has total minutes duration of 210. The average number total day minutes is higher for the customer who churns and the similar pattern has been observed for the Total day charge value i.e., the 50-percentile value of customers who churn is higher than who do not churn. It might be due that both the feature is linearly related having the correlation coefficient value equal to 1. The box plot of total day minutes, total day charges and customer service calls are shown below.







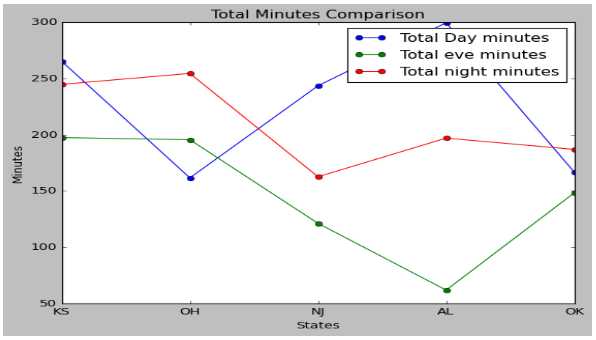
The below figure shows the boxplot of the customer service calls of customers who churns and those who does not. Approximately 25% of customer who churns make calls to customer service more than 4 times but those who does not churn only 5 peoples have call more than 4 times to customer services. The mean value of the customer who churns is twice than who does not.

In next step, we plotted heatmap of all the features and from there it was observed that the correlation value of "Total day minutes", “Total day charges”, “Customer service calls" have higher value comparatively. “Number Vmail messages" and "Total intl calls" only have negative correlation.
- Total day minutes and Total day charge, Total night minutes and Total night charge, Total intl minutes and Total intl charge are

highly correlated with each other have correlation coefficient value equals to 1.

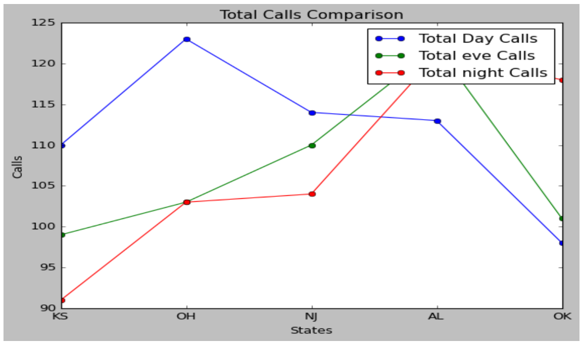
5.2.3 using both categorical and numerical feature:

Further we try to understand the behaviour of over all user by comparing the total minutes spend in a day, evening and night of the customer for first five states respectively as shown in below graph.

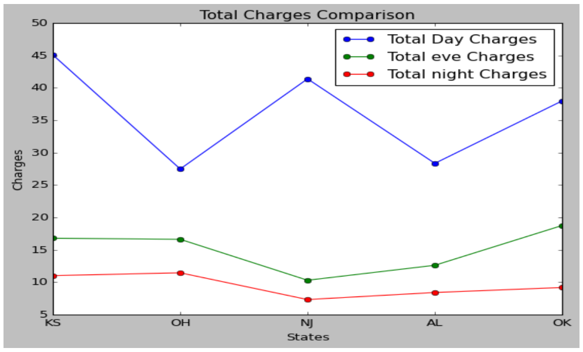


From above figure, we can infer that the Total minutes spent in a day for states OH and OK are in the range of 150 to 175 minutes while the same for state AL are maximum up to 300 minutes and if we check the same for evening, maximum time depleted is up to 200 minutes by the states KS and OH on the contrary with state AL who spends hardly less than 50 minutes. For Night time, all the states consumed estimated duration is in the range of 150 to 260 minutes.

Similar study for calls and charges has also been done for these five states. The graphs were shown below:

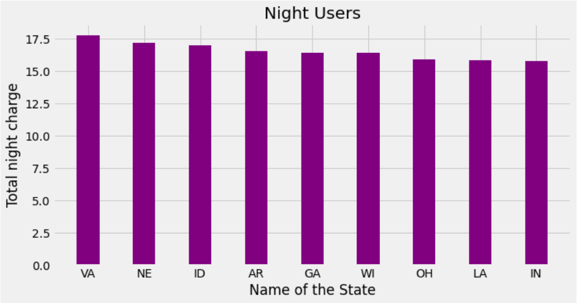


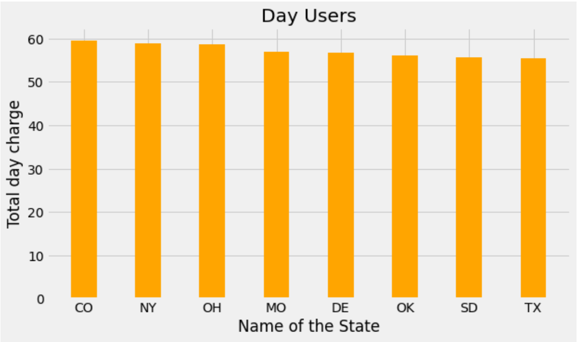
From the plot above, we can predict the maximum calls done by state OH are in the day time. On the other hand, maximum calls done by state AL are in the evening and night time. In addition to this, Customers in State KS prefers minimum number of calls in night time.



From the above line graph, it can be inferred that the customers in the above states have to pay more charges if they want to call in the day time. While if they choose to call in the night time, they have to pay minimum wage as compared to day.

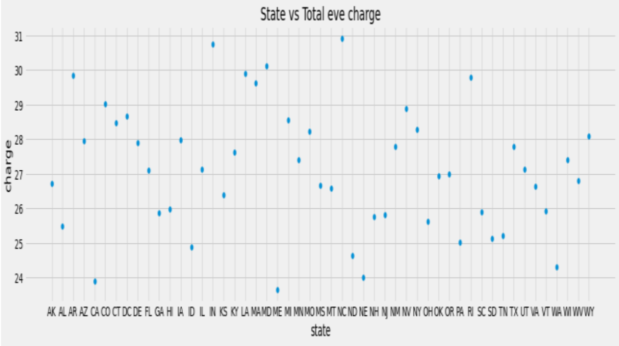
Similarly, we find out the states having maximum day and night calls which is shown below.





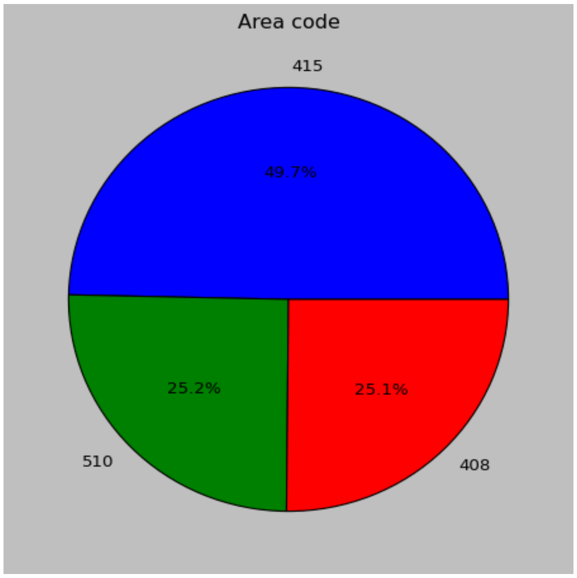
The states which are common in both is OH.

Then we find out the total eve charges for all the states as shown below:



From the above scatter graph, we can anticipate that the total evening charges applicable on states IN and NC are maximum, while for the states AR, LA, MA, MD and RI are comparatively higher than the average charges applicable but other states. States CA, ME, NE and WA are paying minimal amount for the total evening charges.

Based on Area code we visualise that from which area code most of the customer belong using pie chart below.



This Pie chart show that approximately 50% of the total states comes under area code '415' and 25.2% of states comes under code '510' and rest 25.1% comes under area code '408'.
This seems that majority of the customers of the are from this Area code '415'.

- Majority customers >> '415'
- Minority customers >> '408'

6. Observation:

After performing the exploratory data analysis of the above dataset, the following observations were made:

- categories having churn rate higher than company churn rate are Area code (510, 408), customer having no voice mail plan activated.
- out of 51, 21 states have churn rate higher than the company churn rate.
- Customer having international plan have the highest churn rate among all the category explored which is 42.41 % i.e., customers having international plans, 137 out of the total 483 churns.
- States named TX, MD and MI have maximum churn rate.
- Total day minutes and Total day charges of person who churns are higher than who do not churn.
- more than 75% of people who churn have 0 vmail messages, in fact very few of them have vmail messages more than 0.
- Approximately 25% of people who churns calls customer service more than 4 times but only 5 people from the people who do not churn makes customer service calls more than 4 times. Also, the mean value of customer service calls by people who churn is approximately twice than people who do not churn.
- Total day minutes and Total day charge, Total night minutes and Total night charge, Total intl minutes and Total intl charge are highly correlated

with each other have correlation coefficient value equals to 1 and their plot is linear. so, they are directly proportional to each other.

- Total customers who are not happy with the service and churned are 483 out of 3333 results in the churn rate of 14.49 % for the company.

7. Conclusion:

These are some conclusion or recommendations that we are arrived after studying the dataset.

- The company needs to survey in the Area code 510,408 who have higher churn rate than company average to understand the problem and similar kind of survey can also be done for the states having churn rate higher than the company average churn rate.
- Company needs to review their international plan because the customers subscribing to international plan have the highest churn rate observed of about 42.41%.
- Company needs to relook their vmail messages plan.
- The customer service of the company needs to be improved because it seems they are not able to solve the problem of users efficiently.

References:

- Alma Better
- Geeksfoorgeeks
- Stack overflow
- GitHub