Telecom Churn Analysis

Mahesh Y. Ladekar
Data Science Trainee,
AlmaBetter, Bangalore

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. In this paper, we explore and analyze the Orange Telecom's Churn dataset for identifying the key factors responsible for customer churn. The detailed exploratory data analysis is carried out and suggested the actions to be taken for avoiding these churns.

Keywords: Churn rate, number of calls, voice-mail messages, call charges.

1. Introduction

With the enormous increase in the number of customers using telephone services, the marketing division for a telecom company wants to attract more new customers and avoid contract termination from existing customers (churn rate). For the telecom company to expand its clientele, its growth rate (number of new customers) must exceed its churn rate (number of customers existing). Some of the factors that caused existing customers to leave their telecom companies are better price offers, faster internet services, and a more secure online experience from other companies.

A high churn rate will adversely affect a company's profits and impede growth. Our churn prediction would be able to provide clarity to the telecom company on how well it is retaining its existing customers and understand what are the underlying reasons that are causing existing customers to terminate their contract (high churn rate).

The telecom company can use our analysis to measure if it is providing a useful product compared with the product provided by its competitors. Since the cost of acquiring new customers is much higher than retaining its existing customers, the company can use the churn rate analysis to provide discounts, special offers, and superior products to keep current customers.

2. Problem Statement

Given, 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 or not.

To extract actionable insights from the dataset. I listed all the questions that came to mind below after assessing the dataset, and I tried to investigate all of them to find the insights:

- What is overall churn rate?
- What are the call charges in different area code at different time slot?

- What is the churn rate in various area codes? In which area code, maximum churn will be?
- What is churn rate according to states in the data?
- What is the effect of voice-mail message feature on churn?
- What is churn rate by charges at different time slots?
- Are the international call affects the churn rate?
- What is correlation between different features?

3. Features in Dataset:

STATE: 51 Unique States in United States of America

Account Length: Length of The Account

Area Code: 415 relates to San Francisco, 408 is of San Jose and 510 is City of Okland

International Plan: Yes - Indicate 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 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 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 Morning.

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

Dataset information obtained using python script as:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   State                   3333 non-null   object
 1   Account length          3333 non-null   int64
 2   Area code               3333 non-null   int64
 3   International plan      3333 non-null   object
 4   Voice mail plan         3333 non-null   object
 5   Number vmail messages   3333 non-null   int64
 6   Total day minutes       3333 non-null   float64
 7   Total day calls         3333 non-null   int64
 8   Total day charge        3333 non-null   float64
 9   Total eve minutes       3333 non-null   float64
 10  Total eve calls         3333 non-null   int64
 11  Total eve charge        3333 non-null   float64
 12  Total night minutes     3333 non-null   float64
 13  Total night calls       3333 non-null   int64
 14  Total night charge      3333 non-null   float64
 15  Total intl minutes      3333 non-null   float64
 16  Total intl calls        3333 non-null   int64
 17  Total intl charge       3333 non-null   float64
 18  Customer service calls  3333 non-null   int64
 19  Churn                   3333 non-null   bool
dtypes: bool(1), float64(8), int64(8), object(3)
memory usage: 498.1+ KB
```

4. Data Pre-processing:

Given dataset has 3333 entries and with 20 columns as features of dataset.

We observed that account length column doesn't make any sense in analyzing the churn rate. Hence, we remove this column.

The column “International plan” and “Voice-mail plan” has Dtype object with Yes/No string. We replace these string values with Yes=1, No=0 in original column.

We need this values in numerical form for further analysis in dataset to wrk with this feature if needed to develop any ML model for prediction.

We sort/divide the complete data in churn customer data frame and non-churn customer data frame to understand the details of features in dataset.

5. Exploratory Data Analysis:

5.1 Overall Churn Rate:

To check the overall churn rate, we plot the pie chart (shown in Figure 1) for getting churn rate true or false. We can see that our dataset is not balanced at all i.e., True is approximately 15% and False is approximately 85%. So, we analyze the data with other features while taking the target values separately to get some insights.

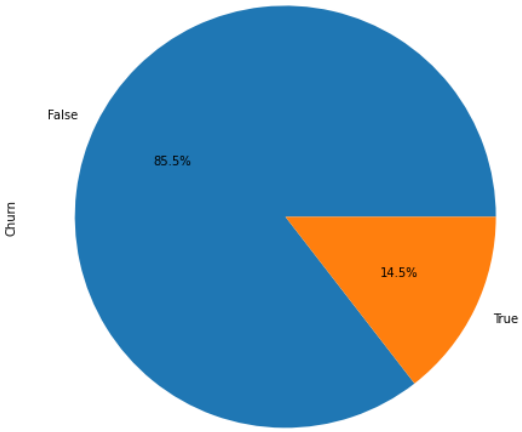

Figure 1: Pie chart showing overall churn rate in complete dataset

5.2 Churn Rate by Area Code:

To find the churn rate area-code-wise from given dataset is carried out in this part using bar plot. Initially, we plot the bar chart of showing call charges at different time slot (i.e. day, evening, night) in churn and not churn customer dataset. It is found that the call charges are more in area code 415 shown in Figure 2,3. Then we plotted the total churn in different area code using bar plot shown in Figure 4, which shows that the area code 415 has more churn rate. This is due to the more charges in area code 415 which means there are more calls or may be more customers. The calls frequency or customers are more but bandwidth is not enough to accommodate these calls. Hence this may lead to more churn in area code 415. It is needed to improve the network service as well as bandwidth of spectrum or optimize the bandwidth appropriately.

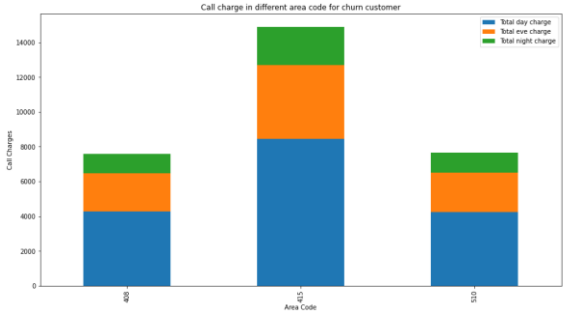

Figure 2: Call charges in churn customers

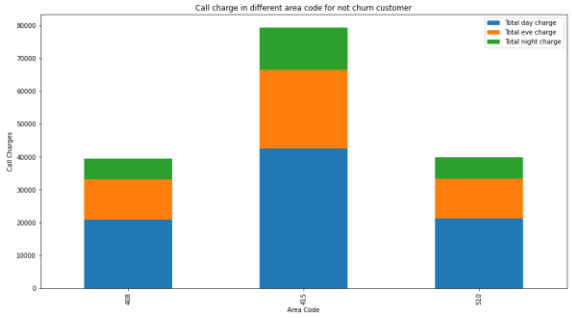

Figure 3: Call charges in not churn customers

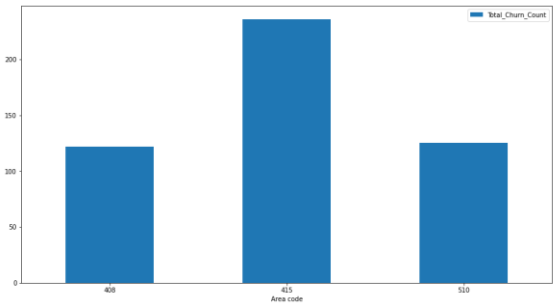

Figure 4: Total churn in different area code

5.3 Churn Rate by States:

This analysis is required to understand the churn rate by states and need to implement different strategy in these states for improving telecom networks to avoid customer churn. This can be shown in Figure 5 as follow:

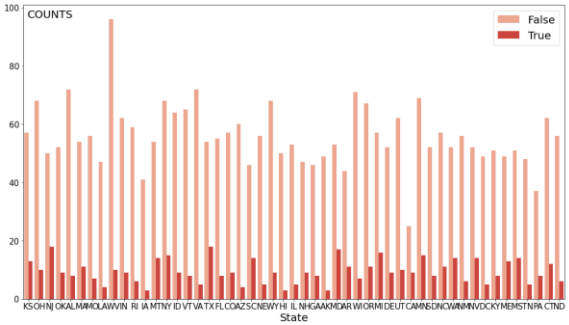

Figure 5: Churn rate by states

States with high churn rate: NJ, TX, MD. In these states network upgradation is required to avoid the churn while the states with low churn rate are IA, HI, AK. No need to upgrade in this area.

5.4 Checking Effect of Voice-Mail Message Feature Over Churn:

To analyze this feature, we utilized the box plot as shown in Figure 6. We have noticed for Voice-Mail Message Feature when there are more than 15 voice-mail messages then certainly there is a churn. To retain the customer, we suggest telecom provider to:

1] Setting up a limit on Voice-Mail service strictly no more than 15 voice mails.

2] Quality Drop in Voicemail after 20 voice mails may be one reason for customer churn.

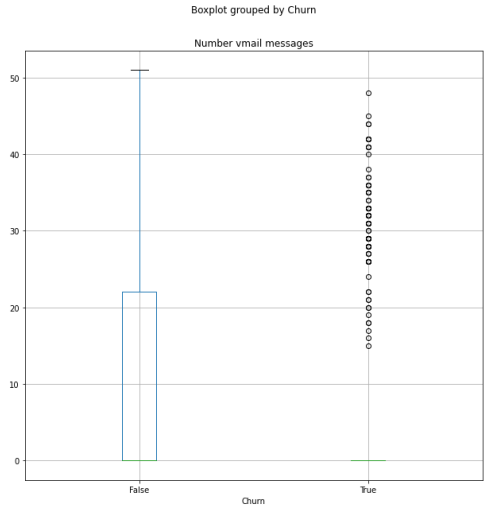

Figure 6: Box plot showing effects of voice mail message over churn

5.5 Churn Rate by Charges:

This analysis is carried out using kde plot using seaborn library.

A kernel density estimate (KDE) plot is a method for visualizing the distribution of observations in a dataset, analagous to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions.

For day charges, evening charges and night charges are analyzed for identifying the churn in which time slot is more.

If day charges are more than 40 means, there is always churn shown in Figure 7. While in evening and night shown in Figure (8,9), it is difficult to say that about churn based on

charges. Because the churn and not churn customer density is same in night and evening slot. Hence, it is needed to improve the network during evening and night-time slot.

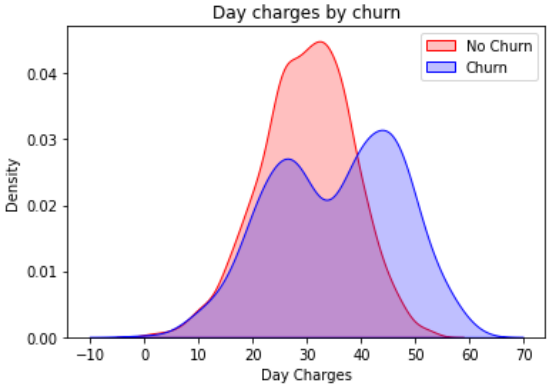

Figure 7: Churn density on day charges

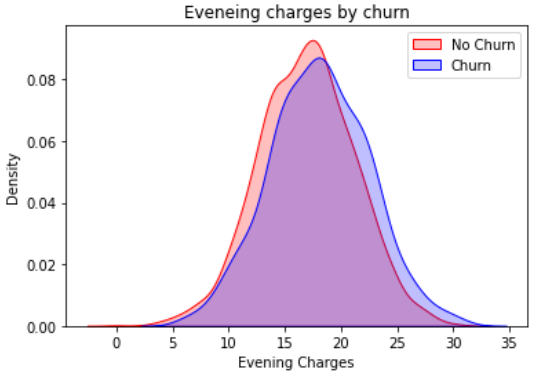

Figure 8: Churn density on eve charges

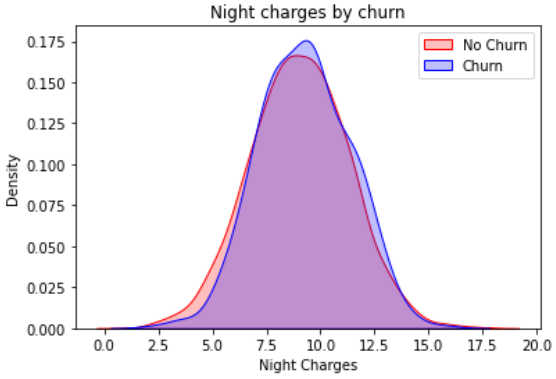

Figure 9: Churn density on night charges

5.6 Churn Rate Due to International Calls:

This feature of international calls is also contributed to churn of customer. It is shown by box plot showing the effect of international calls on customer churn in Figure 10. Users who make the International Call tend to spend more than 10 minutes on an average and box plot suggests that if call goes more than average time, there is churn.

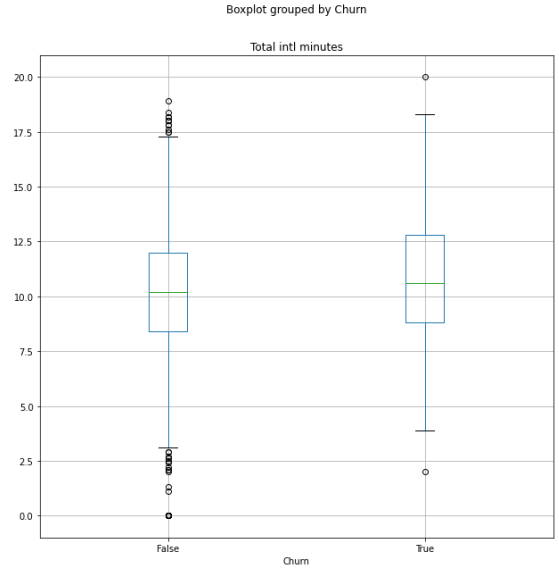

Figure 10: Box plot showing effect of international calls/minutes on churn

These calls clearly indicate that clients without International Plan Suffer and May Leave the Operator.

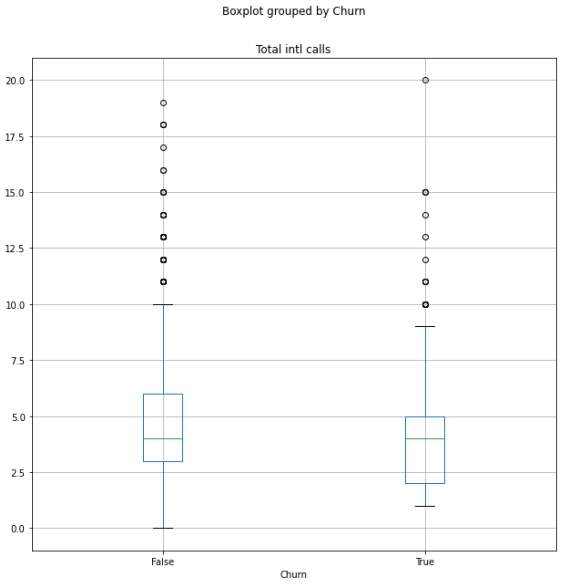

Figure 11: Box plot showing effect of international calls on churn

5.7 Data Correlation Heatmap

A correlation heatmap is a heatmap that shows a 2D correlation matrix between two discrete dimensions, using colored cells to represent data from usually a monochromatic scale. The color of the cell is proportional to the number of measurements that match the dimensional value. This makes correlation heatmaps ideal for data analysis since it makes patterns easily readable and highlights the differences and variation in the same data.

Figure 12 shows the data correlation heatmap of given dataset in which Green = Good (low correlation), Red = Bad (high correlation) between the independent variables.

Independent variables giving total minutes and total charges of day, evening, night and international calls are showing high correlation. It means they are exactly correlated with each other.

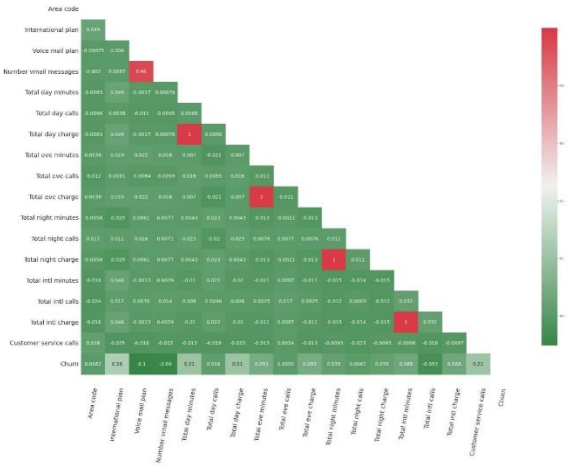

Figure 12: Data correlation Heatmap

6. Conclusion:

The detailed insights of Orange Telecom Churn dataset are given in this document. In this analysis, churn of customer is analyzed using columns/features like area code, states-wise, voice mail messages, international calls, and call charges. The area code 415 (San Francisco) has more churn as compared to the other area code. The states (NJ, TX, MD) are having high churn rate while states like IA, HI, AK have low churn rate. During voice mail message feature analysis, it is found that whenever there are more than 15 voice-mail messages then certainly there is a churn. International calls also affect the churn rate.

References:

1. https://towardsdatascience.com/
2. https://www.kaggle.com/
3. https://exploratory.io/
4. https://www.analyticsvidhya.com/
5. https://seaborn.pydata.org/