Exploratory Data Analysis on Telecom Churn Data

Sai Krishna Vamshi Devarasetty
Data Science Trainee
Alma Better

Abstract:

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

Exploratory Data Analysis helps us in understanding the data and in exploring some questions related to reason for the churning of users based on the data provided.

Keywords: Numpy, Pandas, EDA, Data Frames, Visualizations

1.Problem Statement

Data provided by Orange S.A., who provide telecommunication service. We want to figure out the reasons of why their customers are churning from their service.

- **State**: The user state from which the user is using telecom service.
- **Account Length**: Duration of how long the user has been active with their service.
- **International Plan**: Did user opted for International plan ? (Yes/No)
- **Voicemail Plan**: Did user has active Voice mail plan? (Yes/No).
- **Number vmail messages**: Number of voice mail messages sent by user.
- **Total day minutes, Total day calls, Total day charge**: These 3 variables indicate number of minutes, number of calls and the amount user has been charged during day time.
- **Total eve minutes, Total eve calls, Total eve charge**: These 3 variables indicate number of minutes, number of calls and the amount user has been charged during evening time.
- **Total night minutes, Total night calls, Total night charge**: These 3 variables indicate number of minutes, number of calls and the amount user has been charged during night time.
- **Total Intl minutes, Total Intl calls, Total Intl charge**: These 3 variables indicate number of minutes, number of calls and the amount user has been

charges for international service usage

- **Customer service calls**: Number of customer service calls user has made.
- **Churn**: Did user churned or not? (True/False)

Questions we want to answer are:

- Do churned users using service more during any specific duration (Day, Evening, & night) compared to Non Churned ?
- Does users using service at particular duration has any impact on their usage of other durations ?
- Do area have any significant impact on churned users ?
- Does State variable play any important role in for churning ?
- Did opting for international plan resulted in more churning ?
- Did opting for voice mail plan resulted in more churning ?
- Do customers making more service calls churning ?

2. Introduction

The data set provided is a cleaned customer dataset, it has no missing values. Except State, International plan, Voice mail plan, and Churn all other columns in dataset are numerical (either int64 or float64 type). Even though Area Code is numerical, it is not ordinal. State is the only variable which is string. International plan and Voice mail plan are also given as type string (Yes/No) but we need to considered them as Boolean for analysis. Churn is a Boolean variable with value True being users have churned and False implies who didn't churn. Out of 3333 users 483 are churned and 2850 didn't churned.

3. Exploratory Data Analysis

3.1 Churned vs Non Churned

To compare this we created two data frames ChurnedDF & NonChurnedDF , then plotted histogram plot

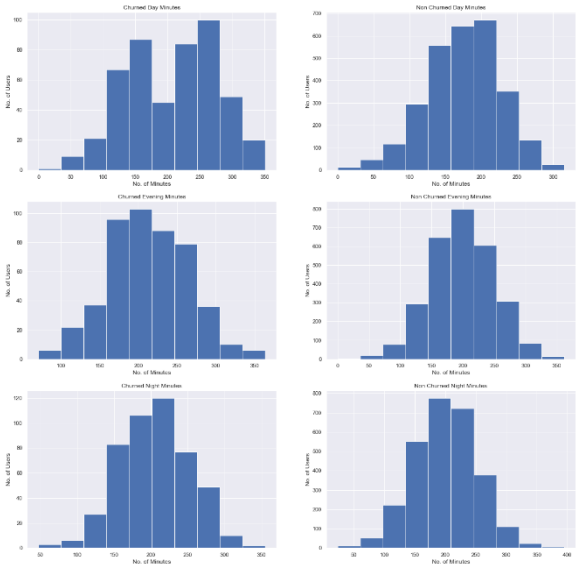

Figure 3.1a

The figure 3.1a clearly indicates that 60% of churned users are above avg. number of minutes.

3.2 Correlation among duration of days

Now we calculated correlation value among duration of days.

	Total day minutes	Total eve minutes	Total night minutes
Total day minutes	1.000000	0.007043	0.004323
Total eve minutes	0.007043	1.000000	-0.012584
Total night minutes	0.004323	-0.012584	1.000000

Figure 3.2a

	Total day calls	Total eve calls	Total night calls
Total day calls	1.000000	0.006462	-0.019557
Total eve calls	0.006462	1.000000	0.007710
Total night calls	-0.019557	0.007710	1.000000

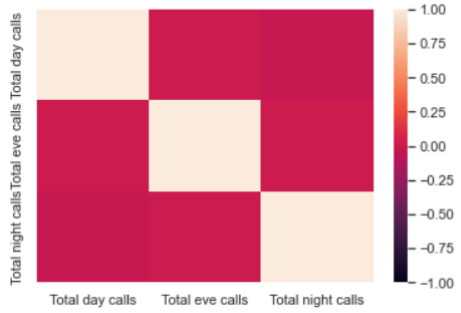

Figure 3.2b

	Total day charge	Total eve charge	Total night charge
Total day charge	1.000000	0.007036	0.004301
Total eve charge	0.007036	1.000000	-0.012601
Total night charge	0.004301	-0.012601	1.000000

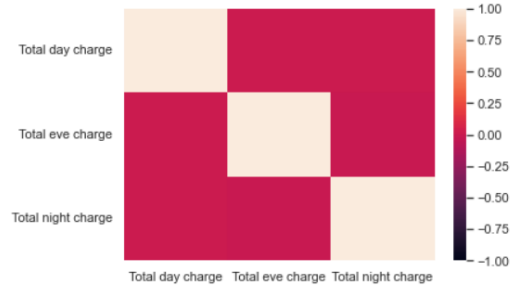

Figure 3.2c

So here when we calculated correlations among day, evening and night for minutes, calls and charges, we found out the correlation is almost 0, which clearly indicates service of one duration has no impact on other duration.

3.3 Area Code based Analysis

Here we tried to do some analysis based an area code and want to see if it played any role in churning. We figured out there are 3 area codes 415, 408, & 510.

```
AreaCodeWiseChurnAndNonChurn = {
    'Churn': {
        415: DataFrame of 415 Area Code with Churn equal to True,
        408: DataFrame of 408 Area Code with Churn equal to True,
        510: DataFrame of 510 Area Code with Churn equal to True
    },
    'NonChurn': {
        415: DataFrame of 415 Area Code with Churn equal to False,
        408: DataFrame of 408 Area Code with Churn equal to False,
        510: DataFrame of 510 Area Code with Churn equal to False
    }
}
```

Figure 3.3a

For analysis I created a nested dictionary of data frames as described in figure 3.3a. Further we plotted boxplot to see distribution for all 3 area codes as shown in figures 3.3b(area code 415) , 3.3c(area code 408) & 3.3d(area code 510). In all images left 3 boxplots are for Churned users and the right 3 are for Non Churned users within that area code.

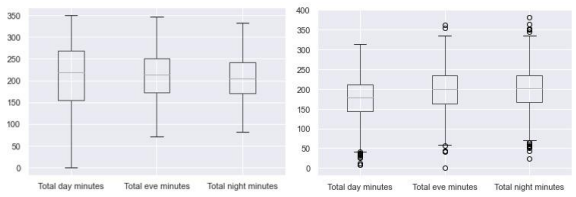

Figure 3.3b

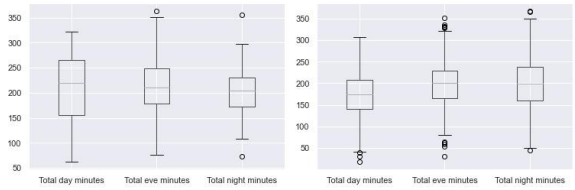

Figure 3.3c

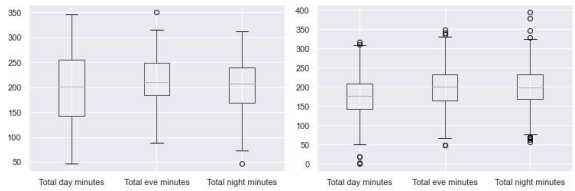

Figure 3.3d

In all 3 are codes, the median (50th percentile) of day and evening minutes is more in Churned users compared to Non Churned and for night it's comparable (Churned is slightly high). And also if we see IQR (Inter Quartile Range) is spread more for Churned users.

3.4 State Based Analysis

Here we did Analysis based on states users belong to. There were 51 states in total in given dataset.

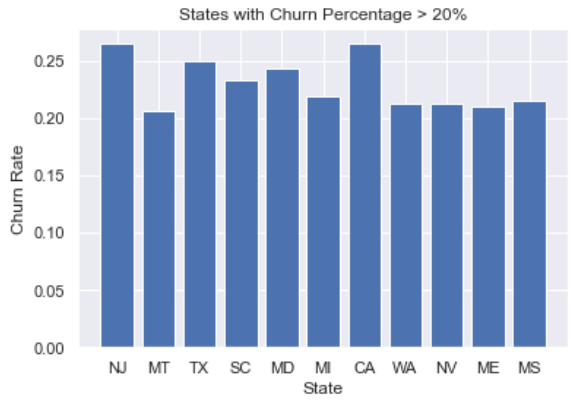

Figure 3.4a

In the above plot we can see all 11 states having churn rate more than 20%. Among them states NJ and CA have highest Churn rate 26.47%.

3.5 International plan based Analysis

Here we are going to perform analysis to figure out if opting International Plan has any impact on Churn rate.

```
International plan  Churn
No                  False    2664
                    True      346
Yes                 False     186
                    True      137
Name: Churn, dtype: int64
```

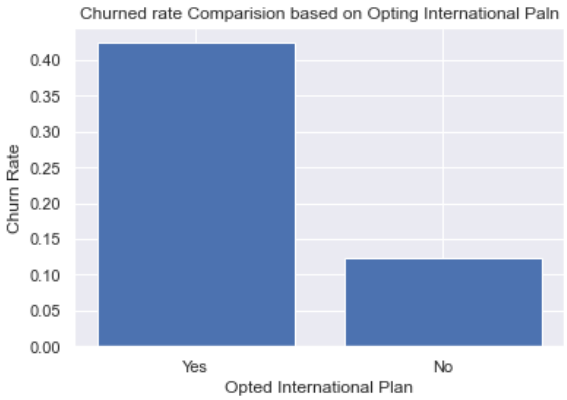

Figure 3.5a

It is clearly evident that users who have opted for international plan has Churned at 40% rate, it indicates their international service is not at all good.

3.6 Voice mail plan based analysis

Here we are going to do same analysis as International Plan to see if opting voice mail plan has shown any impact

```
Voice mail plan  Churn
No               False    2008
                 True      403
Yes              False     842
                 True       80
Name: Churn, dtype: int64
```

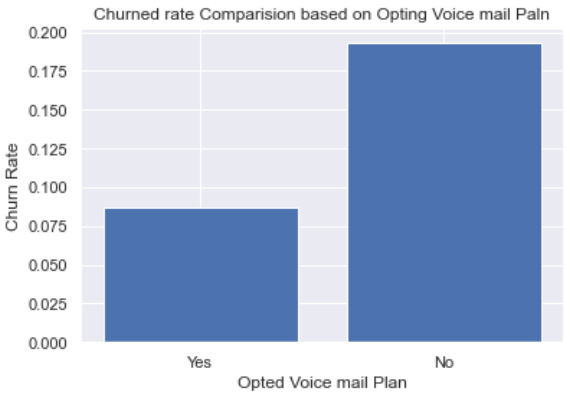

Figure 3.6a

To out surprise people who have opted for voice mail plan are more likely to continue with their telecom service.

3.7 Analysis based on No. of Customer Service Calls

Finally we want to perform an analysis to see if users who called customer service have churned.

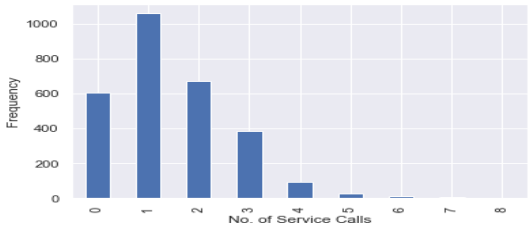

Figure 3.7a

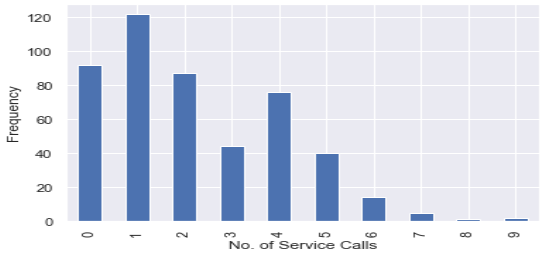

Figure 3.7b

Figure 3.7a, 3.7b indicates the frequency plot of Number of service calls made by Non Churned and Churned users respectively. It clearly shows users who have made more than or equal to 4 customer service calls are likely to Churn

4. Conclusion:

That's it! We reached the end of our analysis. The following are our observations:

- User with International Plan tend to churn more frequently.
- 11 states out of 51 states in the given dataset has more than 20% churn rate.
- Users with voicemail plan have very low churn rate when compared to counter group.
- There's almost no(near to zero) correlation in the service usage among day, evening and night time.
- Users with 4 or more customer service calls Churned more.

References-

1. Numpy, Pandas, Matplotlib & seaborn documentation.
2. Alma Better recorded classes
3. Articles on Towards Data Science.