# TELECOM CHURN ANALYSIS

Aishani Barman Roy, Nakul Pradeep,

Kavya Sharma, Baliram, Shubham

DATA SCIENCE TRAINEES

ALMABETTER BANGALORE

## ABSTRACT

Nowadays, the telecom industry faces fierce competition in terms of customer satisfaction. With the introduction of newer technology, telecom companies' services have expanded from just calls to providing data and web services. The churn rate, which is the rate of attrition or customer churn, can be explained as the rate at which customers stop doing business with an entity. In telecom business, churn occurs when a subscriber leaves the service. According to a Harvard Business Review article (Gallo, 2014), the cost of acquiring a customer is five to twenty-five times higher than the cost of retaining an existing one. Furthermore, increasing retention by 5% can result in an increase in profits of 25% to 95%. With newer companies entering the market, customers will have more freedom to switch telecom providers, thus it is becoming increasingly important to focus on retaining existing customers. So, to thrive in market, telecom companies must innovate, provide better services and grow their customer base.

## INTRODUCTION

Customer churn is one of any industry's major concerns, so the topic of churning is addressed in this study. In this project work, we have tried to analyze what features are driving the churn rates mostly and how each of these features relate to each other and to get a basic understanding of the data which will help us and our business entity. We know from previous studies that the cost of getting a new customer is significantly higher than the cost of keeping an existing one. Churn rate can be defined by the percentage of customers who cease subscribing to a service or the percentage of employees who leave a position. Banking, insurance, internet streaming, and telecommunications, to name a few, have all been affected by churn. Although there are numerous reasons for customer churn, service dissatisfaction, overpriced subscriptions, and better alternatives are among the most common.

## PROBLEM STATEMENT

Orange S.A., formerly France Telecom S.A., is a French multinational telecommunications corporation. The Orange Telecom & Churn Dataset consists of cleaned customer activity data (Features), along with a churn label specifying whether a customer cancelled the subscription. We want to explore and analyze the data to discover key factors responsible for customers churn and come up with ways or recommendations to ensure customer retention. If the churn rate of customers of a company goes up to 50% then the company will be shut in two years.

## DATA DESCRIPTION

Based upon the initial assessment, we found no Null or Duplicate value in our dataset and with the help of info() and duplicated() method respectively, we draw the following key insight about our dataset : -

- The dataset has a shape of (3333, 20) which means that it contains 3333 rows and 20 columns.
- Our dataset has 8 columns with float d-type, 8 columns with integer d-type and 3 columns with object d-type and 1 column which is bool type. we found that the data is clean with no missing values.

We have the following column provided to us in the dataset:

- Churn: A binary identifier whether the customer has churned or not, which is the dependent variable.

And other 19 columns which are possibly driving the outcome of churn:

- State: State in which the customer lives in.
- Account length: It is the number of days the account has been active.
- Area code: An identifier to the area the customer lives in.
- International plan: A binary Identifier to whether the customer has opted for an international plan.
- Voice mail plan: A binary identifier to whether the customer has opted for a voicemail plan.
- Number Voicemail messages: Number of voicemail messages sent or received.
- Total day minutes: Total minutes the customer has talked over phone in the daytime.

- Total day calls: Total calls the customer has made over phone during daytime.
- Total day charge: Total money that was charged to the customer in the daytime.
- Total eve. minutes: Total minutes the customer has talked over phone in the evening.
- Total eve. calls: Total calls the customer has made over phone in the evening.
- Total eve. charge: Total money that was charged to the customer in the evening.
- Total night minutes: Total minutes the customer has talked over phone in the night.
- Total night calls: Total calls the customer has made over phone during nighttime.
- Total night charge: Total money that was charged to the customer for nighttime usage of services.
- Total international minutes: Total minutes the customer has talked over phone internationally.
- Total international calls: Total calls the customer has made over the phone internationally.
- Total international charge: Total money that was charged to the customer for international calls.
- Customer service calls: Total customer service calls that were made to the customer.

## WHY IS ANALYSIS REQUIRED?

- ***High competition***: Telecom industry is expanding day by day and there are lots of service provider in a market, so every company wants to improve their services so that they can make good profit.
- ***Customer Satisfaction:*** In any industry the focus of management is to satisfy customer by providing good services. This can be done by improving services, good customer care support, sufficient amount of service centers.
- ***Identifying Weaker section:*** Every company wants to know their weaker area so that they can work on those area and can improve themselves. Data analysis play a vital role in the growth of industry.

## STEPS INVOLVED

- DATA CLEANING

  It is very important to get rid of the irregularities and clean the data after sourcing it into our system. Irregularities are of different types of data.
  - Missing Values

    - Incorrect Format
    - Incorrect Headers
    - Anomalies/Outliers
- UNIVARIATE ANALYSIS

  This is the type of analysis where we try to find insight by doing analysis of single feature column individually with the help of Charts like bar plots, distribution plot, box plot, pie charts, etc. and try to get a picture of each feature in a dataset.
- MULTIVARIATE ANALYSIS

  In this analysis we will try to get insight with the help of finding relationship between two features or columns with the help of scatter plot, bar plot, box plot, etc. We will mostly try to find relationship between churn and other variables. This is a crucial step to get a picture of which features drive the churn mostly.

## PLOT USED FOR VISUALIZATION

1. **Bar Graph:** Bar Graphs are the pictorial representation of data, in the form of vertical or horizontal rectangular bars, where the length of bars is proportional to the measure of data. They are generally used to visualize categorical data. We have used the grouped bar graphs or clustered bar graphs. It is used to represent values for more than one object that shares the same category.
2. **Histogram and Distribution Plots:** A histogram is a bar plot where the axis representing the data variable is divided into a set of discrete bins and the count of observations falling within each bin is shown using the height of the corresponding bar. Histograms aim to approximate the underlying probability density function that generated the data by binning and counting observations. The histogram is represented by a set of rectangles, adjacent to each other, where each bar represents a kind of data.
3. **Box Plot:** A box and whisker plot—also known as box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. It shows data spread, data distribution, central value of our variable with the help of median and help us in outlier detection. The values are of upper and lower bound whiskers can be treated as outliers here.
4. **Pie Chart:** A Pie chart is a type of graph that displays data in a circular graph. The pieces of the graph are proportional to the fraction of the whole in each category. In other words, each slice of the pie is relative to the size of that category in the group. Pie

charts are often used to represent sample data with data points belonging to a combination of different categories.

5. **Correlation Heat map:** A correlation heat map is a type of visualization tool 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 correlation between the variables that cross through row and column at that cell. It is assisted by a color bar making data easily readable and comprehensible.
6. **Violin Plot:** Violin Plot is a method to visualize the distribution of numerical data of different variables. It is like Box Plot but with a rotated plot on each side, giving more information about the density estimate on the y-axis. The advantage of a violin plot is that it can show nuances in the distribution that aren't perceptible in a boxplot. On the other hand, the boxplot more clearly shows the outliers in the data. Violin Plots hold more information than the box plots, they are less popular as their meaning can be harder to grasp for many readers not familiar with the violin plot representation.

## OBSERVATIONS

1. There are no null or duplicate values in our dataset.
2. The total charges and total minutes are perfectly correlated for day, evening and night usages i.e., there is a perfect linear relationship between charges and minutes.
3. Overall Customer Churn rate is 14.5%.
4. The Churn rate is higher for New Jersey, California, Texas and South Carolina while it is lower for Hawaii(HI), Alaska(AK), Arizona(AZ),(Virginia)VA and (Iowa)IA.
5. Regarding stacked bar chart, we could observe that Area code also has no effect on the Churn rate.
6. We have noticed that customers who have used International plan are churned more than those who haven't used International plan.
7. Regarding Voicemail plan, Customers who haven't subscribe a voice mail plan are more churned.
8. The customers who retained service did about 0-20 voicemail messages. The customers who are churned were the ones who did not do any voicemail or the ones who made voicemail messages more than 20.
9. With help of box plot we could observe that users spending more minutes of call during daytime are churned more.

10. Across total day calls, mostly 100 calls are made by both churned and non-churned customers. Churned customers total day call median is slightly more than that of non-churned customers.
11. The total evening minutes of churned or non-churned customers are approximately same.
12. The total evening call distribution is not a factor for churning.
13. The total night minutes and night calls distribution of Churned and non-churned is almost similar.
14. The total international minutes spent by customer across churned population is slightly more than that of non-churned population.
15. The distribution of total international calls of non-churned is greater than that of churned. However, the median calls made by churned customers are greater than non-churned customers.
16. We can see from the violin plot that the people who retained the service is when the customer service call is 1 for increase in calls the people retained reduces. While in case of churned we can see that it increases to maximum at 1 call then decreases and further increases when customer service calls are 4 then the customer churned decreases.

## CONCLUSIONS

1. Company should introduce a longer validity prepaid plan with discounted prices so customers using more data get comparatively cheaper plan and at the same time retained, thus reducing churn rate.
2. Company should focus on their quality of services as users start to churn with increasing number of customer service calls.
3. Some states where churn rate is high like New Jersey, California, Texas, South Carolina should receive more focus and the company should implement new strategy here.
4. Company should offer attractive benefits to gain new users and retain them
5. Company should offer better benefits to customers with international plans to retain them.
6. With voicemail plan there is a (16.7%) churn rate compared to people with no voice mail plan (8.67%). Introducing voicemail offers for customers with no voicemail plan can reduce churn rate.
7. Effects of different factors are to be tested statistically to draw a more rigid conclusion.
8. The total night call distribution is like total evening call distribution, also the night minutes and evening minutes follow a similar pattern (which can help us reduce feature

by making a feature including evening and nighttime values). Implementing this can reduce model complexity and can help reduce curse of dimensionality on applying dataset to a machine learning technique.

REFERENCES:


- Alma Better Recorded Classes
- Stack overflow
- GeeksforGeeks
- Analytics Vidhya