# Telecom Churn Analysis

**Ruchika Nayak,**
**Data science trainee,**
**AlmaBetter, Bangalore**

## Abstract:

The cab platforms adjust their prices using a specific algorithm which is real time and dynamic known as “Surge Pricing” or “Dynamic Pricing”. We were provided with three such already classified labels in our data set.

Our experiment can help understand what could be the reason for the classification of such labels by feature selection, data analysis and prediction with machine learning algorithms taking into account previous trends to determine the correct classification.

***Keywords:machine learning,surge pricing,dynamic pricing,classified labels***

## 1.Problem Statement

The main objective is to explore and analyze the data to discover key factors responsible for customer churn and come up with ways/recommendations to ensure customer retention.

## 2. Introduction

The cab platforms adjust their prices using a specific algorithm which is real time and dynamic known as **“Surge Pricing”** or **“Dynamic Pricing”**. This algorithm automatically raises the price of a trip when the demand increases more than the supply.

The surge algorithm generally outputs a multiplier which is adjusted along with the base fare, the price per mile and the price per minute to generate the final price. This price is communicated to the riders and the ride is initiated when they confirm the price shown. This surge multiplier is kept discrete and may range from 1.2 to the maximum allowed by the government based on geography.

Our goal here is to build a predictive model, which could help Sigma Cabs in predicting the surge pricing type proactively.

## 3. Customer Churn

Customer churn means shifting from one service provider to its competitor in the market. Customer churn is one of the biggest fears of any industry, particularly for the telecom industry. With an increase in the number of telecom service providers in South Asia, the level of competition is quite high. Although there are many reasons for customer churn, some of the major reasons are service dissatisfaction, costly subscription, and better alternatives. The telecom service providers strive very hard to sustain in this competition. So to sustain this competition they often try to retain their customers than acquiring new ones as it proved to be much costlier. Hence predicting churn in the telecom industry is very important. To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

## 4. Reasons for customer churn

The reasons for customer churn are:

- Service quality
- Content and feature availability
- Lower cost substitutes from competitors
- Negative customer service

## 5. Data description

The telecom churn dataset contains 3333 rows and 20 columns out of which 16 are numerical columns.

- ➢ **Numerical columns:** Account length, Area code, Number vmail messages, Total day minutes, Total day calls, Total day charge, Total eve minutes, Total eve calls, Total eve charge, Total night minutes, Total night calls, Total night charge, Total intl minutes, Total intl calls, Total intl charge, and Customer service calls

- ➢ **Categorical columns:** State, International plan, Voice mail plan, Churn

## 6. Steps involved:

- **Data exploration**

  After reading the dataset which is in csv format we inspected the rows and data type of each columns. We have also explored number of unique values of each column and viewed some basic statistical details like percentile, mean, standard deviation etc. of numerical columns.

- **Null values and Duplicates**

  Our dataset doesn’t contain any null values and duplicate rows.

- **Univariate analysis**

  Our main motive through this step is to visualize each column to get some insights. For this we have used various visualization tools like:

  - ➢ Bar plot
  - ➢ Histogram
  - ➢ Distribution plot
  - ➢ Pie plot

- **Analyzing the effect of numerical features on churn**

  Our main motive through this step is to analyze which numerical feature affects churn the most. For this we have used kernel density estimate (KDE) plot.

- **Analyzing the effect of categorical features on churn**

  Our main motive through this step is to analyze which categorical feature affects churn the most. For this we have used bar plot and cat plot.

- **Exploring correlation between variables**

  With the help of heatmap we have explored the variables which are correlated with each other.

- **Outlier detection**

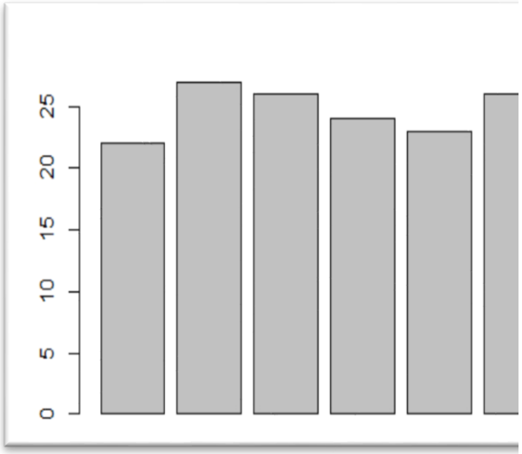
  With the help of boxplot, we have detected the presence of outlier in different numerical features

## 7. Graph Plots Used

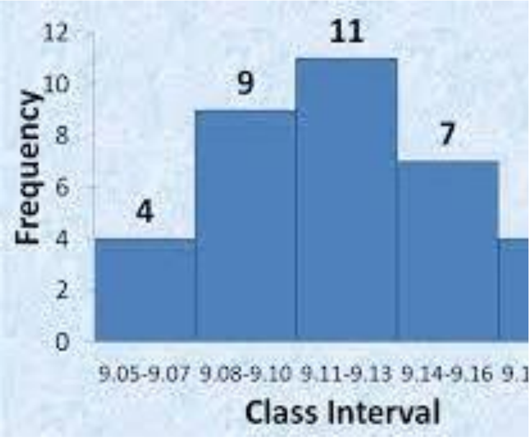
For our data visualization and analyzation we have used the following the following graphical presentation.

1. **Bar plot:** It is a graphical presentation that shows the category of data with rectangular bars with lengths and height that is proportional to the values which they represent. It describes the

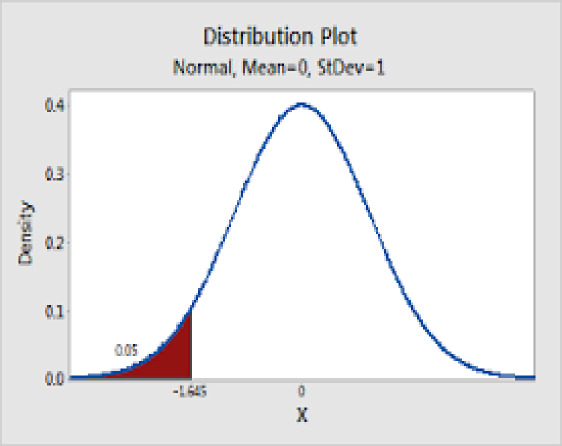
comparisons between the discrete categories.



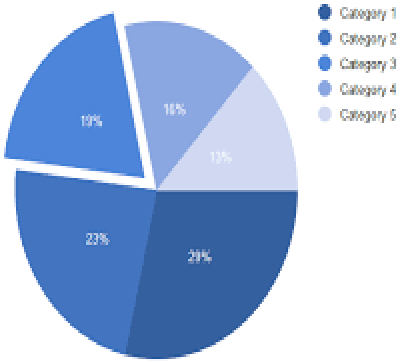
2. **Histogram:** Histogram is a graphical representation that organizes a group of data points into user-specified ranges. Its appearance is similar to bar graph. The hist plot condenses a data series into an easily interpreted visual by taking many data points and grouping them into bins.



3. **Distribution Plot:** A distribution plot or distplot shows the variation in the data distribution. It visually asses the distribution of sample data by comparing the empirical distribution of the data with the theoretical values.

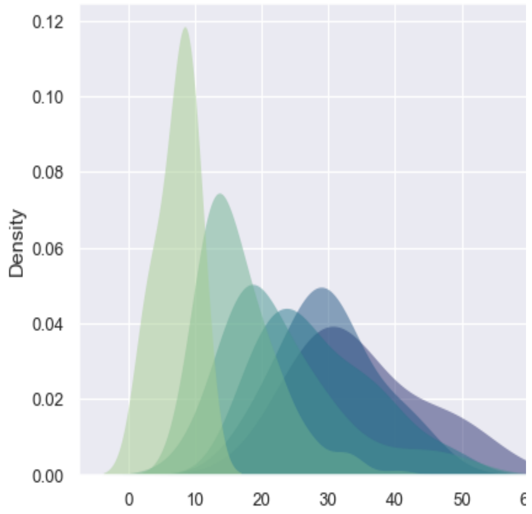


4. **Pie Plot:** A pie plot or pie chart is circular statistical graphic, which is divided into slices and each slice represents the count or percentage of the observations of a level for the variable.

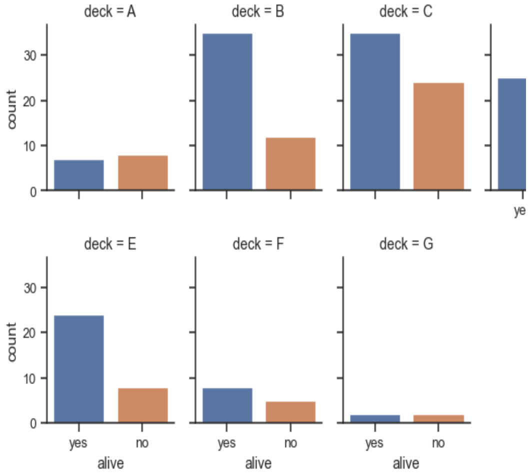


5. **Kernel Density Estimate Plot:** A kernel distribution estimate or KDE plot is a method for visualizing the distributions of observations in a dataset. Relative to histogram, KDE can produce a plot that is less

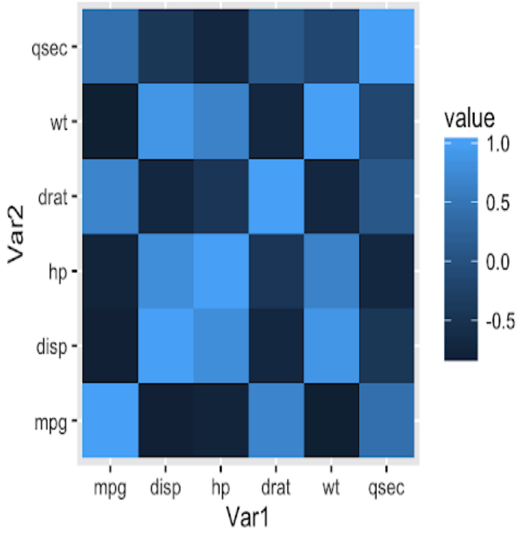
cluttered and more interpretable .



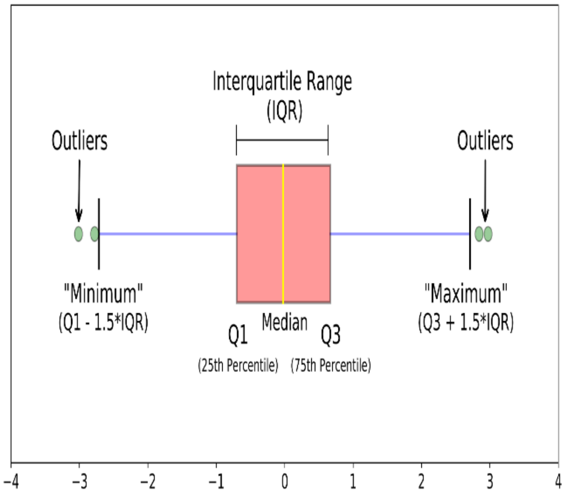
6. **Cat Plot:** It shows the frequencies of the categories of the categorical variables. This function helps us to show the relationship between a numerical and one or more categorical variables.



7. **Correlation Heatmap:** It’s a heatmap that shows a 2D correlation matrix between two discrete dimensions, using colored cells to represent data from usually a monochromatic scale. It demonstrate the linear relationship of variables between each other where rows and columns shows the variables.



8. **Box Plot:** A box and whisker plot also called 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. We have used boxplot mainly to show the outliers.



**8. Conclusion:** That's it! We reached the end of our exercise.
Starting with loading the data , checking the null values , exploring the whole dataset, visualizing it with different types of plots and graphs and from those visualization analyzing the interpretations and finally we concluded that 14% of customers of the Telecom Company have churned. We have also discovered the key factors for customer churn and suggested some ways to ensure customer retention.

**References-**

1. Python
2. GeeksforGeeks
3. Analytics Vidhya