***Project Name - Churn reduction***

***Debarun Banerjee***

***11 March,2019***

Contents

[Chapter 1 3](#_Toc3171178)

[1.1 Introduction 3](#_Toc3171179)

[1.2 Data 3](#_Toc3171180)

[Chapter 2: Methodology 5](#_Toc3171181)

[1. Data Manipulation: 5](#_Toc3171182)

[2. Exploratory Data Analysis 6](#_Toc3171183)

[2.1Scatter Plot: 6](#_Toc3171184)

[2.2 Pie Chart: 8](#_Toc3171185)

[2.3 Histogram: 11](#_Toc3171186)

[2.4 Bar-chart: 14](#_Toc3171187)

[2. Data Processing: 16](#_Toc3171188)

[3. Variable Summary: 16](#_Toc3171189)

[4. Correlation Matrix: 17](#_Toc3171190)

[5. Visualizing data with Principal Components: 18](#_Toc3171191)

[6. Model Building: 19](#_Toc3171192)

[6.1 Logistic Regression Model: 19](#_Toc3171193)

[6.2 Univariate Selection 21](#_Toc3171194)

[6.3 Decision Tree: 22](#_Toc3171195)

[Chapter 3: Conclution: 23](#_Toc3171196)

[Model Selection: 23](#_Toc3171197)

[Appendix: 24](#_Toc3171198)

[Complete R-code: 24](#_Toc3171199)

[Reference: 31](#_Toc3171200)

**Project Description -** Churn (loss of customers to competition) is a problem for companies because it is more expensive to acquire a new customer than to keep your existing one from leaving. This problem statement is targeted at enabling churn reduction using analytics concepts.

# Chapter 1

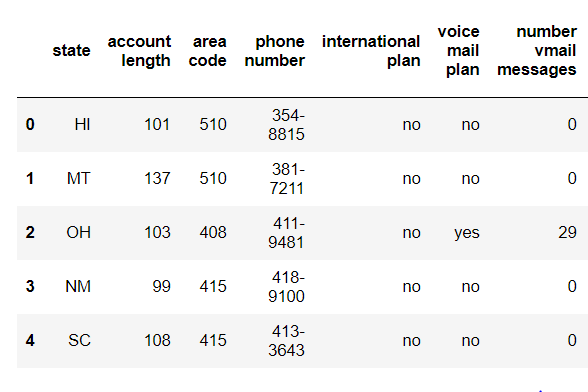
## 1.1 Introduction

Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers. Cell phone service companies, Internet service providers, pay TV companies, insurance firms, often use customer attrition analysis and customer attrition rates as one of their key business metrics, because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients

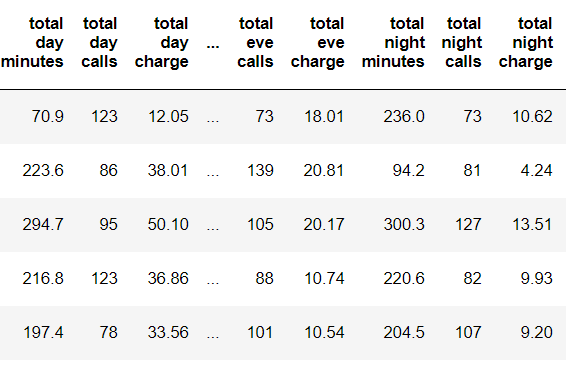
## 1.2 Data

Our task is to build models which will predict the Customer Churn based on given data. Predictive analytics use churn prediction models that predict customer churn by assessing their propensity of risk to churn. Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn. Given below is a sample of the data set that we are using to predict the churn:

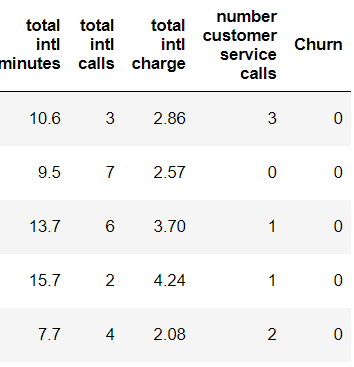
**Table 1.1: Customer Churn Test Data (Columns: 1-7)**



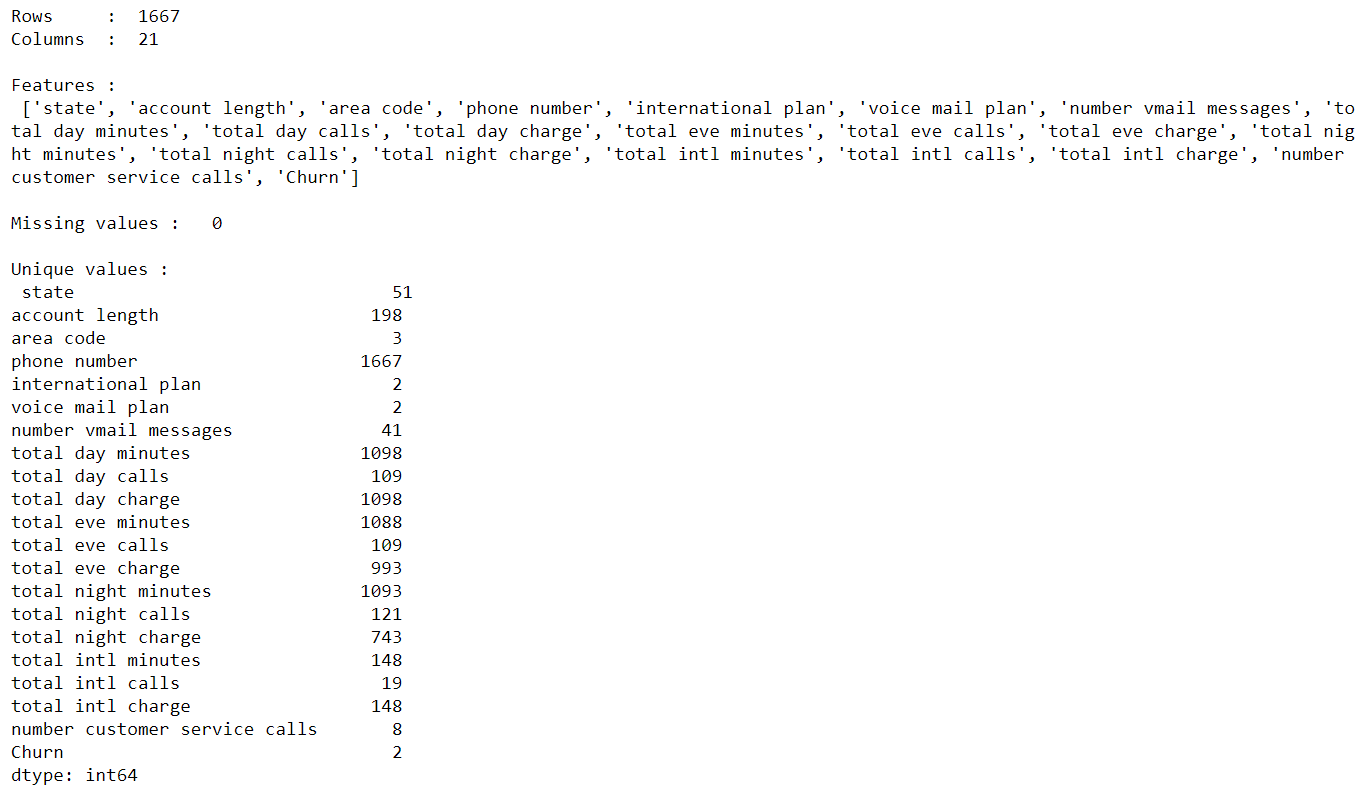
**Table 1.2: Customer Churn Test Data (Columns: 8- 15)**



**Table 1.3: Customer Churn Test Data (Columns: 16- 21)**



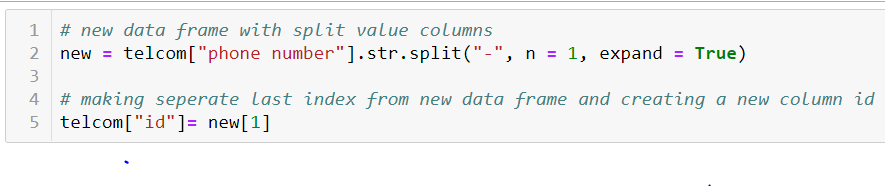
As you can see in the table below we have the summary of our data, where we have 20 features using which we must correctly predict if the customer is going to churn or not:



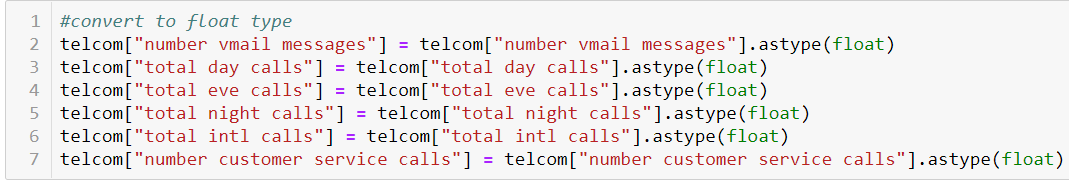
# Chapter 2: Methodology

## 1. Data Manipulation:

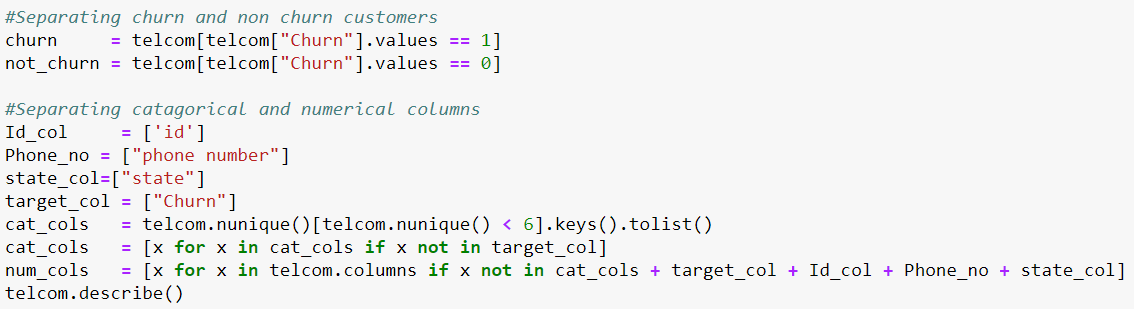
* We have observed that there are 21 column and 1667 rows in our dataset.
* We consider the column “Phone number” as unique column but this is a string. So, we have split the column using separation delimiter operator and created one new column “id” which contains the last part of the unique number as id.



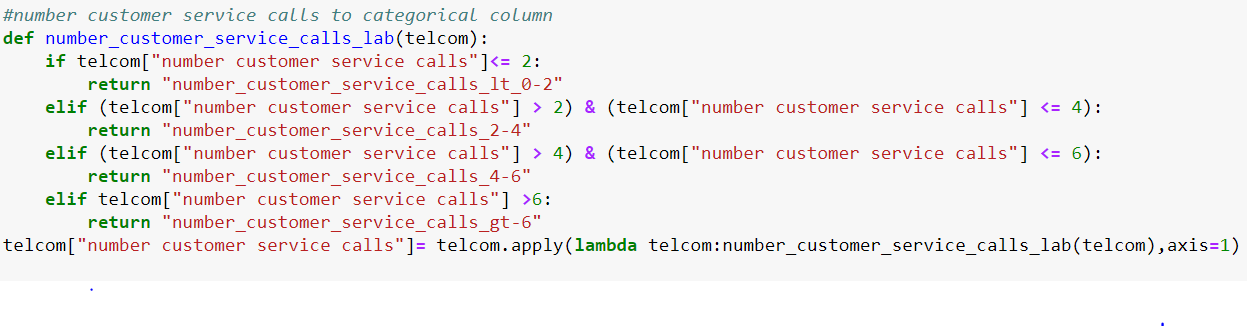
* We don’t have any missing values in our data set.
* We have converted the datatypes as required in our dataset.



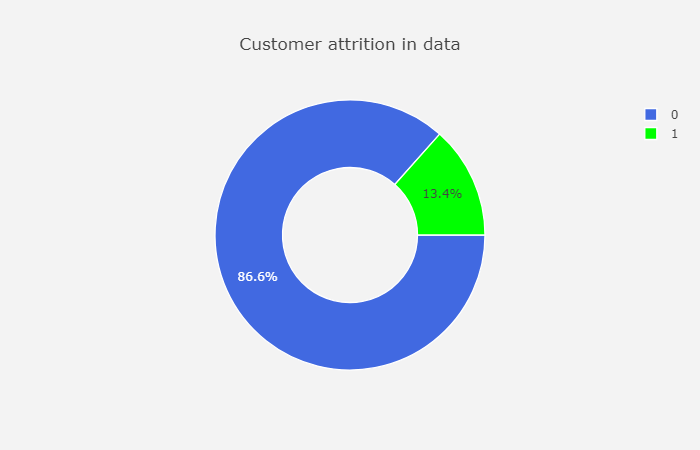
* Categorize churn column to 0 or 1 from Boolean value.
* Split Churn and not churn and connect to each feature and visualize and separating categorical and numerical columns.



* Converted some of the numerical columns (number vmail messages, total day calls, total eve calls, total night calls, total intl calls, total eve calls) to categorical column using the concept of Bins. We have used “lambda” function which is a python inline function to apply the definition to all the values and create Bins accordingly.



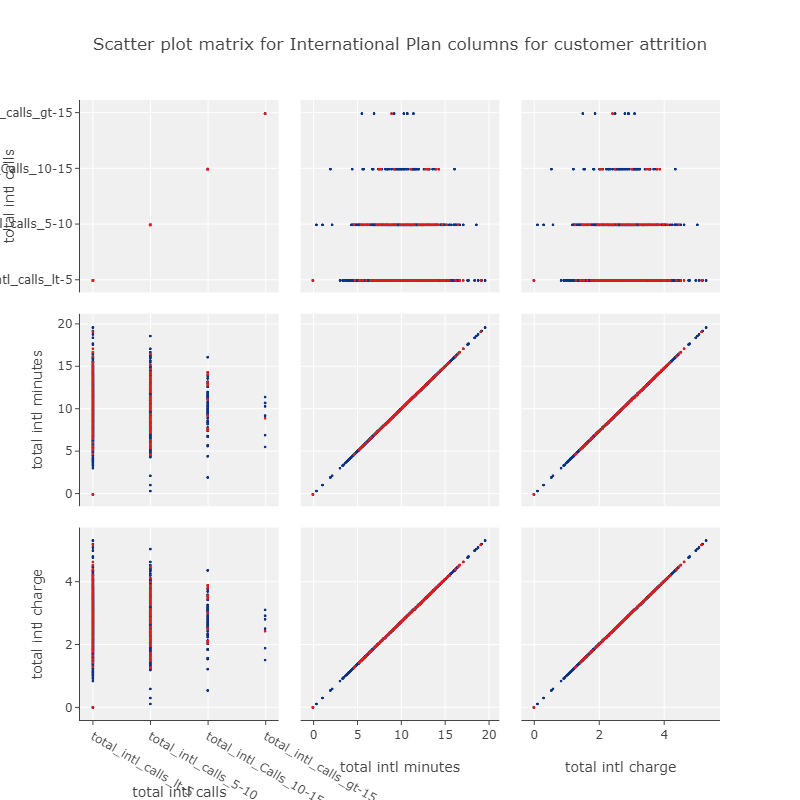
2. Exploratory Data AnalysisExploratory data analysis is an approach to analysing data sets to summarize their main characteristics using visual methods. As a part of Exploratory Data Analysis, we have created Pie chart for our Categorical column, scatter plot to visualize the numerical columns. Given below is the Pie Chart of our target variable where we can see 13.4% of our customer are going to churn.



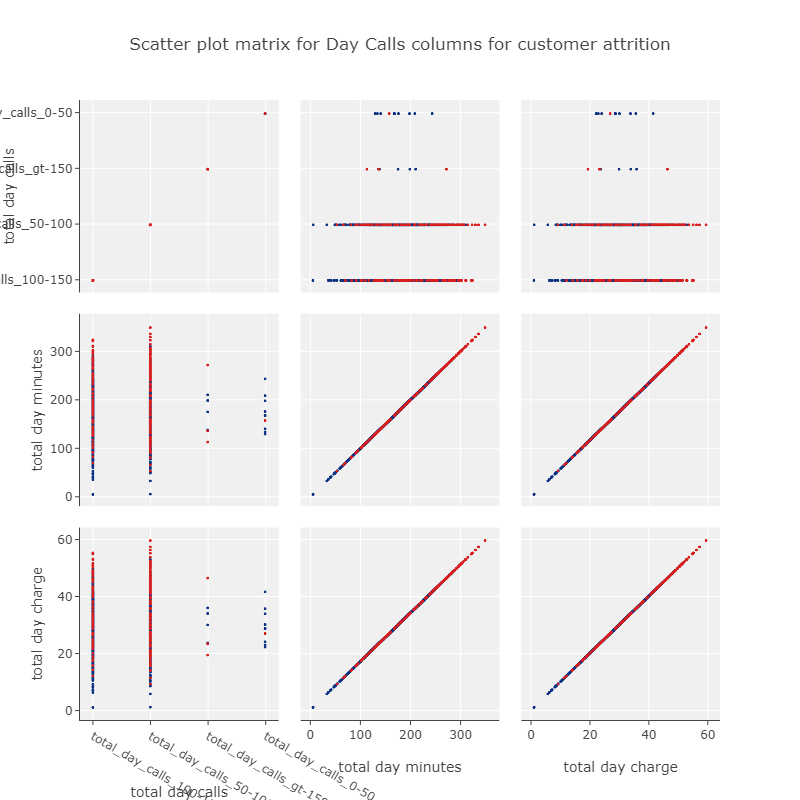
### 2.1Scatter Plot:

The *scatter plot* displays values of two numerical variables as *Cartesian coordinates* in 2D space. In our dataset we have used scatter plot to see the distribution in our numerical columns. Given are the sample scatter plot of our test data.

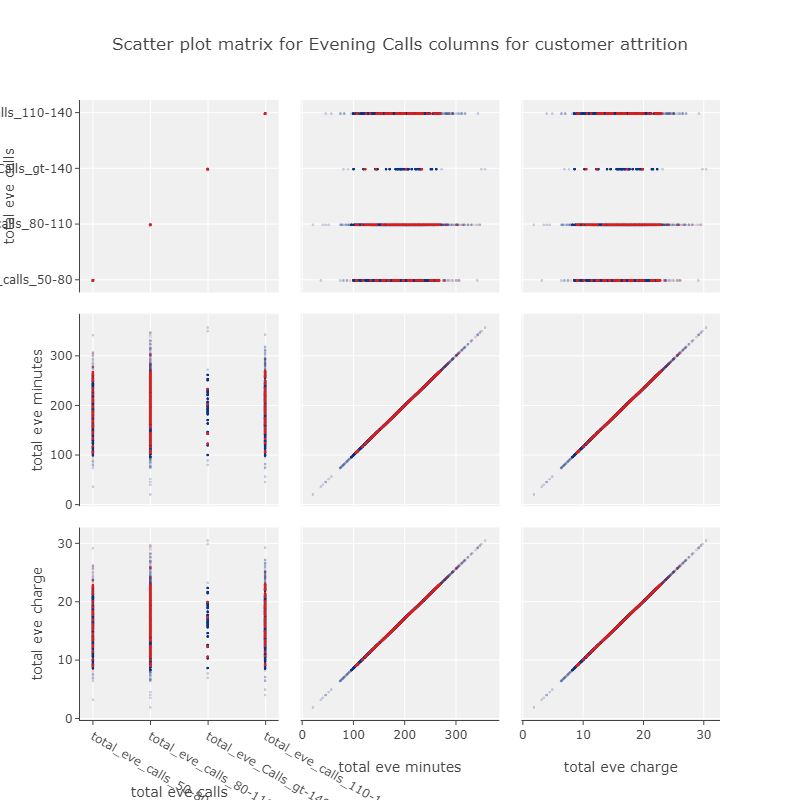
* **Scatter Plot for International Call Columns on Customer Churn:**



* **Scatter Plot Matrix on total day calls columns on Customer Churn:**



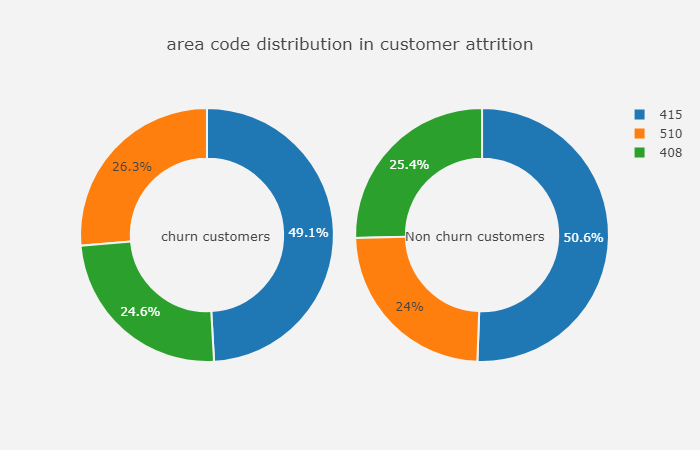
* **Scatter Plot Matrix on Total Evening Calls Columns on Customer Churn:**



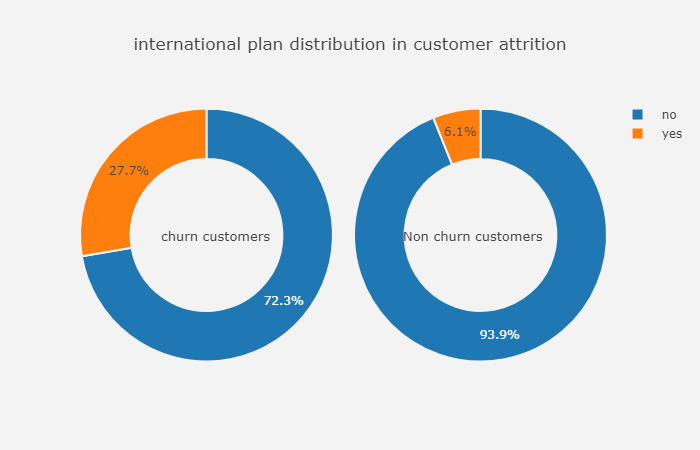
### 2.2 Pie Chart:

Pie Chart is divided into slices to illustrate numerical proportion or percentages of numerical columns towards target variable. We have created Pie chart in our categorical columns of dataset to visualise the percentage of each category for target variable. Given below some sample of Pie chart we have prepared for our test data set.

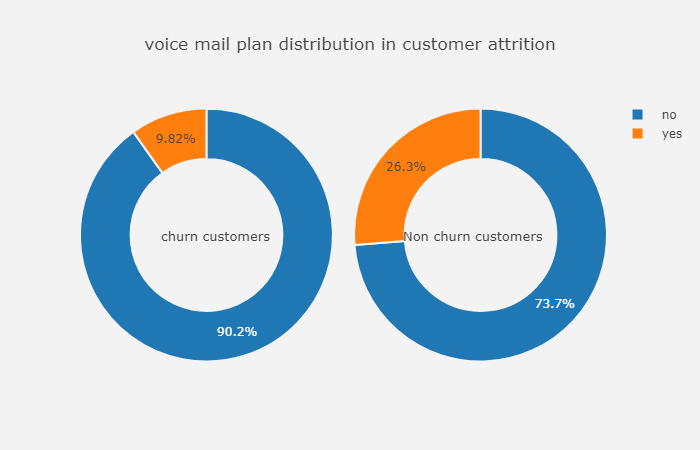
* **Pie Chart of categorical column (area code) on Customer Churn:**



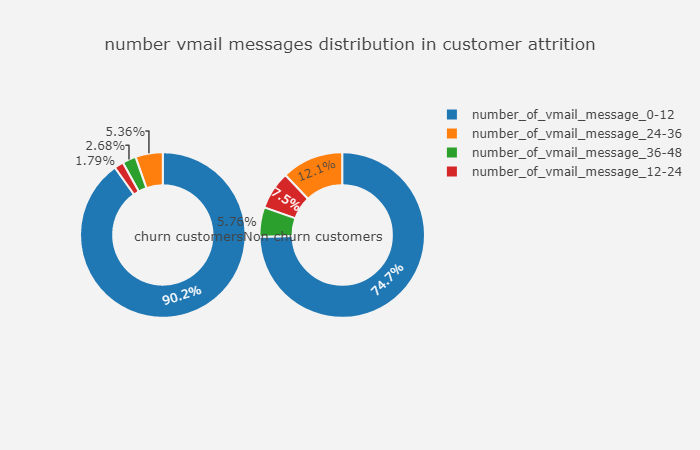
* **Pie Chart of categorical column (international plan) on Customer Churn:**



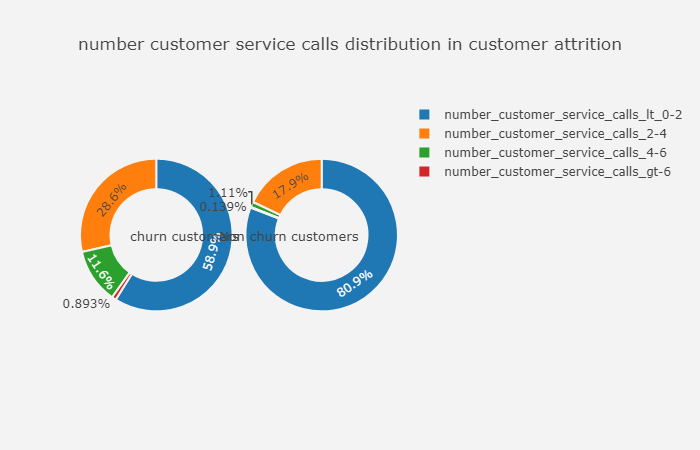
* **Pie Chart of categorical column (international plan) on Customer Churn:**



* **Pie Chart of categorical column (number of vmail message) on Customer Churn:**



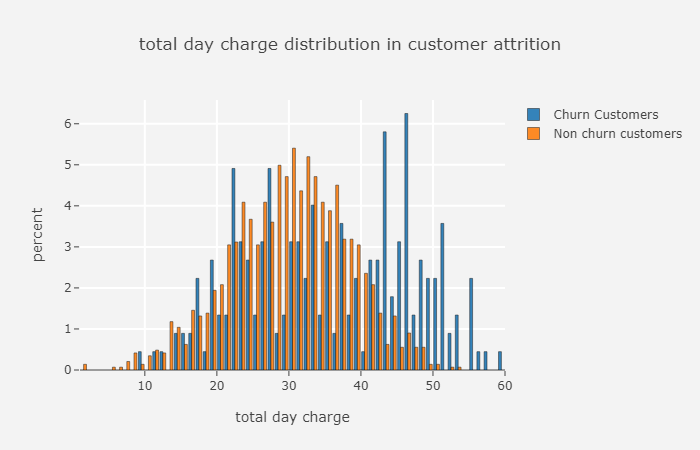
* **Pie Chart of categorical column (number of customer service call) on Customer Churn:**



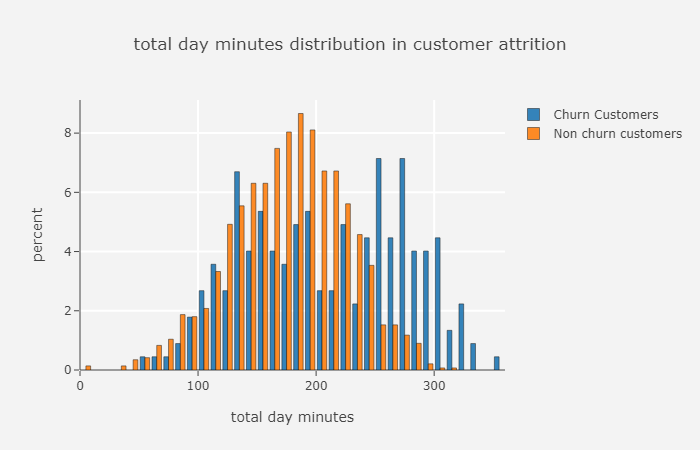
### 2.3 Histogram:

A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable. We have created Histogram for our numerical data to see the percentage of variable within its range based on churn and non-churn category. Given below are some sample histogram we have plotted on top of our numerical columns.

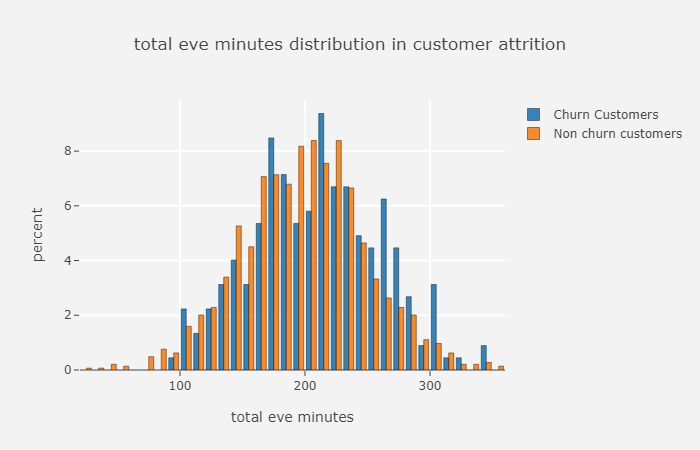
* **Histogram of Numerical columns (total day charge) on customer churn:**



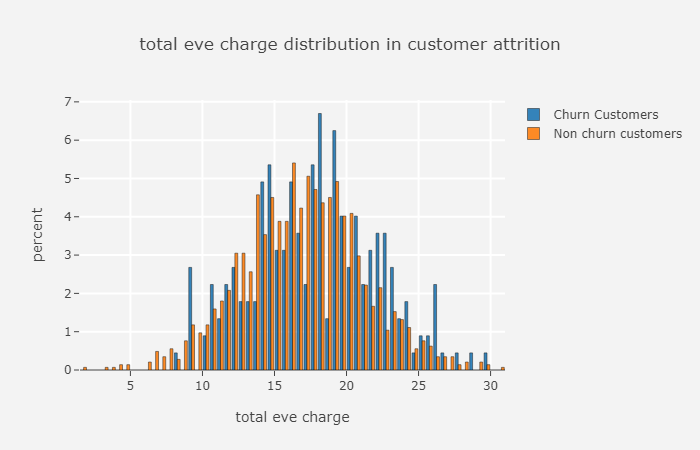
* **Histogram of Numerical columns (total day minute) on customer churn:**



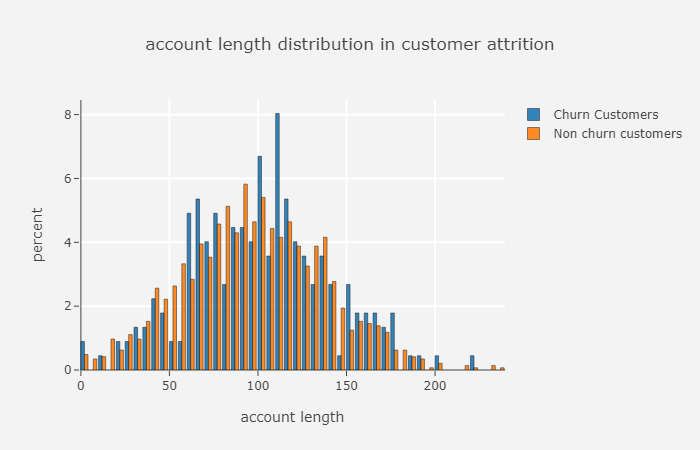
* **Histogram of Numerical columns (total evening minute) on customer churn:**



* **Histogram of Numerical columns (total evening charge) on customer churn:**



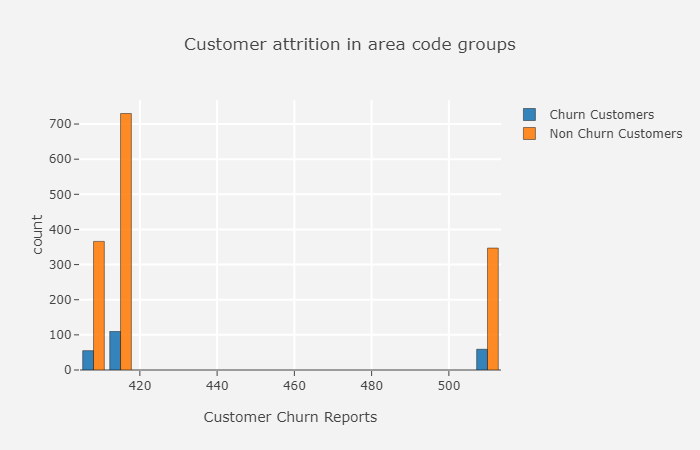
* **Histogram of Numerical columns (account length) on customer churn:**



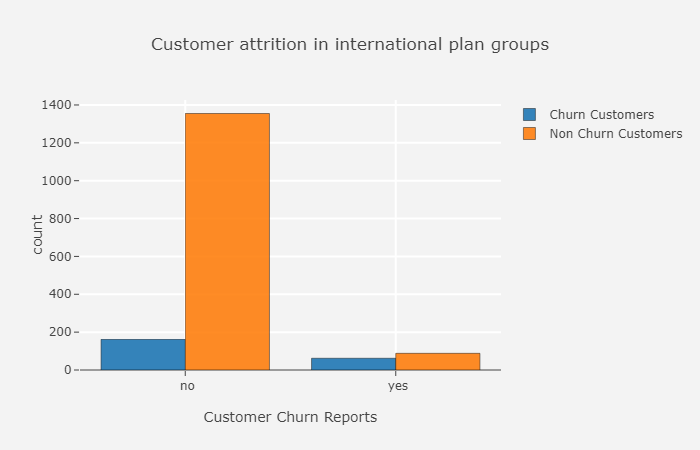
### 2.4 Bar-chart:

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. We have plotted bar-charts for our categorical variable where we can see the count in y-axis and the categories in the x-axis for churn and non-churn data. Given below are some examples of Bar chart we have plotted based on our test data.

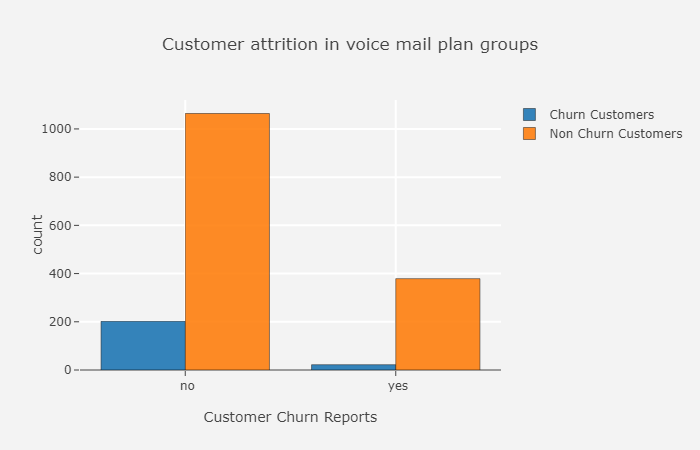
* **Bar-chart for categorical column (Area Code) for customer attrition data:**



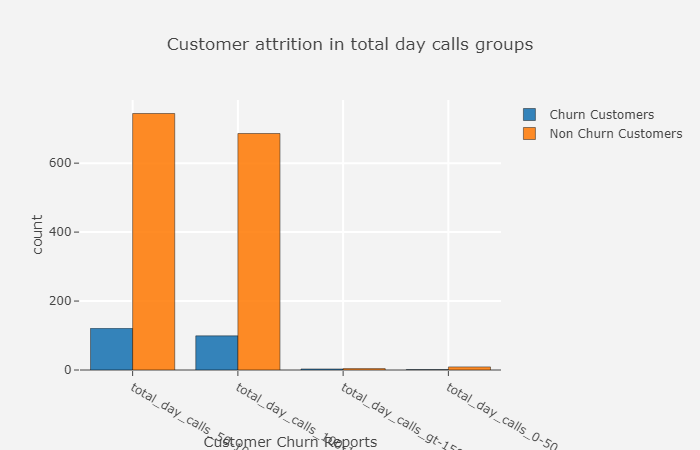
* **Bar-chart for categorical column (International Plan) for customer attrition data:**



* **Bar-chart for categorical column (Voice Mail Plan) for customer attrition data:**

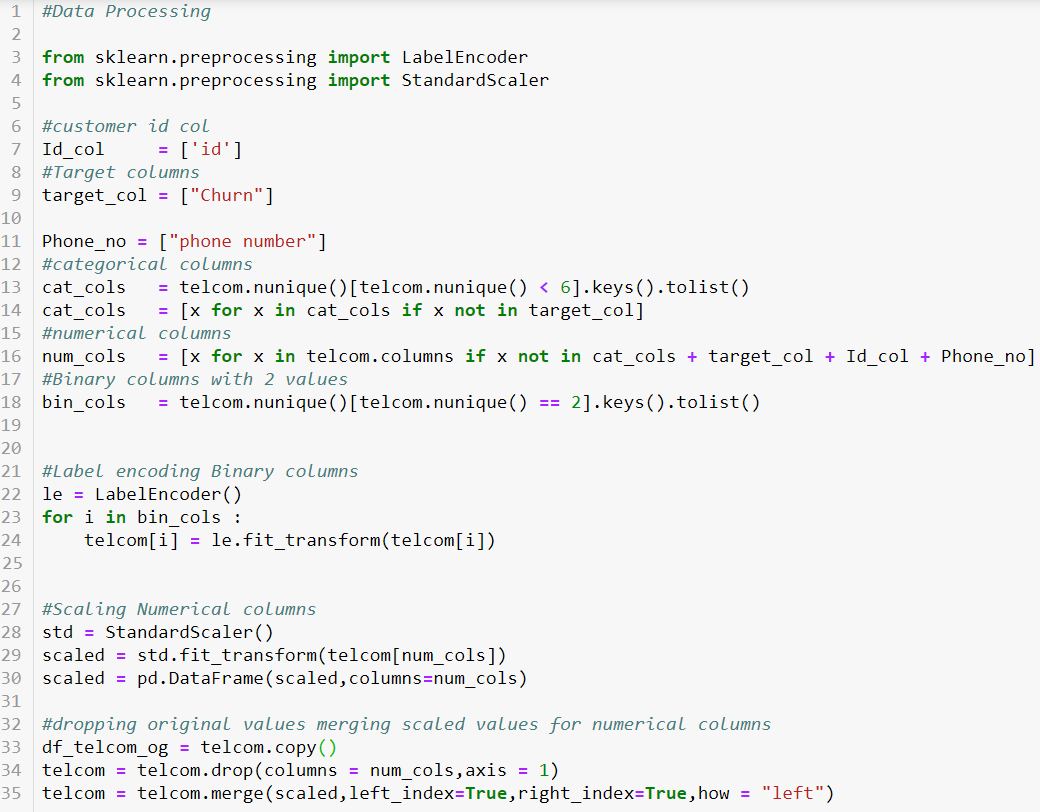


* **Bar-chart for categorical column (Voice Mail Plan) for customer attrition data:**



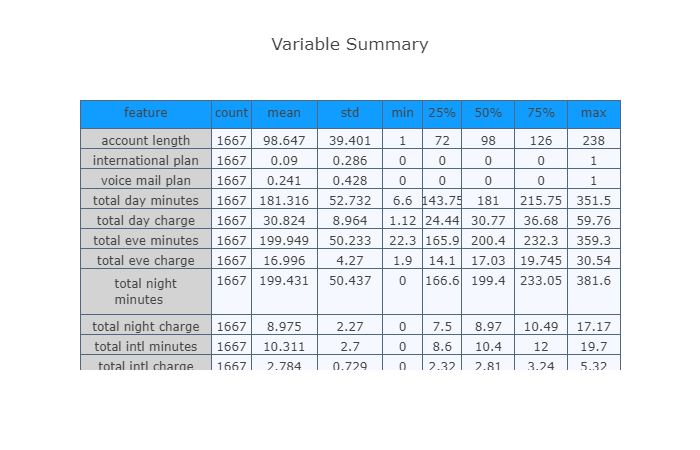
## 2. Data Processing:

Before going to create any model or selecting feature, we need to process some algorithm which require scale data. Hereby we are specifying the ID column, target column, numerical columns, bin columns (which contains 2 categories, named as Binary column). We have imported Labelencoder and StandardScaler from sklearn package and applied Label encoding for Binary columns and Scaled the numerical columns. We have created a copy of telcom data which contains now scaled numerical variables and labelled categorical variables. This dataset is now prepared for further operations.



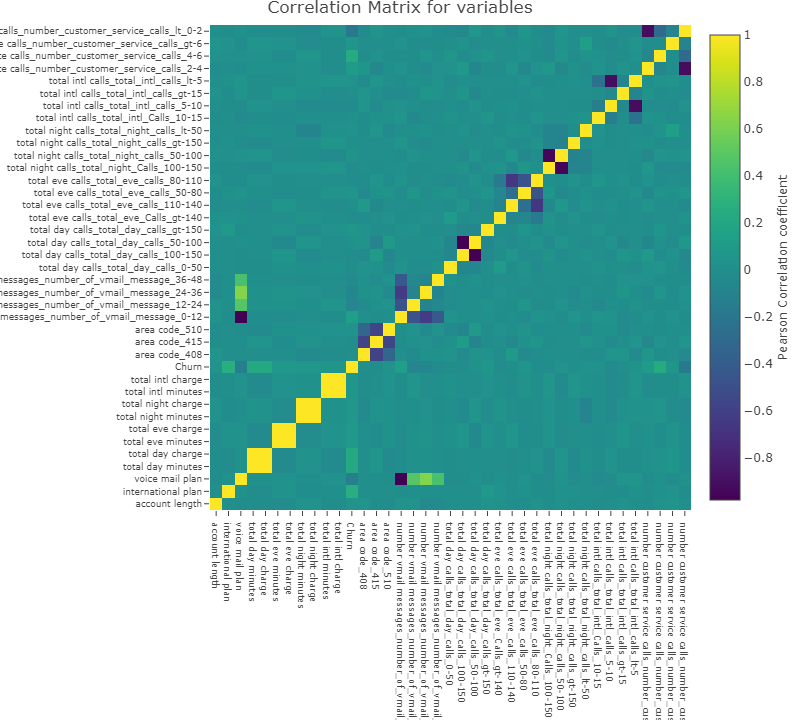
## 3. Variable Summary:

We have listed down summary of the variable we have in our telcom dataset. This summary includes Features, Count, Mean, Standard Deviation, Min, 25%, 50%,75%, Max in a table format.



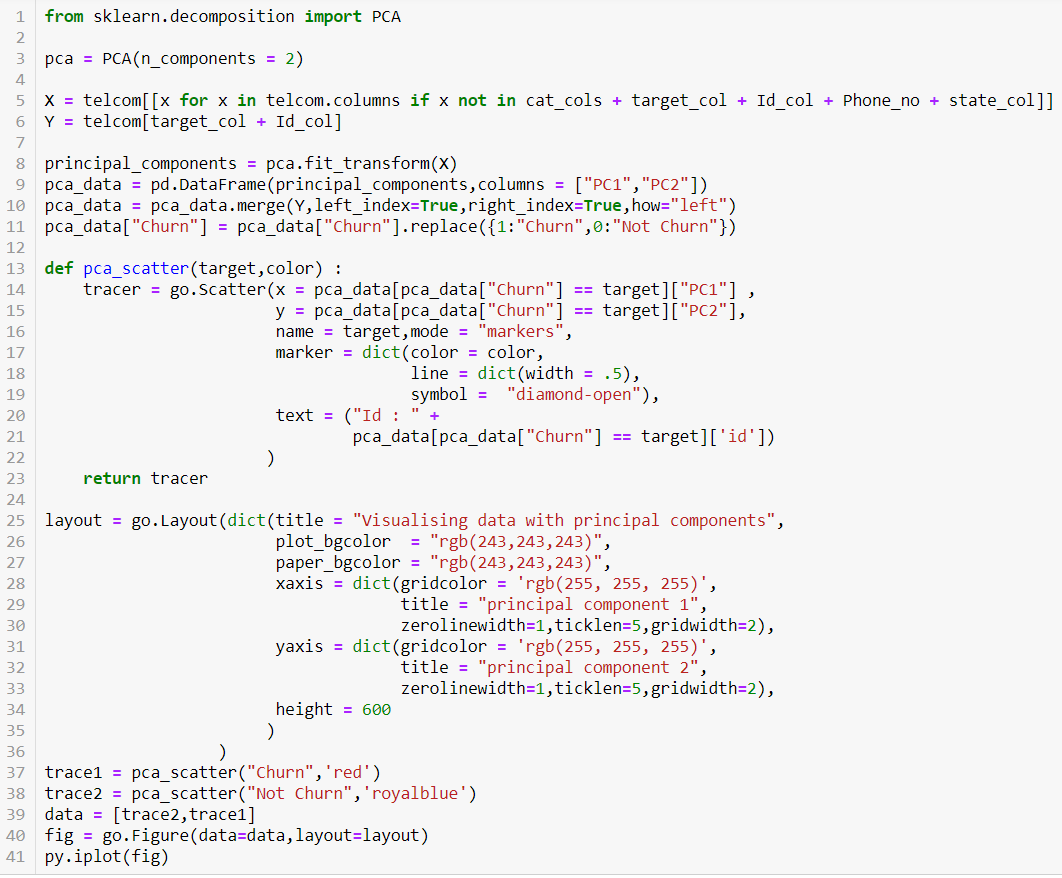
## 4. Correlation Matrix:

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease. We basically apply correlation matrix for numerical variables. In our given dataset we have calculated correlation matrix where we can compare the r value and predict how variables are correlated.

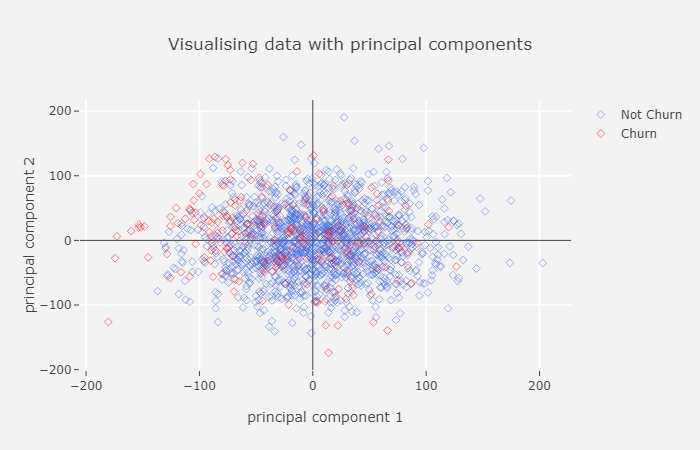


## 5. Visualizing data with Principal Components:

* PCA is basically Principal Component Analysis which is used to reduce the dimensionality in the dataset. Suppose we have 21 columns in our dataset, among those all columns are not necessary. In that case, we use PCA and see the importance of each feature and decommissioned those variables which has no contribution towards our prediction.
* In the given picture below, we have called PCA from sklearn package and defined X and Y axis as PC1and PC2. Then we have called the function pca.fit and passed the value of X.



* We have created a scatter plot of our component PC1 and PC2 and compare the variable feature importance of each column.



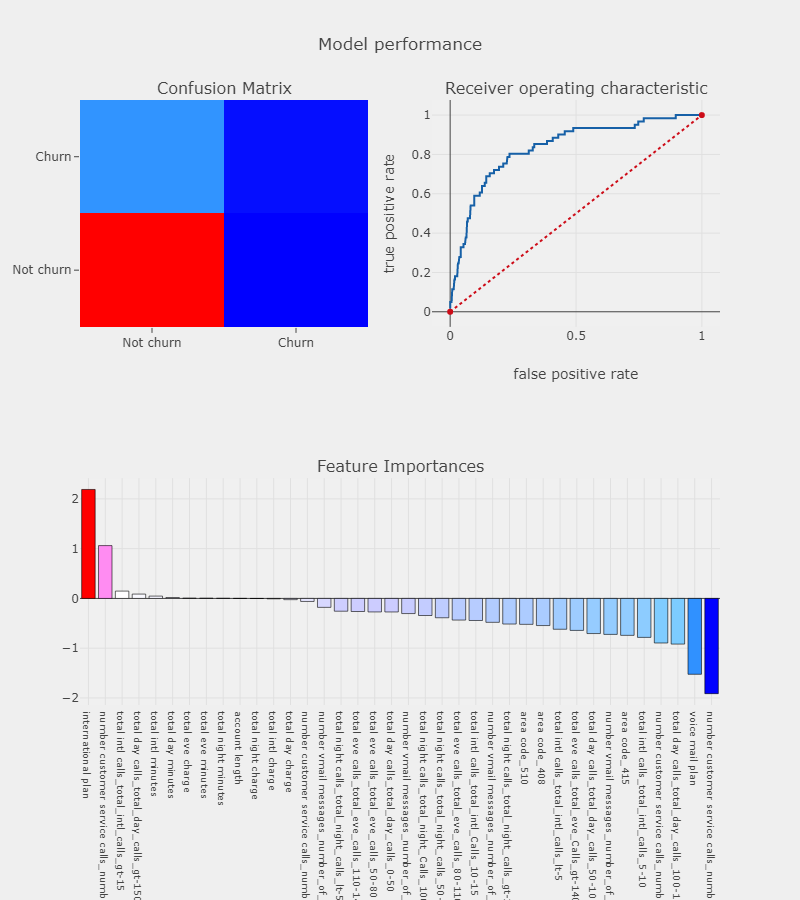
## 6. Model Building:

### 6.1 Logistic Regression Model:

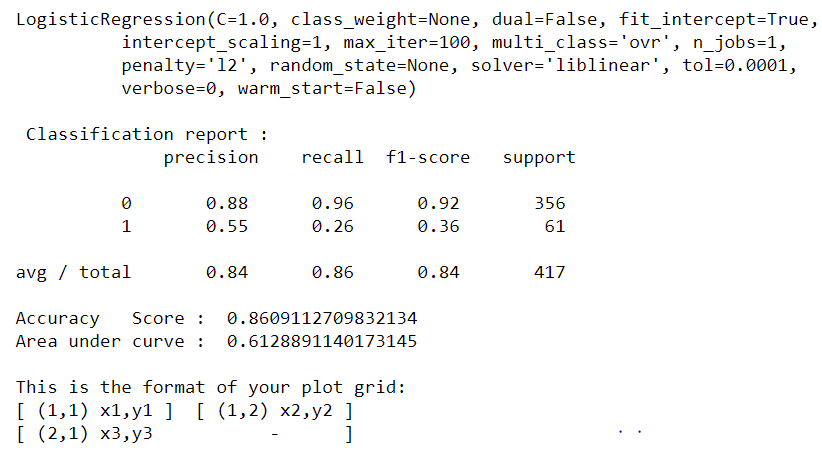
* We have used sklearn library which holds all those functions we require to build the model. “Train\_test\_split” helps to split the data to train and test category based upon given percentage.
* Then have imported the LogisticRegression function, **which** is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
* Then we have calculated Confusion Matrix, which is a technique for summarizing the performance of a classification algorithm.
* We have imported Accuracy Score, classification report to get f1 score, p-value to judge the model performance.
* By getting the value of these matrix, we can understand if there is any imbalance in our data. Because accuracy doesn’t hold well, if our data is imbalances. In that extend we need to use SMOTE to reduce the biasness in dataset.
* As we can see in the below picture that we have splitting .25 part of our data that in 25% as our test data and providing the random state which is basically an offset in dataset.



* We have created a generic function called “telecom\_churn\_prediction” and passing the attribute in it. We have provided the required column as train\_x and getting prediction for test\_x.



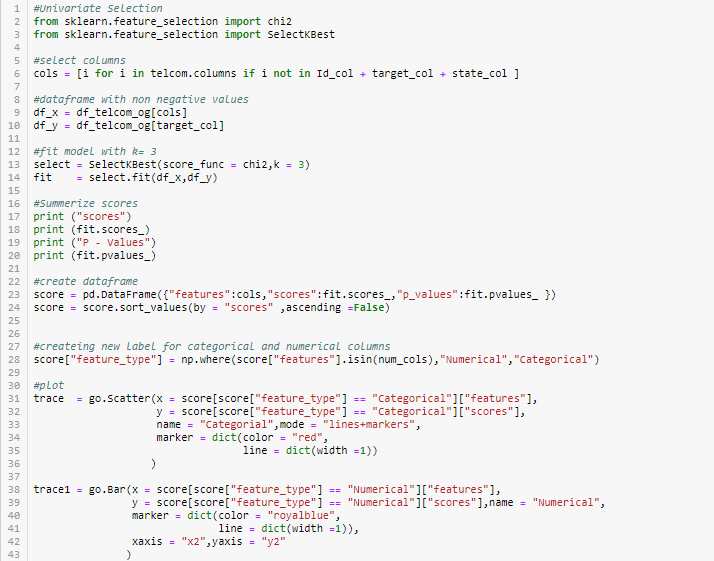
* Given above is the plot where we can see the AUC-ROC curve and score for each feature in the Feature Importance matrix. Also, we have plotted the Confusion Matrix, we have build based upon our data.
* In the below screenshot we can see the accuracy is 86% which is good as per industry standards.



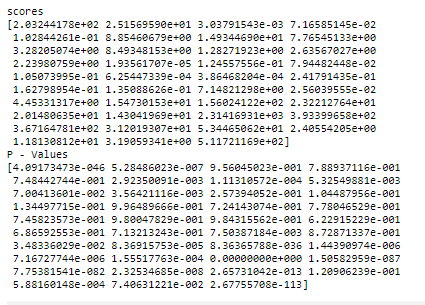
* F score is the harmonic mean of precision and recall. From the F-score we can understand that we have a healthy dataset which is 92 % and 36%. If we see our accuracy is good but our F score is not satisfying in that case, we need to use SMOTE to reduce the biasness in our data.

### 6.2 Univariate Selection

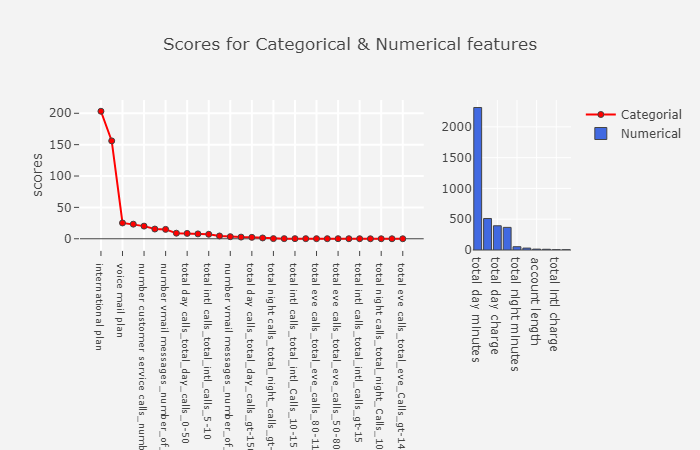
* Feature Extraction with Univariate Statistical Tests (Chi-squared for classification) uses the chi squared (chi^2) statistical test for non-negative features to select the best feature.



* Here we are passing the data frame, which we have created with scaled and labelled data named as “df\_telcom\_og” and getting the non-negative values for our required columns.
* Then we have plotted the score and P-value for each column we have.

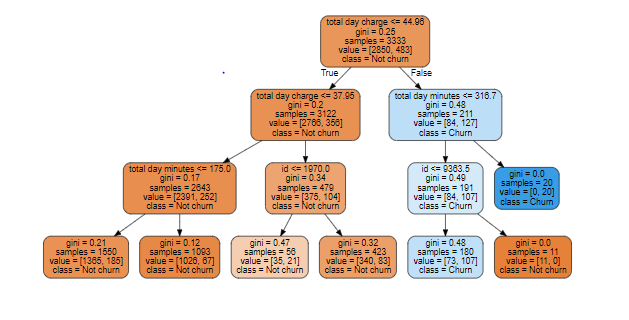


* Then we have plotted the same in a Bar plot for categorical variable and scatter plot for numerical column to visualize the importance of each features.



### 6.3 Decision Tree:

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Decision tree works upon control statement. Here in this project we have created Decision tree model for Churn Prediction Classification model and plotted the tree as well based on Gini index.



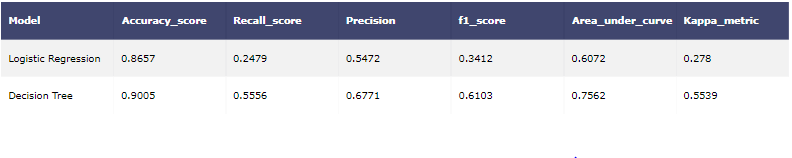
# Chapter 3: Conclution:

### Model Selection:

* Now we can see the models we have prepared and compare which one is giving us the optimum result based on certain parameters.
* We have created Confusion matrix for evaluating our models as we can see the output in given picture below.

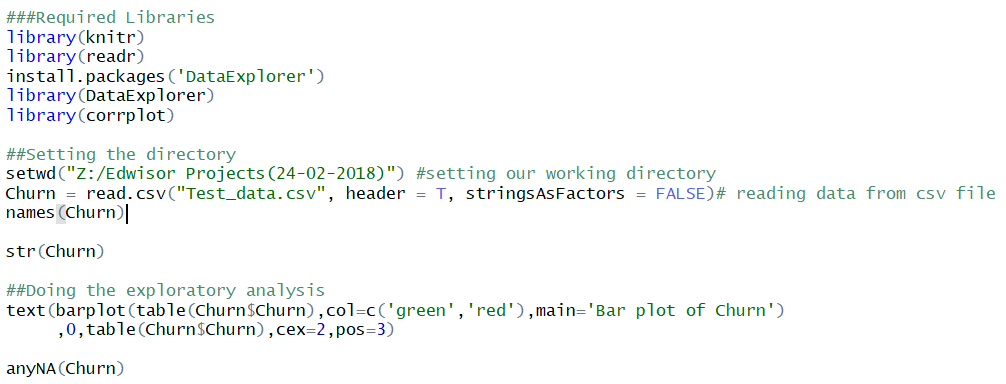


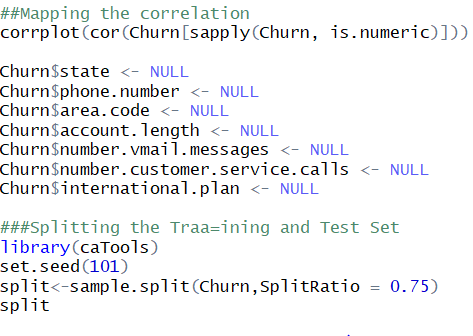
* We have used Accuracy, Recall, Precision, f1 score, AUC, Kappa\_metric as parameters for evaluating our model. Here in the given result below we can see we have got the higher accuracy for Decision Tree model than Logistic Regression Model.
* Now we can tune (by changing the parameters, pruning for underfitting data and SMOTE for overfitting data) our data and parameters for accurate measures.

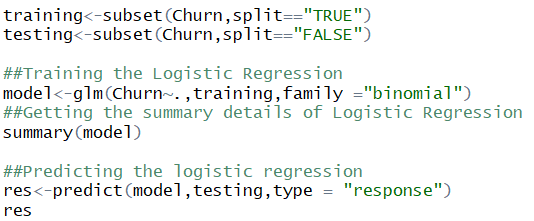


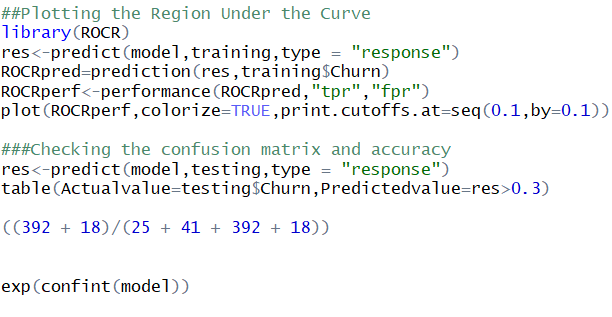
# Appendix:

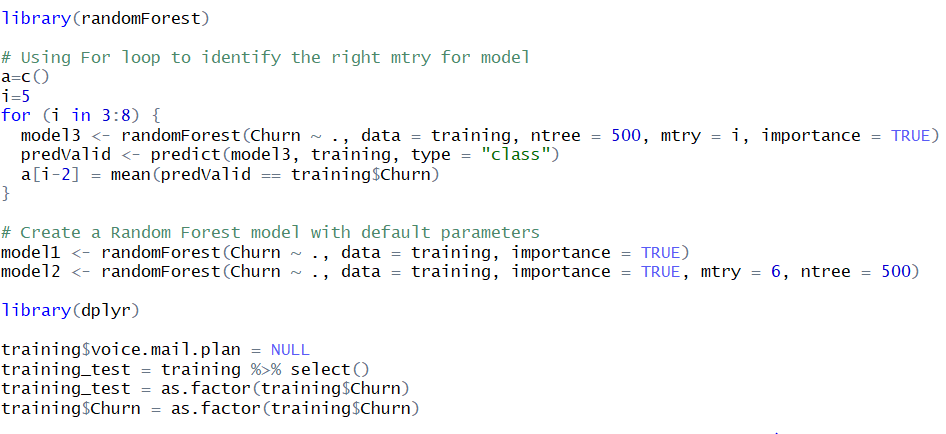
# Complete R-code:

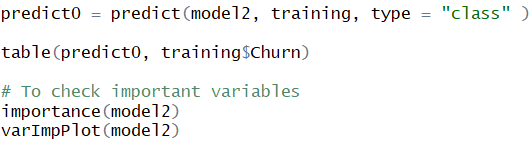


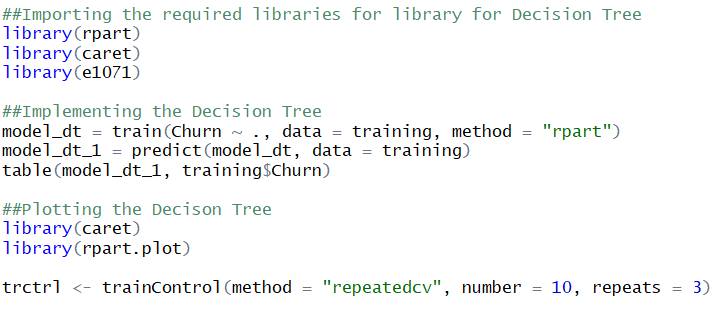


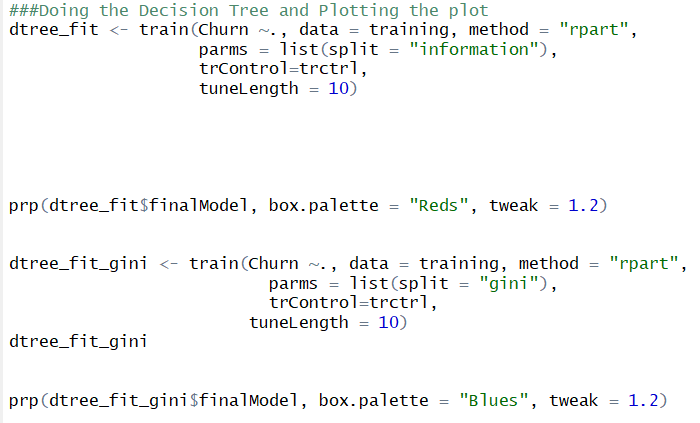


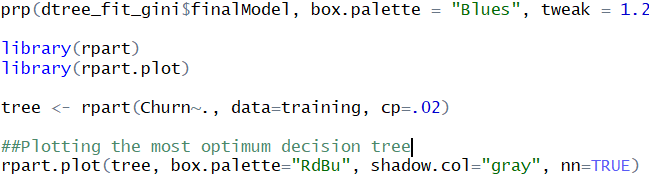




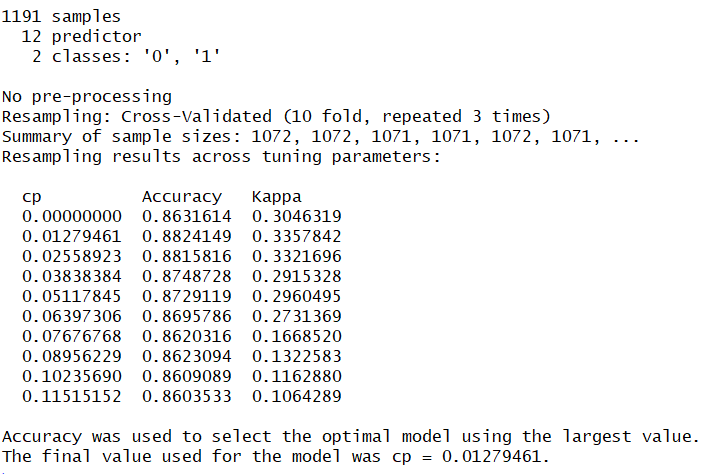


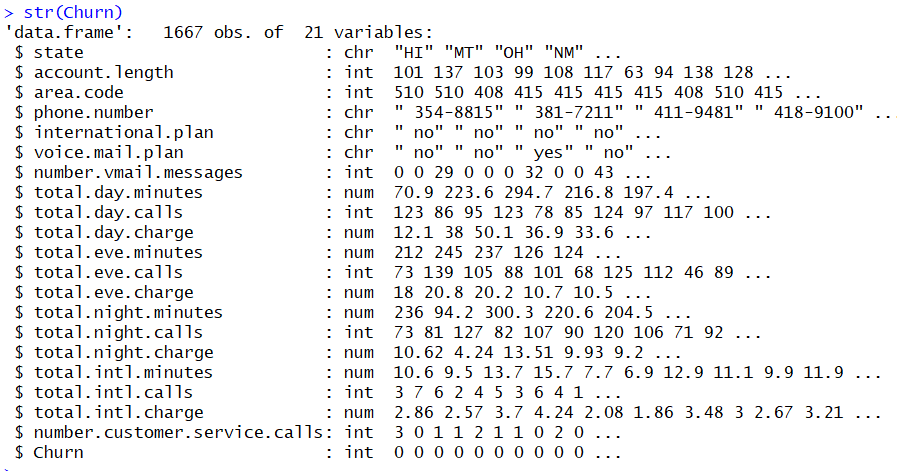


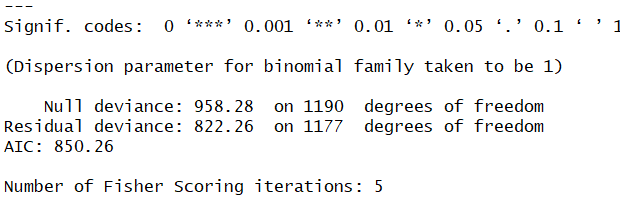


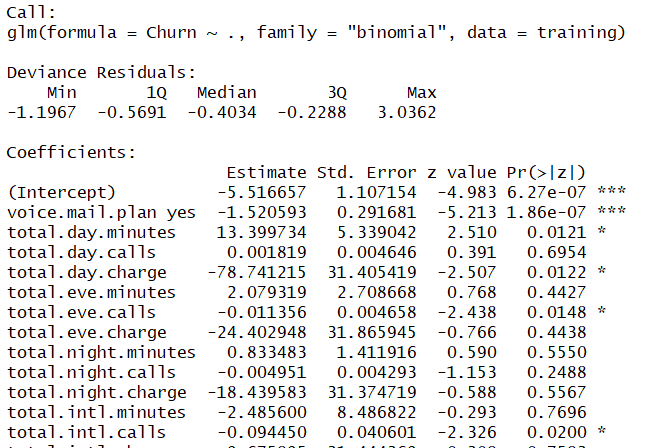


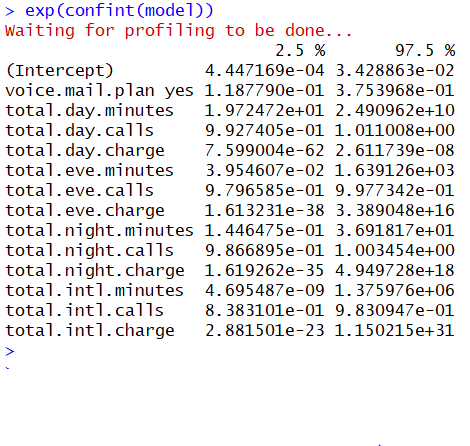
Outputs:

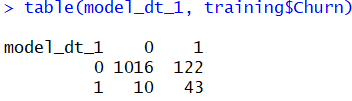


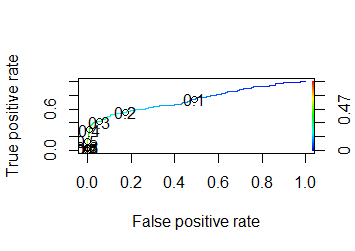


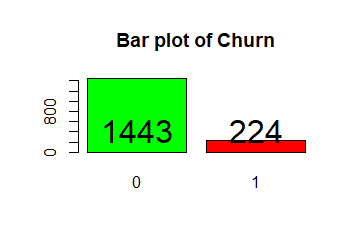


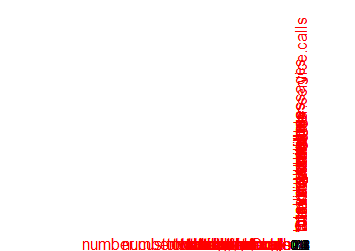


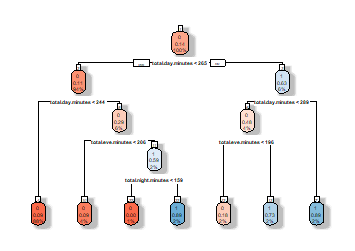


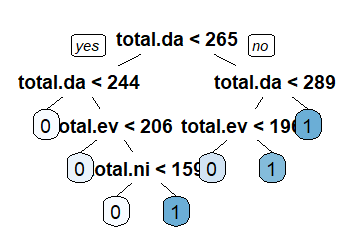


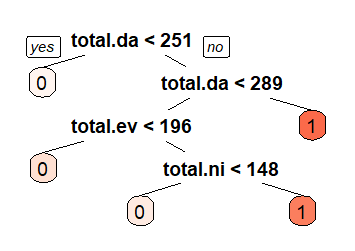












# Reference:

* <https://api.rpubs.com/>
* <https://github.com/>
* <https://stats.stackexchange.com/>
* <https://towardsdatascience.com/>
* https://www.analyticsvidhya.com/