|  |
| --- |
|  |
| CP G1 Group Assignment |
| Predictive Modeling - Telecom Customer Churn Prediction Assessment |

Amit Sharma

Amitesh Bajpai

Sahil Sachdeva

Kushal Maheshwari

Neeraj Singh

Contents

[1. Project Objective 3](#_Toc24826894)

[2. Exploratory Data Analysis – Step by step approach 3](#_Toc24826895)

[Assumptions made before starting our analysis: 3](#_Toc24826896)

[2.1 Environment Set up and Data Import 3](#_Toc24826897)

[2.1.1 Installation of necessary packages and calling libraries 3](#_Toc24826898)

[2.1.2 Set up working directory 4](#_Toc24826899)

[2.1.3 Import and read the dataset 4](#_Toc24826900)

[2.2 Variable Identification and Transformation 4](#_Toc24826901)

[2.3 Missing Value/Negative Value Identification 6](#_Toc24826902)

[2.4 Outlier Detection and Treatment 6](#_Toc24826903)

[2.4.1 Outlier Detection 6](#_Toc24826904)

[2.4.2 Outlier Treatment 7](#_Toc24826905)

[2.5 Exploratory Data Analysis 8](#_Toc24826906)

[2.5.1 Univariate Analysis 8](#_Toc24826907)

[2.5.2 Multivariate Analysis 10](#_Toc24826908)

[2.5.3 Correlation between variables 15](#_Toc24826909)

[2.5.4 Interpreting the business problem 16](#_Toc24826910)

[3. Logistic Regression 17](#_Toc24826911)

[3.1 Logistic Regression with all variables 17](#_Toc24826912)

[3.2 Logistic Regression after exclusion of correlated variables 18](#_Toc24826913)

[3.3 Logistic regression model after removal of insignificant variables 19](#_Toc24826914)

[3.4 Confusion matrix for final logit model 20](#_Toc24826915)

[3.5 ROC curve and AUC 20](#_Toc24826916)

[3.7 K-S 21](#_Toc24826917)

[3.8 Gini 22](#_Toc24826918)

[4. Actionable Insights and Recommendations 22](#_Toc24826919)

# 1. Project Objective

The objective of the report is to explore the dataset named “**Cellphone (1).xlsx**” in R and based on this past data, we need to build a model which can predict whether a customer will cancel their service in the future or not. This exploration report will consist of the following:

* Detailed Exploratory Data Analysis
* Multicollinearity check and summarization of problem statement for business stakeholders
* Logistic Regression model
* Model Performance comparison to select best model
* Actionable Insights for the business stakeholders

# 2. Exploratory Data Analysis – Step by step approach

We shall follow these steps to go through our data exploration activity:

1. Environment Set up and Data Import
2. Variable Identification and Transformation
3. Outlier detection and treatment
4. Missing/Negative value identification and treatment
5. Data Visualization
   1. Univariate Analysis
   2. Multivariate Analysis
   3. Correlation Matrix

## Assumptions made before starting our analysis:

* Our sample size is large enough to yield reliable estimates of correlations among the variables.
* Our sample size is large enough to build, train and test models.

## 2.1 Environment Set up and Data Import

### 2.1.1 Installation of necessary packages and calling libraries

The below mentioned libraries are necessary for our analyses in R.

* Readxl
* Car
* carData
* Caret
* ggplot2
* pROC
* psych
* corrplot
* caTools
* ineq
* scales
* graphics
* stats
* lattice

### 2.1.2 Set up working directory

We need to set up our working directory, install all the necessary packages and import the cellphone dataset in question, to proceed with our study.



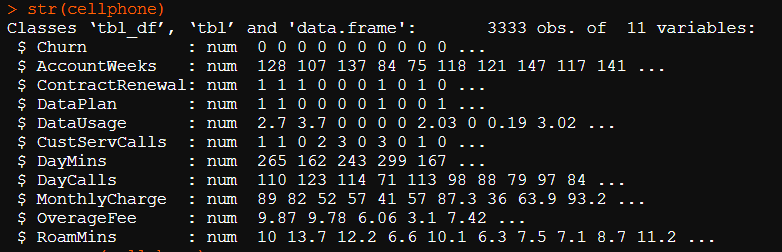
### 2.1.3 Import and read the dataset

The given dataset is in .xlsx format. Hence, the command ‘read.excel’ is used for importing the file. There are two sheets in the file, and we are interested in Sheet #2 and hence we import only the 2nd sheet.

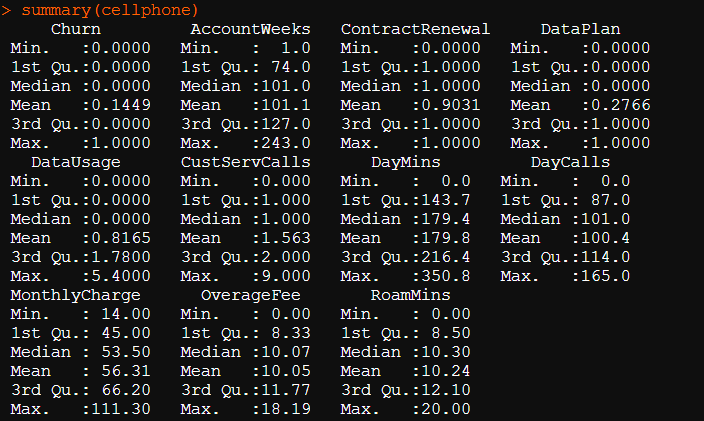


## 2.2 Variable Identification and Transformation

The “str” function shows us the structure of the dataset.



The “summary” function performs a Univariate analysis on our dataset.



The summary shows us that the mean and median are very close for the continuous variables. Also, there are no negative values in our dataset. Also, we can conclude below points.

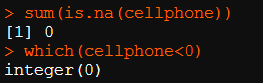
Churn is a target variable as it predicts whether a customer will cancel their service in the future or not.

Churn, ContractRenewal and DataPlan are categorical variables so need to convert them to factor.



## 2.3 Missing Value/Negative Value Identification

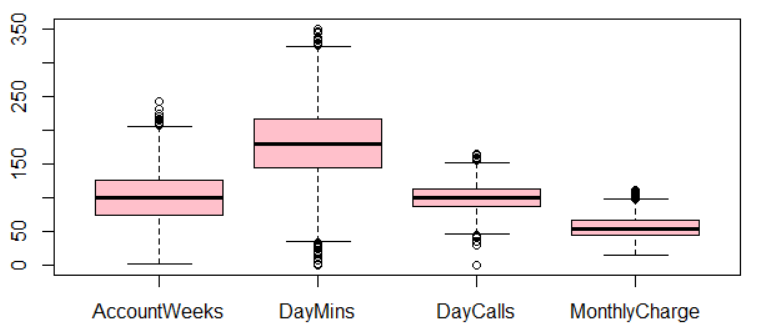
There are no missing or negative datapoints in the file as seen from our analysis below.

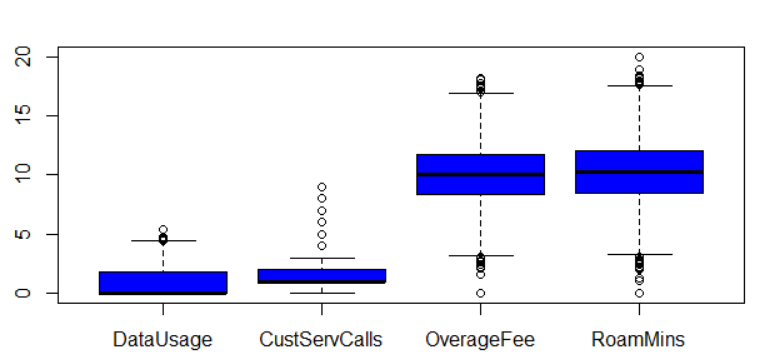


## 2.4 Outlier Detection and Treatment

### 2.4.1 Outlier Detection

We check for outliers by plotting boxplots for all continuous variables.

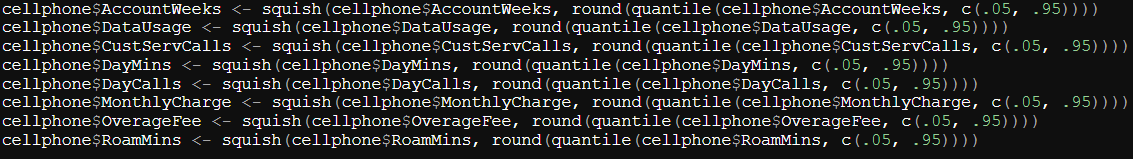




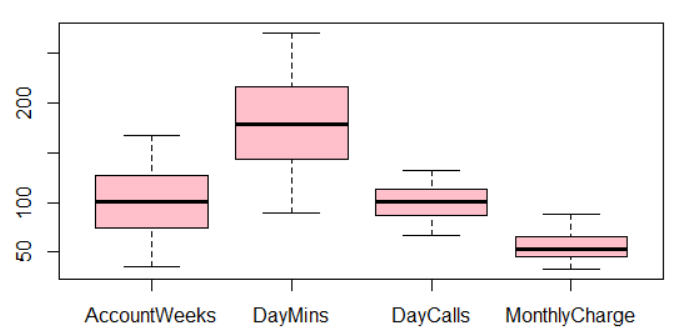
As observed, all the above variables have outliers.

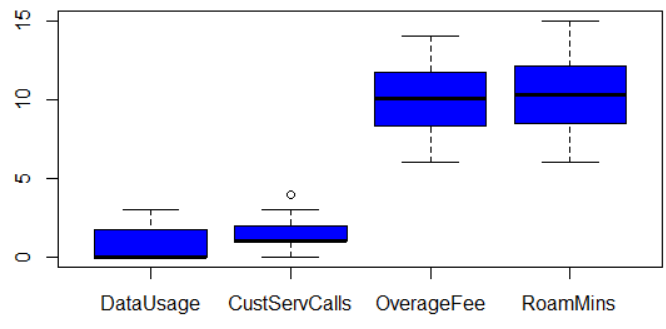
### 2.4.2 Outlier Treatment

Outliers are treated by the flooring and capping outliers’ methods using 5th and 95th percentiles. For this we use squish method from package scales.



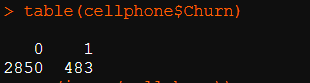
After the treatment, only few outliers are left in the variable *Customer Service calls*, as seen in the new boxplots.





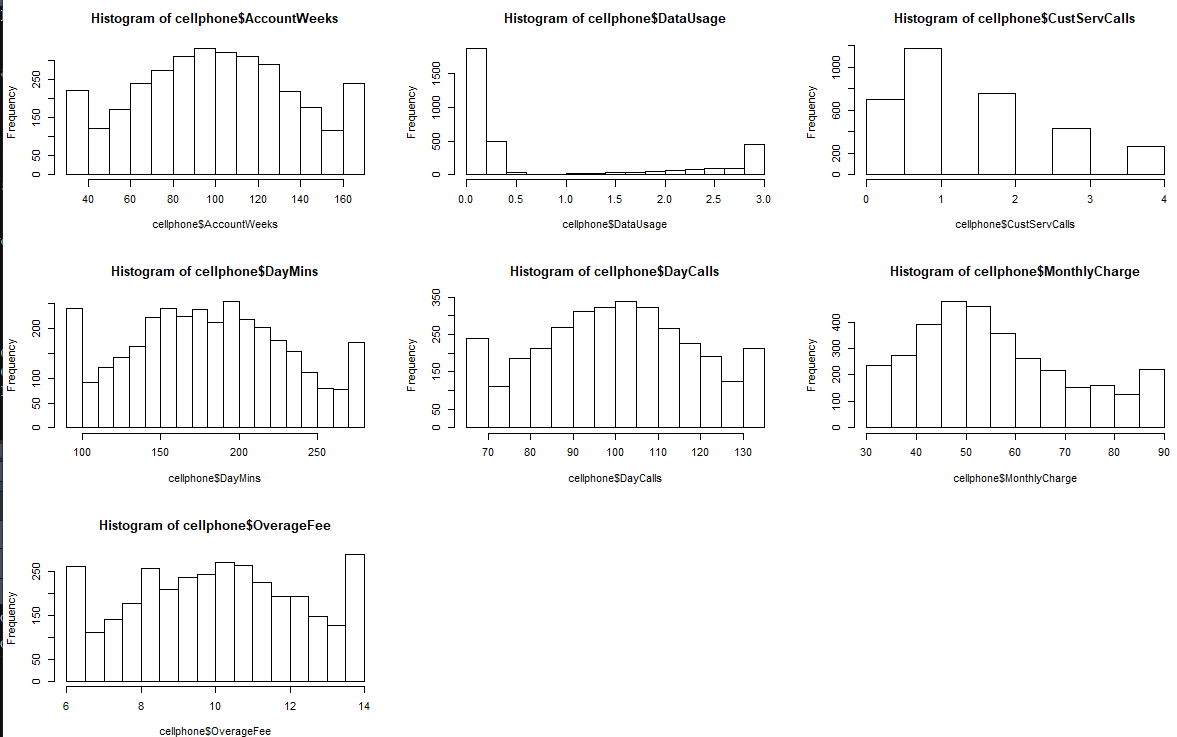
## 2.5 Exploratory Data Analysis

As seen from below results, in our dataset, 483 customers churned, i.e., cancelled service, which is around 14.5% of the population.

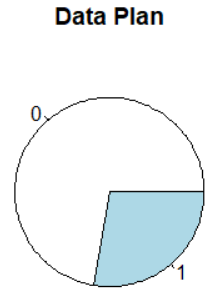
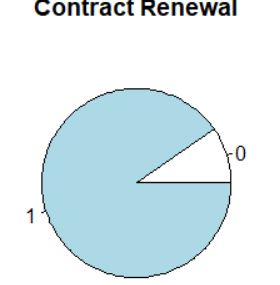
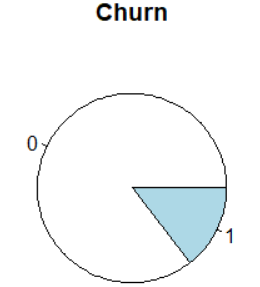


### 2.5.1 Univariate Analysis

We plot histograms for all the continuous variables as seen below.



For categorical variables we use pie charts.



Churn is target variable. 1-Cancelled service customers are 483 customer which is 14% of the dataset provided and 2850 would continue.

Contract Renewal help us understand that around 3000 has renewed the contract and only 10% has not renewed yet.

Only 28% has data plan and good amount of customer which continue without Data plan.

Most customers (almost 90%) recently renewed contracts. 72% customers use less than 1 GB of data in a month.

Account weeks, DayMins, OverageFee, Roam Mins features are almost normally distributed.

Most data of Data Usage, CustServCalls is towards 0 to 2 and right side distributed.

Day calls, data is low between 0 to 50 then histogram is normally distributed.

Average customers have active service from 100+ weeks. Most of the customers in the data set have active accounts from 50 to 150 weeks.

Clearly the Contract Renewal distribution is imbalanced. It seems the people are more interested in renewal of contract with the company.

This can be plus point for the company. 90% of customers have renewed their contract. This is very clear that the customers who have not renewed their contract are more likely to churn.

Around 30% of customers only have data plan. Customers are more likely to churn if they do not have data plan.

Most of the customers have not called or called only 3 times to customer care.

Majority of customers use between 100-300 minutes in a day. The average day time mins per month is normally distributed.

Most of the customers have calls between 50-150 in a day.

Most of the customers have monthly charge between 0 to 100.

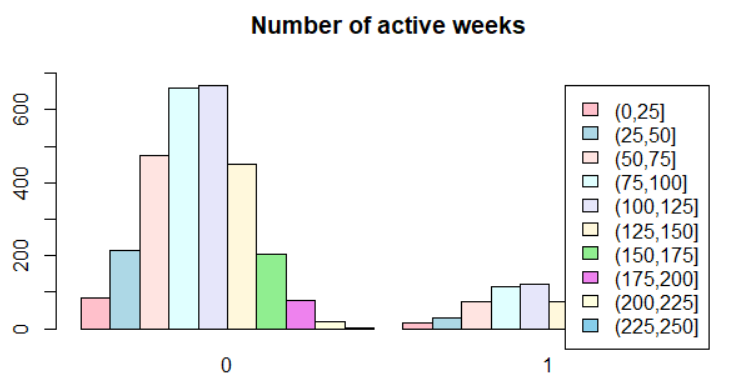
Most of the customers are paying overage fees in bracket 6-12 per month.

90% of customers have roaming mins between 5 to 15min.

### 2.5.2 Multivariate Analysis

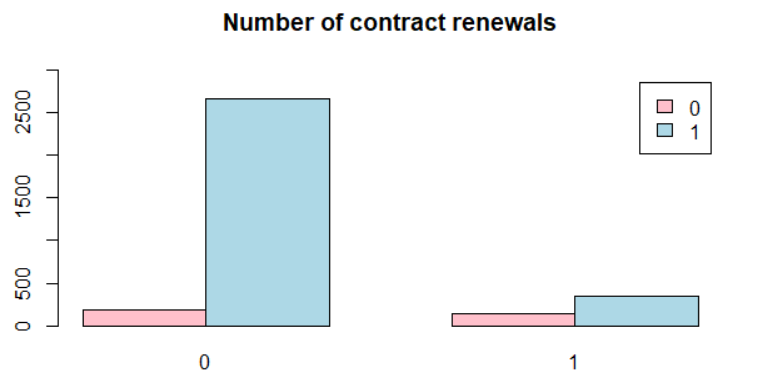
The following plots compares the trend of independent variables when the dependent variable, Churn is 0 and 1.

Account Weeks vs. Churn



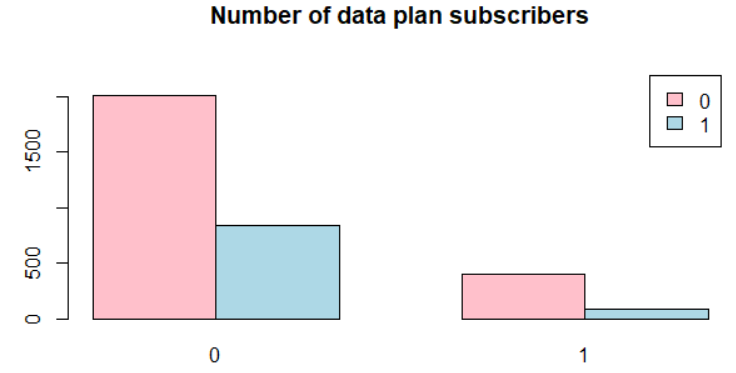
The Churn rate is expected to decrease with increase in active account weeks, but it appears not to be so.

Contract Renewals vs. Churn



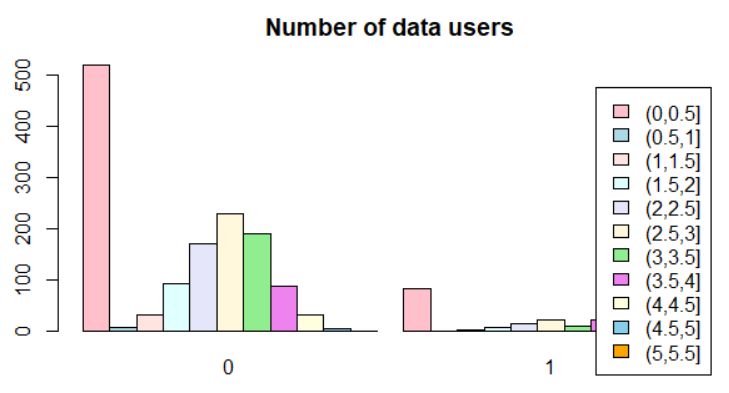
Even though contracts are renewed, more than 11% customers cancelled service. When contracts were not renewed, more than 42% customers cancelled service.

Data Plan subscriptions vs. Churn



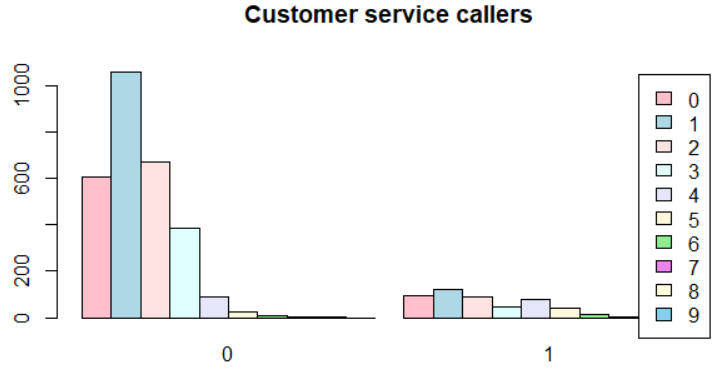
Customers are more likely to cancel service if they do not have a data plan subscription.

Data Usage vs. Churn



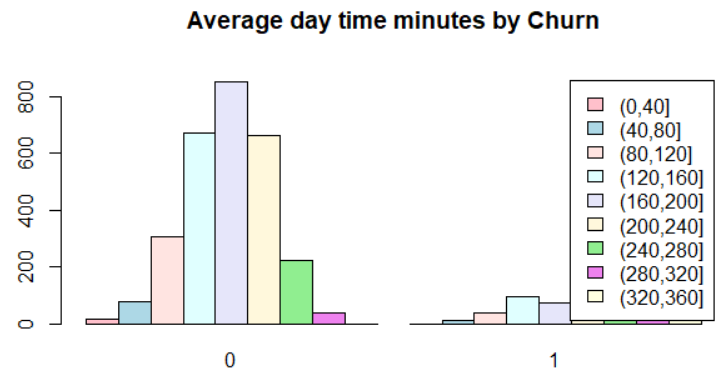
Customers who have less than 0.5 GB of data usage are more likely to churn services.

Calls to customer service vs. Churn



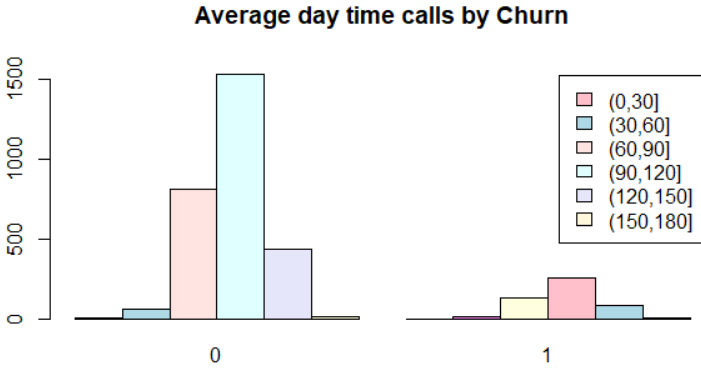
Churn rate is increasing when customers are making 4 or more calls to customer service.

Day time minutes (avg. per month) vs. Churn



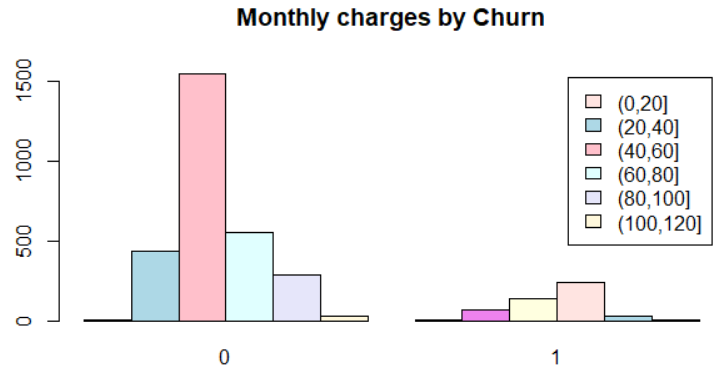
Churn rate is likely to increase if daytime minutes are more than 240 minutes.

Day time calls (avg. per month) vs. Churn



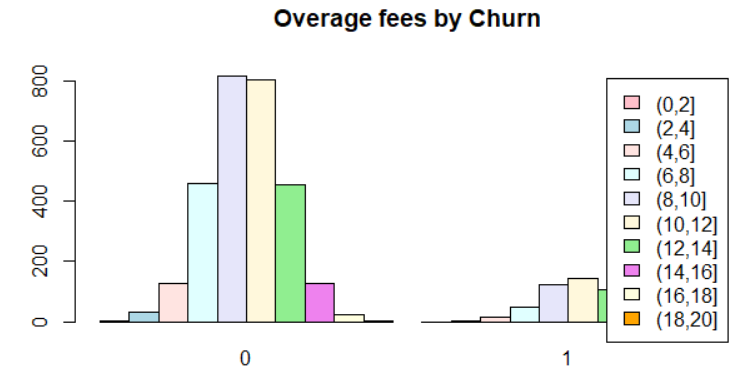
There is no visible churn pattern when compared to daytime calls.

Monthly charges vs. Churn



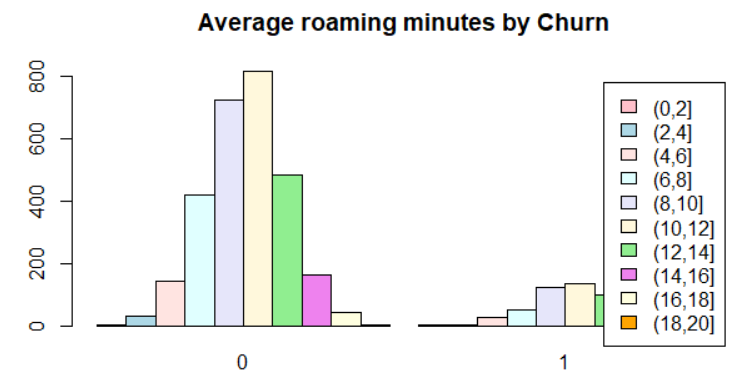
Customers with monthly bill between 60 and 80 are more likely to churn.

Overage Fees vs. Churn



There is no visible churn pattern when compared to overage fees.

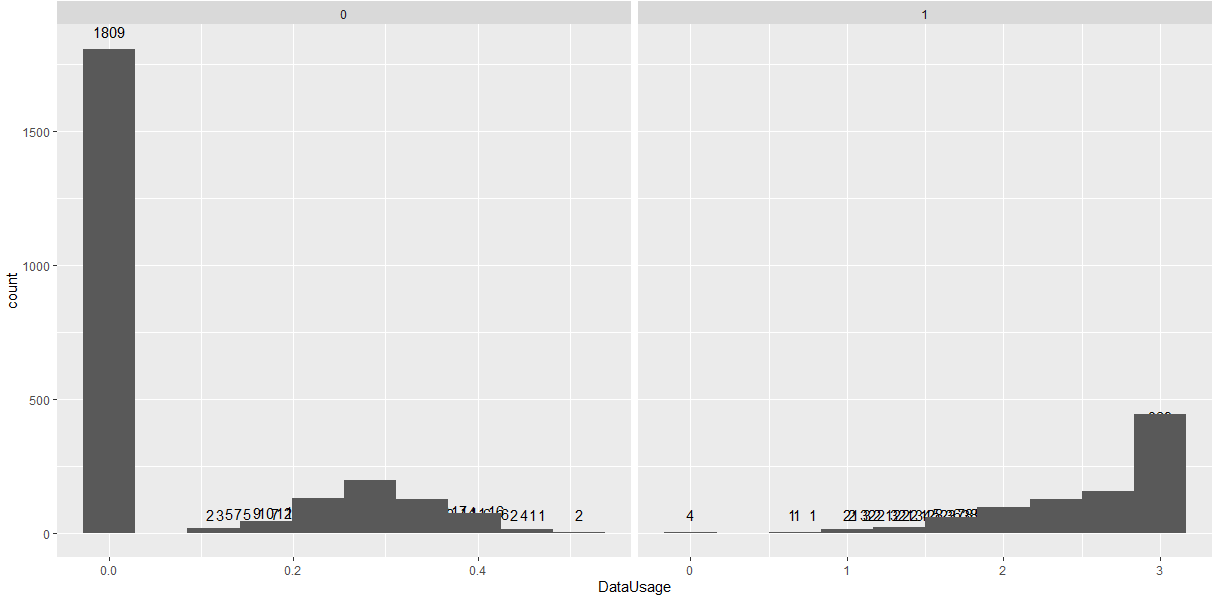
Roaming minutes vs. Churn



There is no visible churn pattern when compared to roaming minutes.

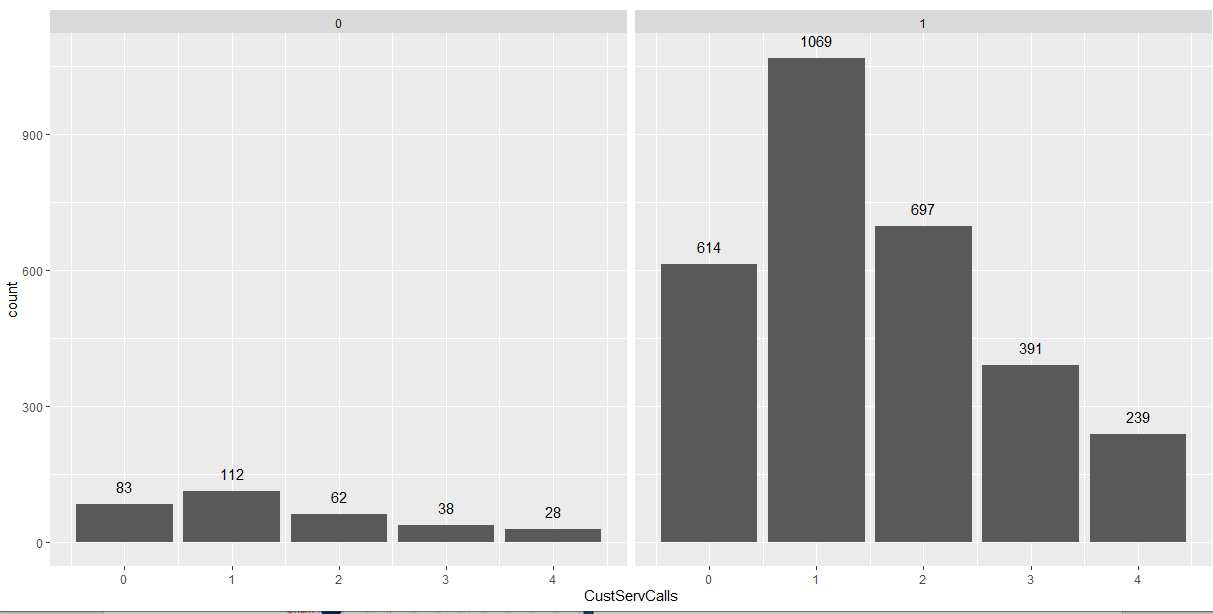
We also examine some other variables with each other, as follows.

Data Usage by Data Plan subscriptions (0 or 1)



There are 4 cases where Data Usage is 0, even though there is a data plan subscription. There are 602 cases where there is no data plan subscription, but there is some amount of data usage.

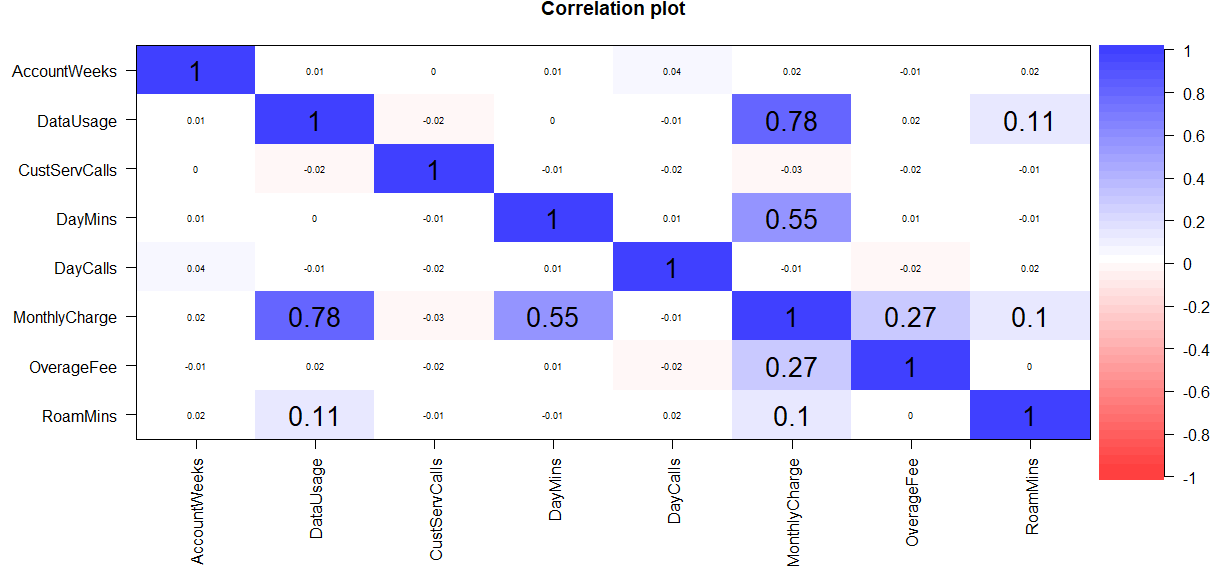
Customer Service Calls vs Contract Renewal



There are 240 cases where customer has not renewed contracts even after many calls to customer service. There are 83 cases where contract has not been renewed and there have been no calls to customer service either.

### 2.5.3 Correlation between variables

Let's look at the correlation between all the variables before we proceed further with creating Logistic Regression model and treat highly correlated variables accordingly to build the regression model.



Data Usage and Data Plan are highly corelated. Monthly Charge is also highly correlated with Data Usage, Data Plan and Day Mins. Churn does not seem to be highly corelated with any of the variables. Churn has maximum correlation with Contract Renewal, Customer Service Calls and Day Mins.

### 2.5.4 Interpreting the business problem

Objective is to reduce customer churn for Telecom service provider by identifying the potential churn candidates beforehand and take proactive actions to make them stay.

Customer churn is subscriber’s cancelling their mobile or fixed line services or porting out to another service provider.

Customer churn is a loss to service provider as its difficult to win back the lost customer and it’s bad marketing if he/she is leaving due to bad customer or user experience.

There are several reasons due to which customer switch from the existing network provider

- Network issues

- Billing issues

- Competitive offers from another Telecom Service Provider

- Financial hardship

- Bad user experience

- High roaming charges

- Overage charge

In this case study we will try to build model as per existing data for the customer who has already churned and their behavior before they actually switched to another operator and what was the reason behind the same so that we can identify potential customer who may get churn in future and service provider can take necessary steps to stop same.

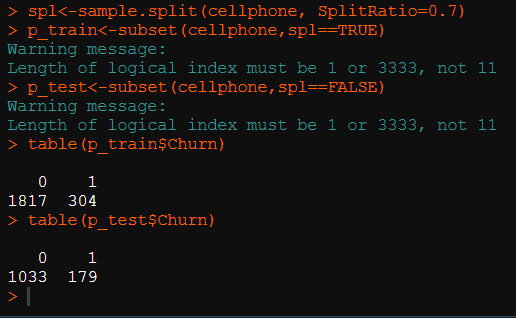
# 3. Logistic Regression

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. GLM does not assume a linear relationship between dependent and independent variables. The dependent variable need not be normally distributed.

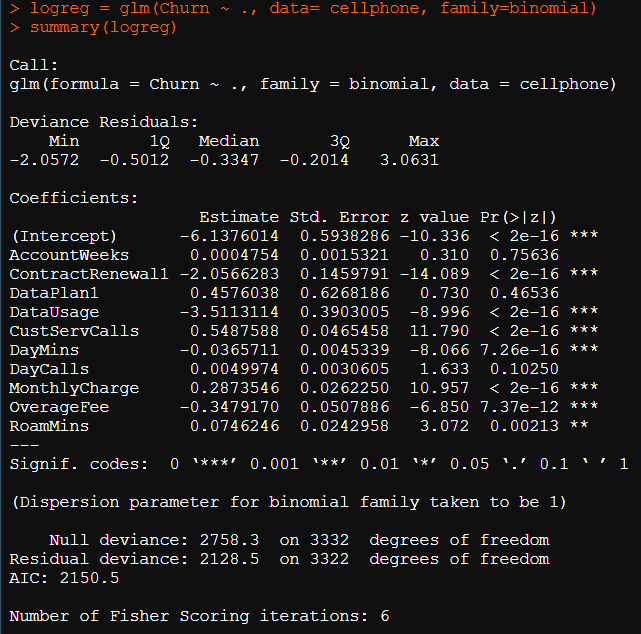
We will start with Logistic Regression Analysis as it will give us clear insight that what are those variables which are significant in building predictive model so that we can achieve more precision by eliminating irrelevant variables so before proceeding we will be splitting available customer data into Train and Test data set and then performing logistic regression.

## 3.1 Logistic Regression with all variables

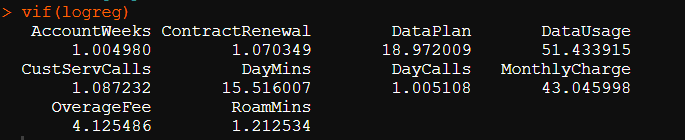
The first step is to split the data into training and test sets. We choose a split of 70% and 30%.



Next, we build the Logistic Regression model.



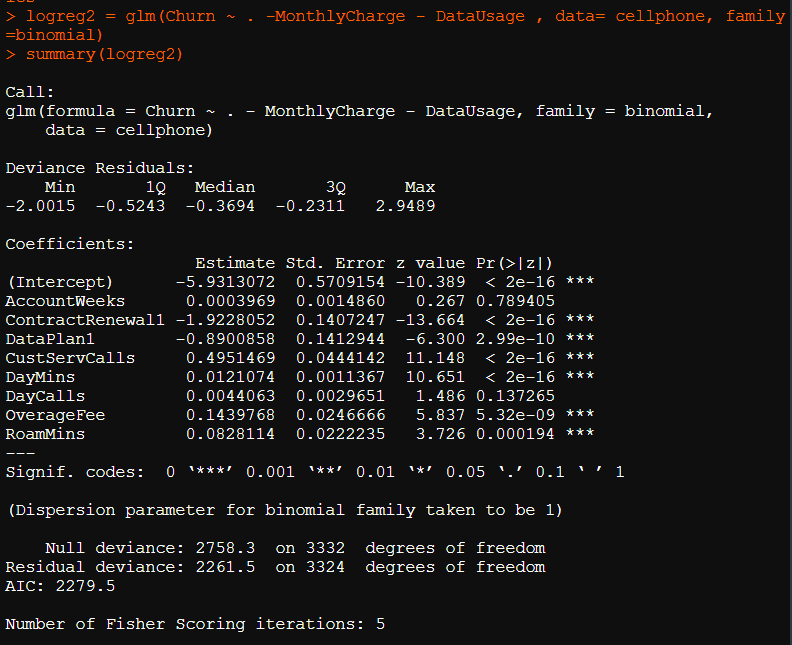
We check for Multicollinearity by using VIF function. Here are the results.



The multicollinearity has caused the inflated VIF values for correlated variables, making the model unreliable for model building. We will exclude Data Usage and Monthly charges from our model as they are heavily correlated with other independent variables, as seen above in VIF results.

## 3.2 Logistic Regression after exclusion of correlated variables

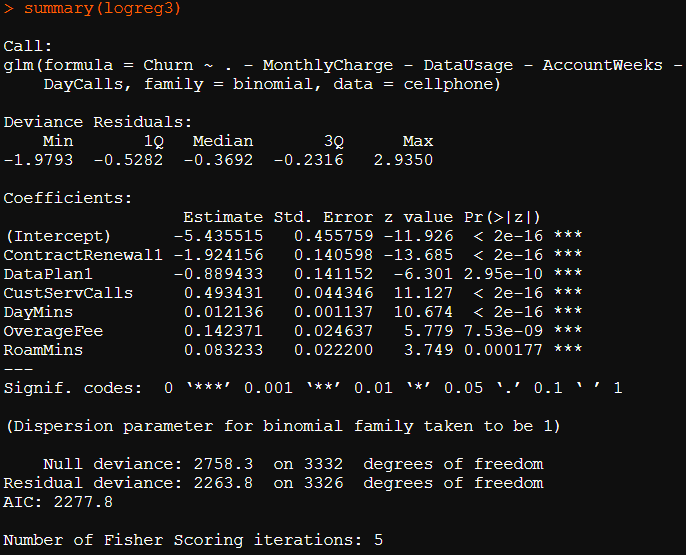
We will not use the MonthlyCharge and DataUsage variables as it is inflating the correlation result and will create a new model.

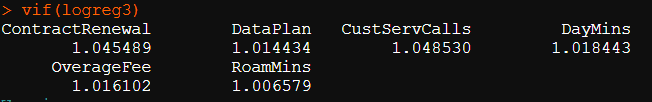


Account Weeks and Day Calls are insignificant variables for the model, as seen from above. Hence, we will be removing them as well.

## 3.3 Logistic regression model after removal of insignificant variables

Next, we move on to removing the insignificant variables and fitting the logit model again on the new dataset.



Now based on this new built model we can see that all values are significant, and we can verify the same by checking the multicollinearity as well.

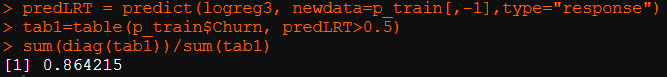
Now we can see that VIF values are within range and all variables are significant and results are making more sense and are in line with the results which we obtained from EDA.

## 3.4 Confusion matrix for final logit model

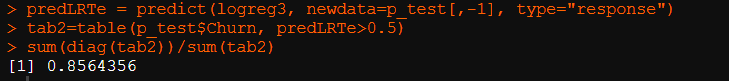
We will start model evaluation on train and test data by executing below code and will see that how accurate we were able to predict that customer will churn or not.

*Calculating Confusion Matrix on Train Data*

We are predicting classification of 0 and 1 for each row and then we are putting our actual and predicted into a table to build confusion matrix to check that how accurate our model is by executing below R Code.



*Calculating Confusion Matrix on Test Data*

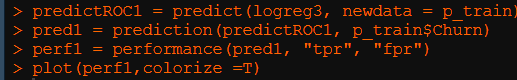


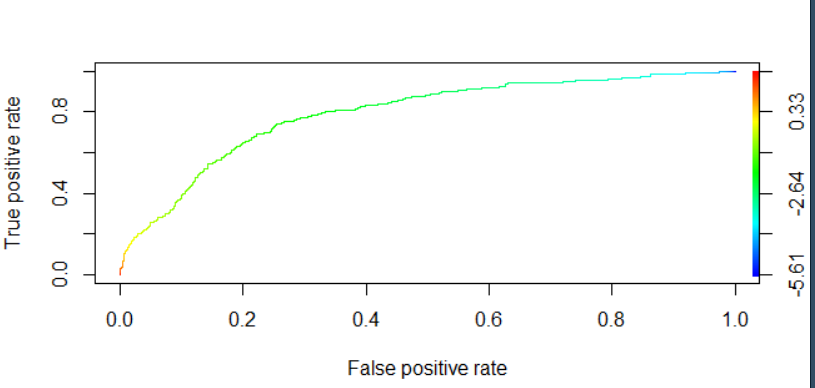
From Confusion matrix we can clearly see that our Train data is 86.42% accurate in predicting and Train data confirms the same with 85.64% of accuracy. We can see there is a slight variation but that is within the range so we can confirm that our model is good model.

## 3.5 ROC curve and AUC

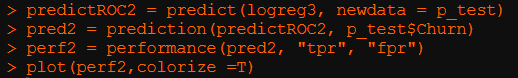
ROC curve is the plot of true positive rate (sensitivity) against false positive rate (1-Specificity) using different cutoff points.

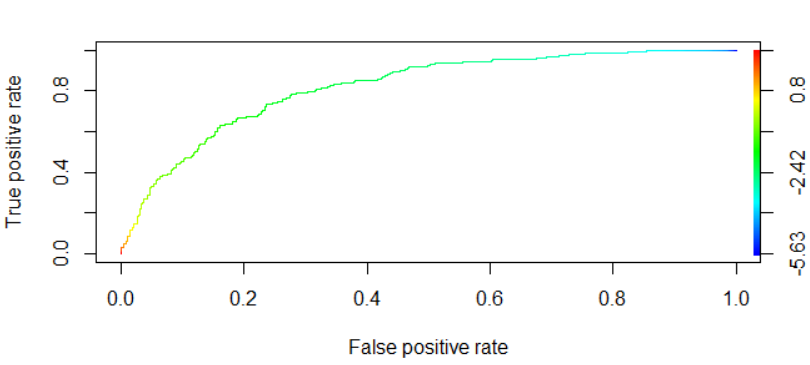
*Calculating ROC on Train Data*





*Calculating ROC on Test Data*

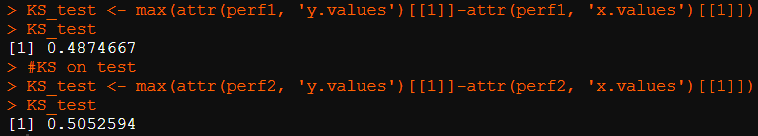




From the plot we can see that plot is covering large area under the curve and we are able to predict on the True Positive side. In Train data true positive rate is 79.06% and in test data it’s 81.82%. So, there is no major variation in our Test and Train data, and this proves that our model is more stable. These are good figures for AUC. Higher the area under curve, better the prediction power of the model.

## 3.7 K-S

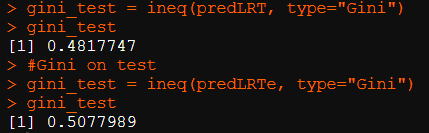
K-S will measure the degree of separation between the people likely to churn and people who will continue using services from current operator. By executing below code on Train and Test model, we will be able to see K-S Analysis result. K-S = The degree of separation between the customers cancelling and not cancelling service.



From K-S analysis we can clearly see that our Train data can distinguish between people likely to churn and people who will continue using services from current operator 48.74 on Train and 50.52% on Test accuracy. We can see there is a slight variation but that is within the range so we can confirm that our model is ok. Higher K-S value implies better model.

## 3.8 Gini

This is a coefficient derived from the ROC curve. Gini is the ratio between area between the ROC curve and the diagonal line & the area of the above triangle. Higher the Gini coefficient, better is the model.



From Gini analysis we can clearly see that our Train data covering maximum area of churn and non-churn users with 48.17% and test data with 50.77% of accuracy. We can see there is a slight variation but that is within the range so we can confirm that our model is ok.

# 4. Actionable Insights and Recommendations

1. Logistics Regression Model can predict the Customer Churn with over 82% accuracy.

2. MonthlyCharge, DataUsage, DayCalls and AccountWeeks, shows that these variables do not have significant effect on customer churn. So, these attributes can be ignored in building further models. That may give better accuracy.

3. Customer Service calls have highest correlation with customer churn. From bivariate analysis also, as the customer service calls to customers increases, the churn rate increases. This is a very useful observation that more customer service calls irritate customers and leads to churn. This area is something the company should focus to stop churn rate.

4. Contract Renewal also have high negative correlation. It is contributing to customer churn. This is also a very meaningful insight to the company to focus more on contract renewal. Customers with no contract renewal have higher churn rate. So, company should focus on providing offers to increase customers attraction towards the renewing contracts of the plan.

5. Daily Mins i.e. average daytime minutes per month is also a key feature which is contributing the customer churn. Company should focus on providing discounts and variety of plans and offers so that customers stick to the companies plan and churn rate can be reduced.

6. The Churn rate does not decrease with increase in active account weeks. Even though contracts were renewed, more than 11% customers cancelled service. Customers are more likely to churn if they do not have a data plan subscription or who have less than 0.5 GB of data usage. Churn rate is increasing when customers are making 4 or more calls to customer service or if daytime minutes are more than 240 minutes. Customers with monthly bill between 60 and 80 are more likely to churn as well.

7. Telecom Company should use focus on group of customer whose Day Mins are between 100 to 300, Overage fee is between 5 to 15 and Customer service calls are 1 to 3 for renewal of service by sending mails , calls etc to provide best results.

