Predictive Modelling

Check for multicollinearity and Building a predictive model.

[1. Introduction 2](#_Toc52131492)

[2. Why do we require a logistic model for Churn prediction? 2](#_Toc52131493)

[3. Exploratory Data Analysis 3](#_Toc52131494)

[3.1 Basic Data Summary 3](#_Toc52131495)

[3.2 Outliers check and their treatment 6](#_Toc52131496)

[3.3 Analysis of dataset 8](#_Toc52131497)

[3.3.1 Churn with AccountWeeks 8](#_Toc52131498)

[3.3.2 Churn with DataUsage 9](#_Toc52131499)

[3.3.3 Churn with DayMins 9](#_Toc52131500)

[3.3.4 Churn with DayCalls 10](#_Toc52131501)

[3.3.5 Churn with MonthlyCharge 10](#_Toc52131502)

[3.3.6 Churn with OverageFee 11](#_Toc52131503)

[3.3.8 Churn with ContractRenewal & Churn with DataPlan 11](#_Toc52131504)

[3.4 Multicollinearity Check 12](#_Toc52131505)

[4. Training and Test set 13](#_Toc52131506)

[5. Building Logistic Regression 13](#_Toc52131507)

[5.1 Logistic Regression Model - Iteration 1 13](#_Toc52131508)

[5.2 Multicollinearity effect - Iteration 1 15](#_Toc52131509)

[5.3 Logistic Regression Model - Iteration 2 (only relevant variables) 16](#_Toc52131510)

[5.4 Multicollinearity effect - Iteration 2 (only relevant variables) 18](#_Toc52131511)

[5.5 Assigning probabilities and class to the training set 18](#_Toc52131512)

[6. Model Performance Measures 20](#_Toc52131513)

[6.1 Model Performance Measures – Training Data 20](#_Toc52131514)

[6.2 Assigning scores and class to the testing set 22](#_Toc52131515)

[6.3 Model Performance Measures – Testing Data set 22](#_Toc52131516)

[6.4 Comparing model performance training and testing set. 24](#_Toc52131517)

[7. Actionable Insights and Recommendations 25](#_Toc52131518)

# Introduction

A sample dataset of postpaid cellphone customers was provided. The data has information about customer usage behaviour, contract details, and payment details. The data also indicates which were the customers who cancelled their service. Based on this past data, the assignment requires to build a model which can predict whether a customer will cancel their service in the future or not using logistic regression techniques. The assignment also requires us to provide some recommendations and actionable insights based on the prediction.

The R codes in the document are given in a grey box.

The R output in the document is given in a white box.

# Why do we require a logistic model for Churn prediction?

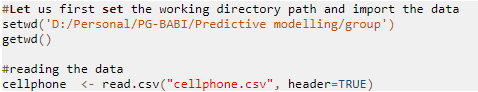
The business problem we are trying to solve here is predicting if a customer will discontinue the service and switch to another operator. An early indication of such customers will help in the mobile company taking preventive actions. A logistic model is useful to predict a resultant variable which are binary in nature here, whether the customer will continue the service or switch to another one.

# Exploratory Data Analysis

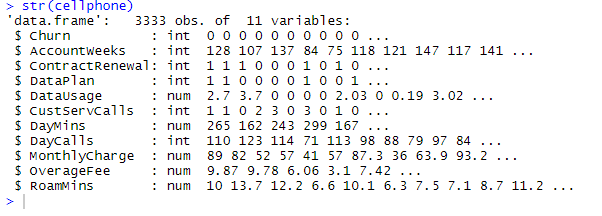
## 3.1 Basic Data Summary

We begin with understanding the data and summary statistics.

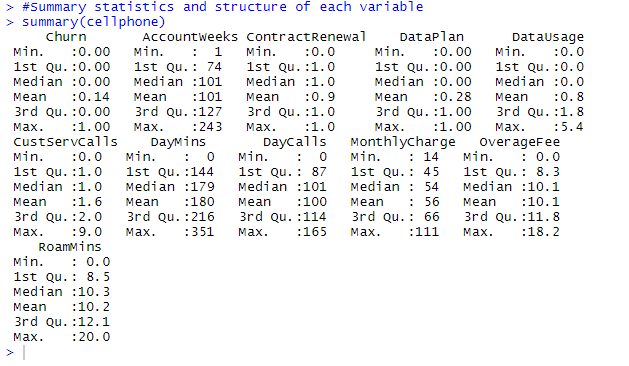
Following code is executed to load the dataset.



To understand the data and its structure, the following command is executed in R which returns the summary descriptive statistics and structure of each variable in the dataset.





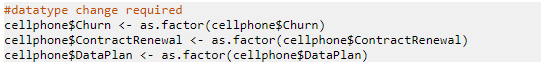


The dataset consists of 3333 observations and 11 variables. The dataset includes customer’s contract details, their usage as well as payment details.

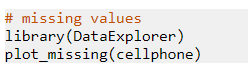
From the min-max values for these variables in the summary of descriptive statistics, one can understand the following:

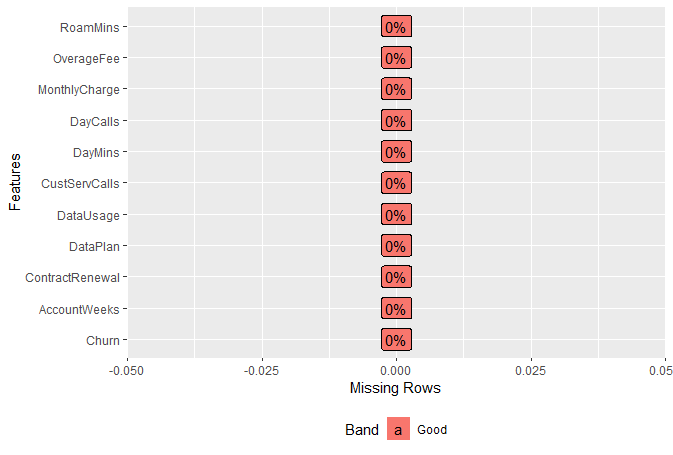
* The variables Churn, Contract Renewal and Data Plan are binary variables, hence they need to be converted to factor variables.
* Whereas the DayMins, DayCalls, MonthlyCharge, Overage Fee and RoamMins are averaged figures over a period of time.
* Churn = 1 is an indicator for the customers who have cancelled the service.

Let us convert the 3 variables to factor variables.



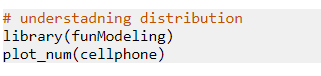
Finding out missing values is an essential part of EDA, if there are some then those need to be treated.

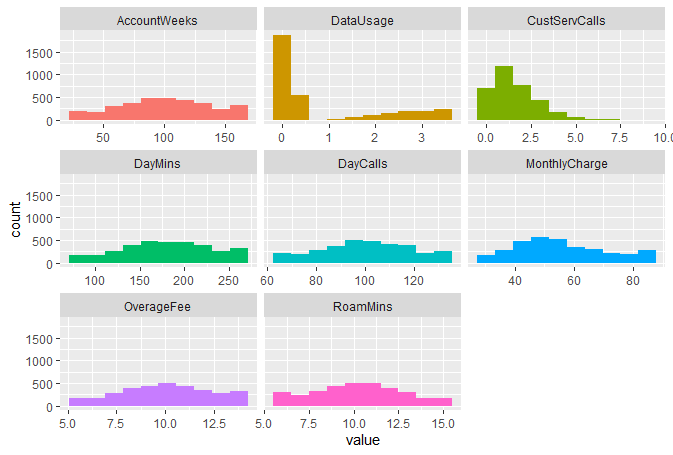




From the figure above and also from summary stats, we see that there are no missing values.

Let’s seed how numeric data are distributed.

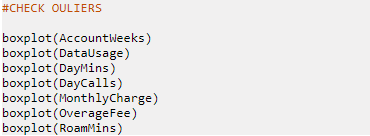




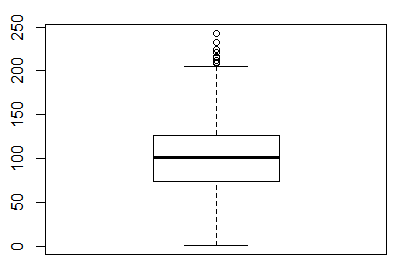
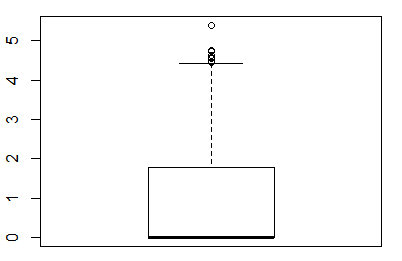
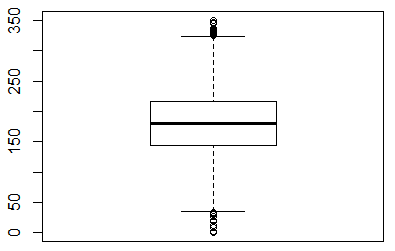
The plot above shows that apart from Data Usage and Customer Service calls, other variables follow a normal distribution.

## 3.2 Outliers check and their treatment

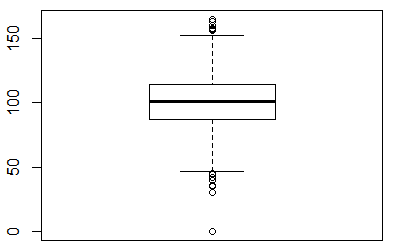
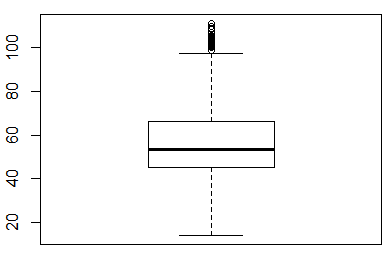
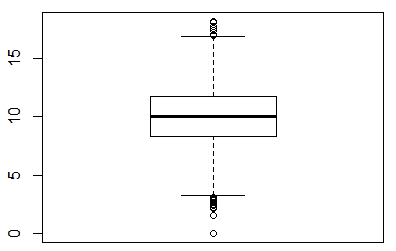
In order to check for outliers, we create a boxplot of each of the variables.



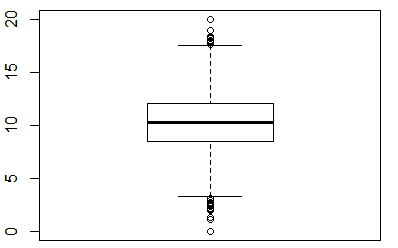
Boxplot output for each variable is as shown below:

AccountWeeks DataUsage DayMins

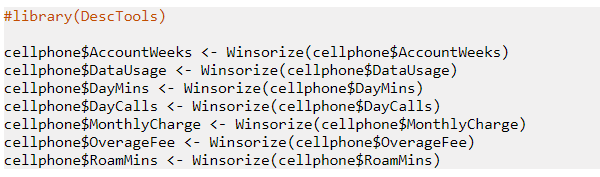
  

DayCalls MonthlyCharge OverageFee

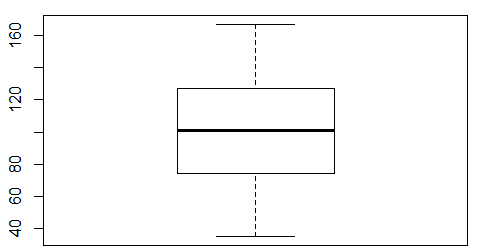
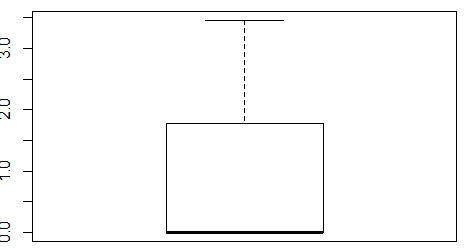
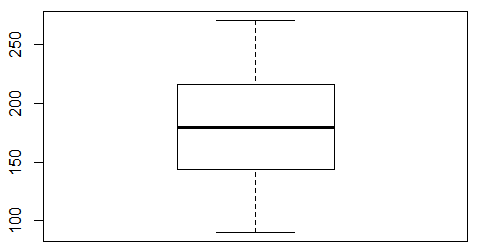


RoamMins

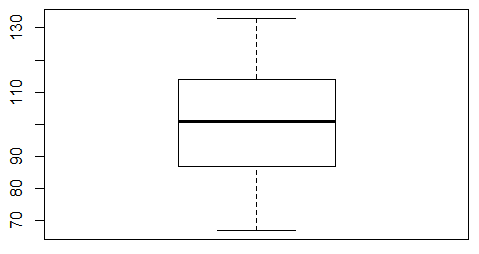
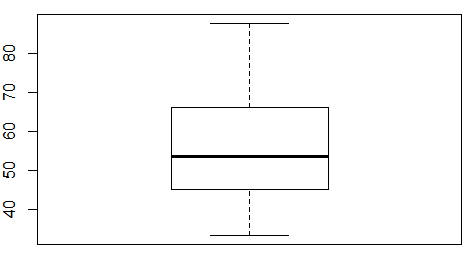
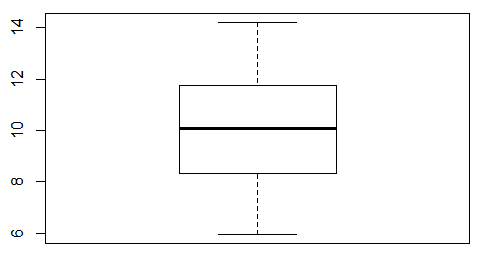
It is observed from the boxplot that all variables have outliers. To remove outliers, we use Winsorize() function in R which replaces the extreme values by the less extreme values. The default parameters replace for 5% percentile values for low and 95% percentile for high values respectively.



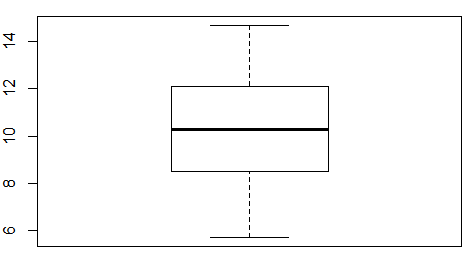
We plot the boxplot again to verify that the outliers have been actually treated.

AccountWeeks DataUsage DayMins

DayCalls MonthlyCharge OverageFee



RoamMins

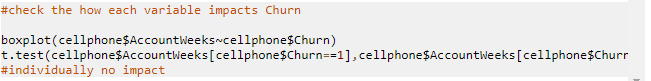
Thus, we have successfully treated the outliers in the dataset.

## 3.3 Analysis of dataset

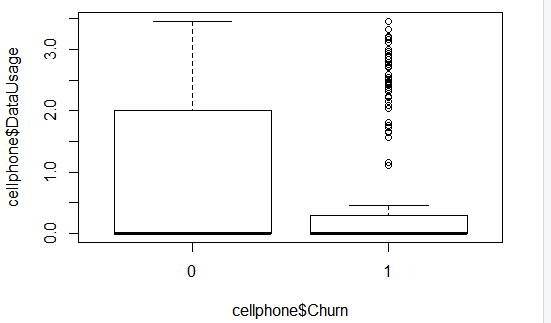
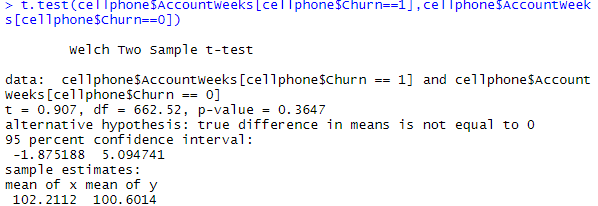
For further analysis, we require to check how each variable impacts the Churn.

### 3.3.1 Churn with AccountWeeks

R code:



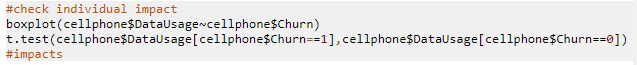
Boxplot and t-test output:

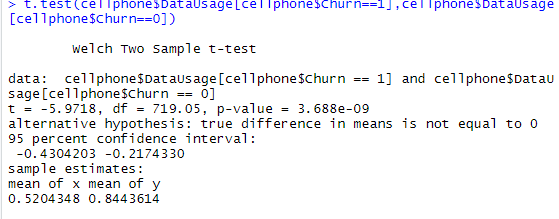
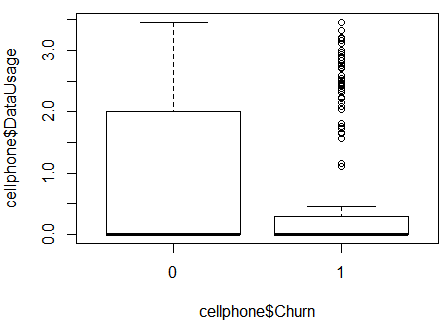
 

From the t-test, we can see that the p-value is 0.3647 is more than alpha of 5%, which means that the null hypothesis cannot be rejected and we can conclude that “Account Weeks” does not have an impact on “Churn”. This can also be seen from the boxplot.

Similarly, we will continue hypothesis testing on other independent variables.

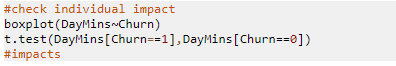
### 3.3.2 Churn with DataUsage

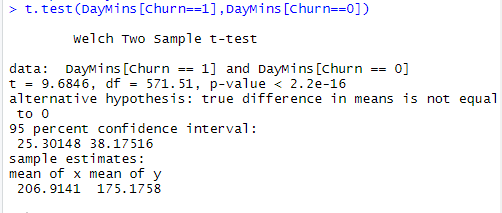




The p-value from the t-test is significantly less than alpha of 5%. Hence we can conclude the monthly data usage of the customer has significant impact on whether he continues the service or not.

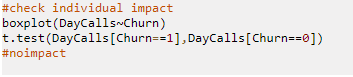
### 3.3.3 Churn with DayMins

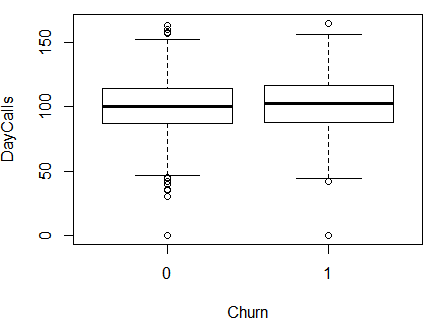
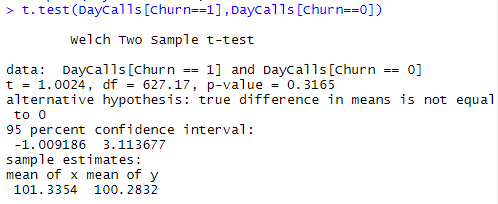




From the p-value, we observe that the daytime minutes of usage also has a significant impact on customer churn.

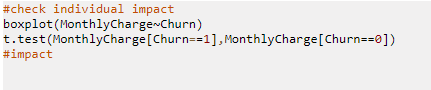
### 3.3.4 Churn with DayCalls

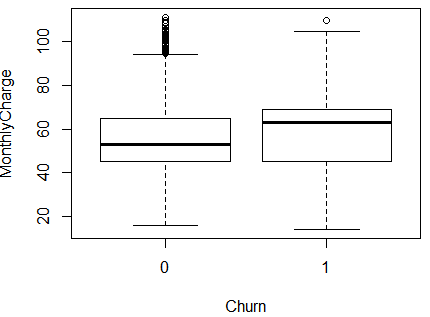
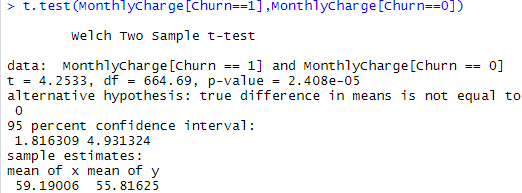


P-value of t-test is within the 95 percent confidence interval, hence this variable does not affect the customer churn individually.

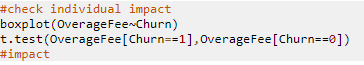
### 3.3.5 Churn with MonthlyCharge

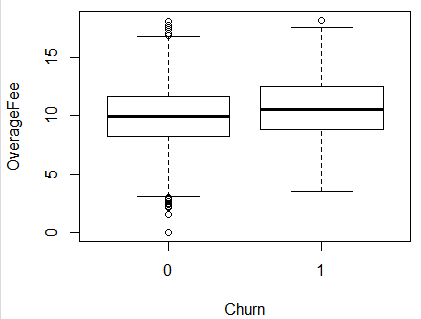
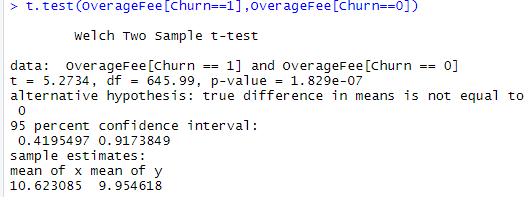


Here, we observe that the monthly bill of the customer has a direct impact on customer churn.

### 3.3.6 Churn with OverageFee

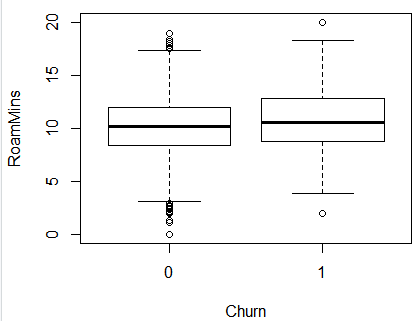
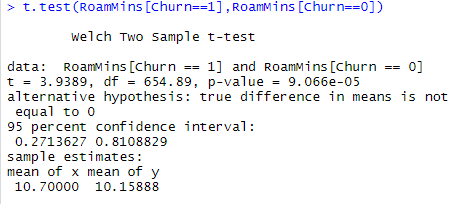


It is observed that Overage fee captured for last 12 months also affects the customer churn.

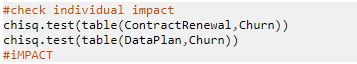
3.3.7 Churn with RoamMins

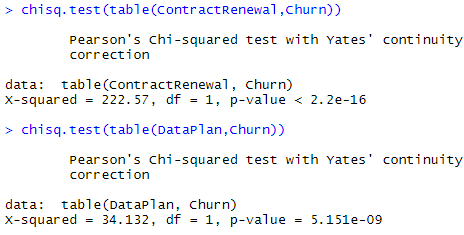


Due to low p-value of the t-test, we can conclude that this variable also has a significant effect over the customer churn.

### 3.3.8 Churn with ContractRenewal & Churn with DataPlan





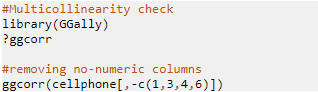
We have performed chi-square test for the 2 categorical variables which suggests that both these variables have a significant impact on whether the customer will cancel the service or not.

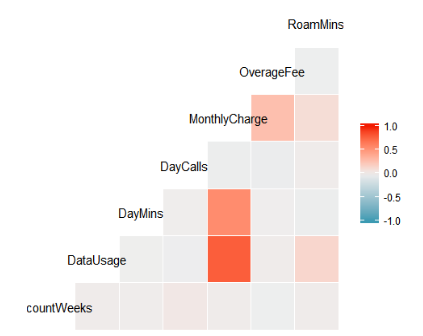
## 3.4 Multicollinearity Check

The correlation amongst variables can be checked only for numeric variables, hence we remove these and plot a correlation plot

These variables are Churn, ContractRenewal, Data Plan and Data usage.

Plotting the correlation:



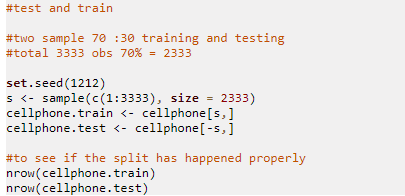


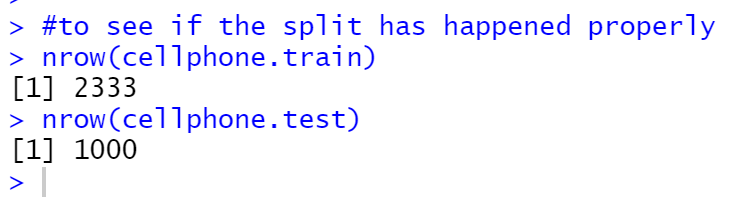
It can be observed from the correlation matrix that Data usage and DayMins are correlated with the Monthly Charge, which is quite obvious as more usage, more will be the monthly charge.

# Training and Test set

Before we perform the logistic regression, let us split the dataset into training and testing data with 70:30 ratio, that is, 2333 observations for the training sample.

Following R code shows the splitting of data and to check if it is split properly.





# Building Logistic Regression

Logistic regression is applied to predict the categorical variables, here, Customer churn.

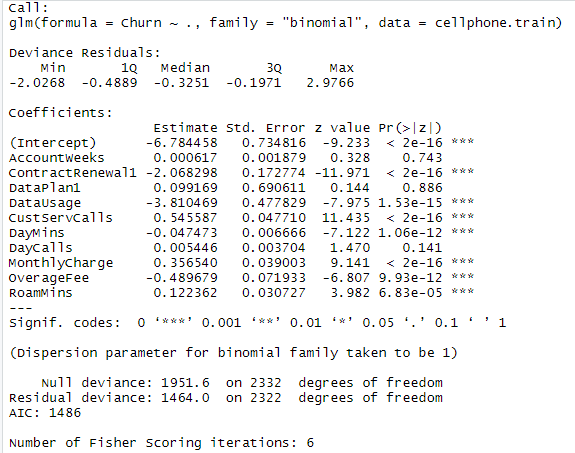
It will help us in understanding which are the variables that impact the churn and how.

## 5.1 Logistic Regression Model - Iteration 1

Let us perform logistic regression using all the independent variables on training set.



The summary of model.1 that we have built is as follows:



Following interpretations can be made from the above summary:

* Below mentioned variables are not that significant for predicting the customer churn as their p-value less than the alpha of 5%.
  + AccountWeeks
  + DataPlan
  + DayCalls
* The prediction equation can be formed as:

Logit (Churn) =

e -6.784458 + 0.000617(x1) -2.068298(x2) + 0.099169(x3) -3.810469(x4 )+ 0.545587(x5) -0.047473(x6) +0.005446(x7) +0.356540(x8) --0.489679(x9) +0.122362(x10)

1+ e -6.784458 + 0.000617(x1) -2.068298(x2) + 0.099169(x3) -3.810469(x4 )+ 0.545587(x5) -0.047473(x6) +0.005446(x7) +0.356540(x8) --0.489679(x9) +0.122362(x10)

Where,

X1 = AccountWeeks

X2 = ContractRenewal if value is “1”

X3 = DataPlan if value is “1”

X4 = DataUsage

X5 = CustServCalls

X6 = DayMins

x7 = DayCalls

X8 = MonthlyCharge

X9 = OverageFee

X10 = RoamMins

Let’s interpret the equation

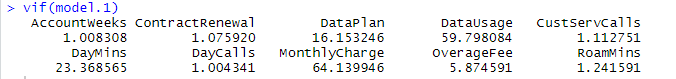
* + If the customer has renewed the contract (indicated by 1) the odds for Churn (cancelling the service) decreases by 2.068298 times (i.e 67% probability) keeping other variables at the same level.
  + If the Data Usage increases by one unit, the odds for Churn (cancelling the service) decreases by 3.810469 times (i.e 79% probability) keeping other variables at the same level.
  + Likewise, we can interpret the equation for other variables.
  + We observe that DataUsage, DayMins and OverageFee influence customer churn negatively, which means more the usage and no.of contract renewals, less is the chance that the customer will cancel the service.
  + Whereas, higher the customer service calls and the monthly charge, greater is the chance that the customer will move to another service provider.
* We have a null deviance of 1951.6 on 2332 dof. Including the independent variables, it is decreased to 1464 on 2322 dof, which is a significant reduction in deviance.

## 5.2 Multicollinearity effect - Iteration 1

From the correlation plot, we know that few variables have medium -to high correlation. For a regression model multicollinearity, independent variables makes the model unstable.

Let’s check “Variation Inflation Factor “ (VIF) of each dependent variable to see how large its standard error of the estimated coefficient is with respect to other dependent variable.





MonthlyCharge has a relatively high VIF compared to other variables.

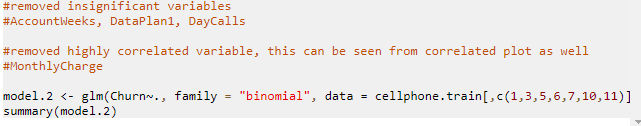
## 5.3 Logistic Regression Model - Iteration 2 (only relevant variables)

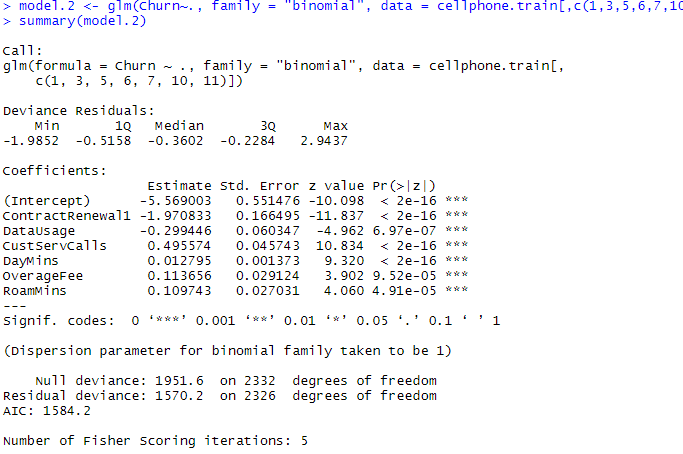
From the exploratory data analysis, the first logistic regression model and VIF output we have observed that some variables do not impact Churn, some are insignificant to predict Churn, there is multicollinearity effect which may render the model unstable.

The following variables will be removed in the next logistic regression model.

|  |  |
| --- | --- |
| Variables Removed | Reason |
| AccountWeeks | Insignificant Variable in Model 1, failed to reject t-test null hypothesis |
| DataPlan | Insignificant Variable in Model 1 |
| DayCalls | Insignificant Variable in Model 1, failed to reject t-test null hypothesis |
| MonthlyCharge | High VIF, and correlated with Data usage and DayMins |

Only the following relevant variables will be retained in our next model





Following interpretations can be made from the above summary:

* All variables are important in this logistic regression model
* The prediction equation can be formed as:

Logit (Churn) =

e -5.569003 -1.970833(x1) -0.299446(x2) + 0.495574(x3) + 0.012795(x4 )+0.113656(x5) +0.109743(x6)

1+ e -5.569003 -1.970833(x1) -0.299446(x2) + 0.495574(x3) + 0.012795(x4 )+0.113656(x5) +0.109743(x6)

Where

X1 = ContractRenewal if value is “1”

X2 = DataUsage

X3 = CustServCalls

X4 = DayMins

X5 = OverageFee

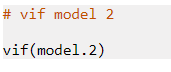
X6 = RoamMins

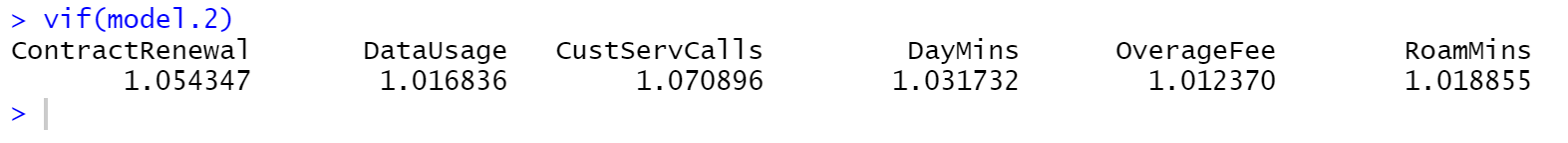
Let’s interpret the equation

* + If the customer has renewed the contract (indicated by 1) the odds for Churn (cancelling the service) decreases by 1.97 times (i.e 66% probability) keeping other variables at the same level.
  + If the Customer Service Call increases by one unit, the odds for Churn (cancelling the service) increases by 0.495574 times (i.e 33% probability) keeping other variables at the same level.
  + Likewise, we can interpret the equation for other variables.
  + We observe that DataUsage, influences customer churn negatively, which means more the usage less is the chance that the customer will cancel the service.
  + Whereas, higher the customer service calls, day minutes, monthly charge and roaming minutes greater is the chance that the customer will move to another service provider.
* We have a null deviance of 1951.6 on 2332 dof. Including the independent variables, it is decreased to 1570 on 2326 dof, which is a significant reduction in deviance.

## 5.4 Multicollinearity effect - Iteration 2 (only relevant variables)

Let’s check multicollinearity of this model by calculating “Variation Inflation Factor “ (VIF) for model 2





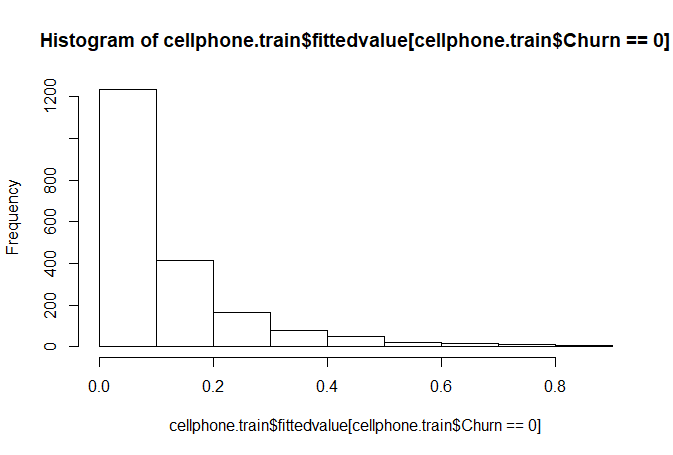
We observe that VIF for all variables is small, hence we can conclude that model 2 does not have multicollinearity effect.

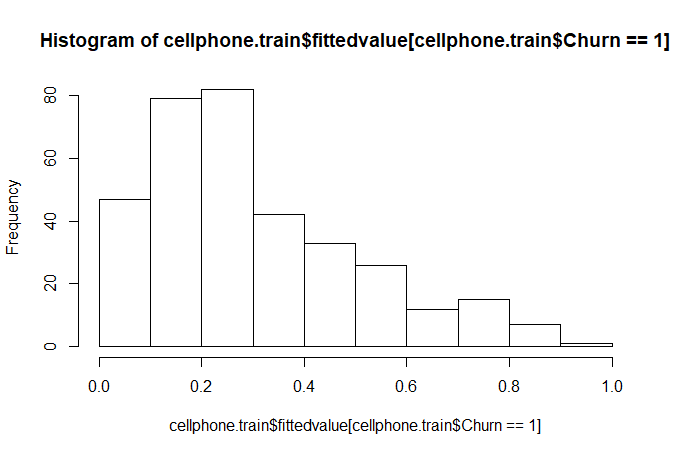
## 5.5 Assigning probabilities and class to the training set

Based on the logistic regression model worked initiation 2 let us calculate probabilities for the training set

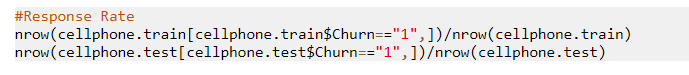


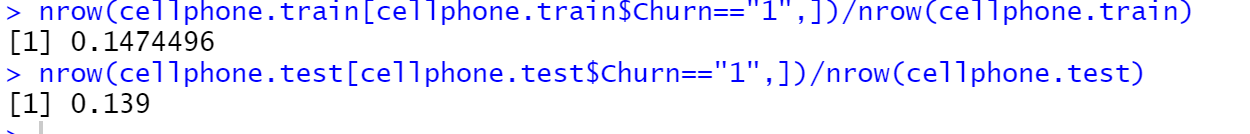
The next step will be to assign class based on the threshold, let us first understand how is the estimated probability distribution for target variable .



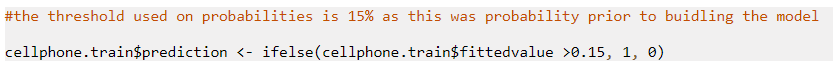


We observe from the plot that probabilities is largely less than 0.1 for Churn = 0, and greater than 0.1 for Churn = 1

Let’s calculate prior probabilities and use it as a threshold for arriving at the clas



We observe the prior probability for both training and test data is approximately 14%. We can use 0.15 as threshold for arriving at the class as this was response rate prior to the building model.

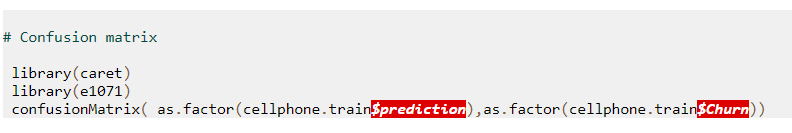


# Model Performance Measures

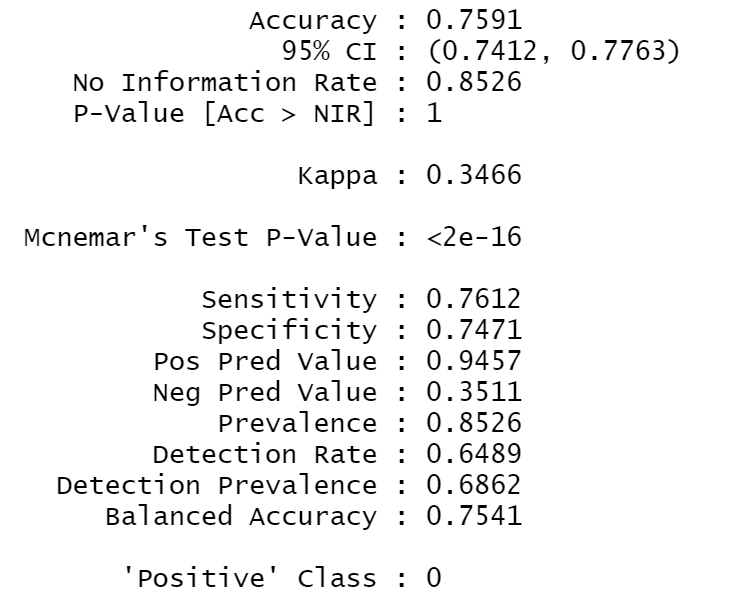
The logistic regression model built by retaining only relevant variables seems to be a good start, we would calculate the model performance measures for that model first on training set then on the testing set.

## 6.1 Model Performance Measures – Training Data

Confusion matrix will be used as the primary model performance measure. We will load the required libraries.

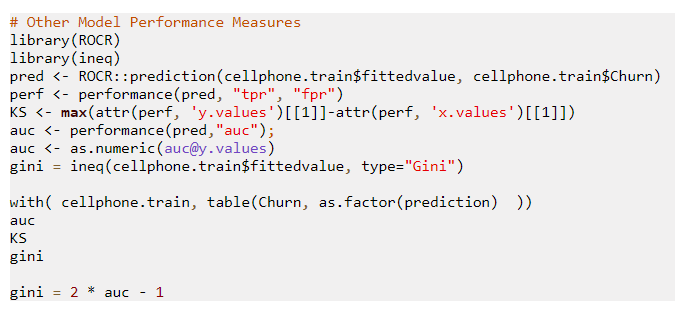


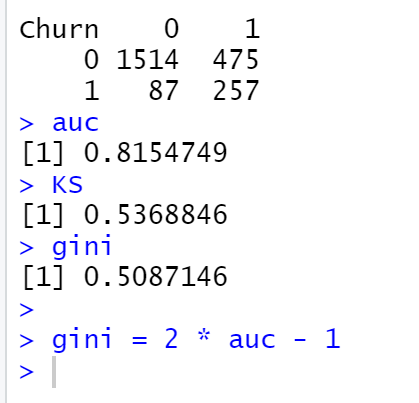
Confusion Matrix and Statistics



The accuracy of the trained dataset is 76% which seems to be fair. The sensitivity and specificity is 76% and 75% respectively. We will test another performance measure as well.

Accuracy of the model is the predictive power of the model, the higher the number better is the predictive power.





AUC stands for Area under the curve. AUC gives the rate of successful classification by the logistic model, which is 82 %. Higher the better.

K-S is a measure of the degree of separation between the positive and negative distributions, here it is 54%. The higher the value the better the model is at separating the positive from negative cases. KS value of more than 40% is considered to be a good model.

Gini coefficient of 50% indicated there is inequality in the data.

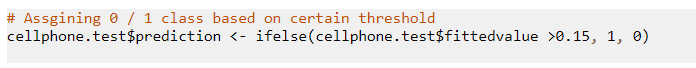
All the model performance measure indicates that the model is a fair model. The next step is to see how they fare in the testing model.

## 6.2 Assigning scores and class to the testing set

To predict the scores of the testing data set, we would use the predict function on testing data set.

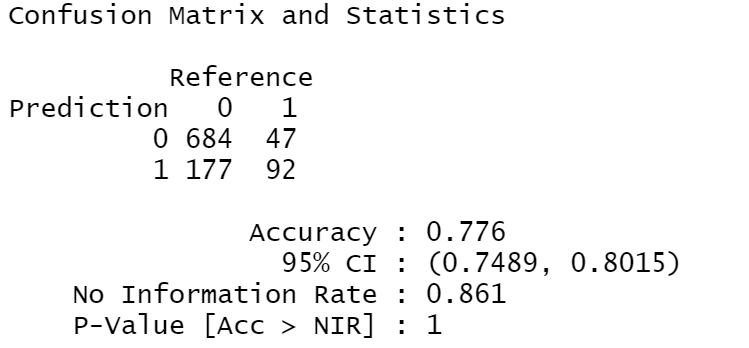


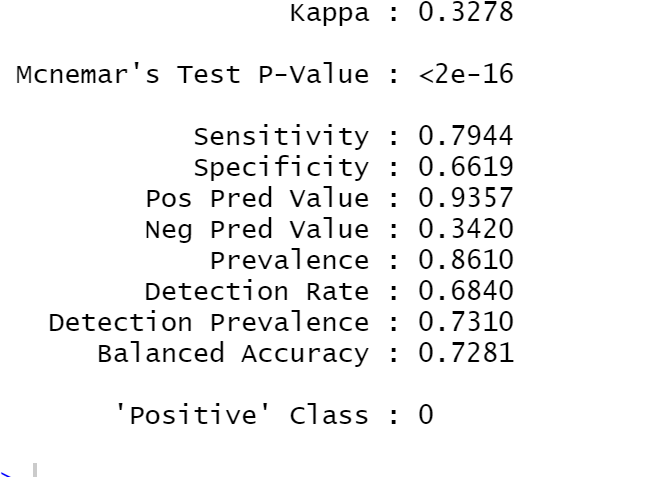
Assign class with the same threshold as taken for the training set.



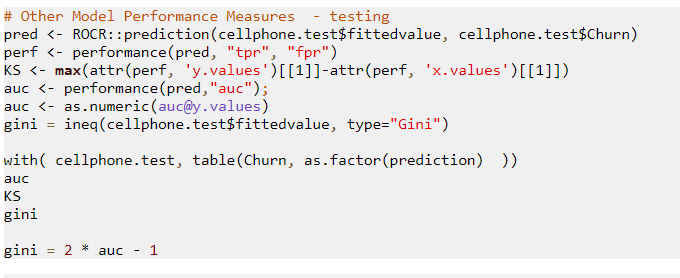
### 6.3 Model Performance Measures – Testing Data set

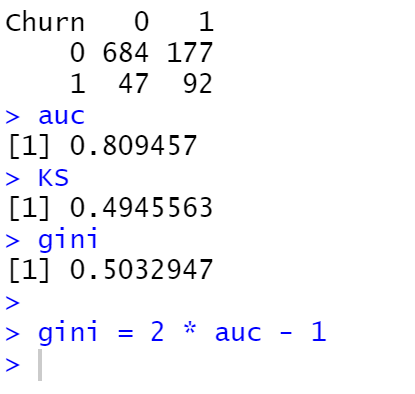
Confusion Matrix 





The accuracy of the testing dataset is 78% which is similar to the training data set. We will test another performance measure as well.





The area under the Curve (AUC) is 80% KS is 49% and gini coefficient is 0.5 seems to be fair for this model on the training set and comparable to training model.

### 6.4 Comparing model performance training and testing set.

The following table gives a comparative matrix on model performance measure of training and testing data set.

|  |  |  |
| --- | --- | --- |
| Performance Metric | Training Data | Testing Data |
| Accuracy Classification Score | 0.76 | 0.78 |
| Area Under the Receiver Operating Characteristic Curve | 0.82 | 0.80 |
| KS | 0.54 | 0.49 |
| Gini Coefficient | 0.50 | 0.50 |

The model performance measure of testing data set as compared to the training data set is similar, which seem to suggest it is s good model and has not overfitted.

# Actionable Insights and Recommendations

The recommendation to the mobile company will be to use the model to target customers that are predicted with a high probability of cancelling the service. Such high-risk customers can be offered incentives by providing an alternative data plan so that they avoid switching to other operators. If churn rate is still high, check for differences with competitors pricing and schemes and remodel their mobile plans.

The company’s bandwidth to work on flagged accounts, drives the threshold applied to the predicted probability which classifies a customer as high risk. Assigning a higher threshold leads to better accuracy in prediction.

The mobile company should filter mobile plans based on the requirement of each customer and should increase subsidiaries to reduce the churn rate. It should also provide its customers higher value added services.

Logistic model focuses on identifying high risk customers more accurately Thus focusing on these customers could help reduce the churn rate.