Mini Project - Telecom Customer Churn

Name : Numer P

Table of Contents

1. Project Objective

The main objective of the report is to explore the Cell Phone Dataset (“Cellphone.xlsx”) and in R and generate insights about the data set. This exploration report will consist of the following,

- ❖ Importing dataset in R
- ❖ Understanding the structure of Dataset
- ❖ Graphical exploration
- ❖ Descriptive Statistics
- ❖ Predictive Modelling

2. Assumptions

The Customer Churn is performed under the predictive modelling for the better understanding of the customers satisfaction to continue the services of the Telecom Brand. The Dataset is explored with the customer usage on Data and the daily calls made by the customer and predicted the customer retention for the telecom company.

The Dataset is explored in the numeric data values for the Day calls, Data Usage, Customer Service received calls from the customer, daily mins of using the churn, weeks used by the customer and the category variables helps to understand the customer cancelled the service and the customer are not availing the data and the customer renewed their service. The predictive modelling shows the customer retention and the telecom company takes decision to whether provide offers or discounts or price cut-off for the customer to existing the same network usage.

3. Exploratory Data Analysis – Step by Step Approach

1. Environment Set up and Data Import
2. Variable Identification
3. Univariate Analysis
4. Bivariate Analysis
5. Variable Transformation
6. Feature Exploration

3.1 Environment Setup and Data Import

3.1.1 Install necessary packages and Import Libraries

This section is used to install packages and invoke the associated libraries. Having all packages at the same places increase code readability.

3.1.2 Setup Working Directory

Setting a working directory on starting of the R session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes etc. related to the project.

Please refer Appendix A for source code.

3.1.3 Import and read the dataset

The given dataset is in .xlsx format. Hence the command 'read_xlsx' is used for import the dataset.

Please refer Appendix A for source code.

3.2 Variable Identification

- setwd() used for setup working directory to export data and files from the folder or location in PC.
- getwd() used to identify the location was correctly entered or not.
- Library function is used to load the installed packages like ggplot2, dplyr, rpivotTable, readxl, readr, mice, car, cardata, psych, lattice, nfactors, data.table, tidyverse, broom, ggally, ROCR, pROC, caret, class, MASS, pscl, e1071, blorr.
- read_xlsx function is used to load the excel files in the path.
- attach function is used to reduce the reusability of variable name to enter each time.
- str function is used to check the category variables formats.
- summary function is used to get the summarised value like length, class and basic statistics values with quartile ranges.
- dim function is used to find the total observations and variables.
- Split function is used to split the data into train and test.
- Predict function is used to predict the train dataset and the validation dataset.
- glm is used for the predictive modelling for the logistic regression.
- Loglik is the function used to predict the likelihood of the predicted dataset.

3.2.1 Variable Identification – Inferences

#getwd()

It shows the working directory dataset

#library(mice)

(mice)takes out cbind and rbind the variables of two different datasets.

#library(readr)

(readr) helps in reading the rectangular datasets while the datasets are in table format.

#library(readxl)

(readxl) helps in reading the files in excel formats.

#library(dplyr)

(dplyr) helps in filter the datasets and intersect the datasets.

#library(ggplot2)

(ggplot2) helps is visualise the datasets in boxplot, histogram and graphical representation.

#read_csv

csv file was imported from the path and shows the variables

#str

str shows the variables along with class of the data. It shows some samples to understand the data.

#class

class function describes the full file in data.frame format. As the files includes category in season variables, it shows the values as character format.

#attach

Variable is attached to reduce the reusability in following syntax.

#dim

It shows the number observations and the variables associated in the file.

#summary

It produces the results as summarised format for each variable.

#names

It gives the column names from the dataset.

#boxplot

Boxplot graph is used for the dataset to find the outliers and the values are grouped.

#hist

This function is used to plot the histogram for the variables.

#glm

It is used for the generalized linear models and specifies the linear prediction for the variables and its error.

#vif

The variance inflation factor function is used for the calculation of variance in the linear models.

#lrtest

The generic function used for the prediction for the likelihood of the generalized linear models.

#loglik

The function is used for the likelihood prediction of the variables.

#predict

The predict function is used for the fitting the model function used for the datasets.

#table

The table is used for the cross classifying the variables in the contingency tables.

#cor

The correlation function is used to understand the correlation between the variables.

#naivebayes

This function is used for computing the posterior probability of the categorical variables with the predictor numeric variables.

#knn

The K Nearest Neighbour function is computed using the Euclidean distance of the variables and vectors in the train dataset using the test set values.

#blr_step_aic_both

The function is used to predict the vectors with building the variables for the information based on the categories in the datasets.

#blr_rsq_mcfadden

The function is used to find the pseudo r squared value for the datasets.

#confusionmatrix

Calculated for the cross validation in the observed and predicted values from the models.

3.3 Univariate Analysis

Univariate analysis is the analysis of data of one variable at time and it involves whether the datasets are descriptive or inferential statistics.

1. How does the data looks like, Univariate and bivariate analysis. Plots and charts which illustrate the relationships between variables.

```
> Cellphone=read_xlsx("Cellphone.xlsx",sheet = 2,col_names = TRUE)
```

```
> head(Cellphone,5)
```

# A tibble: 5 x 11									
	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	128	1	1	2.7	1	265	110	89
2	0	107	1	1	3.7	1	162	123	82
3	0	137	1	0	0	0	243	114	52
4	0	84	0	0	0	2	299	71	57
5	0	75	0	0	0	3	167	113	41
# … with 2 more variables: OverageFee <dbl>, RoamMins <dbl>									

Cell Phone dataset is taken into the EDA and the variables are displayed with the first row values.

```
> tail(Cellphone,3)
```

# A tibble: 3 x 11									
	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	28	1	0	0	2	181	109	56
2	0	184	0	0	0	2	214	105	50
3	0	74	1	1	3.7	0	234	113	100
# … with 2 more variables: OverageFee <dbl>, RoamMins <dbl>									

The tail dataset is displayed with the three rows of the variables.

> summary(Cellphone)

Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls
Min. :0.0000	Min. : 1.0	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.000
1st Qu.:0.0000	1st Qu.: 74.0	1st Qu.:1.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:1.000
Median :0.0000	Median :101.0	Median :1.0000	Median :0.0000	Median :0.0000	Median :1.000
Mean :0.1449	Mean :101.1	Mean :0.9031	Mean :0.2766	Mean :0.8165	Mean :1.563
3rd Qu.:0.0000	3rd Qu.:127.0	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.7800	3rd Qu.:2.000
Max. :1.0000	Max. :243.0	Max. :1.0000	Max. :1.0000	Max. :5.4000	Max. :9.000

DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
Min. : 0.0	Min. : 0.0	Min. : 14.00	Min. : 0.00	Min. : 0.00
1st Qu.:143.7	1st Qu.: 87.0	1st Qu.: 45.00	1st Qu.: 8.33	1st Qu.: 8.50
Median :179.4	Median :101.0	Median : 53.50	Median :10.07	Median :10.30
Mean :179.8	Mean :100.4	Mean : 56.31	Mean :10.05	Mean :10.24
3rd Qu.:216.4	3rd Qu.:114.0	3rd Qu.: 66.20	3rd Qu.:11.77	3rd Qu.:12.10
Max. :350.8	Max. :165.0	Max. :111.30	Max. :18.19	Max. :20.00

The summary of the dataset is explained with no missing values are in the variables.

> str(Cellphone)

Classes ‘tbl_df’, ‘tbl’ and 'data.frame': 3333 obs. of 11 variables:
$ Churn : num 0 0 0 0 0 0 0 0 0 0 ...
$ AccountWeeks : num 128 107 137 84 75 118 121 147 117 141 ...
$ ContractRenewal: num 1 1 1 0 0 0 1 0 1 0 ...
$ DataPlan : num 1 1 0 0 0 0 1 0 0 1 ...
$ DataUsage : num 2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...
$ CustServCalls : num 1 1 0 2 3 0 3 0 1 0 ...
$ DayMins : num 265 162 243 299 167 ...
$ DayCalls : num 110 123 114 71 113 98 88 79 97 84 ...
$ MonthlyCharge : num 89 82 52 57 41 57 87.3 36 63.9 93.2 ...
$ OverageFee : num 9.87 9.78 6.06 3.1 7.42 ...
$ RoamMins : num 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...

In which, Churn, Contract Renewal and Data Plan variables are categorized with the customer continues the service, customer continued contract, customer have data respectively and it is vice versa. The other variables are numeric data which helps to find the prediction of the customers using the telecom service.

```
> attach(Cellphone)
> View(Cellphone)
> Cellphone$Churn=factor(Cellphone$Churn,levels = c(0,1),
+                   labels = c("Customer Using Service","Customer Cancelled Service"))
> Cellphone$ContractRenewal=factor(Cellphone$ContractRenewal,levels = c(0,1),
+                   labels = c("Customer not renewed","Customer renewed"))
> Cellphone$DataPlan=factor(Cellphone$DataPlan,levels = c(0,1),
+                          labels = c("Customer with no data","Customer have data"))
```

The vector is attached for the future reference and the dataset is displayed with the all the variables.

The variables churn, contract renewal and data plan are changed to factor variables as they are containing the levels like 0, 1. The factor variables are used to predict the numeric data in the models. The values are created and separated for the values based on the category variables and numeric variables.

```
> summary(Cellphone)
```

Churn	ContractRenewal	DataPlan
Customer Using Service :2850	Customer not renewed: 323	Customer with no data:2411
Customer Cancelled Service: 483	Customer renewed :3010	Customer have data : 922

DataUsage	CustServCalls	DayMins	DayCalls	AccountWeeks
Min. :0.0000	Min. :0.000	Min. : 0.0	Min. : 0.0	Min. : 1.0
1st Qu.:0.0000	1st Qu.:1.000	1st Qu.:143.7	1st Qu.: 87.0	1st Qu.: 74.0
Median :0.0000	Median :1.000	Median :179.4	Median :101.0	Median :101.0
Mean :0.8165	Mean :1.563	Mean :179.8	Mean :100.4	Mean :101.1
3rd Qu.:1.7800	3rd Qu.:2.000	3rd Qu.:216.4	3rd Qu.:114.0	3rd Qu.:127.0
Max. :5.4000	Max. :9.000	Max. :350.8	Max. :165.0	Max. :243.0

MonthlyCharge	OverageFee	RoamMins
Min. : 14.00	Min. : 0.00	Min. : 0.00
1st Qu.: 45.00	1st Qu.: 8.33	1st Qu.: 8.50
Median : 53.50	Median :10.07	Median :10.30
Mean : 56.31	Mean :10.05	Mean :10.24
3rd Qu.: 66.20	3rd Qu.:11.77	3rd Qu.:12.10
Max. :111.30	Max. :18.19	Max. :20.00

The variables are created with their values. The category variables are created with factor values and the numeric variables are used to predict the models using the category variables.

```
> str(Cellphone)
```

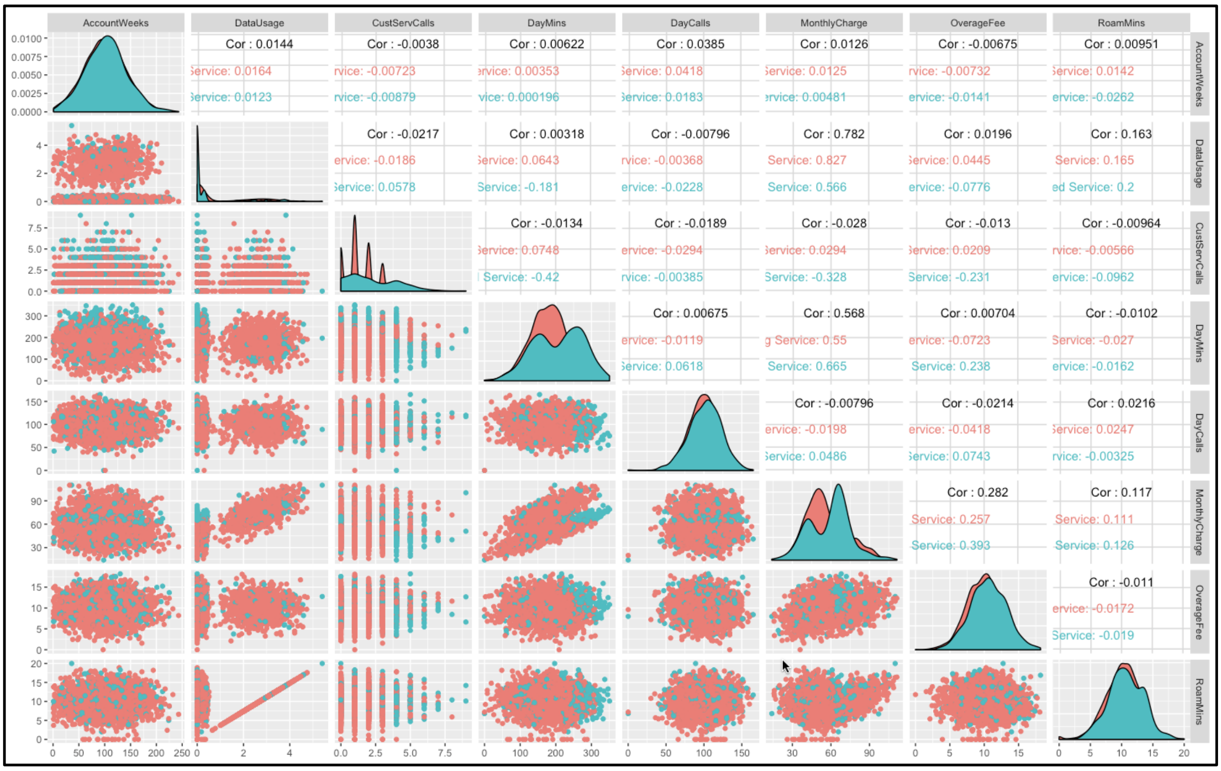
```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':3333 obs. of  11 variables:
 $ Churn          : Factor w/ 2 levels "Customer Using Service",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ AccountWeeks   : num  128 107 137 84 75 118 121 147 117 141 ...
 $ ContractRenewal: Factor w/ 2 levels "Customer not renewed",..: 2 2 2 1 1 1 2 1 2 1 ...
 $ DataPlan       : Factor w/ 2 levels "Customer with no data",..: 2 2 1 1 1 1 2 1 1 2 ...
 $ DataUsage      : num  2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...
 $ CustServCalls  : num  1 1 0 2 3 0 3 0 1 0 ...
 $ DayMins        : num  265 162 243 299 167 ...
 $ DayCalls       : num  110 123 114 71 113 98 88 79 97 84 ...
 $ MonthlyCharge  : num  89 82 52 57 41 57 87.3 36 63.9 93.2 ...
 $ OverageFee     : num  9.87 9.78 6.06 3.1 7.42 ...
 $ RoamMins       : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
```

The factor variables are explained as,

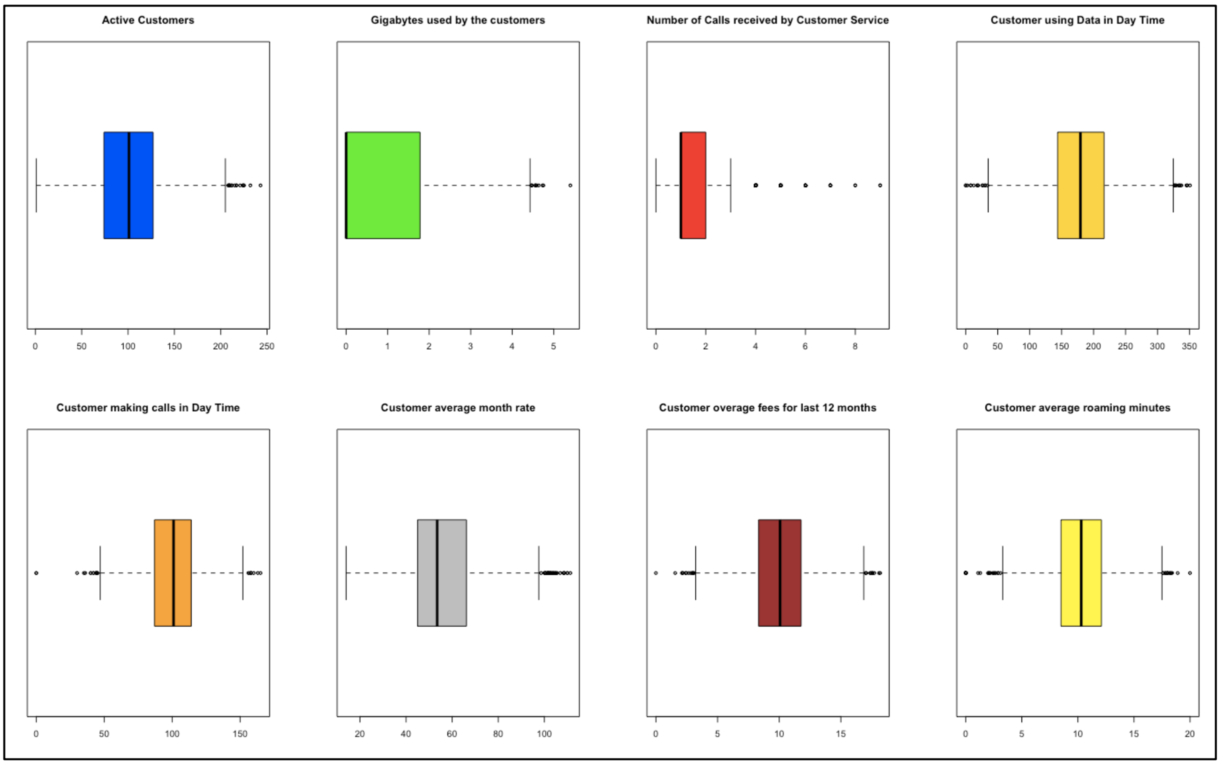
- Churn – “1” as customer cancelled the service and “0” as customer is using the service.
- Contract Renewal – “1” as customer renewed the contract and “0” as customer not renewed the contract.
- Data Plan – “1” as customer have the data plan and “0” as customer with no data.

```
>ggpairs(Cellphone[,c("AccountWeeks","DataUsage","CustServCalls","DayMins","DayCall”,
+                 "MonthlyCharge","OverageFee","RoamMins")],
+       ggplot2::aes(colour=as.factor(Cellphone$Churn)))
```

- The ggpairs are used to display the various correlation plots is the single chart for the better understanding of the scatterplots and the correlation plot charts.
- The plot shows that the variables Day Mins and Monthly Charge are highly correlated with the values 0.568 and next to that the variables overage fees and the monthly charge is correlated with the value 0.282.
- The least correlation are computed for the variables Customer Service calls and the Data Usage of the customer is -0.0217. The correlation values are helps to understand the values to maintained in the models and the variables day mins and monthly charge is picked for the model interpretations and the analysis for the customer using service.



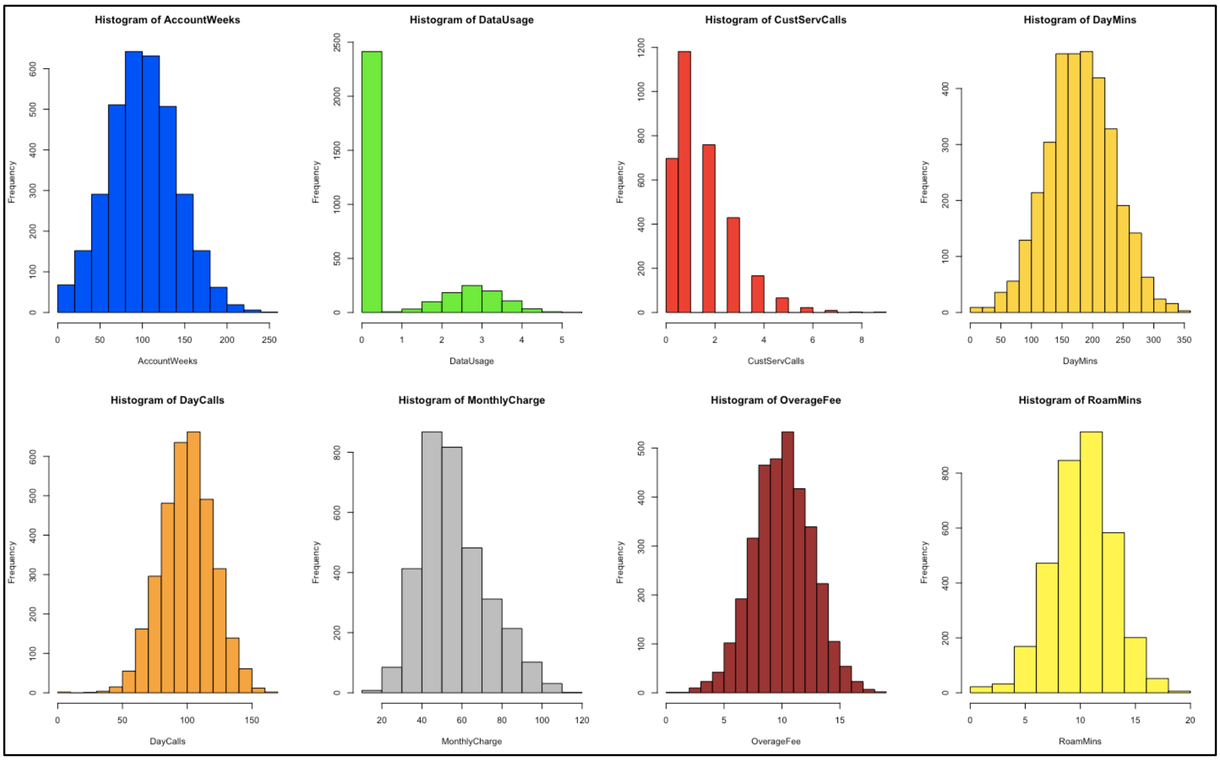
```
> par(mfrow=c(2,4))
> boxplot(AccountWeeks,horizontal = TRUE,
+       main = "Active Customers",col = "blue")
> boxplot(DataUsage,horizontal = T,
+       main = "Gigabytes used by the customers",col = "green")
> boxplot(CustServCalls,horizontal = T,
+       main = "Number of Calls received by Customer Service",col = "red")
> boxplot(DayMins,horizontal = T,
+       main = "Customer using Data in Day Time",col = "gold")
> boxplot(DayCalls,horizontal = T,
+       main = "Customer making calls in Day Time",col = "orange")
> boxplot(MonthlyCharge,horizontal = T,
+       main = "Customer average month rate",col = "grey")
> boxplot(OverageFee,horizontal = T,
+       main = "Customer overage fees for last 12 months",col = "brown")
> boxplot(RoamMins,horizontal = T,
+       main = "Customer average roaming minutes",col = "yellow")
```



The boxplots are used to find the outliers in the variables. The variables are taken as follows,

- The Gigabytes used by the customers in maximum of 5.4 GB and the least used data is 0 GB.
- The Monthly rate is maximum rate of 111.3 and the minimum values is 14.
- The Overage fees for the customer is 18 and the minimum fees is 0.
- The customer using data in the day time is maximum of 350.8 GB and the values are higher when compared to the day time calls made by the customer is 165
- The Customer service centre have handled the maximum of 9 calls from the customers

```
> par(mfrow=c(2,4))
> hist(AccountWeeks,col = "blue")
> hist(DataUsage,col = "green")
> hist(CustServCalls,col = "red")
> hist(DayMins,col = "gold")
> hist(DayCalls,col = "orange")
> hist(MonthlyCharge,col = "grey")
> hist(OverageFee,col = "brown")
> hist(RoamMins,col = "yellow")
```



2. Look out for outliers and missing values

```
> any(is.na(Cellphone))
[1] FALSE
> Cellphone=na.omit(Cellphone)
```

Checked the variables are created with any missing values and it shows that the missing values are not created. If there is any error in the checking, the missing values are also omitted from the datasets.

```
> dim(Cellphone)
[1] 3333   11
```

The total observations after treated missing values the observation are counted as 3333 observations and 11 variables.

```
> ct.data=subset(Cellphone,select = c(Churn,ContractRenewal,DataPlan))
> num.data=subset(Cellphone,select = -c(Churn,ContractRenewal,DataPlan))
> dim(ct.data)
[1] 3333    3
> dim(num.data)
[1] 3333    8
```

The variables are separated as per the categorical and numerical values.

```
> names(ct.data)
```

```
[1] "Churn"          "ContractRenewal" "DataPlan"
```

The Categorical variables are created by the levels like “customer using service”, “customer not using service” in Churn and “customer renewed service”, “customer not renewed the service” in contract renewal and “customer have data plan”, “customer don’t have data plan” in data plan variable.

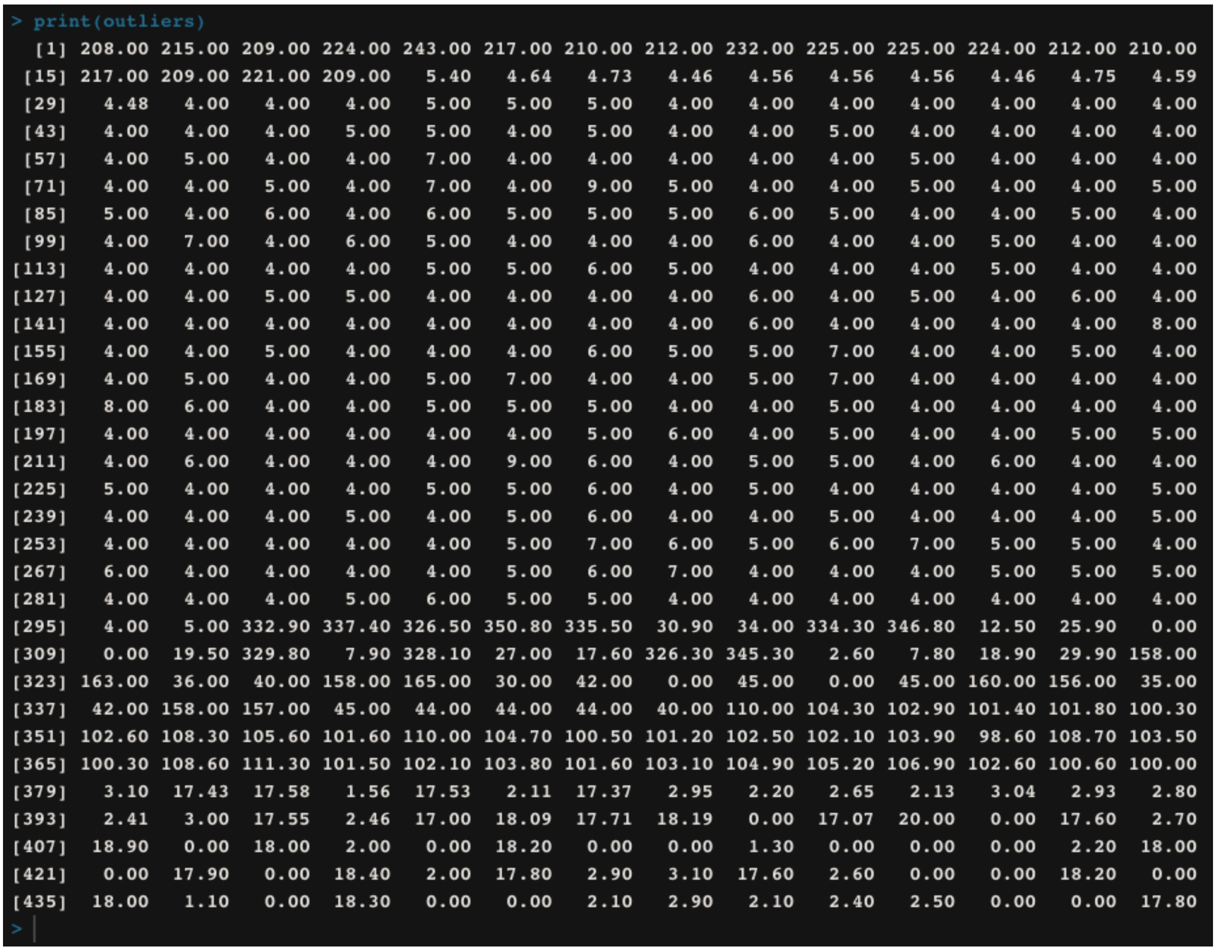
```
> names(num.data)
```

```
[1] "AccountWeeks"  "DataUsage"     "CustServCalls" "DayMins"       "DayCalls"      "MonthlyCharge"
[7] "OverageFee"    "RoamMins"
```

The numerical variables are Account Weeks, Data Usage, Customer Service Calls, Day Mins, Day calls, Monthly Charge, Overage Fee, Roaming Mins.

```
> outliers = boxplot(num.data,plot = FALSE)$out
> print(outliers)
```

The outliers are managed by the maximum values in each variables.



```
> print(outliers)
  [1] 208.00 215.00 209.00 224.00 243.00 217.00 210.00 212.00 232.00 225.00 225.00 224.00 212.00 210.00
 [15] 217.00 209.00 221.00 209.00   5.40   4.64   4.73   4.46   4.56   4.56   4.56   4.46   4.75   4.59
 [29]   4.48   4.00   4.00   4.00   5.00   5.00   5.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00
 [43]   4.00   4.00   4.00   5.00   5.00   4.00   5.00   4.00   4.00   5.00   4.00   4.00   4.00   4.00
 [57]   4.00   5.00   4.00   4.00   7.00   4.00   4.00   4.00   4.00   4.00   5.00   4.00   4.00   4.00
 [71]   4.00   4.00   5.00   4.00   7.00   4.00   9.00   5.00   4.00   4.00   5.00   4.00   4.00   5.00
 [85]   5.00   4.00   6.00   4.00   6.00   5.00   5.00   5.00   6.00   5.00   4.00   4.00   5.00   4.00
 [99]   4.00   7.00   4.00   6.00   5.00   4.00   4.00   4.00   6.00   4.00   4.00   5.00   4.00   4.00
[113]   4.00   4.00   4.00   4.00   5.00   5.00   6.00   5.00   4.00   4.00   4.00   5.00   4.00   4.00
[127]   4.00   4.00   5.00   5.00   4.00   4.00   4.00   4.00   6.00   4.00   5.00   4.00   6.00   4.00
[141]   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00   6.00   4.00   4.00   4.00   4.00   8.00
[155]   4.00   4.00   5.00   4.00   4.00   4.00   6.00   5.00   5.00   7.00   4.00   4.00   5.00   4.00
[169]   4.00   5.00   4.00   4.00   5.00   7.00   4.00   4.00   5.00   7.00   4.00   4.00   4.00   4.00
[183]   8.00   6.00   4.00   4.00   5.00   5.00   5.00   4.00   4.00   5.00   4.00   4.00   4.00   4.00
[197]   4.00   4.00   4.00   4.00   4.00   4.00   5.00   6.00   4.00   5.00   4.00   4.00   5.00   5.00
[211]   4.00   6.00   4.00   4.00   4.00   9.00   6.00   4.00   5.00   5.00   4.00   6.00   4.00   4.00
[225]   5.00   4.00   4.00   4.00   5.00   5.00   6.00   4.00   5.00   4.00   4.00   4.00   4.00   5.00
[239]   4.00   4.00   4.00   5.00   4.00   5.00   6.00   4.00   4.00   5.00   4.00   4.00   4.00   5.00
[253]   4.00   4.00   4.00   4.00   4.00   5.00   7.00   6.00   5.00   6.00   7.00   5.00   5.00   4.00
[267]   6.00   4.00   4.00   4.00   4.00   5.00   6.00   7.00   4.00   4.00   4.00   5.00   5.00   5.00
[281]   4.00   4.00   4.00   5.00   6.00   5.00   5.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00
[295]   4.00   5.00 332.90 337.40 326.50 350.80 335.50  30.90  34.00 334.30 346.80  12.50  25.90   0.00
[309]   0.00  19.50 329.80   7.90 328.10  27.00  17.60 326.30 345.30   2.60   7.80  18.90  29.90 158.00
[323] 163.00  36.00  40.00 158.00 165.00  30.00  42.00   0.00  45.00   0.00  45.00 160.00 156.00  35.00
[337]  42.00 158.00 157.00  45.00  44.00  44.00  44.00  40.00 110.00 104.30 102.90 101.40 101.80 100.30
[351] 102.60 108.30 105.60 101.60 110.00 104.70 100.50 101.20 102.50 102.10 103.90  98.60 108.70 103.50
[365] 100.30 108.60 111.30 101.50 102.10 103.80 101.60 103.10 104.90 105.20 106.90 102.60 100.60 100.00
[379]   3.10  17.43  17.58   1.56  17.53   2.11  17.37   2.95   2.20   2.65   2.13   3.04   2.93   2.80
[393]   2.41   3.00  17.55   2.46  17.00  18.09  17.71  18.19   0.00  17.07  20.00   0.00  17.60   2.70
[407]  18.90   0.00  18.00   2.00   0.00  18.20   0.00   0.00   1.30   0.00   0.00   0.00   2.20  18.00
[421]   0.00  17.90   0.00  18.40   2.00  17.80   2.90   3.10  17.60   2.60   0.00   0.00  18.20   0.00
[435]  18.00   1.10   0.00  18.30   0.00   0.00   2.10   2.90   2.10   2.40   2.50   0.00   0.00  17.80
> 
```

3.4 Bivariate Analysis

Bivariate Analysis is used to analyse the two variables and find the relationship between them. This analysis will help in identifying the association and strength of the variables. The analysis is used find the regression, distributions and plots.

As the Cell phone Dataset is mostly focus on customer churn, the categorical variables are selected and the numeric variables are selected and multicollinearity is measured and the insights are computed.

3. Check for multicollinearity & treat it

```
> lr.reg=glm(Churn~AccountWeeks,data = Cellphone,family = binomial)
> summary(lr.reg)
```

The Generalized Linear Method is used to compute the logistics regression for the account of weeks that the customers are actively using the telecom network.

The Deviance residuals shows the values in the 2.0169 is the 100% value for the error showing in the relation between churn and the customer active state.

Min	1Q	Median	3Q	Max
-0.6041	-0.5658	-0.5566	-0.5452	2.0169

The coefficients of the linear regression is taken the z value of 0.34 which is significant to analyse the data and the churn

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.894953	0.135634	-13.971	<2e-16	***
AccountWeeks	0.001179	0.001234	0.955	0.34	

The deviance value are measured by the degrees of freedom with values of 3332 in 2758.3 for NULL deviance and 3331 in 2757.4 for residual deviance. The AIC is measured with 2761.4 which is maximum likelihood for the churn and the customer active period. The values are significant, hence the customers will continue the service as per the logistic regression model.

```
> lr.reg=glm(Churn~DataUsage,data = Cellphone,family = binomial)
> summary(lr.reg)
```

The model is interpreting the values for the churn and Data Used by the customer.

Min	1Q	Median	3Q	Max
-0.6012	-0.6012	-0.5853	-0.4422	2.4047

The model is interpreting the data with the error value of 2.4 and the values are significant in the z value and the significance rate is higher for the data used by the customers.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.61888	0.05594	-28.941	< 2e-16	***
DataUsage	-0.22506	0.04531	-4.967	6.80E-07	***

Both the churn and the Data Usage of the customers are highly significant to predict the future existence of the customers in the linear method. The AIC is measured with 2734.5 for the likelihood of the data for the regression with DOF = 3331.

```
> lr.reg=glm(Churn~CustServCalls,data = Cellphone,family = binomial)
> summary(lr.reg)
```

Min	1Q	Median	3Q	Max
-1.476	-0.5799	-0.482	-0.3991	2.2671

The linear model is developed for the churn and the customer service calls and received the higher value of 2.2671.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.49016	0.08631	-28.85	<2e-16	***
CustServCalls	0.39617	0.03456	11.46	<2e-16	***

The customer service is highly significance with the churn as the values are intercepted with the customer satisfaction for existing the current telecom service in future. The deviance values are measured with the z value of 11.46 is highly correlated in the value. Hence the customer can have the highest possibilities for the retention in customer service.

The AIC value is calculated with 2631.2 with the DOF is 3331 and the values are highly correlated for the interpretation of the customer churn with customer service calls.

```
> lr.reg=glm(Churn~DayMins,data = Cellphone,family = binomial)
> summary(lr.reg)
```

Min	1Q	Median	3Q	Max
-1.0241	-0.6001	-0.4902	-0.3738	2.8102

The residuals value of the Daily Minutes are calculated as 2.8102 for the 100% error predicted in the variables.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.929289	0.202823	-19.37	<2e-16	***
DayMins	0.011272	0.000975	11.56	<2e-16	***

The daily mins are calculated for the higher significance with churn as the data is mostly connected with the customer satisfaction of service provided. The highest correlation leads in analyse the customer return back to the telecom service.

The AIC value is measured with 2618.3 with DOF of 3331 in the fisher scoring iterations of 5. Hence, the daily minutes of the gigabytes used by the customer is satisfied with the values in the each variable.

```
> lr.reg=glm(Churn~DayCalls,data = Cellphone,family = binomial)
> summary(lr.reg)
```

The variables used for the regression day calls make by the customer.

Min	1Q	Median	3Q	Max
-0.6031	-0.5665	-0.5563	-0.5443	2.0792

Standard error of the interception shows the value of 0.25 which is less error to predict the relation between churn and the day time calls. The z value is 0.287 which makes the less significant between the churn and day calls variables.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.039138	0.253579	-8.041	8.88E-16	***
DayCalls	0.00262	0.002458	1.066	0.287	

The AIC value is calculated as 2761.2 in the DOF is 3331 with the fisher iterations is 4.

```
> lr.reg=glm(Churn~MonthlyCharge,data = Cellphone,family = binomial)
> summary(lr.reg)
```

The maximum residuals values are measured as 2.188 with the deviance error.

Min	1Q	Median	3Q	Max
-0.7498	-0.5707	-0.5366	-0.5043	2.1888

z-value shows the relation is highly significant and the values are measured in the values for the less standard error is 0.177 and the relation can be used in the future predicting value. Interception values are calculated for the estimated value is 0.012 and the customer churn are measured with the z-value of 4.16 for the predicted values.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.468836	0.177192	-13.93	< 2e-16	***
MonthlyCharge	0.012072	0.002902	4.16	3.19E-05	***

The higher significant value is maintained in the AIC value is 2745.3 with DOF is 3331 and the fisher iterations is 4. The NULL deviance value is 2758.3 in the relation is calculated by the DOF is 3332.

```
> lr.reg=glm(Churn~OverageFee,data = Cellphone,family = binomial)
> summary(lr.reg)
```

The regression values are measured with higher deviance.

Min	1Q	Median	3Q	Max
-0.8069	-0.5874	-0.5366	-0.4781	2.2644

The maximum standard error is calculated with the value is 2.2644 and the higher significance is measured with the z value as the overage fees for the customer used data is significantly increasing in the standard error.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.8568	0.21318	-13.401	< 2e-16	***
OverageFee	0.10513	0.01971	5.335	9.56E-08	***

The higher z value shows that 9.56E is negatively significant with the customer churn and the customer is less chance of continue the service with the significant values are connected in the standard error for the overage fees of the variables.

The AIC value is 2733.4 is calculated with the DOF is 3331 and the fisher iteration value is 4.

```
> lr.reg=glm(Churn~RoamMins,data = Cellphone,family = binomial)
> summary(lr.reg)
```

The regression value is measured with the values are calculated in the error value of maximum is 2.2190.

Min	1Q	Median	3Q	Max
-0.7338	-0.5814	-0.5463	-0.4995	2.219

The z value shows the negative significance is maintained in the relation is calculated with the regression for the roaming minutes and the churn is showing the less chance to continue the service based on the z value.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.51472	0.19778	-12.715	< 2e-16	***
RoamMins	0.07091	0.01803	3.932	8.41E-05	***

The roaming minutes is negatively significant with the churn and the customers are based on the not using the service as the AIC value is 2746.6 with DOF is 3331 and the fisher iterations values is measured as 4. The z value is maintained by the error rate of 0.018 and the interception value is calculated for 0.19 and relation between churn and the dataset of roaming minutes is less significant for the regression.

```
> model=glm(Churn~.,data = Cellphone,family = binomial)
> summary(model)
```

The multicollinearity is shown with the values and the numeric variables are selected for the analysis based on the customer churn in the regression.

Min	1Q	Median	3Q	Max
-2.0058	-0.5112	-0.3477	-0.2093	2.9981

The max error for the deviance is measured with 2.9981 and the model shows that the interceptions is classified with all variables for the customer churn.

		Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)		-5.951025	0.54868	-10.846	< 2e-16	***
AccountWeeks		0.0006525	0.00139	0.47	0.638112	
ContractRenewal	Customer renewed	-1.985517	0.14361	-13.826	< 2e-16	***
DataPlan	Customer have data	-1.184161	0.53637	-2.208	0.027262	*
DataUsage		0.3636565	1.92318	0.189	0.850021	
CustServCalls		0.5081349	0.03897	13.04	< 2e-16	***
DayMins		0.0174407	0.03248	0.537	0.591337	
DayCalls		0.0036523	0.00275	1.328	0.184097	
MonthlyCharge		-0.027553	0.19091	-0.144	0.885244	
OverageFee		0.1868114	0.32569	0.574	0.566248	
RoamMins		0.0789226	0.02205	3.579	0.000345	***

The significant values are measured as category variables are contract renewal and data plan. The numeric variables are highly significant with the values and the measured values are calculated as the higher z-values. The customer churn is measured with higher values from the active accounts, data usage, day time minutes of using data, monthly charge and overage fees.

The values are not significant but the higher z values shows the variables have higher correlation to predict the data in the customer satisfaction for the telecom service.

The contract renewal and data plan variables are selected with the values are predicting with the customers renewed the service and the customer have data plan respectively. The variables are suitable for the categorical distribution and the values are measured in the numeric variables are predicting the customer satisfaction.

The AIC value is calculated with 2210.4 with DOF is 3322 in the fisher iteration values is 5.

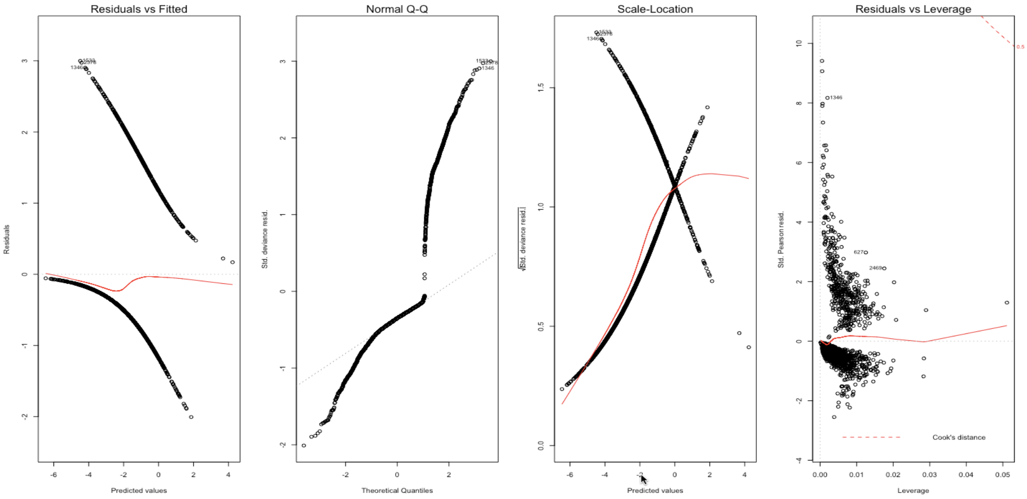
```
> car::vif(model)
```

AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins
1.003246	1.058705	14.087816	1601.163095	1.08125	952.539781

DayCalls	MonthlyCharge	OverageFee	RoamMins
1.004592	2829.804947	211.716226	1.193368

The variance inflation factor is used to predict the analysis of the customer churn and the values are calculated in the basis of the variance. The higher variance inflation rate is measured with the higher value is 211.71 and the lowest inflation rate are measure in the active customer usage, contract renewal, customer service calls, roaming minutes and day time calls are measured in the variance factors.

Residuals and fitted values are showing the various factors are based with different variables. The variables are calculated with the line of the fitted values are not merged for the line and the values are fitted in the values above 1 and the values are mentioned in the normal Q-Q are measured with the values are higher as per the standard deviation. The scale location is intercepting on the point 1 and the values are interpreted for the squares of the standard deviation.



4. Summarize the insights you get from EDA

```
> par(mfrow=c(1,3))
> for (a in names(ct.data)) {
+   print(a)
+   print(round(prop.table(table(Cellphone$Churn,ct.data[[a]])),digits = 3)*100)
+   barplot(table(Cellphone$Churn,ct.data[[a]]),
+           col = c("blue","yellow"),
+           main = names(ct.data[a]))
+ }
```

[1] "Churn"

	Customer Using Service	Customer Cancelled Service
Customer Using Service	85.5	0
Customer Cancelled Service	0	14.5

The table shows that the customer using service is about 85.5% and the customer not using service is 14.5%.

[1] "ContractRenewal"

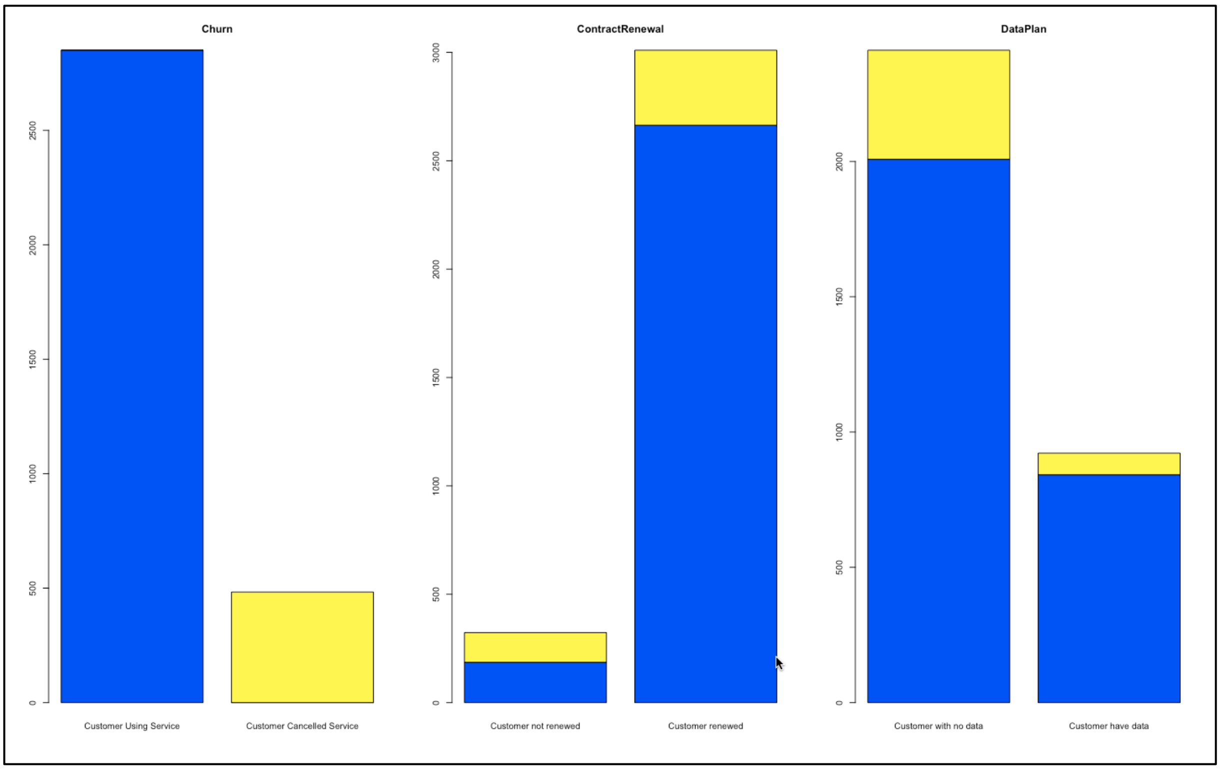
	Customer not renewed	Customer renewed
Customer Using Service	5.6	79.9
Customer Cancelled Service	4.1	10.4

The table is showing the true positive value is measured with 5.6% and the values are false positive rate of 79.9% the values are showing the customer renewed the service to the telecom provider.

[1] "DataPlan"

	Customer with no data	Customer have data
Customer Using Service	60.2	25.3
Customer Cancelled Service	12.1	2.4

The data plan is showing the customers using service are not with data usage of 60.2% and the values measured with the customer using data and availing service is 25.3%. This shows that most of the customers are looking for the various data plan as per customer convenience.



The graphs explained the various customer service using the service for the telecom provider. Contract renewal shows that the values in the customer renewed are not using the service of 10.4% and the telecom company may be profitable for the customer not using the service. The customer with no data plan is using the service at 60.2% and the company facing some issues in using the network for the company charges and increasing the values for the customer have data and not using the service is about 2.4% which is lesser than the customer have data. Hence, the telecom company is making the profit in the customer cancelled service for long period.

Insights

- The customer using the service are renewed the service with 79% and using the telecom data is 25% shows that the customers are engaged in providing network of the company.
- The Customer churn is the basic factor used in predicting the nature of the observed variables and promote the usage of telecom service provider.
- Logistic Regression shows that the customer churn are making the highest significance in numeric data as per the z values and the values are enriched with the data of Z>0.5.
- The customer using the service can be increased for the telecom provider with the most probability of chance in retention in customer usage.
- Multicollinearity shows the values are related to the family binomial and the values selected for the customer renewed service and the customer using data are measured for the positive rates, hence the significance rates for the category variables in the customer field of using the day time calls, day time usage of data.

- The multiple regression of the data shows the relation between churn and numerical variables are created for the negative significance, since it controls the churn the customer service are entitled to use the same network for the next year.
- The histogram of the numeric variables are predict the churn datasets should to organized in the positive and the negative rates for the customer dataset.
- The telecom service provider can retain the customers at the average rate above 50% and the customers are regained with new discounts and the changes in data plan will increase the usage of the customers of the particular telecom.
- Outliers and missing values are treated with the customer churn and the values are arrange with the full model for the numeric and categorical values.
- The outliers defined in the model doesn’t affects the customer churn and the customer usage in the data and the variables are promoting the values for the higher return in the active customers are using the data and calls.

5. Logistic Regression

```
> library(rms)
```

```
> vif(glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial))
```

The customer churn is predicted with the monthly charge and day time usage of the customers. As the logistic regression for the values are measured with the generalized linear model.

MonthlyCharge	DayMins
1.504358	1.504358

Variance inflation factor shows the values are equal in making the logistic regression model and the logistic regression is predicted for this variables.

```
> cor(MonthlyCharge,DayMins)
```

[1] 0.5679679

The correlation between the variables are highly correlated and having the positive relationship between the variables.

```
> summary(glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial))
```

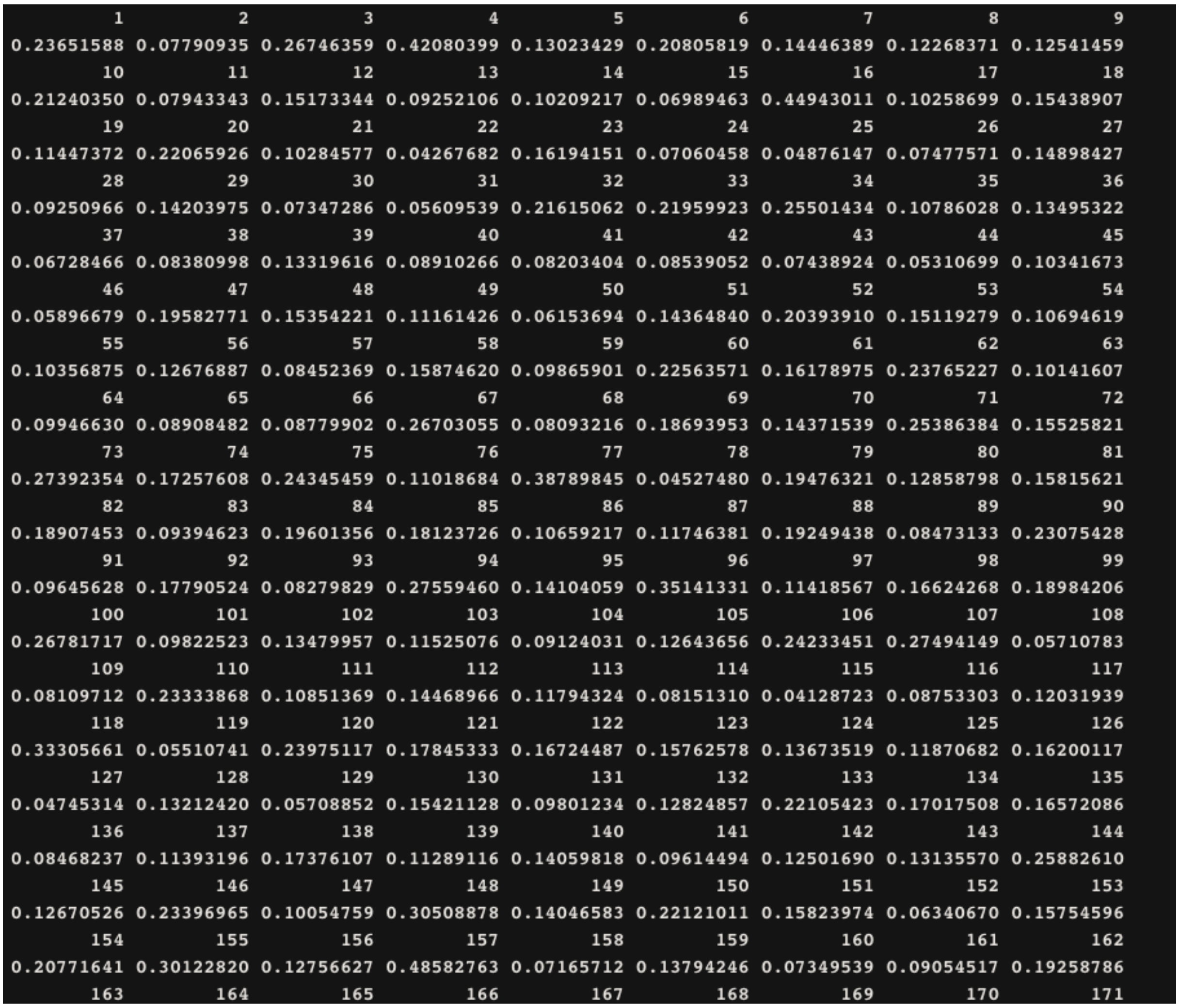
Min	1Q	Median	3Q	Max
-1.0839	-0.5981	-0.4862	-0.3692	2.7642

The logistics regression for the customer churn is based on the variables for the error in the maximum value of 2.7642 and the values are measured with the higher significance values of the basic models.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.626079	0.223301	-16.239	< 2e-16	***
MonthlyCharge	-0.012291	0.003973	-3.093	0.00198	**
DayMins	0.013384	0.001197	11.18	< 2e-16	***

The NULL deviance calculated as 2758.3 on degrees on freedom on 3332 and the residual deviance is 2604.2 for the degrees of freedom of 3330 and the AIC is 2610.2

```
> logistic.churn=glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial)
> logistic.churn$fitted.values
```



```
         1          2          3          4          5          6          7          8          9 
0.23651588 0.07790935 0.26746359 0.42080399 0.13023429 0.20805819 0.14446389 0.12268371 0.12541459 
        10         11         12         13         14         15         16         17         18 
0.21240350 0.07943343 0.15173344 0.09252106 0.10209217 0.06989463 0.44943011 0.10258699 0.15438907 
        19         20         21         22         23         24         25         26         27 
0.11447372 0.22065926 0.10284577 0.04267682 0.16194151 0.07060458 0.04876147 0.07477571 0.14898427 
        28         29         30         31         32         33         34         35         36 
0.09250966 0.14203975 0.07347286 0.05609539 0.21615062 0.21959923 0.25501434 0.10786028 0.13495322 
        37         38         39         40         41         42         43         44         45 
0.06728466 0.08380998 0.13319616 0.08910266 0.08203404 0.08539052 0.07438924 0.05310699 0.10341673 
        46         47         48         49         50         51         52         53         54 
0.05896679 0.19582771 0.15354221 0.11161426 0.06153694 0.14364840 0.20393910 0.15119279 0.10694619 
        55         56         57         58         59         60         61         62         63 
0.10356875 0.12676887 0.08452369 0.15874620 0.09865901 0.22563571 0.16178975 0.23765227 0.10141607 
        64         65         66         67         68         69         70         71         72 
0.09946630 0.08908482 0.08779902 0.26703055 0.08093216 0.18693953 0.14371539 0.25386384 0.15525821 
        73         74         75         76         77         78         79         80         81 
0.27392354 0.17257608 0.24345459 0.11018684 0.38789845 0.04527480 0.19476321 0.12858798 0.15815621 
        82         83         84         85         86         87         88         89         90 
0.18907453 0.09394623 0.19601356 0.18123726 0.10659217 0.11746381 0.19249438 0.08473133 0.23075428 
        91         92         93         94         95         96         97         98         99 
0.09645628 0.17790524 0.08279829 0.27559460 0.14104059 0.35141331 0.11418567 0.16624268 0.18984206 
       100        101        102        103        104        105        106        107        108 
0.26781717 0.09822523 0.13479957 0.11525076 0.09124031 0.12643656 0.24233451 0.27494149 0.05710783 
       109        110        111        112        113        114        115        116        117 
0.08109712 0.23333868 0.10851369 0.14468966 0.11794324 0.08151310 0.04128723 0.08753303 0.12031939 
       118        119        120        121        122        123        124        125        126 
0.33305661 0.05510741 0.23975117 0.17845333 0.16724487 0.15762578 0.13673519 0.11870682 0.16200117 
       127        128        129        130        131        132        133        134        135 
0.04745314 0.13212420 0.05708852 0.15421128 0.09801234 0.12824857 0.22105423 0.17017508 0.16572086 
       136        137        138        139        140        141        142        143        144 
0.08468237 0.11393196 0.17376107 0.11289116 0.14059818 0.09614494 0.12501690 0.13135570 0.25882610 
       145        146        147        148        149        150        151        152        153 
0.12670526 0.23396965 0.10054759 0.30508878 0.14046583 0.22121011 0.15823974 0.06340670 0.15754596 
       154        155        156        157        158        159        160        161        162 
0.20771641 0.30122820 0.12756627 0.48582763 0.07165712 0.13794246 0.07349539 0.09054517 0.19258786 
       163        164        165        166        167        168        169        170        171 
```

The values are fitted for the variables and the missing values are treated by the omissions of the fitted values in the variables.

```
> exp(0.01338)
```

[1] 1.01347

The log values is computed for the fitted values in the predicted variables. The values are used for the general base value of the customer churn.

```
> churn.predicted=ifelse(logistic.churn$fitted.values<0.20,
+                     "Customer with no service","Customer with Service")
> table(Churn,churn.predicted)
```

The table for the comparison is for the variables are created for the customer with service is 448.

churn.predicted

Churn	Customer with no service	Customer with Service
0	2402	448
1	263	220

Customer service is based for predicted values in the fitted values and the values are predicted below 0.20 of the fitted values.

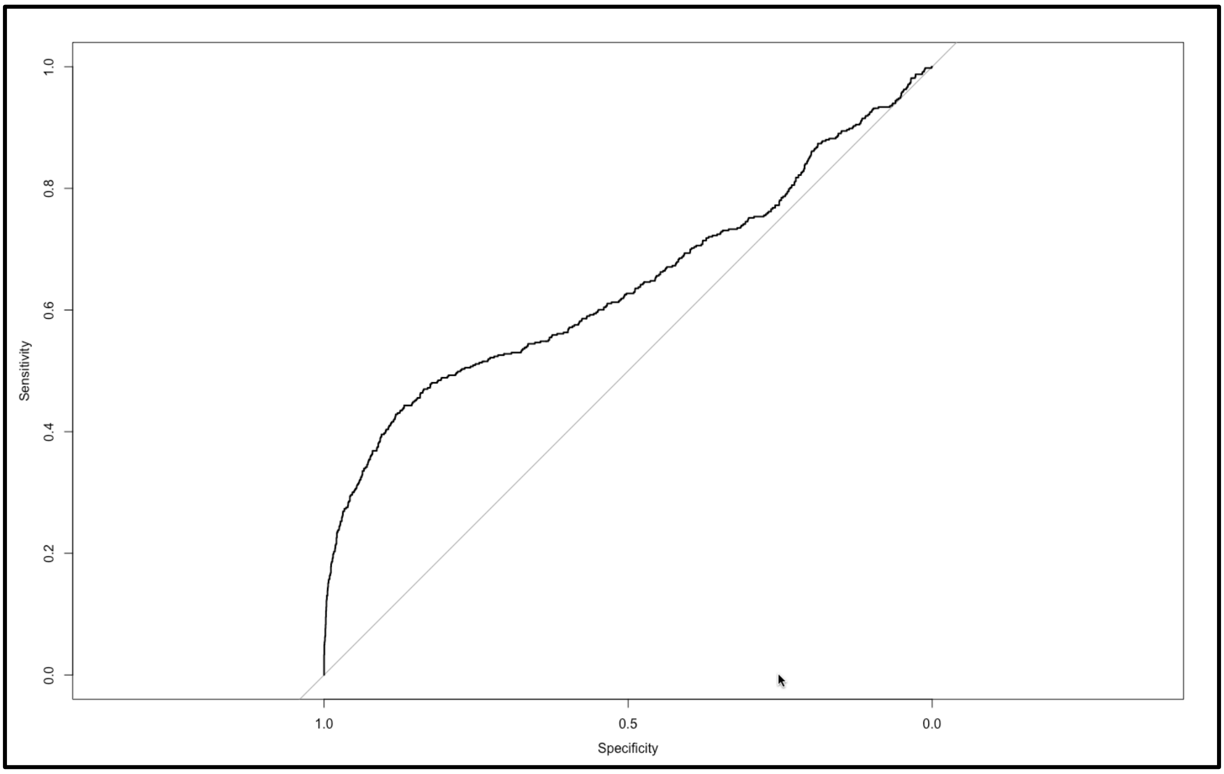
The values are showing the general fitted values in the basic variables to predicted the customer using service is 220 and the values are fitted with 85%.

```
> pROC::roc(Churn,logistic.churn$fitted.values)
```

Area under the curve: 0.6413

The ROC is used to compute the classifiers with the specificity and the sensitivity values for the area under the curve is total value of 64%, the values are showing this ROC to find the groups under the area and the performance is established for the customer using service and the customer not using service.

```
> par(mfrow=c(1,1))
> pROC::plot.roc(Churn,logistic.churn$fitted.values)
```



The ROC curve shows that the sensitivity line is above the curve and the values are computing the values are higher between 0.4 and the specificity is decreased on the value on 1.

6. KNN – K Nearest Neighbour

```
> library(class)

> dim(Cellphone)

[1] 3333   11

> set.seed(500)

> index=sample(3333,2333)

> ktrain=Cellphone[index,]

> dim(ktrain)

[1] 2333   11

> ktest=Cellphone[-index,]

> dim(ktest)

[1] 1000   11
```

The KNN is computed by using the train and test datasets using the splits of the original dataset. The K nearest neighbour is the function used to predict the variables which are nearer to the customer churn in the variables for the prediction of the nearest variable to the customer churn. The test data are created with 2333 observations and 11 variables, test and validating data is created with the 1000 observation and 11 variables.

```
> names(ktrain)

 [1] "Churn"          "AccountWeeks"    "ContractRenewal" "DataPlan"        "DataUsage"

 [6] "CustServCalls"   "DayMins"         "DayCalls"        "MonthlyCharge"   "OverageFee"

[11] "RoamMins"

> kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=5)

> table(ktest$Churn,kchurn)
```

	Customer Using Service	Customer Cancelled Service
Customer Using Service	824	25
Customer Cancelled Service	101	50

The customer using service is based on the values are predicting with 82.4% are interested for the customer organized in the values for the KNN.

```
> kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=6)
```

```
> table(ktest$Churn,kchurn)
```

	Customer Using Service	Customer Cancelled Service
Customer Using Service	823	26
Customer Cancelled Service	103	48

The k value is 6 and the values are computed as the 82.3% in the customer service and the data re predicted with the train and the validation is based on the customer not using service.

```
> kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=7)
```

```
> table(ktest$Churn,kchurn)
```

	Customer Using Service	Customer Cancelled Service
Customer Using Service	830	19
Customer Cancelled Service	103	48

The customer using service is increased by 83% in the k value is 7, then prediction is based on the for is 19%.

```
> kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=8)
```

```
> table(ktest$Churn,kchurn)
```

The k-test is predicted for the values with the customer is having 16% of cancelled service in for the telecom service.

	Customer Using Service	Customer Cancelled Service
Customer Using Service	833	16
Customer Cancelled Service	101	50

```
> kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=9)
```

```
> table(ktest$Churn,kchurn)
```

	Customer Using Service	Customer Cancelled Service
Customer Using Service	832	17
Customer Cancelled Service	103	48

The K value of 9 shows that increase in the cancelled service is about 17% and the values are increased on predicting the test datasets.

```
> kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=10)
```

```
> table(ktest$Churn,kchurn)
```

	Customer Using Service	Customer Cancelled Service
Customer Using Service	833	16
Customer Cancelled Service	103	48

The k value shows the customer cancelled service for the 16% and about of 4.8% is about total false and the groups are nearer to the values.

```
> kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=11)
```

```
> table(ktest$Churn,kchurn)
```

	Customer Using Service	Customer Cancelled Service
Customer Using Service	833	16
Customer Cancelled Service	102	49

K value = 11 shows that the variables are nearer of 83.3% in the data and the K values are predicted for the nearest value.

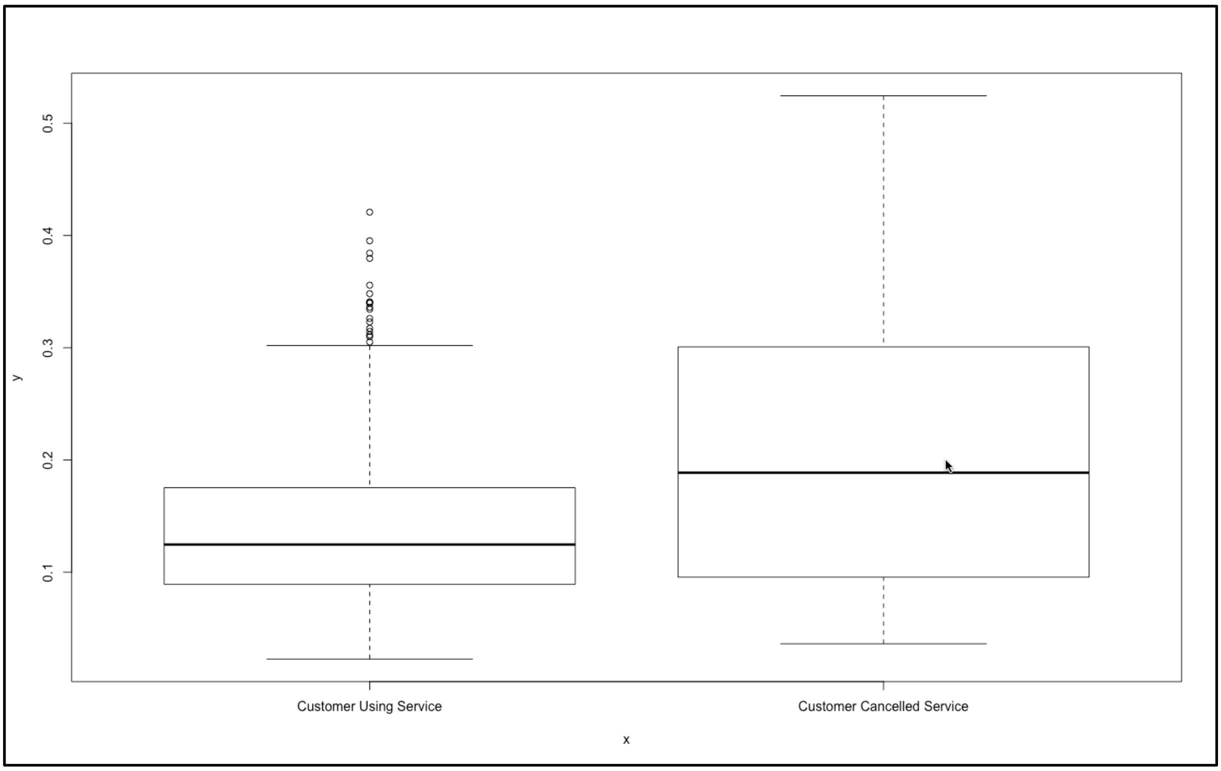
```
> kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=21)
```

```
> table(ktest$Churn,kchurn)
```

	Customer Using Service	Customer Cancelled Service
Customer Using Service	834	15
Customer Cancelled Service	99	52

The distance is calculated for the Euclidean distance in the variables and the classification is measured as KNN. The k value is 21 and it is higher expensive when compared the lower the values. The K values is predicted to the 5% of the values are cancelled the service.

```
> plot(ktest$Churn,
+     predict(glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial),
+           newdata = ktest,type = "response"))
```



This plots shows that the customer using service is not higher when compared with the customer not using the service. The plots shows that the customer using the service are with the outliers of the 0.1% in the top 75% and the values are calculated in the increasing factors for the outlier in the customer cancelled service.

The testing factor is the basic function in measuring the customer using service as per K training set.

```
> pROC::roc(ktest$Churn,
+         predict(glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial),
+                     newdata = ktest,type = "response"))
```

Area under the curve: 0.6508

The KNN value is predicted with the total of 65% making the customer to use the service in the factors for the next period. The K value is predicting the higher factor to validating the value in the variables.

7. Naïve Bayes

The Naïve Bayes is predicted after the assumption of the Linear Discriminant Analysis.

```
> library(MASS)
```

```
> lda.churn=lda(Churn~MonthlyCharge+DayMins,data = Cellphone,CV=TRUE)
```

The linear discriminant analysis is used to predict the values are enhanced for the customer satisfaction towards the telecom service. The linear discriminant analysis is shown and predicted for the analysis using the variables monthly charges and the day time usage of the customers.

The variables are predicted with the factors and numeric variables to the values are practicing in the linear values as posterior and the before period of the variables.

```
> lda(Churn~MonthlyCharge+DayMins,data = Cellphone)
```

Customer Using Service	Customer Cancelled Service
0.8550855	0.1449145

The probability of the customer using the service is predicted with 85% which is favoured to the telecom company to get back the customer averagely. The prior probabilities are assuming the customer churn data for the prediction.

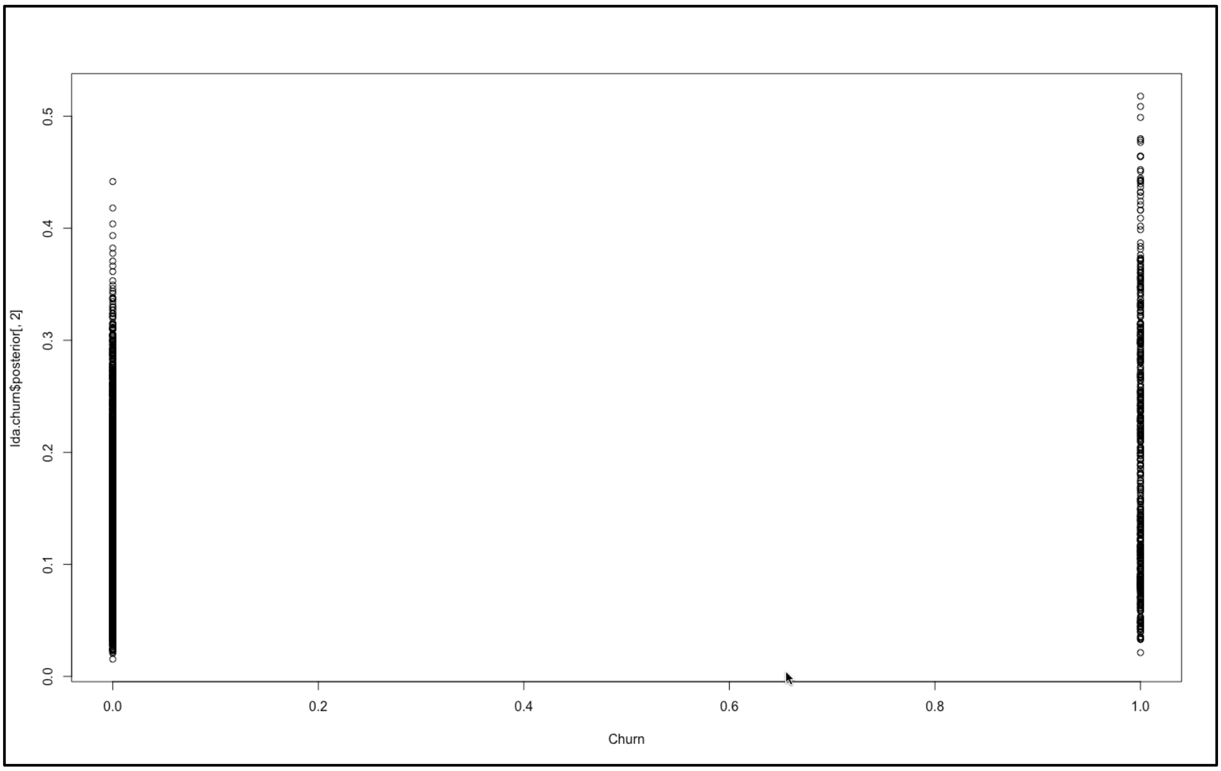
	MonthlyCharge	DayMins
Customer Using Service	55.81625	175.1758
Customer Cancelled Service	59.19006	206.9141

The monthly charges predicted by the customer churn is 59% cancelled the service and availing the monthly charges while 55% of the customers are making the use of the service and they are making the day time usage of the 175% which helps the telecom company to earn the profit but the factors on the customer cancelled service is using the network on day time which is higher than the customer using service. Hence, the company has to need an action in customer service.

	LD1
MonthlyCharge	-0.0191671
DayMins	0.02145467

The coefficients are measured as positive value of 2% value in the Day time calls and the linear discriminants shows that the customer churn is based on the day time usage and the monthly charge is inversely to the customer churn and the values are predict the opposite position of the customer using service.

```
> plot(Churn,lda.churn$posterior[,2])
```



The posterior value of the linear discriminant is shown the basic formulation of the customer churn to allows the value of the predicting variables in the plots.

```
> churn.predicted=ifelse(lda.churn$posterior[,2]<0.15,

+                 "Customer Using Service","Customer Cancelled Service")
```

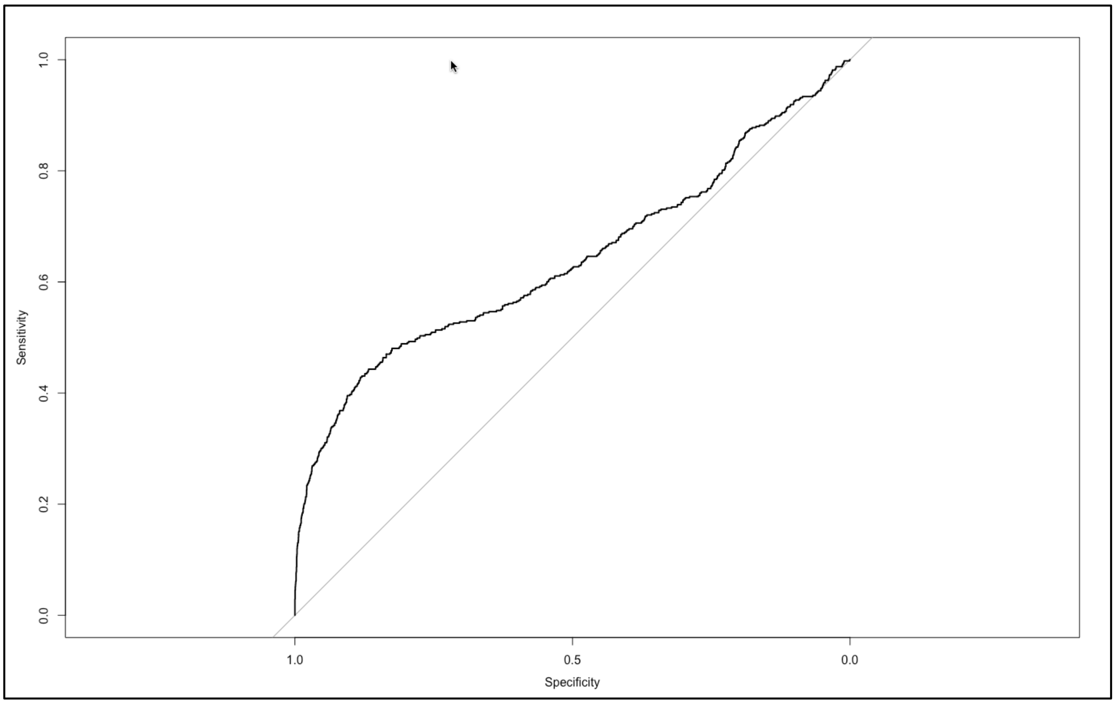
The predicted variables is based on the customer churn, hence the posterior values are calculated for the linear discriminant analysis.

```
> table(Churn,churn.predicted)
```

Churn	Customer Cancelled Service	Customer Using Service
0	1057	1793
1	265	218

The true positive rate shows the values are highly predicted for the customer churn and the values are predicting the general measures with higher customer using service on 1793 and the values are cross validating with the customer prediction in the general prediction of the churn.

```
> pROC::plot.roc(Churn,lda.churn$posterior[,2])
```



The plot shows the similar to the logistics regression and the values are maintaining the total population of the customer churn and the prediction values are measuring the total specificity and the sensitivity for the values in the predicted variables. The linear discriminant analysis is predicting the values are not accurate in the customer churn.

```
> nb.churn=naiveBayes(Churn~MonthlyCharge+DayMins,data = Cellphone)
```

```
> nb.churn
```

Customer Using Service	Customer Cancelled Service
0.8550855	0.1449145

The customer churn is predicted with the Naïve Bayes for the A prior probabilities in which the customer are likely to continue the services are 85.5% and the 14.9% are cancelled services.

Y	[,1]	[,2]
Customer Using Service	55.81625	16.43901
Customer Cancelled Service	59.19006	16.06548

Monthly charge of the customer using the service are measured as the values are prior with using the service is 55.8% and the posterior probability is showing the 16.4% for the customer usage in the telecom service for the higher chance in using the network.

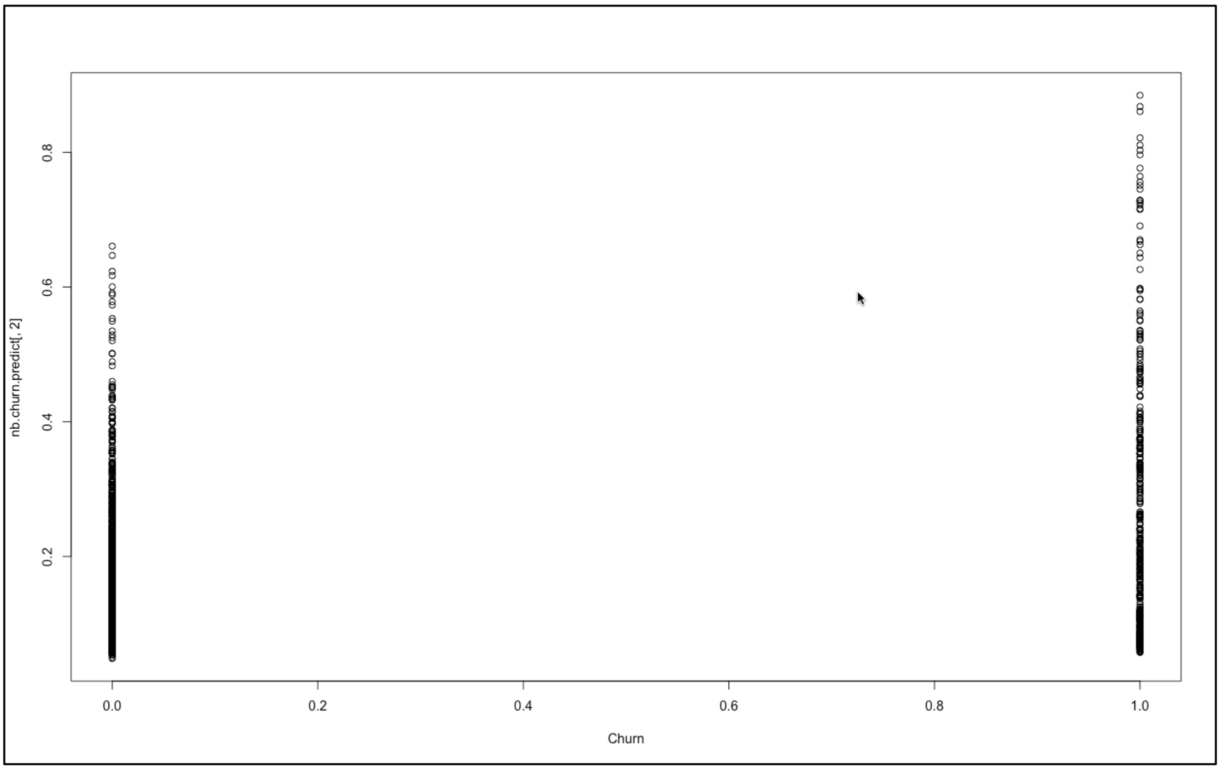
Y	[,1]	[,2]
Customer Using Service	175.1758	50.18166
Customer Cancelled Service	206.9141	68.99779

The Naïve Bayes is predicting the day time usage of the customer are using the service with the value of 175 and the posterior probability are measured for the 50% of the customer churn and 68% of the probabilities are showing the customer not using the service for the prediction of the variables.

```
> nb.churn.predict=predict(nb.churn,type = "raw",newdata = Cellphone)
```

```
> plot(Churn,nb.churn.predict[,2])
```

The prediction is computed for the Naïve Bayes in the customer churn with the whole datasets as the Naïve Bayes for the prediction in the validating of the full datasets. The prediction are same for the test and train datasets in the Naïve Bayes.



The prediction of the customer churn is available for the customer predicting in the posterior probability and the higher classification is value for the customer using in the various models for the general classification of the models and the variables for the customer is predict with not using the service for the telecom service.

8. Model Comparison using Model Performance metrics & Interpretation

```
> library(caTools)

> set.seed(100)

> splits=sample.split(fullmodel$Churn,SplitRatio = 0.60)

> train=subset(fullmodel,splits==T)

> test=subset(fullmodel,splits==F)
```

The model comparison is calculated for the Logistic Regression for the variables predicting in the customer churn for the higher probability for the various values like AUC, ROCR and Gini.

```
> prop.table(table(fullmodel$Churn))
```

Customer Using Service	Customer Cancelled Service
0.8550855	0.1449145

The values for the customer churn is based on the customer using service and the customer cancelled the service from the telecom company.

It shows that 85.5% of the customers are using the service and 14.4% are cancelled the service from the telecom service.

```
> cellphonemodel=glm(Churn~.,data = train,family = binomial)
```

```
> summary(cellphonemodel)
```

Min	1Q	Median	3Q	Max
-1.8932	-0.4974	-0.338	-0.1944	2.987

The deviance residuals for the Train Datasets are measured with the values is 2.987 which is the higher probability of the customer churn with using the service.

		Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)		-6.2404402	0.7178307	-8.693	<2e-16	***
AccountWeeks		-2.0145034	0.18779	-10.727	<2e-16	***
ContractRenewal	Customer renewed	-1.0770316	0.7190972	-1.498	0.1342	
DataPlan	Customer have data	-0.0006898	0.001844	-0.374	0.7083	
DataUsage		1.4184436	2.5311665	0.56	0.5752	
CustServCalls		0.5545877	0.0511533	10.842	<2e-16	***
DayMins		0.0372775	0.0427776	0.871	0.3835	
DayCalls		0.0071152	0.0035688	1.994	0.0462	*
MonthlyCharge		-0.1397107	0.2514072	-0.556	0.5784	
OverageFee		0.3869361	0.4284924	0.903	0.3665	
RoamMins		0.0609767	0.028744	2.121	0.0339	*

The regression values for the variables are highly significant in the contract renewal, customer service calls, day calls and roaming calls. The variables are intercepted with the z value are cleared with the z value is calculated as negative value in the customer churn.

The AIC value is measured with the 1306.5 for the DOF is measured with the 1248.5.

```
> car::vif(cellphonemodel)
```

```
ContractRenewal      DataPlan   AccountWeeks     DataUsage  CustServCalls       DayMins

       1.070557     14.293775       1.011734   1516.499449       1.087330    970.825029

       DayCalls MonthlyCharge     OverageFee      RoamMins

       1.011047   2723.122055     222.527854      1.172910
```

The variance inflation factor is showing the factors that computed in the variables for the higher relation in the customer churn to make an significance values for the variable day mins, data usage, monthly charge in the customer service.

```
> lrtest(cellphonemodel)
```

The likelihood ratio is used to provide the relations between the variables for the higher liking values to be predicted for the chi-square test for the relations between the DF is in full variables and 1st variables for the higher with 371.25 in the customer service.

	#Df	LogLik	Df	Chisq	Pr(>Chisq)	
1	11	-642.25				
2	1	-827.87	-10	371.25	< 2.2e-16	***

The log values of the variables are varied in selecting the all variables and selecting the one variables to the factors.

```
> pR2(cellphonemodel)["McFadden"]
```

```
McFadden

0.2242158
```

The McFadden value is 0.2242158 and shows the values are created with the higher possibilities to create the models in the values for the squared values of the variables for the prediction in the computed values.

```
> logLik(cellphonemodel)
```

```
'log Lik.' -642.2518 (df=11)
```

The log value of the model is computed as -642.2518 and the negative values are produced with the degrees of freedom with 11.

```
> trainmodel1=glm(Churn~1,data = train,family = binomial)
```

```
> 1-(logLik(cellphonemodel)/logLik(trainmodel1))
```

```
'log Lik.' 0.2242158 (df=11)
```

The log value shows the squared McFadden predicted values and the validating data can be predicted in the cell phone service.

```
> logLik(trainmodel1)
```

```
'log Lik.' -827.8743 (df=1)
```

The likelihood of the train model shows the value of -827.8743 which is rate of the train validation and it can be used for the squared value of the log in the customer churn.

```
> trainpredict=predict(cellphonemodel,newdata = train,type = "response")
```

```
> table(train$Churn,trainpredict>0.5)
```

	FALSE	TRUE
Customer Using Service	1659	51
Customer Cancelled Service	227	63

The predicted train datasets shows the values are created with the customer using service with the false positive rate is 1659 and the true positive rate of the 51, which is various divided as well in the customer cancelled service from the telecom.

```
> (1659+51)/nrow(na.omit(train))
```

```
[1] 0.855
```

The train model predicted for the customer using service is about 85% and the train validation shows the customer is averagely using the network and to be continuing the service of the telecom.

```
> testpredict=predict(cellphonemodel,newdata = test,type = "response")
```

```
> table(test$Churn,testpredict>0.5)
```

	FALSE	TRUE
Customer Using Service	1111	29
Customer Cancelled Service	161	32

The predicted table of the validation datasets is showing the results for the validation data with the predicted value above 0.5.

```
> (1111+29)/nrow(na.omit(test))
```

```
[1] 0.8552138
```

The validation data predicted produces the 85% of the customers are using the service and the variables are predicted usage of the customer in the telecom network.

#AUC and ROC for Train and Test Dataset

```
> roctrainpredict=prediction(trainpredict,train$Churn)
```

```
> as.numeric(performance(roctrainpredict,"auc")@y.values)
```

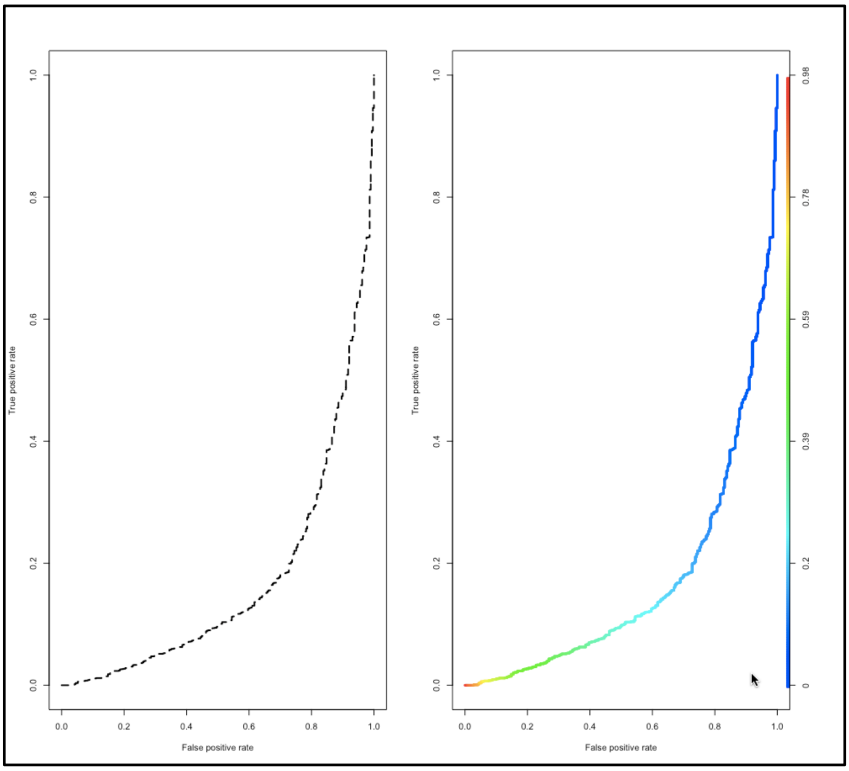
```
[1] 0.1703347
```

The performance is the validation of the train predicted value. The AUC – Area under curve is used to predict the values as 17% and the AUC predicted performance can be achieved with the customer not using service of the telecom.

```
> perf=performance(roctrainpredict,"tpr","fpr")
```

```
> plot(perf,col = "black",lty=2,lwd=2)
```

```
> plot(perf,lwd=3,colorize=TRUE)
```



The plot shows the graphs of the auc line is predicted for the performance of the train datasets in the true and positive rate of the customer using the service.

```
> roctestpredict=prediction(testpredict,test$Churn)

> as.numeric(performance(roctestpredict,"auc")@y.values)
```

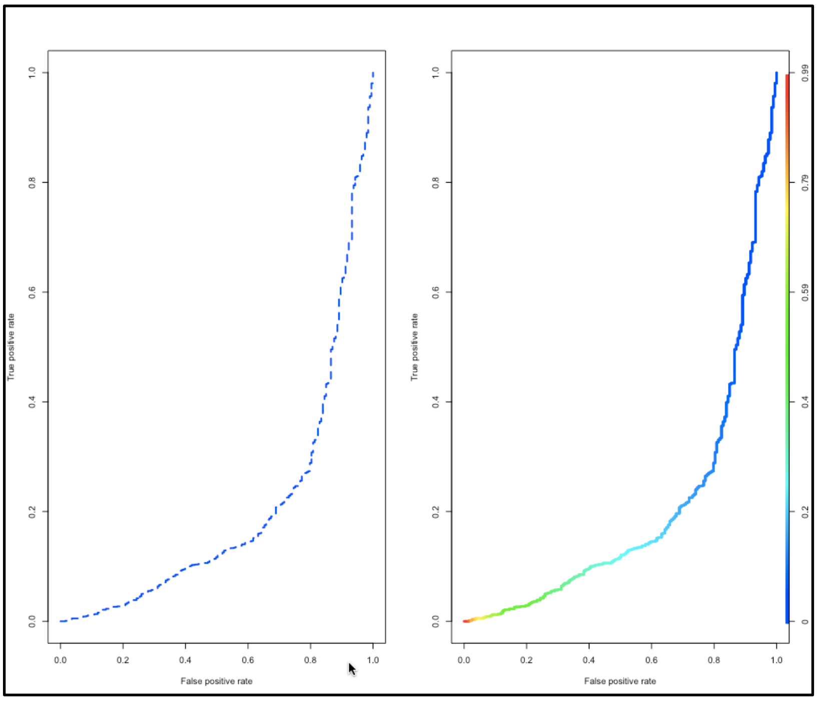
```
[1] 0.2026816
```

The performance measure of the test dataset is showing the values predicted for the performance of the customer churn and the predicted the value of 20% in predicting the measures of the validating the predicted performance.

```
> perf=performance(roctestpredict,"tpr","fpr")

> plot(perf,col = "blue",lty=2,lwd=2)

> plot(perf,lwd=3,colorize=TRUE)
```



The plots showing in the auc curve is predicted as the line is cut-off to the value on the 0.2 and the values are performance of the y values in the test predicted values. The performance values are measured with the nearer values as predicted in the train and test datasets from the customer churn.

```
> blr_step_aic_both(cellphonemodel,details = FALSE)
```

The step wise selection is predicted for the customer churn and the values are measured for the variables used on the 10 observations. In the step wise prediction is used for the variables removed or the variables selected for the step wise selection of the cross validation for the full model observation.

Stepwise Summary

Variable	Method	AIC	BIC	Deviance
ContractRenewal	addition	1561.82	1573.022	1557.82
CustServCalls	addition	1456.639	1473.441	1450.639
DayMins	addition	1356.441	1378.844	1348.441
DataPlan	addition	1328.968	1356.972	1318.968
OverageFee	addition	1306.682	1340.287	1294.682
RoamMins	addition	1302.846	1342.052	1288.846
DayCalls	addition	1300.948	1345.755	1284.948

The deviance values is predicted for the values with same predicted values for the day time calls and the overage fees for the stepwise summary for the values predicted in the AIC and BIC values are measured from the customer churn values.

```
> finalcellphonemodel=glm(Churn~ContractRenewal+CustServCalls+DayMins+DataPlan

+                 +OverageFee+RoamMins+DayCalls,data = train,family = binomial(link = "logit"))

> summary(finalcellphonemodel)
```

The final model is predict for the function using logit and the log with the linear discriminant of the values are measured with the logistics and linear values.

Min	1Q	Median	3Q	Max
-1.9181	-0.4973	-0.3369	-0.1922	2.9934

Maximum deviance values are predicted for the maximum values are measured with 2.9934.

The co-efficient are measured for the values are predicted for the variables with higher significance to the lesser significance for the measured variables.

		Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)		-6.390481	0.67374	-9.485	< 2e-16	***
ContractRenewal	Customer renewed	-2.007461	0.187236	-10.722	< 2e-16	***
CustServCalls		0.553399	0.05102	10.847	< 2e-16	***
DayMins		0.013514	0.001402	9.638	< 2e-16	***
DataPlan	Customer have data	-1.017914	0.192172	-5.297	1.18E-07	***
OverageFee		0.149627	0.029185	5.127	2.95E-07	***
RoamMins		0.063028	0.026696	2.361	0.0182	*
DayCalls		0.007007	0.003561	1.968	0.0491	*

The variables – Contract renewal, customer service calls, day time usage, data plan, overage fees are measured with higher significance for the values predicted with the customer churn for the final cell phone model in the z values. The z values are the co-efficient are measured with the other variables for the predicted values.

The AIC code is predicted with 300.9 and the values are measured with the values predicted for the various variables.

```
> print(exp(finalcellphonemodel$coefficients))
```

(Intercept)	ContractRenewalCustomer renewed	CustServCalls
0.00167745	0.1343293	1.73915432
DayMins	DataPlanCustomer have data	OverageFee
1.01360584	0.36134778	1.1614007
RoamMins	DayCalls	
1.06505683	1.00703197	

The coefficient are measured for the values with customer renewed the service and the customer have data for the prediction on the customer churn for the customer cancelled the service. The values are measured for the exponential values which are predicted in the values with the step wise prediction for the higher expectation in expensive values.

```
> blr_rsq_mcfadden(finalcellphonemodel)
```

```
[1] 0.2239475
```

```
> blr_rsq_mcfadden_adj(finalcellphonemodel)
```

```
[1] 0.2142842
```

The McFadden values are predicted for the squared values of the final model of the customer churn with the values of 22% and the adjusted McFadden value is computed as 21% are predicted for the variables.

```
> pR2(finalcellphonemodel)
```

llh	llhNull	G2	McFadden	r2ML	r2CU
-642.4738769	-827.8742608	370.8007677	0.2239475	0.1692284	0.3005711

the pseudo squared shows the predicted final model is produced the variables are predicting McFadden, G2 and the various predicting variables. The values are predicting the most of the Null deviance values in negative squared of the log function in the final model.

```
> myroc=roc(train$Churn,trainpredict)
```

```
> coords(myroc,"best",ret = "threshold")
```

```
  threshold

1 0.1770099
```

Threshold is the measures of the values are predicting the full mode for the variables in train predicted function with 17% of the values.

```
> myroc1=roc(test$Churn,testpredict)

> coords(myroc1,"best",ret = "threshold")
```

```
  threshold

1 0.1273552
```

Threshold is measuring the predicted values for the validating datasets to the value of 12% in the various prediction of the final model.

```
> table(train$Churn,trainpredict>0.177)
```

	FALSE	TRUE
Customer Using Service	1393	317
Customer Cancelled Service	79	211

The table is compared for the train predicted for the value is measures the values for the customer using service and the customer cancelled service in the threshold level above 0.177 and the measures in the predicted values.

```
> (1393+317)/nrow(na.omit(train))
```

```
[1] 0.855
```

the values are predicted with 85.5% of the regression value measured for the values in the positive rates.

```
> table(test$Churn,testpredict>0.127)
```

	FALSE	TRUE
Customer Using Service	825	315
Customer Cancelled Service	39	154

The table is compared with the customer using the service and the customer renewed service for the prediction in the values are measured above the value of threshold points is above 0.127.

```
> (825+315)/nrow(na.omit(test))
```

```
[1] 0.8552138
```

The predicted test values are measured for the higher related to the final cell phone model is validated with 85.5% in the values.

```
> KS=max(attr(perf,'y.values')[[1]]-attr(perf,'x.values')[[1]])

> KS
```

```
[1] 0
```

The KS chart shows the value is predicted as 0 and the values are negative with the models. Hence, the model is predicted from the performance of the variables in the attributes of the churn data.

```
> AUC=performance(roctrainpredict,"auc")

> AUC
```

The object and the class of the performance predict variables are measured for the null value and the values measured for the performance measures is Area Under the Curve for the customer data. The AUC value is predicted for the performance of 17% in the measures for the values predicted for the customer predicted variables.

```
> AUC=performance(roctestpredict,"auc")

> AUC
```

The object and class variables are predicted for the various class and the models are measured for the AUC in the performance values is 20% and the values are measured for the null value and the y variables are predicted for the AUC.

```
> GINI=ineq(train$Churn,type = "Gini")
```

```
> GINI
```

```
[1] 0.1082751
```

The GINI values predicts the variables are measured for the train set with the performance measures of 10.8%.

```
> GINI=ineq(test$Churn,type = "Gini")
```

```
> GINI
```

```
[1] 0.1081627
```

The GINI co-efficient is measured for the variables for predicting in the validation data sets for the values are measured for the various predictions with the same 10.8%.

```
> confusionMatrix(test$Churn,sample(test$Churn))
```

Prediction	Customer Using Service	Customer Cancelled Service
Customer Using Service	975	165
Customer Cancelled Service	165	28

The predicted confusion matrix for the validation dataset is measured for the percentage occurs with various model performances.

Accuracy: 0.7524
95%CI: (0.7284,0.7754)
No Information Rate: 0.8552
P-Value[Acc>NIR]: 1
Kappa: 3e-04
Mcnemar's Test P-Value: 1
Sensitivity: 0.8553
Specificity: 0.1451
Pos Pred Value: 0.8553
Neg Pred Value: 0.1451
Prevalence: 0.8552
Detection Rate: 0.7314
Detection Prevalence: 0.8552
Balanced Accuracy: 0.5002

The sensitivity and specificity, kappa values are mentioned with the most predicted variables are measuring the values for the accuracy with 75% of the validation data are measuring for the various predicted with various variables.

```
> confusionMatrix(train$Churn,sample(train$Churn))
```

Prediction	Customer Using Service	Customer Cancelled Service
Customer Using Service	1462	248
Customer Cancelled Service	248	42

The values are measured for the customer using service and the customer cancelled service is matrix by the train and the train churn data.

Accuracy: 0.752
95%CI: (0.7325,0.7708)
No Information Rate: 0.855
P-Value [Acc>NIR]: 1
Kappa:-2e-04
Mcnemar's Test P-Value:1
Sensitivity: 0.8550
Specificity: 0.1448
Pos Pred Value: 0.8550
Neg Pred Value: 0.1448
Prevalence:0.8550
Detection Rate:0.7310
Detection Prevalence:0.8550
Balanced Accuracy:0.4999

The accuracy values is predicted with the values are measuring the sensitivity and the specificity are measured for the highest prediction with the Kappa values for the various models. The measuring information rate will be measured with the value of 85.5% in the values are predicted for the confusion matrix in the values.

Interpretation

- Customer Data is splits into train and test datasets for the model performance measures and train data are measured using the various model performance like AUC, GINI, ROC and KS for the various dataset.
- The AUC values are predicted for the model performance for the train datasets is computed by 17% and the values predicted for the test datasets is predicted by 20% which is inferring the customer service is availing the service. The customers are splitting the variables in the total model for the variables with relation to the churn.
- Interpretation of the variables are existing the values with the customer not using the service for the various prediction on the various values. The validation of the customer churn is predicting the linear models for the cell phone churn to produce the variables with increasing to the negative growth i.e. the values are measured as the customer is not availing the service.
- Gini impurity is the co-efficient is based on the values predict for the customer churn in the specificity and sensitivity is making the test dataset and train datasets is measuring the various variables with 10.8%. The Gini-Coefficient is proving that customers are not using the service.
- The KS chart is created for the measuring the variables co-efficient of customer churn is measured with the values of interpreting with the customer churn. The customer churn is based on the customer variables and achieved with the values are measured as the customer are not using the services with the various interpretations.
- Model interpretation of the variables are predicted with customer churn and the model like Gini impurity is showing the results like 10.8% which is least value of the customers are engaged for the service and the customer is interpreting with the less chance to increase the customer usage of the telecom service. Hence, the customer using service are not having the best vision to increase the usage of the telecom service.

9. Interpretation & Recommendations from the best model

The best model predicted for the customer churn is AUC as the model is predicting the both test and train model. The AUC values for the prediction of the train model is 17% and the validation is predicted with 20% for the customer churn related to the variables.

The variables predicted for the interpretation are increased for the best model for the interpretation is predicted with the customer churn is not sufficient in predicting the customer using the service. The customers are predicted based on the variables interpreted for the logistics regression. The validation of the AUC brings the 20% accuracy in the data for the higher probability to increase the various variables in the each models.

The model predicting the Gini impurity are based on the interpretation of the samples picked for the analysis. The variables are predict the data in the prediction level at 10% accuracy in train and validation datasets.

The best model in predicting the systematic procedures for the customer churn is based on the interpretation of the KS which is enriched with the exponential factors for the values are predicted and the customer are cancelled the service.

The model interpretation gives the best accuracy for the various variables in AUC, GINI and KS are providing the customer churn is not responsible for the factors providing the customer validated in the variables. The models predicted are interpreting the customers are not using the service of the telecom service in the models.

The recommendations for the model performance are correlated for the customer not using service in the cell phone dataset in the various variables. The customer churn with the variables are measured with the predict of the 60% which is averagely monitored for the model and the customers are not using the service from the telecom service.

Model interpretations are measuring the models for the variables in the linear discriminant analysis for the customer churn and the customer variables are predicting the values with 58% of the customers are using the service and average of the customer churn are not using the service and the variables are measured with the general analysis for the prediction of the models.

The logistics regression shows that the customer churn is based on the values to consolidate with the customer is not using service at the predicted scores of the customer data to product the customer valid usage of the data form the telecom company.

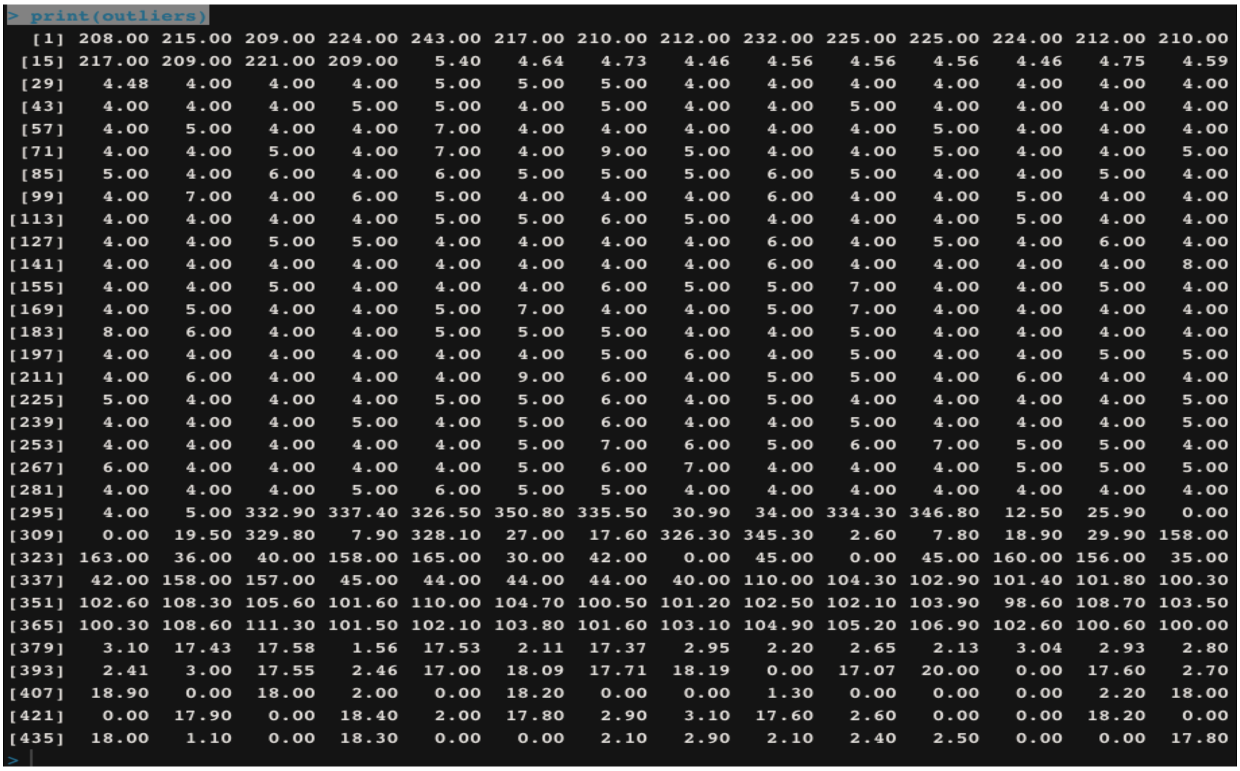
The models created for the customer churn are predicted from the cost effective of the average use of the data used by the customer hence the recommendations are brings with the data usage of the customer services.

3.5 Outlier Identification

The outlier identification for the temperature is found as the maximum values in the variables.

```
> outliers = boxplot(num.data,plot = FALSE)$out
```

```
> print(outliers)
```



```
> print(outliers)
  [1] 208.00 215.00 209.00 224.00 243.00 217.00 210.00 212.00 232.00 225.00 225.00 224.00 212.00 210.00
 [15] 217.00 209.00 221.00 209.00   5.40   4.64   4.73   4.46   4.56   4.56   4.56   4.46   4.75   4.59
 [29]   4.48   4.00   4.00   4.00   5.00   5.00   5.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00
 [43]   4.00   4.00   4.00   5.00   5.00   4.00   5.00   4.00   4.00   5.00   4.00   4.00   4.00   4.00
 [57]   4.00   5.00   4.00   4.00   7.00   4.00   4.00   4.00   4.00   4.00   5.00   4.00   4.00   4.00
 [71]   4.00   4.00   5.00   4.00   7.00   4.00   9.00   5.00   4.00   4.00   5.00   4.00   4.00   5.00
 [85]   5.00   4.00   6.00   4.00   6.00   5.00   5.00   5.00   6.00   5.00   4.00   4.00   5.00   4.00
 [99]   4.00   7.00   4.00   6.00   5.00   4.00   4.00   4.00   6.00   4.00   4.00   5.00   4.00   4.00
[113]   4.00   4.00   4.00   4.00   5.00   5.00   6.00   5.00   4.00   4.00   4.00   5.00   4.00   4.00
[127]   4.00   4.00   5.00   5.00   4.00   4.00   4.00   4.00   6.00   4.00   5.00   4.00   6.00   4.00
[141]   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00   6.00   4.00   4.00   4.00   4.00   8.00
[155]   4.00   4.00   5.00   4.00   4.00   4.00   6.00   5.00   5.00   7.00   4.00   4.00   5.00   4.00
[169]   4.00   5.00   4.00   4.00   5.00   7.00   4.00   4.00   5.00   7.00   4.00   4.00   4.00   4.00
[183]   8.00   6.00   4.00   4.00   5.00   5.00   5.00   4.00   4.00   5.00   4.00   4.00   4.00   4.00
[197]   4.00   4.00   4.00   4.00   4.00   4.00   5.00   6.00   4.00   5.00   4.00   4.00   5.00   5.00
[211]   4.00   6.00   4.00   4.00   4.00   9.00   6.00   4.00   5.00   5.00   4.00   6.00   4.00   4.00
[225]   5.00   4.00   4.00   4.00   5.00   5.00   6.00   4.00   5.00   4.00   4.00   4.00   4.00   5.00
[239]   4.00   4.00   4.00   5.00   4.00   5.00   6.00   4.00   4.00   5.00   4.00   4.00   4.00   5.00
[253]   4.00   4.00   4.00   4.00   4.00   5.00   7.00   6.00   5.00   6.00   7.00   5.00   5.00   4.00
[267]   6.00   4.00   4.00   4.00   4.00   5.00   6.00   7.00   4.00   4.00   4.00   5.00   5.00   5.00
[281]   4.00   4.00   4.00   5.00   6.00   5.00   5.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00
[295]   4.00   5.00 332.90 337.40 326.50 350.80 335.50  30.90  34.00 334.30 346.80  12.50  25.90   0.00
[309]   0.00  19.50 329.80   7.90 328.10  27.00  17.60 326.30 345.30   2.60   7.80  18.90  29.90 158.00
[323] 163.00  36.00  40.00 158.00 165.00  30.00  42.00   0.00  45.00   0.00  45.00 160.00 156.00  35.00
[337]  42.00 158.00 157.00  45.00  44.00  44.00  44.00  40.00 110.00 104.30 102.90 101.40 101.80 100.30
[351] 102.60 108.30 105.60 101.60 110.00 104.70 100.50 101.20 102.50 102.10 103.90  98.60 108.70 103.50
[365] 100.30 108.60 111.30 101.50 102.10 103.80 101.60 103.10 104.90 105.20 106.90 102.60 100.60 100.00
[379]   3.10  17.43  17.58   1.56  17.53   2.11  17.37   2.95   2.20   2.65   2.13   3.04   2.93   2.80
[393]   2.41   3.00  17.55   2.46  17.00  18.09  17.71  18.19   0.00  17.07  20.00   0.00  17.60   2.70
[407]  18.90   0.00  18.00   2.00   0.00  18.20   0.00   0.00   1.30   0.00   0.00   0.00   2.20  18.00
[421]   0.00  17.90   0.00  18.40   2.00  17.80   2.90   3.10  17.60   2.60   0.00   0.00  18.20   0.00
[435]  18.00   1.10   0.00  18.30   0.00   0.00   2.10   2.90   2.10   2.40   2.50   0.00   0.00  17.80
> 
```

3.6 Variable Transformation/ Feature creation

The variables transformation are created with the categorical variables and the numeric variables are diversified for the various prediction in the customer churn.

```
> Cellphone$Churn=factor(Cellphone$Churn,levels = c(0,1),

+                  labels = c("Customer Using Service","Customer Cancelled Service"))

> Cellphone$ContractRenewal=factor(Cellphone$ContractRenewal,levels = c(0,1),

+                  labels = c("Customer not renewed","Customer renewed"))

> Cellphone$DataPlan=factor(Cellphone$DataPlan,levels = c(0,1),

+                         labels = c("Customer with no data","Customer have data"))
```

The category variables are measured with the factor levels for the customer using service, customer renewed service and customer have and customer with data.

4. Conclusion

Telecom Company’s customer data are predicted for the customer using service and customer not using service. The predicted business model of the customer data are showing results with the customers are looking for the better month rates and the data plans for their comfort of using the telecom network service. Customer churn are predicted the data and the mobile usage is decreasing and the average customer nearly 40% are cancelled the service and the customer churn is need to be increased in the various customer satisfaction of the telecom providers. Hence, the customer using the telecom service are in place of cancel their service in future from the telecom usage.

5. Appendix

```
setwd("/Users/numerp/Documents/PGP-BABI/Module 5 Predictive Modelling/Project 4")
getwd()
library(readr)
library(readxl)
library(dplyr)
library(psych)
library(car)
library(carData)
library(ggplot2)
library(mice)
library(lattice)
library(nFactors)
library(scatterplot3d)
library(data.table)
library(tidyverse)
library(broom)
library(GGally)
Cellphone=read_xlsx("Cellphone.xlsx",sheet = 2,col_names = TRUE)
head(Cellphone,5)
tail(Cellphone,3)
summary(Cellphone)
str(Cellphone)
attach(Cellphone)
View(Cellphone)
Cellphone$Churn=factor(Cellphone$Churn,levels = c(0,1),
              labels = c("Customer Using Service","Customer Cancelled Service"))
Cellphone$ContractRenewal=factor(Cellphone$ContractRenewal,levels = c(0,1),
              labels = c("Customer not renewed","Customer renewed"))
Cellphone$DataPlan=factor(Cellphone$DataPlan,levels = c(0,1),
                    labels = c("Customer with no data","Customer have data"))
summary(Cellphone)
str(Cellphone)
any(is.na(Cellphone))
Cellphone=na.omit(Cellphone)
dim(Cellphone)
ggpairs(Cellphone[,c("AccountWeeks","DataUsage","CustServCalls","DayMins","DayCalls",
             "MonthlyCharge","OverageFee","RoamMins")],
     ggplot2::aes(colour=as.factor(Cellphone$Churn)))
ct.data=subset(Cellphone,select = c(Churn,ContractRenewal,DataPlan))
num.data=subset(Cellphone,select = -c(Churn,ContractRenewal,DataPlan))
dim(ct.data)
dim(num.data)
outliers = boxplot(num.data,plot = FALSE)$out
print(outliers)
names(ct.data)
names(num.data)
par(mfrow=c(2,4))
boxplot(AccountWeeks,horizontal = TRUE,
      main = "Active Customers",col = "blue")
boxplot(DataUsage,horizontal = T,
      main = "Gigabytes used by the customers",col = "green")
boxplot(CustServCalls,horizontal = T,
      main = "Number of Calls received by Customer Service",col = "red")
boxplot(DayMins,horizontal = T,
      main = "Customer using Data in Day Time",col = "gold")
boxplot(DayCalls,horizontal = T,
      main = "Customer making calls in Day Time",col = "orange")
boxplot(MonthlyCharge,horizontal = T,
      main = "Customer average month rate",col = "grey")
boxplot(OverageFee,horizontal = T,
      main = "Customer overage fees for last 12 months",col = "brown")
boxplot(RoamMins,horizontal = T,
      main = "Customer average roaming minutes",col = "yellow")
par(mfrow=c(2,4))
hist(AccountWeeks,col = "blue")
hist(DataUsage,col = "green")
hist(CustServCalls,col = "red")
hist(DayMins,col = "gold")
hist(DayCalls,col = "orange")
hist(MonthlyCharge,col = "grey")
hist(OverageFee,col = "brown")
hist(RoamMins,col = "yellow")
#Based on Customer Cancelled Service or not
lr.reg=glm(Churn~AccountWeeks,data = Cellphone,family = binomial)
summary(lr.reg)
lr.reg=glm(Churn~DataUsage,data = Cellphone,family = binomial)
summary(lr.reg)
lr.reg=glm(Churn~CustServCalls,data = Cellphone,family = binomial)
summary(lr.reg)
lr.reg=glm(Churn~DayMins,data = Cellphone,family = binomial)
summary(lr.reg)
lr.reg=glm(Churn~DayCalls,data = Cellphone,family = binomial)
summary(lr.reg)
lr.reg=glm(Churn~MonthlyCharge,data = Cellphone,family = binomial)
summary(lr.reg)
lr.reg=glm(Churn~OverageFee,data = Cellphone,family = binomial)
summary(lr.reg)
lr.reg=glm(Churn~RoamMins,data = Cellphone,family = binomial)
summary(lr.reg)
model=glm(Churn~.,data = Cellphone,family = binomial)
summary(model)
car::vif(model)
```

```
par(mfrow=c(1,4))
plot(model)
par(mfrow=c(1,3))
for (a in names(ct.data)) {
  print(a)
  print(round(prop.table(table(Cellphone$Churn,ct.data[[a]])),digits = 3)*100)
  barplot(table(Cellphone$Churn,ct.data[[a]]),
          col = c("blue","yellow"),
          main = names(ct.data[a]))
}
fullmodel=cbind(ct.data,num.data)
names(fullmodel)
names(Cellphone)
#Split train and test dataset
library(caTools)
set.seed(100)
splits=sample.split(fullmodel$Churn,SplitRatio = 0.60)
train=subset(fullmodel,splits==T)
test=subset(fullmodel,splits==F)
prop.table(table(fullmodel$Churn))
prop.table(table(train$Churn))
prop.table(table(test$Churn))
#Train Dataset
cellphonemodel=glm(Churn~.,data = train,family = binomial)
summary(cellphonemodel)
car::vif(cellphonemodel)
library(lmtest)
lrtest(cellphonemodel)
library(pscl)
pR2(cellphonemodel)["McFadden"]
logLik(cellphonemodel)
trainmodel1=glm(Churn~1,data = train,family = binomial)
1-(logLik(cellphonemodel)/logLik(trainmodel1))
logLik(trainmodel1)
#Prediction on train dataset
trainpredict=predict(cellphonemodel,newdata = train,type = "response")
table(train$Churn,trainpredict>0.5)
(1659+51)/nrow(na.omit(train))
#Prediction on test dataset
testpredict=predict(cellphonemodel,newdata = test,type = "response")
table(test$Churn,testpredict>0.5)
(1111+29)/nrow(na.omit(test))
#Logistics Regression
library(rms)
vif(glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial))
cor(MonthlyCharge,DayMins)
summary(glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial))
```

```
logistic.churn=glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial)
logistic.churn
logistic.churn$fitted.values
exp(0.01338)
churn.predicted=ifelse(logistic.churn$fitted.values<0.20,
                "Customer with no service","Customer with Service")
table(Churn,churn.predicted)
pROC::roc(Churn,logistic.churn$fitted.values)
par(mfrow=c(1,1))
pROC::plot.roc(Churn,logistic.churn$fitted.values)
#Linear Discriminant Analysis
library(MASS)
lda.churn=lda(Churn~MonthlyCharge+DayMins,data = Cellphone,CV=TRUE)
lda.churn
lda(Churn~MonthlyCharge+DayMins,data = Cellphone)
plot(Churn,lda.churn$posterior[,2])
churn.predicted=ifelse(lda.churn$posterior[,2]<0.15,
                "Customer Using Service","Customer Cancelled Service")
churn.predicted
table(Churn,churn.predicted)
pROC::plot.roc(Churn,lda.churn$posterior[,2])
table(Churn,lda.churn$class)
#Naive Bayes
library(e1071)
nb.churn=naiveBayes(Churn~MonthlyCharge+DayMins,data = Cellphone)
nb.churn
nb.churn.predict=predict(nb.churn,type = "raw",newdata = Cellphone)
nb.churn.predict
plot(Churn,nb.churn.predict[,2])
#KNN
library(class)
dim(Cellphone)
set.seed(500)
index=sample(3333,2333)
ktrain=Cellphone[index,]
dim(ktrain)
ktest=Cellphone[-index,]
dim(ktest)
names(ktrain)
kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=5)
table(ktest$Churn,kchurn)
kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=6)
table(ktest$Churn,kchurn)
kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=7)
table(ktest$Churn,kchurn)
kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=8)
table(ktest$Churn,kchurn)
```

```
kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=9)
table(ktest$Churn,kchurn)
kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=10)
table(ktest$Churn,kchurn)
kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=11)
table(ktest$Churn,kchurn)
kchurn=knn(ktrain[,c(7,9)],ktest[,c(7,9)],ktrain$Churn,k=21)
table(ktest$Churn,kchurn)
glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial)
plot(ktest$Churn,
    predict(glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial),
          newdata = ktest,type = "response"))
pROC::roc(ktest$Churn,
        predict(glm(Churn~MonthlyCharge+DayMins,data = Cellphone,family = binomial),
                        newdata = ktest,type = "response"))
#AUC and ROC for Train and Test Dataset
library(ROCR)
roctrainpredict=prediction(trainpredict,train$Churn)
as.numeric(performance(roctrainpredict,"auc")@y.values)
perf=performance(roctrainpredict,"tpr","fpr")
plot(perf,col = "black",lty=2,lwd=2)
plot(perf,lwd=3,colorize=TRUE)
roctestpredict=prediction(testpredict,test$Churn)
as.numeric(performance(roctestpredict,"auc")@y.values)
perf=performance(roctestpredict,"tpr","fpr")
plot(perf,col = "blue",lty=2,lwd=2)
plot(perf,lwd=3,colorize=TRUE)
library(blorr)
blr_step_aic_both(cellphonemodel,details = FALSE)
finalcellphonemodel=glm(Churn~ContractRenewal+CustServCalls+DayMins+DataPlan
                   +OverageFee+RoamMins+DayCalls,data = train,family = binomial(link = "logit"))
summary(finalcellphonemodel)
print(exp(finalcellphonemodel$coefficients))
print(exp(cellphonemodel$coefficients))
blr_rsq_mcfadden(finalcellphonemodel)
blr_rsq_mcfadden_adj(finalcellphonemodel)
pR2(finalcellphonemodel)
library(pROC)
myroc=roc(train$Churn,trainpredict)
coords(myroc,"best",ret = "threshold")
myroc1=roc(test$Churn,testpredict)
coords(myroc1,"best",ret = "threshold")
table(train$Churn,trainpredict>0.177)
(1393+317)/nrow(na.omit(train))
table(test$Churn,testpredict>0.127)
(825+315)/nrow(na.omit(test))
library(ineq)
```

```
KS=max(attr(perf,'y.values')[[1]]-attr(perf,'x.values')[[1]])
KS
AUC=performance(roctrainpredict,"auc")
AUC
AUC=performance(roctestpredict,"auc")
AUC
GINI=ineq(train$Churn,type = "Gini")
GINI
GINI=ineq(test$Churn,type = "Gini")
GINI
library(caret)
confusionMatrix(test$Churn,sample(test$Churn))
confusionMatrix(train$Churn,sample(train$Churn))
```