Mini Project 4 – Predictive Modeling

Submitted by

Sita K

BACP Batch (Dec 19-May 20)

Great Learning

May 8, 2020

Table of Contents

1. Project Objectives 4
2. Assumptions 4
3. Data Analysis – Approach 4
4. Exploratory data analysis 4
a. Check Data Structure 4
b. Check Missing Values 5
c. Plot data to see the distribution 6
i. Univariate Analysis for continuous and categorical variables 6
ii. Bivariate analysis 11
d. Multicollinearity 14
5. Split the data 16
6. Logistic Regression 16
7. KNN 19
8. Naïve Bayes 19
9. Comparison of Models 20
10. Inference 20

Appendix A 21

Table of Figures

FIGURE 1: DATA STRUCTURE 5
FIGURE 2: DATA SUMMARY 5
FIGURE 3: BOXPLOT OF ACCOUNTWEEKS VARIABLE 6
FIGURE 4: BOXPLOT OF DATAUSAGE VARIABLE 6
FIGURE 5: BOXPLOT OF DAYMINS VARIABLE 7
FIGURE 6: BOXPLOT OF DAYCALLS VARIABLE 7
FIGURE 7: BOXPLOT OF MONTHLYCHARGE VARIABLE 8
FIGURE 8: BOXPLOT OF OVERAGEFEE VARIABLE 8
FIGURE 9: BAR CHART OF CHURN VARIABLE 9
FIGURE 10: BAR CHART OF CONTRACTRENEWAL VARIABLE 9
FIGURE 11: BAR CHART OF DATAPLAN VARIABLE 10
FIGURE 12: BAR CHART OF CUSTSERVCALLS VARIABLE 10
FIGURE 13: BOXPLOT ACCOUNTWEEKS VS CHURN 11
FIGURE 14: BOXPLOT OF DATAUSAGE VS CHURN 11
FIGURE 15: BOXPLOT OF DAYMINS VS CHURN 12
FIGURE 16: BOXPLOT OF DAYCALLS VS CHURN 12
FIGURE 17: BOXPLOT OF MONTHLYCHARGES VS CHURN 13
FIGURE 18: BOXPLOT OF OVERAGEFEE VS CHURN 13
FIGURE 19: CORRELATION PLOT 15
FIGURE 20: VIF VALUES 15
FIGURE 21: FINAL VIF 15
FIGURE 22: LOGISTIC REGRESSION MODEL 16
FIGURE 23: ROC PLOT FOR LOGISITIC REGRESSION 17
FIGURE 24: CONFUSION MATRIX - LOGISTIC REGRESSION 17
FIGURE 25: ROC PLOT FOR TESTDATA- LR 18
FIGURE 26: CONFUSION MATRIX FOR TESTDATA-LR 18
FIGURE 27: CONFUSION MATRIX – KNN 19
FIGURE 28: CONFUSION MATRIX - NAIVE BAYES 20
FIGURE 29: VARIABLE IMPORTANCE 21

1. Project Objectives

Customer Churn is a burning problem for Telecom companies. In this project, the data has information about the customer usage behaviour, contract details and the payment details. The data also indicates the customers who have cancelled their services. Based on this past data, we need to build a model which can predict whether a customer will cancel their service in the future or not and provide recommendations to the management.

1. Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it.
2. Data Split: Split the data into test and train, build classification model using Logistic Regression, KNN and Naïve Bayes
3. Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model
4. Final Model: Compare all the model and write an inference which model is best/optimized.
5. Inference: Basis on these predictions, what are the business insights and recommendations

2. Assumptions

- The data provided is conclusive and contains the required data

3. Data Analysis – Approach

1. Environment data setup and data import
2. Calculating the required values using inbuilt functions
3. Apply scaling to the data if required
4. Split the data and build various models
5. Compare the models
6. Provide various recommendations to management

For environment data setup, R's inbuilt packages were used. Also for setting up working directory '**setwd**()' function was used. The given dataset is in .csv format, so we can use **read.csv** function to import the data. All the R commands are in Appendix A.

4. Exploratory data analysis

a. Check Data Structure

The result shows us that there are 3333 observations with 11 variables in the dataset, out of which all are numeric.

```
'data.frame':   3333 obs. of  11 variables:
 $ Churn          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AccountWeeks   : int  128 107 137 84 75 118 121 147 117 141 ...
 $ ContractRenewal: int  1 1 1 0 0 0 1 0 1 0 ...
 $ DataPlan       : int  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  : int  1 1 0 2 3 0 3 0 1 0 ...
 $ DayMins        : num  265 162 243 299 167 ...
 $ DayCalls       : int  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 ...
```

Figure 1: Data Structure

Now let us check the summary of the dataset.
From the below table, by looking at the median and the mean numbers, it gives us an idea that of the data. None of the data seems to be skewed.We will plot the data to see further.

```
     Churn          AccountWeeks    ContractRenewal     DataPlan     
 Min.   :0.0000   Min.   :  1.0   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.: 74.0   1st Qu.:1.0000   1st Qu.:0.0000  
 Median :0.0000   Median :101.0   Median :1.0000   Median :0.0000  
 Mean   :0.1449   Mean   :101.1   Mean   :0.9031   Mean   :0.2766  
 3rd Qu.:0.0000   3rd Qu.:127.0   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :243.0   Max.   :1.0000   Max.   :1.0000  
   DataUsage      CustServCalls      DayMins         DayCalls      MonthlyCharge   
 Min.   :0.0000   Min.   :0.000   Min.   :  0.0   Min.   :  0.0   Min.   : 14.00  
 1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:143.7   1st Qu.: 87.0   1st Qu.: 45.00  
 Median :0.0000   Median :1.000   Median :179.4   Median :101.0   Median : 53.50  
 Mean   :0.8165   Mean   :1.563   Mean   :179.8   Mean   :100.4   Mean   : 56.31  
 3rd Qu.:1.7800   3rd Qu.:2.000   3rd Qu.:216.4   3rd Qu.:114.0   3rd Qu.: 66.20  
 Max.   :5.4000   Max.   :9.000   Max.   :350.8   Max.   :165.0   Max.   :111.30  
   OverageFee       RoamMins    
 Min.   : 0.00   Min.   : 0.00  
 1st Qu.: 8.33   1st Qu.: 8.50  
 Median :10.07   Median :10.30  
 Mean   :10.05   Mean   :10.24  
 3rd Qu.:11.77   3rd Qu.:12.10  
 Max.   :18.19   Max.   :20.00  
```

Figure 2: Data Summary

b. Check Missing Values

There are no missing variables in the data set

c. Plot data to see the distribution

i. Univariate Analysis for continuous and categorical variables

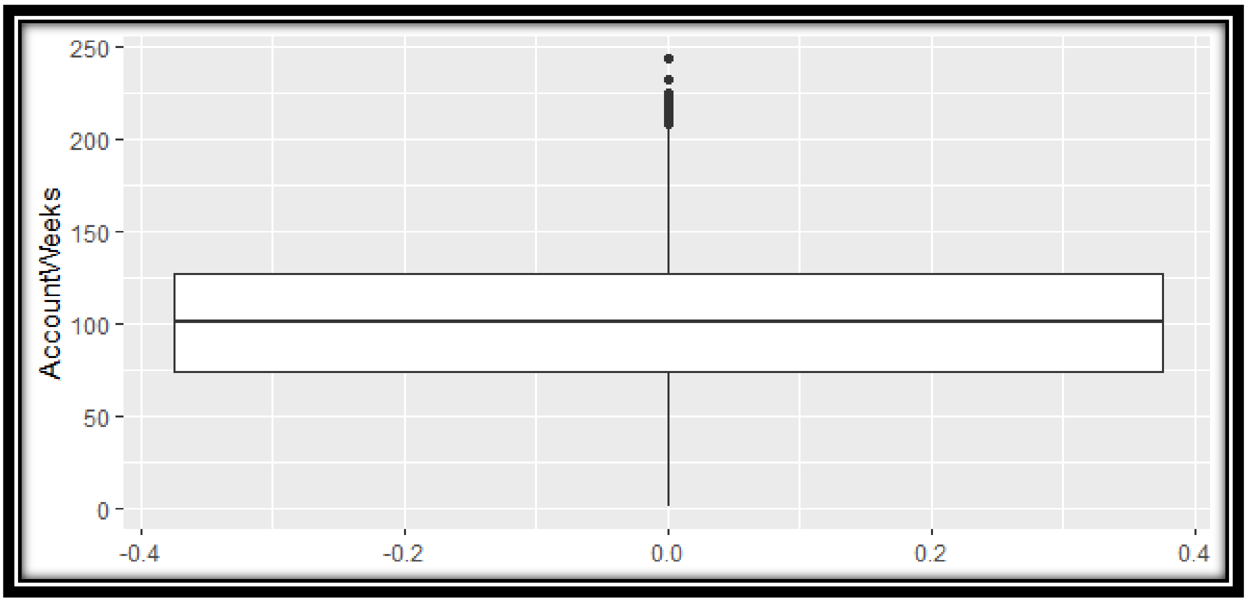


Figure 3: Boxplot of AccountWeeks variable

Accountweeks has some outliers.

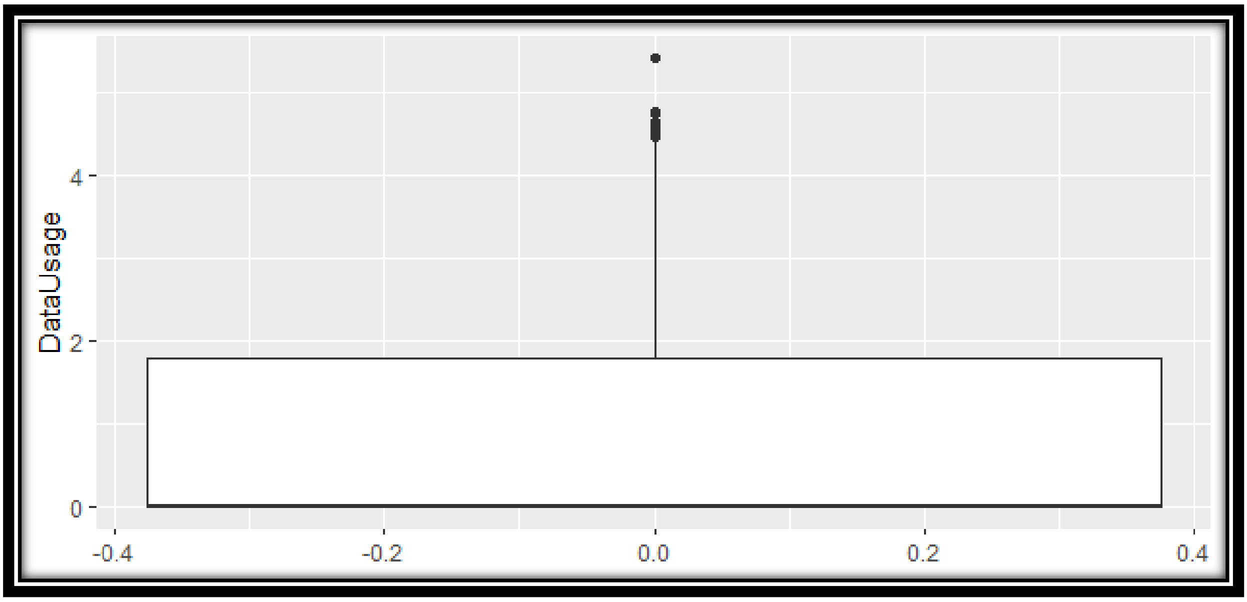


Figure 4: Boxplot of DataUsage variable

DataUsage is right-skewed and has some outliers.

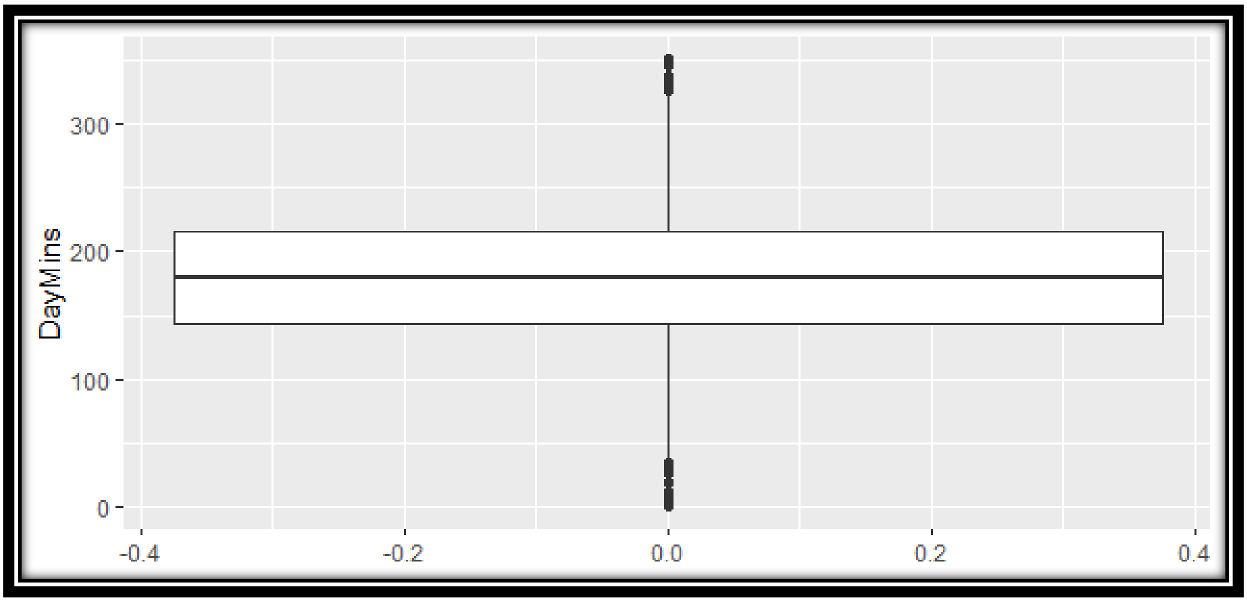


Figure 5: Boxplot of DayMins variable

DayMins has some outliers.

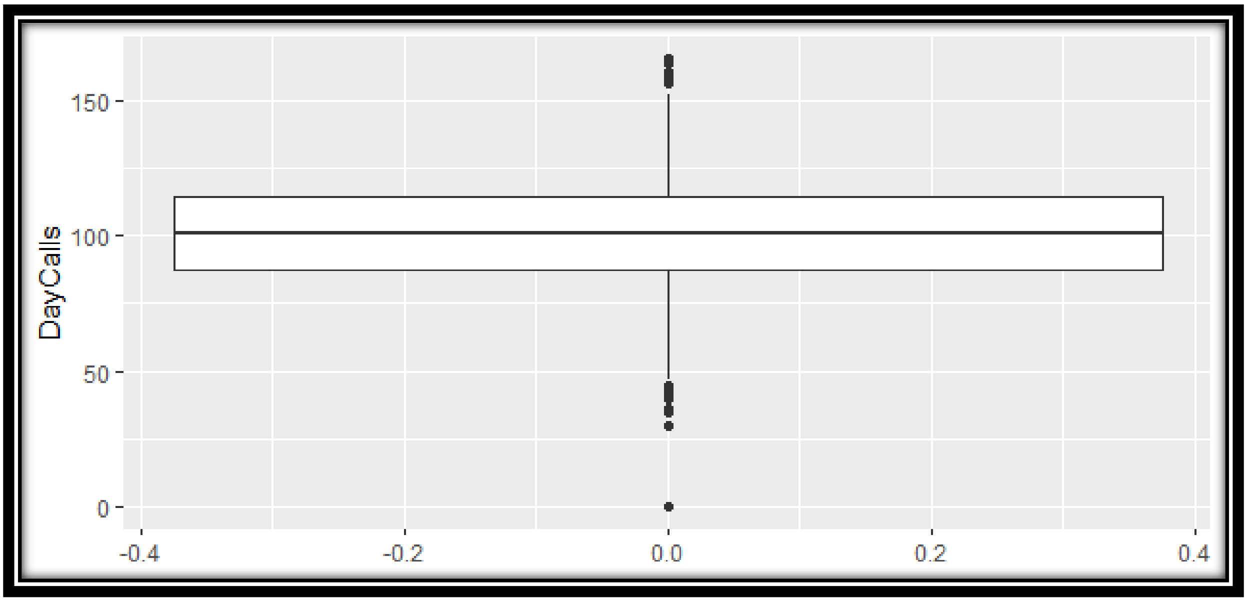


Figure 6: Boxplot of DayCalls variable

DayCalls is slightly left-skewed and has some outliers.

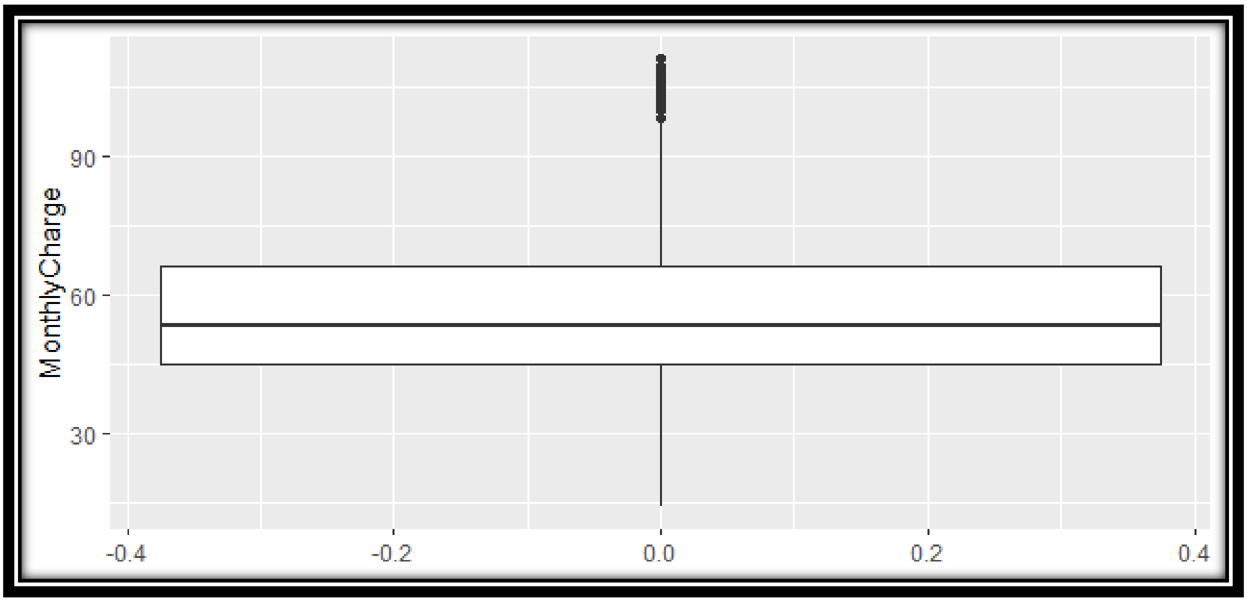


Figure 7: Boxplot of MonthlyCharge variable

MonthlyCharge is slightly right-skewed and has some outliers.

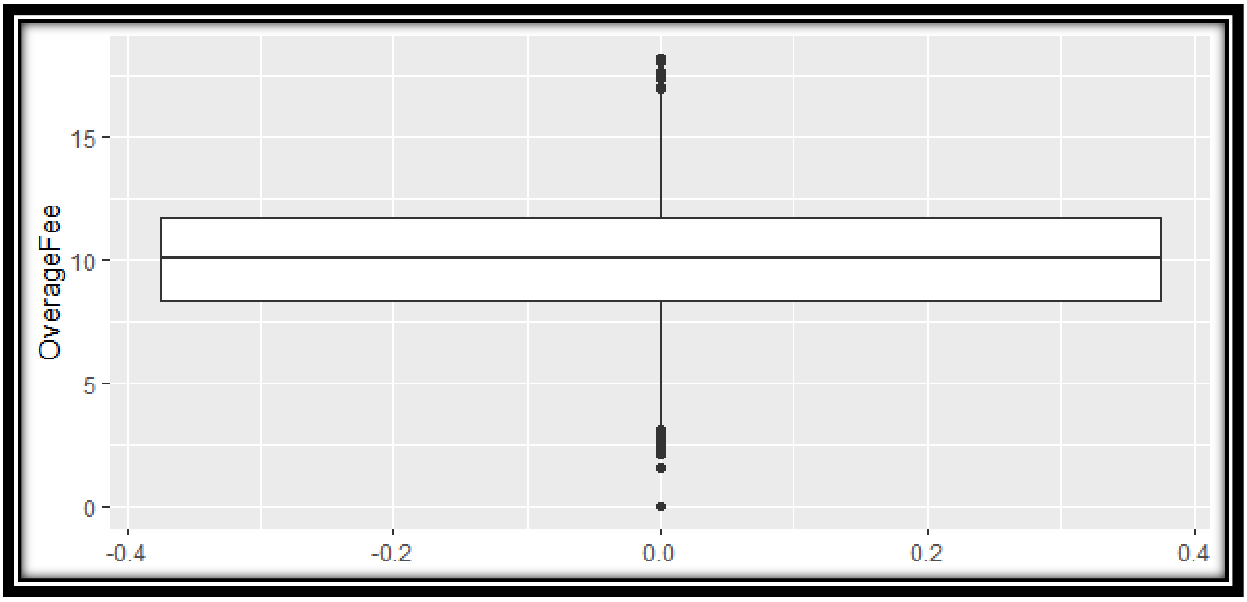


Figure 8: Boxplot of OverageFee variable

OverageFee has some outliers.

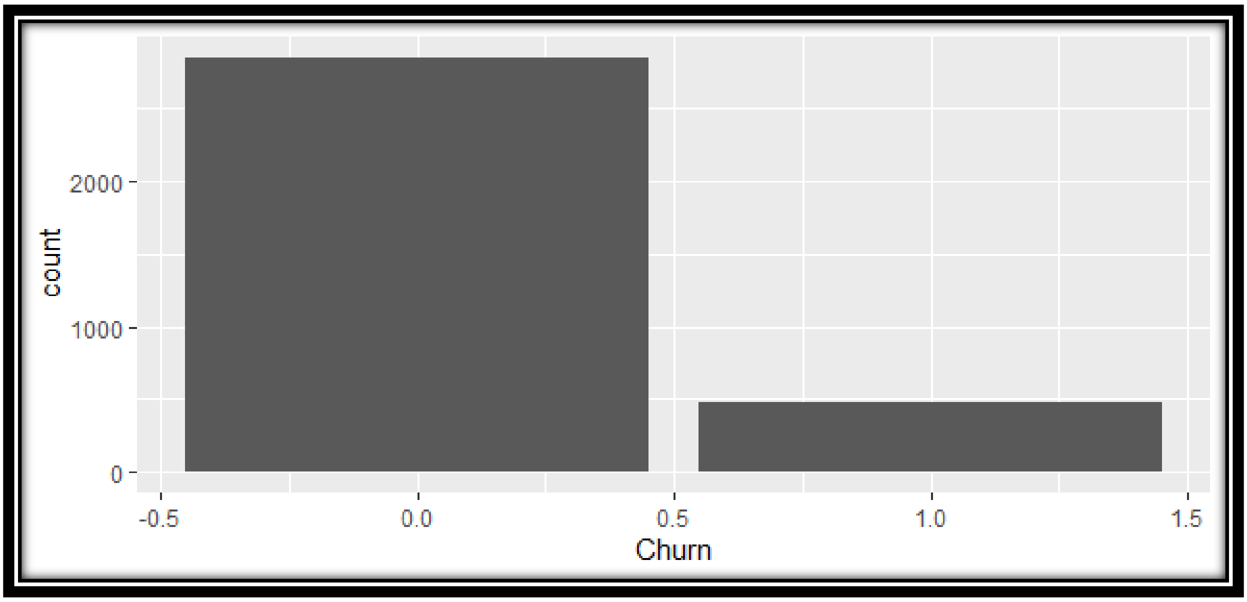


Figure 9: Bar Chart of Churn Variable

There are 2850 cancelled cases and 483 not cancelled ones. That is about 85.5% Churn Ratio which is not good.

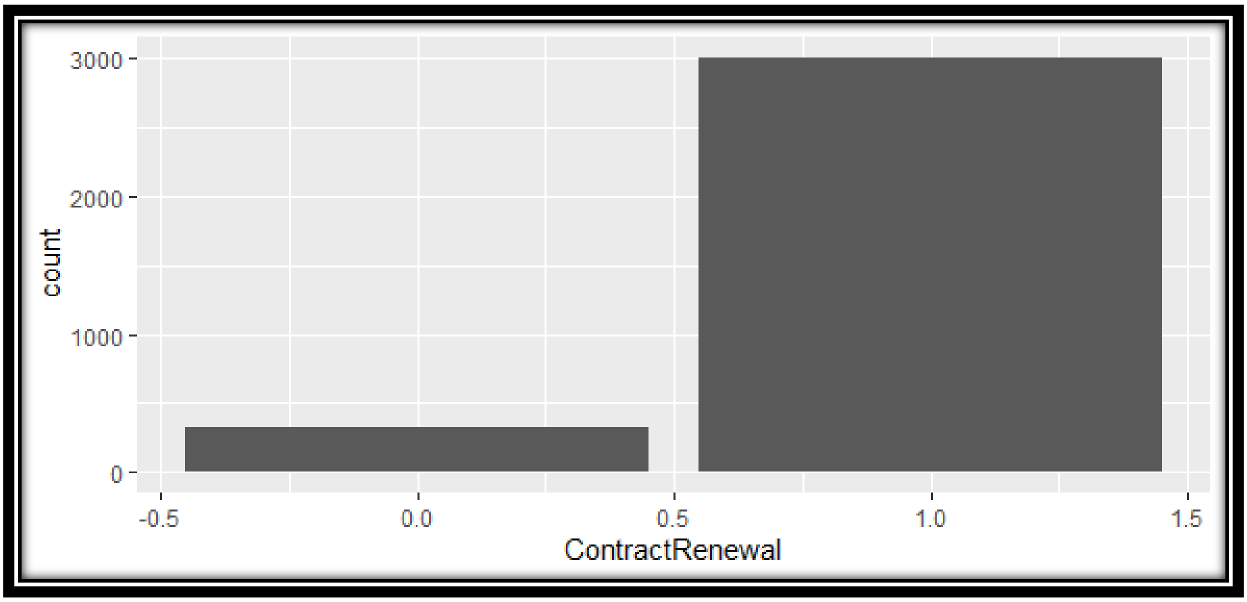


Figure 10: Bar Chart of ContractRenewal Variable

Around 3010 out of 3333 cases did a contract renewal.

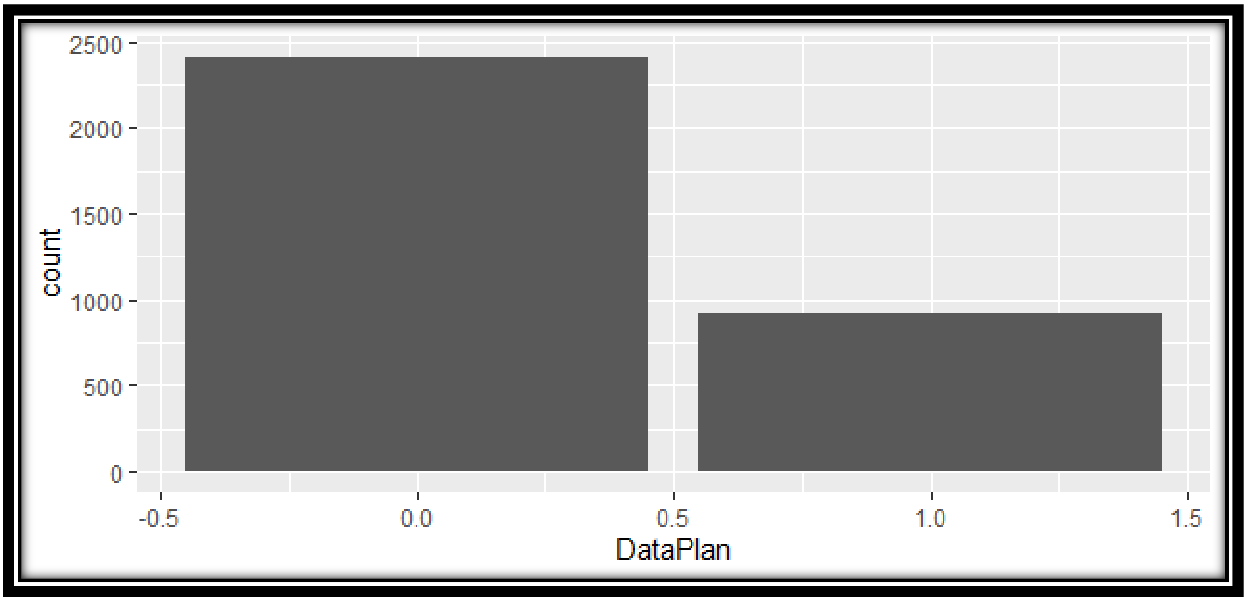


Figure 11: Bar Chart of DataPlan variable

Cases who have not opted for Data Plan are more.

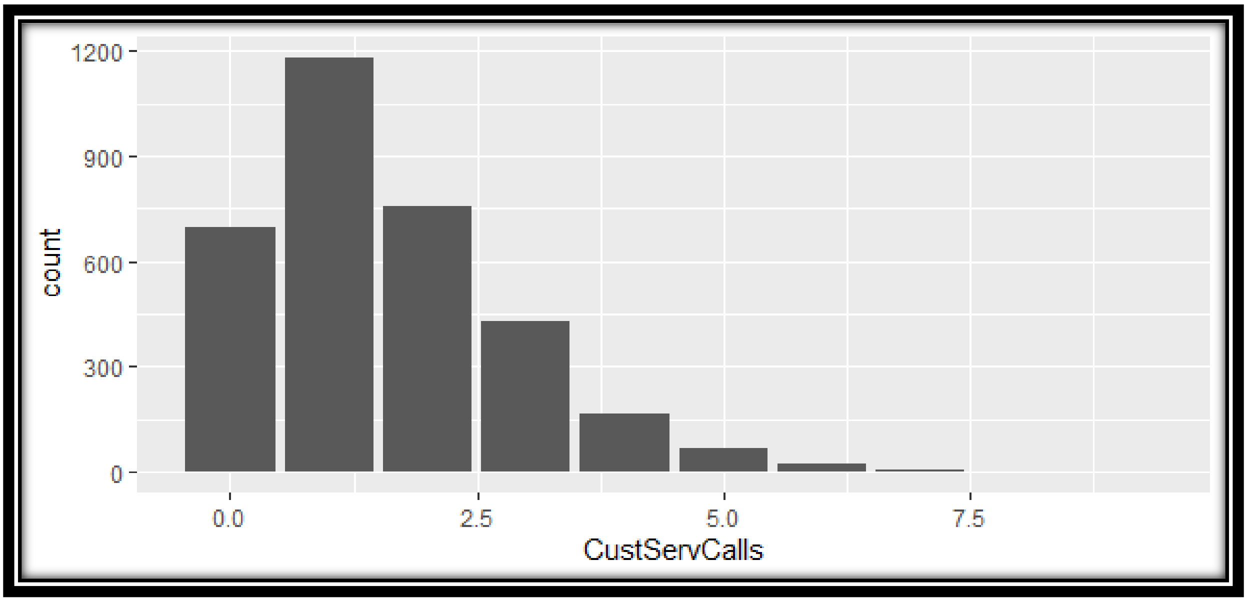


Figure 12: Bar Chart of CustServCalls variable

Most of the customers have called atleast more than once.

ii. Bivariate analysis

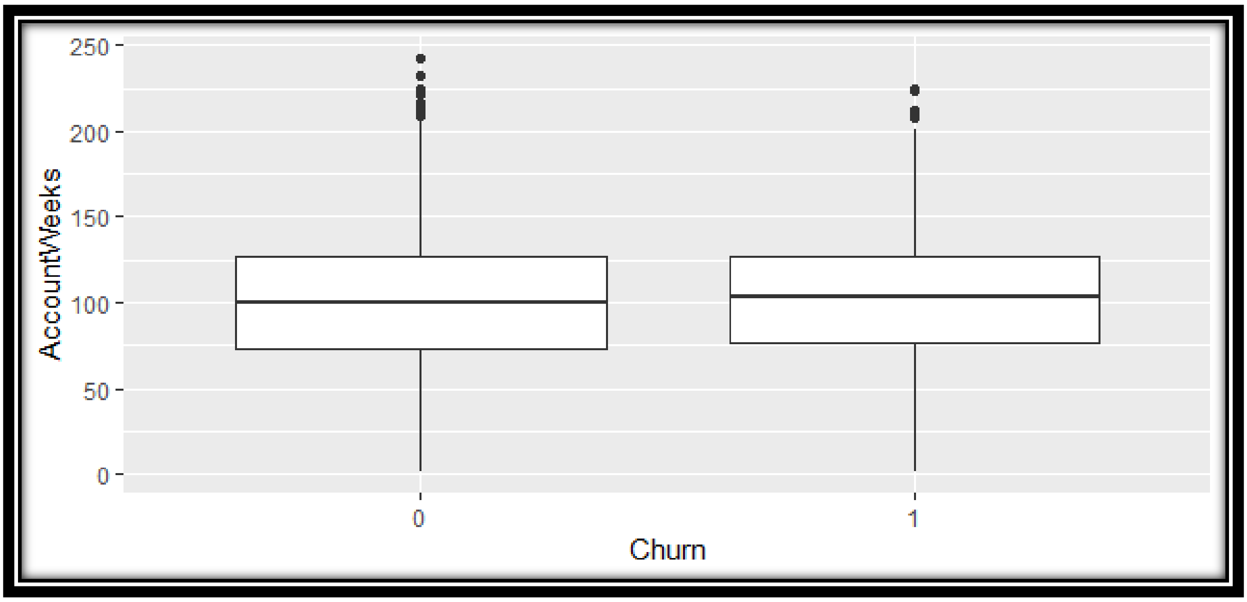


Figure 13: Boxplot AccountWeeks vs Churn

Observation - Median of AccountWeeks for customers with Churn is almost equal to customers with no claim.

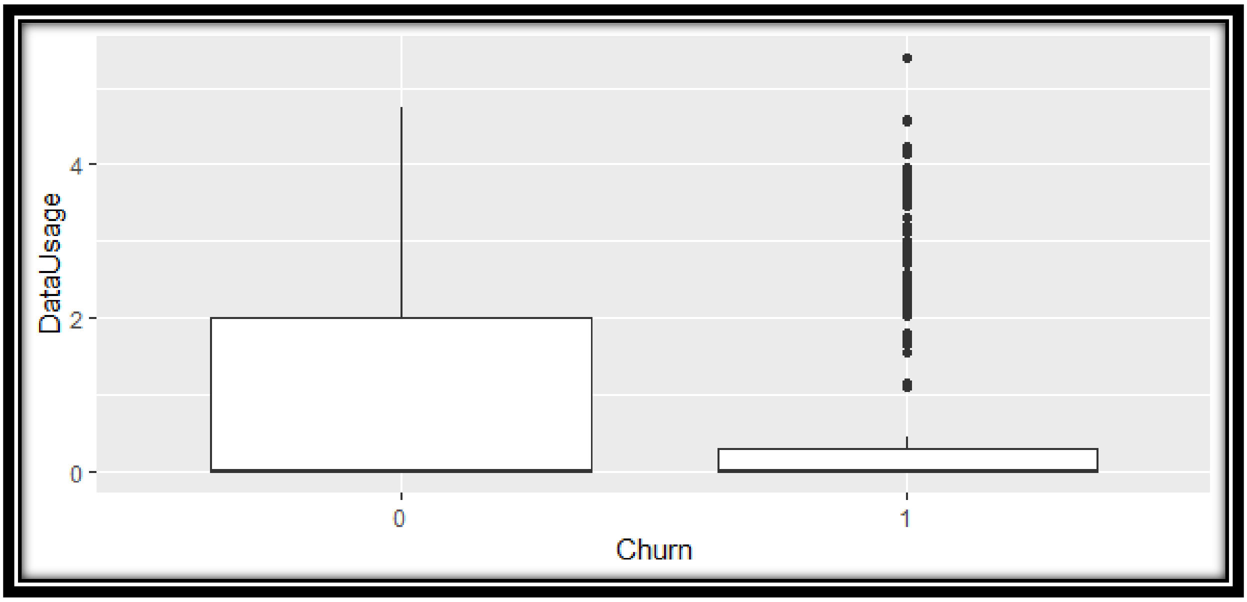


Figure 14: Boxplot of DataUsage vs Churn

Observation – We can see DataUsage with no Churn is on higher side. Thus higher the DataUsage more probability of no Churn.

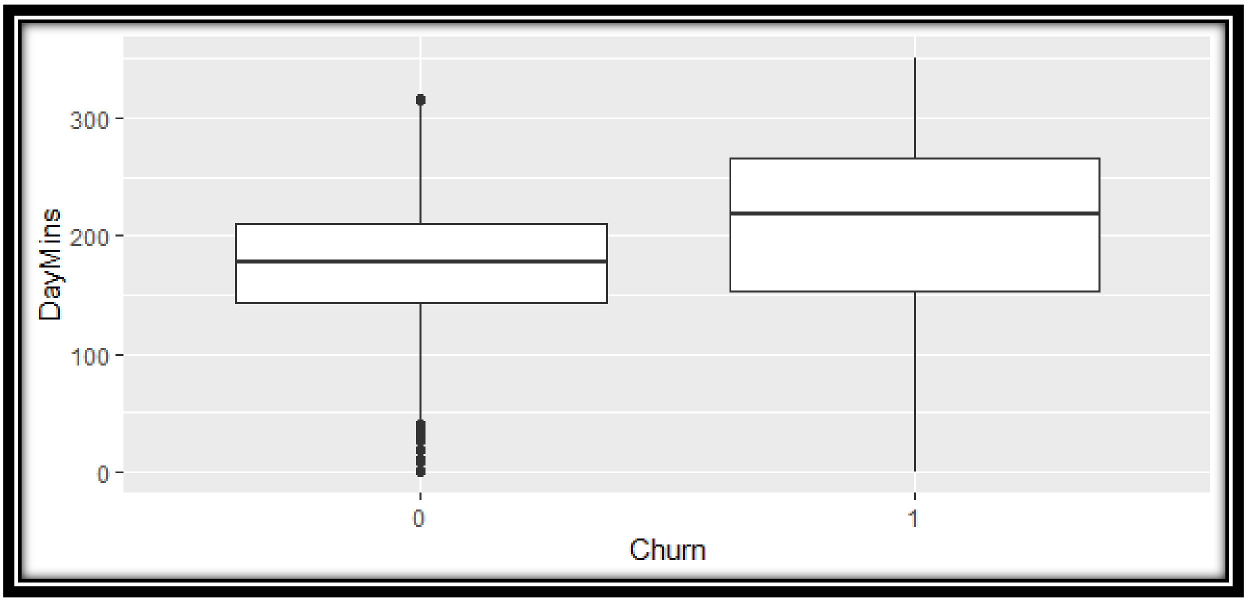


Figure 15: Boxplot of DayMins vs Churn

Observation – Higher the daymins more probability of Churn

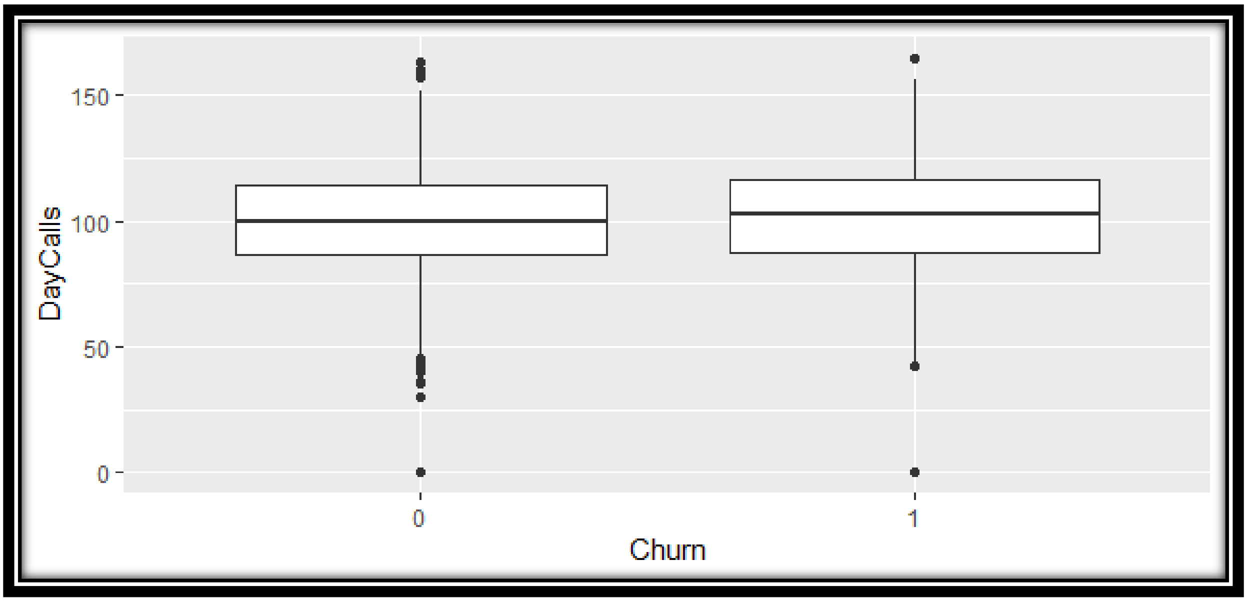


Figure 16: Boxplot of DayCalls vs Churn

Observation – DayCalls seems to be same for both cases.

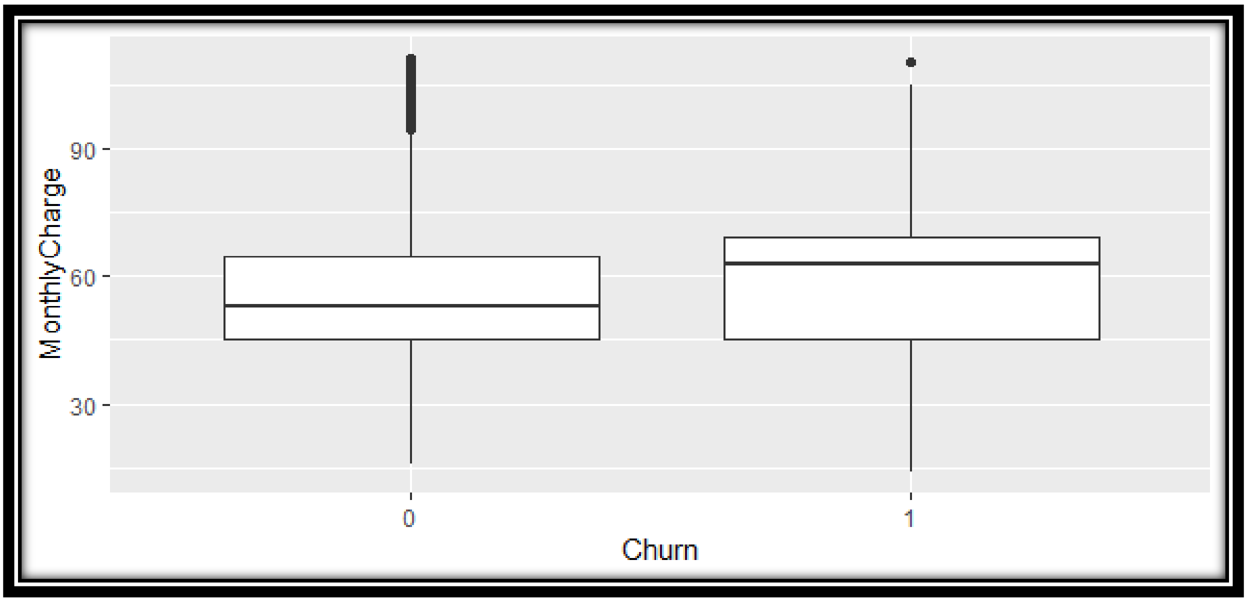


Figure 17: Boxplot of MonthlyCharges vs Churn

Observation – MonthlyCharges seems to be same for both cases.

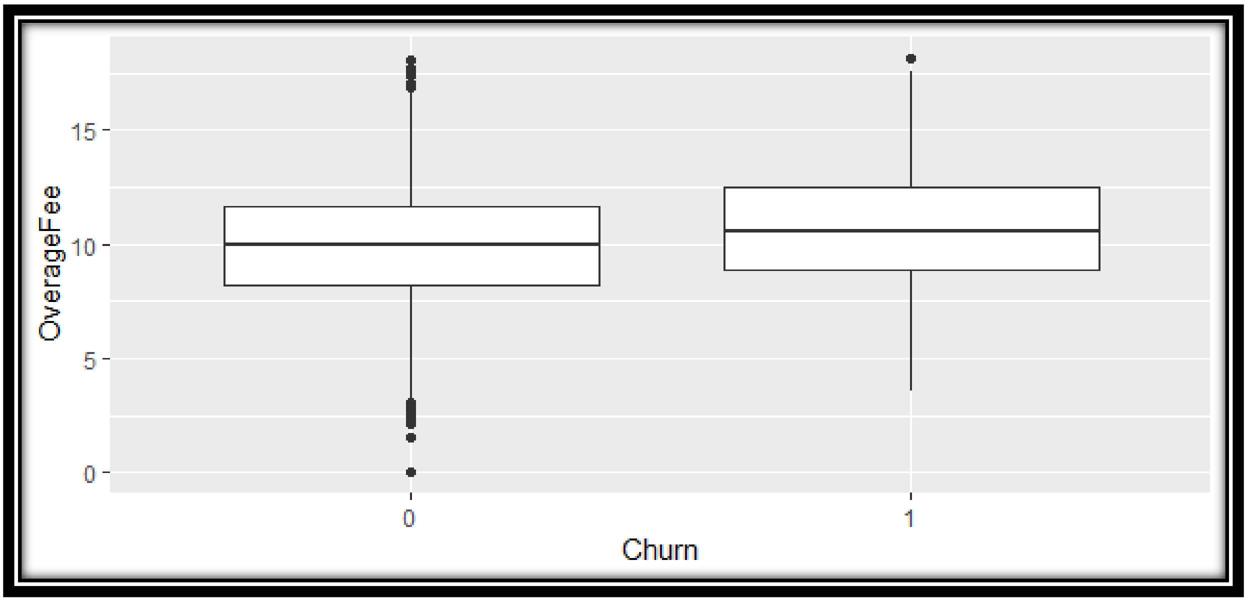


Figure 18: Boxplot of OverageFee vs Churn

Observation – OverageFee seems to be same for both cases.

```
     ContractRenewal    0    1
Churn
0                     186 2664
1                     137  346
> summary(mytable1)
Call: xtabs(formula = ~Churn + ContractRenewal, data = mydata)
Number of cases in table: 3333
Number of factors: 2
Test for independence of all factors:
        Chisq = 225.05, df = 1, p-value = 7.145e-51
```

```
      DataPlan    0    1
Churn
0              2008  842
1               403   80
> summary(mytable2)
Call: xtabs(formula = ~Churn + DataPlan, data = mydata)
Number of cases in table: 3333
Number of factors: 2
Test for independence of all factors:
        Chisq = 34.78, df = 1, p-value = 3.697e-09
```

```
      CustServCalls    0    1    2    3    4    5    6    7    8    9
Churn
0                    605 1059  672  385   90   26    8    4    1    0
1                     92  122   87   44   76   40   14    5    1    2
> summary(mytable3)
Call: xtabs(formula = ~Churn + CustServCalls, data = mydata)
Number of cases in table: 3333
Number of factors: 2
Test for independence of all factors:
        Chisq = 342.7, df = 9, p-value = 2.243e-68
        Chi-squared approximation may be incorrect
```

Chi-square test indicates Contract Renewal, Data Plan and Customer service calls are significant variables.

d. Multicollinearity

To check for multicollinearity, we will plot the correlation plot.

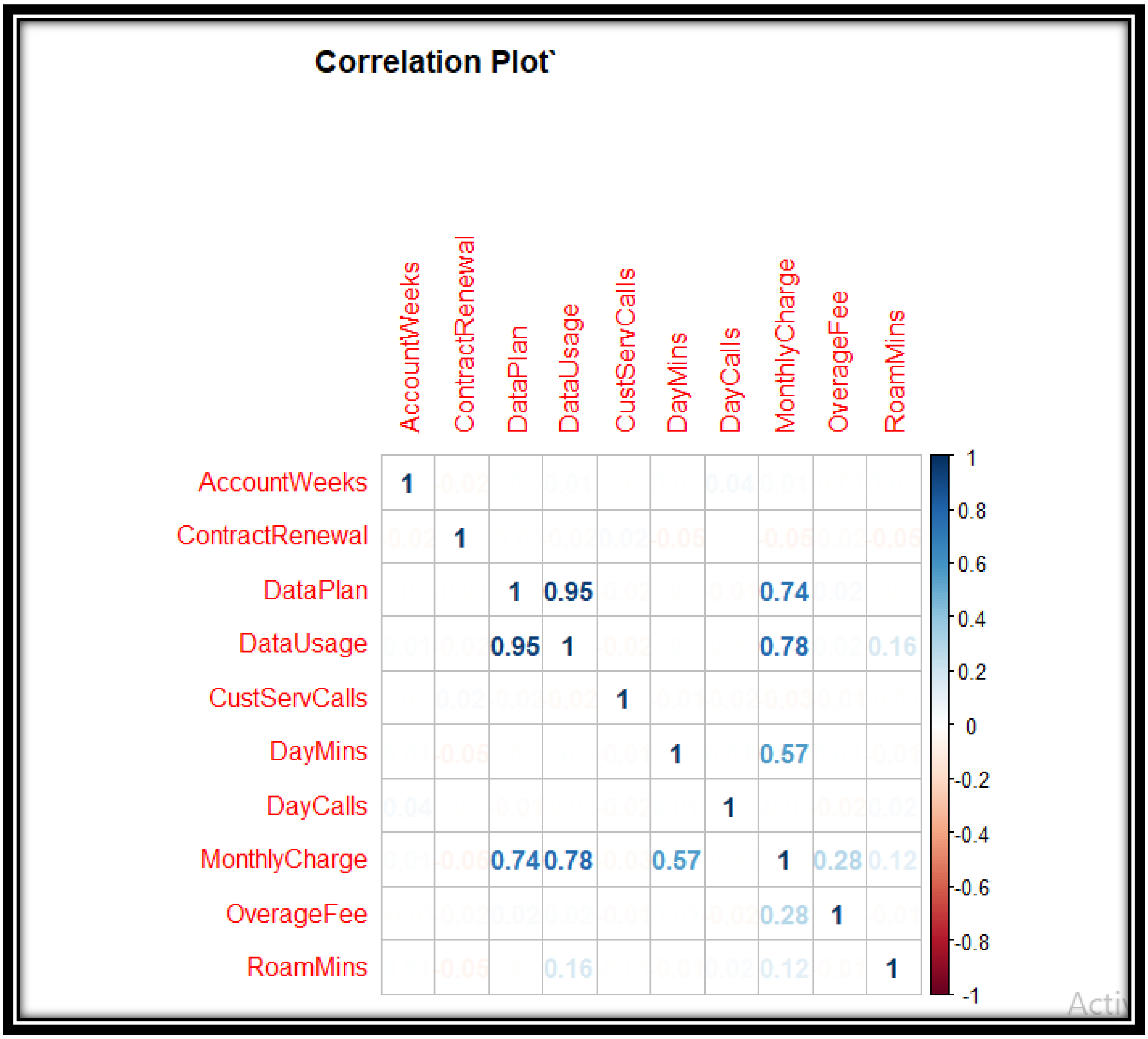


Figure 19: Correlation Plot

We can see that DataUsage and DataPlan are highly correlated. Also MonthlyCharge and DataPlan, DataUsage are also related. DayMins and MonthlyCharge have a slight correlation. We will check the vif values also.

```
 AccountWeeks ContractRenewal        DataPlan       DataUsage   CustServCalls 
     1.003791        1.007216       12.473470     1964.800207        1.001945 
      DayMins        DayCalls   MonthlyCharge      OverageFee        RoamMins 
  1031.490608        1.002935     3243.300555      224.639750        1.346583 
```

Figure 20: Vif values

We will remove the columns with high vif values like MonthlyCharge and DataUsage and that will reduce the vif values.

```
 AccountWeeks ContractRenewal        DataPlan   CustServCalls         DayMins 
     1.002227        1.006143        1.000937        1.001659        1.002862 
     DayCalls      OverageFee        RoamMins 
     1.002901        1.001646        1.002987 
```

Figure 21: Final Vif

5. Split the data

We have divided the dataset into test and train with 30:70 ratio respective.

Train data has 14% Churn ratio

Test data has 14% Churn ratio.

Observation - We can see almost equal representation in both training and testing set for dependent variable.

6. Logistic Regression

First we need to convert Churn into a factor variable before applying logisitic regression. Then form the model using **glm** function and get its summary

```
Call:
glm(formula = Churn ~ ., family = "binomial", data = traindata)

Deviance Residuals:
     Min         1Q     Median         3Q        Max
-2.057590  -0.503084  -0.334602  -0.192577   3.074374

Coefficients:
                   Estimate  Std. Error   z value   Pr(>|z|)
(Intercept)     -6.57631995  0.65224065 -10.08266 < 2.22e-16 ***
AccountWeeks     0.00143972  0.00167907   0.85745    0.39120
ContractRenewal -1.94777875  0.17062509 -11.41555 < 2.22e-16 ***
CustServCalls    0.55380686  0.04804412  11.52705 < 2.22e-16 ***
DayMins          0.01854978  0.00169471  10.94572 < 2.22e-16 ***
DayCalls         0.00282085  0.00330143   0.85443    0.39287
MonthlyCharge   -0.03446080  0.00609973  -5.64957 1.6085e-08 ***
OverageFee       0.22057282  0.03030646   7.27808 3.3860e-13 ***
RoamMins         0.10782491  0.02459255   4.38445 1.1628e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1930.419  on 2332  degrees of freedom
Residual deviance: 1502.109  on 2324  degrees of freedom
AIC: 1520.109

Number of Fisher Scoring iterations: 6
```

Figure 22: Logistic Regression Model

Now we will predict the output and form the confusion matrix.

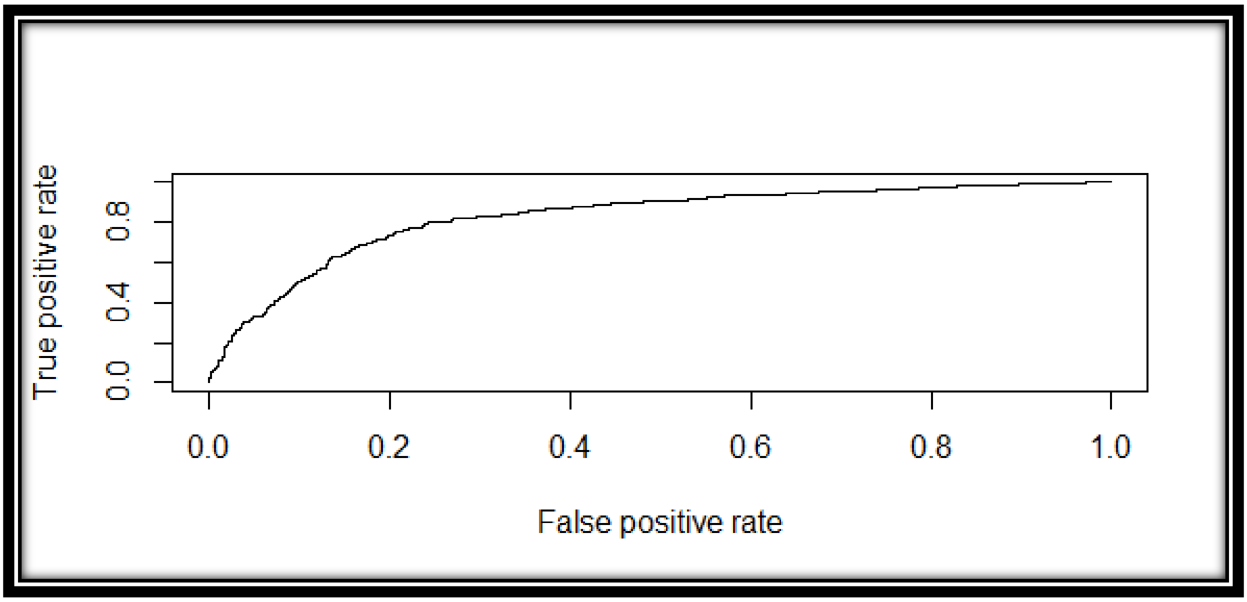


Figure 23: ROC Plot for Logisitic Regression

Area under the curve is around 82.94% and confusion matrix is

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1593   87
         1  402  251

               Accuracy : 0.7904
                 95% CI : (0.7733, 0.8068)
    No Information Rate : 0.8551
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3901

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7426
            Specificity : 0.7985
         Pos Pred Value : 0.3844
         Neg Pred Value : 0.9482
             Prevalence : 0.1449
         Detection Rate : 0.1076
   Detection Prevalence : 0.2799
      Balanced Accuracy : 0.7705

       'Positive' Class : 1
```

Figure 24: Confusion Matrix - Logistic Regression

Accuracy is good, but sensitivity and specificity is ok. Positive Pred Value is also 38% which is not that good. KS score is around 56% and Gini is around 66% which are low. We will try for testing data also.

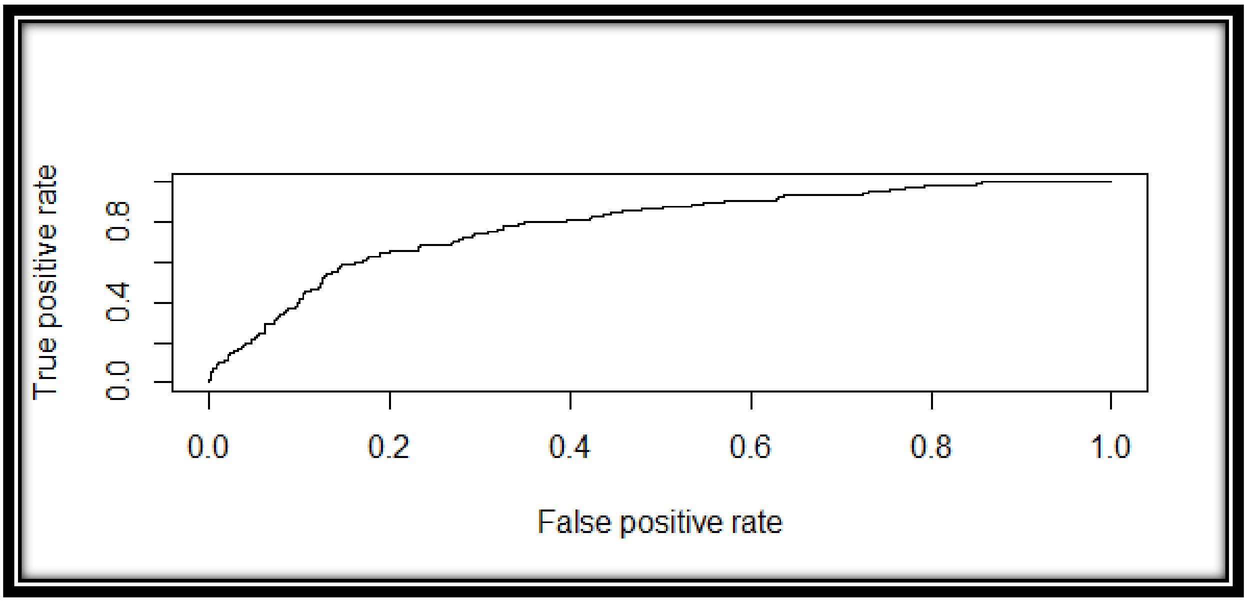
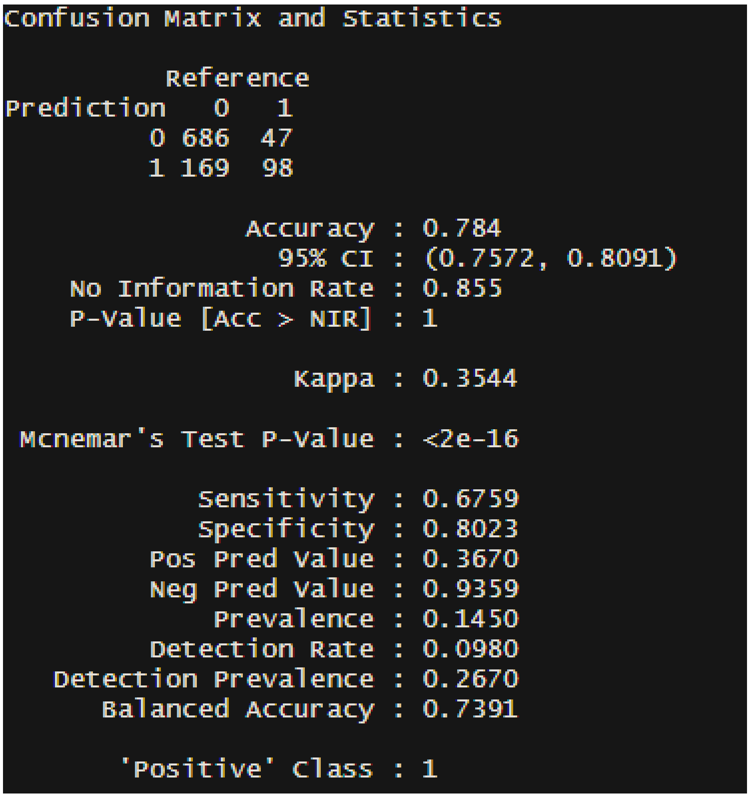


Figure 25: ROC plot for testdata- LR

Area under the curve is around 78.80% which is lesser than the train data.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 686  47
         1 169  98

               Accuracy : 0.784
                 95% CI : (0.7572, 0.8091)
    No Information Rate : 0.855
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3544

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6759
            Specificity : 0.8023
         Pos Pred Value : 0.3670
         Neg Pred Value : 0.9359
             Prevalence : 0.1450
         Detection Rate : 0.0980
   Detection Prevalence : 0.2670
      Balanced Accuracy : 0.7391

       'Positive' Class : 1
```

Figure 26: Confusion Matrix for testdata-LR

Here also Model performs poorly with less sensitivity and Pos Pred value. KS score is 48% and Gini is around 58% which are low.

7. KNN

In KNN, it’s mandatory to scale the data. After scaling the data, we can split it up into test and train data.

We will apply the model to test data first.

After trying various k values, the maximum accuracy is achieved with k=7. Accuracy is around 88.26%.

Confusion matrix is given below.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 793  91
         1  23  64

               Accuracy : 0.8826
                 95% CI : (0.8607, 0.9022)
    No Information Rate : 0.8404
    P-Value [Acc > NIR] : 0.0001164

                  Kappa : 0.4678

 Mcnemar's Test P-Value : 3.494e-10

            Sensitivity : 0.41290
            Specificity : 0.97181
         Pos Pred Value : 0.73563
         Neg Pred Value : 0.89706
             Prevalence : 0.15963
         Detection Rate : 0.06591
   Detection Prevalence : 0.08960
      Balanced Accuracy : 0.69236

       'Positive' Class : 1
```

Figure 27: Confusion Matrix – KNN

Accuracy and Specificity is high for this model but sensitivity is very low. Pos Pre Value is also high.

8. Naïve Bayes

For naïve bayes, we use the same data after converting y variable i.e. Churn into a factor. Here the accuracy comes around 85.79%. Confusion Matrix is given below.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 770  92
         1  46  63

               Accuracy : 0.8579
                 95% CI : (0.8343, 0.8792)
    No Information Rate : 0.8404
    P-Value [Acc > NIR] : 0.0725125

                  Kappa : 0.3979

 Mcnemar's Test P-Value : 0.0001278

            Sensitivity : 0.40645
            Specificity : 0.94363
         Pos Pred Value : 0.57798
         Neg Pred Value : 0.89327
             Prevalence : 0.15963
         Detection Rate : 0.06488
   Detection Prevalence : 0.11226
      Balanced Accuracy : 0.67504

       'Positive' Class : 1
```

Figure 28: Confusion Matrix - Naive Bayes

Accuracy and specificity values are high but sensitivity and pos pred values are low.

9. Comparison of Models

Model	Accuracy	Sensitivity	Specificity
Linear Regression	79	74.26	79.85
KNN	88.26	41.3	97.2
Naïve Bayes	85.6	40.6	94.4

Accuracy is higher for KNN and Naïve Bayes for threshold of 0.5 while for LR it's less for threshold of 0.165. Sensitivity and Specificity is higher for LR model while they are very less for KNN and Naïve Bayes.

If Accuracy is not that important, then I think LR model is the best model among these. But if accuracy is important, then KNN is a better model because its specificity is high hence predicting how many customers are still with the mobile network.

10. Inference

From the summary of LR model we can know that which all variables are significant. We can also use the variable importance function to know the same.

```
> varImp(LR)
                    Overall
AccountWeeks      0.8475843
ContractRenewal1 11.4578678
DataPlan1         5.8850640
CustServCalls    11.5640278
DayMins           9.8472238
DayCalls          0.8645190
OverageFee        5.8367159
RoamMins          3.8144578
```

Figure 29: Variable Importance

The top 3 variables are Contract Renewal, Customer service calls and Day Mins. So customers opting for Contract Renewal with more Day Mins are less likely to cancel the service of the telecom company.

Also Customer Service Calls are a bit indicator regarding their usage and satisfaction and should be considered to understand whether the customer is likely to continue the service or not.

So business can think about offering more discounts and more talk time for customers who are opting for contract renewal which will encourage them to do so. Also more lucrative data plans and less fees for overage might also encourage customers to stay back.

Appendix A

R code is attached along with the report.

Predictive_analysis_ Project_4.R